IRIS: A GOAL-ORIENTED BIG DATA BUSINESS ANALYTICS FRAMEWORK

by

Eunjung Park

APPROVED BY SUPERVISORY COMMITTEE:

Lawrence Chung, Chair

Farokh B. Bastani

Latifur Khan

Zhiqiang Lin

Dedicated to my husband Jungho and my two kids, Sieun and Eunsung.

# IRIS: A GOAL-ORIENTED BIG DATA BUSINESS ANALYTICS FRAMEWORK

by

EUNJUNG PARK, BE, MS

DISSERTATION

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY IN

SOFTWARE ENGINEERING

THE UNIVERSITY OF TEXAS AT DALLAS

May 2017

# ACKNOWLEDGMENTS

"The Sovereign LORD is my strength; he makes my feet like the feet of a deer, he enables me to go on the heights. (Habakkuk 3:19)"

I would like to thank God who guided me to study at UT Dallas and meet many good people. Among the good people, first of all, I am very thankful to Dr. Lawrence Chung who was my supervisor and has taught me with patience. He always said that the starting point of research is problems which are very basic but hard to find, and the final goal of research is to just help our society. Now, I understand the meaning of his statement and will follow his guidance in my future research.

I also thank my committee members, Dr. Farokh B. Bastani who was always kind, Dr. Latifur Khan who deepened my knowledge on big data and Dr. Zhiqiang Lin who made me think more in the perspective of computer science. Additionally, to our lab members, Dr. Tom Hill, Shin Yi Lin, Haanmo Johng, Kirthy Kolluri, Sung Soo Ahn, Seungtack Baek, Ronaldo Pinheiro Goncalves Ju and Sangwoo Moon for their advice and continuous help.

I am always thankful to my husband, Jungho and two kids, Sieun and Eunsung, who were with me on this long journey regardless of whether I was happy or sad. In addition, I thank my dad and mom, father-in-law and mother-in-law, my sisters and brothers who always supported me in the long distance from Korea.

Finally, I give thanks again to God who prepares my future in advance.

March 2017

IRIS: A GOAL-ORIENTED BIG DATA BUSINESS ANALYTICS FRAMEWORK

Eunjung Park, PhD
The University of Texas at Dallas, 2017

Supervising Professor: Dr. Lawrence Chung

Big data analytics is the hottest new practice in Business Analytics today. However, recent industrial surveys find that big data analytics may fail to meet business expectations because of lack of business context and lack of expertise to connect the dots, inaccurate scope and batch-oriented Hadoop system. In this dissertation, we present *IRIS* – a goal-oriented big data analytics framework for better business decisions, which consists of a conceptual model that connects a business side and a big data side, providing context information around the data, an evidence-based evaluation method which enables to focus the most effective solutions, a process on how to use IRIS framework and an assistant tool using Spark, which is a real-time big data analytics platform. In this framework, problems against business goals of the current process and solutions for the future process are explicitly hypothesized in the conceptual model and validated on real big data using big analytics queries. As an empirical study, a shipment decision process is used to show how IRIS can support better business decisions in terms of comprehensive understanding both on business and data analytics, high priority and fast decisions.

Additionally, at the core of Big Data lies data, which is essential for supporting business analytics in gaining insights about business practices towards making better business decisions. The quality

of business analytics inevitably depends on the kinds of individual data and relationships between the data, which should all be defined in a data model. A poor data model can lead to omissions or commissions of important business considerations, likely resulting in bad business decisions. However, there is little work on systematically and rationally developing a big data model for better supporting business analytics, especially in the presence of a ***v****ariety* of sources and types of data that are increasingly becoming available and useful. In this dissertation, we propose three notions of big data model quality – relevance, comprehensiveness and relative priorities with a goal-oriented approach to building such qualities in a big data model. In this goal-oriented approach, alternatives in big data models are explored and selected for validating potential problems and solutions, while also achieving business goals. An empirical study has been conducted on the shipping decision process of a world-wide retail chain, to gain an initial understanding of the applicability of this approach.

Finally, many software systems are being developed to help with business processes, which typically involve a number of (human) tasks in achieving organizational goals. However, aligning a software system well with its intended business process has been challenging, since the tasks in a business process usually lack formal definitions and can be performed via multiple different allocations of resources. In this dissertation, we propose a goal-oriented transformational approach to deriving use cases, as requirements on the software system, from a business process which is modeled in BPMN (Business Process Model and Notation). In this approach, a business process is modeled not only in terms of the functionally-oriented BPMN but also non-functional business goals, and the target software requirements are also modeled in terms of functionally-oriented use cases together with non-functional requirements. Those tasks to be performed by a software system

are transformed into use cases, in consideration of multiple alternative interpretations of business tasks, different allocations of software functionality and the granularity of the target requirements guided via similarity and granularity. Additionally, an intermediate model is utilized in the 2-step transformation process to deal with the ontological gap and the many-to-many relationships between the source and the target. This process is facilitated by context-aware transformation rules and a supporting tool. A study of a quote flow business process shows that our goal-oriented transformational approach helps produce more cohesive, correct and comprehensive use cases.

TABLE OF CONTENTS

LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1

# INTRODUCTION

## 1.1 Motivation

Big data analytics is a technology which helps turn hidden insights in big data into business value by using advanced analytics techniques in order to support better business decisions, as the hottest new practice in Business Analytics today. According to some surveys ([1, 2]), about 80% of CEOs or executive teams view that big data analytics initiatives have the potential to drive business value such as creating new revenue streams, improving operational efficiency or cutting cost, and over 80% of the participant organizations do ongoing projects. However, according to another survey [3], 55% of big data projects do not get completed, and many others fall short of their business objectives.

The key reasons that big data analytics projects fail can be summarized as the followings by reviewing surveys (e.g., [1, 2, 3]) and several articles (e.g., [4, 5]). 1) Lack of business context around the data and lack of expertise to connect the dots, which hinder comprehensive understanding of business and analytics results for right decisions. 2) Inaccurate scope with business silo, which means that big data analytics are done regardless of business objectives under rare communications between departments in an organization. 3) Batch-oriented Hadoop system which is inadequate for real-time data processing to support fast business decision-making. 4) Hard to derive actionable business insights or requirements from data.

Additionally, the quality of big data analytics can only be as good, or as bad, as the quality of the big data it uses. The quality attributes of big data, together with relationships between data, should therefore be defined in a big data model. With a good quality big data model, big data analytics

can accurately identify important business concerns, trend opportunities, and useful business insights, which, in turn, can lead to good business decisions. Yet, no guidelines are available for how to develop a high-quality big data model in a systematic and rational way.

Finally, with the advances in Information Technology (IT), use of software systems increasingly has become prevalent and critical in more efficiently and effectively carrying out the various tasks and activities in the fast changing business domain. However, aligning a software system well with its intended business process has been challenging, due to the difficulty in firstly understanding and precisely modeling a business process and secondly coming up with a requirements specification of a system for supporting the business process. Use of well-known notations helps: BPMN (Business Process Model and Notation) for precise modeling of a business process, and UML use case for modeling software requirements; but the second difficulty still remains. There are several issues: 1) Since there do not exist formal definitions of the models, different interpretations are possible which can lead to transformations that deviate from the original meaning; 2) A business activity can be performed either by people or system functionality; 3) The granularity of a use case is not necessarily the same as that of a business activity or task; and 4) Neither BPMN nor UML use cases consider Non-Functional Requirements (NFRs).

## 1.2 Solution Overview

To address the problems, we propose IRIS[1] – a novel big data analytics framework using Spark in a goal-oriented approach for better business decisions, i.e., supporting comprehensive understanding of business and data, high priority and fastness. This framework provides a

[1]IRIS, in Greek mythology, is a goddess who symbolizes a bridge between heaven and earth – in our adaptation, "connecting the dots".

conceptual model for big data analytics which connects the dots between important concepts of business side such as business goals, problems, solutions, business processes and those of a big data side such as big analytics and big queries to bridge the gap between the two. It helps not only in comprehensive understandings of current and future business, but also communications between stakeholders by helping analyzers explicitly model the concepts. Moreover, in the spirit of goal-orientation, alternatives in problems with the as-is business process and solutions for the to-be business process can be hypothesized in the conceptual model and validated by analyzing big data, and significant ones are selected after trade-off analysis through our evidence-based evaluation method. This process helps analyzers focus on the most effective solutions within the given time and budget. To support our concepts, we implemented IRIS assistant tool which integrates Eclipse Modeling Framework (EMF) for the conceptual model and Apache Spark for a big data analytics which enables real-time processing in a distributed and clustered computing platform, leading to fast business decisions.

Moreover, we suggest how to model big data for supporting business analytics. In detail, this dissertation proposes a systematic approach to big data modeling. The approach specifically tackles three aspects of big data model: *relevance, comprehensiveness* and *relative priorities.* These aspects stipulate that the data and relationships between the data should be *relevant* to their use, *comprehensive* enough to cover business decision-making and *prioritized* so that their importance is clear. In order to attain these three data model qualities, this dissertation then proposes a goal-oriented approach that helps *build such qualities into* a big data model. Our ultimate goal is to provide this approach as a service over the Internet so that it can be used to support big data analytics.

Finally, we propose how to derive requirements from the to-be processes by goal-oriented big data business analytics. More specifically, we present GO-BP2UC, a goal-oriented framework for transforming software-allocated elements of business processes in BPMN with NFRs into Use Cases with NFRs. GO-BP2UC is intended to facilitate the exploration of alternatives for dealing with multiple interpretations of the same elements in a business process, and selection among the alternatives through a trade-off analysis, using similarity and granularity as the selection criteria, in spirit of other general goal-oriented approaches (e.g., [6, 7, 8, 9]). Similarity is further refined into Ontological- and Contextual- Similarity. Ontological Similarity means how well concepts in a source model are matched to concepts in a target model, which can be evaluated by using a taxonomy of ontologies. Contextual Similarity means how well the contextual information – allocation and granularity of a concept in a source model – is reflected into a target model, which can be automatically transformed by rules with conditions. Granularity means the size of a unit of target element including single or clustered ones. As for the BPMN and Use Case Models, we augment them with NFRs. Also, an Intermediate Model (IM) is used, which is composed of intermediate entities and relationships, for helping to deal with the ontological gap and the many-to-many relationships between the source (i.e., BPMN augmented with NFR) and the target (i.e., Use Case Model augmented with NFR). To support the transformation process, an assistant tool is implemented as a proof of concept.

### 1.3 Contribution

First of all, with respect to the IRIS: goal-oriented big data business analytics framework, the main contributions are four which assist to turn analytics from big data into business value: 1) a conceptual model connecting a business and a big data side in a goal-oriented approach for big

data analytics; 2) an evidence-based evaluation method for high priority; 3) a process on how to use IRIS; and 4) a real-time-based supporting tool for fastness.

Second, as for IRIS-BigDM (Big Data Modeling): a goal-oriented big data modeling for business analytics, our suggestion helps relevant, comprehensive and highly important big data model.

Finally, when it comes to IRIS-BP2UC: Deriving Requirements from business processes in a goal-oriented transformational approach, analyzers can elicit quality of software system requirements in terms of cohesiveness, correctness, comprehensiveness which are aligned business goals.

## 1.4 Dissertation Outline

In Chapter 2, Related work will be described and Chapters 3 to 5 present the core of the IRIS Framework, i.e., IRIS: A Goal-oriented big data business analytics Framework, IRIS-BigDM: A Goal-oriented big data modeling for business analytics and IRIS-BP2UC: Deriving Requirements from business processes in a goal-oriented transformational approach, respectively. In Chapters 6 and 7, IRIS assistant tool implementation and IRIS framework application in two empirical studies will be discussed. In Chapter 8, we will evaluate our solution with other previous work. In Chapter 9, we will conclude this dissertation with future work.

# CHAPTER 2

# RELATED WORK

## 2.1 Goal-Oriented Requirements Engineering (GORE)

As Goal-Oriented Requirements Engineering (GORE) is a main stream among requirements engineering communities, according to [10], it is useful to elicit requirements by the repeated asking of “why”, “how” and “how else” questions, to relate requirements to organizational and business context and to deal with conflicts.

**NFR Framework for Representing Non-Functional Requirements Objectives**

Concerning the representation for both FRs and NFRs, there are several goal-oriented frameworks, including KAOS [6], the NFR Framework [7] and i* [8], each with its own emphasis and characteristics. NFR framework considers how non-functional requirements such as security or usability can be dealt using softgoal notion which has no clear-cut criteria whether it is achieved or not. In NFR framework, softgoals are expressed in the form of Type [Topic]. While Type is a non-functional part, Topic is a functional part. There are three kinds of Softgoals, i.e., NFRSoftgoal for non-functional requirements, Operationalizing Softogoal for a concrete mechanism and Claim Softgoal for a justification. These Softgoals are further refined using Satisficing relationship, good enough satisfy, toward another Softgoal such as fully or partially positively (MAKE or HELP), or fully or partially negatively (BREAK or HURT) and Decomposition relationships such as AND or OR. To check the final goal achievement, leaf goals are checked if it is satisficed or not, and it can be evaluated by a bottom-up label propagation mechanism, with a label – Satisficed, Weakly Satisficed, Weakly Denied, Denied, Conflict, or

Undetermined. As Figure 2.1 shows, the Softgoals can form a Softgoal Interdependency Graph (SIG) with Satisficing or Decomposition relationships.

For example, as Figure 2.1 shows, G1: Security [Campus Navigation] and G2: Usability [Campus Navigation] are NFR softgoals of disabled person and it can be further decomposed into sub-softgoals, i.e., G1.1: Confidentiality [Campus Navigation] for G1 with eql (equal), and G2.1: Familiarity [Campus Navigation] and G2.2: Easy to Use [Campus Navigation] for G2 with an AND relationship. To achieve the NFR softgoals, operationalizing softgoals can be identified. OG1.1.1: ID/Password Authentication, OG1.1.2: Locally Keep Search History, OG2.1:Google Map UI or OG2.2:One Functionality per One UI Page are candidate operationalizing softgoals. While OG1.1.1: ID/Password Authentication positively affects the Confidentiality [Campus Navigation] (MAKE contribution) and at the same time it has negative contribution to G2.2: Easy to Use [Campus Navigation], G1 positively affect the both NFR Softgoals. Thus, G1 is selected as a solution. OG2.1: Google Map UI and OG2.2: One Functionality per One Page positively affect the G2.1: Familiarity [Campus Navigation] and G2.2: Easy to Use [Campus Navigation] with MAKE contributions. Thus, G2.1 and G2.2 are selected as solutions.

The G1:Security and G2:Usability NFR Softgoals are satisficed by using label propagation mechanism which evaluate from bottom to top.

**PIG for Representing Problems**

A problem is a phenomenon from which stakeholders are suffering to achieve a goal. There are several different kinds of models for problem representation and root cause analysis, including notably FTA (Fault Tree Analysis) [11], Fish Bone Diagram [12], and PIG (Problem Interdependency Graph) [13]. While FTA is suitable when information is available about AND/OR

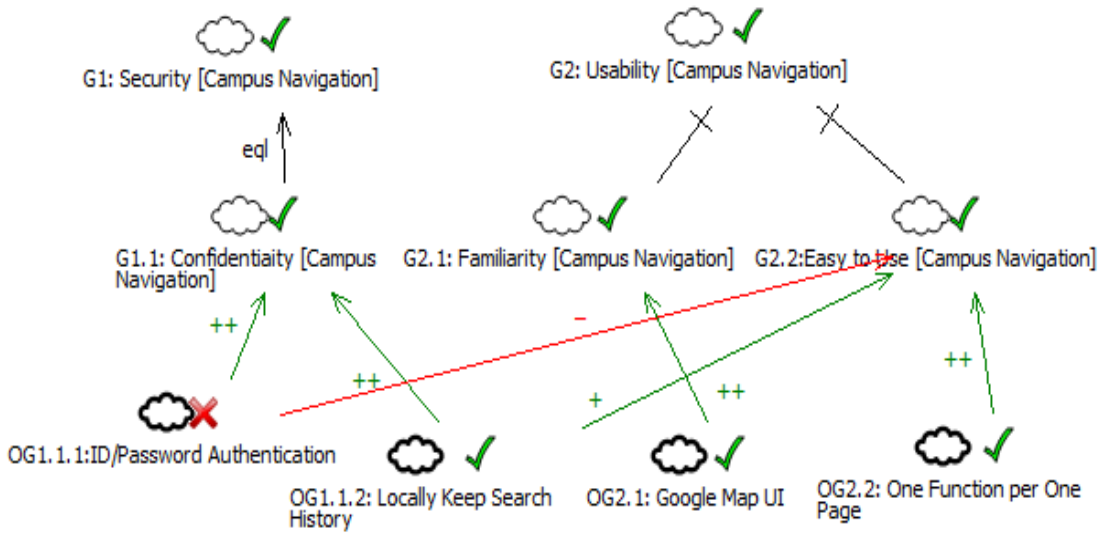


Figure 2.1. An example of Softgoal Interdependency Graph (SIG)

logical relationships among root causes, Fish Bone Diagram is adequate when uncertainties exist about relationships among root causes. We adopt PIG, since it accommodates both conventions, and additionally it closely resembles SIG, offering richer ontology such as NFR Softproblem and Operationalizing Softproblem which do not have clear-cut definitions or criteria, together with the same kinds of satisficing relationships (MAKE, HELP, HURT and BREAK). This ontology enables requirements engineers to make richer reasoning using the label propagation procedure.

For example, in Figure 2.2, similar to SIG, P1: Ineffective [Campus Navigation] that disabled person is suffering from is a Softproblem and it can be further decomposed into P1.1: Inaccurate [Campus Navigation] and P1.2: Inefficient [Campus Navigation]. Based on the description of the running example for campus navigation system, the root causes can be identified such as OP1.1.1: Insufficient indoor information, OP1.1.2: Incorrect verbal direction, OP1.1.3: Complicated campus layout, OP1.1.4: Unfamiliar campus layout and OP1.1.5: Difficulty of Move impairments. While both NFR Framework and PIG can express not only certain relationship such as logical

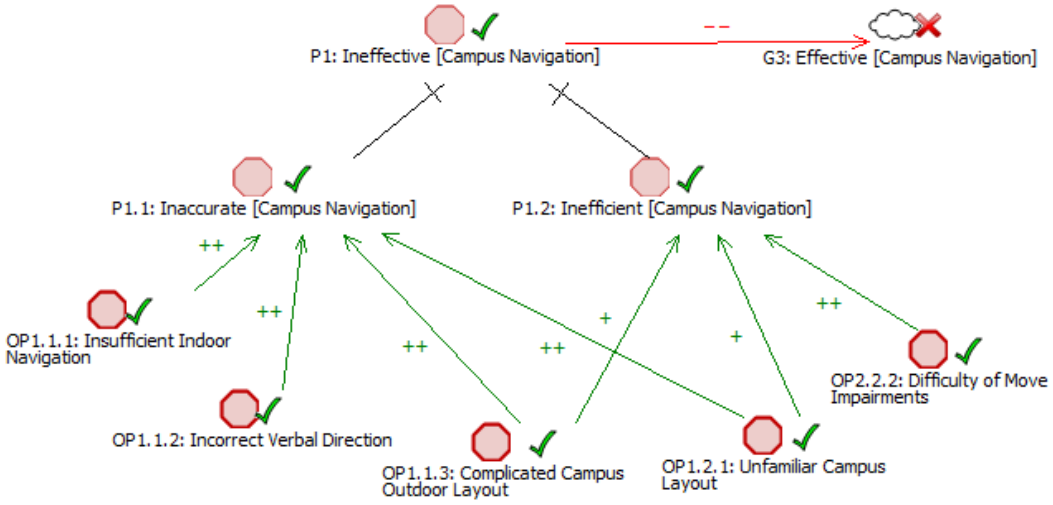


Figure 2.2. An example of Problem Interdependency Graph (PIG)

AND or OR relationships, but also uncertain relationship such as satisficing relationships, i.e., MAKE, HELP, HURT and BREAT, it is hard for them to express sequential relationships such as sequence flows. Moreover, since they are conceptual model, they have a possibility of the mismatch between a model and the reality that actual data reflects.

## 2.2 Big Data and Analytics

**Big data** is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them. According to [14], it can be characterized in terms of 4Vs, i.e., Volume, Velocity, Variety and Voracity, and each characteristic has several challenges to address.

**Volume** refers to a large amount of data and the most challenging aspect is Scalability to handle a growing number of data. Hadoop platform including HDFS and MapReduce or NoSQL database which supports distributed computing can address this issue. **Velocity** describes the increasing rate at which data flows into an organization. The main challenge is to deal with streaming data

processing. Spark which supports real-time processing for the streaming data can solve the problem. **Variety** means various degrees of structure with in the source data such as structured, unstructured and semi-structured data, text and multimedia, which needs effective mechanism for linking diverse data in inner structure. **Veracity** is uncertain or imprecise data because of unstructured data. As a result, data quality, data cleansing, master data management and data government remain critical disciplines when working with Big Data.

**Big Data Storage**

- **HDFS (Hadoop Distributed File System)**

  For large size of data, different kinds of file systems are needed, and HDFS [15] is one type of file system to deal with big data storage. Unlike traditional file system which has 512 bytes, block size is large, default 128MB, and flexible, thus it can deal with scalability of data. It consists of name node (the master) and data node (workers) which are for managing filesystem namespace, and storing and retrieving data blocks respectively. For addressing the problem of a distributed file system, data loss due to Single Point Of Failure (SPOF), HDFS has the mechanism for high availability such as data replication, QJM (Quorum Journal Manager) and failover controller. HDFS is a part of Hadoop for big data processing.

- **NoSQL Database**

  A NoSQL database [16] which originally refers to “non SQL” or “not only SQL” provides a mechanism for storage and retrieval of data other than the tabular relations used in relational databases. This kind of databases are suitable for handling huge amounts of data with schema-free, providing easy replication and eventual consistent.

  - **Cassandra**

Apache Cassandra [17] as a column-oriented database is a distributed database management system designed to handle large amount of data across many commodity servers, providing high availability with no single point of failure and scalability.

- **MongoDB**

  MongoDB [18] is a cross-platform document-oriented database program which uses JSON-like documents with schemas. It stores documents in collections which are similar to tables in relational databases, but don't require the same schema.

**Big Data Analytics Framework**

Big data analytics computing frameworks such as Hadoop process large scale data with a parallel and distributed algorithms on commodity clusters. MapReduce and Spark are main streams in industry. While MapReduce is a batch-oriented processing system in which intermediate (shuffle) files for sorting are stored on disks during the processing, Spark is a real-time processing system centered on RDD (Resilient Distributed Dataset) in which Shuffle files are stored in memory, thus, it is faster than MapReduce [19].

- **MapReduce**

  MapReduce is a programming model for big data processing with a parallel, distributed algorithm on a cluster. It consists of a Map part which generates key value pairs and sorts them, and a Reduce part which performs a summary operation on the sorted key value pairs. While doing MapReduce work, intermediate shuffle files are stored on disk which leads to more disk I/O, hence long processing time.

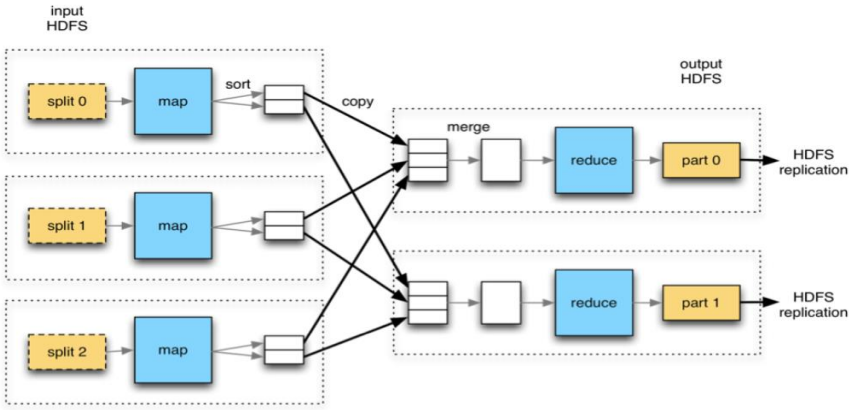


Figure 2.3. MapReduce architecture on HDFS

- **Spark**

  Spark provides an application programming interface centered on RDD which is an abstract data structure and a read-only multiset of data distributed over a cluster of machines. It has diverse libraries for processing big data such as Spark SQL, Spark Streaming, MLlib, GraphX and Packages which can be implemented with Scala, Java, Python or R. Moreover, Spark can connect not only SQL database, but also NoSQL database. While MapReduce stores intermediate results on disks, Spark store them in memory which results in fast processing.

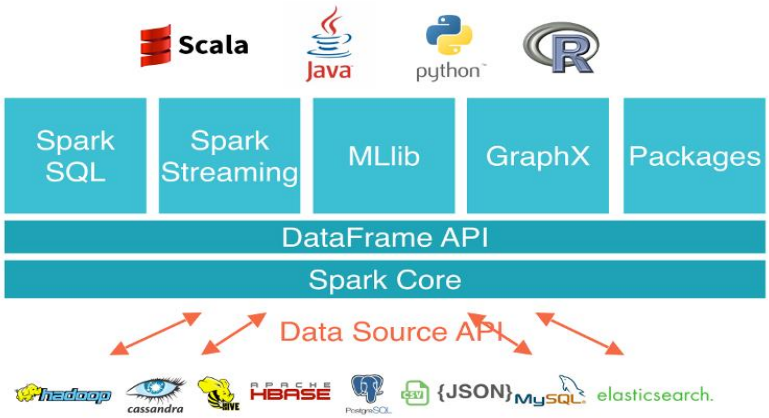


Figure 2.4. Spark architecture

### Data Mining and Machine Learning

Data mining [21] is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources are heterogeneous which can be databases, data warehouses or web systems. For mining patterns, diverse machine learning algorithms and statistics are used such as characterization and discrimination, classification and regression, clustering analysis or outlier analysis.

Big data and data mining have strengths that they provide hidden insights from huge amount of data, but they rarely consider business processes and hard to analyze impact on business goals. Hence, to create **Value** which refers to transformation from big data to business value as another characteristics of big data, big data or data mining themselves are not enough.

## 2.3 Business Processes

A Business Process is a collection of activities to create more business values which a means to achieve business goals. There are many kinds of business process models such as BPMN (Business Process Model and Notation) [22], Workflow, Activity diagram in UML, Petri-net, IDEF0, etc. Among them, BPMN is a standard notation to represent a business process as well as data objects such as inputs or outputs of an activity.

### BPMN

BPMN's basic element categories are flow objects, connecting objects, swim lanes and artifacts. Flow objects such as events, activities which describe the kind of work which must be done, gateways which represent conditions. As connecting objects, sequence flow which shows orders of activities, message flow representing flow across organizational boundaries, association that is used to associate artifact to a flow object. Swim lanes such as Pool and Lane are used to organize

and categorize activities. Artifacts give more information such as data objects which represent required data in an activity, group which is used to make a group of different activities, and Annotation. Figure 2.5 shows a sales demand forecasting process diagram represented by BPMN. In Figure 2.5, three participant, i.e., Forecasting Group, Sales Group, Marketing Group are working together to forecast sales demand. This process starts in the Forecasting group with Gather Demand History Data (Task). After the Forecasting Group gathers the data, Demand History will come out as an output. The data will become an input of Forecast Demand. After forecasting sales demand, the result will be sent to Sales Group who reviews forecast. After this, changes from Sales Group will be incorporated. Again, the updated demand forecast data will be sent to Marketing Group, then after review, marketing changes are incorporated.

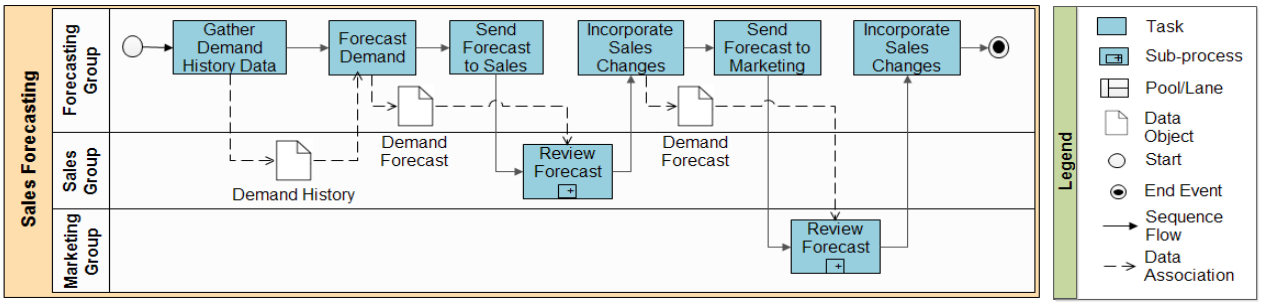


Figure 2.5. An example of sales demand forecasting process with BPMN

## 2.4 Goal + Business Analytics/ Business Intelligence

### BIM (Business Intelligence Model)

In the perspective of applying a goal-orientation approach to business analytics, BIM (Business Intelligence Model) [23] is similar to our approach, but ours goes beyond. BIM which focuses more on business strategy level cannot explicitly model problems and detailed business processes. Moreover, it does not consider big data which be characterized as 5V–Volume, Velocity, Variety, Veracity and Value, and it supports only descriptive analytics. On the other hand, our proposal

emphasizes on business process level and applied advanced analytics techniques to big data which enables not only descriptive, but also predictive analytics. For an evaluation method, in BIM, indicators which reflect business performance data are propagated only in the entity, but in our proposal, diverse analytics results are used as claims for a goal achievement which can be applied to entity and relationships. Figure 2.6 shows an example of BIM.

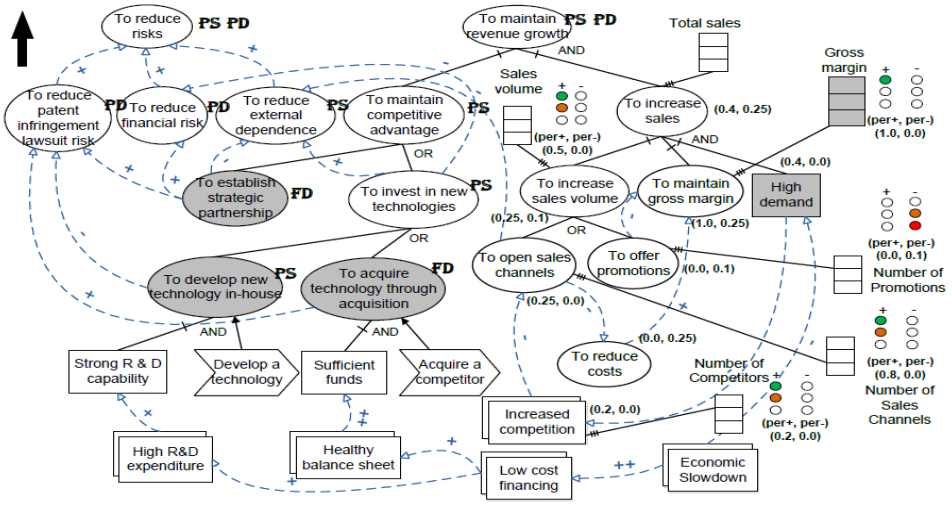


Figure 2.6. An example of Business Intelligence Model (BIM)

**FBCM (Fact Based Collaboration Modeling)**

FBCM (Fact-Based Collaboration Modeling) [24] which provides a method to align software requirements to business strategy by using BSC (Balanced Score Card) is similar to IRIS in the view of business goal-awareness and KPI utilization. However, while FBCM does not explicitly model problems and root causes, ours can do, which enables it to focus on critical solutions. Moreover, FBCM does not provide how much a business element contributes to other elements and how to evaluate business goal achievement, but IRIS has diverse positive and negative satisficing relationships between modeling elements, and an evidence-based reasoning method for evaluation of a business goal. Figure 2.7 shows an example of FBCM.

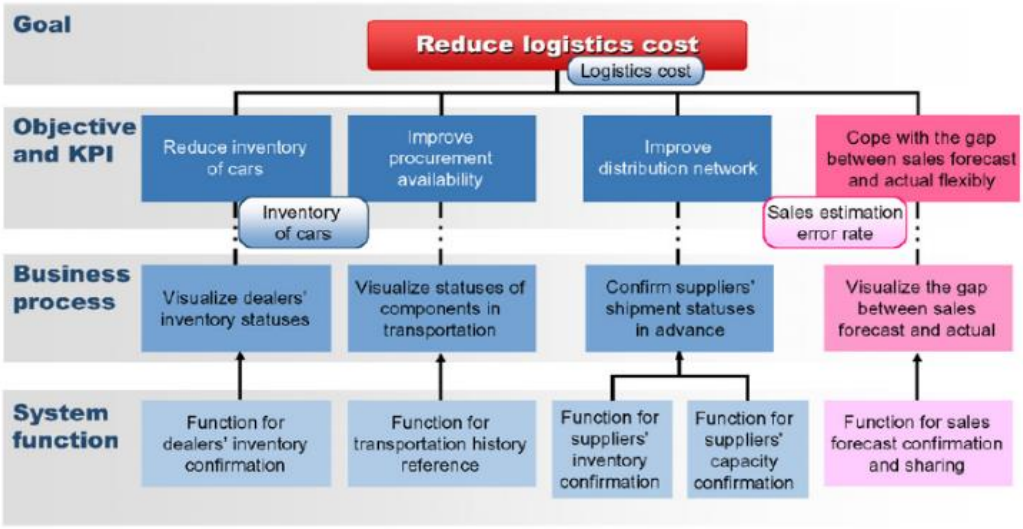


Figure 2.7. A goal analysis tree from FBCM

**GOMA (Goal-Oriented big data Modeling Approach)**

GOMA [25] also applied a goal-oriented approach to big data modeling. It suggests phenomenon modeling to represent insights gained from big data for evaluating business goal achievement and validating business problems, which adopts a temporal logic. It also models business goals and problems, but we extend it. IRIS considers business processes as a mediator between business and big data to provide business context. Moreover, all modeling concepts such as goal, problems, solutions, priority and relationships are validated by big data analytics or big queries. Additionally, our suggestion provides a GO-BigBAM which is a conceptual modeling language to connect big data with business and an assistant tool to analyze big data including descriptive and predictive analytics by using Spark. Figure 2.8 shows a portion of phenomenon model for the bank churning example. Figure 2.9 shows Phenomenon P1 and P2 to validate CRR > 95% is not achieved.

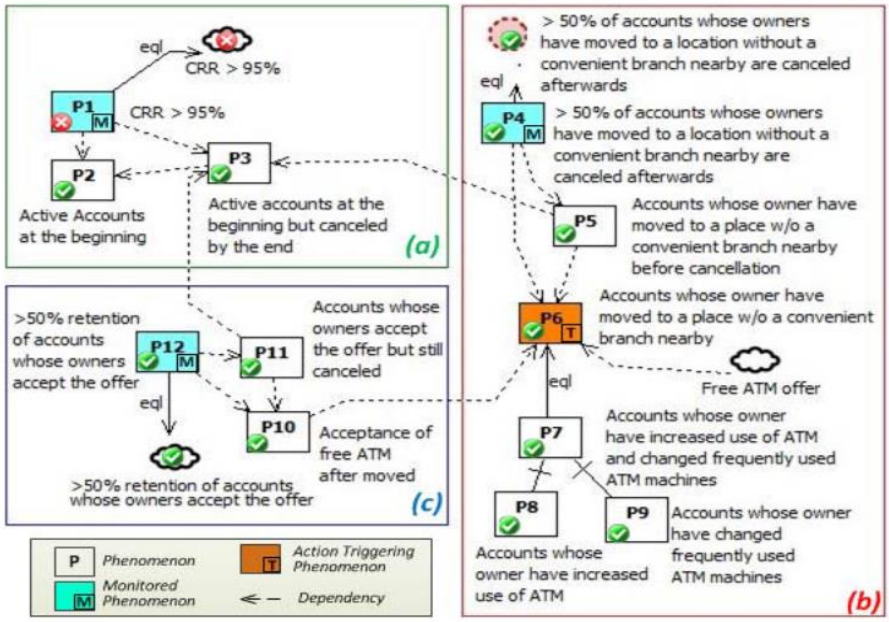


Figure 2.8. A portion of the Phenomenon Model for the bank churning example

Phenomenon P1

```
Description:
  Customer retention is higher than > 95%.
Input:
  P2, P3
Output:
  CRR(P3,P2)
Validation:
  CRR(P3,P2) > 0.95
Function:
  CRR({canceled},{baseline}):
    1 - count(canceled)/count(baseline)
```

Phenomenon P2

```
Description:
  Active accounts at the beginning of a reporting period.
Input:
  { Account.id as acct_id,
  interval(Account.open_date,Account.close_date) as A },
  interval(ReportingPeriod.begin_date,
    ReportingPeriod.end_date) as R
Output:
  { acct_id, A }
Selection:
  not (R older A) and not (R after A)
Validation:
  count(Output) > 0
```

Figure 2.9. Example of phenomenon model

**URN-base BPM**

While URN-base BPM [26] focuses more on KPI modeling and evaluation, ours focus more on explicitly modeling insights on problems and solution for business process reengineering by using a big data processing which supports a distributed parallel computing. Moreover, not only entities, but also relationships of modeling elements can be validated by big data. Additionally, our model uses Type [Topic] to express problems and solutions which enables more precise location of problems and solutions along with business goal layer, business process layer, big data analytics layer of concepts. We integrate BPMN for business process modeling and whole-part of business processes enables to do root cause analysis. We also use interactive queries for business performance measurement.

**BAF (Business Analytics Framework)**

Business Analytics Framework (BAF) [27] is a conceptual modeling framework for business analytics. It utilizes business analytics to answers for business questions and provides three different kinds of views: Business View, Analytics Design View, Data Preparation View. Using these views, it helps to find requirements analytics of data analytics system. Figure 2.10 shows BAF.

The big differences between BAF and ours are that our framework supports distributed parallel computing process on big data, can show alignment relationships from business goals – business processes – big data analytics. Additionally, in our framework, insights means deep understanding for solving problems or doing solutions in business processes to achieve business goals. Big data analytics validates not only problems or solutions, but also elements of alignment views including entities and relationships.

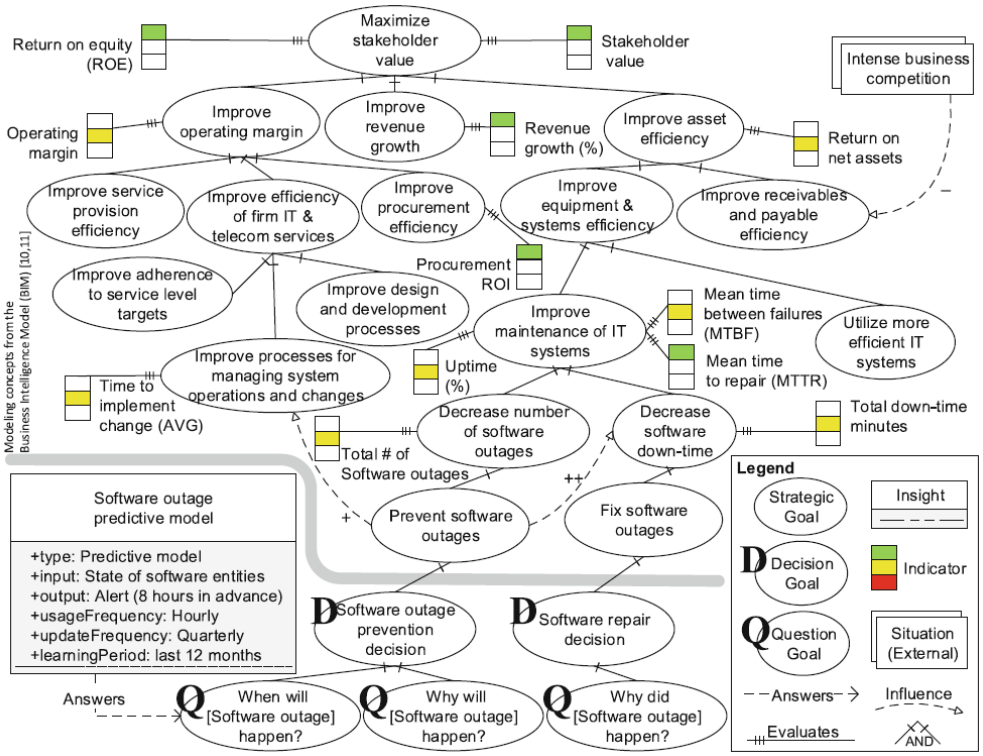


Figure 2.10. Business view for business analytics framework

## 2.5 Goal + Business Process

In the perspective of the combination of goal and business process, i* [8] applied goal model to business process reengineering (organizational modeling). [28] enriches goal models for business processes, then makes a diverse process specifications, among them select one, then transform BPEL. While they focus more on modeling perspective and evaluation in which business processes are aligned with business goals, ours consider both modeling and data part, especially, using big data.

## 2.6 Business Process + Data

Process Mining [29] and Process Analytics [30] utilize process event log data and modeled process and real process confirm. Using these technologies, analyzers can find business process bottle neck

or structural problems of business process and help enhance business processes. The most important difference is that they do not consider business goals and alternatives, i.e., goal-orientation. Thus, it is hard to diagnose how business processes are aligned with business goals. Moreover, they also deal with big data, but they do not use a distributed parallel processing. Ours uses distributed parallel processing which provides extensibility and high availability.

## 2.7 Business Process Reengineering

As for the business process and reengineering, Business Process Management (BPM) [31], BPMN [22] and Business Process Reengineering (BPR) [32] offer diverse technologies and tools. Many of them seem to be mostly about business processes and a variety of KPIs as their scope. We adopt key ideas from these business process-related areas, and extend them.

# CHAPTER 3

# IRIS: A GOAL-ORIENTED BIG DATA BUSINESS ANALYTICS FRAMEWORK

In this chapter, we will describe how important concepts such as business goals, business processes, business problems and solutions and big data are connected with each other to support better business decisions in terms of comprehensive understanding, high priority and fastness to achieve business goals.

## 3.1 A Running Example: A Shipment Decision

To illustrate the key concepts of goal-oriented big data analytics framework, as well as for the later empirical study, a real process of a worldwide fashion retailer on shipping decision is used [33]. The company, we say ABC, has 2,000 stores across 88 countries and also online stores, offering considerably more products than similar companies – about 11,000 distinct items annually compared with 2,000 to 4,000 items for its key competitors, through shipments of items – twice weekly – from the warehouses to its world-wide stores. The volume of data to be collected is huge. As shown in Figure 3.1, for its global distribution, ABC's headquarters sends a weekly offer to each store, with a maximum quantity the store can request for each of the items. Using this data, together with some other data, such as its sales history and local inventory, the manager of the store manually decides the weekly shipment quantity and sends a shipment request to the global warehouse team. This team aggregates all the requests that come from all over the world and reconciles shipment quantities, if there is not enough inventory to fulfill all the requests. The team decides on a shipment schedule, using their previous experience, and ships items according to the

shipment schedule. When a store receives the requested items from the global warehouse team, its store manager displays them according to the company's policy on displays.

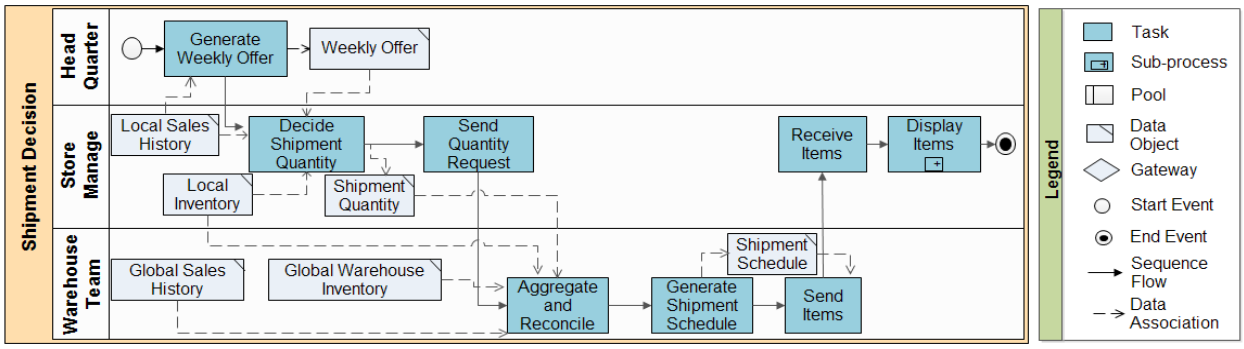


Figure 3.1. AS-IS business process of shipping decision process in BPMN

## 3.2 Design Rationales of Adopted Concepts

The aim of our framework is to help reengineer business process by finding business problems in current (as-is) business models which are against business goals and transforming the as-is toward the next (to-be) business process models which are aligned with business goals, both of which are supported by the evidences from big data. To achieve the purpose, we adopted several concepts and the following are rationales of design decisions.

**Business Process Models**

A business process model which can represent sequences of activities in order to service customers is essential to achieve business goals. There are several business process models such as BPMN (Business Process Model and Notation), Workflow, Activity Diagram in UML, Petri-net and IDEF0. Among the models, BPMN is a standard notation for modeling business processes and it can express not only the flows of activities, but also those of data objects such as inputs or outputs of an activity which can offer clues related to data for analytics. Thus, we adopt BPMN [22] model.

### Goal-Oriented Requirements Models

Goal-oriented requirements models treat all requirements as goals and the abstract goals are refined into more concreate sub-goals by exploring alternatives and selecting among alternatives with trade-off analysis. The identified sub-goals have contribution relationships toward its parent goals. This concept is suitable to express the alignment of business goals, business process and big data analytics which need explorations of diverse alternatives and selections to achieve business goals. There are several goal models such as KAOS [6], i* [8] and NFR (Non-Functional Requirements) Framework [7]. Among them, NFR framework has strong points to represent both functional and non-functional requirements in any abstraction level. Furthermore, Type [Topic] expression which Type represents non-functional side and Topic for functional side allows requirements to be refined into more detailed requirements respectively. This expression along with BPPMN enables root cause analysis. Additionally, NFR Framework can represent not only logical AND/OR relationships, but also relative Satisficing relationships such as Make (++), Help (+), Hurt (-) or Break (--). Thus, it is good to express functional and non-functional requirements traceability. PIG (Problem Interdependency Graph) is similar to NFR Framework and it is used to present problems. PIG along with NFR Framework are also used to provide insights on problems and solutions on as-is and to-be business processes respectively.

### Big Data Processing Frameworks

For big data processing, there are several kinds of frameworks such as MapReduce and Spark which are main streams in industry. They support to process large amounts of data in fast time with a parallel and distributed computing paradigm on commodity clusters. While MapReduce stores intermediate processing files (shuffle) on disks for sorting and reducing, which increase the

number of disk I/O, Spark stores the shuffle files in memory. Thus, the performance of Spark is faster than that of MapReduce [34]. Moreover, while MapReduce provides only the functions regarding to mapping and reducing, Spark offers diverse libraries such as Spark SQL and Spark MLlib which enable interactive analytics which our framework needs.

### 3.3 Goal-Oriented Big data Business Analytics Model (GO-BigBAM) Ontology

In the GO-BigBAM conceptual model for analytics, important business concepts such as business goals, problems and solutions are explicitly represented in order to avoid omissions of important concerns, which facilitates communication between stakeholders reducing ambiguities and eventually enables analyzers to focus on critical business problems and solutions. Additionally, this model includes BPMN (Business Process Model and Notation) [22] which can express both flows of business activities and data, playing a role for a mediator between business and big data. Moreover, GO-BigBAM embraces concepts for big data and big queries to help validate modeled business concepts by the analyzed results from big data. This model connects the dots of those diverse and important concepts in different abstraction levels for big data analytics.

Figure 3.2 and Table 3.1 show the main concepts and relationships along with its diagrammatic convention. We adopted several models and tightly integrated these into GO-BigBAM. Concepts for business such as Business Goal, Solution and Problem are adopted from NFR Framework [7], PIG (Problem Interdependency Graph) [13], so business goals and problems are treated as Softgoals and Softproblems respectively which have no clear-cut criteria, together with Satisficing Contributions which means satisfy goals in a good enough manner. According to NFR Framework, business goals, problems or solutions can be represented by Type [Topic] (e.g., Revenue Lift [ABC

Inc.]), in which Type is for non-functional requirements and Topic is for a functional target domain in this model, a business process which is represented by BPMN.

In this model, every element is a concept, which itself can consist of more refined concepts, hence allowing for recursive decompositions of any concept and the relationships between parents and child are Satisficing Contribution such as MAKE ( ), HELP ( ), BREAK ( ) and HURT ( ) and Correlation relationships such as Conflict or Harmony. Additionally, since a business processes is a means to achieve the business goals, the relationship is also represented by Satisficing Contribution. Moreover, the relationship between business and big data concepts is also Satisficing Relationship because big data results can be used as a claim to validate business modeling elements. A problem, for example, *Low Hit Rate [Clearance Pricing Decision]* is an insight, which makes a negative contribution towards achieving a goal, *Achieve (Forecast Hit Rate > 25%)*, while a solution, *Accurate Prediction Model [Clearance Pricing Decision]* also is an insight, but with a positive contribution towards the goal, which will be validated by big data analytics. KPI (Key Performance Indicator) measures a business process and quantifiable performance goal evaluates KPI.

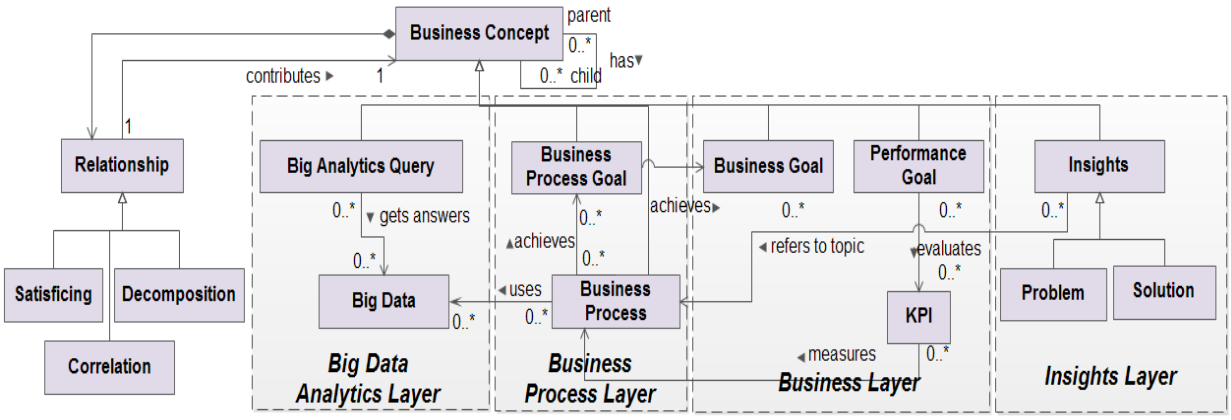


Figure 3.2. Goal-Oriented Big data Business Analytics Model in IRIS framework

Table 3.1. Definition of concepts in GO-BigBAM

| Name | Definition | | Notation |
|---|---|---|---|
| Business Goal | A statement of what a business wishes to accomplish | |  |
| Business Process Goal | A statement of what a business process is intended to accomplish | |  |
| KPI | Key Performance Indicator | |  |
| Performance Goal | A measurable goal to achieve a business or a business process goal | |  |
| Satisficing Contribution | Positive | MAKE, HELP, SOME PLUS towards a parent goal |  |
| | Negative | BREAK, HURT, SOME MINUS towards a parent goal |  |
| Correlation | Conflict | A conflicting relationship which was not intended to achieve a goal |  |
| | Harmony | A synergetic relationship which was not intended to achieve a goal |  |
| Satisficing Label | Satisficed, Weakly Satisficed, Weakly Denied, Denied, Conflict, Undecided | |  |
| Insight | The result of apprehending the inner nature of thighs, and there are two kinds of Insight in our framework | | - |
| Problem | A Phenomenon, which makes some negative contribution towards achieving a Business (Process) Goal | |  |
| Solution | A Phenomenon, which makes some positive contribution towards achieving a Business (Process) Goal | |  |
| Business Process | A collection of inter-related business activities or tasks. The Business process-specific ontology is adopted from BPMN [22] and simplified. | | BPMN Notation |
| Big Data | Data which has characteristics of high volume, high velocity, high variety, and high veracity | |  |
| Big Query | Query for Big Data which can execute in a forma of SQL regardless of underlying big data platform | |  |
| Big Analytics | Analytics results from Big Data such as correlation or prediction analysis | |  |

### 3.4 GO-BigBAM Methods

In our framework, we will provide diverse analytics methods, i.e., Big Analytics, Big Query, Insights Hypothesizing/Validation, Evidence-based Reasoning Method and Solution Selection Method. These methods enable big data analysis to find the most critical business problems and the best solutions to achieve business goals by validating hypothesized problems and solutions in a GO-BigBAM model.

**Evidence-based Reasoning Method**

By using GO-BigBAM model, analyzers can hypothesize a number of diverse candidate problems and solutions in the form of more specific ones, but they should find real problems and solutions, and select the most critical ones which help achieve final business goals among the alternatives.

For this reason, we provide an evidence-based qualitative evaluation method supported by big queries or analytics. The overall concept of reasoning is similar to that of NFR Framework [7], but the difference is that the real data from a big query or analytics is used to validate a model as a supporting claim. This means that models by our framework can overcome the limitations of conceptual modeling, that is, the mismatch between conception and reality, by supporting big data query or big analytics. The evidence acts as Claim Softgoal in NFR Framework which follows the evaluation rules from NFR Framework.

According to the NFR Framework, the evaluation procedure consists of two steps, individual impact- and collection impact-analysis. In the individual impact analysis, the label of leaf node is assigned to one of four labels, i.e., satisficed (✓), denied (✗), conflicting ( ), undetermined ( ). According to the individual impact evaluation catalog in Table 3.2, parent label will be determined. Parents label includes weakly satisficed ( ), weakly denied ( ). Each individual element (an

entity or a relationship) will be evaluated by an evidence of data, then the impact analysis will follow as Figure 3.3.(a) shows.

Once the individual impact analysis finishes, the collection impact analysis starts. As Figure 3.3. (b) shows, if both individual labels with HELP contribution are satisficed, then parents are also satisficed. Also, if one label with HELP is satisficed and the other one with BREAK is denied, then there will be conflict, and if one label with HELP is denied and the other one with BREAK is satisficed, then it is denied. This evaluation process will be done until the final goal, which is called a label propagation procedure.

Table 3.2. Individual impact evaluation catalog

| Child label | Parent label given child contribution type | | | |
|---|---|---|---|---|
| | BREAK | HURT | HELP | MAKE |
| ✗ | $w^+$ | $w^+$ | $w^-$ | ✗ |
| ↘ | ↘ | ↘ | ↘ | ↘ |
| ? | ? | ? | ? | ? |
| ✓ | ✗ | $w^-$ | $w^+$ | ✓ |

Figure 3.3 shows examples of our evidence-based quantitative evaluation which reflects the above rules. In this example, big queries ( ) and KPIs ( ) are used as evidences.

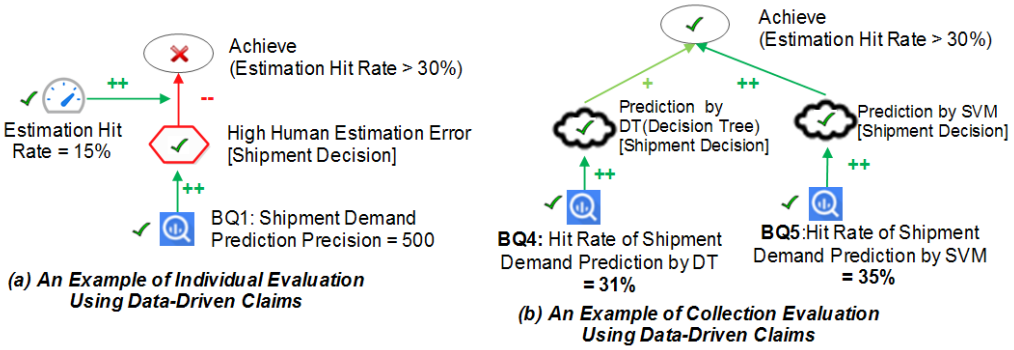


Figure 3.3. Examples of Evidence-based Reasoning Method

### Insight Hypothesizing Method

A business process is a whole-part structure similar to onion rings, so there are several kinds of hypothesizing Phenomenon including Problems or Solution using this structure. Top-down is a hypothesizing method which starts from the outer most elements to the inner most elements, Bottom-up is from inner most to the outer most and Hybrid is a way to integrate Top-down and Bottom-up.

### Insight Validation Method

This method plays a role to validate problems or solutions by using the results from Big Analytics Method or Big Query Method. For this method, Performance goal is used in the form of "Achieve (KPI, Operator, Target Value)", where Operator = {<, =, >} to check whether a goal is achieved or not and Performance goal has Threshold Offset which allows a flexible evaluation range.

Problem/Solution Validation Algorithm constitutes of two stage: deciding goal-achievement status using analytics values and deciding validation label according to the goal-achievement status. The algorithm is like the following.

```
/* 1. Decide Goal-Achievement Status by Analytics_Value*/
If Operator = ">" then
  If Analytics_Value >= Target_Value – Threshold_Offset then
      Goal_Achievement = True
  Else Goal_Achievement = False
  End If
Else If Operator = "<" then
  If Analytics_Value <= Target_Value + Threshold_Offset then
      Goal_Achievement = True
  Else Goal_Achievement = False
   End If
Else If Operator = "=" then
  If ((Target_Value – Threshold_Offset) <= Analytics_Value <= (Target_Value + Threshold_Offset) then Goal_Achievement = True
  Else Goal_Achievement = False
  End If
End If

/* 2. Decide Validation Label by Goal-Achievement */
If Validating_Type = Problem then
    If Goal-Achievement = False then
        Problem.Label = Satisficed
    Else
        Problem.Label = Denied
```

```
    End If
Else If Validating_Type = Solution then
        Solution.Label = Satisficed
    Else
        Solution.Label = Denied
End If
```

**Solution Selection Method**

Selection method is for finding the best solution among the alternatives. An option in potential solutions has its relative importance which are inherited from a parent goal which is traceable back, and several contribution links towards problems. In this method, the best solution will be selected as the maximum value among sums of Relative Importance Value (RIV) * Relative Contribution Value (RCV) for each option. The formula will be the following. Table 3.3 shows the values of RIV and RCS respectively.

Then, to rank alternatives, Priority Ranking Value (PRV) can be calculated in consideration of Relative Importance Value (RIV) and Relative Contribution Value (RCV) of a candidate option according to Table 3.3.

$$PRV(O) = \left|\sum_{i=1}^{n} RIV(OC_i) * RCV(OC_i)\right| \quad (1)$$

where O is an potential option, n is the number of contribution links towards parents of O, O = $\{OC_1, OC_2, \ldots, OC_n\}$ and $OC_i$ is a contribution link.

If PRV of a candidate option is the highest among the values for all the candidates, then it will be selected first as the most important problem or solution.

Table 3.3. Relative importance and contribution assignment.

| **Relative Importance Value (RIV)** | | **Relative Contribution Value (RCV)** | | | |
|---|---|---|---|---|---|
| !! | 1.5 | ↑-- | -3.0 | ↑++ | 3.0 |
| ! | 1.2 | ↑- | -2.0 | ↑+ | 2.0 |
| No Priority | 1.0 | ↑S- | -1.0 | ↑S+ | 1.0 |

## Big Analytics Methods

Big analytics methods return analytics results using Spark Machine Learning Library (Spark MLlib) to predict or describe data. The analytics results can be saved back to Big Data Platform or can be used to retrieve them by Big Query Methods. Among the methods that Spark MLlib provides as Figure 3.4 shows, we have implemented Decision Tree and Support Vector Machine (SVM) in our tool.

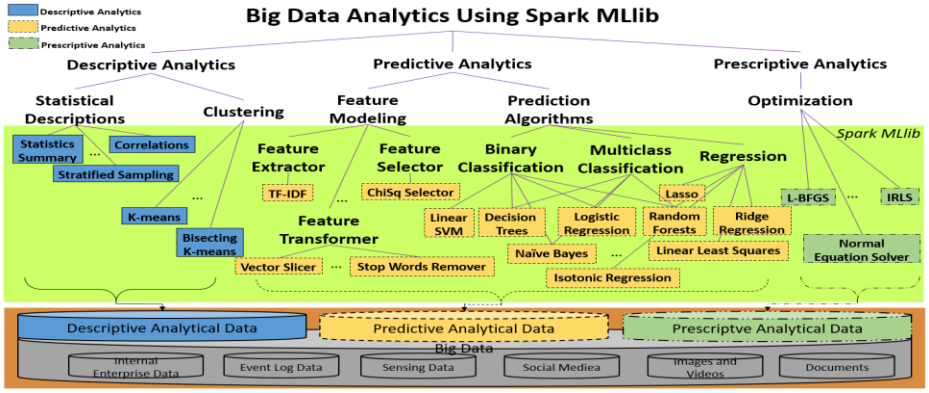


Figure 3.4. Big data analytics catalog in Spark

## Big Query Methods

Big Query Methods can also be used to get results from Big data by using Spark SQL. While Big Analytics Method is for processing big data, Big Query is for retrieving data, either processed- or non-processed data.

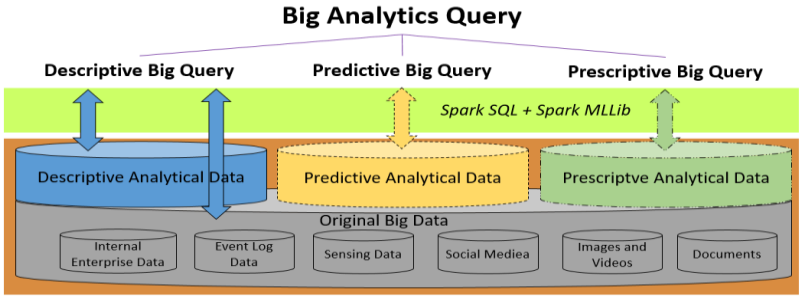


Figure 3.5. Big data query using Spark SQL

## 3.5 IRIS Action Process

In IRIS, the process for using big data business analytics is as depicted in Figure 3.6. It consists of four stages. In business analytics, data gathering and cleansing are needed but these are beyond the scope of this dissertation.

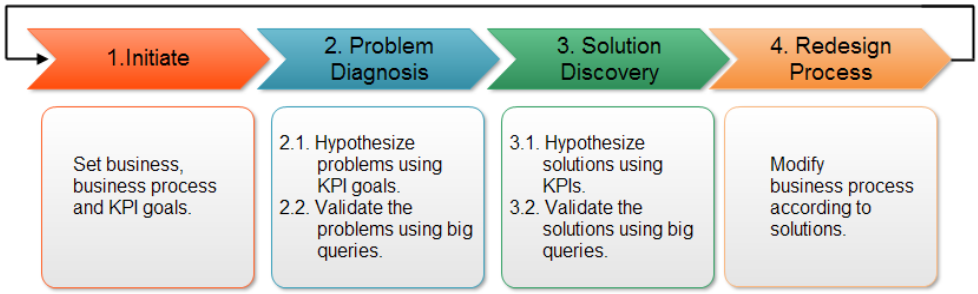


Figure 3.6. IRIS process for big data business analytics

**1. Initiate:** This is to establish business goals, refine them in terms of business process goals, and further refine them in terms of KPI (Key Performance Indicator) goals. As will be shown in Figure 3.7 in the next section, for example, Zara had *Revenue Lift* as a business goal, and, to achieve the goal, it needed *Effective [Clearance Pricing Decision]* as a business process goal, which in turn needed *Reliable [Clearance Pricing Decision]* and *Timely [Clearance Pricing Decision]* as more specific business process goals. These two are potentially conflicting with each other, and will be shown as such later in Figure 3.8. A business process goal can be expressed as a KPI goal, in the form: Achieve [KPI] – e.g., *Achieve (Forecast Hit Rate > 25%)* or *Achieve (Processing Time for clearance pricing decision < 15 days).*

**2. Problem Diagnosis:** This is to find high-level problems and their sub-problems (root-causes). A problem is a phenomenon, which negatively contributes to achieve a business or business process goal. For example, *Low Hit Rate [Clearance Pricing Decision]*, which could be refined to be more specific, in terms of Demand Prediction and Markdown Prediction may BREAK *Achieve*

*(Forecast Hit Rate > 25%)*. Finding a problem is a two-step process: 1) hypothesizing a phenomenon to be a problem seeing the AS-IS process presented in a business context model, 2) validating the hypothesized problem, using appropriate big queries in terms of the negated KPI of the goal:

Hypothesize Problem: KPI Goal × Business Process → [Hypothesized Problem =

~Achieved (KPI Goal) × Business Process]

For example, *Hypothesize Problem (Achieve (Forecast Hit Rate > 25%), Clearance Pricing Decision) = [(~Achieved (Achieve(Forecast Hit Rate >25%))), Clearance Pricing Decision)]*.

Validate Problem: Hypothesized Problem × Big Query × Big Data

→ {True, False}

If the result of Validate Problem is True,

Change Status of Problem: Hypothesized Problem → Validated Problem.

For example, a big query can be produced for the hypothesized problem *[~Achieved (Achieve(Forecast Hit Rate > 25%)), Clearance Pricing Decision)]*. The query processor can take Hypothesized Problem, Big Query expressed in SQL syntax and Big Data including platform information as inputs, and determine True as the output, which signifies that the Hypothesized Problem is validated to be a problem. If the result of the Validate Problem is True, it will be changed into Validated Problem, i.e. *Low Hit Rate [Clearance Pricing Decision]* with the ✓ symbol. The Validated Problem can be refined into more specific causes by finding related activities among sub-elements in the current process in a top-down manner. In this case, business analyzers can find out *Predict Demand Manually* and *Estimate Time to Sell (ETS)* are related to the problem and again they are hypothesized and validate by using big queries. As a result, *Low*

*Hit Rate [Predict Demand Manually]* and *Low Hit Rate [Estimate Time to Sell (ETS)]* become Validated Problems. Alternately, it is possible to hypothesize in a bottom-up manner which starts from sub-elements to the top-elements, or hybrid which combines the top-down and the bottom-up. Hybrid method will be described in Section 3.6.

The same KPI may be used exactly as is in finding a problem, or instead a leeway may be exercised in setting the threshold – e.g., instead of using 25%, a lower value, such as 22%, or a range, such as {22% .. 26%}, could be used.

**3. Solution Discovery:** This is to find solutions to the problems of the AS-IS business process, towards a new TO-BE process. Similarly to the case with finding a problem, finding a solution also takes a two-step process: hypothesizing an action to be a solution in a business context model, and then validating it using big queries. For the steps:

Hypothesize Solution: Validated Problem × Business Process →

[Hypothesized Solution = KPI Goal × Business Process]

Validate Solution: Hypothesized Solution × Big Query × Big Data

→ {True, False}

If the result of Validate Solution is True,

Change Status of Solution: Hypothesized Solution → Validated Solution.

For example, given *Low Hit Rate [Predict Demand Manually]* as a Problem and clearance pricing decision process as its context, the use of *Clearance Pricing Prediction by Big Data (Social Media Fashion Trend, Online and Offline Sales Data)* could be hypothesized to be a solution. Validation of a solution is done, similarly to the way a problem is validated. One different thing is that before the big queries are executed, the big data can be predicted by using IRIS prediction function which

utilizes Spark ML (Machine Learning) Library such as Decision Tree or SVM (Support Vector Machine).

**4. Redesign Process:** The solutions, that have been discovered in the previous phase, now need to be realized in a TO-BE process. The parts of the AS-IS process that need to be modified are the Topic parts of every Type [Topic] expression of the solutions. For example, the *Adjust Decision* group of the AS-IS process, which appears as the topic of a solution, needs to be modified for *Timely [Clearance Pricing Decision]*. There can be many different ways for this modification, and typically it would require some intervention by subject matter experts.

Throughout the process, alternatives are ranked, according to their relative importance, in problems, solutions, KPIs, queries, and the specific changes to be made to the AS-IS process. For example, multiple big queries can be ranked, for business process KPI goals:

$$\text{Rank Queries: Big Query} \times \text{KPI Goal} \rightarrow \text{Big Query X Rank}$$

In other words, given a set of queries Big Query = $\{q_1, q_2, \ldots, q_i, \ldots, q_m\}$ and another set of KPI goals KPI Goal = $\{g_1, g_2, \ldots, g_j, \ldots, g_n\}$, Rank Queries will yield $\{(q1, r1), (q2, r2), \ldots, (q_i, g_j), \ldots, (q_m, r_n)\}$, where r stands for some ranking (order). The higher value means higher ranking. Overall, evaluation of each alternative is done, in consideration of the priority of the goals it is associated with.

Also throughout the process, a label propagation procedure [7, 13] can be used to determine the degree to which a lower-level phenomenon affects its upper-level problems and goals. In IRIS, this procedure works in consideration of big data analytics as evidences, that have to do with the determination of a phenomenon as a problem or a solution.

### 3.6 IRIS in Action: Shipment Decision Process

We applied IRIS framework – its model, method, process and tool – to the shipment decision process to answer the research questions. The road map for this application starts with an as-is shipment decision process diagnostics (Figure 3.7), followed by a consideration of alternative potential solutions, and tradeoffs among them (Figure 3.8), for the problems of the as-is process and ends with next business actions for to-be process.

**1. Initiate**

ABC has *Increase Global Revenue [ABC Inc.]* as a business goal and a measurable performance goal is identified to see if the business goal is achieved or not, in this case *Achieve (Revenue > 10%).* To achieve the performance goal, there are many candidate business processes goals as hypotheses such as *Effective [Clearance Pricing Decision]* or *Effective [Shipment Decision]*. *Effective [Shipment Decision]* can be further refined into *Match Delivery Requirements [Shipment Decision]* and *Minimize Maintenance Cost [Shipment Decision]*. To select the most effective target process, analyzers validate the effectiveness of processes by using big queries or big analytics, in this case big analytics such as BA1 and BA2. We assume after execute big analytics, the results come out as Figure 3.7. After execution of big data analytics, they know Shipment Decision Process has more contribution than Clearance Pricing Decision Process. Additionally, *Match Delivery Requirements [Shipment Decision]* is more critical (denoted by "!!") than *Minimize Maintenance Cost [Shipment Decision]*, validated by using BA3 and BA4. These goal refinements are done during "Initiate" stage of the IRIS process.

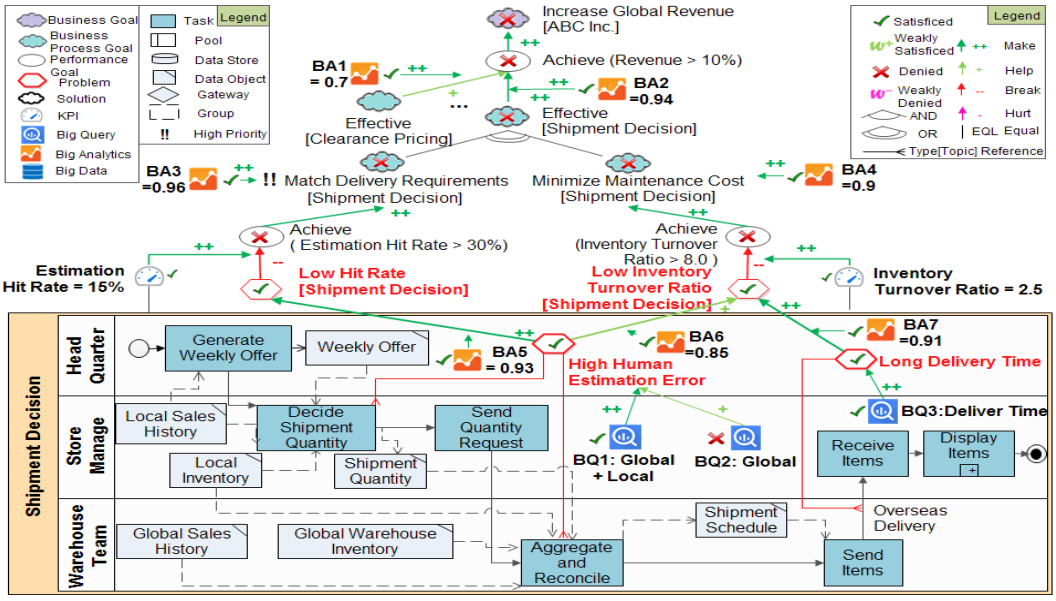


Figure 3.7. AS-IS clearance pricing decision process diagnostics

- BA1: /* correlation between clearance decision and revenue */
- BA2: /* correlation between shipment and revenue */
- BA3: /* correlation between estimation hit rate and revenue */
- BA4: /* correlation between maintenance cost and revenue */

**2. Diagnose Problems**

Finding problems with the as-is business process consists of two steps: hypothesizing a problem using GO-BAM model and validating it by using big queries or big analytics. In the first step, analyzers hypothesize candidate problems. In Figure 3.7, the problems are denoted as *Low Hit Rate [Shipment Decision]* which BREAK *Achieve (Estimation Hit Rate > 30%)*, and *Low Inventory Turnover Ratio [Shipment Decision]* which BREAK *Achieve (Inventory Turnover Ratio > 8.0)*. As root causes of the former problem, *High Human Estimation Error [Demand Shipment Quantity, Aggregate and Reconcile]* is hypothesized and for the later problem, *Long Delivery Time [Overseas Delivery]* is also hypothesized as a root cause. While *High Human Estimation Error* is associated with the two activities, i.e., *Demand Shipment Quantity* and

*Aggregate and Reconcile*, *Long Delivery Time [Overseas Delivery]* is associated with a sequence flow, i.e., *Overseas Delivery*.

In the second step, the hypothesized problem gets validated through one or more big queries or big analytics. There can be many candidate queries or big analytics for validating problems. For example, regarding to the previously hypothesized problem, *High Human Estimation Error [Demand Shipment Quantity, Aggregate and Reconcile]*, analyzers can consider two statistical big queries: BQ1, for global inventory and local estimation, and BQ2, for local inventory. Both queries can be generated by using Data objects which are connected with corresponding activities *Demand Shipment Quantity* and *Aggregate and Reconcile*, i.e., *Local Sales History*, *Shipment Quantity*, and *Global Sales History*, *Shipment Schedule* respectively. BQ1 and BQ2 are shown below as SQL queries which are successfully executed in our IRIS assistant tool. Among the two queries, BQ1 is chosen because considering global and local estimation is more suitable than only global estimation.

BQ1: Demand Difference of Global + Local Estimation

BQ2: Demand Difference of Global Estimation

```
/* shipment demand prediction difference on global inventory */
SELECT a.category, avg(prdt_dmd - real_dmd) as prediction_diff
FROM ( SELECT category, prdt_dmd
        FROM   shipment_schedule_history
        WHERE  sales_year=2015
            GROUP BY category ) a,
    ( SELECT  category, sum(sales_count) AS real_dmd
        FROM     global_sales_history
        WHERE    sales_year = 2015
        GROUP BY category) b
WHERE a.category = b.category;

 /* shipment demand prediction difference on local inventory */
SELECT a.category, avg(prdt_dmd - real_dmd) as prediction_diff
FROM ( SELECT category, prdt_dmd
        FROM   shipment_quantity_history
        WHERE  sales_year=2015
            GROUP BY category ) a,
    ( SELECT  category, sum(sales_count) AS real_dmd
```

```
        FROM     local_sales_history
        WHERE    sales_year = 2015
        GROUP BY category) b
WHERE a.category = b.category;
```

Let us suppose the answer to the selected BQ1, demand differences on global and local, are 400 and 500 respectively, and the threshold is 200, then it will True (hence, the ✓ symbol), the hypothesized problem is likely to be a validated problem (hence, the ✓ symbol); otherwise, not a validated problem. Once *High Human Estimation Error* is validated, analyzers should validate the relationship between *Low Hit Rate [Shipment Decision]* and *High Human Estimation Error* to see if the root cause really affects the parent problem using BA5, 6.

BA5: /* correlation between estimation hit rate and demand difference */

BA6: /* correlation between inventory turnover ratio and demand difference */

For the other hypothesized problem, *Long Delivery Time [Overseas Delivery]*, BQ3 can be used. Similarly, Data objects of activities such as *Send Items* which are connected with → *Overseas Delivery* are used to generate the query statement.

BQ3: /* time for overseas delivery */

```
SELECT a.store_code, a.category, avg(delivered_date- scheduled_date)
FROM ( SELECT store_code, category, request_date, scheduled_date
        FROM   shipment_schedule_history
        WHERE  sales_year=2015
            GROUP BY store_code, category, request_date) a,
    ( SELECT  store_code, category, request_date, delivered_date
        FROM     local_inventory_history
        WHERE    sales_year = 2015
        GROUP BY store_code, category, request_date ) b
WHERE a.store_code = b.store_code and a.category = b.category
            and a.request_date = b.request_date;
```

BA7: /* correlation between long delivery time and inventory turnover ratio */

Let us suppose that BQ3 also validates *Long Delivery Time* as True. Since both *High Human Estimation Error* and *Long Delivery Time* are satisficed (✓), according to the evidence-based evaluation method in previous section, their parents *Low Hit Rate [Shipment Decision]* and *Low Inventory Turnover Ratio [Shipment Decision]* are also satisficed and the final business goal *Increase Global Revenue [ABC Inc.]* is denied using the closed world assumption [7, 13].

**3. Discover Solutions**

Discovering solutions to the problems with the as-is process, towards a to-be business process, also consists of two steps: hypothesizing a solution and validating it. In the first step, for example in Figure 3.8, given *Low Hit Rate [Shipment Decision]* as a problem, *Prediction by Decision Tree [Aggregate and Reconcile]* and *Prediction by SVM [Aggregate and Reconcile]* which have BREAK contribution toward *High Human Estimation Error[Decide Shipment Quantity, Aggregate and Reconcile]* could be hypothesized to be solutions. For *Long Deliver Time [Overseas Delivery]*, *Prediction by SVM [Overseas Delivery]* is hypothesized as a solution.

Big queries are also used to validate the hypothesized solutions and we show only descriptions, e.g.:

***BQ4:*** /* the accuracy of shipment demand prediction on data predicted by Decision Tree*/

***BQ5:*** /* the accuracy of shipment demand prediction on data predicted by SVM */

***BQ6:*** /* the accuracy of shipment demand prediction on history data predicted by SVM */

***BQ7:*** /* the accuracy of overseas delivery prediction on data predicted by SVM */

Similarly, if the solutions are validated by big queries or big analytics, according to the evidence-based evaluation method, analyzers can check if the final goal is achieved or not with degrees such as Satisfice or Weakly Satisficed.

As other alternatives, the validated solutions can be combined together to fully achieve the final goal. For example, the combination of both solutions, i.e., ⬭ *Prediction by SVM [Aggregate and Reconcile]* and ⬭ *Prediction by SVM [Overseas Delivery]* can be another alternative solution.

To select the most effective solution, analyzers need to rank the solutions. According to formula (1), each solution calculates the Priority Rank Value (PRV). For example, while PRV of only the ⬭ Prediction by DT [Aggregate and Reconcile] is |(1.5*-2)| =3 (denoted S✓ in Figure 3.8.(a)), PRV of the above combined solution can be calculated |(1.5* -3) + (1.0*-3)| = 7.5 (denoted S✓ in Figure 3.8.(b)). For each selection, the final business goal achievement will be Weakly Satisficed $W^+$ and Satisficed✓ respectively. Thus, the combination solution will be selected.

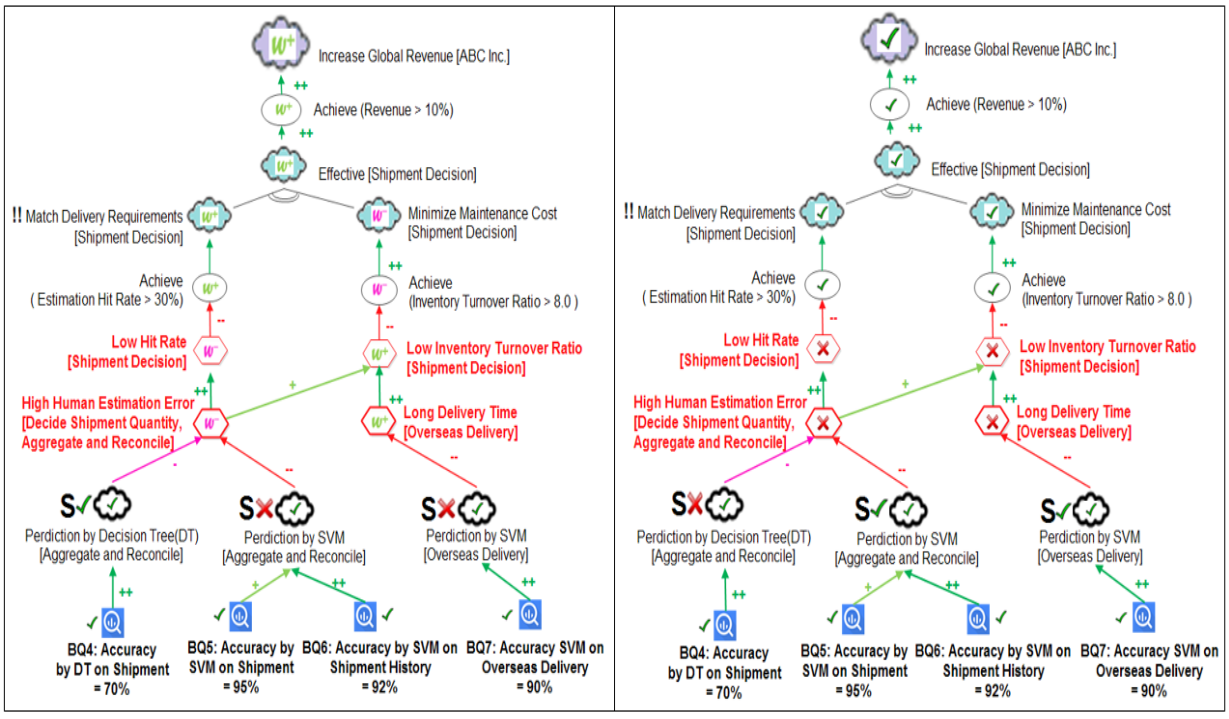


Figure 3.8. Alternative potential solutions, and tradeoffs among them, for the problems of the AS-IS Process

**4. Find Next Actions**

Using the above final solution, analyzers can change business processes as the next action items. For example, for the solution, *Prediction by SVM [Aggregate and Reconcile]*, *Aggregate and Reconcile* can be changed into *Predict Shipment Demand by SVM* and *Reconcile*. For the other solution, *Prediction by SVM [Overseas Delivery]*, *Predict Overseas Delivery* can be newly added.

## 3.7 Experiment Data Results

When it comes to fastness (RQ3), we run our program on customer data for demand prediction by using Decision Tree, SVM machine learning algorithms, and correlation analysis which Spark provides, and executed BQ2, and BA1. Execution environment is Processor (2.5G), RAM (8.0G), 64-bit OS and Spark V.1.6.0 is executed in a standalone mode with multiple worker nodes. Data size is total 68M. We measured processing time for each case, and the result are shown in Figure 3.9. Figure 3.9.(a) is the results of processing time including training on prediction by SVM and Decision Tree respectively, (b) is the results of processing time of Big Query and Big Analytics for correlation.

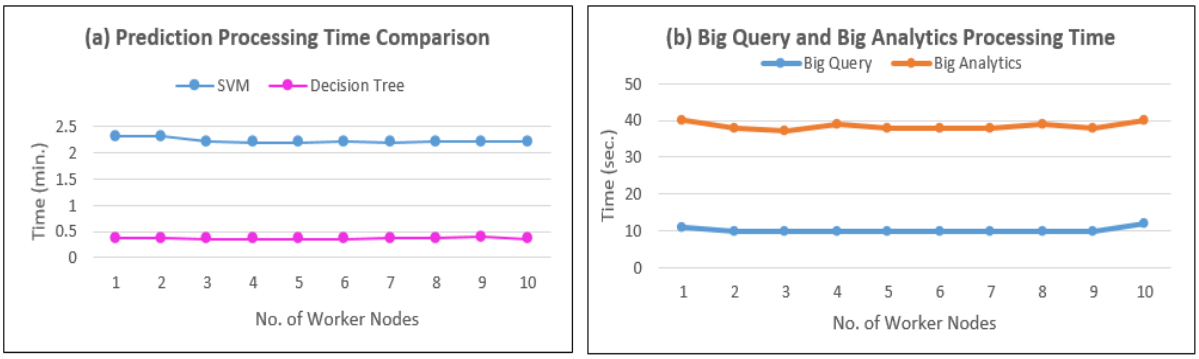


Figure 3.9. (a) Comparison of prediction performances between SVM and Decision Tree (b) Big query and big analytics processing time in IRIS assistant tool using Spark

While average of prediction processing time by SVM was about 2.3 min., average by Decision Tree was about 0.5 min. Also, while average of big query processing time was about 10 sec., average of big analytics was about 38 sec. The reason is that BA1 includes the procedure mapping from string data to float to analyze in Spark, it took more time. The tasks were processed within reasonable time.

# CHAPTER 4

# A GOAL-ORIENTED BIG DATA MODELING TO SUPPORT BUSINESS ANALYTICS

Thought it emerged only recently, big data has quickly been embraced by all walks of life, including businesses, governments and academia. The big data fever is unprecedented, due to the promises, hypes and hopes built around it. Yet, behind the façade of big data is the simple notion of data – the data that are characterized by many Vs (***v***olume, ***v***elocity, ***v***ariety, ***v***eracity and possibly more) and that require new technologies to unleash its power.

Business analytics (hereafter, *BA*) [35] is one of the key areas that can benefit greatly from big data. With new business insights gained from big data, BA can help make better business decisions and improve business processes. Yet, the quality of BA can only be as good, or as bad, as the quality of the big data it uses. The quality attributes of big data, together with relationships between data, should therefore be defined in a big data model. With a good quality big data model, BA can accurately identify important business concerns, trend opportunities, and useful business insights, which, in turn, can lead to good business decisions.

Yet, no guidelines are available for how to develop a high-quality big data model in a systematic and rational way.

To address this problem, this paper proposes a systematic approach to big data modeling. The approach specifically tackles three aspects of big data model: *relevance, comprehensiveness* and *relative priorities.* These aspects stipulate that the data and relationships between the data should be *relevant* to their use, *comprehensive* enough to cover business decision-making and *prioritized* so that their importance is clear. In order to attain these three data model qualities, this chapter then proposes a goal-oriented approach that helps *build such qualities into* a big data model. Our

ultimate goal is to provide this approach as a service over the Internet so that it can be used to support big data analytics.

## 4.1 Related Work

The key distinctive of our goal-oriented approach are: addressing the quality of a big data model, in terms of relevance, comprehensiveness and prioritization; three comprehensiveness dimensions for modeling big data in consideration of a variety of types and sources of data; an ontology, for explicitly identifying the key vocabulary for communication and modeling, which includes goals, problems and solutions, business analytics, as well as the three comprehensive dimensions of a data model; and a rational and systematic process for modeling big data, with the help of a (prototype) tool support.

Research in the area of data mining is relevant to our work, in identifying multiple groups of data for obtaining more accurate prediction results [36]. By and large, data mining aims at extracting patterns and knowledge from a large volume of data. Although the focus in data mining usually is not on modelling data but on techniques, including machine learning in AI, the notion of interestingness at least roughly seems to be related to our notion relevance and/or prioritization. For example, the notion of interestingness seems to have to with determining the relevance and importance of the data or links/web sites and prioritization of such links/websites.

The area of Database is relevant to our work since our work is also concerned with modeling data. A key difference between the two has to do with the notion of Big Data in our work. Although work in the area of databases is not about Big Data, the notion of virtual repository, through a federated approach, is applicable to our work. A virtual repository is intended to offer an easy access to (usually geographically) distributed, and often times independent and heterogeneous,

databases, through a single (virtual) database with a single set of mechanisms for accessing the database [37]. Creating a virtual database involves integrating often times incompatible database schemas of a native and a number of foreign database schemas. Our notions of "internal" and "external" respectively are similar to "native" and "foreign", at least roughly. The difference lies in the comprehensiveness of "naïve" and "foreign", i.e., by and large, "native" and "foreign" respectively cover only internal-offline-proper and external-offline-proper of our comprehensive dimensions of big data. From database research, we adopt the three abstraction principles of classification/instantiation, generalization/specialization and aggregation/decomposition [38].

Research in Big Data [39] has seen significant advancements on techniques for the "how" aspect of efficiently managing huge amounts of data (e.g., Hadoop [40], MapReduce [41], and machine-learning techniques [42]), but not so much for "what" kinds of data to manage and "why". IRIS is intended for addressing the "what" and "why" questions in modeling (virtual) big data, not in an ad hoc manner, but rationally and systematically.

The areas of business process management (e.g., [43]) and business process reengineering (e.g., [44]) offer methodologies, techniques and tools (e.g., [45]). The majority of these address business processes and a variety of KPIs [46], and some of them cover BA. We adopt key ideas from these areas, and go beyond, in a complementary manner, to provide a disciplined methodology for developing a big data model.

More recently, in the area of requirements engineering, some proposals have started to appear on big data or BA (e.g., [47]), e.g., for business process improvement in terms of diagnosing problems with a business process. Our work is similar, but additionally considers use of big data analytics in finding not only problems but also exploring solutions too. More specifically, our work

incorporates business processes, problems and goals, and big data models, in rationally and systematically "connecting the dots" among them all, in modeling big data for supporting BA.

## 4.2 Overview of GO-BigDM

In general, a goal-oriented approach means exploring alternatives and selecting among them, through tradeoff analyses, towards achieving goals. This way, a goal-oriented approach helps rationalize decisions – be they about a business or modeling, in a systematic manner, while also helping establish traceabilities among the key (ontological) concepts. Goal-oriented approaches have long been used in AI problem solving and many other areas (e.g., [6], [7], [8], [48], [49]). A variety of applications have benefitted from the use of a goal-oriented approach (See, for example, [50], [51], [52], [53]).

IRIS takes a goal-oriented approach to modeling big data for supporting BA. In this approach, goals are explicitly represented, and problems and solutions are explored, hypothesized and validated, with the use of big data, which, in turn, necessitates the exploration of, and selection among, alternatives in big data models. Throughout this process, forward traceabilities are established among business goals, problems, solutions, and big data models.

The three qualities of a big data model are also presented as goals, whose achievement depends on how well the big data model helps identify and solve business problems, in ultimately achieving the relevant business goals. In other words, the relevance, comprehensive and relative priorities of data should become all backward traceable.

More specifically, IRIS offers an ontology (i.e., vocabulary for its particular universe of discourse), which explicitly shows the key individual concepts and relationships between them. This ontology includes goals, problems, solutions, business KPIs (key performance indicators), and also big data

modelling concepts. Additionally, this ontology also includes concepts for the qualities of a big data model.

In particular, the ontology includes three comprehensiveness dimensions – internal vs. external, offline vs. online, and proper vs. analytical. Internal data is the (conventional historical) data that pertains to the particular business organization, while external data comes from outside sources, such as social networking sites or online marketplaces. The offline-online dimension is intended for considering not only the conventional offline data but also online data, which is increasingly becoming available and useful. The proper-analytical dimension is intended for maintaining and processing not only application-specific data but also data about the results of analytics. These three dimensions are intended to help capture a variety of not only different types of data but also different sources of data. Incorporation of theses varieties leads to a virtual big data model, which helps explore new or external opportunities, that is not possible by simply analyzing the conventional, offline historical data.

In IRIS, the notion of virtuality is treated as an important property of a big data model, but potentially with significantly increased cost for collecting, maintaining, processing, transmitting, analyzing, visualizing and understanding the data, especially if its volume is huge and/or its velocity has to be extremely fast. Relevance and prioritization here come in to play an important role in deciding what to include, and what not, as well as in deciding how much efforts and resources to allocate to the different types and sources of big data.

The ontology also includes three kinds of relationships between the entities of a big data model, which correspond to the three conventional organizational dimensions (or abstraction principles) – classification/instantiation, generalization/specialization, and aggregation/decomposition [38].

These three dimensions are used not only for structuring a large amount of (i.e., big) data but also for exploring relevant data that might be potentially useful in supporting BA.

The three qualities of a big data model, together with the three organizational dimensions, are intended to help avoid omissions or commissions of potentially important data.

### 4.3 Data Quality Definitions

There can be good, as well as bad, data models. A data model would be good, if it can effectively support BA, in particular, in finding business problems and solving them; otherwise, bad. For big data modelling, the key emphasis in IRIS lies in, among other Vs, the notion of ***v****ariety*, hence the need for accommodating a ***v***ariety of not only different types of data but also sources of data. This is important, especially in the emergence of social networking sites, online marketplaces, web analytics, Internet of Things, sensor networks, etc., that are increasingly becoming part of every walks of life.

But then, the growth in the ***v****olume* of data seems remarkably large and also in their ***v****elocity*, making the cost and technical feasibility issues become more important than ever before. The availability of more data might mean, the more complete analysis, but at an (possibly prohibitively) increased cost for managing the data.

There may be many different notions of *data model quality* (See, for example, [54]), but, for a "*big" data model*, we propose three notions of quality:

- *Relevance* of data, for including data that seems potentially relevant to validating problems and solutions, and excluding data that does not – this would help avoid prohibitive increase in cost for collecting, maintaining, processing, transmitting, analyzing, visualizing and understanding the potentially tremendous ***v***olume of data.

- *Comprehensiveness* of data, for a ***v****ariety* of different types and sources of data.
- *Relative priorities* of data, for determining how to allocate limited amount of resources, for example, when the ***v***olume of data is huge or the ***v***elocity at which data may arrive needs to be extremely fast.

In the presence of a huge ***v***ariety, and also possibly ***v***olume and ***v***elocity, of data, these big data qualities can help decide what to manage and what not? For example, concerning Zara's shipping decision, they can help answer a number of questions, such as:

- Should information about the skirts Kate wore at a Queen's party be part of the data, hence the virtual data model?
- Should the online game trend be considered?

How much effort should be allocated to collecting data about a competitor's price change vs. data about hottest keywords in Internet search?

**Relevance**

This aspect is about the utility of a data element, which can be used as a criteria in determining if the data element should be considered in supporting BA. This notion is useful for helping to prevent commissions. For example, for Zara's shipping decision on ladies' apparel, a social network recommendation on games is unlikely to be useful, hence irrelevant.

A data model element, *e*, is said to be relevant, if there is a (traceability) link between *e* and some problem or solution. More specifically, a data model (element), *e*, is said to be *d-distance relevant*, if the number of links that lie between *e* and some nearest problem or solution is *d*. The smaller the distance, the more relevant *e* is. A data model (element), *e*, is also said to be *satisficing relevant*, if *e* makes a contribution towards validating either a hypothesized problem or solution. The more

positive the contribution is, the more relevant *e* is. These two notions of big data relevance are depicted in Figure 4.1. In Figure 4.1, each $C_i$ denotes a class or an entity in EERD [55]. Here, $C_1$'s distance to Problem is smaller than $C_3$'s, hence more relevant. Now, $C_1$'s distance is the same as $C_2$'s distance, but $C_1$'s contribution is stronger than $C_2$'s contribution, hence more relevant. The same holds for the Solution too.

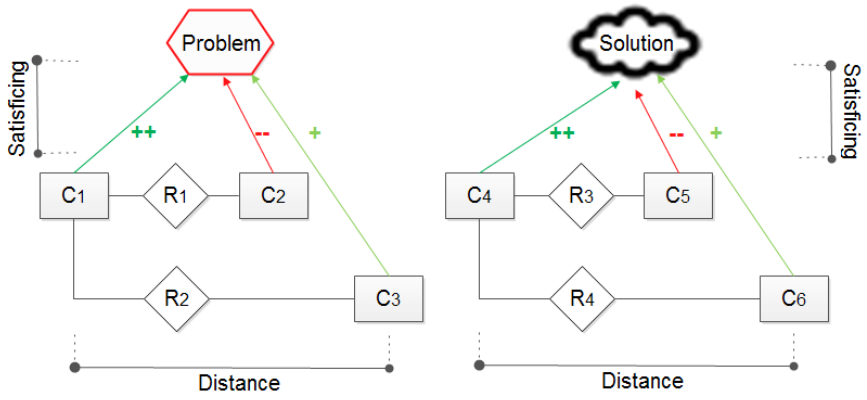


Figure 4.1. Distance relevance and satisficing relevance

**Comprehensiveness**

This aspect is to accommodate a ***variety*** of types and sources of data that are increasingly becoming available and useful into a big data model, for better serving BA. This notion is useful to help prevent omissions of potentially important data. For example, for Zara's shipping decision on ladies' apparel, external data in the form of a social network recommendation on popular ladies' apparel products is likely to be useful, hence relevant and included in the big data model.

Besides social networking datasets, there are also other sources of potentially relevant data too, for example, online shopping and advertisement statistics, which could help avoid missing out trend opportunities, due to the consideration of only the historical data that pertains to a particular business. The three comprehensive dimensions of big data, as shown in Figure 4.2, are intended to

help capture and utilize data from a *variety* of sources and in many varying types, in carrying out big data analytics.

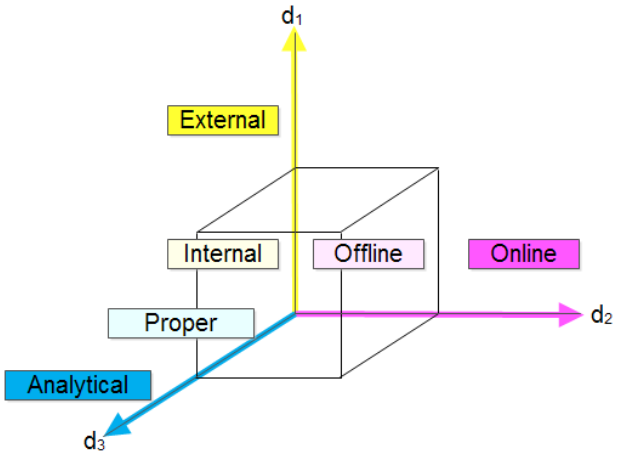


Figure 4.2. Three big data comprehensive dimensions

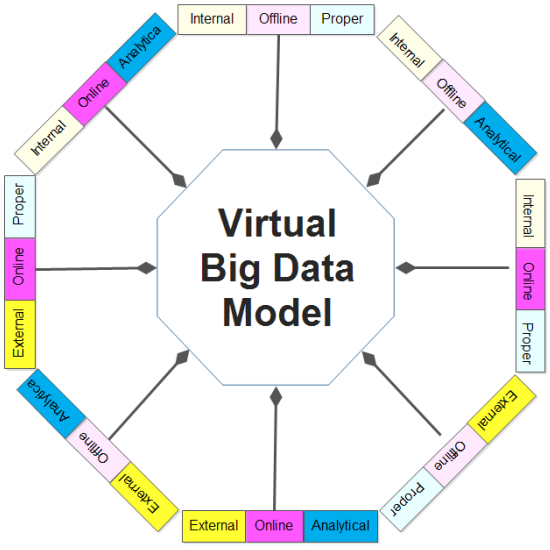


Figure 4.3. The resulting eight square units

Briefly,

- *Internal/External dimension*: to consider not only data that is *internal* to a business (e.g., in the business's local datacenter) but also data that is external to the business (e.g., through a national

repository or a social networking site);

- *Offline/Online dimension*: to consider not only the traditional offline data but also online data that is increasingly becoming prevalent; and
- *Proper/Analytical dimension:* to consider not only ordinary data (e.g., Sales History or Local Inventory Level – say, of first-order data) but also data that has been generated, through analytics, from such ordinary regular data (e.g., Sales Trend or Local Inventory Level Trend – say, of second-order data and higher).

These three dimensions, then, would yield eight (8) different sources, as the result of the cross product: {Internal, External} X {Offline, Online} X {Proper, Analytical}.

For the sake of brevity, we will sometimes use $d_1$ to refer to the internal/external dimension, $d_2$ to refer to the online/offline dimension and $d_3$ to refer to proper/analytical dimension. So, each of the eight (8) square units of the three dimensions then can be denoted as $d_1$-$d_2$-$d_3$. For example, "Internal-Offline-Proper" refers to those offline, ordinary historical data of a traditional system that is internal to a particular business. For another example, "External-Online-Analytical" refers to those analytical data that come from external, online sources, such as social networking sites or marketplaces.

Now, we introduce the three organizational dimensions, as in Figure 4.4, which are intended to help structure a large ***v***ariety of big data possibly in a huge ***v***olume arriving at a fast ***v***elocity. All these three dimensions can be used to explore, and relate, data in the same or across different dimensions of big data comprehensiveness.

Briefly,

- *Classification/Instantiation dimension:* This is to relate instances to classes. For example,

suppose that Kate's skirt becomes a hot keyword in Internet search (e.g., on Most Read or Trend Now). In this case, Kate's skirt is likely to be an instance of the class of order items which Zara has.

- *Generalization/Specialization dimension*: This is to relate data through superclass-subclass relationships. For example, online shopping is growing as an overall trend in the world, which can be considered to be more general than Zara's own potential online shopping trend. Then, the overall trend may be reflected in Zara's online clothes sales trend, hence consequently affecting shipping decisions on clothes, concerning both offline and online (This is a kind of deductive reasoning, hence likely to be sound). Now let us suppose that the world-wide trend in children's overalls is growing. This is a more special case than Zara's overall sales on all items, and predicting that Zara's overall sales on all items will also go up, hence consequently increase in shipping quantities, will more likely be invalid. (This is a kind of abductive reasoning, hence likely to be unsound).
- *Aggregation/Decomposition dimension:* This is to associate data of different classes (in some literature, in the name of attributes or properties). For example, hot keywords from an external search engine can be combined together with internal online feedback. This combined entity may be referred to as a fashion hot trend, and could be used in rating Zara's products. On the other hand, the increasing popularity of computer games may be relevant to Zara's shipping decision, but only very remotely.

Besides these three organizational dimensions, other techniques can also be used for relating data, especially across different dimensions. For example, simple keyword-based matching or possibly more reliable semantic (similarity) network could also be used.

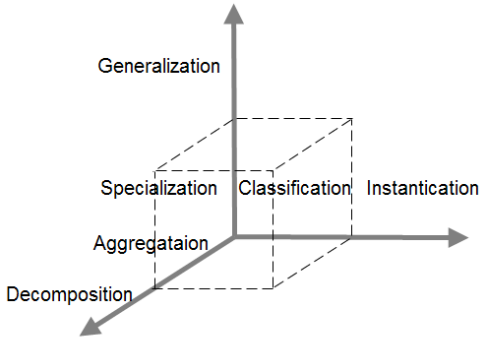


Figure 4.4. Three organizational dimensions

**Prioritization**

This aspect is useful in determining what data to incorporate into the virtual big data, how much efforts should be put on obtaining the data, how much resources to allocate to the data, etc. The priorities of data should reflect the priorities of what the data is intended for. In IRIS, big data is intended for supporting BA, concerning validating potential problems and solutions.

Roughly speaking, the priorities of data are inherited from their respective potential problems and solutions. However, while priorities are propagated downward, the priorities of lower-level refinements (e.g., KPIs, parts of a data model and big queries) can change in either direction. In particular, a tradeoff analysis can lead to the change in the priorities of those operationalizations that overall make strongly negative contributions to achieving higher-level goals – in this case, strongly positive contributions to realizing problems.

Continuous pruning of the potentially large space of problems, KPIs, data models and queries should take place, according to their relative priorities. Refinements/expansion of the (parts of the) data model then should proceed, according to the priorities associated with them – i.e., higher ones before lower ones. There are various prioritization schemes that are available in the literature (see, for example, [56], [57]).

## 4.4 Ontology for Big Data Modeling (GO-BigDM)

Figure 4.5 is an ontology for Big Data Modeling and Table 4.1 shows the definition of ontology and its notation.

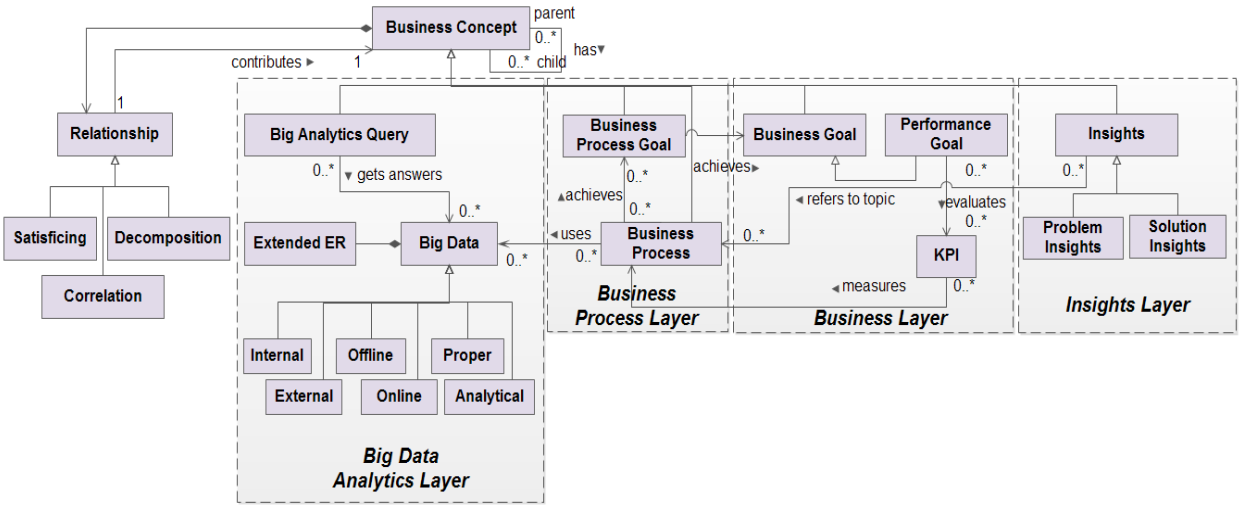


Figure 4.5. GO-BigDM ontology for big data modelling

Table 4.1. GO-BigDM ontology definition and notations

| Name | Definition | | Notation |
|---|---|---|---|
| Big Data | Data which has characteristics of high volume, high velocity, high variety, and high veracity | | |
| | Internal | data generated inside of a business | I |
| | External | data generated outside of a business | E |
| | Offline | data which is not connected with web | Off |
| | Online | data which is connected with web | On |
| | Proper | data which is not processes | P |
| | Analytical | data which is processes for analytics | A |
| Extended Entity Relationship | A notation for modeling entities and relationships, extended with generalization/specialization | | EERD [55] |

## 4.5 IRIS Process for GO-BigDM

In IRIS, the process for developing a big data model for supporting BA consists of four steps, as

depicted in Figure 4.6. This process is not like a water-fall process, but rather incremental, iterative and interleaving.

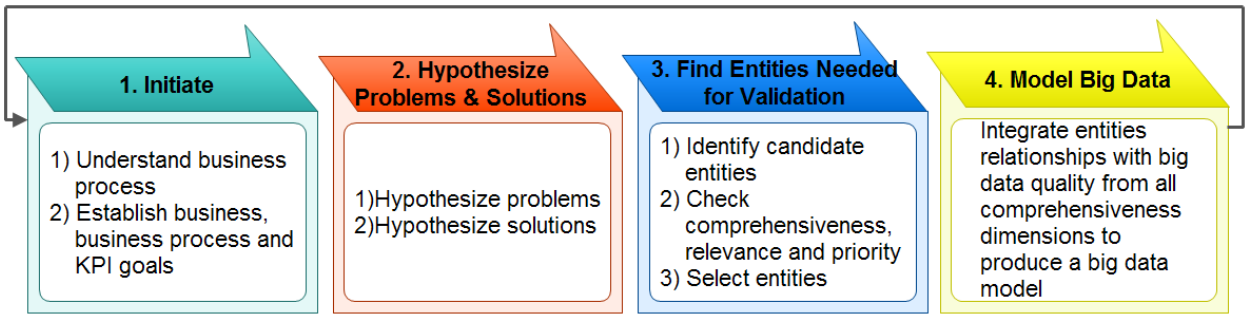


Figure 4.6. GO-BigBM process for big data modelling

1. **Initiate:** This step is for 1) understanding the business process, together with the current data model, which acts as the context for carrying out BA; and 2) establishing business goals, refining them in terms of high-level business process goals, and further refining them in terms of KPI (Key Performance Indicator) goals.
2. **Hypothesize Problems & Solutions:** This is to hypothesize high-level potential problems and their sub-problems (root causes), as well as potential solutions.
3. **Find Entities Needed for Validation:** This is to find candidate entities and relationships between them, which might be needed in validating potential problems and potential solutions, and to select among them. This takes three steps: 1) identify and bring candidate entities from the three comprehensiveness dimensions; 2) check the comprehensiveness, relevance and priority of the candidate entities, according to the potential problems and solutions that correspond to the entities; and 3) select entities and relationships, according to their quality aspects.
4. **Model Big Data:** This is to integrate entities and relationships from every possible dimension to produce a (final) virtual big data model. It involves integrating the selected entities and

relationships from Step 3 into Internal-Offline-Proper. Here, the integration utilizes the three organizational dimensions to find relationships between entities-and-relationships of Internal-Offline-Proper and those from other sources that might need to be refined during integration. Here, the virtual big data model is represented using EERD (Enhanced Entity Relationship Diagram).

After Step 4, big queries would need to be run against a big database, which corresponds to the virtual big data model, but this is beyond the scope of this dissertation.

Throughout the process, alternatives are ranked, according to their relative importance, not only for the entities and relationships in big data model, but also early on for goals, problems, solutions, KPIs and big queries. Also throughout the process, a label propagation procedure [7, 13] can be used to determine the degree to which a lower-level phenomenon affects the upper-level problems and goals the phenomenon is relevant to.

## 4.6 IRIS in Action for GO-BigDM

IRIS's goal-oriented process can help develop a virtual big data model to support BA for Zara's shipping decision process, which was shown earlier in Figure 3.1, via the four steps in Figure 4.6 of the previous section.

**1. Initiate**

Let us suppose that Zara's innovation team wants to use big data analytics through a virtual big data model. So, they understand the business process for making shipment decisions, as earlier shown in Figure 3.1, and the current operating data model, as shown in Figure 4.7, which is used for making these shipment decisions. Afterwards, the team finds and establishes one or more business goals, here, *Increase Global Revenue* (the top portion in Figure 4.8). There are many

ways to increase revenue, including effective marketing and effective shipment decision, which the team considers important, hence treating *Effective [Marketing]* and *Effective [Shipment Decision]* as goals to be achieved. Between these two, the team considers that Shipment Decision is a more critical factor to increasing revenue, so the team wants to first find out if there is any problem with the current shipment decision making process. So, the team refines *Effective [Shipment Decision]* in terms of two business process goals, i.e., *Customer Satisfaction [Shipment Decision]*, and *Minimize Maintenance Cost [Shipment Decision]*. These two goals are expressed in measurable terms, here, KPI goals – *Achieve (Demand Hit Rate > 25%)* and *Achieve (Global Inventory Rate < 10%)* respectively.

As in Figure 4.8, Local Inventory, Sales History, and Weekly Offer entities are needed for determining Local Shipment Quantity. Likewise, Local Shipment Quantity, Sales History, Global Warehouse Inventory and Local Inventory are needed for determining Global Shipment Quantity. The data model is represented using EERD, which essentially is ERD augmented with generalization and specialization. So, data objects in BPMN become entities in ERD. This data model represents only an Internal-Offline-Proper data model from the perspective of a virtual big data model.

**2. Hypothesize Problems and Solutions**

The team now hypothesizes potential problems with the current shipment decision process and their sub-problems (root causes) and potential solutions. It can be done from the perspective of the KPI goals, here *Achieve (Demand Hit Rate > 25%)* and *Achieve (Global Inventory Rate < 10%).* Zara's team considers two potential problems, *Low Hit Rate [Shipment Decision]* and *High Inventory Rate [Shipment Decision]*, as the most likely and important problems. While the team

thinks more deeply about the shipment process, they could conjecture more specific potential problems, such as Time Pressure, Decision by Observation or Exceed Request for Top Selling Item. As with problems, potential solutions can also be hypothesized – here, big data models for Demand and Inventory.

If only the Internal-Offline-Proper data model is used, then two queries, BQ1 and BQ2, might be considered for the purpose of validation (in Figure 4.8). Suppose BQ1 involves more entities than BQ2, hence yielding a more accurate prediction than BQ2. Then, between the queries, the former would have a stronger contribution (++) than the latter (+) towards validating the potential problems they are relevant to, namely, Time Pressure, Decision by Observation and Exceed Request for Top Selling Item.

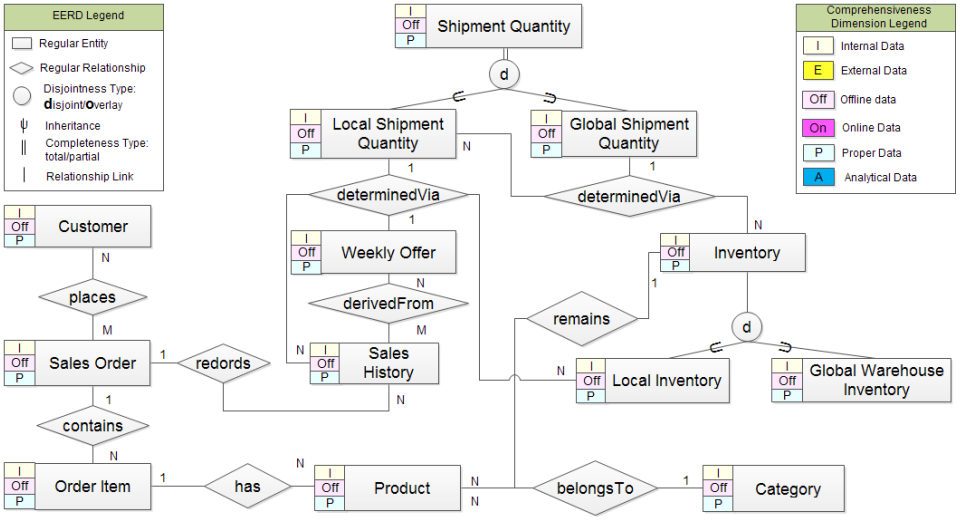


Figure 4.7. Extended Entity Relationship Diagram (EERD) for Internal-Offline-Proper Shipment Data (Step 1)

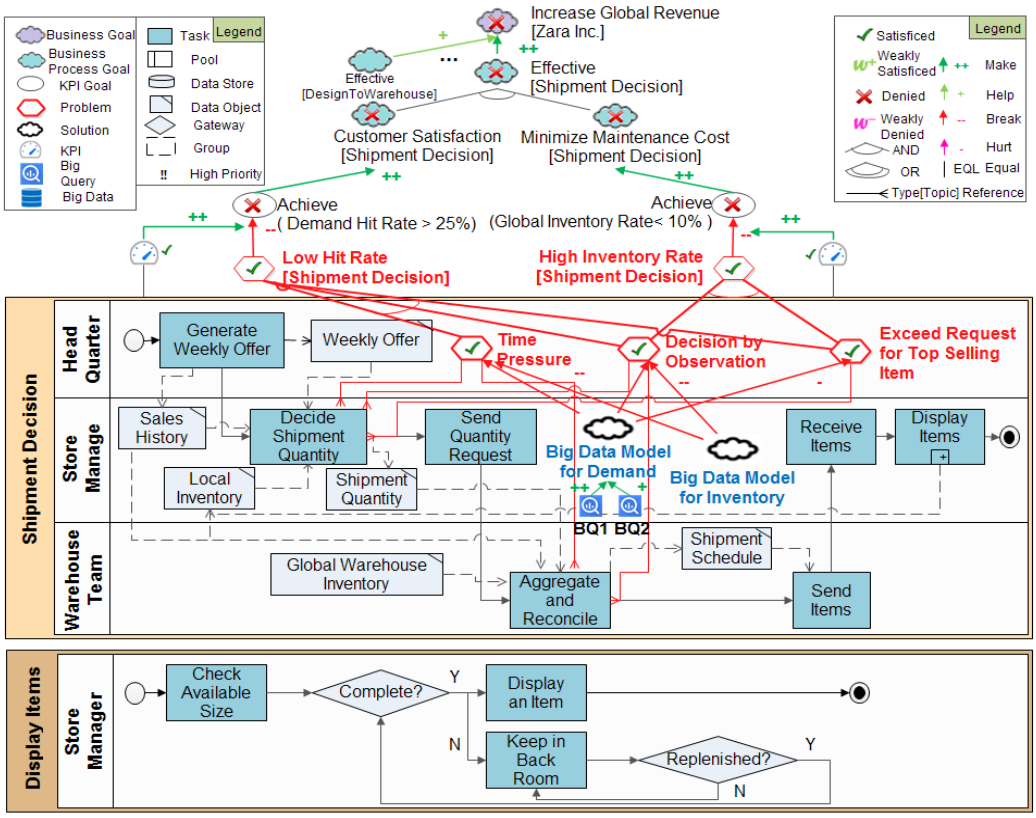


Figure 4.8. Diagnostics for Shipment Decision (Steps 1 & 2)

### 3. Find Entities Needed for Validation

For validating the potential problems, let us assume that Zara's innovation team thinks that only the consideration of the current shipment decision process will be enough. So, some candidates are extracted from the Internal-Offline-Proper Shipment Data Model, in consideration of the current process, as shown on the left-hand side of Figures 4.9 to 4.11. Here, the input-output relationships between entities, concerning the business process tasks they are associated with, are preserved.

On the other hand, let us assume that the team thinks it is important to explore new and external opportunities in validating potential solutions. So, using the Internal-Offline-Proper for validating potential problems as the frame of reference, the team identifies a variety of other types and sources

of entities in the three comprehensiveness dimensions, here External in Figure 4.9, Online in Figure 4.10 and Analytical in Figure 4.11.

For example, the team considers the external Amazon Sales History as being relevant to the internal Sales History (this is Zara's own), in Figure 4.9. Since these two can be subclasses of some higher-level class, say Sales History, what is applicable to one may also be applicable to the other. So, the chain of entities associated with Sales History, here Local Shipment Quantity and Global Shipment Schedule, are copied and associated with Amazon Sales History.

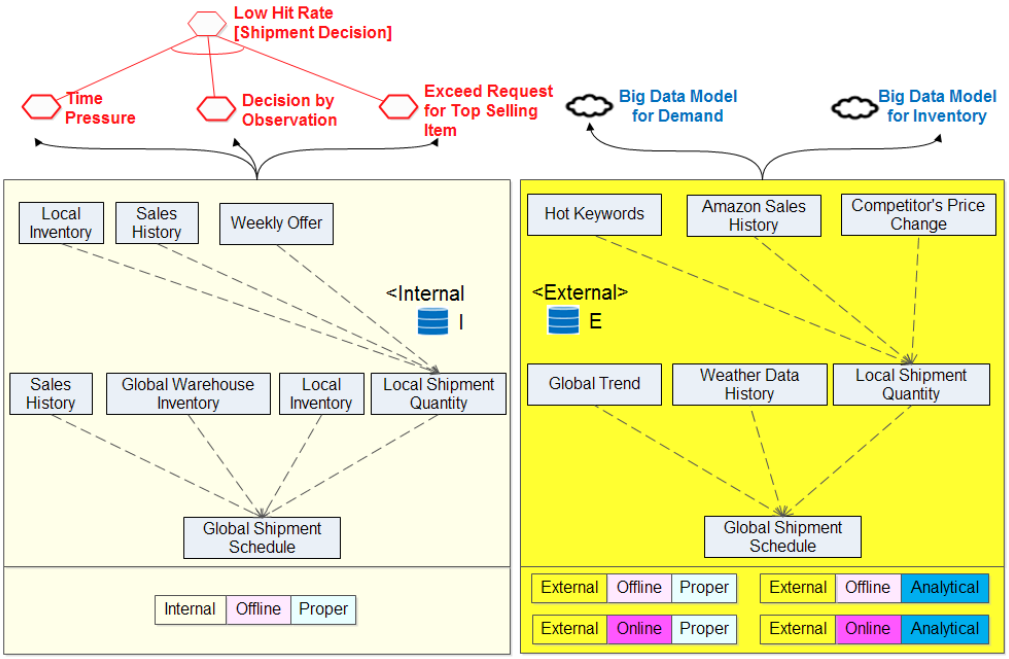


Figure 4.9. Candidate Entities from Internal-External Dimension (Step 3.1)

Similarly, the team considers the offline Sales History (again Zara's own) and the online Online Sales History are relevant to each other, in Figure 4.10, since these two can again be subclasses of some class, say Sales History. So, the chain of entities associated with the offline Sales History are copied and associated with Online Sales History.

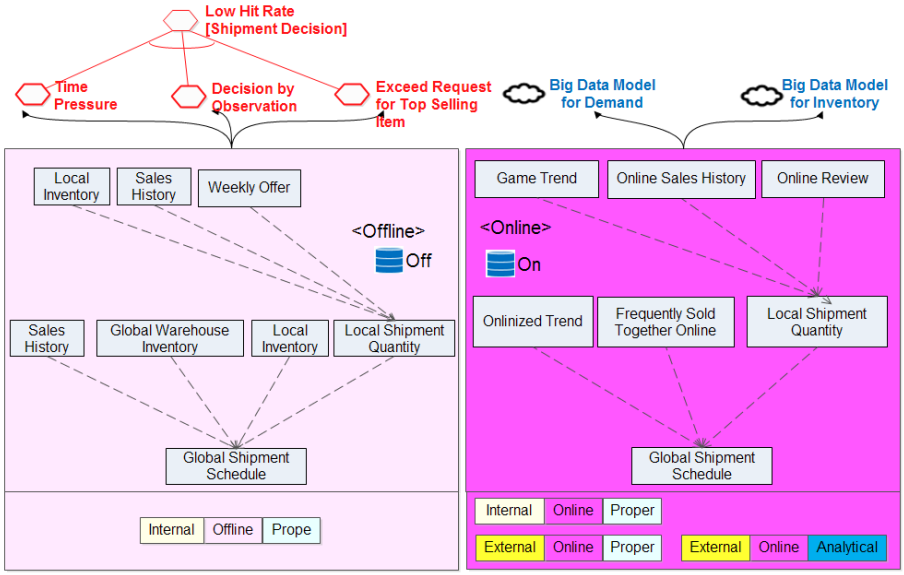


Figure 4.10. Candidate Entities from Online-Offline Dimension (Step 3.1)

Concerning Figure 4.11, the team thinks the Proper Sales History (this is Zara's own) and the analytical Items Frequently Sold Together are relevant to each other, since Items Frequently Sold Together (e.g., hats and scarfs are frequently sold together) can influence Zara's decision on shipping quantities of items that are frequently sold together. Hence, here again, Local Shipment Quantity and Global Shipment Schedule are copied and associated with Items Frequently Sold Together. Also, Global Warehouse Trend is more general, concerning Global Warehouse Inventory, hence likely to be incorporated into a virtual data model later.

An external entity and internal entity are not always relevant to each other, if they cannot share the same superclass or be associated with each other. For example, Competitor's Price Change in Figure 4.9 is not relevant to any internal entity, hence not likely to be useful.

Zara's team now checks the comprehensiveness, relevance and priority of the candidate entities, according to the potential solution that corresponds to the entities, so as to filter the candidate entities to find the most relevant and high priority entities. The team collects, for the potential

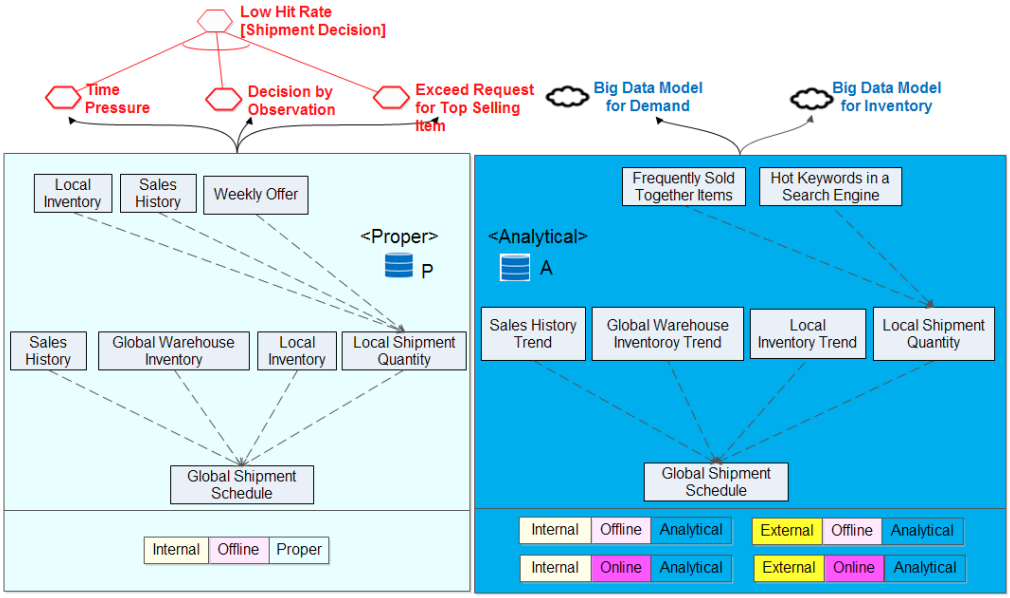


Figure 4.11. Candidate Entities from Proper-Analytical Dimension (Step 3.1)

solution, all the relevant entities – each entity is shown in a row, as in Tables 4.2 and 4.3. Although these tables here are used in validating a potential solution, similar tables can be used in validating other potential solutions and potential problems. Then, for each entity, the team checks the comprehensiveness dimensions the entity or other entities (that it is related to) belong to. For example, Online Feedback has an Internal, Online, Proper attributes. Then, the entity's Relevance is indicated, in terms of its distance and satisficing relevance. Here, the team uses one of H(igh), M(edium) and L(ow), but the ranking scheme can vary according to the characteristics of the particular business process and analytics. For example, in Figure 4.8, Shipment Quantity has 1-distance relevant to the problem with Time Pressure. Now, in the Internal-Offline-Proper Shipment Data model, in Figure 4.8, Local Shipment Quantity, which is a subclass of Shipment Quantity, is derived from Weekly Offer, Sales History and Local History, hence, these four entities are treated as a cluster and 1-distance relevant, whereas Customer is 3-distanc relevant. However Online Game Trend is deemed almost unrelated to any of the internal entities, hence the symbol ∞.

The Priority of the entity is related to the priority of the business goal, business process goal, and KPI goal that the entity is relevant to. The priority scheme again can be determined, as needed. Here, a scale of 0 to 10 is used, where the higher the value, the higher the priority. Zara's innovation team can have a diverse combination of options to be included in the virtual big data model, such as Opt1 (option 1), which includes all those entities that have 1 as their priority value, or Opt2 (option 2) which includes those with either 1 or 2 as their priority value. Here, the heuristic for deciding what entity to include (O) and what not (X) can depend on the number of check (√) symbols, the relevance values and priority values.

Zara's team then selects entities and relationships, according to their quality aspects. The options include Internal-Offline-Proper entities that are needed in validating the potential solution and, more importantly, entities from every other dimension that are relevant to them. This table representation can be thought of as another way, for representing options and their contributions towards achieving big data model qualities, which is complementary to a graph representation, as shown in Figure 4.12. In the graph representation, Cost is additionally shown, which conflicts with the comprehensiveness of big data model quality, hence used in carrying out a tradeoff analysis between the options. The result of this tradeoff analysis is shown to be the selection of option 1.

**4. Build Big Data Model**

Zara's team carries out two discrete activities for building a big data model. Firstly, they integrate those entities and relationships from every possible dimension , which were selected in the previous step, into Internal-Offline-Proper to produce a virtual big data model. Here, the team utilizes the three organizational dimensions to find relationships between entities-and-relationships of Internal-Offline-Proper and those from other dimensions that might need to be refined during

integration. The virtual big data model is represented using EERD, as in Figure 4.13.

For example, Global Shipment Quantity in Figure 4.13 is related to Global Warehouse Trend as an initial relationship. Global Warehouse Trend is shown as a superclass of Global Warehouse Inventory. For another example, Online Feedback and Hot Keywords have some similar traits, so they can be combined together. Then, the resulting composite entity can be used to rate Product. Figure 4.13 shows the final virtual big data model, which consists of existing entities and newly added entities. In this model, some of the Internal-Offline-Proper entities, as well new entities, are associated with a priority value (in green color). However, those entities in the original Internal-Offline-Proper data model that have not yet been involved in validating any potential problem or solution are not, but later may, if they are involved in validating some other problem or solution.

## 4.7 Discussion

Zara's decision-making process for its shipping has been used not only for the purpose of illustrating the key concepts of IRIS's goal-oriented approach to modeling big data, but also for a

Table 4.2. Selection of entities for demand prediction according to their quality characteristics (Steps 3.2 and 3.3)

| Quality Aspects & Entity Candidate Selection Entities for Demand Prediction | Quality Aspects | | | | | | | | | Entity Selection | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Comprehensiveness | | | | | | Relevance | | Priority | Opt1 | Opt2 |
| | Dimension1 | | Dimension2 | | Dimension3 | | DR | SR | | | |
| | Internal | External | Offline | Online | Proper | Analytical | | | | | |
| Online Feedback | √ | | | √ | √ | | 2 | H | 9 | O | O |
| Price Change | √ | | √ | | √ | | 5 | H | 8 | X | O |
| Competitor's Price Change | | √ | √ | | √ | | 5 | M | 7 | X | X |
| Items Frequently Sold Together Online | √ | | | √ | √ | | 3 | M | 7 | X | X |
| Amazon Sales Data | | √ | | √ | √ | | 2 | L | 5 | X | O |
| Online Game Trend | | √ | | √ | | √ | ∞ | L | 7 | X | X |
| Hottest Keyword in Search Engine | | √ | | √ | | √ | 2 | M | 9 | O | O |
| Onlinized Trend | | √ | | √ | | √ | 5 | M | 9 | O | O |
| Local Shipment Quantity | √ | | √ | | √ | | 1 | H | 9 | O | O |

Table 4.3. Selection of entities for inventory prediction according to their quality characteristics (Steps 3.2 and 3.3)

| Quality Aspects & Entity Candidate Selection Entity for Inventory Prediction | Quality Aspects | | | | | | | | | Entity Selection | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Comprehensiveness | | | | | | Relevance | | Priority | Opt1 | Opt2 |
| | Dimension1 | | Dimension2 | | Dimension3 | | DR | SR | | | |
| | Internal | External | Offline | Online | Proper | Analytical | | | | | |
| Onlinized Trend | | √ | | √ | | √ | 5 | H | 10 | O | O |
| Online Sales History Trend | √ | | | √ | √ | | 2 | H | 9 | X | O |
| Global Warehouse History Trend | √ | | √ | | √ | | 2 | H | 10 | O | O |
| Weather Data Analysis | | √ | | √ | | √ | 3 | H | 10 | O | O |

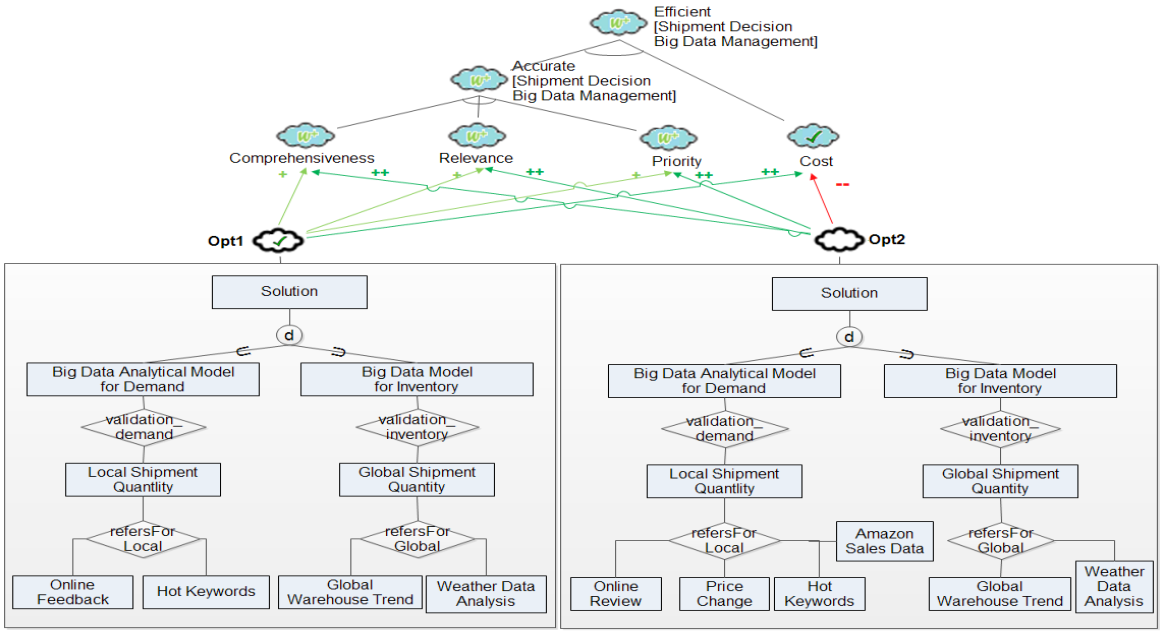


Figure 4.12. Initial Tradeoffs and Selection between Data Entities wrt. Conflicting Goals (Step 3.3.)

(partial) basis of an empirical study. We feel the current study, albeit its small size, shows that IRIS supports the modeling of big data as a service for carrying out business analytics for Zara's shipment decision making.

We also feel that IRIS helps the process of modelling big data to become mostly traceable, if not all, while helping explore and select among alternatives in problems, solutions, business analytics

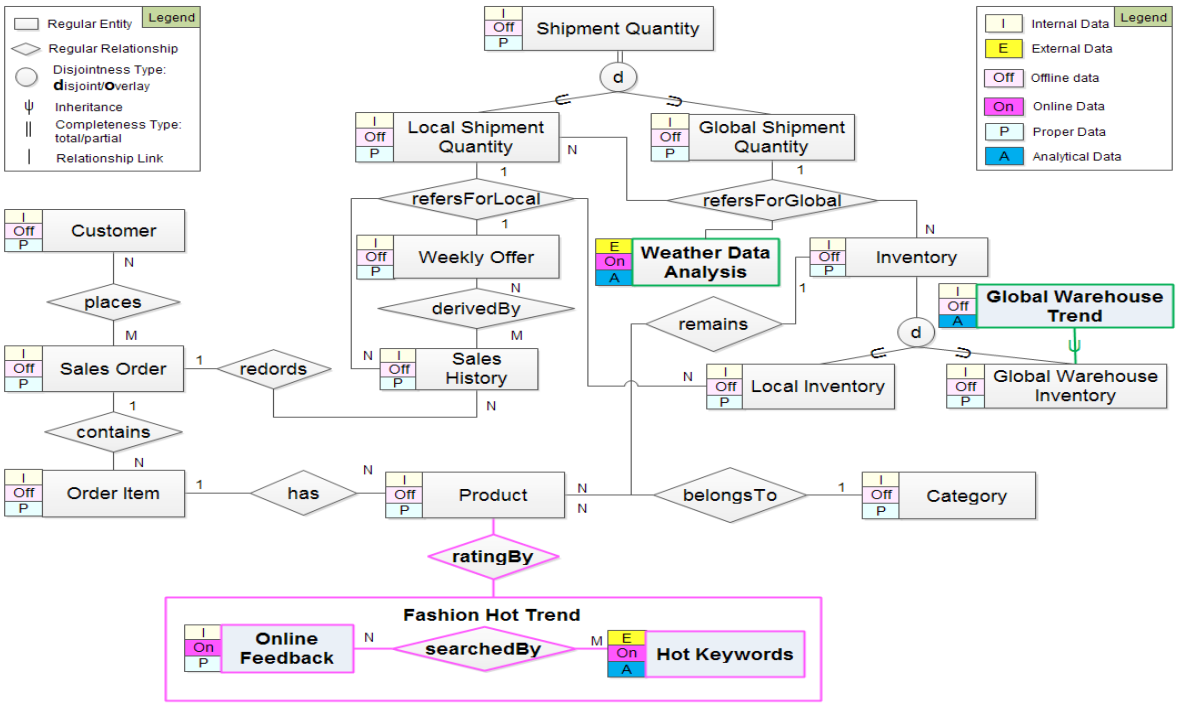


Figure 4.13. Virtual Big Data Model, Consisting of Entities and Relationships from Multiple Sources, for Shipment Decision (Step 4)

and big data models. This, we feel, would help justify, and boost the level of confidence in, the quality of the resulting big data model. However, we have not shown that the potential problems and solutions indeed turn out to be real, key problems and solutions. This would require running big data on a real platform using real data, and, afterwards, monitoring the various real phenomena that are related to either problems or solutions – this seems difficult, if not infeasible in reality.

In this study, we wanted to see if IRIS can help with "connecting the dots" in modeling big data, from two different perspectives: 1) between, on the one hand, business goals, potential problems and solutions, and big data, on the other; and 2) among a variety of different types and sources of data. For the former, we feel that IRIS's approach helped traceably link the various concepts. For the latter, we feel that the three notions of big data quality helped. In particular, the notion of the

comprehensiveness dimension helped consider incorporating newer technologies, including online and external sources of data, such as social networking and online marketplaces, so as not to omit potentially important business opportunities.

Concerning the notion of relevance, among the three notions of big data model quality, measuring the distance between entities and potential problems and solutions was relatively easy with internal data, but not so easy with external data. This was because external data can contain (highly) unstructured data and individual pieces of data, which would need to be first classified, to be related to some internal entities (classes).

IRIS's goal-oriented approach is intended for a systematic and rational process. Our observation in this regard is that this approach helped "connect the dots" among many important concepts, including business goals, potential problems and solutions with a business process, KPIs, analytics, and their alternatives. We thought this should be possible but initially were unsure of how, and now feel that IRIS's approach helps turn this possibility into more of reality.

We feel that IRIS's approach helps hypothesize and validate problems with the business process and solutions, which involves the use of KPIs, which in turn require the use of a big data model. In particular, the three organizational dimensions helped structure and explore data not only in the same but also across different dimensions of comprehensiveness. Additionally, the notion of distance in relevance helped relate data across different dimensions, hence helping avoid omissions or commissions of data.

In the beginning of our empirical study, we had not thought of the use of the three dimensions in exploring potentially relevant data. We thought that they would be useful in organizing a large volume of data only. Now we feel that we have learned that the three dimensions also help

systematically explore potentially useful data.

In order to have a sense of the technical feasibility, in particular, concerning the timely decision making, we ran big queries in Vertica, which is an analytic database management tool and supports both SQL and several types of NoSQL databases. The queries were run on a real big data platform of a consulting company, which maintains a large volume of data for a business whose characteristics are similar to Zara's. We could obtain answers to the queries within seconds.

# CHAPTER 5

# DERIVING USE CASES FROM BUSINESS PROCESSES

With the advances in the Information Technology (IT), use of software systems increasingly has become prevalent and critical in more efficiently and effectively carrying out the various tasks and activities in the rapidly changing business domain.

However, aligning a software system well with its intended business process has been challenging, due to the difficulty in firstly understanding and precisely modeling a business process and secondly coming up with a requirements specification of a system for supporting the business process. Use of well-known notations helps: BPMN (Business Process Model and Notation) for precise modeling of a business process, and UML use case for modeling software requirements; but the second difficulty still remains. There are several issues: 1) Since there do not exist formal definitions of the models, different interpretations are possible which can lead to transformations that deviate from the original meaning; 2) A business activity can be performed either by people or system functionality; 3) The granularity of a use case is not necessarily the same as that of a business activity or task; and 4) Neither BPMN nor UML use cases consider Non-Functional Requirements (NFRs).

In this chapter, we present GO-BP2UC, a goal-oriented framework for transforming software-allocated elements of business processes in BPMN with NFRs into Use Cases with NFRs. GO-BP2UC is intended to facilitate the exploration of alternatives for dealing with multiple interpretations of the same elements in a business process, and selection among the alternatives through a trade-off analysis, using similarity and granularity as the selection criteria, in spirit of other general goal-oriented approaches (e.g., [6, 7, 8, 9]). Similarity is further refined into

Ontological- and Contextual- Similarity. Ontological Similarity means how well concepts in a source model are matched to concepts in a target model, which can be evaluated by using a taxonomy of ontologies. Contextual Similarity means how well the contextual information – allocation and granularity of a concept in a source model – is reflected into a target model, which can be automatically transformed by rules with conditions. Granularity means the size of a unit of target element including single or clustered ones. As for the BPMN and Use Case Models, we augment them with NFRs. Also, an Intermediate Model (IM) is used, which is composed of intermediate entities and relationships, for helping to deal with the ontological gap and the many-to-many relationships between the source (i.e., BPMN augmented with NFR) and the target (i.e., Use Case Model augmented with NFR). To support the transformation process, an assistant tool is implemented as a proof of concept.

Section 2 describes related work. Section 3 presents Go-BP2UC – a goal-oriented framework for transforming business processes in BPMN into Use Cases. Section 4 shows the use of Go-BP2UC in a study of insurance quote flow, while Section 4 introduces a supporting tool. Section 5 describes an evaluation.

### 5.1 Related Work

Our transformational approach adopts the notion of MDA (Model Driven Architecture), which is an approach to producing and maintaining systems by fully or partially automating the transformation from a high-level abstraction model into an executable information systems [58]. The automation includes Computation-Independent Model (CIM) for domain models, Platform-Independent Model (PIM) for logical system models, Platform-Specific Model (PSM) for

implementation models, and transformations between them. In our case, a business process model in BPMN [22] belongs to CIM, and a use case model in UML Use Case [59] to PIM.

The idea of transforming a business process in BPMN to Use Cases has previously been used [60, 61, 62, 63], although previous work deals with only one-to-one transformations of the functional aspect, and also without considering the contextual information. Our approach also takes a goal-oriented approach, whereby the transformation can be more rationally carried out in consideration of more alternatives in Intermediate Model. That is, business process activities such as Task can be condensed or Sub-process can be split into Clustered Business Activity Units, together with appropriate intermediate relationships, and a selection will be made using Similarity and Granularity as the criteria. When it comes to transformation rules, patterns have been used for extracting use cases [62, 64, 65, 66], but without considering the contextual information; In our approach, rules take into account the environmental context such as the system allocation information and the granularity of a source element, not only for the functional but also non-functional sides. As for the non-functional side, transformation rules are also proposed for transformation a BPMN model augmented with security goals in [67], but our rules are for any types of non-functional goals and requirements. Go-BP2UC can be considered complementary to transforming a BMPN into a BPEL executable [22], in the sense that Go-BP2UC is for defining the "what" of a software system, while translation into a BPEL in terms of a sequence of executable statements is about the "how" of a software system, and Go-BP2UC facilitates handling multiple possible interpretations of BPMN elements, while also dealing with non-functional goals.

## 5.2 Overview

Go-BP2UC adopts a goal-oriented approach to transform business processes in BPMN with NFR into Use Cases with NFR, which facilitates rational decision making through the exploration of alternatives, for dealing with multiple possible transformations of business processes, and selection among the alternatives through a trade-off analysis. One of the reasons for this approach is because of the absence of complete formal semantics for either BPMN or Use Case, multiple different interpretations are possible for the same transforming element. Explorations of alternatives help deal with such multiplicity. This is important since there is an ontological gap between BPMN process concepts and UML Use Case concepts. The ontological gap includes differences in the granularity of a BPMN Task, which is an atomic activity that is included within a Process [22], and of a Use Case, which is a set of complete actions performed by a system yielding an observable result for one or more actors [59]. This means that one or more than one Task can be transformed into one Use Case. Another important kind of ontological gap has to do with relationships, such as “Include”, “Extend” and “Generalization” in Use Case, which BPMN does not offer. The other reason is that business activities are done by human or systems, thus, all elements in business processes can be transformed, but sometimes not, which means business activities are allocated to a supporting system.

Go-BP2UC also helps deal with non-functional requirements, by considering augmentation of both BPMN processes and UML Use Cases with non-functional descriptions which are enabled by NFR integrated models respectively. For reference, Go-BP2UC handles two categories of non-functional goals – one for an application-specific business process model, such as Speed, and the

other for the transformation process, including Similarity and Granularity. Figure 5.1 depicts the overall Go-BP2UC process.

As Figure 5.1 shows, first of all, a software engineer select elements in business process in BPMN augmented with NFR as software-allocated elements. Then, they are transformed into alternative Intermediate Models, in which related elements are clustered into Clustered Business Activity Units (CBPUs) and intermediate relationships are established. That is why these Intermediate Models can help deal with the differences in the ontology and granularities between the BPMN and Use Case. One Intermediate Model is selected among the alternatives, if it satisfices Similarity and Granularity, and then each element or a combination of elements in the Selected Intermediate Model can be transformed into a Use Cases augmented with NFR, with the help of transformation rules which reflect ontological- and contextual- similarity as conditions implemented by QVT (Query-View-Transformation). Transformation rules cover the diverse transformation possibilities of not only entity elements, but also relationships, in consideration of the context information i.e., system allocation or granularity for both functional and non-functional elements.

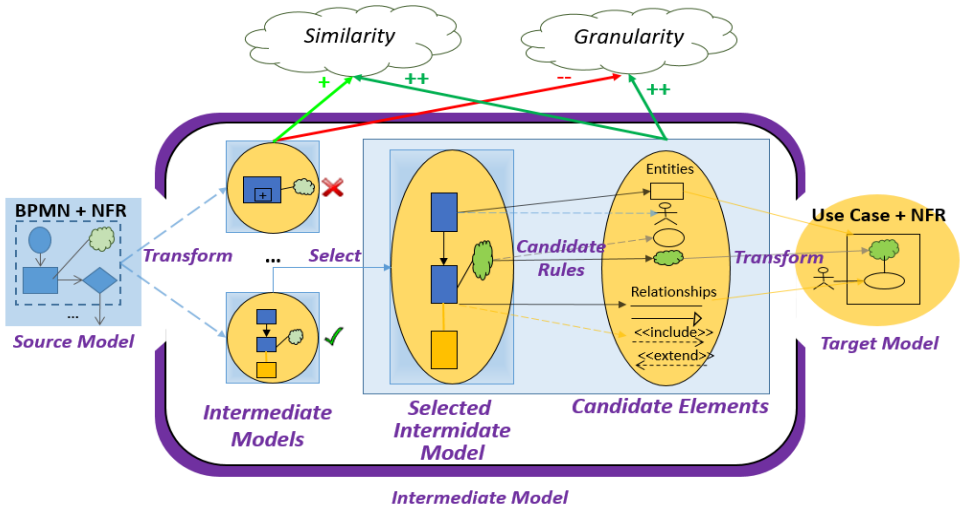


Figure 5.1. Overview of the GO-BP2UC Transformation Process

## 5.3 Ontology of Intermediate Model for GO-BP2UC

Ontology is explicitly specifiable key concepts and their relationships in a knowledge domain that analyzers want to represent [68]. In this framework, we define ontologies for Intermediate Model which is in between NFR augmented- BPMN and Use Case Model, and Figure 5.2 and Table 5.1 show ontology for the Intermediate Model in the form of a metamodel. Although the extended-BPMN and Use Case Model are available, we will show them in section 6.4, a supporting tool.

The ontologies of Intermediate Model consist of two parts, functional- and non-functional- sides. In the functional side, *Clustered Business Activity Unit (CBAU)* which can be transformed into a *Use Case*, *Actor*, *System Boundary*, or a combination of them is a complete set of actions that can be performed by a system as a cohesive unit for clustering related Business Activities and is used to coherently cluster BPMN elements. The notions of *Primary Participant* and *Secondary Participant* both correspond to the notion of *Participant* including *Pool* and *Node* in BPMN; among these two kinds, only a *Primary Participant* will be associated with a *Use Case* in the target model. Regarding the functional side relationships, *Flow Relationship* is a simplified relationship of *Sequence Flow, Data Flow* and *Message Flow* in BPMN. Interestingly, *Condition Relationship* is a set of new intermediate relationships that are not in BPMN for representing a conditional relationship between *CBAUs* which bridges BPMN model and Use Case Model.

Similarly, on the non-functional side, *Clustered Softgoal* plays a role for clustering and bridging gaps between business level- and system level- *Softgoals*. The non-functional notation for the Intermediate Model is adopted from [7]. *Clustered NFR softgoals* represent non-functional goals or requirements to be satisficed (i.e., in a good-enough sense), *Clustered Operationalizing Softgoals* are functional requirements for satisficing NFRs, where a *Clustered Softgoal* can be

expressed as Type [Topic]. While type can represent non-functional part, Topic can do functional part.

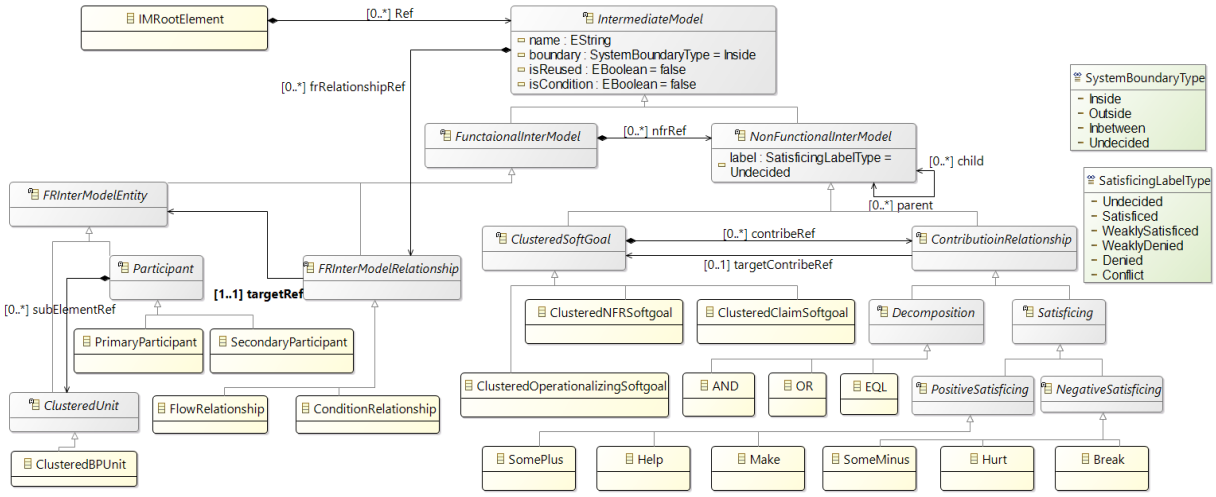


Figure 5.2. A metamodel for the Intermediate Model

Table 5.1. Ontology for the Intermediate Model

| Category | Ontology Name | Definition | Notation |
|---|---|---|---|
| Functional Side | Clustered Business Activity Unit (CBAU) | a complete set of actions that can be performed by a system as a cohesive unit for clustering related Business Activities | |
| | Primary Participant | a participant who directly interacts with a target software system | |
| | Secondary Participant | a participant who indirectly interacts with a target software system | |
| | Flow Relationship | a relationship that has the flow from source to destination | |
| | Condition Relationship | a relationship that represent a condition between CBAUs such as inclusion, extension, generalization or qualification | |
| Non-Functional Side | Clustered Softgoal | a group of NFR Softgoals close together | |
| | Clustered Operationalizing Softgoal | a group of Operationalizing Softgoals to satisfice Clustered NFR Softgoals | |
| | Contribution Relationship | positive or negative relationship toward a parent Clustered Softgoal | ++ / -- |

## 5.4 Transformation Method

Transformation method is intended to help ensure that the target Use Case Model is semantically similar to the original BPMN. In Go-BP2UC, transformation is not entirely automatic but involves humans, here the term requirements engineer, in the method. The goals of the transformation process are Similarity and Granularity.

**Step 1. Set Transformation Contextual Information:** During the transformation process, the requirements engineer provides contextual information for dealing with alternative transformation possibilities. For Ontological Similarity, contextual information concerns whether a Pool/Node is a Primary Participant or a Secondary Participant, which is used to determine the type of the target actor (here, respectively, Primary Actor or Secondary Actor), and whether an element in a Pool/Node is Target System Inside or Outside (here, respectively, a system or environment). For granularity, contextual information also considers the degree of preferred granularity, High, Medium and Low. In high granularity, a sub-process is a transformation unit and in low granularity, a task is, and in medium, both a sub-process and a task will be considered as a transformation unit. For a NFR Softgoal of a process, it is needed to make sure if a NFR Softgoal has one or more Operationalizing Softgoal which is a concrete mechanism to achieve the NFR Softgoal.

**Step 2. Build Intermediate Models:** For Ontological- and Contextual- Similarity on transformation, a diverse set of Intermediate Models can be considered for the source, in terms of the intermediate functional and non-functional concepts in Table 5.1. For bridging the gap between the source and the target ontologies, including differences in their semantics and granularities, *Clustered Business Activity Units (CBAUs*) are used as collections of inter-related elements, such as *Sub-process*, *Tasks,* or *Groups* by a role. A *CBAU* is similar to the concept of "step" in [69],

where a step refers to a sequence of *Tasks* that can be performed without interruption by the same role (e.g., the same *Node* or *Pool*), but more flexible than the "step", because, in Go-BP2UC, *CBAUs* can be merged or split into other *CBAUs*, in a manner to satisfice high *Similarity*. A transformation unit can be transferred into Intermediate Models using rules in List 1, which consist of *CBAUs*, by using two kinds of relationships, Condensed Subsumption and Division.

- **Condensed Subsumption** means that ontologically co-related elements is subsumed to a CBAU in a property preserving manner. It is similar to "isA", but, unlike "isA", the members of a *CBAU* are not necessarily homogeneous.
- **Division** means that a business process element splits into several *CBAUs*. For example, a *Sub-process* in BPMN can be divided into more than one *CBAUs.*

In this step, *Condition Relationship* can be used as part of an Intermediate Model to help identify the target relationships in Use Case, such as *Include, Extend, or Generalization*, that BPMN does not have.

As for non-functional elements, if a functional source element subsumes another functional source element, where the subsuming element acts as the Topic of a *Clustered Softgoal* (i.e., in its Type [Topic] expression) in an Intermediate Model.

**Step 3. Select an Intermediate Model:** The requirements engineer selects an Intermediate Model among multiple Intermediate Models, using Ontological- and Contextual- *Similarity* and Intended *Granularity* as the selection criteria in this stage.

Selected IM = $IM_i$, where $i = \max_{1 \le i \le n}$( $Sim_{ontology}$(Source, $IM_i$) + 1.5*$Sim_{granularity}$(Granularity($IM_i$), Intended Granularity) (1)

where n is the number of Intermediate Models(IM), and Granularity ∈ {High, Medium, Low}.

As for the Ontological Similarity, we can use the formula (3) which utilizes (2) from [70].

Let a taxonomy, which classifies all the ontologies in BPMN, Intermediate Model, Use Case Model and NFR Framework in Figure 5.3, be augmented with a function p : C → [0, 1] , such that for any c ∈ C, P(c) is the probability of encountering an instance of a concept c according to [70]. In it, Similarity is defined as followings:

$$Sim(c1, c2) = \max_{c \in S\,(c1,c2)} [-\log P(c)], \qquad (2)$$

where S(c1, c2) is the set of concepts that subsume both c1 and c2. This means that if c1 is-a c2, c1 is a subclass of c2, then p(c1) ≤ p(c2) and if the probability increases, its informativeness decreases. Thus, the more abstract a concept, the lower its information content. By utilizing formula (1) and (2), we formally define Ontological Similarity as followings.

$$\text{Sim}_{\text{ontology}}(\text{Source, Target}) =$$

$$\frac{\sum_{i=1}^{n} Sim\ Concept\ (ei,ei')}{N} + \frac{\sum_{j=1}^{m} Sim\ Relationship\ (rj,rj\prime)}{M} + \frac{\sum_{i,j=1}^{n} Sim\ Concept\ (ei,rj')}{N}, \quad (3)$$

where ei ∈ Source Entity , rj ∈ Source Relationship and ei´ ∈ Target Entity , rj´ ∈ Target Relationship. In this formula, each concept and relationship in source and target are compared, and also concept in source and relationship in target are compared.

$$\text{Sim}_{\text{granularity}}(\text{Granularity}(IM_i), \text{Intended Granularity}) = \{1, 0.5\}, \qquad (4)$$

if Granularity($IM_i$) = Intended Granularity then 1, else 0.5.

Figure 5.3 shows a taxonomy according to the intended aims of an element in BPMN, Intermediate Model, Use Case Model and NFR Framework in order to measure Similarity. This taxonomy can be differently defined, so it needs agreement, but in this chapter, we defined as Figure 5.3 as an example.

**Step 4. Automatically Select Adequate Transformation Rules for the Selected Intermediate Model to Use Case:** In this step, the (selected) Intermediate Model is transformed into Use Case Model with NFR. While the source model is transformed by using rules in List 1, diverse alternatives rules can be considered and the selection criteria is Ontological- and Contextual-Similarity. In List 1, the second part shows examples of transformation rules from IM to Use Case Model with NFR for both functional and non-functional elements.

In the rules, contextual information such as primary /secondary participant, system inside/ outside, which condition relationship, the number of reused module, existence of non-functional goal can be expressed with condition in a rule. When the conditions are met, the rule will be selected automatically.

**Step 5. Transform from Intermediate Model to a Target Model:** Using the selected rules, transformation will occur from Intermediate Model to a Target Model, i.e., NFR augmented Use Case Model.

## 5.5 Transformation Rules

List 1 shows some examples of transformation rules from BPMN with NFR to Intermediate Model and from Inter Mediate Model to Use Case Model with NFR for both functional and non-functional elements which are implemented by QVT.

Entity elements, as well as relationships, can be transformed into a diverse set of target elements, in consideration of the appropriate contextual information. Concerning rules for a functional element, Clustered Business Activity Unit (CBAU) can be transformed into a Use Case, Actor or even subsumption by another Use Case, depending on where the CBAU is located in the BPMN process.

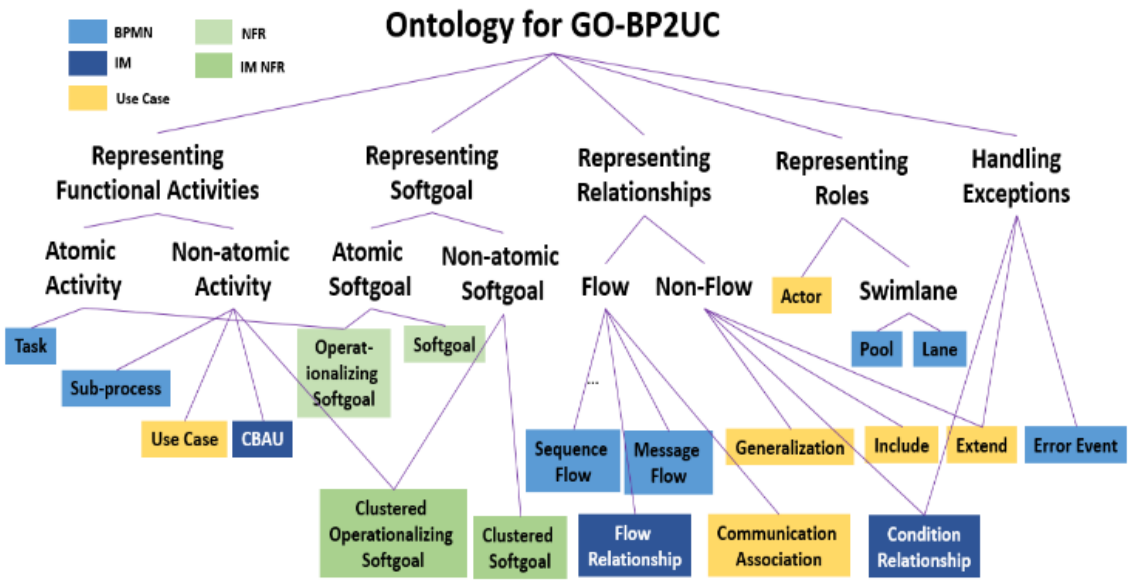


Figure 5.3. A taxonomy for ontology of GO-BP2UC

If a CBAU is performed (contained) by a Primary Participant and it is Target System Inside, it can be transformed into a Use Case. However, if it is performed by a Secondary Participant and Target System Outside, the CBAU can be subsumed by the Actor which corresponds to the Secondary Participant. Concerning rules for non-functional elements, NFR Softgoals should be operationalized in terms of functional services, hence getting transformed into Clustered Operationalizing Softgoal, then into a Use Case.

List 1: Examples of Transformation Rules Implemented using QVT

```
**********************************************************************************************************************************************************************
An Example of Transformation Rules for BPMN Model Augmented with NFR to Intermediate Model
**********************************************************************************************************************************************************************
abstract mapping Pool::Pool2BaseParticipant():Participant {          name := self.name; boundary := self.boundary;}
mapping Pool::Pool2Participant():Participant disjuncts Pool::Pool2PrimaryParticipant, Pool::Pool2SecondaryParticipant{}
mapping Pool::Pool2PrimaryParticipant():PrimaryParticipant inherits Pool::Pool2BaseParticipant when {self.boundary = SystemBoundaryType::Inside;}
{              subElementRef := self.objectsRef[Activity]-> map Activity2ClusteredBPUnit();}
mapping Pool::Pool2SecondaryParticipant():SecondaryParticipant inherits Pool::Pool2BaseParticipant when {self.boundary = SystemBoundaryType::Outside;}
{              subElementRef := self.objectsRef[Activity]-> map Activity2ClusteredBPUnit(); }
abstract mapping OperationalizingSoftgoal::Softgoal2BaseClusteredSoftgoal():ClusteredSoftGoal {          name := self.name;          boundary := self.boundary;  }
mapping OperationalizingSoftgoal::Softgoal2ClusteredSoftgoal():ClusteredSoftGoal
```

*disjuncts OperationalizingSoftgoal::Softgoal2ClusteredOperationalizingSoftgoal, OperationalizingSoftgoal::Softgoal2Dummy{}*

*mapping OperationalizingSoftgoal::Softgoal2ClusteredOperationalizingSoftgoal():ClusteredOperationalizingSoftgoal*

*when {self.boundary = SystemBoundaryType::Inside and self.label = SatisficingLabelType::Satisficed} {name := self.name; label := self.label; boundary := self.boundary;}*

*mapping OperationalizingSoftgoal::Softgoal2Dummy():Dummy when {self.boundary = SystemBoundaryType::Outside and self.label = SatisficingLabelType::Satisficed}*

*{name := self.name; }...*

**********

***An Example of Transformation Rules for Intermediate Model to Use Case Model Augmented with NFR***

**********

*abstract mapping ClusteredBPUnit::ClusteredBPUnit2BaseUMLElement():UMLElement { name := self.name; }*

*mapping ClusteredBPUnit::ClusteredBPUnit2UMLElement():UMLElement disjuncts ClusteredBPUnit::ClusteredBPUnit2UMLElement, ClusteredBPUnit::ClusteredBPUnit2Actor*

*{}*

*mapping ClusteredBPUnit::ClusteredBPUnit2Usecase():UseCase inherits ClusteredBPUnit::ClusteredBPUnit2BaseUMLElement when {self.boundary =*

*SystemBoundaryType::Inside;}{}*

*mapping ClusteredBPUnit::ClusteredBPUnit2Actor():Actor inherits ClusteredBPUnit::ClusteredBPUnit2BaseUMLElement when {self.boundary =*

*SystemBoundaryType::Outside;}{}*

*mapping ClusteredOperationalizingSoftgoal::Softgoal2ClusteredOperationalizingSoftgoal():UseCase*

*when {self.boundary = SystemBoundaryType::Inside and self.label = SatisficingLabelType::Satisficed}*

abstract mapping ConditionRelationship::ConditionReltionship2BaseDiretedRelationship():DirectedRelationship {name := self.name;}

mapping ConditionRelationship::ConditionRelationship2Relationship():DirectedRelationship

disjuncts ConditionRelationship::ConditionRelationship2Include, ConditionRelationship::ConditionRelationship2Extend {}

mapping ConditionRelationship::ConditionRelationship2Include():Include inherits ConditionRelationship::ConditionReltionship2BaseDiretedRelationship

when {self.boundary = SystemBoundaryType::Inside and self.targetRef.isReused= true}{}

mapping ConditionRelationship::ConditionRelationship2Extend():Extend inherits ConditionRelationship::ConditionReltionship2BaseDiretedRelationship

when {self.boundary = SystemBoundaryType::Inside and self.isCondition= true}{} ...

**********

## 5.6 GO-BP2UC IN ACTION

In this section, we show a study of a business process for an insurance quote flow – an adaptation of the one in [71], with some simplification but also with an extension with non-functional goals. Figure 5.4 shows that, when a Prospect wants to contract an Insurance Company, an Agent makes a Quote for the Prospect. The Agent receives Lead which is information about a new customer, and makes an appointment to gather more details.

Once the Agent completes the personal information, the system will calculate the total quote for the Prospect. If the Quote is adequate, the Prospect will sign; Otherwise, he/she will decline and explore other options (here, only a normal case). Finally, a formal policy will be made and given

to the Prospect. Collecting missing information can be time-consuming, so Speed is added as a non-functional requirement to be addressed. Figure 5.4 shows the whole business process, in terms of both functional and non-functional requirements.

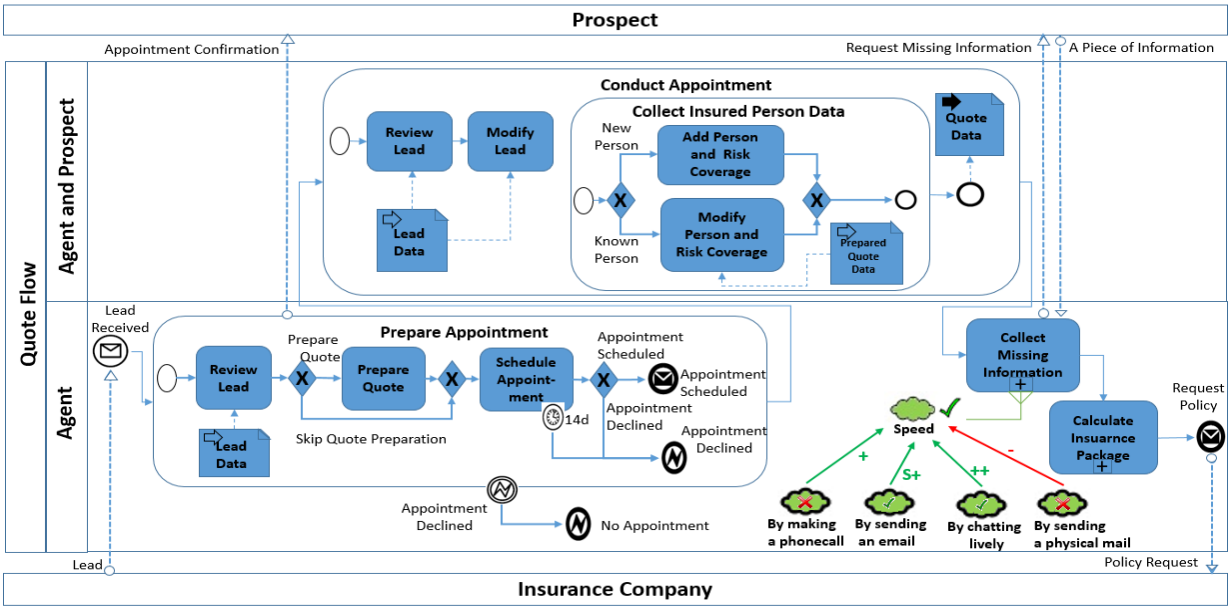


Figure 5.4. An Insurance Quote Flow Business Process in BPMN

**Step 1. Set Transformation Contextual Information:** As mentioned in Section 5.4, a software engineer should initially provide contextual information to the transformation engine. For example, there are three Participants/Pools – *Prospect*, *Quote Flow*, and *Insurance Company*, and in Quote Flow, there are two Nodes – *Agent* and *Agent and Prospect.* Among them, Quote Flow which includes *Agent* and *Agent and Prospect* is Primary Participant since it directly uses the target system. Others are Secondary Participant. Additionally, what business activities will be Inside or Outside of the target system should be specified. In this example, all the activities in Quote Flow are set as System Inside, but we will explain the case that Sub-Process(Conduct Appointment) is not allocated in the next step5. For the degree of preferred granularity, we assume that Medium is

set. Before the transformation, Speed (NFR Softgoals) are operationalized into *By sending an email* and *By chatting lively* (Operationalizing Softgoals).

**Step 2. Build Intermediate Models:** Once the contextual information is set, as in Figure 5.5, the source model is transformed into diverse Intermediate Models (IMs). For understanding, IM1 shows when the granularity is set Low and Model2 and 3 is set High and Medium respectively. In IM1, each Task in each Sub-Process in the source model becomes a Clustered Business Activity Unit (CBAU) respectively without any *Condensed Subsumption* or *Division*. In IM2, each Sub-Process becomes a CBAU. In IM3, as the granularity is set as Medium, a transformation unit will be flexible. A Sub-Process, *Prepare Appointment*, is divided into two CBAUs, *Review Lead* and *Prepare Appointment* with a Condition Relationship with inclusion condition because *Review Lead* is a reused activity in *Conduct Appointment*. The *Prepare Appointment* is again divided into two CBAUs, *Prepare Appointment* and *Cancel Appointment*, for handling a normal case and an exceptional case respectively. The two CBAUs are connected to each other through a Condition Relationship, *Appointment Declined* as in Figure 5.5. Other activities are similar.

**Step 3. Select an Intermediate Model:** A selection will be made using the formula (1) in Section 5.4. To simplify the calculation of Similarity, we use relative distances between ontologies in Figure 5.3.

Selected IM = $IM_{i,}$ where $i = \max_{1 \le i \le 3}$( $Sim_{ontology}$ (QFBP, $IM_i$) + 1.5*$Sim_{granularity}$ (Granularity($IM_i$), Intended Granularity)).

Quote Flow Business Process (QFBP) and $IM_i$ consist of many sub-elements, we can calculate $Sim_{ontology}$ (QFBP, IMi) as average of each transformation. For convenience, let's think about Sub-Process (Prepare Appointment) in Figure 5.4 as $QFBP_1$.

**IM1** = (Sim $_{Concept}$(Task(Review Lead), CBAU(Review Lead)) + Sim $_{Concept}$(Task(Prepare Quote), CBAU(Prepare Appointment)) + Sim $_{Concept}$(Task(Schedule Appointment), CBAU(Schedule Appointment)) / 3 + (2*Sim $_{Relationship}$(Sequence Flow, Flow Relationship))/2 + 0 + 1.5*0.5= -log((4+4+4)/3 + (2*2)/2+0) + 0.75= -log(6)+ 0.75 ≈ -0.78 + 0.75 = -0.03.

**IM2** = Sim $_{Concept}$ (Sub-Process(Prepare Appointment), CBAU(Prepare Appointment))/1 + 0 + 0 + 1.5*0.5 = -log(2)+0.75 ≈ −0.3+0.75 = 0.45.

**IM3** = (Sim $_{Concept}$ (Sub-Process(Prepare Appointment), CBAU(Prepare Appointment)) + Sim $_{Concept}$ (Sub-Process(Prepare Appointment), CBAU(Review Lead) + Sim $_{Concept}$ (Sub-Process(Prepare Appointment), CBAU(Cancel Appointment))) / 3 + (Sim $_{Relationship}$(Sequence Flow, Flow Relationship) + Sim $_{Relationship}$(Sequence Flow, Condition Relationship(null))/2 + Sim $_{Concept}$ (Error Event(Appointment Declined), Condition Relationship (Appointment Declined))/1 + 1.5*1 = -log(4.5)+1.5 ≈ 0.85.

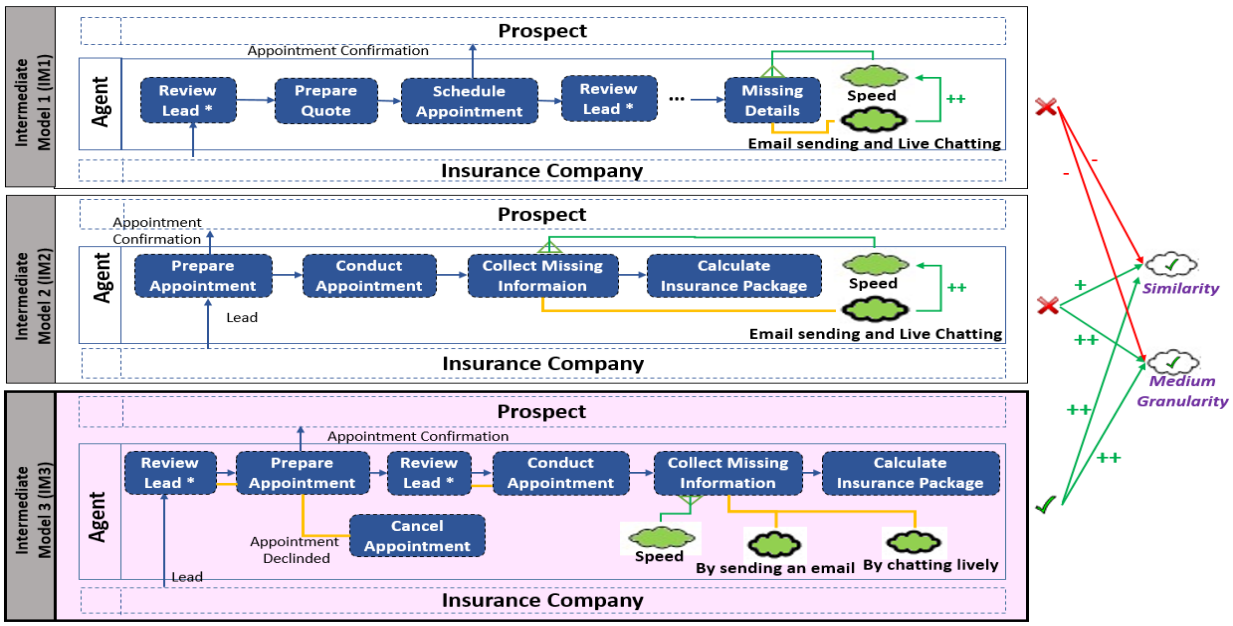


Figure 5.5. Candidate Intermediate Models mapped from an Insurance Quote Flow Business Process with Medium Granularity

According to the calculation, IM3 is selected. This result is also applied to the whole business process in Figure 5.5.

**Step 4. Automatically Select Adequate Transformation Rules for Elements of the Selected Intermediate Model:** Each element of the Selected Intermediate Model can be transformed into different Use Case elements. Figure 5.6 shows examples of consequences of candidate transformation from Intermediate Model to Use Case Model, where the rule is selected from the transformation function in the bottom line of List 1.

For example, No.1 in Figure 5.6, two CBAUs (Prepare Appointment and Cancel Appointment), which are connected with Appointment Canceled (Condition Relationship), can be transformed into two Use Cases (Prepare Appointment with Appointment Declined Extension Point and Cancel Appointment) and Extend Relationship. Alternatively, the part can be transformed into the two Use Cases with Include Relationship. In this case, Ontological Similarity from formula (3) is the same because they have the same probability in the Taxonomy in Figure 5.3. However, Contextual Similarity is different. For example, Condition Relationship in the Intermediate Model has a condition, i.e., Appointment Declined. Among the alternatives in List 1 from Intermediate Model to Use Case Model, the rule "Is_Condition is true and System_Boundary is true" is selected. This selection is done automatically.

Similarly, No.2 in Figure 5.6, Review Lead and Conduct Appointment is associated with a Condition Relationship which has no condition. This can be transformed into two Use Cases (Review Lead and Conduct Appointment) with Extend or Include. When we consider the Contextual Similarity, since Review Lead appears two times in the whole IM1 and is system inside, Include is more adequate than Extend. Consequently, Include Relationship is chosen.

No3. Shows how the NFR Softgoals can be transformed into Use Case Model augmented with non-functional requirements. In Figure 5.3, Collect Missing Information has a NFR Softgoal, .i.e., Speed, for which we introduced Send an email or By chatting lively, and Collect Missing Information. The Condition Relationship can be transformed into Include, Extend, Generalization, or Association. We chose the Generalization relationship since the Operationalizing Softgoals are specific mechanism to achieve their original NFR Softgoal, Save Time as shown in Figure 5.6.

**Step 5. Map from Intermediate Model to a Target Model:** After the rule selection, the transformation will be done from Intermediate Model to Use Case Model augmented with NFR using the selected rules. In Figure 8.1 in Chapter 8, the Use Case Diagram shows the result of the transformation. If Sub-Process(Conduct Appointment) is not allocated to the target system Use Case(Conduct Appointment) and Use Case(Review Lead) will not be included in this result diagram, which are represented with blurred images.

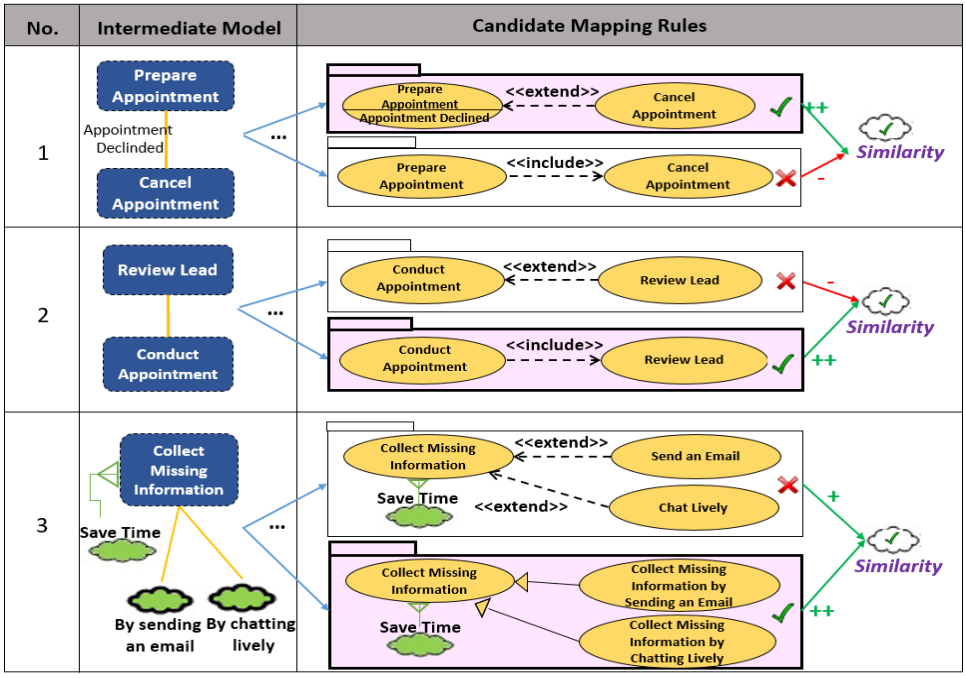


Figure 5.6. Examples of candidate transformation rules

# CHAPTER 6

# IRIS ASSISTANT TOOL IMPLEMENTATION

We have implemented IRIS assistant tool to support our suggestion. As Figure 6.1 shows, IRIS consists of two part, business modeling part and big data part in the perspective of components. For the business modeling part, we defined our own modeling language for helping evidence-based business process reengineering, GO-BigBAM, which helps diagnose as-is processes according to business goals and transform them into to-be processes with the insights supported by big data. The modeling language is composed of models which have layered styles and diverse views such as Business Goal-Process-Big Analytics Query Alignment View, Transformational Insights View and Big Analytics Query View. EMF (Eclipse Modeling Framework) [72] and Sirius [73] are used to implement business modeling part. The communication between the two components is done by Big Analytics Queries which can be semi-automatically generated in Business Modeling Part. When it comes to the big data part, there is an execution engine for Big Analytics Queries utilizing Spark [74] which processes big data in real-time on distributed parallel computing and provides libraries for analytics such as Spark SQL or Spark MLlib. Spark also allows to process diverse data format including HDFS (Hadoop Distributed File System), SQL database and NoSQL database that offers means other than the tabular relations used in relational databases. For big data storage, we utilized Cassandra [75] which supports high availability with no single points of failure as one of NoSQL databases.

By using this modeling framework, analyzers can find the most critical problems among diverse alternatives in as-is business processes according to their business goals and discover the most effective solutions for to-be, which means a goal-oriented approach and the procedures are

supported by the evidences from the results of Big Analytics Queries. The selection of queries is also done by goal-orientation approach. To solve the identified problems, analyzers can also find diverse alternative solutions for to-be business processes similar to the problem finding. Those insights on problems and solutions for BPR are utilized to generate a new TO-BE process, which can be also modeled by our language.

For GO-BigDM, IRIS assistant consists of two parts: extended BPMN part and big data part. The extended BPMN part is intended for modeling a business process, which acts as the context for carrying out BA. The extension results from integrating BPMN, SIG (Softgoal Interdependency Graph) and PIG (Problem Interdependency Graph). The big data part, in turn, consists of the EER and machine-learning parts. Both the extended BPMN part and the EER part are implemented using the metamodelling facility of EMF (Eclipse Modeling Framework) and visualized using Sirius. The big data part is intended for modeling a (virtual) big data and for helping with analysis and prediction, using a big data platform and big queries. For some types of NoSQL databases, the query expressive power may be limited. For example, despite Cassandra's claimed scalability, CQL (Cassandra Query Language) per se at the moment does not seem to allow for aggregate functions or qualifiers without using keys. However, with the help of Spark SQL, that aggregate functions or qualifiers are possible.

For GO-BP2UC, the rules for transformations from the new TO-BE process to Use Cases for a software system is implemented by QVT [76]. Both the source model of BPMN augmented with NFR, the intermediate model and the target model of Use Case augmented with NFR, are defined with metamodels.

## 6.1 IRIS Architecture

**1. Conceptual Model (GO-BigBAM):** Business Modeling Part provides GO-BigBAM for explicitly modeling current and next alignment statuses of business processes and insights on problems and solutions along with big data. As Figure 6.1 show, the concepts of this model has three layers, i.e. Business Goal

Layer, Business Process Layer and Big Data Analytics Layer. As Figure 6.1 and Figure 3.2 show, Business Layer for Business Goal, Performance Goal and KPI, Business Process Layer for Business Process Goal

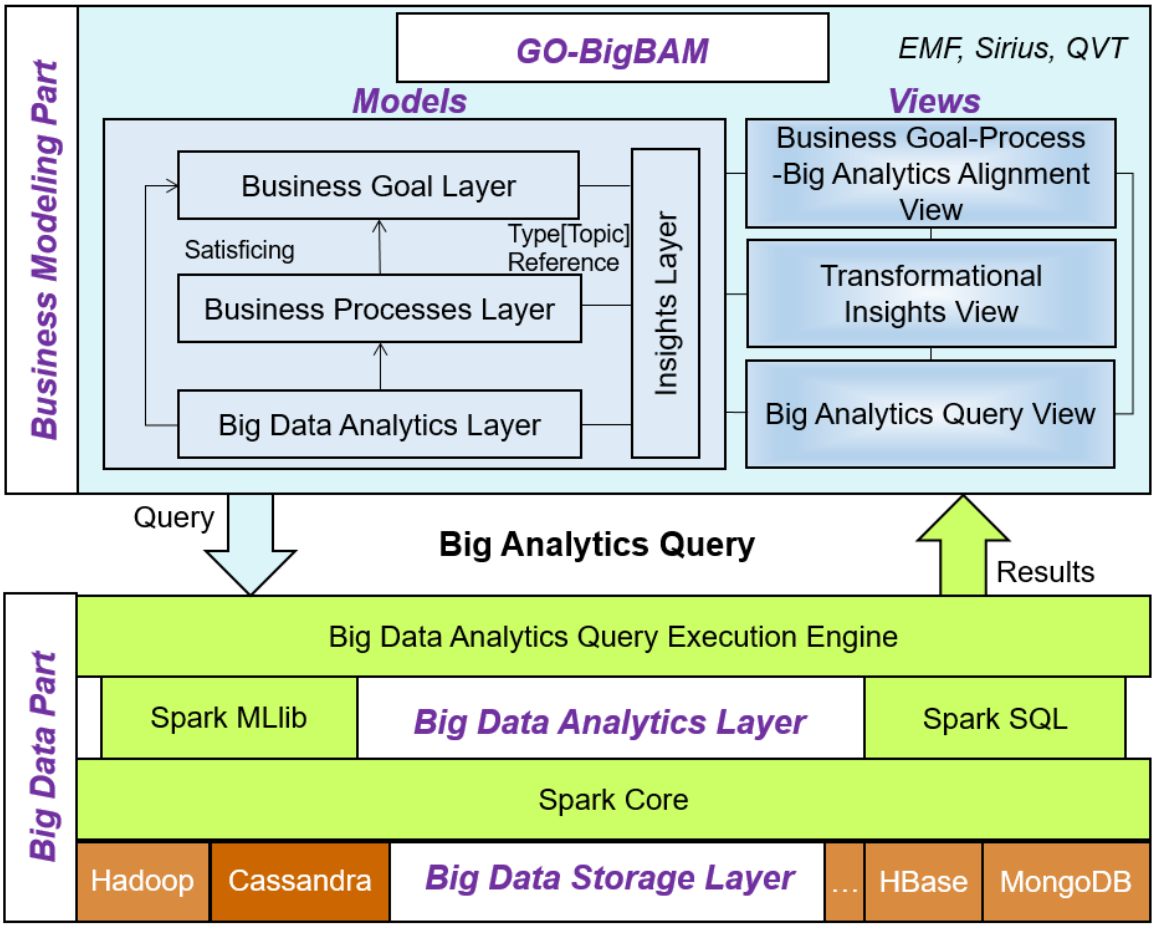


Figure 6.1. The architecture for IRIS assistant tool using Spark

and Business Process, Big Data Analytics Layer for Big Analytics Query are horizontal layers and Insight Layer for Problem and Solution is a vertical layer. Each horizontal lower layer has relationships such as Satisficing Relationship positively or negatively towards its higher layer, which can reflect well not only goal-orientation approach, but also the alignment of business activities toward business goals. Additionally, vertical layer has Type [Topic] Reference relationship which enables to represent problems and solutions in any horizontal layer. In this relationship, Topic can be any element in any layer. Thus, Type and Topic can be refined into more specific ones.

These ontological concepts are viewed by diverse views, i.e., Business Goal-Process-Big Analytics Query View, Transformational Insights View and Big Analytics Query View. In the next subsections, we will describe each view with meta-models and corresponding examples.

**Business Goal-Process-Big Analytics Alignment View**

This view is used to figure out how well business processes of an organization are aligned with their business goals using big data analytics. Figure 6.2 shows the elements and their relationships in this view. Colored classes are extended elements of NFR Framework.

In this view, as Business Concept is a parent of all entities such as Goal, Business Process, Big Data Analytics and Insight, it can present hierarchical relationships such as Satisficing (Make, Help, Hurt and Break), Correlation (Conflict or Harmony) or logical relationships such as Decomposition (AND or OR) between parent and child. It also has a satisficing label attribute for representing how much a Business Concept is satisfied in a qualitative manner, i.e. ✓Satisficed, $w^+$Weakly Satisficed, $w^-$Weakly Denied, ✗Denied or ↘Conflict. By using those concepts, we can express a Goal-Orientation approach which explores alternatives and selects among the

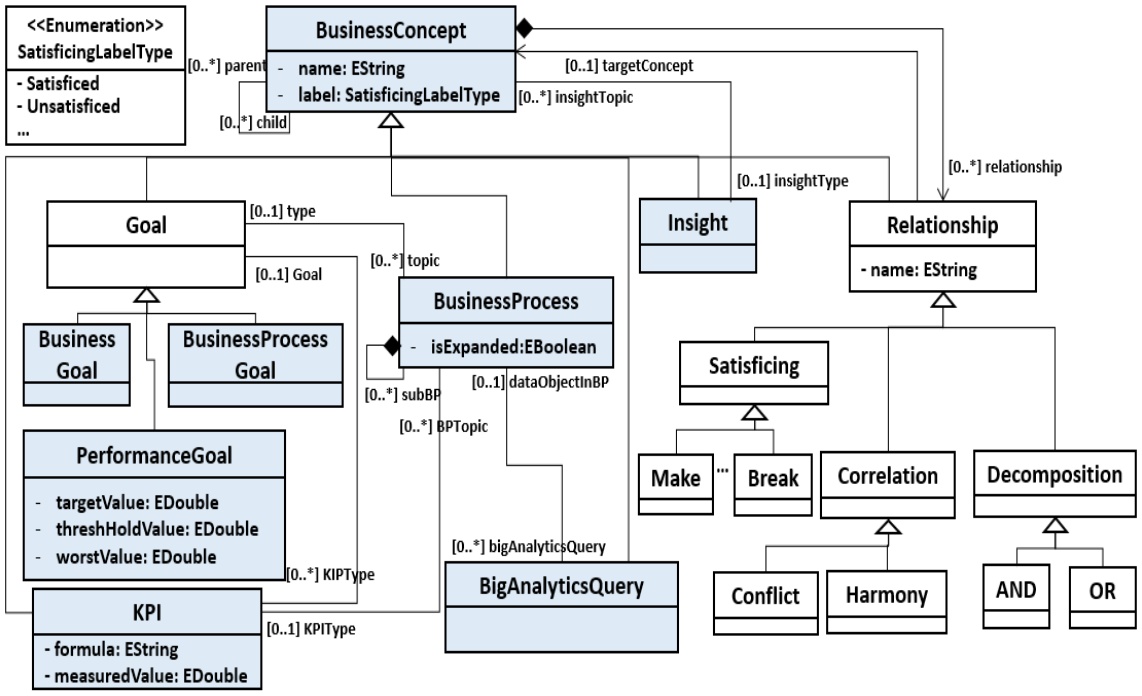


Figure 6.2. Business Goal-Process Alignment View

alternatives with a trade-off analysis. Thus, not only alignment between layers but also alignment within a layer can be expressed.

As Goal is a statement of what a subject tries to accomplish, according to whether the subject is an organization or a business process, it will be Business Goal or Business Process Goal respectively. Additionally, for qualitative measurements, Performance Goal evaluates KPIs by using targetValue, threshHodValue and worstValue which is adopted from [78]. As KPI is used to measure current business performance, it can be connected to Business Process or Goal with Type [Topic] reference. In this framework, Type [Topic] reference relationship is treated as an attribute. By evaluating KPI, a satisficing label for goal achievement is determined. When measuring KPI, Big Analytics Query is used. In the next subsection, we will explain Big Analytics Query View in detail.

As Business Process is a collection of inter-related business activities or tasks, the Business process-specific ontology is adopted from BPMN [22]. Business Process can be represented with or without expanded form of sub process. Figure 6.3 shows an example of Business Goal-Process-Big Analytics Query Alignment View using the initial understanding of automobile logistics case. Reduce Logistics Cost is a business goal and its performance goal is Reduce Logistics Cost > 3000 with a KPI, Logistic Cost. This representation can be abbreviated as a direct connection of a business goal and a KPI without a performance goal such as the case of Reduce Inventory of Cars with Inventory of Cars in Figure 6.3. In later example, to avoid complex diagram, we will use the abbreviation. To achieve the business goal, Reduce Inventory of Cars with Make is relatively better than Improve Distribution Network with Help. Reduce Complete Cars [Inventory] and Visualize Status [Inventory] are business process goals to achieve Reduce Inventory of Cars. Increasing [BTO] and Accurate [Sales Forecasting] have Make contribution to Reduce Complete Cars [Inventory]. To represent of validation of the modeled elements, each element can be supported by Big Analytics Query.

**Transformational Insights View**

This view is for providing insights for transformation from as-is processes to to-be processes and Figure 6.4 shows internal elements and relationships. Insight is the result of apprehending the inner nature of things, and there are two kinds of Insight in our framework: Problem Insight and Solution Insight for BPR.

While Problem is an Insight on a phenomenon which makes some negative contribution towards achieving a Business Goal or Business Process Goal, Solution is an Insight on a phenomenon which makes some positive contribution towards a Business Goal or Business Process Goal. Since

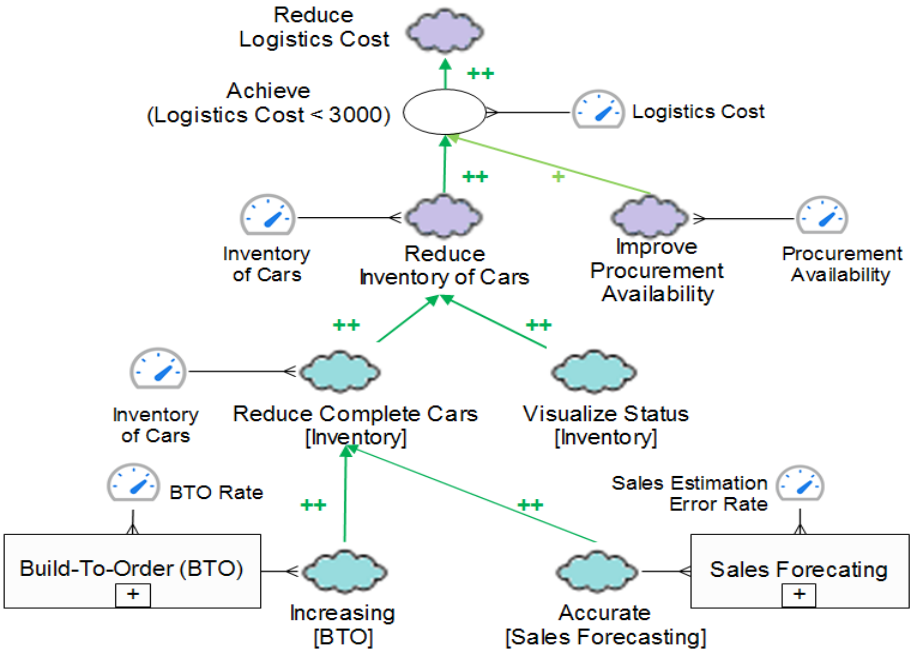


Figure 6.3. An example of Business Goal-Process-Big Analytics Alignment View without a validation by a big analytics

these Insights are also inherited from Business Concept, the relationship between parent and child can be possible with Satisficing, Correlation and Decomposition. Moreover, since Insight has the insightTopic attribute to BusinessConcept which not only entities, but also relationships inherit, it can have Topic [Topic] reference which allows refinements into more specific topic elements. Thus, more specific analysis on problems or solutions can be expressed. When validating Problem Insight and Solution Insight, Big Analytics Query can be used.

Figure 6.5 shows the example problem insight of Transformational Insight View. In J's case, filed observers reported that Aggressive [Sales Forecasting] causes Forcible Sales to Clean up Inventory. The relationship between the problems can be represented by Make of parent and child and Aggressive [Sales Forecasting] references Sales Forecasting with Sales Estimation Error Rate.

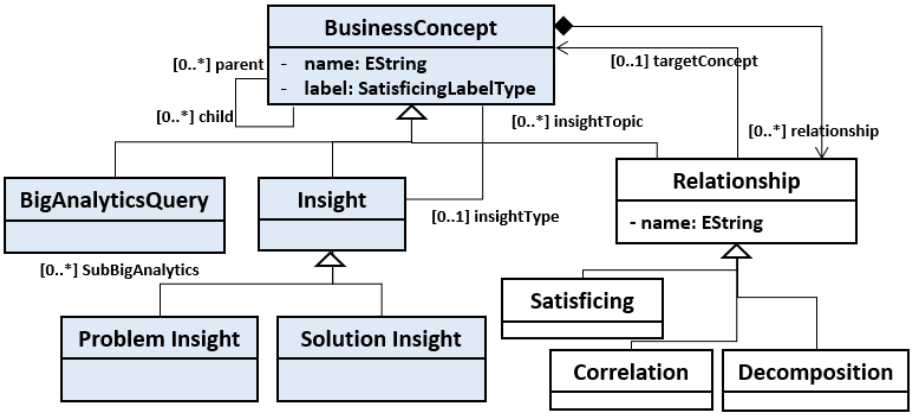


Figure 6.4. Transformational Insight View

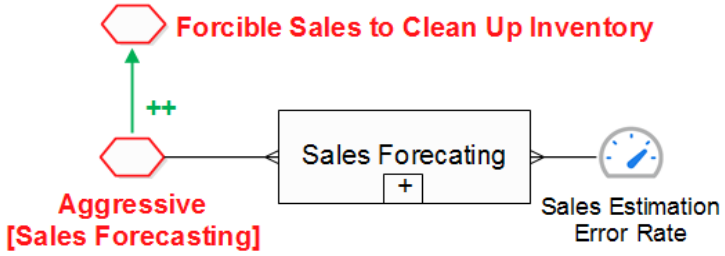


Figure 6.5. An example of Transformational Insight View without a validation by big analytics query

**Big Data Analytics Query View**

This view is used to show how a big data analytics query is organized in a query tree to validate modeled Business Concepts. Figure 6.6 shows the relationships between the elements of this view. We adopted data preparation from [79] and colored classes present different part from it. The most important difference is all of Business Concepts, i.e., not only entities but also relationships, are validated by Big Analytics Query with Satisficing Relationship. Moreover, Entity and Attributes are specialized according to big data storages. Big Analytics Query View makes a query tree showing the order of query for not only general query but also analytics algorithms which enables to make a semi-automatic query statement in standard SQL format.

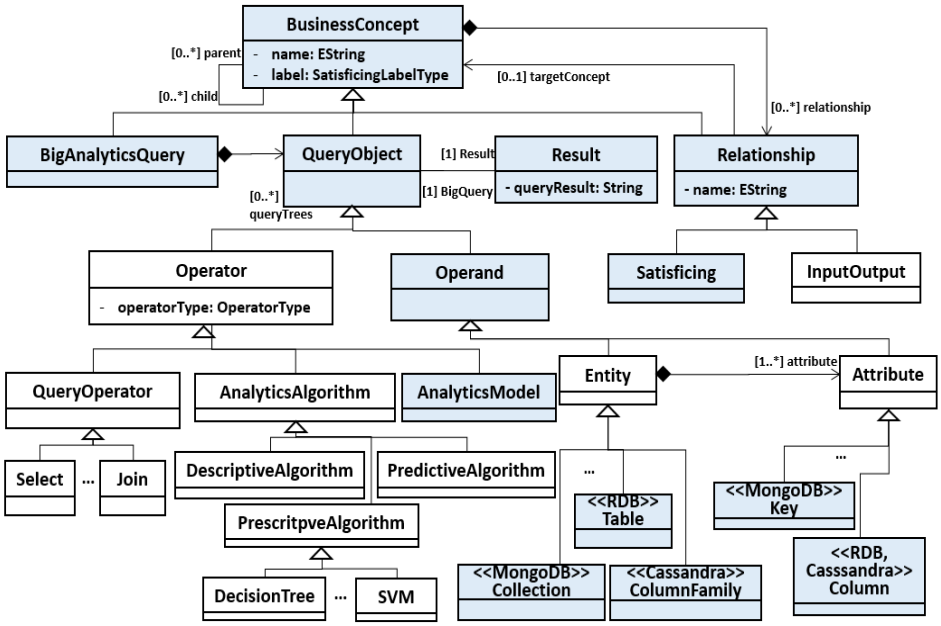


Figure 6.6. Big Data Analytics Query View (partial)

Big Analytics Query is composed of several Big Query Trees and both elements are inherited from Business Concept. Thus, they have parent and child relationships with Satisficing. The parent can be Business Concept and its subclasses such as Goal, Business Process or Insight. Additionally, there is an InputOutput relationship for Big Query Tree. In BigQueryTree, there are two kinds of elements, i.e., Operator and Operand. Operator is to represent query or analytics algorithms for big data analytics. For example, there are Select or Join for query operation, and prediction algorithms such as Decision Tree or SVM (Support Vector Machine) for analytics algorithms. We will describe diverse analytics algorithms in the next section which Spark provides.

Regarding to Operand, these will be leaf nodes of Big Query Tree, and Entity or Attribute will come as an Operand. The Entity and Attribute can be different depending on which database will be used. For example, in traditional RDBMS, Entity and Attribute will be matched with Table and Column respectively. However, in NoSQL DBMS for Big Data storage such as Cassandra or

MongoDB, they are different. As Figure 6.6 shows, in Cassandra, Entity will be ColumnFamily and in MongoDB, Collection for Entity and Key for Attributes. This Big Query Tree can be used to select the most efficient query among diverse alternatives and generate Big Data Analytics Query statement.

Figure 6.7 shows an example of J's case with Big Analytics Query View. To decide the most effective Business Goals, analyzers can use two different Big Analytics Queries; BAQ1 for Correlation between Inventory Cars and Logistics Cost, BAQ2 for Correlation between Procurement Availability and Logistics Cost. BAQ1 ○ GroupBy CarType of InventoryCarSensingTable then save into InventoryKPIs.InventoryCars attribute. InventoryCars and LogisticsCost are inputs of ○Correlation operation. BAQ2 is similar. To give an example, we assume (this data is not given in [80]) that in normal case without no problems the Result of BAQ1 is 0.9 and the Result of BAQ2 is 0.7. According to [81], if 0.8< correlation coefficient (r) <1, the strength of correlation is very strong, and if 0.6 < r < 0.79, the strength of correlation is strong. Since BAQ1 and BAQ2 support Make and Help relationship respectively, both of the queries and supporting relationships have ✓Satisficed labels.

**2. Big Data Analytics Platform:** This layer enables Big Analytics Queries which are semi-automatically generated in Big Analytics Query View to be executed in Big Analytics Query Execution Engine using Spark. The Execution Engine utilizes Spark which provides diverse libraries regarding to big data analytics such as Spark SQL and Spark MLlib. Spark SQL allows to execute query statements on structured or semi-structured data with a standard SQL format, and Spark MLlib offers diverse machine learning algorithms and statistical functions. In the next subsections, we will describe how the Execution Engine works, what analytics methods Spark

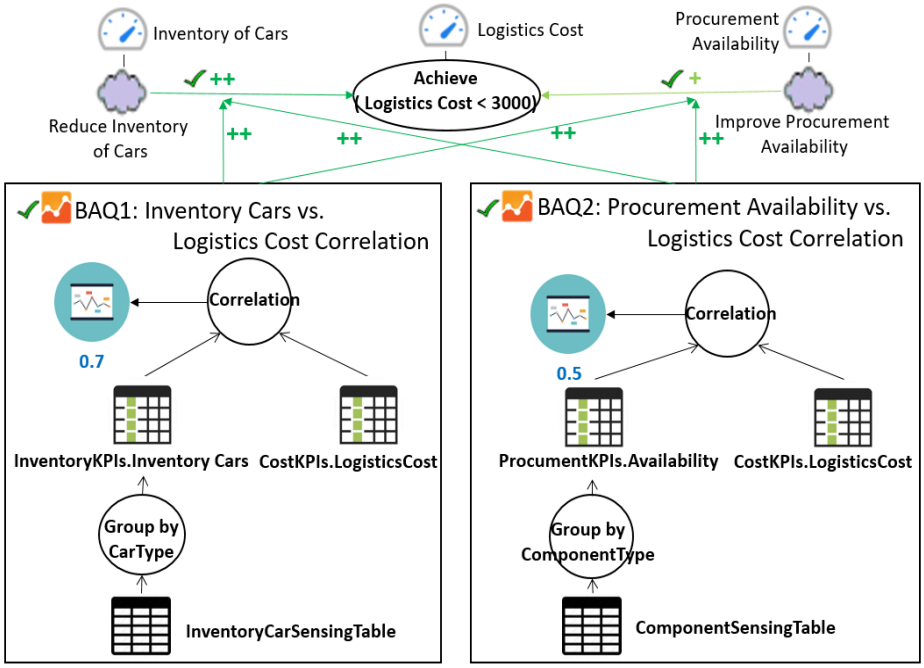


Figure 6.7. An example of Big Analytics Query View

provides, how our framework utilizes them.

Big Data Analytics Query Execution Engine gets a Big Analytics Query as input and returns back the results of it. According to the kind of the input query, the engine uses only Spark SQL or only Spark MLlib or both of them.

**Big Analytics Query Execution Engine**

Big Data Analytics Query Execution Engine gets a Big Analytics Query as input and returns back the results of it. According to the kind of the input query, the engine uses only Spark SQL or only Spark MLlib or both of them.

Basically, Spark works with RDD (Resilient Distributed Datasets) which is an immutable distributed collection of objects, in order to deal with big data. Each RDD is split into multiple partitions, which may be computed on different nodes of the cluster [82]. Spark SQL uses extension of RDD model, called a DataFrame which contains an RDD of Row objects. DataFrame

has many transformational operations. Additionally, Spark MLlib provides diverse machine learning algorithms and statistical functions for descriptive, predictive and prescriptive analytics. In this engine, by utilizing Spark MLlib, analytical data can be stored back to Big Data Storage Layer, and Big Data Analytics Query can retrieve the analytical data. In the next subsection, we will describe more detailed one.

**Big Analytics Query Catalog**

Regarding to Big Analytics Query Methods, IRIS provides two methods using Spark, i.e., Big Data Analytics Method and Big Data Query Method. While Big Data Analytics Method generate analytical data by using Spark MLlib, Big Data Query Method retrieves data not only from original proper data, but also from analytical data. Figure 3.4 shows what kind of big data analytics methods are provided in Spark MLlib. The analytics methods can be categorized into descriptive, predictive and prescriptive analytics. Statistical Descriptions such as Statistic Summary or Correlations and Clustering such as K-means can belong to Descriptive Analytics. For predictive analytics, there are diverse feature modeling methods and predictive algorithms such as Decision Tree or Linear SVM. When it comes to predictive analytics, Spark MLlib provides Normal Equation Solver. By using these methods, Big Data Analytics Query Execution Engine generate analytical data and save back to Big Data Storage Layer. According to analytics types, descriptive, predictive and prescriptive analytical data will be created in Big Data Storage Layer. According to what kind of data a query statement uses, Big Data Query method can be categorized into Descriptive- for describing real world events, Predictive- for future probabilities and trends and Prescriptive- for best actions Big Analytics Query. Descriptive Big Query uses Original Proper Data or Descriptive

Analytical Data which was generated by Descriptive Analytics Method. Other Big Analytics Query also applied the same as Figure 3.5 shows.

**Big Data Storage:** This layer is for storing not only proper, but also analytical data. Spark can support not only SQL Database such as traditional RDBMS which is based on relations between tables, but also NoSQL Database such as Cassandra or MongoDB which provides a mechanism to store and retrieve data other than the tabular relations.

In IRIS, among several NoSQL Databases, Cassandra is selected as Big Data Storage which is popular in industry and academia. Cassandra is a highly scalable, high-performance distributed database designed to handle large amount of data across many commodity servers, providing high availability with no single point of failure. Cassandra does not provide expensive table join and aggregation operations, but it depends on data nesting or schema de-normalization to enable complex queries to be answered by only accessing a single table. Thus, it has high performance. However, Spark SQL makes join possible programmatically.

## 6.2 GO-BigBAM Metamodel

Figure 6.8 partially shows a metamodel for GO-BigBAM which is defined with ECore which EMF provides for defining metamodel. This metamodel contains ontology for GO-BigBAM. The metamodel largely consists of big data part, business process part and Softgoal/Softproblem part, and they are tightly integrated together. Softgoal and Softproblem are connected with big data and business process part by Type[Topic] relationship. Additionally, Figure 6.9 and 6.10 are screen shots of GO-BigBAM with BIRT [77] which is a reporting tool and with diverse views respectively.

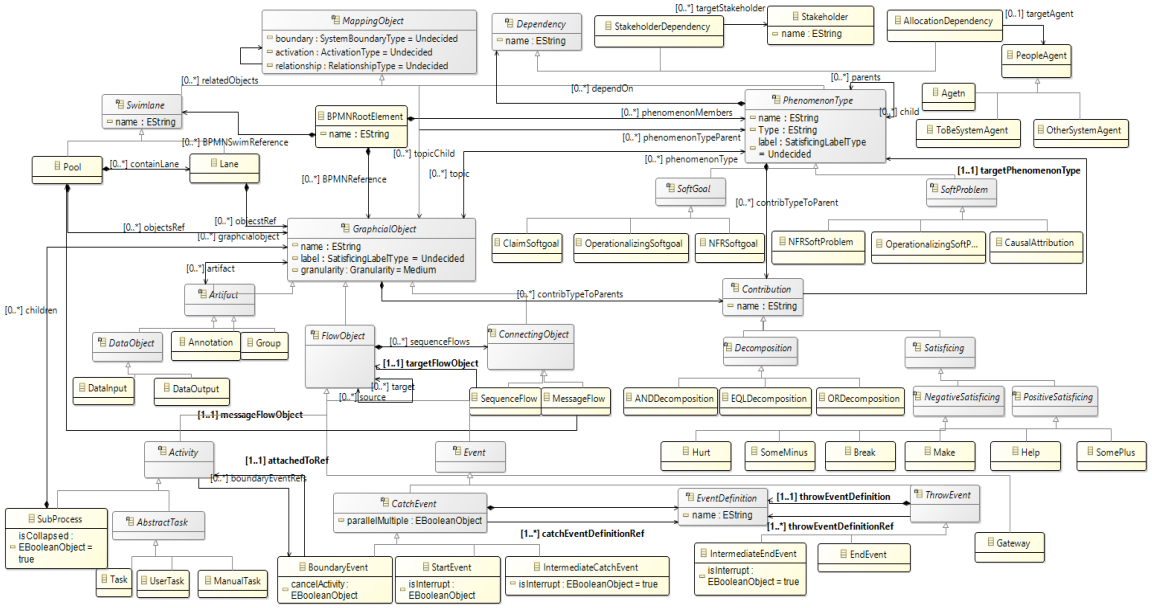


Figure 6.8. A metamodel for GO-BigBAM (partial)

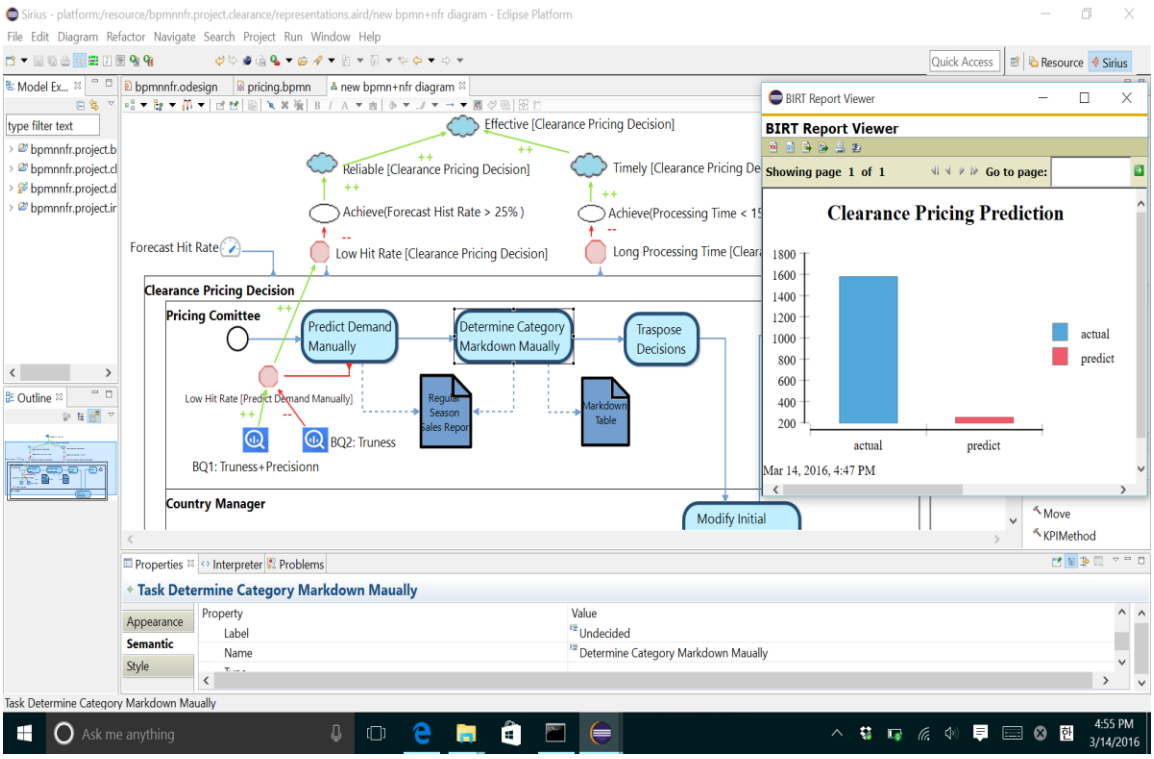


Figure 6.9. IRIS assistant: a tool for GO-BigBAM with BIRT

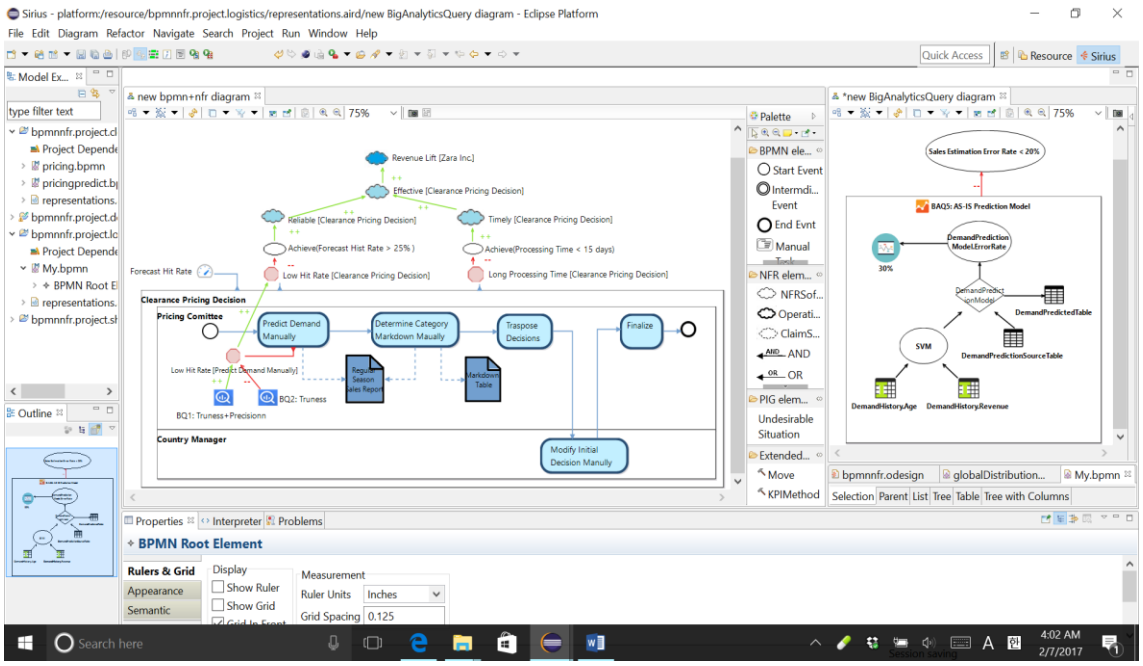


Figure 6.10. IRIS assistant: a tool for GO-BigBAM with diverse views

## 6.3 GO-BigDM Metamodel

Figure 6.11 partially shows a metamodel for GO-BigDM and Figure 6.12 is a screen shot of GO-BigDM.

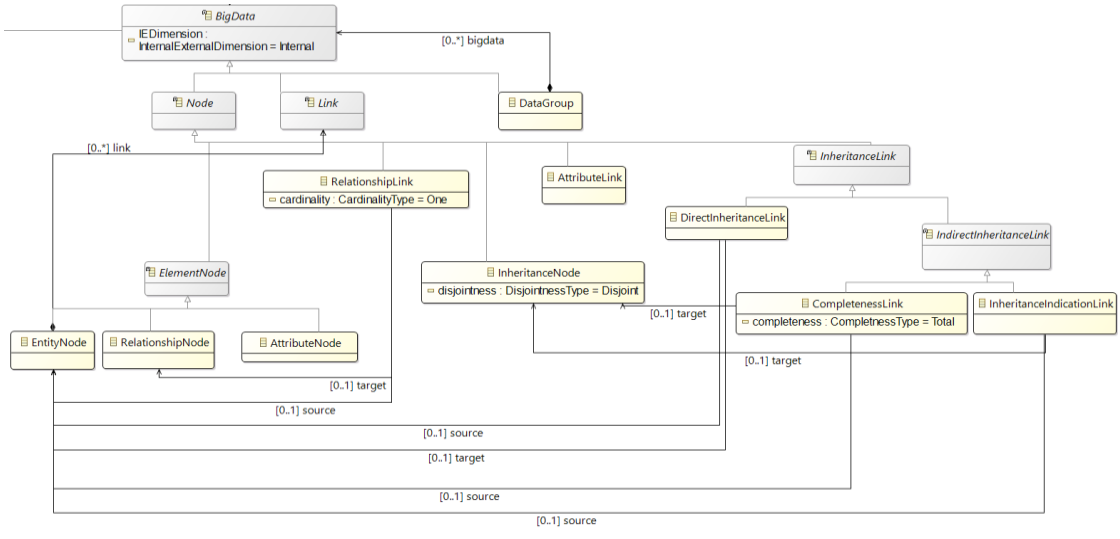


Figure 6.11. A metamodel for GO-BigDM (partial)

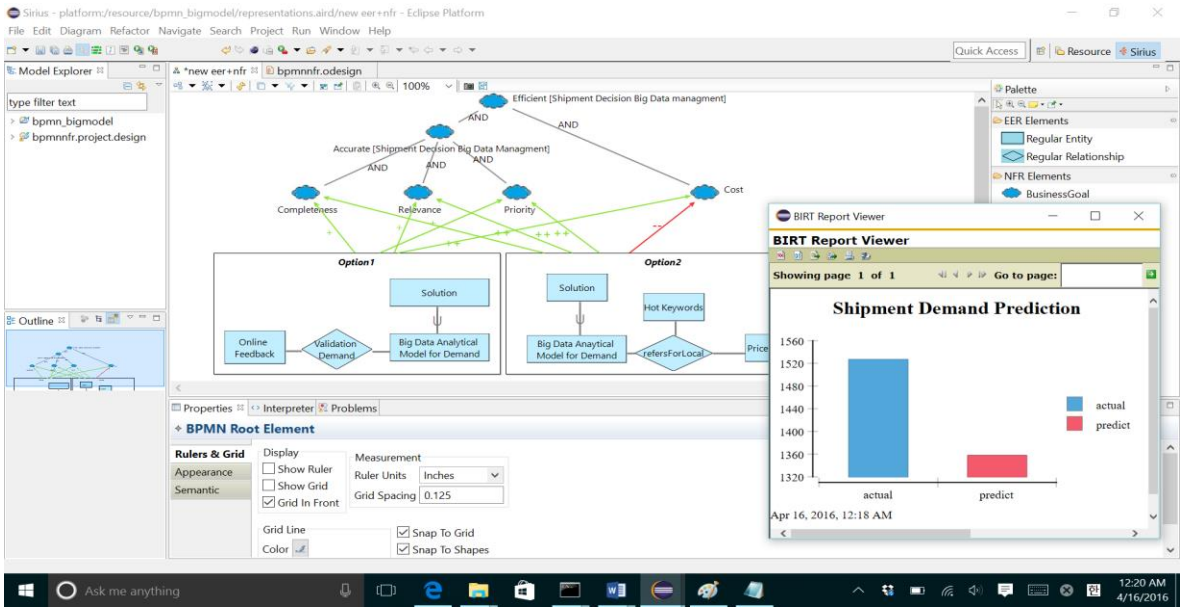


Figure 6.12. IRIS assistant: a tool for GO-BigDM with BIRT

## 6.4 6.4 GO-BP2UC Metamodel

Figure 6.13 is a metamodel for intermediate model in GO-BP2UC and Figure 6.14 is its screen shot of a view.

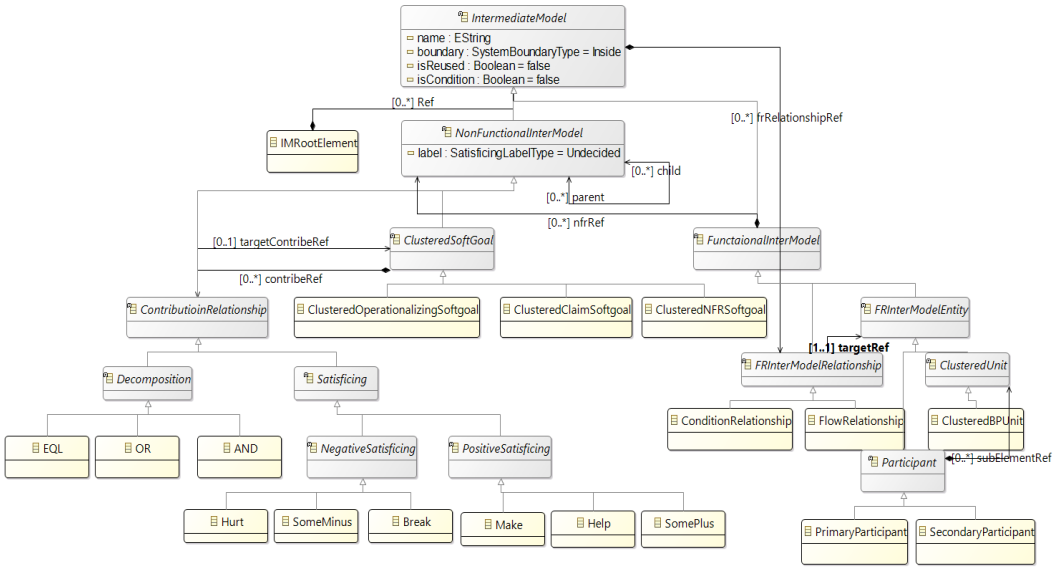


Figure 6.13. A metamodel for Intermediate Model in GO-BP2UC

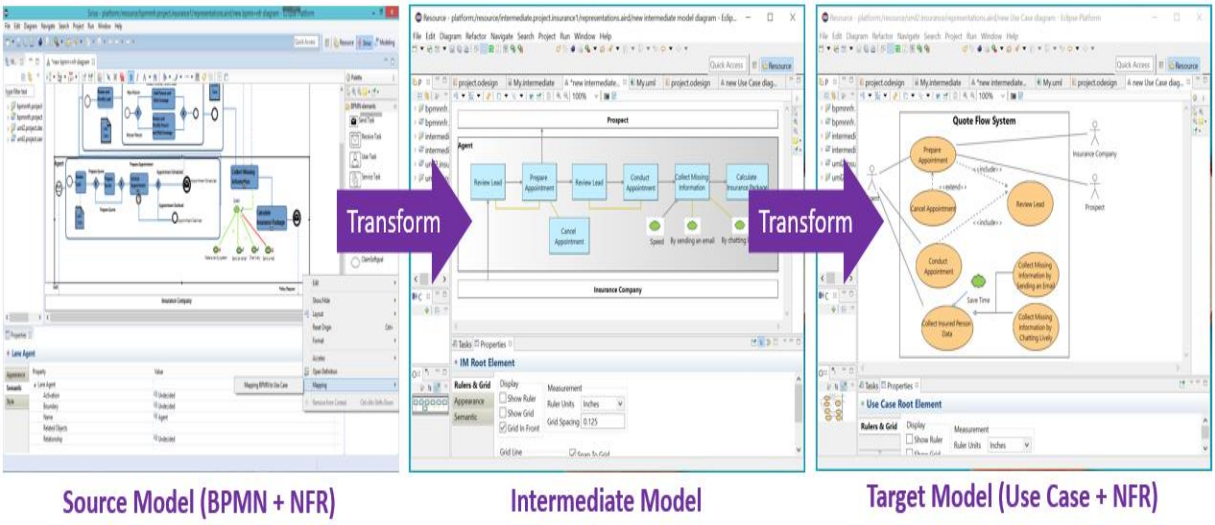


Figure 6.14. IRIS assistant: a tool for GO-BP2UC

# CHAPTER 7

# APPLYING IRIS FRAMEWORK: EMPIRICAL STUDIES

## 7.1 Automobile Industry Logistics Case

### Example Description

In this section, we describe a case of automobile logistics from [80] as a running example to illustrate of our key concepts in IRIS for BPR as well as for later discussing the applicability of it. An automobile company, J which needs to manage statuses of suppliers' shipment and a lot of components to complete a car, has been trying to *Reduce Logistics Cost* as a business goal and used *Logistics Cost* as a KPI (Key Performance Indicator) for the measurement of the goal performances. The company considered that among many others, *Reducing Inventory of Cars* would be helpful to achieve the business goal and set *Inventory of Cars* as a KPI for the measurement. They also set *Reduce complete of cars of Inventory* and *Visualize dealers' inventory statuses* as business process goals to achieve the business goal. Meanwhile, regarding to production, J has been used a forecasting process for accurate sales forecasting in which several departments were involved such as planning, sales and marketing departments. Additionally, it has been increasing the amount of products by build-to-order (BTO) in which production is initiated only after demand is known and each item is delivered directly to the customer after production is completed.

The company expected the activities to positively affect the reduction of logistics cost. However, field observers reported that *too much expectations for new products and aggressive sales forecast by the planning section cause forcible sales of salespersons to clean up inventory*. To deal with

the problem, they added a new business process goal, *Cope with the gap between sales and actual during sale forecasting* measured by Sales estimation error rate.

To validate our proposal, we apply IRIS to the automobile logistics case which was FBCM used as case study [80]. Our study can show how IRIS can help BPR using Big Data with a goal-oriented approach in the perspective of enhancing reliability which can find hidden conflict relations, preciseness which can provide more specific requirements and traceability which can visualize sources and destinations. In the end, we evaluate our approach by comparing with previous work, then explain threats to validity.

**Step 1. Hypothesizing Problems.** J's business goal is to *Reduce Logistics Cost*. To achieve the goal, *Reduce Inventory of Cars* and *Improve Procurement Availability* can be candidates as in Figure 6.3 and with the validation of Satisificing relationships as in Figure 6.7, *Reduce Inventory of Cars* is selected as a target business goal. As Figure 14 shows, analyzers thought that *Increasing [BTO]* and *Accurate [Sales Forecasting]* Make contribution to *Reduce Complete Cars [Inventory]*. Figure 7.1 is an evolved version of initial version of Figure 6.3. After the analyzers knew that too much expectations for new products and aggressive sales forecast by the planning section cause forcible sales of salespersons to clean up inventory, they hypothesized *Cope with the Gap Between Sales and Actual flexibly [Sales Forecasting]* can solve the problem. However, they don't know which part they should change.

**Step 2. Find Business Problems.** Analyzers measured the KPIs and they found out that *Reduce Logistic Cost* and *Reduce Inventory of Cars* were *Denied*. As Figure 7.2 shows, to see the labels of goal achievement using Business Goal-Process-Big Analytics Query Alignment View, when they set *Increasing [BTO]* is *Satisficed* and *Accurate [Sales Forecasting]* is

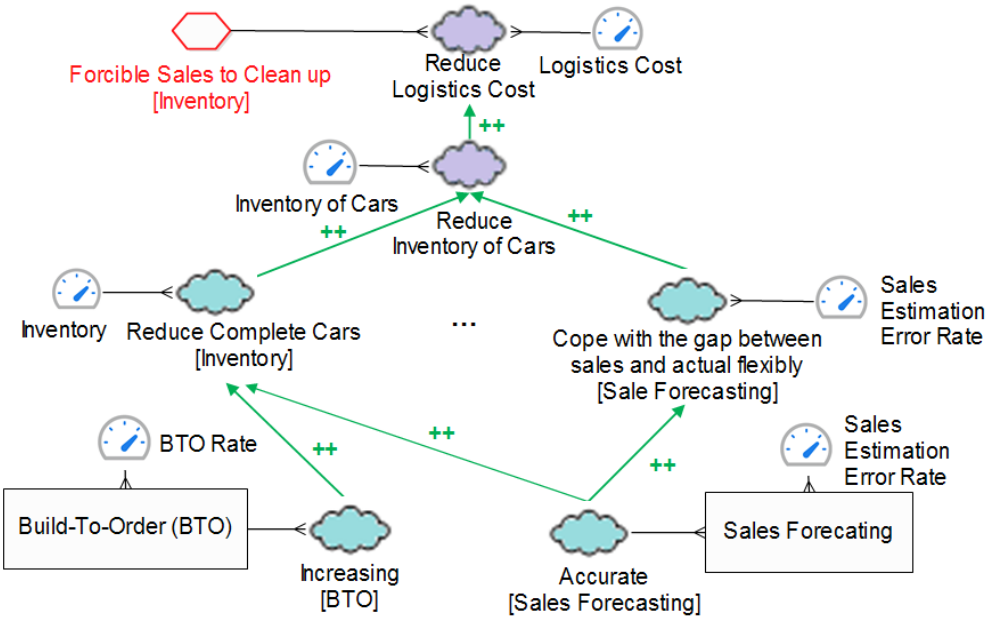


Figure 7.1. Hypothesizing a problem

*Denied, Reduce Complete Cars [inventory*] is in *Conflict* according to label propagation mechanism of NFR Framework [7]. To find out what caused this problem, they analyzed correlations of different combinations between KPIs. Among them, the correlation between Sales Estimation Error Rate and *Logistics Cost* by BAQ3 was strong positive relation (0.9) which means that when *Sales Estimation Error Rate* increase, *Logistics Cost* also increase, and the correlation between *BTO Rate* and *Sale Estimation Error Rate* by using BAQ4 was positive relation (0.7) which means that when *BTO Rate* increases, *Sales Estimation Error Rate* also increases. These data are provided by [80]. Using the view, they explicitly modeled the hidden correlations. When the positive Correlation between *Increasing [BTO]* and *Cope with the Gap between Sales and Actual Flexi*bly was established, they discovered that *Satisficed Increasing [BTO]* and *Denied Accurate [Sales Forecasting*] also caused *Conflict* situation. Analyzers felt this can be a cause of the problem, and explicitly added *Overlooking*

*Impacts by BTO Sales [Sales Forecasting]* as a root cause of *Aggressive [Sales Forecasting]* respectively in initial understanding of problems in Figure 7.1 as Figure 7.2 shows. To find out more detailed root cause problems, analyzers look into the Sales Forecasting process. To show an example, we assume that J used the process of Figure 7.3. They found out that *Gather Demand History Data*, *Forecast Demand* and *Incorporate Sales Changes* tasks are related to the problem. Regarding to the *Forecast Demand*, they investigate how the forecasting was processes. When they check Forecast Demand as Figure 7.4, they found out *DemandByBTO.Amount* is not included in the feature of the prediction model. Thus, they can add *DemandByBTO.Amount* as input feature of the prediction model as the to-be prediction model of Figure 7.4.

**Step 3. Explore Business Solutions.** Using Transformational Insight View, analyzers can see more simplified problems and solutions without business processes as Figure 7.5 shows.

They explored diverse alternative solutions to solve the business problems and establish a contribution relationships. For example, for *Overlooking Impacts by BTO Sales [Forecast Demand]*, while *Forecast Including BTO Rate [Forecast Demand]* *Break* the problem, *Decision Tree Model Excluding BTO Rate [Forecast Model]* *Make* it, which means it cannot be a solution because it makes worse. Thus, the former will be selected as a solution. When a selection is made, data analysis can be used as a supporting evidence. In this case, BAQ6 (to-be prediction model descried in Figure 7.4) can be used for supporting the solution.

**Step 4. Evaluate and Select a Business Process.** Business solutions identified in Step3 can be combined together to obtain diverse solutions to solve the problems, which leads to potential

alternative to-be processes. For example, in Figure 7.5, while *To-Be Process1* combines *Include Sales Data by BTO [Gather Demand History Data]* and *Change Forecast Model*

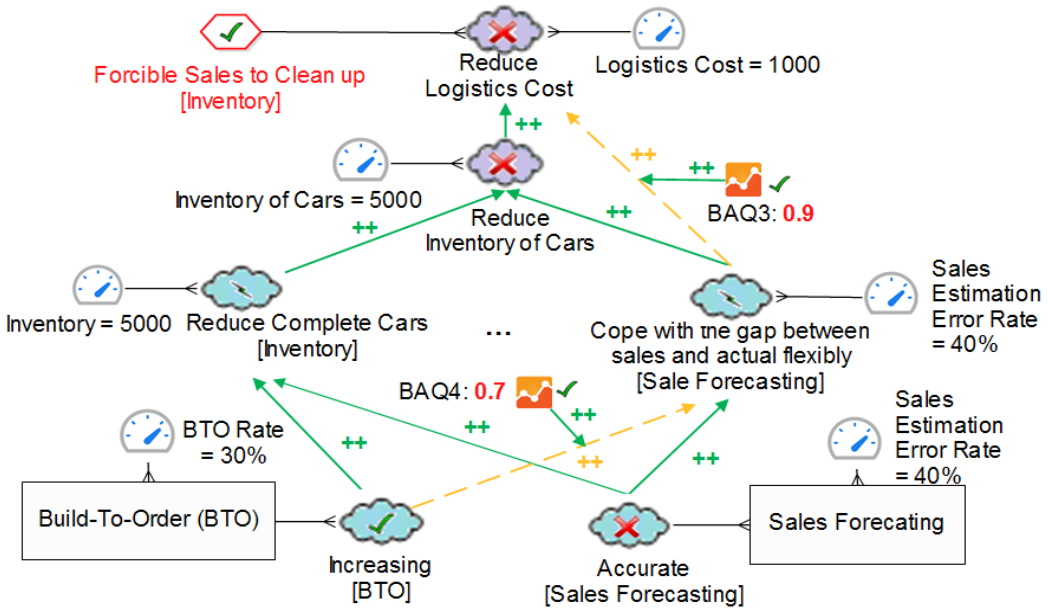


Figure 7.2. Validating the problems

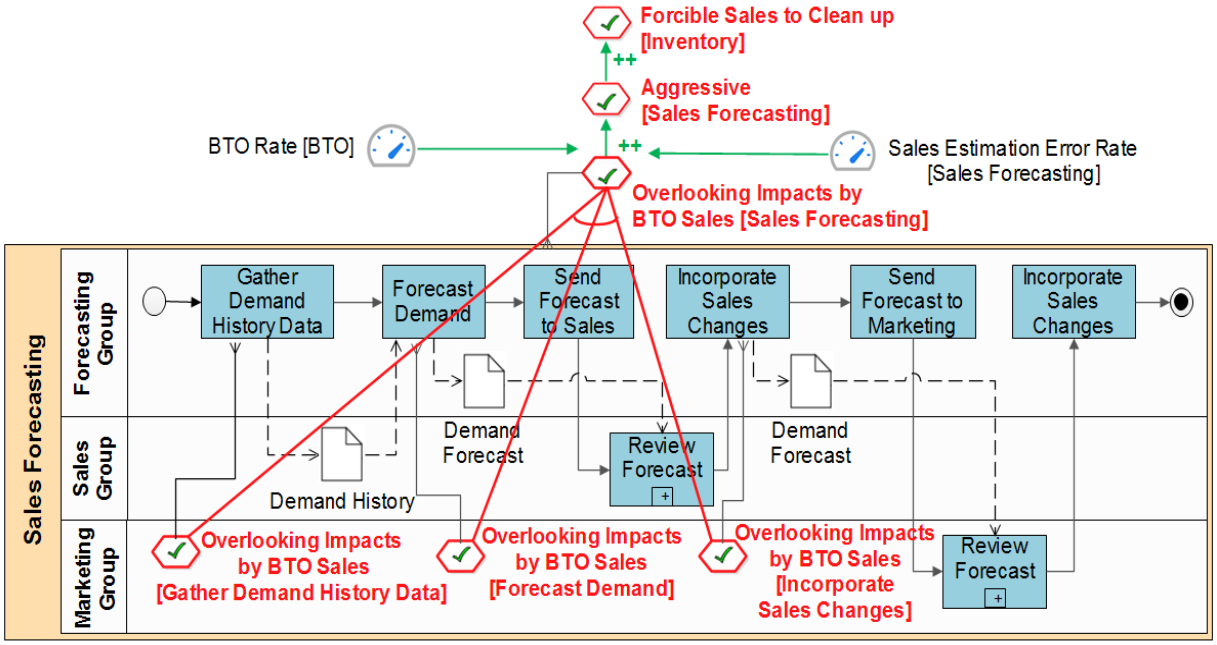


Figure 7.3. A root cause diagnosis in sales forecasting

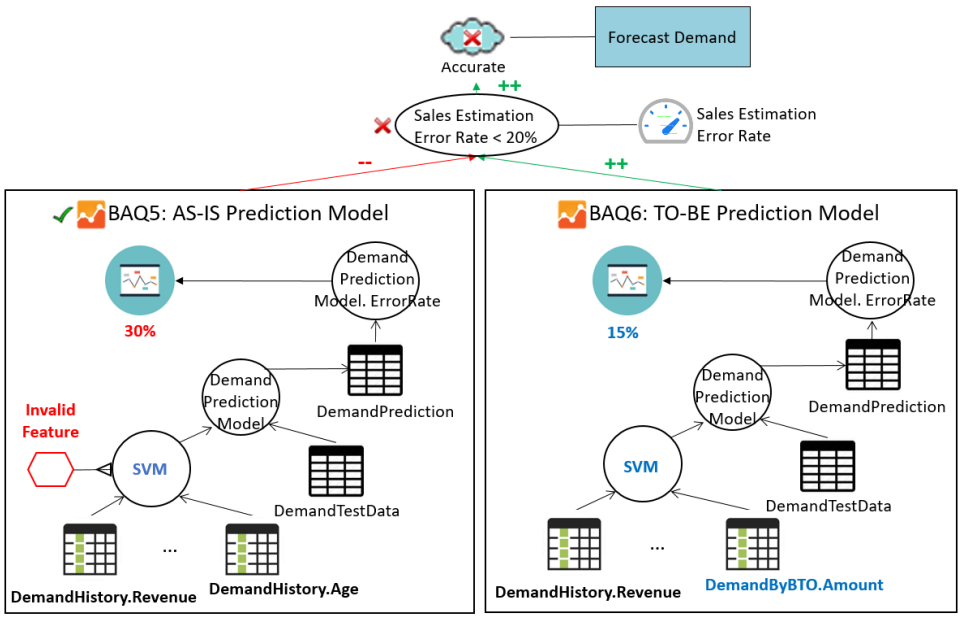


Figure 7.4. An example of Big Analytics Query View

*Including BTO Rate [Forecast Demand]*, *To-Be Process2* additionally combined *Visualize the Sales Change by BTO [Incorporate Sales Changes]*.

Although the evaluation result of *To-Be Process1* is not shown in Figure 7.6, according to the label propagation mechanism, it will *Weakly Satisfice* to *Reduce Logistics Cost*. On the other hand, since *To-Be Process2* solves all the problems found in Step3, it will fully *Satisfice* *Reduce Logistics Cost*. Thus, *To-be process2* will be selected as the final TO-BE process. In Figure 7.6, elements with thick lines will be changing parts.

## 7.2 Clearance Pricing Decision Process

### Example Description

As a running example, Figure 7.7 is about a demand and a pricing prediction process models with BPMN for the clearance pricing of a company, Z [83]. The decision process consists of two sub-processes: one for *Determining Initial Markdown* (discount) *Category* before a clearance starts,

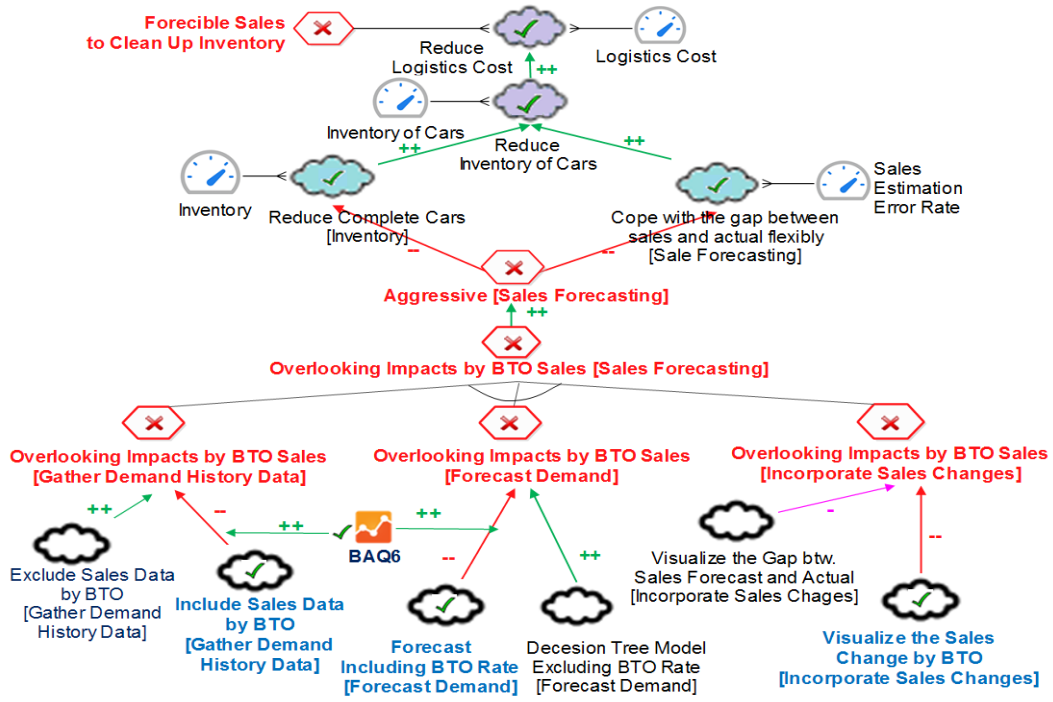


Figure 7.5. Solution finding of sales forecasting

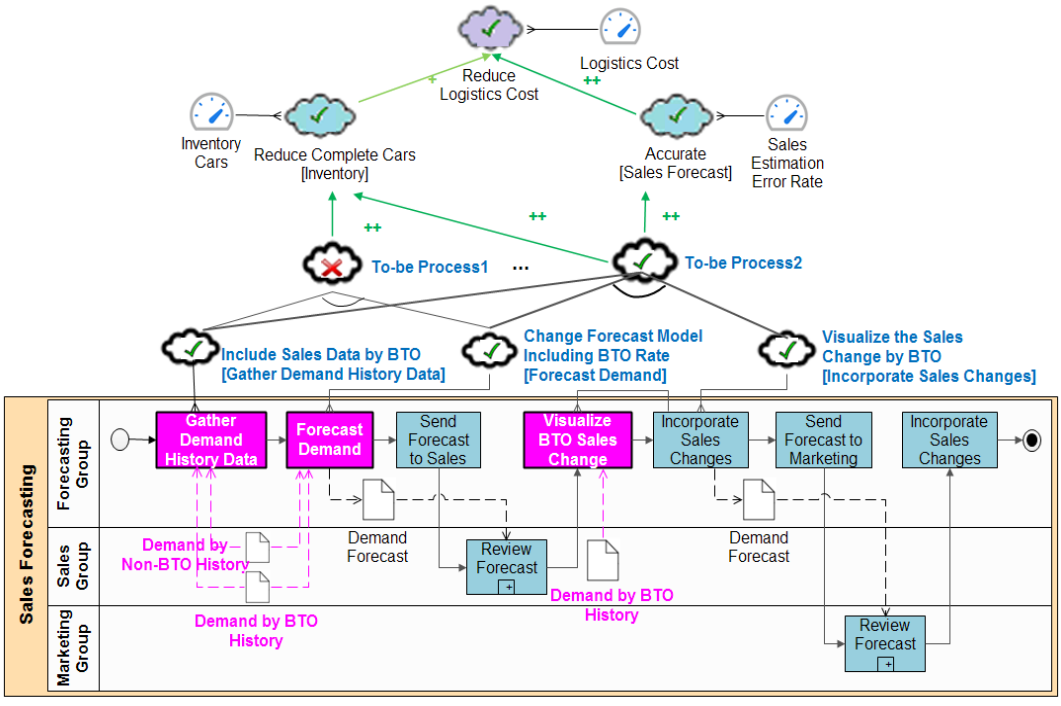


Figure 7.6. Solution finding of sales forecasting

and the other for *Updating the Markdown Category* during the clearance. *Determining the initial markdown* starts with a task *Predict Demand Manually*, which involves reviewing unsold inventory and sales performance during a regular season, and then the participant *Pricing Committee* comes in to reach the initial *Final Markdown*. This first sub-process takes about 1 month. During clearance sales, each *Country Manager* *Estimates Time to Sell (ETS)*, using the sales average of the previous three weeks of *Weekly Clearance Sales Reports* and their own personal experiences. If sales is slower than predicted (i.e., the *ETS* is greater than the actual time remaining), the initial *Final Markdown* may be considered risky, which can lead to *Further Markdown Manually*.

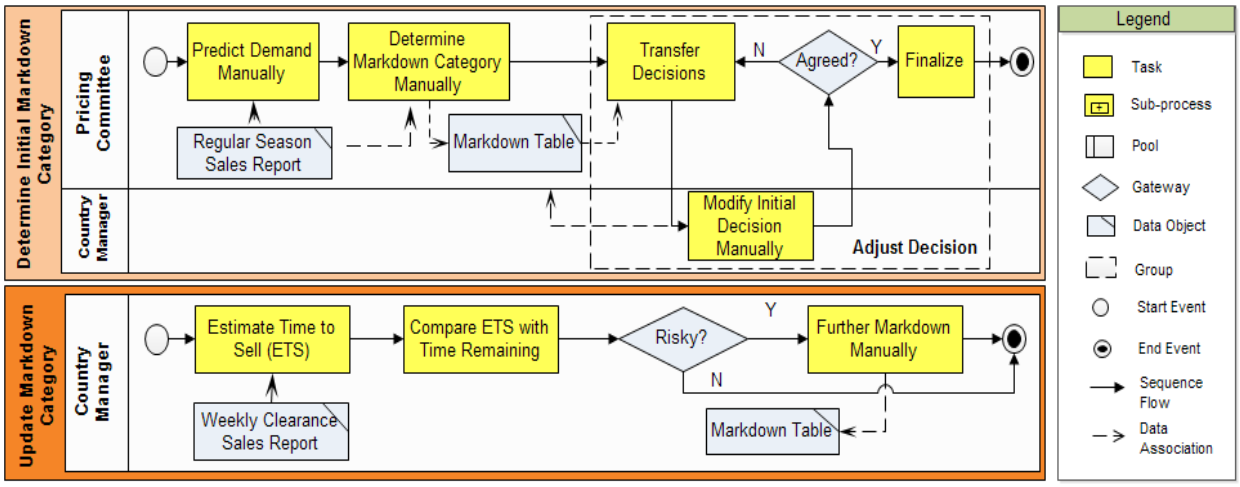


Figure 7.7. AS-IS business process of clearance pricing decision in BPMN

We applied IRIS – its ontology, process and tool – to the clearance pricing decision process. The road map for this application starts with an AS-IS clearance pricing decision process diagnostics (Figure 7.8), followed by a consideration of alternative potential solutions, and tradeoffs among them, for the problems of the AS-IS Process (Figure 7.9), and ends with two alternative TO-BE processes (Figure 7.10 and 7.11).

They show the overall traceability from a business goal to a business process goal, and then to a business process, together with business process problems and solutions and their corresponding big queries. A particular focus in this section lies in finding problems and solutions, using IRIS's complementary and goal-oriented approach, in exploring, and selecting among, alternatives.

**1. Initiate.** As mentioned in the previous section, Zara has *Revenue Lift* as a business goal, and, to achieve the goal, it needs *Effective [Clearance Pricing Decision]* as a business process goal, for which the reliability of the decision, *Reliable [Clearance Pricing Decision]*, and *Timely [Clearance Pricing Decision]* are more specific business process goals and the former is critical (denoted by "!!").

A business (process) goal can be expressed as a KPI goal, in the form: Achieve [KPI] – e.g., *Achieve (Revenue Increase > 10%), Achieve (Forecast Hit Rate > 25%) or Achieve (Processing Time < 15 days).* Here, the *Forecast Hit Rate* is not for a single item of products but all items collectively. These goal refinements are done during "Initiate" stage of the IRIS process.

**2. Problem Diagnosis.** Finding problems with the AS-IS business consists of two steps: hypothesizing a problem in a business context model and validating it by using big queries. In the first step, to find problems, we used a hybrid method (the combination of a top-down and a bottom-up) as we mentioned. In Figure 7.8, the problems are denoted as *Low Hit Rate* [*Clearance Pricing Decision]*, which BREAK *Achieve (Forecast Hit Rate > 25%), and Long Processing Time* [*Clearance Pricing Decision]*, which BREAK *Achieve (Processing Time < 15days).* The former problem is refined into *Low Hit Rate [Predict Demand Manually]* and *Low Hit Rate [Predict Markdown Manually]*, since business analysts figure out the clearance pricing decision depends on the *Demand Prediction* and the *Markdown Prediction* by seeing this AS-IS process model.

While *Low Hit Rate [Predict Demand Manually]* is associated with *Predict Demand Manually* and *Estimate Time to Sell (ETS)*, *Low Hit Rate [Predict Markdown Manually]* is associated with three tasks, *Determine Category Markdown Manually*, *Modify Initial Decision Manually* and *Further Markdown Manually*. The latter problem is also refined into *Long Processing Time [Adjust Decision]*, where *Adjust Decision* is denoted with a dotted line.

In the second step, the hypothesized problem gets validated through one or more big queries. Concerning the Forecast Hit Rate KPI above, would be:

Hypothesize Problem (*Achieve* (*Forecast Hit Rate* > 25%), Clearance Pricing Decision) =

[~*Achieved (Achieve (Forecast Hit Rate* >25%)), Clearance Pricing Decision].

For the hypothesized problem, we need big queries and big data. Big queries are just database queries, but, due to the notion of the 5***V***s, there are different types of NoSQL queries, aside from standard SQL queries. However, big queries derived from the corresponding business process KPI can be executed in a SQL format on Spark with Scala. In Figure 7.8 and 7.9, BQ1 to BQ8 are big queries, of which BQ1 and BQ2 successfully run on our tool, whose answer about the value of the particular KPI will indicate if the corresponding hypothesis is valid or not.

There can be many candidate queries for validating problems. For example, regarding to the previously hypothesized problem, *Low Hit Rate [Predict Demand Manually]*, we may consider two statistical big queries: BQ1, for both the trueness and precision of the prediction, and BQ2, for only the trueness of the prediction. According to the ISO 5725 [84], trueness refers to how close the predicted value is to the actual, while precision refers to how big, or small, the deviations are among the predicted values. Concerning queries, trueness uses the current demand prediction against the actual historical demands only, but precision uses the current demand

prediction against the average of the accumulated trueness data about the historical predictions. Either one or both could be used, but, if a choice needs to be made, which one is better?

Given {BQ1, BQ2} and {*Achieve (Forecast Hit Rate > 25%)*}, Rank Queries could yield {(BQ1, *Achieve (Forecast Hit Rate > 25%),* 2), (BQ2, *Achieve (Forecast Hit Rate > 25%),* 1). Between the two queries, BQ1 is more suitable (hence, ++ contribution) for validating the problem *Low Hit Rate [Predict Demand Manually]*, since it can yield more reliable answer.

If the answer to the selected BQ1 is True (hence, the ✓ symbol), the hypothesized problem is likely to be a validated problem (hence, the ✓ symbol); otherwise, not a validated problem. Similarly, BQ3 and BQ4 can be used to validate a hypothesized problem *Low Hit Rate [Predict Markdown Manually]*. If needed, big queries can be used to quantitatively estimate the *Forecast Hit Rate* (e.g., *Forecast Hit Rate = 18%).*

Let us suppose that BQ3 also validates *Low Hit Rate [Predict Markdown Manually]* as a problem. Since both *Low Hit Rate [Predict Demand Manually]* and *Low Hit Rate [Predict Markdown Manually]* are satisficed (✓), their parent *Low Hit Rate [Clearance Pricing Decision]* also is, according to the mechanism of label propagation using the closed world assumption [7].

To give a flavor of big queries, BQ1 and BQ2 are shown below as SQL queries which are successfully executed in our IRIS assistant tool:

BQ1: /* **trueness** + **precision** of demand prediction */

BQ2: /* **only trueness** of demand prediction */

```
/* trueness of demand prediction */
SELECT a.category, (prdt_dmd - real_dmd) as trueness
FROM ( SELECT category, prdt_dmd
```

```
        FROM   markdown_list
        WHERE  sales_year=2011) a,
    ( SELECT  category, sum(sales_count) AS real_dmd
        FROM     sales_records
        WHERE    sale_type='c'
        AND      sales_month >= 7 AND      sales_month <= 10
        AND      sales_year = 2011
        GROUP BY category) b
WHERE a.category = b.category;
/* precision of demand prediction */
SELECT a.category, avg(prdt_dmd - real_dmd) as precision
FROM ( SELECT category, sales_year, prdt_dmd
    FROM   markdown_list ) a,
   ( SELECT   category, sales_year, sum(sales_count) AS real_dmd
    FROM     sales_records
    WHERE    sale_type='c'
    GROUP BY category, sales_year) b
WHERE a.category = b.category AND a.sales_year = b.sales_year
GROUP BY a.category;
```

Due to space limitation, actual queries are omitted from the description of other queries, but only brief remarks are given below:

***BQ3:*** /* **trueness** + **precision** of markdown prediction */

***BQ4:*** /* **only trueness** of markdown prediction */

Similarly, *Long Processing Time [Adjust Decision]* may be hypothesized as a problem and validated by big queries:

***BQ5***: /* each task processing time + communication time for a business process */

***BQ6:*** /* the processing time of the whole process of determine initial markdown category */

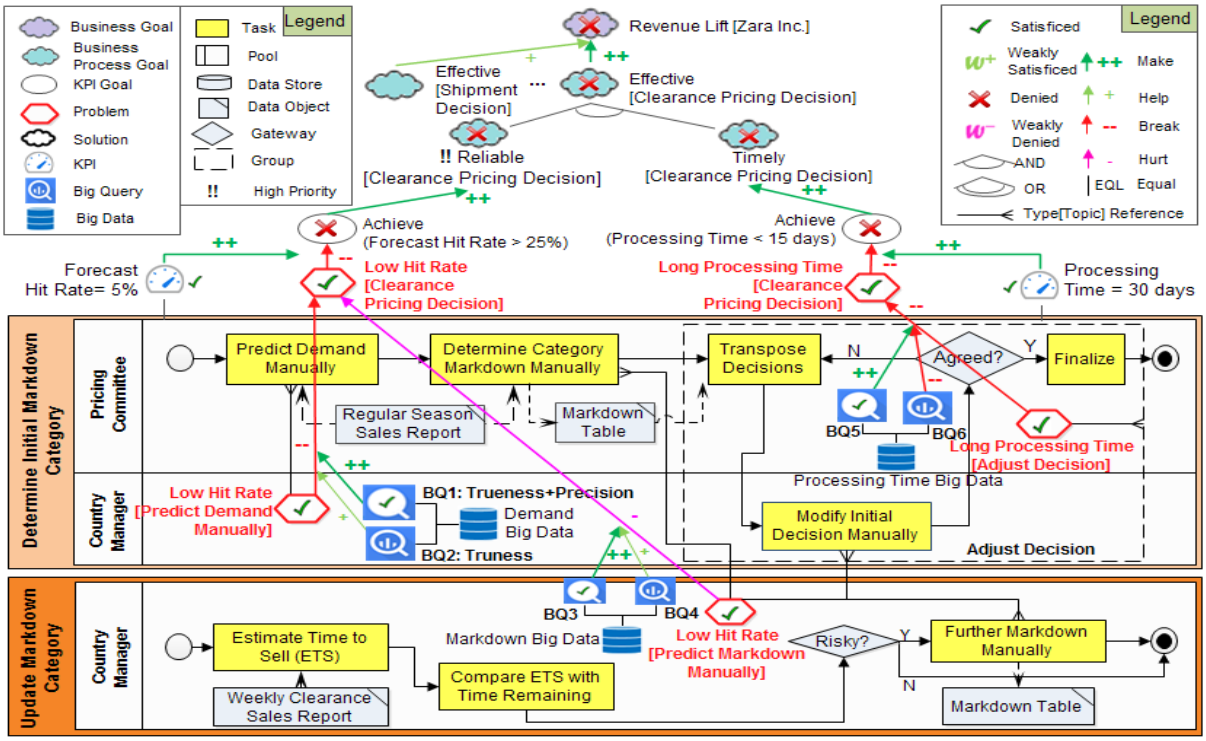


Figure 7.8. AS-IS clearance pricing decision process diagnostic

For our study, big queries were not run on Zara's real data, but data for another real retail business, whose characteristics are quite similar to Zara's, in order to determine big data queries can indeed be processed, and in a timely manner. The particular business is a world-wide department store, with over 47,000,000 as the number of order items and over 390,000,000 as the number of orders

- representing the outstanding status of the current period. Due to the proprietary nature of the business, the field and table names have been sanitized.

**3. Solution Discovery.** Finding solutions to the problems with the AS-IS process, towards a TO-BE business process, also consists of two steps: hypothesizing a solution and validating it. In the first step, as in Figure 7.9, for example, given *Low Hit Rate [Predict Demand Manually]* as a business process problem and Zara's clearance pricing decision process as its context, use of a *Big Data of Social Media Fashion Trend, Online & Offline sales [Prediction]* which MAKE both *reliable* and *timeliness* prediction could be hypothesized to be a solution. For validating the hypothesized solution, big queries are again used, e.g.:

***BQ7:*** /* clearance pricing prediction by big data: social media fashion trend + online & offline sales */

***BQ8:*** /*clearance pricing prediction by big data: offline sales */

There can be other alternatives to this solution, from the perspective of price prediction as a reliable prediction, e.g., use of just *Human Expertise* or an *Analytical Model* prediction which Zara selected. A solution can also be considered from a timeliness perspective – e.g., *Moderate Dis-intermediation [Adjust Decision]* vs. full *Dis-intermediation [Adjust Decision]*. Exploration of potential solutions then can be done in combination of the different perspectives. Each of them then needs to be validated. For predictive queries, an analyzer can use the prediction function in IRIS tool which is implemented using diverse machine learning algorithms before running a query. In order to rank them, here, a tradeoff analysis would be needed, as to how much such alternative, hypothesized solutions can (help) eliminate or alleviate the relevant problems. Such a tradeoff analysis is needed, due to the *Reliability* and *Timeliness* goals early on that are conflicting with

each other. In Figure 7.8, such a tradeoff analysis is shown in term of the different contributions (i.e., different line colors or numbers of pluses and minuses). Then ranking can be done in consideration of the relative importance of the problems, in a number of different ways. For example of one such way, !!*Reliable* and its descendants can be given 1.2 as their importance, and *Timely* and its descendants 1.0. The number of pluses and minuses contribution, together with the relative importance values, could then be used to rank *Clearance Pricing Prediction by Big Data + Moderate Dis-intermediation* (e.g., |(-2 * 1.2) + (-3* 1.2) + (-2 * 1.0) + (-3 * 1.0)| = 10.9). If this is the highest value, among the values for all the hypothesized candidate solutions, it will be validated first, and, if validated, then possibly reified in a TO-BE process.

**4. Redesign Process.** The solutions, that have been highly ranked and validated in the previous phase, need to be reified in a TO-BE process. For example, use of the solution Analytical Models [Prediction] can lead to a TO-BE process 1 (Figure 7.10) and Big Data [Prediction] + Moderate Dis-intermediation [Adjust Decision] to TO-BE process 2 (Figure 7.11). The Topic part of Type [Topic], e.g., the Adjust Decision part of the second option, needs to be modified in moving to a TO-BE process. There can be many different ways for this modification, and typically would require some intervention by subject matter experts [85]. In Figure 7.11, the entire group of tasks of Adjust Decision, which involves two participants, is replaced by a single task Finalize, which involves only one participant. Social media fashion trend and online sales data are added input of Predict Demand by Big Data and Estimate Time to Sell (EST). All the changed parts will be shown in pink.

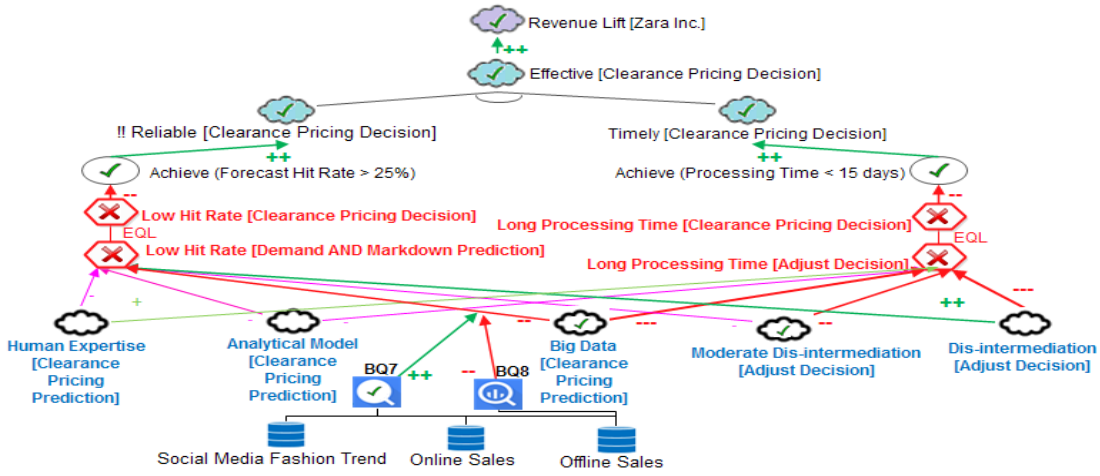


Figure 7.9. Alternative potential solutions, and tradeoffs among them, for the problems of the AS-IS process

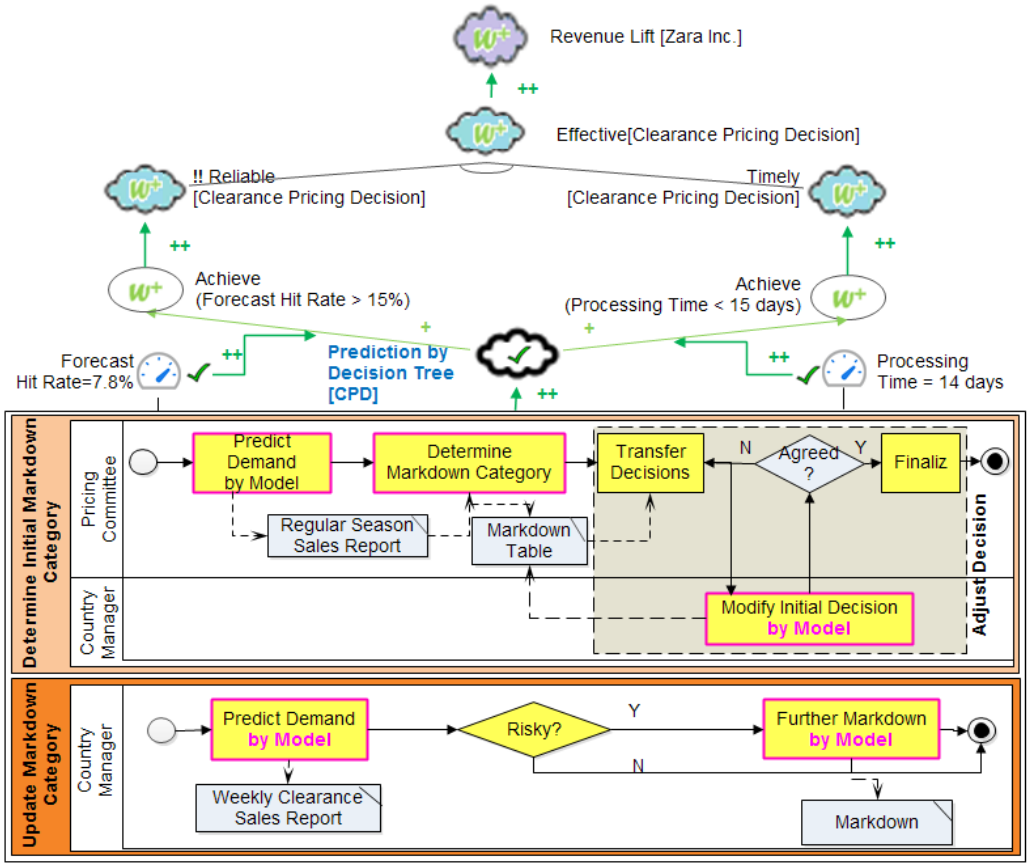


Figure 7.10. TO-BE Process 1: using only an analytic model

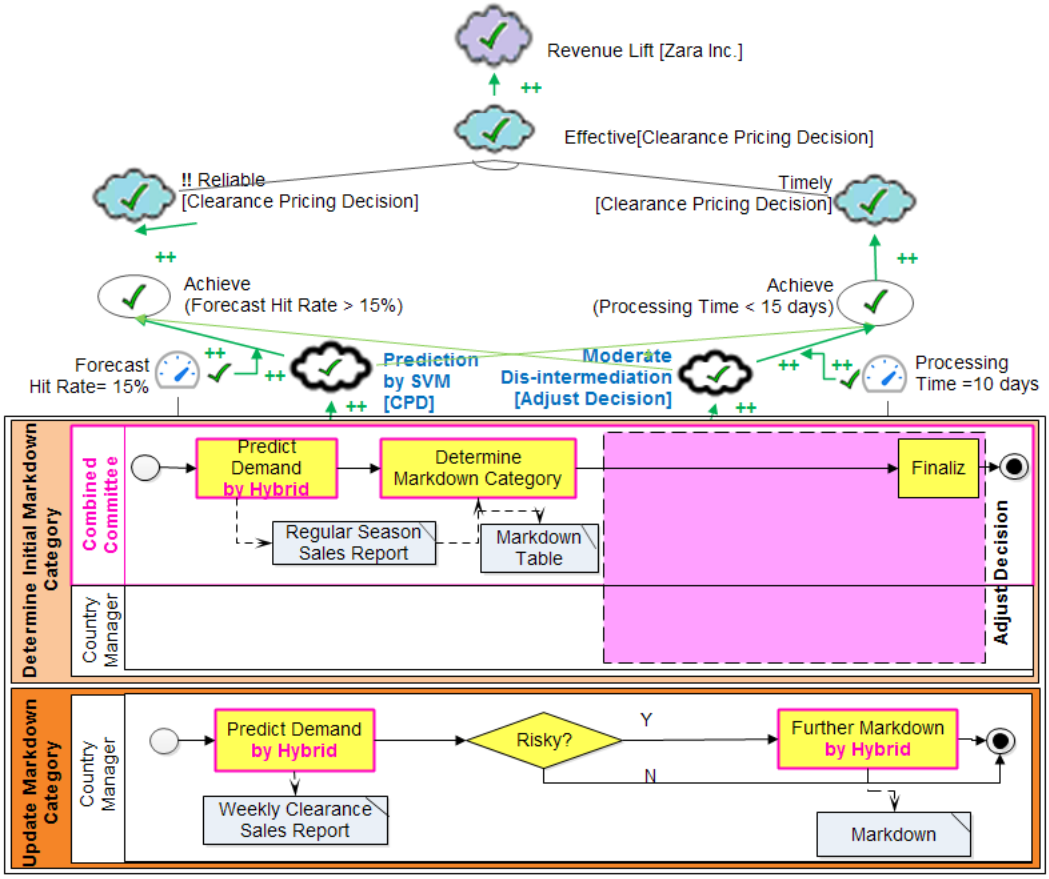


Figure 7.11. TO-BE Process 2: big data prediction + moderate dis-intermediation

# CHAPTER 8

# EVALUATION

In this chapter, we will evaluate our work by comparing with previous work according to whether applied project is the same or not.

## 8.1 Evaluation of GO-BigBAM

### Comparison with the Same Project

The example of automobile logistics comes from [80] in which FBCM methodology was applied. FBCM is a technological approach to defining and validating business requirements that are used to lead functional requirements. In this methodology, it adopted BSC (Balanced Score Card) which evaluates business performance in the perspective of not only finance, but also customer, process and learning & growth. FBCM has strong points to provide diverse points of views for strategy and extract system functions which are aligned with business goals and strategic goals. However, in the following perspective, IRIS can complement FBCM.

**Reliability:** Although FBCM can find causal relationships between business objectives and business process using KPI, it is not trivial to discover conflicting situations between processes because it does not have the concept of Satisficing contribution which can represent positive or negative contribution from child to parent. Moreover, FBCM considers only current status rather than an original status which has no problem. On the other hand, IRIS have Satisficing Relationship which has Make and Help for positive contribution, and Hurt and Break for negative contribution. For example, Build-To-Order (BTO), same with Market-To-Stock (MTO), is a strategy to reduce inventory of completely made cars, which does not need prediction and Market-

To-Stock (MTS) is also a strategy to reduce inventory with demand prediction. According to [86], the finished goods inventory level is lower in the dynamic system in which MTS operation and MTO operation are dynamically adjusted within the amount of the shortfall. However, we can conjecture that J did not use dynamic system before recognizing the inventory problem since members of SCM organizations assumed that coping with the gap between sales forecast and actual flexibly may lead to reducing inventory of complete cars according [80]. As a result, as an organization gets to increase the rate of BTO, the error rate of demand estimation also increased as correlation coefficient (0.7) showed [80]. For this situation, IRIS can detect the conflicting status using Business Goal-Process Alignment View, and moreover, root causes of the problem can be identified by looking into a detailed business process diagram and big query analytics view.

**Preciseness of Solutions**: Though FBCM can find business requirements and system requirements which are aligned with business goals, IRIS can drive more specific and fundamental requirements. For example, according to [80], FBCM elicits a business requirement for business processes, Visualize gap between sales forecast and actual flexibility, and a software requirement for a system, sales forecast confirmation and sharing to solve the problem, too much expectation for new products and aggressive sales forecast by the planning section cause salespersons' forcible sales to clean up inventory. However, it cannot fundamentally solve the problem, i.e., the same problem can arise if prediction is not carefully controlled. However, IRIS can find more specific and problem solving solutions for root causes such as ⬭Include Sales Data by BTO [Gather Demand History Data], ⬭Change Forecast Model Including BTO Rate [Forecast Demand] and ⬭Visualize the Sales Changes by BTO [Incorporate Sales Changes]. This is because IRIS uses not only diverse relationships such as Type [Topic] reference, Satisficing, Conflicting, but also

ontological concepts for Business Processes, explicit problems and solutions and complementary views.

**End-to-End Traceability for Analytics**: While FBCM only shows the traceability from business goals, business solutions, software requirements, IRIS extends the FBCM and each view represent end-to-end traceability in a different point of view. Moreover, in the spirit of Goal-Orientation, IRIS enables analyzers to explore alternatives and select one or more options among them in each different abstraction layer (business goal layer, business process layer and big data analytics layer), and to analyze impacts of the selection, which can show rationale on a decision.

**Comparison with Different Projects**

There are many frameworks similar to our solution such as URN-based BPMS [87], Business Analytics Frameworks [88], Business Intelligence Modeling [23] and Process Mining [45]. As Table 8.1 shows, each tool has its own strengths and weaknesses. While our solution has strong points on transformational insights modeling of problems and solutions and big data parallel processing among other things, the functions for KPI modeling and evaluation is weak. Detailed explanations for the criteria are following. *Confliction Discovery* is whether goal modeling which can detect conflictions is used or not. *Precision of Solutions* is whether detailed business processes are used to find solutions or not. *End-to-End Traceability* is whether a traceable line from business goal, business process, data analytics to data is shown. *Explicit Problems and Solutions* is whether explicit ontology for problems and solutions exists or not. *Big Data Parallel Processing* is whether distributed parallel processing on big data is provided or not. *Process Modeling* is whether detailed business processes are used or not. *KPI Evaluation* is whether each framework provides KPI

evaluation method or not. *Goal Modeling and Evaluation* is whether a framework uses a goal-modeling and evaluation mechanism or not.

Table 8.1. Comparison with existing frameworks (relative strength: . < some+ < + < ++)

| | Confliction Discovery | Precision of Solutions | End-to-End Trace-ability | Explicit Problems and Solutions | Big Data Parallel Processing | Process Modeling | KPI Modeling and evaluation | Goal Modeling and evaluation |
|---|---|---|---|---|---|---|---|---|
| IRIS | ++ | ++ | + | ++ | ++ | ++ | some+ | ++ |
| URN-based BPMS [87] | ++ | + | some+ | . | . | + | ++ | ++ |
| BA Framework [88] | ++ | some+ | ++ | . | . | . | ++ | ++ |
| BIM [23] | ++ | some+ | some+ | . | . | some+ | ++ | ++ |
| Process Mining [45] | . | ++ | . | . | . | ++ | + | . |

## 8.2 Evaluation of GO-BigDM

Zara's decision-making process for its shipping has been used not only for the purpose of illustrating the key concepts of IRIS's goal-oriented approach to modeling big data, but also for a (partial) basis of an empirical study. We feel the current study, albeit its small size, shows that IRIS supports the modeling of big data as a service for carrying out business analytics for Zara's shipment decision making.

We also feel that IRIS helps the process of modelling big data to become mostly traceable, if not all, while helping explore and select among alternatives in problems, solutions, business analytics and big data models. This, we feel, would help justify, and boost the level of confidence in, the quality of the resulting big data model. However, we have not shown that the potential problems and solutions indeed turn out to be real, key problems and solutions. This would require running big data on a real platform using real data, and, afterwards, monitoring the various real phenomena that are related to either problems or solutions – this seems difficult, if not infeasible in reality.

In this study, we wanted to see if IRIS can help with "connecting the dots" in modeling big data, from two different perspectives: 1) between, on the one hand, business goals, potential problems and solutions, and big data, on the other; and 2) among a variety of different types and sources of data. For the former, we feel that IRIS's approach helped traceably link the various concepts. For the latter, we feel that the three notions of big data quality helped. In particular, the notion of the comprehensiveness dimension helped consider incorporating newer technologies, including online and external sources of data, such as social networking and online marketplaces, so as not to omit potentially important business opportunities.

Concerning the notion of relevance, among the three notions of big data model quality, measuring the distance between entities and potential problems and solutions was relatively easy with internal data, but not so easy with external data. This was because external data can contain (highly) unstructured data and individual pieces of data, which would need to be first classified, to be related to some internal entities (classes).

IRIS's goal-oriented approach is intended for a systematic and rational process. Our observation in this regard is that this approach helped "connect the dots" among many important concepts, including business goals, potential problems and solutions with a business process, KPIs, analytics, and their alternatives. We thought this should be possible but initially were unsure of how, and now feel that IRIS's approach helps turn this possibility into more of reality.

We feel that IRIS's approach helps hypothesize and validate problems with the business process and solutions, which involves the use of KPIs, which in turn require the use of a big data model. In particular, the three organizational dimensions helped structure and explore data not only in the same but also across different dimensions of comprehensiveness. Additionally, the notion of

distance in relevance helped relate data across different dimensions, hence helping avoid omissions or commissions of data.

In the beginning of our empirical study, we had not thought of the use of the three dimensions in exploring potentially relevant data. We thought that they would be useful in organizing a large volume of data only. Now we feel that we have learned that the three dimensions also help systematically explore potentially useful data.

In order to have a sense of the technical feasibility, in particular, concerning the timely decision making, we ran big queries in Vertica, which is an analytic database management tool and supports both SQL and several types of NoSQL databases. The queries were run on a real big data platform of a consulting company, which maintains a large volume of data for a business whose characteristics are similar to Zara's. We could obtain answers to the queries within seconds.

### 8.3 Evaluation of GO-BP2UC

We have developed a Use Case Diagram as shown in Figure 8.2, using BPSec Extension (here after, BPSec) by the proposal in [89], and compared it against the Use Case Diagram in Figure 8.1, which results from the use of Go-BP2UC.

Part1 in Figure 8.2 shows the differences in granularity and relationships. Although OMG defines Use Case as a set of complete actions performed by a system [59], the granularity of Use Cases is not definitive. Reflecting this, Go-BP2UC allows for Tasks in BPMN to be transformed into one of the several concepts in Use Cases, split or even merged. For instance, Prepare Appointment or Conduct Appointment in Figure 8.1 by GO-BP2UC are collections of Tasks, while Review Lead in Figure 8.2 by BPSec is an individual Task. Additionally, Exclusive Gateway, which represents alternative options and (Exception Handling) Events, are not handled by BPSec. In contrast, GO-

BP2UC shows Cancel Appointment as an Extend use case for exceptional case. Moreover, Review Lead is commonly used for both Prepare Appointment and Conduct Appointment, hence being included in both by GO-BP2UC. In contrast, each Task in BPMN gets transformed into a Use Case by BPSec – e.g., Use Case (Prepare Quote) and Use Case (Modify Lead) by BPSec directly comes from Tasks in BPMN.

Moreover, in GO-BP2UC, Use Case (Conduct Appointment) and Use Case (Review Lead) will not be included in this result diagram if Sub-Process (Conduct Appointment) is not allocated to the target system, but BPSec has Use Cases corresponding to the BPMN elements such as Use Case (Review Lead) and Use Case (Modify Lead) although they are not assigned to the target system.

Moreover, in GO-BP2UC, Use Case (Conduct Appointment) and Use Case (Review Lead) will not be included in this result diagram if Sub-Process (Conduct Appointment) is not allocated to the target system, but BPSec has Use Cases corresponding to the BPMN elements such as Use Case (Review Lead) and Use Case (Modify Lead) although they are not assigned to the target system.

Part2 shows the difference when dealing with Actors. In BPSec, there is no distinction between Primary Participant that will become a system, and Secondary Participant that will become an environment.

In Figure 5.5 of Chapter 5, Prospect and Insurance Company are connected with Prepare Appointment via Message Flow, which implies that the two participants belong to different organizations. In BPSec, Prospect is directly connected the Schedule Appointment use case, whereas it is not represented by GO-BP2UC. In addition, due to the lack of distinction, it is possible

to have Quote Flow Actor as a Generalized Actor without any association with any Use Case. GO-BP2UC and BPSec show the difference in dealing with Non-Functional Requirements (NFR). While only Security NFR is considered by BPSec, other types of NFRs, such as Speed shown in Figure 5.4 of Chapter 5, can also be transformed by GO-BP2UC. That is, Collect Missing Information by Sending an Email and Collect Missing Information by Chatting Lively are from Operationalizing Softgoals for Save Time, which is a refinement of Speed. Collect Insured Person Data has inheritance relationship with them.

Table 8.2 summarizes a comparison between GO-BP2UC and BPSec, in terms of Cohesiveness, which means how well related elements are connected, Correctness, which means how valid transformations are, and Comprehensiveness, which refers to NFR modeling and transformation capabilities. While use cases can be related to each other in GO-BP2UC, via include, extend and inheritance, but not in BPSec, which can lead to low Cohesiveness. Also, BP-Sec can have 2 invalid actors, such as Prospect and Quote Flow, if activities or participants are partially allocated. Finally, GO-BP2UC has use cases which are derived from business process NFRs, but BPSec does not.

Table 8.2. A comparison btw GO-BP2UC and BPSec in terms of quality of Use Case diagram

| Quality Criteria | Comparison | All Allocation | | Partial Allocation | |
|---|---|---|---|---|---|
| | | GO-BP2UC | BPSec | GO-BP2UC | BPSec |
| Cohesiveness | Base Use Case | 3 | 9 | 2 | 9 |
| | Extend Use Case | 1 | 0 | 1 | 0 |
| | Include Use Case | 1 | 0 | 0 | 0 |
| | Inheritance Use Case | 2 | 0 | 2 | 0 |
| Correctness | Primary Actor | 2 | 4 | 2 (0 invalid actors) | 4 (2 invalid actors) |
| Compre-hensiveness | Use Case from NFR | 2 | 0 | 2 | 0 |

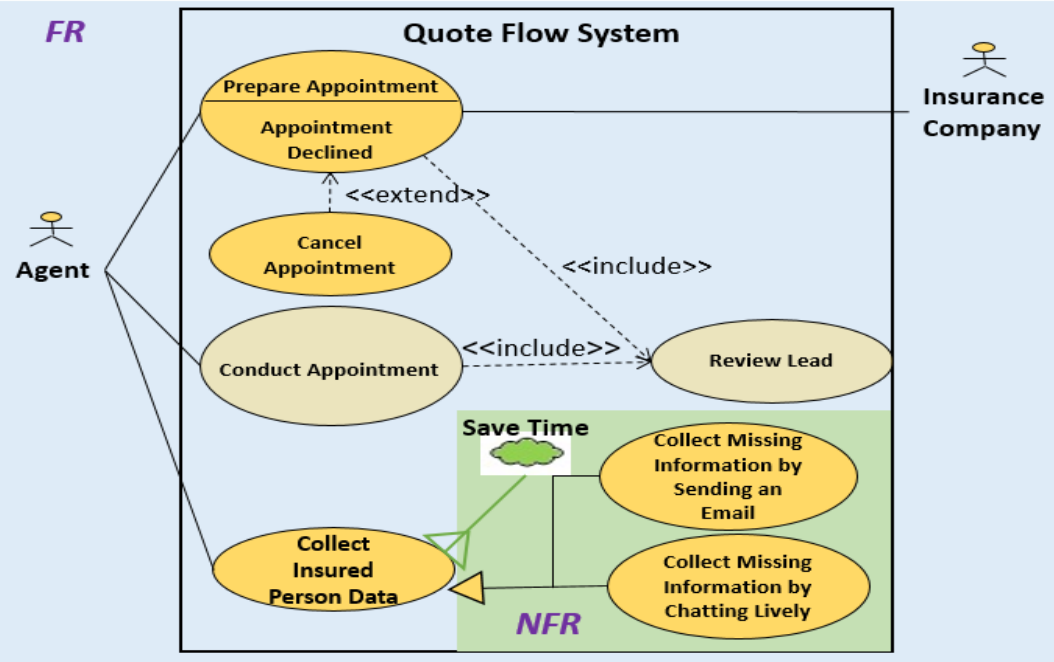


Figure 8.1. A Use Case diagram from GO-BP2UC

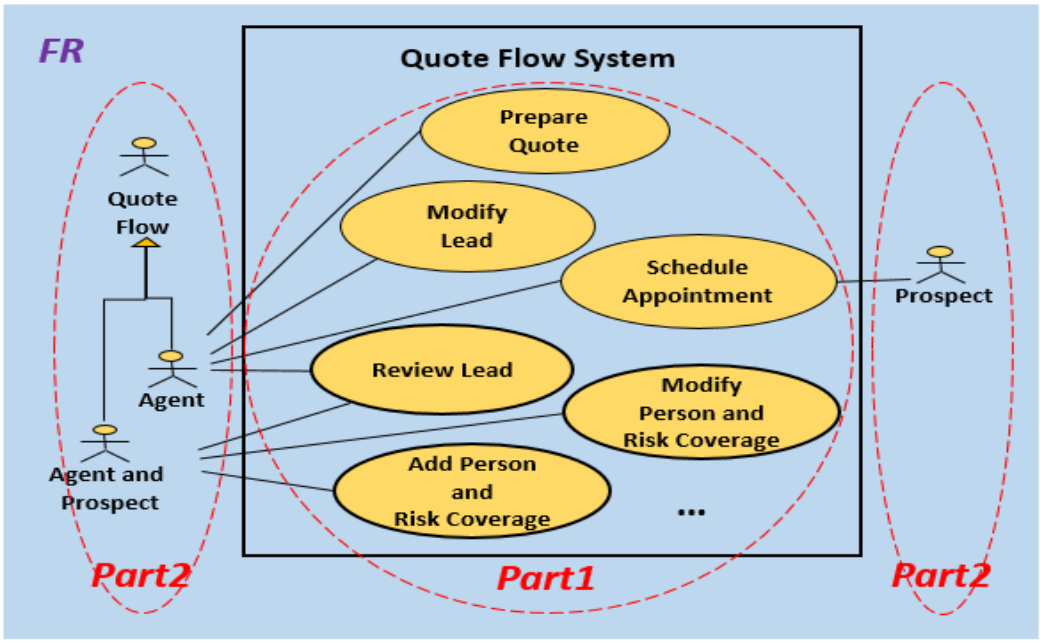


Figure 8.2. A Use Case diagram from BPsec

### 8.4 Threats to Validity

**IRIS: GO-BigBAM**

As internal threats, since we had to use only the results that [80] showed in the dissertation, the result may deviate from the real solution. To mitigate this threat, regarding to the claim, confliction discovery, we referred to the analytic data which existed in the paper [80], i.e., correlation analysis between BTO Rate and Sales Estimation Error Rate. In addition, [86] supports our claim by the discovery that a hybrid production system of MTO and MTS needs adjustment between them and a dynamic hybrid system performed better than static production to reduce logistics inventory. In addition, since we could not access the original data, there was a feasibility issue. To mitigate the threat, we utilized other similar data on demand prediction, and experimented correlation and prediction analysis using our solution. The data size of experiment was relatively small to say big data. Additionally, although [80] provides business process objectives, it does not describe details of detailed business processes, so we utilized general prediction process for demand forecasting from [90]. It may be different from the original situation.

As external threats, we analyzed only automobile industry domain in this paper, so the conclusion may be over generalized. Additionally, our solution can be applied only to the automobile industry.

**IRIS: GO-BigDM**

Our empirical study was based on publicly-available documents, including articles, whitepapers and information on web sites, but without access to any real (proprietary) documents. Hence, we could not use the real database schema that was being used, which inevitably might have led to biased results and conclusions. Our particular empirical study was carried out, involving a real big data platform of a consulting company, one of whose customers was quite similar, in terms of its

characteristics, but still was not real. Neither did our study involve access to the data about over 2,000 stores across 88 countries and around 11,000 items. Hence, our study may deviate from the real for the particular process. Also, saying the same for other kinds of business processes, especially in different application domains, would be an over-generalization.

Concerning the company's shipping decision, we tried to model the particular business process as faithfully to the actual process as possible, but, due to our inability to access the proprietary information, it is likely that turning the informal description of the business process into (semi-)formal BPMN process was not quite faithful. This might have resulted in some unfaithful representation of the actual process, hence the possibility of biased results.

Also, our empirical study, as the phrase suggests, did not involve real software practitioners or big data scientists who are working for the company, although we did seek some external opinion on our earlier work. Hence, the feelings we had may not be shared by the real software practitioners or big data scientists who are working for the company.

Throughout our empirical study, we have used our creative imagination, more than in one regard. One was concerning the extrapolation of a current trend into a future trend. We perhaps have taken a rather optimistic approach, i.e., integration of data from a variety of types and sources is not only possible but also quite probable. But it seems only time will tell if this will hold.

**IRIS: GO-BP2UC**

As internal threats, to compare ours with BPSec, we utilized their rules to our example. There may be misinterpretation while transforming a quote flow business process to a use case diagram. As external threats, we applied only a quote flow business process, so it may be over generalized.

# CHAPTER 9

# CONCLUSION

In this dissertation, we have proposed IRIS – a goal-oriented big data analytics framework. More specifically, IRIS includes GO-BigBAM conceptual model which connects business and big data, an evidence-based evaluation method for selecting the most effective solutions, a process for finding business problems and solutions – firstly by hypothesizing them, and secondly by validating them using big queries or big analytics – and a supporting tool which is implemented on top of Spark, real-time big data analytics framework. Although there are several limitations, at least through an empirical study, IRIS can help with big data business analytics in a value-added manner, i.e., comprehensive understanding on business and analytics, high priority, and fast decisions.

Additionally, we have proposed a goal-oriented approach to modeling big data as a service for supporting business analytics (BA). This goal-oriented approach of IRIS is intended to rationally and systematically help model big data, by exploring, and selecting among, alternatives in potential problems and solutions. Problems and solutions are hypothesized and validated, in consideration of important business goals, using big data analytics. This goal-oriented approach considers three aspects of big data model quality, namely, relevance, comprehensiveness and prioritization. In particular, three dimensions of comprehensiveness are proposed, for accommodating a ***v***ariety of types and sources of data, which are related and organized along the three organizational primitives. More specifically, IRIS's goal-oriented approach to modeling big data includes: 1) an ontology (or essential vocabulary), which explicitly recognizes goals, problems and solutions, business analytics, as well as the three comprehensive dimensions and the organizational

dimensions of a data model; 2) a process for using the ontology and finding and solving business problems and solutions, using big data analytics, which in turn require the use of a virtual big data model; and 3) a (partial, evolving) tool support. Through an empirical study, we feel that we have an initial demonstration that the goal-oriented approach can help boost the level of confidence in the quality of the resulting big data model.

Finally, we have presented a novel goal-oriented framework for transforming business processes in BPMN augmented with NFRs into Use Cases augmented with NFRs, in a traceable manner. This framework allows for exploration of alternative interpretations of the business process elements, utilizing an ontology of Intermediate Models, which facilitate the consideration of the similarities and differences between the source and the target. The framework also offers a set of transformation rules, which incorporate contextual information about the transformation elements, for both functional and non-functional goals. A comparative study has shown that this framework can help produce more cohesive, correct and comprehensive use cases.

Table 9.1 shows the resolutions of challenges which were mentioned in Motivation. Star Mark (*) indicates our contribution. For inaccurate scope, we adopted goal-orientation approach which enables to set business goals, resulting in the effect of reducing scope and helping find the most critical problems and effective solutions. Regarding to lack of business context around the data and lack of expertise to connect the dots, GO-BigBAM enables to connect diverse and important concepts for big data analytics with evidence-based reasoning method for goal achievement evaluation. For batch-oriented system Hadoop isn't enough, IRIS supports real-time processing by using Spark. When it comes to hard to derive actionable business insight from data, GO-BigBAM helps find next actions in business processes to achieve business goals. Moreover, GO-BP2UC

helps to automatically transform to-be business processes into use cases which are requirements of software system. For inadequate data, GO-BigDM helps maintain quality of big data.

Table 9.1. Resolutions of challenges by IRIS

| Challenges | IRIS Framework Solutions |
|---|---|
| Inaccurate Scope | * Set Business Goals for reducing scope<br>* Help find the most critical problems and effective solutions |
| Lack of Business Context Around the Data | * Connecting diverse and important concepts in GO-BigBAM<br>* Evidence(as big data analytics)-based Reasoning Method for Goal Achievement Evaluation |
| Lack of Expertise to Connect the Dots | |
| Batch-oriented system Hadoop isn't enough | Support real-time processing by using Spark |
| Hard to derive actionable business insights or requirements from data | * Helps find next actions in business processes to achieve business goals by GO-BigBAM<br>* Automatic transformation to-be business processes to use cases by GO-BP2UC |
| Inadequate data | * Help maintain quality of big data by GO-BigDM |

We plan to extend the capabilities of IRIS assistant – a prototype tool which is intended to support the use of the IRIS concepts – for example, semi-automatically translating them into big data queries. More studies – empirical and realistic case studies – are needed in a variety of application domains, in order to further determine both the strengths and weaknesses of IRIS. Big size data experiment is needed in a cloud environment experiment with a clustered mode.

## REFERENCES


[1] S. LaValle, E. Lesser, R. Shockley, M.S. Hopkins and N. Kruschwitz, *Big data, analytics and the path from insights to value,* MIT sloan management review, vol.21, 2013.

[2] Informatica and Capgemini, The Big Data Payoff: Turning Big Data into Business Value, 2016.

[3] Infochimps, *CIOs & Big Data: What Your IT Team Wants You to Know*, 2012.

[4] https://hbr.org/2013/02/get-the-maximum-value-out-of-y

[5] http://www.informationweek.com/software/information-management/vague-goals-seed-big-data-failures/d/d-id/1108384?

[6] A. van Lamsweerde, "Requirements Engineering in the Year 00: A Research Perspective," Proceedings of 22nd International Conference on Software Engineering, 2000. pp. 1-15.

[7] L. Chung, B. A. Nixon, E. Yu, J. Mylopoulos, *Non-functional Requirements in Software Engineering*, Kluwer Academic Publishers. 2000.

[8] E. Yu, P. Giorgini, N. Maiden and J. Mylopoulos, *Social Modeling for Requirements. Engineering*, The MIT Press. 2011.

[9] A. I. Antón, M. W. Michael and C. Potts, "Goal Decomposition and Scenario Analysis in Business Process Reengineering*". Advanced Information Systems Engineering*, 1994. pp 94-104.

[10] E. Yu and J. Mylopoulos, "Why goal-oriented requirements engineering," *Proceedings of the 4th International Workshop on Requirements Engineering: Foundations of Software Quality*. Vol. 15. 1998.

[11] L. Xing and S. V. Amari, *Fault Tree Analysis*, Handbook of Performability Engineering, Springer London, 2008. pp. 595-620.

[12] K. Ishikawa, *Guide to Quality Control*. No. TS156. I3713, 1982.

[13] S. Supakkul and L. Chung, "Extending Problem Frames to Deal with Stakeholder Problems: An Agent- and Goal-oriented Approach," *Proceedings of ACM Symposium on Applied Computing.* 2009. pp. 389-394.

[14] http://www.datasciencecentral.com/profiles/blogs/data-veracity

[15] T. White, *Hadoop: The Definitive Guide*, O'Reilly Media, Inc., 2015.

[16] https://en.wikipedia.org/wiki/NoSQL

[17] https://www.tutorialspoint.com/cassandra/

[18] https://docs.mongodb.com/

[19] A. Davidson and A. Or. "Optimizing shuffle performance in spark," University of California, Berkeley-Department of Electrical Engineering and Computer Sciences, Tech. Rep, 2013.

[20] *Apache SPARK*, http://spark.apache.org/

[21] J. Han, J. Pei and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[22] Business Process Model and Notation (BPMN) V2.0.2, OMG, http://www.omg.org/spec/BPMN/2.0.2/

[23] J. Horkoff, D. Barone, L. Jiang, E. Yu, D. Amyot, A. Borgida and J. Mylopoulos, "Strategic business modeling: representation and reasoning," *Software & Systems Modeling*, 2014. pp. 1015-1041.

[24] A. Kokune, M. Mizuno, K. Kadoya and S. Yamamoto, “FBCM: Strategy Modeling Method for the Validation of Software Requirements,” *Journal of Systems and Software*, 2007. pp. 314-327.

[25] S. Supakkul, L. Zhao and L. Chung, "GOMA:Supporting Big Data Analytics with a Goal-Oriented Approach," *IEEE BigData Con.*,2016. pp. 149-156.

[26] A. Pourshahid, D. Amyot and L.peyton, "Toward an integrated user requirements notation framework and tool for business process management." *e-Technologies, Int. MCETECH Conf. on*. IEEE, 2008.

[27] S. Nalchigar, Y. Eric and R. Ramani, "A Conceptual Modeling Framework for Business Analytics." *Conceptual Modeling: 35th International Conference, ER,* 2016. pp.35 - 49.

[28] A. Lapouchnian, Y. Yu and J. Mylopoulos. "Requirements-driven design and configuration management of business processes," *Int. Conf. on Business Process Management*, 2007.

[29] W. M. P. van der Aalst, Process Mining: Discovery, Conformance and Enhancement of Business Processes, Springer Science & Business Media. 2011.

[30] M. z. Muhlen and R. Shpairo, *Business Process Analytics*, Handbook on Business Process Management 2. 2010. pp. 137-157.

[31] W. M. P. van der Aalst, “Business Process Management: A Comprehensive Survey, “*ISRN Soft. Eng.* 2013. pp. 1-37.

[32] W. J. Kettinger, J. T. C. Teng and S. Guha, “Business Process Change: a Study of Methodologies, Techniques, and Tools,” *MIS Quarterly*, 1997. pp. 55-80.

[33] F. Caro, J. Gallien, M.Miranda., J. Torralbo, J. Corras, M.Vazquez, J. Calamonte, J.Correa, “Zara uses operations research to reengineer its global distribution process,” *Interfaces* 40(1), 2010. pp. 71-84.

[34] A. Davidson and A. Or. "Optimizing shuffle performance in spark." University of California, Berkeley-Department of Electrical Engineering and Computer Sciences, Tech. Rep, 2013.

[35] T. H. Davenport, “Competing on Analytics,” *Harvard Business Review* 84(1): 98. 2006.

[36] M. J. Berry and G. Linoff, *Data Mining Techniques: for Marketing, Sales, and Customer Support*. John Wiley & Sons, Inc., 1997.

[37] A. P. Sheth and J. A. Larson, “Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases,” *ACM Computing Surveys*: Vo. 22 No. 3, 1990. Pp. 183-236.

[38] J. M. Smith and D. C. P. Smith, “Database abstractions: aggregation and generalization,” *ACM Transactions on Database Systems (TODS)*: Vol. 2, No. 2. 1977. Pp. 105-133.

[39] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A. H. Byers, *Big Data: The Next Frontier for Innovation, Competition, and Productivity*.2110.

[40] T. White, Hadoop: The Definitive Guide. O’Reilly. 2012.

[41] J Dean, S Ghemawat**, “**MapReduce: simplified data processing on large clusters,” *Communications of the ACM (CACM)*: 51(1). 2008. Pp. 107-113.

[42] E. Alpaydin, *Introduction to Machine Learning*. The MIT Press. 2014.

[43] W. M. P. van der Aalst, “Business Process Management: A Comprehensive Survey, “*ISRN Soft. Eng.* 2013. pp. 1-37.

[44] W. J. Kettinger, J. T. C. Teng and S. Guha, “Business Process Change: a Study of Methodologies, Techniques, and Tools,” *MIS Quarterly*, 1997. pp. 55-80.

[45] W. M. P. van der Aalst, Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer Science & Business Media. 2011.

[46] D. Parmenter, Key performance indicators: developing, implementing, and using winning KPIs. John Wiley & Sons.

[47] P. Alireza, G. Mussbacher, D. Amyot and M. Weiss, "An Aspect-Oriented Framework for Business Process Improvement," In *E-technologies: innovation in an open world*, 2009. pp. 290-305.

[48] D. Amyot, S. Ghanavati, J. Horkoff, G. Mussbacher, L. Peyton and E. Yu, "Evaluating Goal Models within the Goal-Oriented Requirement Language," *Int. Journal of Intelligent Sys*. *25*(8). 2010. pp. 841-877.

[49] Y. Yijun, J. CSP Leite and J. Mylopoulos. "From Goals to Aspects: Discovering Aspects from Requirements Goal Models," *Proc., 12th IEEE Int. Conf. on Requirements Eng., 2004.* pp. 38-47.

[50] D. B. Reuben and M. E. Tinetti, "Goal-oriented patient care? an alternative health outcomes paradigm," *New England Journal of Medicine*, vol. 366, no. 9, pp. 777–779, 2012.

[51] M. Hoeper, I. Markevych, E. Spiekerkoetter, T. Welte, and J. Niedermeyer, "Goal-oriented treatment and combination therapy for pulmonary arterial hypertension," *European Respiratory Journal*, vol. 26, no. 5, pp. 858–863, 2005.

[52] H. Liu, H. Lieberman, and T. Selker, "Goose: a goal-oriented search engine with commonsense," in *Adaptive Hypermedia and Adaptive Web-Based Systems*. Springer, 2002, pp. 253–263.

[53] A. Faaborg and H. Lieberman, "A goal-oriented web browser," *Proc., ACM SIGCHI conference on Human Factors in computing systems*, 2006, pp. 751–760.

[54] R. Y. Wang & D. M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *Management Information Systems (MIS)*: Vol. 12, No. 4, 1996. pp. 5-33.

[55] T. J. Teorey, D. Yang, J. P. Fry, "A Logical Design Methodology for Relational Databases Using the Extended Entity-Relationship Model,**"** *ACM Computing Surveys (CSUR)*: Vol.18, No. 2, 1986. pp. 197-222.

[56] L. Lehtola, M. Kauppinen and S. Kujala. "Requirements Prioritization Challenges in Practice," *Product focused soft. process improvement*. Springer, 2004. pp. 497-508.

[57] A. Perini, F. Ricca and A. Susi, "Tool-Supported Requirements Prioritization: Comparing the AHP and CBRank methods," *Info. and Soft. Technology* 51(6). 2009. pp. 1021-1032.

[58] MDA Revision Guide 2.0, OMG, http://www.omg.org/mda/.

[59] OMG Unified Modeling Language TM (OMG UML) V2.4.1, OMG, http://www.omg.org/spec/UML/2.4.1/.

[60] R. M. Dijkman and S. M. M. Joosten, *Deriving Use Case Diagrams from Business Process Models*, Univ. of. TWENTE, 2002.

[61] P. Liew, K. Kontogiannis and T. Tong. "A Framework for Business Model Driven Development", *Proc.*, 12th *Int. Workshop on Software Technology and Engineering Practice (STEP'04)*, 2005.

[62] J. Berrocal, J. García-Alonso, C.Vicente-Chicote and J.M. Murillo, "A Pattern-Based and Model-Driven Approach for Deriving IT System Functional Models from Annotated Business Models", M. J. Escalona and G. Aragón (eds.) *Info. Sys. Dev*. 2014. pp 319-332.

[63] R. M. Dijkman, S. M. M. Joosten and O. F. Utopics, "An Algorithm to Derive Use Case Diagrams from Business Process Models", *Proc., 6th Int. Conf. on Soft. Eng. and App. (SEA)*, 2002.

[64] S. Štolfa and I. Vondrák. "Mapping from Business Processes to Requirements Specification", *Retrieved on 7th Aug*, 2008.

[65] A. Rodríguez, E. Fernández-Medina and M. Piattini, "Towards Obtaining Analysis-Level Class and Use Case Diagrams from Business Process Models", I. Y. Song and M. Piattini (ed.) *Adv. in Concept. Mod.–Chall. and Opp*. 2008. pp 103-112.

[66] S. Kherraf, É. Lefebvre and W. Suryn. "Transformation from CIM to PIM using patterns and archetypes", *Pro. 19th Australian Conf. on Soft. Eng.*, 2008. pp 338-34.

[67] A. Rodríguez, E. Fernández-Medina and M. Piattini, "Towards CIM to PIM transformation: from secure business processes defined in BPMN to use-cases", G. Alonso and P. Dadam (eds.) *Bus. Proc. Mag*. Springer Berlin Heidelberg, 2007. pp 408-415.

[68] T.R.Gruber, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing", *Int. Journal of Human-Commputer Studies*, 1995.

[69] P. Liew, K. Kontogiannis and T. Tong. "A Framework for Business Model Driven Development", *Proc.*, 12th *Int. Workshop on Software Technology and Engineering Practice (STEP'04)*, 2005.

[70] J.J.Jian, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy", *Int. Conf. Research on Computational Lingustics*, 1997.

[71] BPMN. Example. http://ackermann-consulting.de/business/process/management/bpmn-example/

[72] *EMF*, https://eclipse.org/modeling/emf/

[73] *Sirius*, https://eclipse.org/sirius/index.html

[74] *Apache SPARK*, http://spark.apache.org/

[75] *Apache Cassandra*, http://cassandra.apache.org/

[76] QVT, http://www.omg.org/spec/QVT/

[77] *BIRT*, http://www.eclipse.org/birt/

[78] J. Horkoff, D. Barone, L. Jiang, E. Yu, D. Amyot, A. Borgida and J. Mylopoulos, "Strategic Business Modeling: Representation and Reasoning," *Software & Systems Modeling*, 2014. pp. 1015-1041.

[79] S. Nalchigar, Y. Eric and R. Ramani, "A Conceptual Modeling Framework for Business Analytics." *Conceptual Modeling: 35th International Conference, ER*, 2016. pp.35 – 49.

[80] A. Kokune, M. Mizuno, K. Kadoya and S. Yamamoto, "FBCM: Strategy Modeling Method for the Validation of Software Requirements," *Journal of Systems and Software*, 2007. Pp. 314-327.

[81] Pearson's correlation, http://www.statstutor.ac.uk/resources/uploaded/pearsons.pdf

[82] H. Karau, A. Konwinski, P. Wendell and M. Zaharia, *Learning Spark: Lightning-Fast Data Analysis*, O'REILLY. 2015.

[83] F. Caro and J. Gallien, "Clearance Pricing Optimization for a Fast-fashion Retailer," *Operations Research* 60(6). 2012. pp. 1404-1422.

[84] *Accuracy*, https://en.wikipedia.org/wiki/Accuracy_and_precision

[85] T. H. Davenport and J. E. Short, "The New Industrial Engineering: Information Technology and Business Process Redesign," *Sloan Management Review* 31(4). Summer 1990. pp. 11-27.

[86] Z. G. Zhang, I. Kim, M. Springer, G. Cai, Y. Yu, "Dynamic pooling of make-to-stock and make-to-order operations", Int. Jour. of Production Economics, 2013. pp. 44-56.

[87] A. Pourshahid, D. Amyot and L.peyton, "Toward an integrated user requirements notation framework and tool forbusiness process management." *e-Technologies, Int. MCETECH Conf. on*. IEEE, 2008.

[88] S. Nalchigar, Y. Eric and R. Ramani, "A Conceptual Modeling Framework for Business Analytics." *Conceptual Modeling: 35th International Conference, ER*, 2016. pp.35 – 49.

[89] A. Rodríguez, E. Fernández-Medina and M. Piattini, "Towards CIM to PIM transformation: from secure business processes defined in BPMN to use-cases", *Business Process Manag*ement Springer Berlin Heidelberg, 2007. pp 408-415.

[90] J. T. Mentzer, *Sales Forecasting Management: A Demand Management Approach*, Thousand Oaks, 2005.

## BIOGRAPHICAL SKETCH

Eunjung (Grace) Park received her Bachelor's degree in the School of Computing from Soongsil University, and her Master's degree in Software Engineering from Sogang University in South Korea. She worked in the industry for 15 years as a programmer, a maintainer, an internal instructor for software development methodologies and an auditor of information systems at Hanjin Information Systems and Technology. She got certificates of qualification on Professional Engineering Computer System Application (PECSA) and Korea Certified Information Systems Auditor (KCISA) which were admitted by the Korean government in 2009. After joining The University of Texas at Dallas in August 2013, she has served as a Teaching Assistant for several undergraduate and graduate level courses including (Advanced) Requirements Engineering, Database Design, (Advanced) Software Architecture, Agile, C++ Programming and others, and as a Research Assistant in the summers of 2015 and 2016. Her interest research areas are Big Data, Business Analytics, Requirements Engineering, Business Process Management (BPM), Model Driven Development (MDD), Cloud Computing, IoT (Internet of Things) and IT-Business Alignment.

# CURRICULUM VITAE

## CONTACT INFORMATION


- Name : Grace Eunjung Park
- Email : g.e.park@ieee.org, ejpark095@gmail.com, ejpark95@naver.com


## RESEARCH AREAS

Big Data, Business Analytics, Requirements Engineering, Business Process, Big Data Analytics Tool, Spark, Software/Systems Architecture, Cloud Computing, IoT

## EDUCATION

**Ph.D of Software Engineering** (July 2013 ~ May 2017)
**Dissertation Title:** IRIS: A Goal-Oriented Big Data Business Analytics Framework
- University of Texas at Dallas, Richardson, Texas, U.S.A

**Master of the Graduate School of Information & Technology** (Aug. 2012)
**Thesis Title:** A Hybrid Agile Process for Package Type Development Projects
- Sogang University, Mapo-gu, Seoul, Republic of Korea

**Bachelor of Information Technology** (Feb. 1999)
- Soongsil University, Dongjak-gu, Seoul, Republic of Korea

## AWARDS

**Exemplary Employee of Hanjin Information Systems & Technology** (Nov. 3, 2011)
**Honors Scholarship of Sogang University** (2010, 2011)
**Honors Scholarship of Soongsil University** (1995, 1997)

## CERTIFICATION

**Professional Engineer Computer System Application (PECSA)** (Jun 1. 2009)
- Human Resources Development Service of Korea, Mapo-gu, Republic of Korea

**Korea Certified Information Systems Auditor (KCISA)** (Nov 5. 2009)
- Ministry of Public Administration and Security, Jongno-gu, Republic of Korea

## SKILLS

**Big Data and Analytics Framework:** Spark, Hadoop, MapReduce, Kafka
**Database**: (NoSQL) Cassandra, (SQL) Oracle, MS-SQL, MySQL
**Languages: (Programming)** Java, Scala, C++, C#, C, Python **(Script)** Java Script, JSP, (**Modeling)** UML, BPMN, **(Data Modeling)** ER, EER **(Query)** SQL, CQL, **(Others)** HTML, XML
**OS:** Windows, Unix, Linux
**Middleware:** IBM Websphere

## PUBLICATIONS

**A Modeling Framework for Business Process Reengineering Using Big Data Analytics and a Goal-Orientation**
- The 11th IEEE International Conference on Research Challenges in Information Science (2017)

**A Goal-oriented Big Data Analytics Framework for Aligning with Business**
- The 3rd IEEE International Conference on Big Data Computing Service and Applications (2017)

**IRIS: A Goal-oriented Big Data Business Analytics Framework on Spark for Better Business Decisions**
- The 4th IEEE International Conference on Big Data and Smart Computing (2017)

**Deriving Use Cases from Business Processes: A Goal-Oriented Transformational Approach**
- The 32nd ACM Symposium on Applied Computing (2017)

**Problem-Aware Traceability in Goal-Oriented Requirements Engineering**
- The 29th International Conference on Software Engineering & Knowledge Engineering (2016)

**Silverlining: A Simulator to Forecast Cost and Performance in the Cloud**
- CrossTalk Magazine, The Journal of Defense Software Engineering, USA (2015)

**Management Information Systems** (Jan 1. 2011)
- Seoul Metropolitan Office of Education, Republic of Korea

## PROFESSIONAL EXPERIENCE

**Teaching Assistant** (2013 Fall ~ 2017 Spring), University of Texas at Dallas, U.S
- Courses: Requirements Engineering, Software Architecture, Software Engineering, Agile, Database Design, C++ Programming

**Research Assistant** (2015 and 2016 Summers), University of Texas at Dallas, U.S
- Mapping Business Processes to Use Cases: a Goal-Oriented Approach

**Employee** (Feb 1999 – June 2013)
Hanjin Information Systems & Technology (HIST), Seoul Gangseo-gu, Korea

- **Instructor for internal education courses** (June 2008 – June 2013)
- *The HIST Hybrid Agile Method*: Establishment of the HIST Hybrid Agile Process with Agile and Waterfall Model. An education targeting for members of the Inha Hospital International Medical Center system development project
- *Agile Methods*: basic understanding of Agile Methods, support XP, SCRUM practices and practical training
- *A Component Based Development Method, MARMI-III*: basic understanding of CBD methodologies, MARMI-III process explanation and practical training

- **Quality Assurance** (June 2008 – June 2013)

- *The System Development Project of Inha International Medical Center, QA*: the Hybrid Agile methodology, a VB.NET based development project for the Medical Information system including checkups and hospital back office, Hybrid Agile Process education and coaching
- *The System Development Project of Job Korea Sales Management, QA*: the Information Engineering methodology, a VB.NET based development project for sales management system, project auditing
- *The Integrated Management System Development Project for Water Supply of Korea Water Resources Corporation, QA*: the Information Engineering methodology, a Java based integral information system project for local water supply facilities, project auditing
- *The Shipment Operation Development Project of Glovis, QA*: the Information Engineering methodology, a VB.NET based system development project for shipment operation, project auditing
- *The Groupware Development Project of S-Oil, QA*: Information Engineering Methodology, an ASP.NET based system development project for groupware, project auditing
- *The Administrative Tasks Development Project of Ministry of National Defense, assistant QA*: the National Defense's CBD methodology, a Java based development project, output analysis support
- *The Flight Medical Care Information System Development Project of Korean Airline, QA*: the Information Engineering methodology, a VB.NET based system development project for medical care information system, project auditing
- *Internal Quality Audit for Korean Air System Maintenance Service*: Establishment of a Framework and metrics for SM Service quality improvement
- *ISO9001 Internal Quality Audit*: biannual internal quality audit

- **System Maintenance** (Jan 2001 – May 2008)

- *Topas' Airline Ticket Reservation/Ticketing System*: an airline ticket reservation and ticketing system which travel agencies use, Java based 3-tier System, program development and system maintenance
- *Topas' Portal System*: portal system targeting for travel agencies and travelers, ASP programming language, program development and system maintenance.

- **Development Projects** (Aug 1999 – Dec 2003)

- *Mokpo City's Electronic Tourism System Development Project*: Uni-script language based development project, requirement analysis and program development
- *Top Flight's Intranet System Development Project*: an ASP based development project for Top Flight's Intranet system, requirement analysis and program development