A Big Data Conceptual Model to Improve Quality of Business Analytics

Grace Park[1(✉)], Lawrence Chung[2], Haan Johng[2], Vijayan Sugumaran[3], Sooyong Park[4], Liping Zhao[5], and Sam Supakkul[6]

[1] State Farm, Richardson, USA
g.e.park@ieee.org
[2] University of Texas at Dallas, Richardson, USA
{chung,HaanMo.Johng}@utdallas.edu
[3] Oakland University, Rochester, USA
sugumara@oakland.edu
[4] Sogang University, Seoul, Republic of Korea
sypark@sogang.ac.kr
[5] The University of Manchester, Manchester, UK
liping.zhao@manchester.ac.uk
[6] NCR Corporation, Irving, USA
ssupakkul@ieee.org

Abstract. As big data becomes an important part of business analytics for gaining insights about business practices, the quality of big data is an essential factor impacting the outcomes of business analytics. Although this is quite challenging, conceptual modeling has much potential to solve it since the good quality of data comes from good quality of models. However, existing data models at a conceptual level have limitations to incorporate quality aspects into big data models. In this paper, we focus on the challenges cause by Variety of big data propose IRIS, a conceptual modeling framework for big data models which enables us to define three modeling quality notions – relevance, comprehensiveness, and relative priorities and incorporate such qualities into a big data model in a goal-oriented approach. Explored big data models based on the qualities are integrated with existing data grounded on three conventional organizational dimensions creating a virtual big data model. An empirical study has been conducted using the shipping decision process of a worldwide retail chain, to gain an initial understanding of the applicability of this approach.

Keywords: Big data conceptual model · Big data modeling quality · Goal-oriented big data · Business analytics · Goal-orientation

1 Introduction

Big data has quickly been embraced by all walks of life, including businesses, governments, and academia. The big data revolution is unprecedented, due to the promises and hopes built around it. Yet, behind the façade of big data is the simple notion of data – the

F. Dalpiaz et al. (Eds.): RCIS 2020, LNBIP 385, pp. 20–37, 2020.
https://doi.org/10.1007/978-3-030-50316-1_2

data that is characterized by many Vs (Volume, Velocity, Variety, Veracity, Value, and possibly more) that require new technologies to unleash its power.

Business analytics (hereafter, BA) is one of the key areas that can benefit greatly from big data. With new business insights gained by big data from diverse sources and types, BA helps make better business decisions and improve business processes. Yet, the quality of BA can only be as good as the quality of the big data it uses. With good quality big data, BA can accurately identify important business concerns, trends and opportunities, and useful business insights, which, in turn, can lead to good business decisions.

There are many challenges to ensuring the good quality of big data for business analytics. Among them, this paper investigates challenges caused by Variety of big data because the ability to integrate more diverse sources of data drives successful big data outcomes [22]. However, it is crucial for organizations to adequately select external data and incorporate it with internal data fitting for business purposes instead of just blindly adding more data. According to prior research, since the amount of data is huge, it is hard to evaluate big data quality within a reasonable time [1]. If there is irrelevant data, organizations waste more time and money. Additionally, data from diverse sources bring heterogeneous types of data and complex structures causing integration problems with existing data [2]. Furthermore, as current big data analytics focuses too much on data processing itself, it has limitations to support business concerns [3], thus, it is hard for external data to semantically relate to business concerns not knowing the rationales of its selection.

A conceptual model that helps decide project scope at a high-level abstraction and establish a structural organization has great potential to explore many big data entities and relationships for business analytics without collecting or processing concrete data. Some researchers [4, 5] recognize the potential of conceptual modeling for big data, but they are initial stages for business analytics, especially for diversity; there is little work to prescribe how conceptual modeling for big data pursue good quality of business analytics in the perspective of Variety. According to [6], since design decisions positively or negatively affect quality attributes, the process of selection should be rationalized in terms of quality attributes. However, not only existing conceptual models for big data [9] but also traditional data models [8] have limitations to express the process of rational selections on which data entities and relationships are needed for important business considerations resulting in omissions or commissions of big data. Additionally, there is little work to suggest how the new entities of big data from external sources can be incorporated with existing data causing a more complex structure. Moreover, it is hard to identify the origins of the external big data i.e., where a data entity came from, which leads to breaking a balanced view on data diversity and negatively affects reliability.

To address these problems, this paper proposes IRIS, a systematic approach for the quality of Variety of big data in a conceptual level. First, it defines an ontology at which we adopted a goal-orientation approach for rational selections and incorporate concepts such as business problems, solutions on top of Extended Entity-Relationship (EER) model [7]. Second, this approach selects big data modeling quality attributes that are closely related big data quality to support Variety and defines three quality aspects:

relevance, comprehensiveness, and relative priorities, and integrates them into a selection process as selection criteria. These aspects stipulate that the data and relationships between the data should be relevant to their use, comprehensive enough to support balanced views for business decision-making, and prioritized so that their importance is clear. Finally, this paper utilizes conventional three organizational dimensions: Classification/Instantiation, Generalization/Specialization, and Aggregation/Decomposition to integrate existing data entities and the new external entities of big data creating a virtual big data model that helps new or external opportunities.

The key contributions of this paper are as follows: 1) the new ontology including entities and relationships for big data business analytics in which goal-orientation is integrated with big data model enables to explore diverse data entities and select relatively optimal ones into a virtual big data model dealing with good quality of Variety, and eventually, good quality of business analytics; 2) the three attributes to properly resolve challenges caused by Variety provide selection criteria for the selection process of data providing decision rationales both for quality of data and data model; 3) the three organizational dimensions which handle the integration problem with existing data helps reduce complexity.

This paper is organized as follows. Section 2 and 3 describe related work and a running example respectively. Section 4 presents IRIS, a goal-oriented conceptual model for big data models which consists of ontology, the three qualities and organizational dimensions of big data models. Then, Sect. 5 shows an application of the approach to the running example. Section 6 provides a discussion including the limitations of this approach. In the end, a summary of contributions is given, together with future work.

2 Related Work

Conceptual modeling is a key related area. Conventional data models are ER [8] and EER [7], which enable conceptual data modeling. As big data technologies are prevalent, researches on conceptual modeling for big data are increasing. For example, [9] and [10] provides a big data modeling methodology for Apache Cassandra and schema inference for JSON datasets providing an inference algorithm, respectively, and [11] suggests domain ontology as a conceptual model for biomedical informatics to support data partitioning and visualization. Although these works contribute conceptual modeling for big data, they have limitations to build required qualities into big data models. However, our solution is suitable for addressing the problem since it utilizes a goal-oriented approach, combines business concepts, and the collaboration of three dimensions of quality and organization.

In addition, big data quality is another related area. Some researchers have been investigating to improve the quality of big data. For instance, [12] suggests quality in use model for big data providing three data quality dimensions, i.e., Contextual Consistency, Operational Consistency, and Temporal Consistency. Additionally, [13] suggests how to deal with data granularities for data quality evaluation and analyzes data quality dimensions. [14] addresses the quality of big data at the pre-processing phase to support data quality profile selection and adaptation. The key distinctions are 1) this work focus on big data modeling qualities which affect big data quality, and 2) the

notion of a virtual repository through a federated approach is applied to our work. A virtual repository is intended to offer easy access to (usually geographically) distributed, and oftentimes independent and heterogeneous, databases, through a single (virtual) database with a single set of mechanisms for accessing the database [15]. Creating a virtual database involves integrating oftentimes 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. We also adopt the three abstraction principles of Classification/Instantiation, Generalization/Specialization and Aggregation/Decomposition [16].

Finally, the area of goal-oriented requirements engineering is importantly related to our research. Some research (e.g., [17, 26]) has proposed a conceptual model for business analytics, considering business goals and strategies, although they do not use big data. The former defines three complementary modeling views that help find analytics algorithms and data preparation activities aligned with enterprise strategies. The work is similar to ours from the perspective of using goal-concepts and conceptual models, but our suggestion is more focused on conceptual big data models considering data modeling qualities and organizations. Additionally, our previous work [18, 19] suggests a goal-oriented big data business analytics framework to bridge the gap between big data and business. Although this paper utilizes the basic concepts such as business goals, problems, and solutions, this work is more focused on big data modeling and its quality, exploring diverse data models, and evaluating them using trade-off analysis from the perspective of Variety. Especially, this paper provides guidance on how to quantitatively calculate relevance using structural relevance and semantic relevance.

3 A Running Example: Shipping Decision

We utilize a shipping decision-making process [21] from Zara, Inc. (hereafter, Zara) which is a worldwide fashion retailer as a running example throughout the paper for illustrating the key concepts of IRIS's goal-oriented approach to modeling big data in the perspective of Variety, as well as from the perspective of an empirical study, concerning the applicability of IRIS's approach. Due to the inability to access data inside the company, EER diagrams in this example have been reconstructed based on real case descriptions of [21] and other diagrams are based on our imagination.

For its global distribution, Zara'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.

4 IRIS: A Conceptual Modeling Framework for Quality of Big Data Business Analytics

4.1 IRIS Ontology

The main concepts in IRIS and relationships between them that are needed for the purpose of communication and modeling about big data are depicted in Fig. 1 in which thick lines are extended elements from [18].

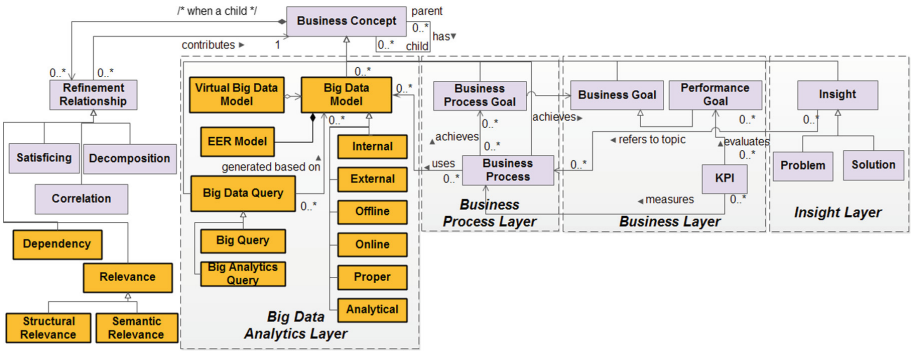

Fig. 1. IRIS ontology for big data modeling.

In IRIS, every concept is treated as a goal which can be refined more specific sub-goals with corresponding relationships. To represent these common characteristics, we define Business Concept as a root element with parent/child relationships. A parent has many children and vice versa. In the case of a child, it has many Refinement Relationships toward its parents but, a Refinement Relationship toward a parent is one since a specific refinement of a child toward a parent is only one. For example, to Increase Global Revenue of Zara *(Business Goal)* as a parent, Effective Shipment Decision *(Business Process Goal)* or Effective Clearance Pricing *(Business Process Goal)* can be explored as children and Weakly Positive of Effective Shipment Decision and Positive of Effective Clearance Pricing are the relationships towards the parent goal respectively. This notion is well explained in [19] about how Big Data Query can evolve from abstract business questions to specific big data analytics query statement using this goal concept. A more specific example of big data models will be shown in Sect. 7.

Additionally, IRIS enables to model both Problem and Solution which are treated as insight on whether a Business Concept respectively makes positive and negative contributions towards achieving a goal. They are validated by the results from Big Data Queries or KPI (Key Performance Indicator). For example, Low Hit Rate on Shipment Decision is identified as a Problem since 15% of Estimation Hit Rate on Shipment Decision (KPI) is lower than 30% of Performance Goal by the result of corresponding big analytics queries.

Concerning the big data part, in particular, Big Data Model is specialized in the three comprehensive dimensions, i.e., Internal-External, Offline-Online, and Proper-Analytical. Any kind of Big Data Model – be it the whole or its parts – can be modeled

with EER Ontology. Here, the whole big data model would be a virtual big data model (Sect. 4 will describe details), consisting of both the native and foreign big data models. Since Big Data Model itself is inherited from Business Concept, it has the ability of Refinement Relationship towards Insight that includes Problem and Solution. In this case, Insight is considered as a goal to achieve. This allows modelers to select a relatively best big data model that helps find insights on business problems and solutions. In IRIS, data modeling elements and data insight elements co-exist. A Big Data Query is generated based on Big Data Model and it also helps select Big Data Model. It is specialized in Big Analytics Query and Big Query, and they are distinguished from whether they need analytics such as machine learning or statistical results or not respectively. Additionally, Dependency and Relevance relationships are identified and they play an important role to see diverse views and find the most relevant data elements.

4.2 Quality Definitions of Big Data Model

In improving the quality of business analytics, it is critical to capture important concepts in a specific business domain, their relationships, and constraints on both the concepts and relationships and represent them in a model. Building an ontology as comprehensively as possible helps improve the quality of the data model by reducing omissions and commissions in the model. However, it does not mean we can always collect entire data for the whole of the ontology. Collecting data for some ontology might be too expensive, too time-consuming, or even impossible in the real world. Due to the issues, improving the quality of the data model can help enhance the quality of data.

The key emphasis in IRIS lies in, among other Vs, the notion of Variety, hence the need for accommodating a variety 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, sensor networks, etc., that are increasingly becoming part of every walks of life. But then, the growth in the volume of data seems remarkably large and also in their velocity, 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 a (possibly prohibitively) increased cost for managing the data.

There are two categories of quality; one for data quality such as accuracy or consistency [20] and the other is data model quality such as understandability or maintainability [25]. Among them, we propose the following three notions of data modeling quality - relevance, comprehensiveness, and relative priorities. The rationales behind the selection are 1) those modeling quality are closely related to data quality, so a given data model can impact data quality [23], 2) they help explore diverse data entities and select ones to achieve good quality of Variety, and 3) this work focuses on data qualities that can be dealt in a conceptual level.

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

- *Relative priorities* of data for determining how to allocate a 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.

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 Fig. 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.

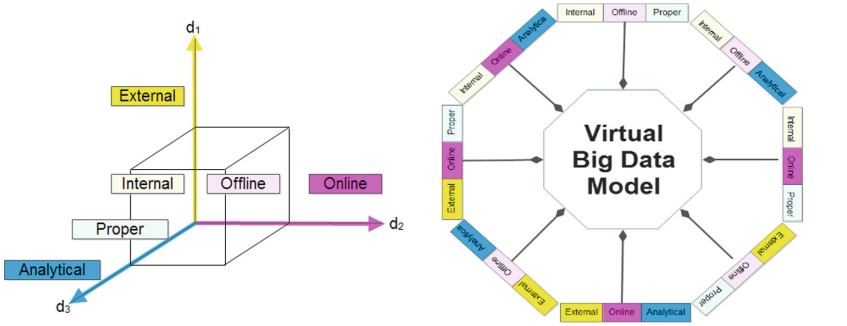

Fig. 2. Three big data comprehensiveness dimensions and the resulting eight squares.

Regarding the perspective of data, there are many dimensions such as structured/unstructured or descriptive/predictive, however, in this paper shows the following three dimensions as an example.

- *Internal/External dimension*: to consider not only data that is *internal* to a business (e.g., in the business's local data center) 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 different sources, as the result of the cross product: {Internal, External} X {Offline, Online} X {Proper, Analytical}. Big data entities explored by utilizing this view are selected through the next Relevance and Prioritization quality attributes.

Relevance

This aspect is about the utility of a data element, which can be used as a criterion 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 link between e and some Problem or Solution. More specifically, a data model (element), e, is said to be structural 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 semantic relevant if e makes a positive or negative 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 Fig. 2 (Fig. 3).

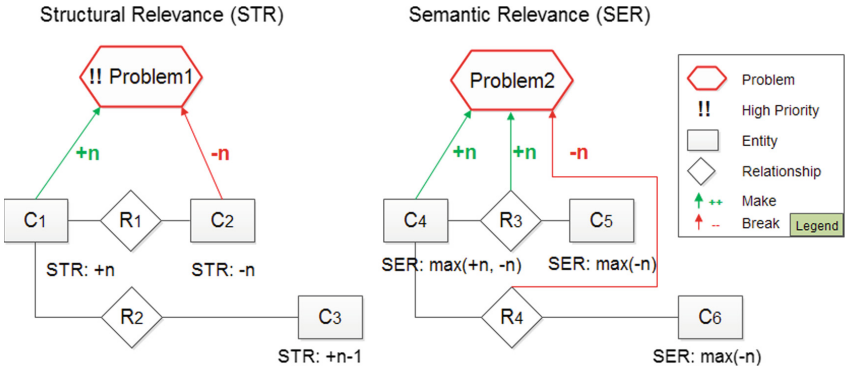

Fig. 3. Structural relevance and semantic relevance.

In Fig. 2, each Ci denotes a class or an entity in EERD (Enhanced Entity Relationship Diagram). For Structural Relevance (STR), C1's distance to Problem is smaller than C3's, hence more relevant. Now, C1's distance is the same as C2's distance, but C1's contribution is stronger than C2's contribution, hence more relevant. Regarding Semantic Relevance (SER), C4's relevance is the maximum value between R3's contribution and R4's one since C4 is semantically related to Problems2 with R3 and R4. More detailed examples will be explained in Sect. 5.

Prioritization

This aspect is useful in determining what data to incorporate into the virtual big data, how much effort 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 which are hypothesized, then modeled.

The priorities of data are inherited from the priorities of their respective potential Problems and Solutions for the data to be intended to validate. For example, in Fig. 2, there are two Problem1 and 2, and C1 or C4 are the most relevant to validate them respectively. However, C1 can be selected as Virtual Big Data Model since Problem1 has a higher priority which is denoted as !!. While priorities are propagated downward, the priorities of lower-level refinements can change in either direction. In particular, a trade-off analysis can lead to the change in the priorities of those operationalizations that overall make strongly negative contributions to achieving higher-level goals. Section 5 will show more detailed examples.

4.3 Organizational Dimensions

Now, we introduce the three organizational dimensions, which are intended to help structure a large variety of big data, possibly in a huge volume arriving at a fast velocity. All these three dimensions can be used to explore, and relate, data in the same or across different dimensions of big data comprehensiveness.

- *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 that 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 worldwide 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 hot fashion trend and could be used in rating Zara's products.

5 IRIS in Action for Quality of Big Data Conceptual Modeling

IRIS's goal-oriented process can help develop a virtual big data model to support the BA for Zara's shipping decision process of the running example in Sect. 2.

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, and the current operating data model, as shown in Fig. 4 and Fig. 5. Afterward, the team finds and establishes one or more business goals, here, Increase Global Revenue (the top portion in Fig. 4). There are many ways to increase revenue, including effective marketing and effective shipment decision, which the team considers important, hence treating Effective [Clearance Pricing] and Effective [Shipment Decision] as goals to be achieved.

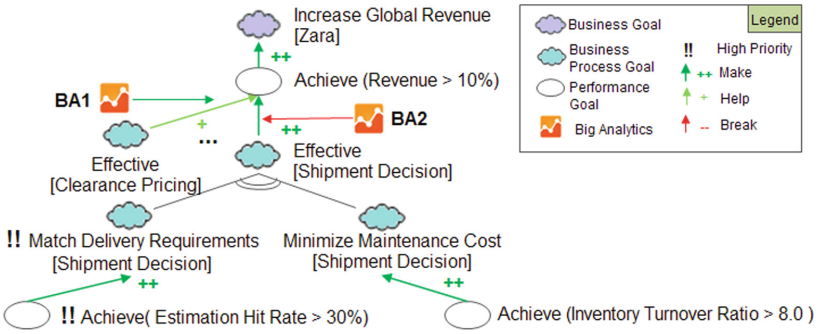

Fig. 4. Diagnostics for shipment decision.

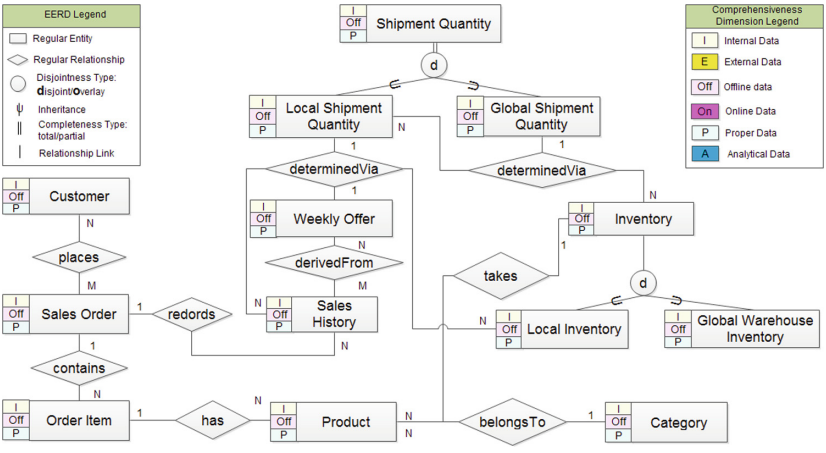

Fig. 5. Internal-offline-proper shipment data with EERD.

Between these two, the team considers Shipment Decision is a more critical factor to increasing revenue, using BA1: correlation between clearance decision and revenue and BA2: correlation between shipment and revenue, so 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.,

!!Match Delivery Requirements [Shipment Decision], and Minimize Maintenance Cost [Shipment Decision]. Among these two goals, !!Match Delivery Requirements [Shipment Decision] has a higher priority by the team's decision. These two goals are expressed in measurable terms, here, Performance goals – !!Achieve (Estimation Hit Rate > 30%) and Achieve (Inventory Turnover Ratio > 8.0) respectively. As the priority of a child goal is inherited from its parent goal, !!Achieve (Estimation Hit Rate > 30%) is a higher priority than the other.

As in Fig. 5, 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/specialization and aggregation/decomposition. This data model represents only an Internal-Offline-Proper data model from the perspective of a virtual big data model.

Finding Internal Data Entities to Validate Hypothesized Problems/Solutions

The team now hypothesizes potential problems with the current shipment decision process and their sub-problems (root causes) and potential solutions. As Fig. 6 shows, it can be done from the perspective of the Performance goals, here !!Achieve(Estimation Hit Rate > 30%) and Achieve (Inventory Turnover Ratio > 8.0). Zara's team considers two potential problems, High Difference of Demand Estimation [Shipment Decision] and Long Delivery Time [Shipment Decision], as the most likely and important problems because of the results from BQ1: Estimation Hit Rate on Shipment Decision (15%) and BQ2: Inventory Turnover Ratio on Shipment (2.5). As with problems, potential solutions can also be hypothesized - here, !!Reliable Prediction Model on Demand [Decide Quantity Request, Aggregate Reconcile] and Reliable Prediction Model on Delivery [Overseas Delivery]. However, the former solution has more importance since it inherits the priority from !!Match Delivery Requirements [Shipment Decision] through !!Achieve(Estimation Hit Rate > 30%).

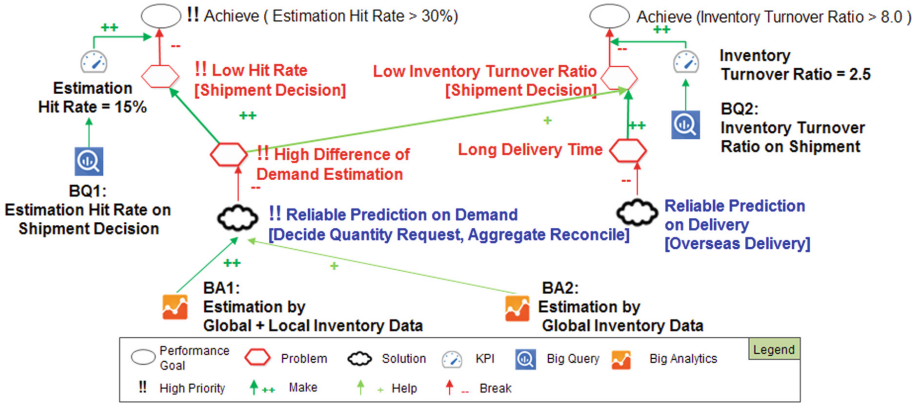

Fig. 6. Hypothesized problems, solutions, and Big Analytics Queries with priority.

If only the Internal-Offline-Proper data model is used, then two queries, BA1: Estimation Adding Global + Local Inventory Data and BA2: Estimation Adding Global Inventory Data, might be considered for the purpose of validation in Fig. 6. The team considered not only global but also local inventory affects the accuracy of the estimation model, so the team added BA1 which uses a superset of the query elements for the estimation model. Then, between the queries, the former would have a stronger contribution (++) than the latter (+) towards validating the potential problem and the solutions they are relevant to, namely, !! High Difference of Demand Estimation [Shipment Decision] and !! Reliable Prediction on Shipment Quantity respectively. More details for finding problems/solutions are in our previous work [18].

To create Big Analytics Query, first of all, they need to include data entities from Internal-Offline-Proper data model. Figure 7 shows an example of which data entities are the most relevant to validate Reliable Prediction on Shipment Quantity using Structural Relevance and Semantic Relevance. For Structural Relevance (STR), Shipment Quantity is the starting point with $+3$ points and STR will be reduced by 1 whenever a relationship will be passed except for inheritance relationship. The STR of Local Shipment Quantity is still $+3$ since it inherits from Shipment Quantity, and the STR of Sales Order is $+1$ since it has two relationship between Local Shipment Quantity. For Semantic Relevance (SER), the relationships of disjoint inheritance, determined Via relationship, and records contribute $+3$, $+3$, and -1 respectively to validate Reliable Prediction on Shipment Quantity. The SER of an entity is the maximum of SERs of related relationships. The total Relevance is the sum of STR and SER. Thus, the least related Sales Order $(+1+(-1))$ is not included.

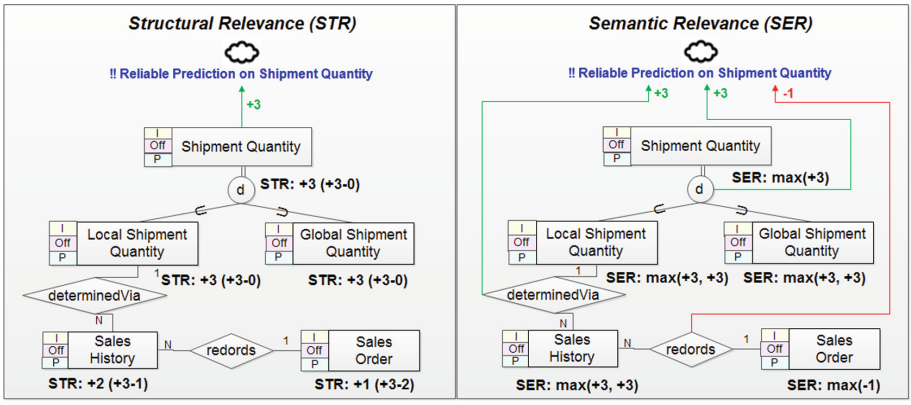

Fig. 7. An example of structural relevance and semantic relevance.

Expanding Existing Data Entities

Let us assume that the innovation team thinks it is important to explore new and external opportunities in validating the potential solutions. By using the Internal-Offline-Proper Dependency Diagram in Fig. 8, the team identifies a variety of other types and sources

of entities in the three comprehensiveness dimensions, here External, Online, and Analytical. The Dependency Diagram in Fig. 8 left box represents the dependent relations between entities which are easily inferred from input-output relationships in BPMN and this is utilized as a reference to explore diverse entities.

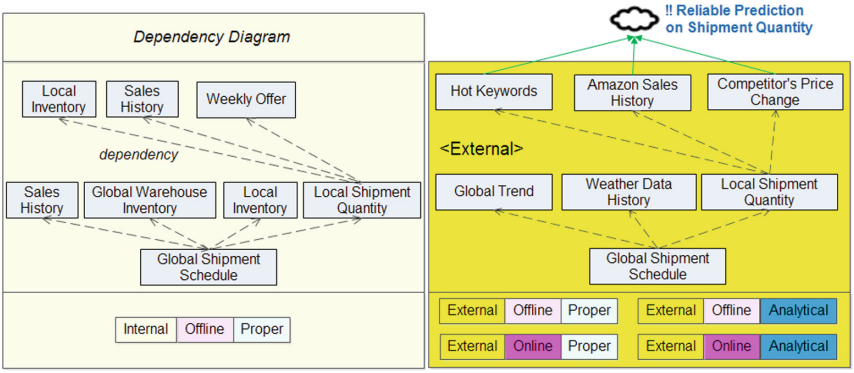

Fig. 8. An example of candidate entities derived from external dimensions.

For example, the team explores the external Amazon Sales History as being relevant to the internal Sales History (this is Zara's own), in Fig. 8 bottom right box. 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 chains of entities associated with Sales History, here Local Shipment Quantity and Global Shipment Schedule, are copied and associated with Amazon Sales History.

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

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 the Global Warehouse Inventory, hence likely to be incorporated into a virtual data model later.

Build an Initial Virtual Big Data Model

Zara's team integrates those entities and relationships from every possible dimension, which were derived 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 Fig. 9.

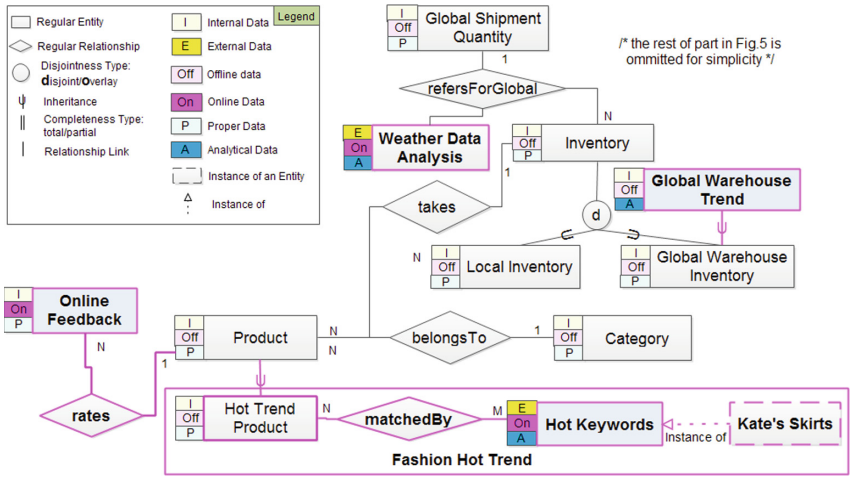

Fig. 9. Virtual big data model, consisting of entities and relationships from multiple sources, for shipment decision.

Figure 9 shows the initial virtual big data model, which consists of existing entities and newly-added entities. In this model, some of the Internal-Offline-Proper entities are integrated with new entities. For example, Global Shipment Quantity in Fig. 9 is related to the Global Warehouse Trend as an inheritance relationship. Global Warehouse Trend is shown as a superclass of Global Warehouse Inventory (Generalization/Specialization dimension). Additionally, Weather Data Analysis is identified as an entity for global shipment quantity (Classification/Instantiation dimension). 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 the Product (Aggregation/Decomposition dimension).

The relevance values of entities are able to be evaluated again based on the initial virtual big data model as Fig. 10 shows. Each entity follows the relevance calculation rules.

Evaluate Quality Attributes and Select Entities

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 solution, all the relevant entities - each entity is shown in a row, as in Table 1. 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 STR and SER relevance. In Fig. 8, it has STR:+3 and SER:+3 relevant to the solution with ☁!! *Reliable Prediction on Shipment Quantity.*

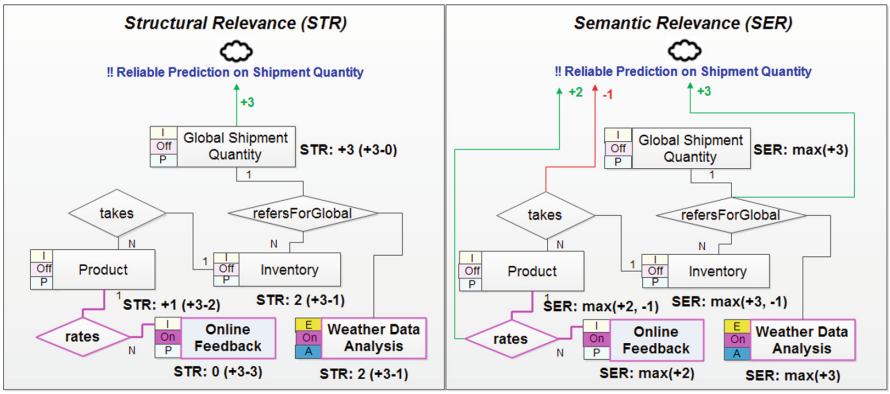

Fig. 10. Virtual big data model, consisting of entities and relationships.

The priority of the entity is related to the priority of parent goals, that is, the business goal, business process goal, and Performance goal. The priority scheme again can be determined, as needed. Here, High, Medium, and Low are used, where their values are 3, 2, and 1 respectively. Zara's innovation team can have a diverse combination of options to be included in the virtual big data model.

Table 1. Selection of entities for shipment prediction their quality characteristics.

Candidate entities	Quality aspects								
	Comprehensiveness						Relevance		Priority
	D1		D2		D3				
	I	E	Off	On	P	A	STR	SER	
Shipment quantity	√		√		√		+3	+3	H (3)
Sales order	√		√		√		+1	−1	L (1)
Online feedback	√			√	√		0	+2	L (1)
Weather data analysis		√		√		√	+2	+3	M (2)

6 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 for big

data but also for the basis of an empirical study. This study shows that IRIS supports the modeling of big data for carrying out business analytics for Zara's shipment decision making. Additionally, a prototype tool is implemented [24].

Additionally, IRIS helps the process of modeling big data conceptually to become mostly traceable, if not all, while helping explore and select among alternatives in problems, solutions, business analytics, and big data models. This would help justify, and boost the level of confidence in, the quality of the resulting big data conceptual model, eventually the quality of business analytics. 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, afterward, monitoring the various real phenomena that are related to either problems or solutions – this seems difficult, if not infeasible in reality.

IRIS's approach helps hypothesize and validate problems with the business process and solutions 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.

Despite the above benefits of IRIS's approach, it has some limitations. First, the empirical study was based on publicly available documents, including articles, white papers, and information on websites, 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. Secondly, 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, we need to evaluate with regard to the utility of the proposed approach in really complex organization contexts. Thirdly, we investigate only the quality attributes which conceptual data model can deal with, so later we need to further research the relationships between them and extend other qualities. Finally, how to make the automatic measurement of the quality criteria, i.e., relevance, comprehensiveness, and prioritization, and to provide more guides for three organizational dimensions in the context of big data remains a challenge. We need further explorations to resolve this challenge.

7 Conclusion

In this paper, we have proposed a goal-oriented approach for a conceptual model of big data to support business analytics. This goal-oriented approach of IRIS is intended to rationally and systematically help model big data at a conceptual level by exploring alternatives and selecting the appropriate ones to validate 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 variety of types and sources of data, which are related and organized

along with 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, big data model; 2) three dimensions of big data model quality; 3) utilization of three organizational dimensions of a data model. Through an empirical study, 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.

There are several lines of future research. One concern offering guidelines and rules for linking a variety of different types and sources of data in different dimensions, towards developing a richer, virtual big data model. We also plan to define rules to automatically measure the three quality attributes, to provide more guidance for three organizational dimensions, to extend the capabilities of IRIS Assistant, concerning the exploration of, and selection among, KPIs and translation into big data queries, and to incorporate the axioms for the comprehensive dimension towards providing some automatic reasoning capability. More studies are needed, be they empirical or case, in a variety of application domains, in order to further determine both the strengths and weaknesses of IRIS.

References

1. Cai, L., Zhu, Y.: The challenges of data quality and data quality assessment in the big data era. Data Sci. J. **14**, 2 (2015). https://doi.org/10.5334/dsj-2015-002
2. Taleb, I., Serhani, M.A., Dssouli, R.: Big data quality: a survey. In: IEEE International Congress on Big Data, pp. 166–173 (2018)
3. Grover, V., Chiang, R.H.L., Liang, T.P., Zhang, D.: Create strategic business value from big data analytics: a research framework. J. Manag. Inf. Syst. **35**, 388–423 (2018)
4. Embley, D.W., Liddle, S.W.: Big data—conceptual modeling to the rescue. In: Ng, W., Storey, V.C., Trujillo, J.C. (eds.) ER 2013. LNCS, vol. 8217, pp. 1–8. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41924-9_1
5. Storey, V.C., Song, I.Y.: Big data technologies and management: what conceptual modeling can do. Data Knowl. Eng. **108**, 50–67 (2017)
6. Mylopoulos, J., Chung, L., Nixon, B.: Representing and using nonfunctional requirements: a process-oriented approach. IEEE Trans. Softw. Eng. **18**(6), 483–497 (1992)
7. Teorey, T.J., Yang, D., Fry, J.P.: A logical design methodology for relational databases using the extended entity-relationship model. ACM Comput. Surv. **18**(2), 197–222 (1986)
8. Chen, P.: The entity-relationship model – toward a unified view of data. ACM Trans. Database Syst. **1**, 9–36 (1976)
9. Chebotko, A., Kashlev, A., Lu, S.: A big data modeling methodology for apache cassandra. In: Proceedings of IEEE International Congress on Big Data, pp. 238–245 (2015)
10. Baazizi, M.A., Lahmar, H.B., Colazzo, D., Ghelli, G., Sartiani, C.: Schema inference for massive JSON datasets. In: Proceedings of Extending Database Technology (2017)
11. Jayapandian, C., Chen, C.-H., Dabir, A., Lhatoo, S., Zhang, G.-Q., Sahoo, S.S.: Domain ontology as conceptual model for big data management: application in biomedical informatics. In: Yu, E., Dobbie, G., Jarke, M., Purao, S. (eds.) ER 2014. LNCS, vol. 8824, pp. 144–157. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-12206-9_12
12. Caballero, I., Serrano, M., Piattini, M.: A data quality in use model for big data. In: Indulska, M., Purao, S. (eds.) ER 2014. LNCS, vol. 8823, pp. 65–74. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-12256-4_7

13. Cristalli, E., Serra, F., Marotta, A.: Data quality evaluation in document oriented data stores. In: Woo, C., Lu, J., Li, Z., Ling, T.W., Li, G., Lee, M.L. (eds.) ER 2018. LNCS, vol. 11158, pp. 309–318. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01391-2_35
14. Taleb, I., Dssouli, R., Serhani, M.A.: Big data pre-processing: a quality framework. In: Proceedings of the IEEE International Congress on Big Data, pp. 191–198 (2015)
15. Sheth, A.P., Larson, J.A.: Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Comput. Surv. **22**(3), 183–236 (1990)
16. Smith, J.M., Smith, D.C.P.: Database abstractions: aggregation and generalization. ACM Trans. Database Syst. (TODS) **2**(2), 105–133 (1977)
17. Nalchigar, S., Yu, E.: Business-driven data analytics: a conceptual modeling framework. Data Knowl. Eng. **117**, 1–14 (2018)
18. Park, G., Chung, L., Khan, L., Park, S.: A modeling framework for business process reengineering using big data analytics and a goal-orientation. In: Proceedings of the 11th International Conference on Research Challenges in Information Science (RCIS), pp. 21–32 (2017)
19. Park, G., Sugumaran, V., Park, S.: A reference model for big data analytics. In: Proceedings of the 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication, pp. 382–391 (2018)
20. Wang, R.Y., Strong, D.M.: Beyond accuracy: what data quality means to data consumers. Manag. Inf. Syst. (MIS) **12**(4), 5–33 (1996)
21. Caro, F., et al.: Zara uses operations research to reengineer its global distribution process. INFORMS J. Appl. Anal. **40**(1), 71–84 (2010)
22. https://sloanreview.mit.edu/article/variety-not-volume-is-driving-big-data-initiatives/
23. https://liliendahl.com/2019/06/13/data-modelling-and-data-quality/
24. https://sites.google.com/site/irisforbigdata/
25. Gosain, A.: Literature review of data model quality metrics of data warehouse. Procedia Comput. Sci. **48**, 236–243 (2015)