Mphasis
Unleash the Next

Big Data Variety And Veracity

Data awareness and quality perspective

Dr. Aravindhan Arunagiri
Manager, Mphasis NEXTlabs

NEXT labs

Summary

Present businesses extend beyond the business paradigms of traditional market. To be competitive, they need to be proactive to market and engineer them to create new business opportunities. Companies resort to Big Data initiative in order to have extended insights of their stakeholders, which includes customers, vendors and investors. Big Data, characterized by its high volume, velocity, variety and veracity of data, leads to complexity in using various analytics methods to gain business insights. In this paper the importance of data awareness regime (discovery and management of relevant data), with respect to various aspects of Big Data, is discussed. A set of Big Data variety implications during Data awareness regime are defined and discussed case by case with suitable examples. The importance of data quality (in terms of fitness to purpose) required to achieve valuable business insights from Big Data is highlighted.

Introduction

Big data analytics is the process of knowledge discovery from the data that is enormous in volume, massive in terms of velocity and generated from variety of sources. In Big Data, variety refers to the data residing in multiple data sources like enterprise transactional data, social network applications data, web logs, user blogs, third party market report, data marts, point-of-sale data, IOT, etc. Within the respective sources, the relevance and veracity of data is maintained. When collective insights are required to be made from even a small set of data sources (say 2 database tables) the present analytics paradigm proves insufficient. With present databases and management systems, the processes of collating the data (even logically) from different tables require minimal criteria, such as equivalence of attributes (data type, domain), attributes names, in both tables. In most of the cases, the representation of same concepts (data) may be different across multiple sources. Hence, it requires human intervention to identify the relevant data and suitably collate the information for analytics. Further storing such collated information requires repeating the whole exercise of data warehousing. The discovery of relevant data from Big Data is complex due to a set of implications resulting from the variety aspect of the Big Data. The following are the five sets of implications called as Big Data variety implications:

[1] **Inconsistent information of objects:** Ambiguous representation of same concept with different identity

[2] **Redundant information:** Occurrence of same entity or concept (may be a person, thing, abstract concept) in multiple data sources

[3] **Incomplete information of objects:** Splitting of information about same entity across multiple sources

[4] **Data awareness regime:** discovering and management of relevant entities (concepts)

[5] **Assurance of data quality** in terms of information quality (fit-to-purpose)

The implication of data quality and awareness, which are demanding factors at enterprise level, are even more challenging when it comes to Big Data context.

Data awareness

In Big Data viewpoint, the data awareness involves identification of relevant entities (objects) that may be residing in single or multiple data sources. The identification of entities also include coherence (relationships) among those identified entities. In Big Data the main objective of adapting variety of data is to get complete information of related objects, (like {person, asset}, {asset, process}, {process, legal}) which may not be available with single data source. The analysis of collective information unleash new business insights (which are initially unanticipated). The following section discusses the variety aspect of data with respect to business analytics and intelligence.

Case [1] Inconsistent information of objects

Most of the times, the same object (a person, thing, and abstract concept) is described by different attributes, in different data sources. So, when the data sources are combined, the information gained, due to attributes combination, may be more or less or unchanged.

Table – 1a

Cust_ID	Name	Asset	Price
2B300	John Doe	John villa	200K
4R100	Bill Gram	Mist casa	150K
3Z122	Jim Rev	Mella	100K

Table – 1b

Name	Asset	Complaint
Joeami07	John villa	Plumbing
Grami	Mist casa	Parking
Dolly	Mella	Electrical

In the above figure, both the tables have information of same person, but have different identity. For example, assume John Doe in *Table 1a* is same person with username Joeami07 in *Table 1b*. Here, though the attribute labels are "Name" in both tables, the ambiguity of object (person) is based on attribute values. Now the same person "John Doe" cannot be identified until the information gained from two tables is manually matched. Collating information of these two tables requires both manual intervention and domain knowledge.

Case [2] Redundant information

Table – 2a

Cust_ID	Name	Asset	Price
2B300	John Doe	John villa	200K
4R100	Bill Gram	Mist casa	150K
3Z122	Jim Rev	Mella	100K

Table – 2b

U_name	Property	Value
Joeami07	John villa	200K
Grami	Mist casa	150K
Dolly	Mella	100K

In some cases, different sources have same data but are represented by different attribute or class names. Refer above tables. Here the ambiguity of objects in different tables is based on attribute labels, provided the values are matching. For example attribute "Asset" in *Table 2a* represents attribute "Property" in *Table 2b* and both have information of name of the property. In this case, matching the objects by attribute list is not possible automatically, unless their values are matching. This leads to retrieving and analyzing the same data redundantly.

Case [3] Incomplete information of objects

This is the most common and important aspect related to "variety" component of Big Data. Every data source has partial information about a particular object and complete information can be achieved by merging the sources accordingly. The Big Data initiatives important driver is to get maximum information of objects by collating the data sources. *Table 1a* has information of customer's name, their respective Customer ID, and Asset holding. Other information, regarding the person's user name and the complaint booking of the property, is shown in *Table 1b*. The information about same object is split across different tables. Collating tables to gain information requires sorting out both the ambiguities by attribute labels and attribute values.

Case [4] Data awareness regime

In handling variety of data in Big Data, the discovery of relevant data (object/entity) for analysis is a critical process. In Case [3], the collation of different data sources gives complete information for analyzing a particular object. Apart from gaining the complete information by merging data sources (by merging logically different data sources), there are also possibilities to discover new (hidden) entities and higher level abstract insights.

Table – 4a

Cust_ID	Name	Asset	Price
2B300	John Doe	John villa	200K
4R100	Bill Gram	Mist casa	150K
3Z122	Jim Rev	Nancy sis.	100K

Table – 4b

U_name	Property	Value	LoanID	Value
Joeami07	Liza villa	100K	230001	John Doe
Grami	Mist casa	200K	320044	Bill Gram
Dolly	Nancy sis.	150K	511011	Jim Rev

For example, the *Tables 4a* and *4b* provide name and asset related information of the same person. *Table 4b* provides the loan information related to the asset of people in *Table 4a*. The person "John Doe" has asset "John Villa" according to *Table 4a*. But in *Table 4b*, "John Doe" has loan against asset "Liza villa". It is interesting to note that John Doe has two assets, with loan taken against one of them. This discovery of new object (here "Liza villa") and its relationship with existing object needs manual intervention of someone with domain expertise to assert them.

Case [5] Data quality

In all the above cases [1] through [4], the data in single source needs to be collated with data in another source, to gain new or refined information of the objects. The above Big Data variety implications are even more complex if there are mixed instances. For example, in *Table 4b* if the Attribute label "Name" is not available then even manually matching the person's name becomes complex. The complexity of collating the data sources increases with existence of one or more Big Data variety implication cases, as mentioned herein.

While the discovery of relevant data requires collating of data sources, assurance of quality of the discovered data entities relies in its fitness to purpose (analysis). The objects discovered in terms of its features (attributes) need to align with Analytics methods and framework used for gaining business insights.

Conclusion

Enterprises are using Big Data analytics to gain enhanced and new business insights. The existing analytics and data platforms are insufficient in terms of handing the variety aspect of Big Data. This variety aspect of Big Data gives multiple implications when we attempt to understand, collate and manage the data (information) gleaned from different sources. The Big Data variety implications also necessitates the maintaining of data quality aspects (in terms of its fitness to various analytics paradigms) which are important to gain valuable business insights.

Dr. Aravindhan Arunagiri

Manager, Mphasis NEXTlabs

Aravindhan is an expert of Logics and Algorithms within Mphasis NEXTlabs. He has hands on experience in various fields of BPM, Big Data, Cognitive Computing and AI with additional pursuits in Robotics. He holds a PhD from Department of Management Studies, Indian Institute of Science, Bangalore.

About Mphasis

Mphasis is a global technology services and solutions company specializing in the areas of Digital, Governance, Risk & Compliance. Our solution focus and superior human capital propels our partnership with large enterprise customers in their digital transformation journeys. We partner with global financial institutions in the execution of their risk and compliance strategies. We focus on next generation technologies for differentiated solutions delivering optimized operations for clients.