International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 06 | June-2017 www.irjet.net p-ISSN: 2395-0072

The Comparison of Big Data Strategies in Corporate Environment

Mr. Sagar Jhaveri[1], Mrs. PratibhaDeshmukh[2]

[1] Student, Dept. of Information Technology, Bharati Vidyapeeth Institute of Management & Information Technology Navi Mumbai, Maharashtra,India
[2]Professor, Dept. of Information Technology, Bharati Vidyapeeth Institute of Management & Information TechnologyNavi Mumbai, Maharashtra,India

---

Abstract - *The growth in an organization's business leads to tremendous increase in the volume of data generated by it. The data may be in a structured, unstructured and/or in semi structured format. Big data is the way how these business organizations handle this data. The aim of this paper is to identify a suitable big data strategy for organizations to handle large volume of data in a fast and efficient manner.To do that existing research in the field of data warehousing and business intelligence has been reviewed and our findings are synthesized in a contingency matrix that may support practitioners in choosing a suitable big data approach. Though any strategies can be beneficial for certain corporate circumstances, the Hybrid approach is a combination of traditional relational database structures and Map Reduce technique which is the strategy that is most valuable for companies pursuing business analytics.*

***KeyWords*: Bigdata, GPS, RDBMS, DFS, MapReduce, Hybrid approach**

1. INTRODUCTION

Big Data – data that contains variety, arriving in increasing volumes with higher Data complexity, growth is being driven by development of millions of intelligent sensors and devices that are transmitting data (Internet of Things) and by other sources of structured data or unstructured data. These platform, tools and software used for this purpose are collectively called "Big Data technologies". Big Data can also be define by the four "V"s: Volume, Velocity, Variety, and Value. It becomes a reasonable test to determine whether you should add Big Data to your architecture. However, to obtain the desired insights, data needs to be sourced, stored, and analysed. During the past years, accessing and processing the collected, voluminous, and heterogeneous of data has become increasingly time consuming and complex [2]. For example, applications enable a great customer experience which is often powered by smart device and also enable the ability to respond in the moment to customer actions. Data scientists and Business analysts are developing a host of new techniques and models to uncover the value provided by this data. Its provided solutions are helping to discover personalized value chains, predict product, consumer trends and reveal product reliability. With a total of 1.8 zeta byte in 2011, the amount of generated data has not yet reached its end: as expected by experts, a global provider of IT market intelligence. The total amount of data collected until the end of 2012 was estimated to be 1.48 times the data collected in previous years, with more than 80% of this data being unstructured. Business increasingly use these data masses provided by billions of networked sensors in smart devices, cashier systems, automobiles, or weather stations to learn more about their customers, suppliers, and operations [3]. This development raises the question of how companies may manage to cope up with the amount of data generated, referred to as big data. The aim of this paper is to provide a set of factors that affects different Big Data strategies which organizations may use for decision making or various implementations. To do that, we have reviewed technologies to find different Big Data strategies as well as factors that are combined both into a matrix that may support the users in choosing a suitable Big Data strategy for their needs.

2. LITERATURE SURVEY

The characteristics of Big Data were first described in 2001, when Laney[4] identified three key attributes of large data as: high variety, volume, and velocity. To date, these characteristics have become the defining features of Big Data. However, various authors and business analyst enlarged these defining characteristics with further aspects such as dedicated storage, management, and analysis techniques. Further changes to the definition include the addition of a fourth V, veracity, emphasizing the aspect as quality of data. Taking these different extensions of the original definition into account, Big Data is defined as a phenomenon characterized by an increase in volume and variety of data that requires advanced techniques and technologies to store, distribute, manage, and analyse data. The economic potential of Big Data is as diverse as the data itself and the key driver for organizations to adopt Big Data analytics. There are four Big Data strategies organizations can use such as: external structured data like Global Positioning System (GPS) or credit history data, internal structured data such as inventory data or Customer Relationship Management (CRM), external unstructured data such as Facebook or Twitter posts, as well as internal

unstructured data such as text documents. All the above categories have particular characteristics that certain Big Data strategies address better than others. Fields of Big Data cover a wide range of industries and businesses. Suggestions range from health care (reduction of costs resulting from over- and under-treatment) with a $300 billion annual potential, to the public sector and government sector with a $250 billion annual potential, over online shopping and marketing (better understanding of consumers with respect to product and price preferences) with a potential of 60% increase in operating margins. These field's potentials are unlocked by the application of different Big Data techniques such as crowd sourcing, data fusion, data integration, analysis of network, modeling of data, together with simulation of data. Thus, the probability and economic potentials of these stratregy is enormous and executives should assess whether and how [8], they could make use of these potentials. By perfectly using the information about the Big Data strategies mentioned in this paper, companies will be able to improve their decision-making and better realize of data which Big Data strategy will be use full for their organization.

3. METHDOLOGY

Big Data is still a new and emerging field of research. Consequently, our understanding of Big Data's basic constituents is still fragmented. By identifying different Big Data strategies and their facilitating conditions, We hope to contribute to the field's knowledge. We rely on the literature review methodology because of its ability to expose theoretically and uncovered research points. We followed established guidelines for literature reviews however, also included industry reports and best practices. This approach has been recommended for newly emerging research themes. This review follows a three-staged literature analysis i.e. data analytics, data warehousing, and also business intelligence. First, we have searched and analysed existing data of this to identify a set of corporate Big Data strategies. We particularly looked at the different categories of Big Data described previously and analysed the capabilities of existing data analysis approaches with respect to how well they can handle the various Big Data categories. Secondly, we have systematically analysed the literature to identify the factors that affects Big Data strategies choice. For that, we have studied different aspects affecting choice in traditional data warehousing environments, evaluated them regarding their relevance for Big Data analytics, and summarized them in matrix. On the basis of this, three groups of factors have been formed, each influencing the strategy decision differently. Finally, we linked the context factors to different types of strategies. Resulting matrix will not only provide assistant for practitioners in choosing a suitable Big Data strategy, but it will also help company with profiles that are associated with different Big Data strategies.

4. CONCEPT

4.1 RDBMS

Big data management firm has started growing over few decades, RDBMS(Relational Data Base Management Systems) was the first technology. Still, most of the backend systems are RDBMS for telecommunication business, online digital data, financial systems, medical. As the amount of data collected, and analysed are increasing tremendously, it is difficult to manage and control . Many organizations have faced problems with limitations of traditional RDBMS architectures. RDBMS[5] tools is little bit complex, hence feels the necessity of alternate tools that can handle such a huge data which is usually referred to as „big data". Advantage of parallel RDBMS is that they can handle and analyze large volumes of data very fast and stable. Considering, the RDBMS capabilities with regards to *variety* and *velocity* of data has several problems. Different business intelligence tools can then be used to analyse and access the data. Since volume of the data to be processed steadily increased day by day, most of the contemporary companies revert to the parallelized RDBMS to handle the large amount of data. Hence, data are to be stored on different machines, tables are partitioned over the nodes in a cluster, and an application layer allows for accessing the different data portions on the different nodes. The main goal of RDBMS architecture is to provide linear speed-up as well as scale-up. This means that twice as much as hardware allows for execution of twice as large tasks in the same elapsed time. For example, HP banks on the usage of massive, parallel SQL databases. They are integrating a large amount of new in-memory analytic functions and new technologies to easily expand or downsize deployments.

4.2 Map Reduce

The second most popularly used strategies for big data analysis is the distributed file system(DFS) and Map Reduce engine. The RDBMS approach works well with the small volume of data, when it comes to handle huge volume of data it is really a difficult task to process that data through traditional approach. The issue of handling huge amount of data has been resolved by using an algorithm Map Reduce. Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment[6].Hadoop framework allows to process and store huge volume of data in a distributed environment across several computers using simple programming models and still they could perform statistical analysis on huge amount of data. Hadoop runs its application using the algorithm called Map Reduce. Map Reduce is a model for processing and generating logical and huge datasets [8]. In general Map Reduce algorithm divides the work into smaller tuples and assigns these tuples to number of computers connected over the network and collects the result to form the final result of dataset. Map Reduce model consists of two

tasks, namely Map and Reduce. The second task is the Reduce task in which it takes the output from a Map task as input and break data into tuples- a smaller set of data and this task will be performed after the Map task. The input data to be analysed is stored in the distributed file system and then processed using the Map Reduce engine. The results(output) are then again stored in the file system and directly streamed to a business intelligence application.

4.3 Hybrid Approach

Both Big Data strategies, traditional RDBMS as well as Map Reduce and DFS, have potentials as well as limitations. Thus, the introduction of a hybrid solution combining the benefits of both approaches seems reasonable and is called for by different authors as well. In general, there are three possible solutions to integrate RDBMS systems and Map Reduce . In a Map Reduce-dominant approach, the Map Reduce technology is extended with relational components leading to efficient processing of large volumes of structured as well as unstructured data, while profiting from typical RDBMS-strengths such as query optimization. However, performance gains in query processing are limited, since the relational components can typically only be integrated on several nodes. RDBMS dominant hybrids have integrated Map Reduce capabilities in their engines to particularly improve processing abilities for unstructured data. Although it has been shown that this approach tends to produce faults under certain circumstances, it is the one most often used by commercial vendors. The last and least frequently adopted approach is a loose coupling of Map Reduce systems and RDBMS. The loose-coupled approach might appear most valuable on first sight because it theoretically allows for connecting any Map Reduce system with an RDBMS. Lacking interface standardization however causes problems such as complicated data transfer or optimization between both systems. Due to this problem, the benefit of "simply" connecting some systems proves fallacious to some degree leading to the just mentioned low adoption rates. In summary, regardless of which approach a company chooses, a hybrid strategy allows to efficiently handle Big Data without struggling with some of the problems depicted for the pure RDBMS or Map Reduce strategy. Still, performance of hybrid systems does not – at least at the moment – exceed that of uncombined strategies. Rather, such systems process and analyse data with large volume, variety and velocity within acceptable performance and fault boundaries. From a financial point of view, the hybrid strategy approach is more expensive than an uncombined strategy. While the potentially high licensing costs can predominantly be ascribed to the RDBMS involved, often more processing power and storage is needed leading to higher hardware expenses. However, researchers as well as software vendors are currently putting efforts in the development of hybrid solutions that become increasingly per formant. Prominent examples are Hadoop DB (Map Reduce-dominant), the Oracle in-database Hadoop, Microsoft, who announced their Poly Base technology in November 2012, or Green plum , who all build up on a traditional RDBMS system (RDBMS-dominant).

4.4 Analytics

Based on the above discussion, we observed that different organizational environments pursue different requirements on a Big Data strategy. To better support practitioners in Big Data strategy choice, We compared the four identified Big Data strategies regarding how well they addressed each of the contingency factors. For instance and as discussed above, when the relevance of Big Data analytics is high in a company, the Map Reduce strategy seems most fruitful. However, also a hybrid solution might be valuable in case it follows a Map Reduce-dominant approach. If in turn an RDBMS-dominant implementation is chosen, the hybrid strategy is only slightly better than a "pure" RDBMS approach. Several patterns are observable for the different factor categories. By contrast, We find both the Map Reduce as well as hybrid strategy being positively associated with most of them. The RDBMS strategy again has turned out to be a suitable approach in nearly all situations as long as expectations towards Big Data processing performance are not too high. In cases in which performance plays a more important role, hybrid approaches comprise a workable trade-off between costs, processing, and analysis performance.

5. CONCLUSION

In this paper, we aim at providing guidance for companies on how to approach the phenomenon of Big Data. Based on a review of practitioner as well as scientific, we have found four Big Data strategies and discussed it regarding factors influencing strategy choice. The eight respective contingency factors can be grouped into three dimensions, namely strategy, resources, and operating environment. Although other authors have already discussed context factors that might influence Big Data strategy choice, a structured analysis of such contingency factors has not been performed so far. We therefore contribute to the still limited research on Big Data by providing a basis for future discussions on the adequacy and success of various Big Data strategies for differing corporate environments. As illustrated by the analysis of the opportunities and challenges of Big Data in this paper, organizational decision makers need to start thinking about whether and how to facilitate Big Data analytics. They therefore benefit from our research by gaining a better understanding of the different factors they should consider before deciding on Big Data solution investments. This scarcity clearly calls for more research, as most of the current research is focused on technological aspects, while only a very limited number also deals with organizational aspects.

REFERENCES

[1]Big Data, *A New World of Opportunities*, NESSI White Paper, December 2012.

[2]http://www.BDStrategies.com/content /rdbms/ed/rdbms.html

[3]Nair, R., and Narayanan, A., *Getting Results from Big Data - a Capabilities-Driven Approach to the Strategic Use of Unstructured Information*, Booz & Company, 2012.

[4] Laney, D., *3d Data Management: Controlling Data Volume, Velocity, and Variety*, Stanford, 2001.

[5] Sethy et al., *International Journal of Advanced Research in Computer Science and Software Engineering* 5(7), July- 2015

[6]Harshawardhan S. Bhosale "A review paper on big data on hadoop" International Journal of Scientific and Research Publications,,October 2014(Vol 4)

[7] Konstantin Shvachko,"*Hadoop Distributed file system*"

[8] Jeffrey Dean and Sanjay Ghemawat "Map Reduce: Simplified Data Processing on Large Clusters" OSDI 2004

[9]Kyong-Ha Lee Hyunsik Choi "Parallel Data Processing with Map Reduce: A Survey" SIGMOD Record, December 2011 (Vol. 40, No. 4)

[10] Lavalle, S., Lesser, E., Shockley, R., Hopkins, M.S., and Kruschwitz, N., "*Big Data, Analytics and the Path from Insights to Value*", Sloan Management Review, 52(2), 2011, pp.
20-31.

[11]http://documentslide.com/documents/ieee-2014-47th-hawaii-international-conference-on-system-sciences-hicss-588e1e4d9f577.html.

[12] http://www.sersc.org/journals/IJDTA/vol9_no1/5.pdf