

# Data Driven DSS in Big Data Environments for power supply organization

<sup>1</sup>Pooja Nahar, Dr. Sanjay Kumar<sup>2</sup>

<sup>1</sup>Ph.D. Scholar Computer Science Jaipur National University Jaipur, Rajasthan, India

<sup>2</sup>Associate Professor Computer Science Jaipur National University Jaipur, Rajasthan, India

<sup>1</sup>Orcid id: 0000-0003-1052-6842

## Abstract

DSS is the area of IS discipline which is focused on supporting and improving management decision making process. In the fast changing world, organizations have to react not only on present but also on future challenges and take decisions in a fast manner. They have to make decisions on high quality data basis. The data driven decision making is being recognized broadly. Managing and maintaining big data will provide more flexible collaborative query results. Big data analytics can greatly improve decisions. Not only the amount of data matters, but what organizations will do with that data also. Big data is used to analyze that lead to better decisions and strategy. For the case company, we have defined a DSS in our previous paper. In this paper we will give the solution of handling, querying, and updating the transactions with large data. This paper proposed a sustainable big data based data driven Decision Support System.

**Keywords**— Decision Support System, Big Data, Hadoop, Map Reduce, HDFS

## INTRODUCTION

As increasing the number of customers, it increases the amount of data that any organization needs to manage. Storing that much of data would've been a problem for any organization, but new technologies like Hadoop have erased this burden. DSS is the area of IS discipline which is focused on supporting and improving management decision making process [1, 2]. It can also be explained as a class of IS, which is a computerized manner and supports the activities of decision making. It is especially developed for aid in better decision making as it supports the solution of a non-structured management problems while the process of decision making is invisible [3]. In early 1990', mainframe based DSS shifted to client-server DSS, where server is linked to the user's computer by a network with the TCP/IP protocol. Now web is the platform of choice for building DSS, because of its flexibility of expansion and easy availability of Internet. WWW is used to communicate decision information and computation. It is platform independent and provides universal access to DSS applications. Public sectors also initiating, encouraging and supporting informatics projects so that they could accelerate managerial and technological

development. DSS is a concept in which more complex problems could be solved and it helps in making decisions in such a complex environment. This supports customer's participation in decision making. When the matter is of velocity and volume of data, it is necessary to use the concept of big data.

It can take data from any source and analyze it to find answers that enable 1) cost reductions, 2) time reductions, 3) new product development and optimized offerings, and 4) smart decision making.

## PROBLEM

The case company will also face the challenges of traditional transaction processing systems while doing their routine tasks. When data will be as large like in terabytes or more then even running a single query will be a big problem for them. Relational database management systems (RDBMS) often have difficulty handling big data. As electricity is the basic need of any citizen, so as the population will increase, the customers will increase for the case company and data as well as. So maintaining and handling this large amount of data will be a big problem.

There are many types of decisions, which need to take in routine basis or in a regular interval. Traditionally decisions were made by some selected peoples like board members or panel members or top level management. They had a face to face meeting and discuss for their decisions with their knowledge [4]. There are some limitations in this method. The decisions may make without having proper knowledge and the cost of taking decisions may be high. Here is the complex decision algorithm to decide whether to disconnect connection of any customer for not paying his monthly / quarterly / yearly bill. In this algorithm many a times user has to check many conditions on a vast amount of data then the system performance and query performance will be effective. In this decision, many customers are there who did not pay their electricity bill. In order to not paying our system will recognize those members, the local decision for this is to disconnect their connections. But many of them may be our prime customers who have been paying their bills regularly, or some of them may not be able to pay their bill because of some tragedy or etc. So to decide whether to give

disconnection orders to all of them or not, is a big problem and taking decisions against that with rules and regulations is complex. Viewing their previous records and answering many questions, need to run many queries on a previous 2 / 3 / 5 years of data. Here we need to manage big data while taking decisions.

## BIG DATA

There is no hard and fast definition of about exactly what size a database needs to be for the consideration of “big”. The main deal is to process this big data. Processing multiple data from multiple sources, (which form a large pool of data), has important consideration for big data. It increases the efficiencies in whatever the organization collected data for. Apache Hadoop is the most common and influential tool to analyze big data. It is completely a open source. Organizations like our case company are used to analyzing internal data and external data. Right data at the right time with right format is used by the managers to support decision making. Big data is a term for data sets that are so large or complex i.e. traditional data processing application software could not deal with large data. Capturing, storing, analyzing, searching, transferring, visualizing, querying etc. are the challenges which traditional data processing application software has to face. Greater access to data and the technologies for managing and analyzing data are changing the world. It provides greater personal data transparency and portability for better decision-making. It makes data smart and actionable [5]. The problem of working with data that exceeds the storage of a single computer could be dealt with big data concept. There are 3 Vs of big data concept, which characterize it- a) Volume – In this; datasets can be orders of magnitude larger than traditional datasets, which demands more thought at each stage of the processing and storage. b) Velocity - It is the speed that information moves through the system from multiple sources and processed and then update the system. c) Variety – Big data has wide range of sources, so the formats and file types are vary significantly. Big data systems can accept and store data in its raw state too. Then, any transformation and changes to this data would happen at the time of processing. For high storage and computational needs of big data, we need computer cluster technique [6, 7].

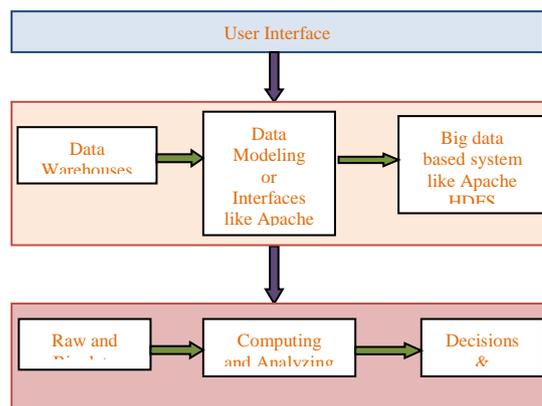
In a big data processing it includes- (i) Ingesting data into the system- It is the process of taking raw data and adding it to the system. Some dedicated ingestion tools are available like Apache Sqoop, which can take existing data from relational databases and add it to a big data system. Analysis, sorting and labeling takes place during ingestion process. This is generally called ETL (Extraction, Transformation and Loading). It captured data as raw as possible for maintaining flexibility.

(ii) Persisting the data in storage- After the ingestion process,

it typically hand the data off to the components that manage storage, so that it can be reliably persisted to disk. This means leveraging a distributed file system for raw data storage. Tools like **Apache Hadoop Distributed File Systems** allow large quantities of data to be written across multiple nodes in the cluster. (iii) Computing and Analyzing the data- Once the data is stored then the system can begin its processing. Data is often processed repeatedly. Batch processing is very common for computing over large data sets, as well as we have real time processing or stream processing too. Splitting, Mapping, Shuffling, Reducing, and Assembling are the steps of Batch Processing. This strategy is used by **Apache Hadoop's MapReduce**. Many other tools are available for computing and analyzing data with big data systems. These can easily be plug into any framework and provide interfaces to interact, like **Apache Hive** provides a data warehouse interface for **Hadoop**, **Apache Pig** provides a high level querying interface, while **Apache Drill**, **Apache Impala**, **Apache Spark SQL** and **Presto** provides SQL-like interactions with data. (iv) Visualizing the results- Visualizing data is very important way to make sense of a large number of data sets.

The data changes frequently of any organization, interactive exploration and visualization of the data in a format conducive to sharing, presenting, or collaborating make sense for understanding.

## ARCHITECTURE OF DATA DRIVEN DECISION SUPPORT SYSTEM FOR BIG DATA



**Figure 1:** (Framework / Architecture)

This introduces the proposed architecture of Data Driven DSS with big data. This integrates big data resources with a web based decision support system. In this system all users would interact with the system through Internet.

From the main layer, users would use the GUI based interface so that they could understand the system easily. Then the data from the data warehouses would be modeled to the storage system of big data i.e. Hadoop Distributed File System. For that it needs some interfaces according to the requirement.

After that raw big data is available, any type of computing and analysis could be done using this. It may help in decision making and provide different results in visualized form. So that it could be understandable by the common person. It would improve the query efficiencies too. It depends on the processing type, whether batch processing / real time processing / any other.

## PROBLEMS AND CHALLENGES IN USING BIG DATA

The analysis of Big Data involves multiple distinct phases, each introduces challenges. People poorly understood the complexities of using clusters. Big Data has to be managed, where data may be noisy, heterogeneous etc. Storing, filtering and compressing these data in such a way that useful information should not be discarded. It needs to handle uncertainty and error. If we want to make this project successful then we have to consider all of these challenges before implementing this concept.

- i) Timeliness- As the name it has 'big', the data sets are very large. So larger data sets need more time to analyze. When the term Velocity comes, it does not meet that speed for which it is meant for [7, 8]. The system which effectively deals with the size, could process data sets faster [9].
- ii) Privacy and security of data- This 'big data' term is basically used where multiple types of data are to be stored and fetched many times with multiple data sources. Means many data sources could be connected, shuffling, mapping or reducing at the end for results. So there is rare chance to have security/privacy of data. Personal data could also be fetched and used for wrong ways [7, 8]. There is a great public fear regarding the misuse of personal data which may link multiple sources. Managing privacy is a technical as well as a sociological problem. These new technologies generally ask to share his/her list of contacts and location; this also generates privacy concern. Even, we do not know how to share private data while limiting disclosure and how to give them fine-grained information to them [9].
- iii) Size- As the name enters in our mind, we first notice 'big'. It has really very big data which it handles. Big means not only large but different type of data also. Previously it was a challenge to handle big data, but now data is still becoming big and bigger. Now because of this getting bigger data, computing resources and CPU speed is not scaling/increasing than its comparison [7, 8]. Managing large and rapidly increasing volumes of data has been a challenging issue [9].
- iv) Data Access and Sharing- Due to large amount of data, governance process does not want to open and share that. Sharing between companies threatens the secrecy of data, so they hesitate to make it available for any company [8].

Even after cleaning and error correction, some incompleteness and errors may likely to remain in the database. It is also a big challenge to manage them while accessing [9].

- v) Technical challenges- When the large data sets concern, there would be a big challenge to handle fault. How the failure/damage would be handled. When it is operating on large clusters, it is very important to manage system failures. To decide data storage devices is also a important task because of data scalability. Big data has a stress on large data sets but maintain quality of data is also a big concern because irrelevant big data has no value for the company. So it should be ensured that which data is important and how to convert information from the raw irrelevant data for better results [8].
- vi) Human Resources and Man Power- Big data needs human collaboration to understand or visualize results. In many complex situations, it needs multiple human experts of different domain to really understand the problem. Big data cannot exclude this very important human integration in the system. So it needs persons with new skills [7, 8]. These skills also need to be developed in other individuals using training programs which should be held by the organization itself [8]. There are many patterns that humans can easily detect but computer algorithms take hard time to find. There is similar value to human input at all stages [9].

## BENEFITS OF USING BIG DATA

'Big data' is very similar to the 'small data' but bigger. Having bigger data consequently requires different tools, techniques etc. to manage this situation. Yes, it is complex and difficult to manage but it has many advantages. Many organizations are accepting big data before their data is not big. It's not only the bigness which impresses them, but many other advantages too [10].

- i. Increases the storage capacity- It drastically increases the storage capacity. When the organization's data exceeded to the volume which any relational data base could not handle, then it needs to shift their database to big data based systems like Hadoop. They can store large data sets as well as variety of data sets. In other words, it can handle different types of data with multiple file formats [7, 8, 10].
- ii. Reduces the cost- Big data technologies like Hadoop provide cost effective solutions; which are utilized for cost reduction. Managing large amount of data with many advantages is less costly instead of using multiple tools for every task [10].
- iii. Time Reduction- New technologies for big data reduces the time for complex and large scale analytical

calculations drastically. This can also run those complex queries and find results faster [10].

- iv. Improving process efficiency- As it provides the best results in an effective reading or understanding manner, this will surely improve the process efficiency. The managers will understand the scenario more appropriately and would take effective decisions [10]. Presentation of the results and its interpretation by non-technical domain experts is crucial to extracting actionable knowledge.

## CONCLUSION

Big data is not only big, but also it is complex to be stored, processed and analyzed. This Big Data is more suitable to store and analyze social media interaction, medical records, government records or genetic sequences. The Big Data concept is emerged from the increasing volumes of internal and external data having 3 V's- Volume, Velocity and Variety. It explores both internal and external information. This paper provides a theoretical analysis and understanding of Big Data along with Decision Support System, thus providing tools and techniques of using Big Data. It has some challenges like; the whole organization must adapt to this new technology and understand the new concept.

## REFERENCES

- [1] David Arnott, Graham Pervan, Peter O'Donnell and Gemma Dodson, "An Analysis of Decision Support Systems Research: Preliminary Results", Decision Support in an Uncertain and Complex World: The IFIP TC8/WG8.3 *International Conference 2004*
- [2] Arnott, D. and Pervan, G. "A critical analysis of Decision Support Systems research", *Journal of Information Technology*, 20, 2, June, 2005, pp67-87
- [3] Sandeep Kautish and Prof (Dr.) M. P. Thapliyal, "Concept of Decision Support Systems in relation with Knowledge Management – Fundamentals, theories, frameworks and practices", *International Journal of Application or Innovation in Engineering & Management (IJAEM)*, ISSN 2319 – 4847 Volume 1, Issue 2, October 2012
- [4] Hong Zhou, Christopher Noble and Julie Cotter, "Big Data Based Intelligent Decision Support System for Sustainable Regional Development"
- [5] Ehrenberg, R.: What's the big deal about Big Data? InformationArbitrage.com blog post, 19 January 2012. <http://informationarbitrage.com/post/16121669634/whats-the-big-deal-about-big-data>
- [6] Qiang Yang, Fellow, IEEE, "Introduction to the IEEE transactions on Big Data", *IEEE transactions on big*

*data*, vol. 1, no. 1, January-March 2015

- [7] Harshawardhan S. Bhosale and Prof. Devendra P. Gadekar, "A Review Paper on Big Data and Hadoop", *International Journal of Scientific and Research Publications, Volume 4, Issue 10*, October 2014 1 ISSN 2250-3153
- [8] Iqbaldeep Kaur, Navneet Kaur, Amandeep Ummat, Jaspreet Kaur, Navjot Kaur," Research Paper on Big Data and Hadoop", *IJCST Vol. 7, Issue 4*, Oct - Dec 2016 ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)
- [9] Agrawal D., Bernstein P., Bertino E., Davidson S., Dayal U., Franklin M., . . . . Widom J. (2012) "Challenges and Opportunities with Big Data: A white paper prepared for the Computing Community Consortium committee of the Computing Research Association" <http://cra.org/ccc/resources/ccc-led-whitepapers/>
- [10] Bijesh Dhyani and Anurag Barthwal, "Big Data Analytics using Hadoop", *International Journal of Computer Applications (0975 – 8887) Volume 108 – No 12*, December 2014