Asian Journal of Convergence in Technology *Volume II, Issue III*
Issn No.:2350-1146, I.F-2.71

Big Data: Trends and Challenges

Prachi P. Abhyankar,

Department of Information Technology
Finolex Academy of Management and Technology, Ratnagiri 415612
p.abhyankar2@gmail.com.

Swati A. Powar

Department of Information Technology
Finolex Academy of Management and Technology, Ratnagiri 415612
swati.powar@gmail.com

Abstract: **Organizations and companies today generate tremendous amount of data that is received and generated from various sources. Big data analytics is not only about handling this tremendous amount of data. It is also considering the volume, structure and quality of data. Big data can be analysed for better decision making and strategic business moves. This paper reviews big data concepts along with the challenges faced by the organizations to handle and store this data.**

Index Terms – Big Data, Challenges in Big Data

I. INTRODUCTION

Data has been increasing massively in the recent years due to technology and hence organizations require efficient techniques to handle this data. Big data is a collection of massive data sets where traditional data management and analysis tools cannot be applied to extract knowledge. Data collection is considered as big data when it is so large that an organization cannot effectively or affordably manage or exploit it using conventional data management tools. Volume, variety and velocity are the three defining properties of big data. Volume corresponds to the amount of data generated by organizations. Variety refers to managing complex data types including structured and unstructured data. Velocity refers to speed of data processing, capture and share.

A. Characteristics of Big Data

The Fig 1. shows three components of big data which are volume, variety, velocity [1, 3, 4, 6].

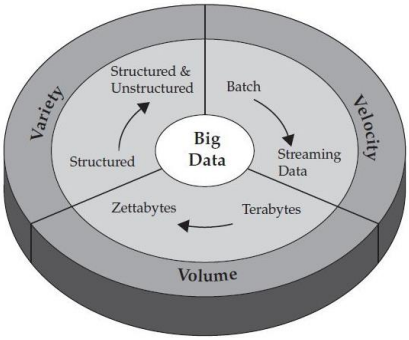

Fig 1. Characteristics of Big Data

Volume: Massive information sets that are generally in petabytes. It is the amount of data collected.
Variety: Data comes in all types of formats, structured, numeric data to unstructured text documents, email, video, and audio, transactional data.
Velocity: Data streams at an unprecedented speed must be dealt with in a timely manner. RFID tags, sensors are driving needs to deal with massive data in near-real time.

B. Big Data Architecture

The architecture of big data consists of different parts where each component of architecture has several alternatives with its own advantages and disadvantages for a particular workload. It is premised on a skill set for developing reliable, scalable, completely automated data pipelines. It requires profound knowledge of every layer in stack beginning with cluster design and spanning everything from Hadoop tuning to setting up top chain responsible for processing the data. It consists of following parts [6]:

Infrastructure as a Service (IaaS): This includes the storage, servers, and network as the base, inexpensive commodities of the big data stack. This stack can be bare metal or virtual (cloud). The distributed file systems are part of this layer.

Platform as a Service (PaaS): The NoSQL data stores and distributed caches that logically queried using query languages form the platform layer of big data. This layer provides the logical model for the raw, unstructured data stored in the files.

Data as a Service (DaaS): The entire array of tools available for integrating with the PaaS layer using search engines, integration adapters, batch programs, and so on in this layer.

Big Data Business Functions as a Service (BFaaS): Specific industries—like health, retail, ecommerce, energy, and banking—can build packaged applications that serve a specific business need and leverage the DaaS layer for cross-cutting data functions.

II. BIG DATA TECHNIQUES

Wide variety of techniques have been developed and adapted to visualize, analyse, manipulate and aggregate big data to make this kind of data volume tractable. Hadoop provides a comprehensive tool set for building distributed

systems including data storage, data analysis and coordination. Conventional data technologies and methods are most of the time slow, expensive and not suitable to handle the storage and the processing of large growing volumes of heterogeneous data. Some of these techniques include [5]:

A/B testing: used to compare different options against a control grouping order to determine what treatments will improve a given objective.

Cluster Analysis: used for classifying objects that splits a diverse group into smaller groups of similar objects.

Ensemble Learning: uses multiple predictive models to obtain better predictive performance.

Network Analysis: used to analyse connections between nodes in a network and their strength.

Machine Learning: makes use of artificial intelligence to automatically learn to recognise complex patterns and make intelligent decisions based on data.

A. Frameworks

Frameworks are based on many concepts like [3]:

Distributed storage: Unlike traditional systems, they store blocks of very large files across multiple nodes. They are designed to run on low-cost hardware and provide a high streaming access to data sets.

Massive Parallel Processing: the multiple time consuming tasks of Big Data applications, are processed in parallel across several servers. MPP helps to avoid copying distant data to execute computations. It executes jobs where data are stored in order to minimize network congestion and to ensure a fast processing.

Fault tolerant and scalability: big data systems are usually based on reliable architecture that has capability to handle additional clusters and handle more data and massive processing.

B. Databases in big data

NoSQL systems are distributed, non-relational databases designed for large-scale data storage and for massively-parallel data crunching across a large number of commodity servers. NoSQL (Not Only SQL) databases provide a cheaper way (than RDMS) to handle the storage and the management of Big Data in distributed environment. Such databases offer different levels of fault-tolerance and data availability. Usually do not support indexing and SQL querying. They are also often slow in handling large queries. In addition, unlike RDBMS, they do not ensure ACID principles for reliable transactions [3]. To address these limitations, the NewSQL was developed for Big Data applications. It constitutes a new relational database management systems based on a distributed architecture. It also ensures good data availability and performance of online transaction processing.

III. CHALLENGES IN BIG DATA MANAGEMENT

There are many different challenges that have to be faced by the organizations. Some of them include [1, 2, 5, and 7]:

1. Hadoop, the framework and set of tools for processing large data sets was originally designed to work on clusters of physical machines. That has now changed. Increasing number of technologies are now available for processing data like Google's BigQuery data analytics service, IBM's Bluemix cloud platform[5].
2. Distributed analytics framework like MapReduce are evolving into distributed resource managers that are gradually turning Hadoop into general purpose data operating system [5]. The ability to run different kinds of queries and operations in Hadoop tends to turn it into low cost, general purpose place to analyse.
3. Unlike traditional database theory, designing the data set before data entry can be skipped and design can be made dynamically. People who use this must be highly skilled. These data sets may lack traditional properties of monitoring access control, encryption, security and tracing the lineage of data from source to destination.
4. With big data, analysts have tremendous amount of data to work with. Also large processing power to handle number of records with numerous attributes. Traditional statistical analysis based methods have their limitations. Speed with which problem can be solved is also affected.
5. Also last but not the least, basic challenge is to prioritize data which means understand which data is noise and which is really useful amongst huge piles of data. The use of cloud computing and virtualization further complicates the decision to host big data solutions outside the enterprise. Organizations struggle to determine how long this data has to be retained, as some data is useful for making long-term decisions, while other data is not relevant even a few hours after it has been generated. With the advent of new technologies and tools required to build big data solutions, availability of skills is a big challenge [1].

IV. SOME POSSIBLE SOLUTIONS

In order to increase the speed, ApacheSpark, a large scale data processing engine and its associated query tool SparkSQL are being tested [5]. Interactive capability and streaming capabilities are also adequate. Alternatives to traditional SQL based relational databases called NoSQL (Not Only SQL) are gaining popularity. Large number of open source NoSQL databases are already present each with its own specialization. For example, a NoSQL product with graph database capability called ArangoDB has ways to

analyse network of relationships than traditional database [5]. Another set of machine learning tools based on neural networks is still evolving called deep learning. Use of in-memory databases to speed up analytical processing is also on rise. HTAP (Hybrid Transaction Processing) allows transactions and analytical processing to reside in the same in-memory database.

Together, these diverse technologies can fulfil almost any big data access, analysis and storage requirement. Knowledge of suitable technology for right type of task is required along with advantages and disadvantages of particular solution in terms of usability, maturity, cost, security etc.

IV. APPLICATIONS

Big Data has been of concern to organizations working in physical sciences (meteorology, physics), life sciences (genomics, biomedical research), government (defence, treasury), finance and banking (transaction processing, trade analytics), communications (call records, network traffic data), and, of course, the Internet (search engine indexation, social networks) [3].

1. Banking: with large amount of information streaming in from countless sources, banking sectors have to face innovative ways to manage big data. While it's important to understand customers and boost their satisfaction it's important to minimize risk and fraud. Big data brings new insights to analyse such situations with new techniques.
2. Government: government agencies can apply analytics to their big data and manage utilities, agencies, traffic congestion or prevention of crime. The issues of transparency and privacy must also be handled.
3. Education: educators armed with data-driven insights can make significant impact on school systems, students and curriculum. Identification of at-risk students, their progress can implement better system for evaluation.
4. Health Care: when big data is managed effectively it can uncover hidden insights to improve patient care. Several applications of big data have been tested to improve public and private medical service and to better support patients and medical practitioners.
5. Manufacturing: manufacturers can boost quality and output while minimizing waste and solving problems faster and making more agile business decisions.
6. Retail: customer relationship building is crucial to retail industry and big data analytics can be used for effectively handling transactions, marketing products.

CONCLUSION

With so many emerging trends around big data analytics, organizations need to create conditions that will allow experimentation. The importance of big data doesn't revolve around how much data is with the organization but what can be done with it. Data can be collected from various sources and analysed to enable cost reductions, time reductions, new product development and optimized offerings and smart decision making. There are left open challenges still to be resolved to enhance the capabilities of big data applications.

REFERENCES

[1] Munesh Kataria, Ms. Pooja Mittal, "BIG DATA: A Review", *International Journal of Computer Science and Mobile Computing*, Vol 3. Issue 7, July 2014, pg 106-110

[2] Anirudh Kadadi, Rajeev Agrawal, Christopher Nyamful, Rahman Atiq*, "Challenges of Data Integration and Interoperability in Big Data", *2014 IEEE International Conference on Big Data*

[3] Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, Samir Belfkih, "An Overview of Big Data Opportunities, Applications and Tools",*IEEE 2015*

[4] Ankit Kumar Tiwari Hemlata Chaudhary Surendra Yadav , "A Review on Big Data and its Security", *IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems* **ICIIECS'15**

[5] www.computerworld.com/article/2690856/trends-in-big-data-analytics.html

[6] Sravanthi Kanchi, Shubhrika Sandilya, Shashank Ramkrishna, Siddhesh Manjrekar, Akshata Vhadgar**, "** *Challenges and Solutions in Big data management - An Overview"*, 3rd International Conference on Future Internet of Things and Cloud, 2015

[7] Laura Wilber, *"A Practical Guide to Big Data: Opportunities, Challenges, Tools"*, 2012 Dassault Systems.