A Survey on Big Data Management

C. Sreedhar, Dr. N. Kasiviswanath

Abstract—**Big data has gained its popularity in the recent years. Industries, organizations in government and private sectors are striving towards handling huge volumes of data produced towards effective business and governance. Data is no more confined to structured data. The data generated by web logs, sensor nodes, climate monitoring devices, smart phones, airlines, multimedia data, tweets and XML data are in different data formats (unstructured, semi-structured). Handling large volumes of data and such different data formats raises the need for proper data management in big data. In this paper, we propose a four-stage method required to manage the data in big data management and finally review an analysis of big data management**

***Index Terms*—About four key words or phrases in alphabetical order, separated by commas.**

I. Introduction

File Management System (FMS) ruled for over years during the evolution of the Computers. Due to the difficulties of typical file-processing systems such as data redundancy, inconsistency, data isolation, atomicity of updates, concurrent access by multiple users, has lead to the evolution of Database Management System (DBMS). DBMS and relational DBMS (RDBMS) played a vital role in solving real world complex problems. Digitization has created new types of large and real-time data from industries and organizations. Due to rapid growth in usage of smart devices, social media websites such as Facebook, Twitter, Youtube, and multimedia data, traditional database systems poses a great challenge in handling such huge volumes of data. This has lead to the evolution of Big Data Management Systems (BDMS). IDC defines Big Data technologies as a new generation of technologies and architectures designed to economically extract value from a very large volumes of wide variety of data by enabling high-velocity capture, discovery and or analysis [1].

Big data are described to be big in three characteristics namely volume, velocity and variety, summarized as 3V's [3] and identified the 3Vs as the biggest challenges for data management. NIST describes big data as the deluge of data which exceed the capacity or capability of current or conventional methods and systems [7]. Big data deals with huge volume, generated by the various data formats such as structured data (data in relational databases or in spreadsheets), unstructured data (email messages, audio and video data) and semi-structured data (HTML, XML tagged text). In the year 2010, more than 1 ZetaBytes (ZB) of data is generated [2]. The size of the data generated within the next few years, it is beyond predictable in practical. The applications of big data includes in banking, financial services, retail, energy, power, healthcare, airlines, manufacturing, telecommunications, pharmaceuticals, life science and many more..

II. Big data Issues and Challenges

A. Review Stage

It is at the conceptual level to describe about the issues and challenges faced by the big data in the organizations. Some of the issues and challenges of big data are discussed in this paper. Public sector and private sector organizations hold and have access to huge volumes of data, which may be considered as wealth of data and have the potential to transform services accurately to the needs of the citizens. Facing the challenges with the ever growing size of the data produced has to be managed in order to maximize the potential of big data. The data generated by the computers will be in the form of homogenous data in general. The data generated by the humans will be in heterogeneity and thus data must be prepared and collected in proper fashion to yield better results and analysis of data. In addition to the big data challenges, big data characteristics such as variety, volume, velocity, veracity and value has an important trends on the mobility of data, consumption and as well as ecosystem capabilities [9].

Big data in healthcare deals with the vast volumes of patient-centric data. Managing medical images and measuring pathological characteristics will be a challenging issue in protecting and securing the patient's data. Big data applications will have unpredictable results as the data scientists reveals new trends and new sequencing in genomics which were hidden previously. Huge data sets in multimedia has become predominant in the sites like youtube and it is the challenging issue to process and reply back the accurate result to the internet user.

Real-time analytics handles huge volumes of streaming data. In some scenarios like business, real-time traffic, it is no longer desirable to wait for prolonged period of duration for the results to be analyzed. Real-time streaming data may contain structured, unstructured data. Scalability plays a challenging issue in such areas. Technological advancement in the processors has increased at faster speeds has brought a new challenge in the scalability. The systematic approach towards data acquisition in order to enhance randomness in sampling the data and reduce bias is not apparent in the collection of big data sets [14]. Storage and transport can also

Manuscript received April 10, 2015

C Sreedhar, CSE Dept., G Pulla Reddy Engineering College, Kurnool.
Dr. N. Kasiviswanath, Prof & HOD, G Pulla Reddy Engineering College, Kurnool

be stated as the major challenge. Current disk technology limits to about 4 terabytes per disk. Though an exabyte of information can be processed on a single computer, it is quite complex to connect directly the required number of disks. IT companies store large amount of data as logs in order to deal with the problems related when error or failure occurs. Traditional database systems cannot handle these logs because of either volume or the nature of the data format. The privacy issues associated with the increased amount of data is an important challenge. Aadhar in India can be described as the best example for the issues and challenges related to big data. Millions of people in the country, the complete details of the citizens are encapsulated in an aadhar. Security and privacy becomes the major challenge in protecting the rights of the citizen and protecting the data into the unauthorized users is a critical concern.

III. BIG DATA MANAGEMENT

Traditional RDBMS tools are not designed to run on clusters of servers on which masses of data sets can be processed in parallel. Big data management is the alternative solution to capture, store, process huge volumes of data across clusters. We propose four major phases in BDM: Data Acquisition, Data storage, Data Processing and Data Delivery. Fig. 1 illustrates the phases of BDM. Each of these phases is described in the next sub sections.

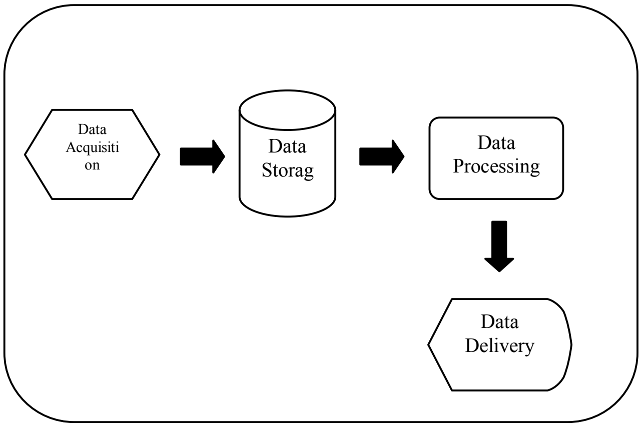

Fig. 1 Big Data Management

A. Data Acqusition

Data preparation, collection and cleansing is a crucial phase in BDM. In this phase, detecting and correcting inaccurate and incomplete data involves complex operations to clean up the data in database and achieve consistency to different sets of data merged from separate databases. Data sources such as logs, sensor networks, airlines produces huge volumes of data and filtering such data becomes a challenging issue. Identifying the useful data and redundant data is the key challenge of data scientists to automatically generate the right metadata which describes what and how the data is recorded, measured and analyzed. Metadata acquisition systems have an important role due to the fact that a processing error at one step can result in wrong analysis and visualization of the output.

B. Data Storage and Processing

Due to diverse complexity in organizations handling big data, there is no single comprehensive big data technology standard to meet all the requirements of big data. Currently, big data includes different entity types, traditional database models and data processing techniques [4]. Due to the limitations of traditional data storage processes, there is a need for novel architectures for storage, processing and integration with analytic systems of big data. Hadoop is one of the frameworks to meet data storage and data processing of huge volumes of unstructured data. Apache Hadoop is a java-based open source platform that enables processing of huge volumes of data across distributed nodes. Hadoop distributed file system (HDFS) is the core component of the Hadoop framework that manages data storage. Data is stored in the blocks of size 64MB on the local disk. MapReduce is the base component of Hadoop framework. With Hadoop, data storage and data processing may be achieved through HDFS and MapReduce respectively.

Traditional RDBMS are designed to handle volumes of data and run simple queries at a faster rate due to the reason that data is indexed so that only small portions of the data need to be examined in order to execute a query. In big data, data cannot be indexed for the reason that data may be in the form of unstructured and semi-structured. In order to execute a query, all the data has to examined. MapReduce is a data processing algorithm which uses a parallel programming implementation that involves distributing a task across multiple nodes running a map function. The map function splits the data into smaller parts and sends to different machines so that all the smaller parts are executed concurrently

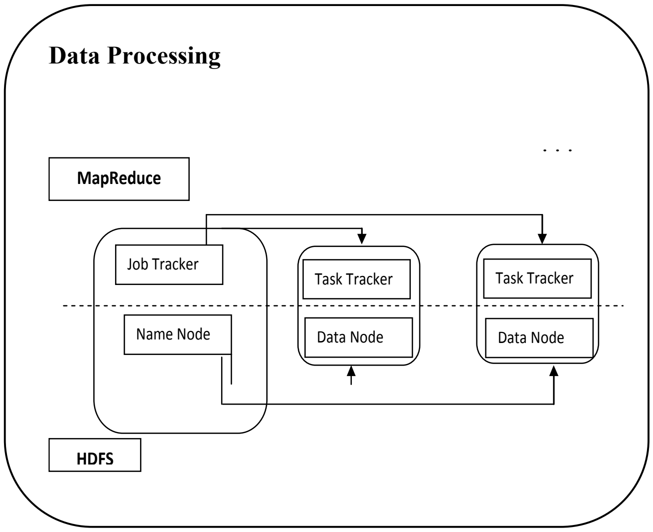

Fig. 2 Hadoop Data Processing

Fig. 2 illustrates the basic components of Hadoop, which supports data storing and data processing through HDFS and MapReduce. MapReduce is the software programming framework that simplifies processing of big data sets across clusters of computers. The map task splits a data set into independent chunks to be processed in parallel. The output of maps are sorted and then submitted to the reduce task. Both the data sets (input) and output are stored in HDFS. Typically the data is processed and stored on the same node, thus making more efficient to schedule tasks. HDFS forms as the primary storage that uses multiple replicas of data blocks, distributes them on nodes across clusters.

C. Data Delivery

The goal of the data scientist is to consider multidisciplinary scenarios by creating meaningful data and to input to the decision-makers with high value information in order to produce the best and possible decisions. The following are the steps for realizing the potential of big data-data collection, processing, managing, measuring, consuming, storing and governing. Data analysis is considered as the challenging issue in big data management. There are two major components that dive the success of data delivery namely data analytics and data visualization. Big data analytics is the process of examining big data sets to explore and unwrap the hidden patterns and other useful information which can be used to make better decisions by the decision-maker.

For effective large-scale data analytics, the process of locating, identifying, cleansing the data has to be done completely in an automated fashion. Analyzing big data supports the researchers, data analysts to make better and faster decisions by using advanced analytics techniques such as machine learning, predictive analytics, data mining, text analytics and NLPs. Data visualization is one of the profound developments in the world of big data. Visualization is a visual representation of the insights gained from data analysis. Data visualization tools allow the organizations to view information in an intuitive and in graphical way. When organizations are capable of analyzing big data, they can be benefited further through visualization, which plays a key role by making the individual analytic components understandable and by making them bring together the results of various decisions. The question arises of how much big data can be viewed and understood straight through visualization techniques. To answer this, there are several factors that depends on the type of data, queries made on the data, size of the display, computational scalability, sharing privileges and the visual layout

IV. Analysis of Big Data Management

A. Figures and Tables

Jinchuan CHEN et al [6], reviews big data challenges from a data management perspective. According to the author, the sources of big data include user generated contents, transactional data, scientific data, web data and graph data and describe five major challenging issues in big data management described in Table 1.

- big data diversity,
- big data segmentation,
- big data integration and cleaning,
- big data indexing and query
- big data analysis and mining

BDM Challenging Issue	Description
Big data diversity	How can organizations manage and benefit with the wealth of data is an challenging issue.
Big data segmentation	How can big data derive the best insight from data and the organizations must choose of what and how to categorize the data.
Big data integration and cleaning	Data integration is the process of standardizing the data definitions and data structures containing multiple data sources. Integrating and cleaning of different data sets with different data formats needs sophisticated algorithms to achieve better results.
Big data indexing and query	Most of the data sets are quite complex to introduce indexing due to the fact that it holds huge volume of data. Indexing and querying techniques are to be applied suitable to big data sets.
Big data analysis and mining	Data Mining and data analysis are the important factors that lead to the better results from big data.

Table 1. Challenges in Big Data Management

Ian Gorton et al [8] describes about NOSQL, as the means of achieving scaling of clusters with low cost, high performance through data replication and data set partitions across the clusters. The author illustrates the classification of data models namely document databases, key-value databases, column-oriented databases and graph databases described in Table 2.

Viet-Trung Tran et al [11] proposes a big data storage solution called Pyramid, which is a specialized array-oriented storage manager that demonstrates substantial scalability improvements. Scalability in data management is a challenging issue in BDM. Pyramid is an array-oriented storage and versioning solution for multi-dimensional data based on the following design principles:

- Array Versioning – Data updates are represented using immutable data and metadata. A new snapshot of the whole array is created, whenever a multi-dimensional array needs to be updated. Incremental snapshots in the form of independent arrays are used in storing the previous snapshots.
- Versioning array-oriented access interface – An interface is

designed to support multi-dimensional data which enables fine-grained versioning access to subdomains.

- Multi-dimensional aware data chunking – In this design principle, multi-dimensional data is stored in an array and arrays are split into chunks and are distributed among the storage elements.

Lock-free and distributed chunk indexing: Distributed quad-tree like structure is proposed to index the chunk layout. This design principle has an advantage of optimization for concurrent updates.

In [10], the authors focus on the importance of Big data and analytics. The main importance of big data consists in the areas to improve efficiency in handling huge volumes of data of different type. Big data can be turned to business advantage and in order to make every decision, there should be a mechanism to manage big data. BDM is best described with Apache Hadoop, one of the fastest growing big-data processing platform which enables distributed and clustered processing of large data sets across clusters of commodity servers [13].

Sattam Alsubaiee et al [12] propose ASTERIX as an approach to meet the challenges of BDM. Currrently this approach is capable of executing parallel queries. ASTERIX uses parallel, semi-structured information management system with the capabilities of ingesting, storing, indexing, querying, analyzing and publishing very large quantities of semi-structured data. Three layered model is used to achieve parallel, semistructured information management system.

- Hyracks layer: It is the bottommost layer and is runtime layer with the job of accepting and managing data-parallel computations requested either direct end uses or by layers above it. Jobs are submitted in directed acyclic graph forms which are made of operators and connectors.
- Algebricks algebra layer: This layer is used for parallel query processing and optimization by providing all traditional relational operators such as select, project and join.
- Parallel information management layer: Provides overview of software components of ASTERIX map to nodes in a shared-nothing cluster which serves as the runtime executor for query execution and storage management operations in ASTERIX.

Abhishek Roy et al [5] develops a system for end-to-end data processing of genomic data sets. This approach makes an attempt to improve efficiency in data quality and parallel processing techniques. Genomic data sets are complex in terms of storing, processing and analyzing. Data processing of genomic data is proposed in three phases:

- Sequencing and pre-analysis: This phase consists of collection of raw image to base calls and data pre-analysis. Raw data in the form of captured images is parsed into short read sequences. Data pre-analysis evaluates the quality of each read sequence and poor quality images are removed and formats the data for the downstream processing.
- Data Processing: Involves short read sequences alignment against a reference genome. After the alignment, genomic variants are detected against reference genome which includes single nucleotide variants, large structure variants. Variation and validation are the crucial steps in data processing phase.

Deep Analysis: Data integration and deep analysis are the two major components which produces a high-level information with biomedical meanings from all available data

Data Model	Description	Examples
Document database	In document database data model, it is assumed that document formats are self-describing and may contain with different data formats.	CouchDB and MongoDB
Key-value database	In key-value databases, key forms vital role in accessing and searching records. This data model replicates the data in order to achieve high scalability.	Riak and DynamoDB
Column-Oriented databases	The extension of key-value data model forms the basis for column-oriented databases, where a column is treated as key-value pair.	HBase and Cassandra
Graph databases	This model organizes data in the form of directed graph. Graph traversal and sub graph matching problems can be solved using this model	Neo4j and GraphBase

Table 2. Classification of data models.

V. Conclusion

Big data includes handling huge volumes of data. The need to process and mange large quantities of data is growing at a faster rate than predicted. Big data management should be capable of adopting new types of data in the next decade, which may pose a great challenge in framing new standards for measuring and managing the data. In order to take big advantage of big data management, big data tools like MapReduce over Hadoop and HDFS, promises to help the organizations better understand the needs of the problems raised from huge volumes of data. Big data presents technology with many boons and challenges and at the same time presents many opportunities to unearth hidden facts in science and technology. Although data management is viewed in terms of software, big data management should be viewed in a holistic way of combination of both hardware and software.

References

[1] John Gantz and David Reinsel, “The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East”. Study report, IDC, December 2012.

[2] Rick Villars, Marshall Amaldas, “IDC White Paper: Rethinking your Data Retention Strategy to Better Exploit the Big Data Explosion, October 2011.

[3] D. Laney, “3-d data management: Controlling data volume, velocity and variety”, META Group Research Note, February, vol. 6, pp. 1-4, February 2001.

[4] Samson Oluwaseun Fadiya, Serdar Saydam, Vanduhe Vany Zira, “Advancing big data for humanitarian needs”, Procedia Engineering 78, pp. 88 – 95, 2014. Available online at www.sciencedirect.com.

[5] Abhishek Roy, Yanlei Dial, Evan Mauceli, Yiping Shen, Bai-Lin Wu, “Massive Genomic Data Processing and Deep Analysis”. In proceedings of the VLDB Endowment, vol. 05, no. 12, 2012.

[6] Jinchuan CHEN, Yueguo CHEN, Xiaoyong DU, Cuiping LI, Jiaheng LU, Suyun ZHAO and Xuan ZHOU, “Big data challenge: a data management perspective”. In Front. Comput. Sci., DOI 10.1007/s11704-013-3903-7, pp. 157-164, 2013.

[7] R.M. Warrd, R. Schmieder. G. Highnam and D. Mittelman, “Big data challenges and opportunities in high-throughput sequencing”, Systems Biomedicine, vol. 01, no. 01, pp. 29-34, March 2013.

[8] Ian Gorton and John Klein, “Distribution, Data, Deployment: Software Architecture Convergence in Big Data Systems”. Software Engineering Institute, Carnegie Mellon University, www.sei.cmu.edu, May 2014.

[9] Cisco White Paper, Cisco Visual Networking Index, “Global Mobile Data Traffic Forecast”, Update 2010-2015, Feb 2011.

[10] Elena Geanina Ularu, Florina Camelia Puican, Anca Apostu, Manole Velicanu, “Perspectives on Big Data and Big Data Analytics”. Database Systems Journal col. III, no. 4. 2012.

[11] Vier-trung Tran, Bogdan Nicolae, Gabriel Antonie, “Towards Scalable Array-Oriented Active Storage: the Pyramid Approach”. https://hal.inria.fr/hal-00640900, pp. 19-25, 2012.

[12] Sattam Alsbaiee, Yasser Altowim, Hotham Altwaijry, Alexander Behm, Vinayak Borkar, Yingyi Bu, Michael Carey, Raman Grover, Zachary Heilbron, Young-Seok Kim, Chen Li, Nichola Onose, Pouria Pirzadeh, Rares Vernica, Jian Wen, “ASTERIX: An Open Source System for Big Data Management and Analysis (Demo)”. In proceedings of the VLDB Endowment, vol. 05, No. 12, 2012

[13] http://en.wikipedia.org/wiki/Apache_hadoop.

[14] Danah Boyd, Kate Crawford, “Six provocations for Big Data”. Available online: http:// papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431, July 2012

C Sreedhar completed his B.E, M.E and currently pursuing PhD in the Computer Science & Engineering stream. 15 papers were published in various National/ International Journals/Conferences. His research areas include Big Data, Wireless Networks, Security.

Dr. N. Kasiviswanath completed his B.E, M.S and PhD in the Computer Science stream. He has 25 research papers published in various National/International Journals/Conferences. His research areas include Wireless Networks, Data Structures.