Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 7, July 2014, pg.106 – 110

REVIEW ARTICLE

BIG DATA: A Review

Munesh Kataria[1], Ms. Pooja Mittal[2]
(M.Tech Student)[1], (Assistant Professor)[2]
[1,2]Department of Computer Science and Applications,
M. D. University, Rohtak-124001, Haryana, India

Abstract: Big data is pool of large and complex data sets so it becomes difficult to process data using database management tools. With the fast evolution of data, data storage and networking collection capacity, Big Data are rapidly growing in all science and engineering domains. The analysis of big data can be difficult as it often involves collection of mixed data based on different patterns or rules. The challenges include capture, storage, search, sharing, analysis, and visualization. The trend to larger data sets is because of the extra information drawn from analysis of a single large set of related data, compared to separate smaller sets with the same total amount of data. Big Data mining is the ability of extracting useful information from huge streams of data or datasets, that because of its velocity, variability and volume. This paper argues applications of Big Data processing model and also Big Data Mining.

Keywords: Big Data, Data Mining, Hadoop, Architecture

I. INTRODUCTION

Data is easier to capture and access through third parties such as Facebook, D&B, and others. Geo location data, social graphs, user-generated content, user's personal information, machine logging data, and sensor-generated data are just a few examples of the array of data captured. It's not surprising that developers find increasing value in leveraging this data to enrich existing applications and create new ones made possible by it. The use of the data is rapidly changing the nature of communication, shopping, advertising, entertainment, and relationship management. Applications that don't find ways to leverage it quickly will quickly fall behind. Scientists regularly face problems because of large data sets in many areas, including meteorology, genomics; complex physics simulations, biological environmental research, internet search, and finance and business informatics. Data sets grow in size in part because they increasingly gathered by widespread information-sensing mobile, remote sensing, software logs, cameras, microphones, radio frequency identification readers, and wireless sensor networks. Big data is difficult to work with using relational databases and desktop statistics, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers". "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management alternatives. Big data usually includes data sets with sizes beyond the ability of commonly-used software tools to capture, curate, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target, from a few dozen terabytes to many petabytes of data in a single data set. With this difficulty, a new platform of "big data" tools has arisen to handle sense making over large quantities of data, as in the Apache Hadoop Big Data Platform.

Big Data

Big data described by the three properties below—occasionally referred to as the three V's but organizations need fourth V i.e. value to build big data job.

Volume: massive information sets that are command of size bigger than data managed in habitual storage and analytical results. Imagine petabytes rather than terabytes. [3]

Variety: complex, variable and heterogeneous data, which produced in formats as different as public media, email, images, video, blogs, and Web explore histories.

Velocity: Data created as a stable with real-time queries for significant information to be present up on claim instead of batched.

Value: Resulting insights that for trends and patterns, difficult analysis based on graph algorithms, machine learning and statistical modeling. These analytics overtake the results of querying, reporting and business intelligence. [4]

II. BIG DATA ARCHITECTURE

Analogous to the cloud architectures, the big data landscape divided into four layers shown vertically in Figure 1

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.

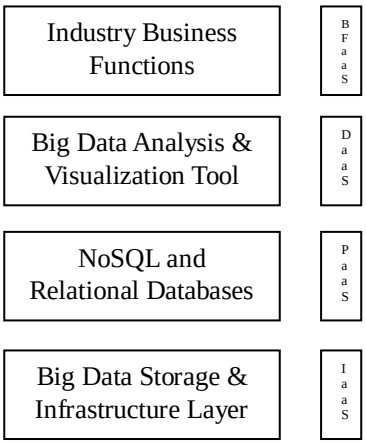

Figure 1: Big Data Architecture Layers

III. BIG DATA ANALYTICS

Without the emerge of new data-warehousing and technologies, there would no big data phenomenon. Data will be more extreme in the future (e.g. with the three Vs) and new techniques needed, making it possible to analyze this data. The last year's the ability to store and analyze data in comparison with the data that produced lagged behind. New data-warehousing and database technologies introduced to address this problem. This section will elaborate on the developments in different (technological) fields making big data analytics possible.[2]

THE RISE OF THE CLOUD

The rise of the cloud plays a significant role in big data analytics and likely this role will increase as the cloud adopted by a growing number of organizations. Cloud computing is a successful model of service oriented computing. It provides services at different levels of IT, for example, Infrastructure as a Service (IaaS), Platform as a service (PaaS) and Software as a Service (SaaS). Some advantages of cloud computing, compared to in-house computing, are:

• Infinite computing resources are available on demand;

• No up-front commitment by cloud users; users can start small but think big;

• Pay for use of resources on a short-term basis (e.g. more resources on peak hours);

These advantages are useful for big data analytics in several ways. To analyze data, there must be data available and as described earlier, data created in a much faster way than ever before. Therefore, a lot of storage space is necessary (especially with the “store and analyze” approach). A significant proportion of data organizations own created by end-

users (such as visitors of the organization's website) and hence, cannot control by the organization itself. This shows the need to easily demand more resources from the cloud provider when required.[4]

THE GLOBAL INTRODUCTION OF NOSQL DATABASES

New forms of databases have developed, giving up at least one constraint of the ACID principle. ACID stands for atomicity (a transaction is "all or nothing"), consistency (the database will be in a consistent state before and after a transaction), isolation (transactions may not interfere with each other) and durability (a transaction is always permanent). Since the amount of data is growing extremely fast compared with how technology evolves (e.g. Moore's law and Kryder's law) and the structure of data itself, scaling databases has become important. Since vertical scaling (e.g. moving the database to a more powerful system or increasing the capacity of the database system) is always limited to the fastest possible system available and relatively expensive, horizontal scaling (e.g. distributing the database, or its functions, across multiple nodes) is often preferred since relatively cheap commodity systems can be used and no physical limitations are in play. However, a horizontal scaling database has disadvantages comparable to many other distributed systems. First of all, not all constraints of ACID can be applied at the same time. A theory, known as the CAP theorem, says that if you want consistency, availability, and partition tolerance, you have to settle for two out of three. Partition tolerance refers to "no set of failures less than total network failure is allowed to cause the system to responding correctly". Since the ACID principle cannot longer be fully fulfilled when using a distributed database, a new alternative has been introduced, known as BASE. BASE stands for "Basic Availability", "Soft-state" and "Eventual consistency" indicating that, rather than requiring consistency after every transaction, it is enough for the database to eventually be consistent. A possible implication of moving from ACID to BASE is two customers buying the same book while there is only one copy available. Since two transactions can occur almost at the same time, both customers think they have bought the book and therefore the organization needs to apologize to one of them. Although this is not an ideal situation, it is better to slowing down their site which will affect all customers.

Most new databases are NoSQL compliant where NoSQL is often defined as "NotOnly SQL" or "Not Relational". In his paper Cattell (2011) identifies NoSQL databases by the following six key features:

- The ability to horizontally scale throughput over many servers (nodes);
- The ability to replicate and to distribute data over many servers (nodes);
- A simple call level interface or protocol;
- A weaker concurrency model than ACID (e.g. BASE);
- Efficient use of distributed indexes and RAM for data storage;
- The ability to dynamically add new attributes to data records.

Most existing NoSQL databases can be categorized in four types of databases, namely key-value stores, document stores, extensible record stores and scalable relational systems (Cattell, 2011). The latter type refers to recent developments which makes it possible to horizontal scale traditional relational databases. However, benchmarks have showed that these databases cannot achieve scaling comparable with "real" NoSQL systems. This in turn indicates that there is a true need for highly scalable NoSQL databases, especially when dealing with complex data (e.g. in terms of the three Vs discussed earlier).

HADOOP, THE OPEN SOURCE HEART OF BIG DATA ANALYTICS

According to Forrester, Hadoop is the nucleus of the next generation enterprise data warehousing by delivering cloud-facing architectures. Created by Doug Cutting, the creator of Apache Lucene, Hadoop provides a comprehensive tool set for building distributed systems, including data storage, data analysis and coordination. Hadoop originates from Apache Nutch, an open source web search engine. After realizing that existing architectures would not scale to the billions of pages on the web, the initiators wrote an open source implementation based on Google's distributed file system, called Nutch Distributed Filesystem (NDFS). In 2004 Google released a paper that introduced MapReduce, a parallel programming model and an associated implementation for processing, analyzing and generating large data sets across a cluster of commodity machines (Dean &Ghemawat,2008), to the public. Nearly a year later all Nutch algorithms ported to useMapReduce and NDFS. In 2006, Nutch became a separate subproject under the name Hadoop and two years later it became a top-level project at Apache, confirming its success. In that year, Hadoop used by many international organizations such asLast.fm and Facebook. For many, Hadoop is a synonym for big data because of its powers to store and handle huge amounts of (unstructured) data within a smaller time frame in an economically responsible way. So the Hadoop ecosystems play a major role in big data analytics. Figure 2 illustrates the "mountain of data" commonly find within organizations. With traditional data analytics, only the peak analyzed and utilized to create value or support value creation. This peak often consists of highly structured data stored in traditional data warehouses. Since the amount of unstructured data is growing rapidly as described earlier, this peak is becoming relatively smaller. With Hadoop, it is possible to store and analyze unstructured data in a much smaller time frame using the power of distributed and parallel computing on commodity hardware. More important, the line indicating the boundary of data that can be utilized and data that cannot, is dropping, leading to a much greater peak and hence, in more possible value. Together with its free license, huge community and open source techniques, many initiatives using Hadoop have emerged, also indicating its success.

Also, many big IT organizations started to distribute their own commercial version of Hadoop by adding enterprise support, additional functionalities and tools and even bundled with specific hardware.

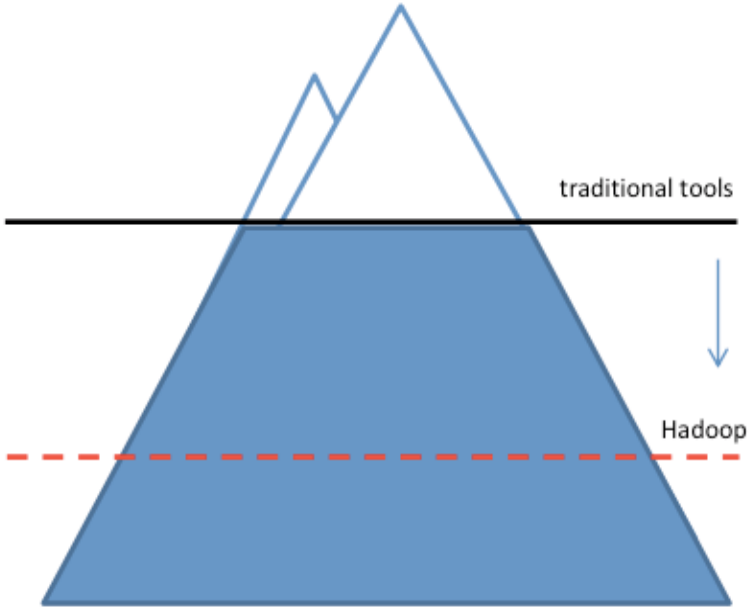

Figure 2: New technologies make it possible to utilize more data (Knulst, 2012 [1])

IN-MEMORY ANALYTICS

The techniques discussed so far are mainly focusing on analyzing huge amounts of unstructured data (both the volume and variety part of big data). Although these techniques both make use of disk storage and memory, another generation of data aware housing runs primary in memory significantly improving the time needed to get the job done, sometimes even 100.000 faster than traditional techniques as Steve Lucas(2012) argued. Hence, in-memory analytics focuses on the velocity part of big data. In-memory analytics can make real time analytics possible, even when dealing with large amounts of data. This in turns means that insights are more valuable, taking account the fact that some knowledge loses value when it becomes "old", especially when also known by competitors. In-memory analytics uses a so called "in-memory database" (IMDB) to perform analysis. Although this technique is not new (it originates from the 1980s) it became an possible to utilize more data (Knulst,2012 [1])

IV. DATA MINING FOR BIG DATA

Data mining includes extracting and analyzing huge amounts of data to discover models for big data. The methods came out of the grounds of artificial intelligence and statistics with a database management.

Searching information from data takes two major forms: prediction and description. It is tough to know what the data shows. Data mining is used to summarize and simplify the data in a way that we can recognize and then permit us to gather things about specific cases based on the patterns. Normally, the objective of the data mining is either prediction or classification. In classification, the thought is to arrange data into sets. For example, a seller might be attracted in the features of those who answered versus who didn't answered to an advertising. There are two divisions. In prediction, the plan is to predict the rate of a continuous variable. For example, a seller might be involved in predicting those who *will* reply to a promotion. Distinctive algorithms used in data mining are as follows:

A. **Classification trees:** A famous data-mining system that is used to categorize a needy categorical variable based on size of one or many predictor variables.

B. **Logistic regression:** An algebraic technique that is a modification of standard regression but enlarges the idea to deal with sorting. It builds a formula that predicts possibility of happening as a role of the independent variables.

C. **Neural networks:** A software algorithm molded after the matching architecture of animal minds. The network includes of output nodes, hidden layers and input nodes. Each unit associated with a weight. Data mentioned to the input node, and by a method of trial and error, the algorithm correct the weights until it reaches a stopping criteria.

D. **Clustering techniques like K-nearest neighbors:** A procedure that identifies class of related records. The K-nearest neighbor technique evaluates the distances between the points and record in the historical data. It assigns record to the set of its nearest neighbor in a data group.

V. BIG DATA CHALLENGES

One of the very basic challenges is to understand and prioritize the data from the garbage that is coming into the enterprise. Ninety percent of all the data is noise, and it is a daunting task to classify and filter the knowledge from the noise. In the search for inexpensive methods of analysis, organizations have to compromise and balance against the confidentiality requirements of the data. The use of cloud computing and virtualization further complicates the decision to host big data solutions outside the enterprise. But using those technologies is a trade-off against the cost of ownership that every organization has to deal with. Data is piling up so rapidly that it is becoming costlier to archive it. 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. A higher level of proficiency in the data sciences required to implement big data solutions today because the tools are not user-friendly yet. They still require computer science graduates to configure and operationalize a big data system.

VI. CONCLUSION

Today many technologies are emerging in the field of Big Data. Hadoop file system is one of them. Apache Hadoop is an open-source software framework that supports data-intensive distributed applications, licensed under the Apache v2 license. It supports the running of applications on large clusters of commodity hardware.

Big data is directed to continue rising during the next year and every data scientist will have to handle a large amount of data every year. This data will be more miscellaneous, bigger and faster. We discussed in this paper several insights about the subjects and what we think are the major concern and the core challenges for the future. Big Data is becoming the latest final border for precise data research and for business applications. Data mining with big data will assist us to discover facts that nobody has discovered before. By the data level, the independent information sources and the range of the data gathering environments, habitually result in data with complex conditions, such as missing unsure values. The vital challenge is that a Big Data mining structure needs to consider complicated interaction between data sources, samples and models along with their developing changes with time and additional probable factors. A system wants to be cautiously designed so that unstructured data can be connected through their composite relationships to form valuable patterns, and the development of data volumes and relationships should help patterns to guess the tendency and future.

REFERENCES

[1] Knulst, " *De stand van Hadoop*", Incentro, 2012.
[2] Russom, " *Big Data Analytics*", TDWI Research, 2011.
[3] Bloem, J. Doorn, M. V. Duivestein, S. Manen & Ommeren, " *Creating clarity with Big Data*", Sogeti, 2012.
[4] Vinayak Borkar, Michael J. Carey, Chen Li, "Inside "*Big Data Management": Ogres, Onions, or Parfaits?*", EDBT/ICDT 2012 Joint Conference Berlin, Germany, 2012.
[5] Domingos, Pedro, and Geoff Hulten. "Mining high-speed data streams."*Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2000.
[6] Wu, Xindong, et al. "Data mining with big data." *Knowledge and Data Engineering, IEEE Transactions on* 26.1 (2014): 97-107.
[7] Sawant, Nitin, and Himanshu Shah. "Big Data Application Architecture." *Big Data Application Architecture Q & A*. Apress, 2013. 9-28.

