Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 2, Number 9; April-June, 2015 pp. 21-24

http://www.krishisanskriti.org/acsit.html

Era of Big Data: Features and Challenges

Aman Gupta[1] and Ajay Indian[2]

[1]MCA Scholar, IICA, Invertis University, Bareilly
[2]IICA, Invertis University, Bareilly
E-mail: [1]guptaaman570@gmail.com, [2]ajay.i@invertis.org

Abstract—*Big Data refers to various forms of large information sets that require special computational platforms in order to be analyzed. In 2001, industry analyst Doung Laney (currently with Gartner), articulated the mainstream of definition of Big Data as the three Vs. Volume, Velocity and Variety. SAS considered two additional dimensions when thinking about Big Data: the Variability and Complexity .Oracle defined Big Data in terms of four Vs – Volume, Velocity, Variety and Value.*

Computing has become global number of devices like cell phones, smart phones, laptops; Personal sensors are creating countless new digital oceans of information. A few years ago we talked about data storage in megabytes and gigabytes but now a day's huge amount of data found on internet which is close to 500 billion gigabytes.All the peoples around the world act as a sensor and generate data at every second either it is our location shown on Facebook through gprs signals or our movement by gathering data from our mobile phones or smartphone devices. The 5C architecture can be used to process big data based on Connection, Conversion, Cyber, Cognition, and Configuration.

1. INTRODUCTION

Big Data is a broad term for data sets so large or complex that they are difficult to process using traditional data processing applications. Big Data is a term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications Data sets are growing in size in part because they are increasingly being gathered by information-sensing mobile devices, remote sensing. Big data includes data sets with sizes outside the ability of commonly used software tools to capture, manage and process data within an elapsed time [17]. Big data is the set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale. Big data can also be defined as "Big data is a large volume unstructured data which cannot be handled by standard database management systems like DBMS, RDBMS or ODBMS."

Data in unstructured form is growing more rapidly. This type of data includes all human information like tweets, Facebook data, Google maps, medical records and images, high definition videos etc. In most of the organizations 70 to 80% of data is in unstructured form and it is very difficult to analyze that data. Wedefine the rapidly grown data in different organizations.

YouTube	. 300 hours of videos are uploaded to YouTube every minute. . Each month, more than 1 billion unique users access YouTube. . Everyday people watch hundreds of millions of hours on YouTube and generate billions of views. The numbers of hour's people are watching on YouTube each month is up 50% year over year.
Facebook	. In every 20 minutes 3 million messages sent. . In every 20 minutes 1 million links are shared. . Total number of monthly active users is 131 billion.
Twitter	. The site has over 645,750,000 users. . The site generates 175 million tweets per day.
Google&LinkedIn	. 1 billion accounts have been created. . Total number of Linked users 313,000,000.

2. FEATURES OF BIG DATA

Big Data can be described in following features:

i. **Volume** – The quantity of data that is generated is very important in this context. It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered as Big Data or not. The name 'Big Data' itself contains a term which is related to size and hence the characteristic.

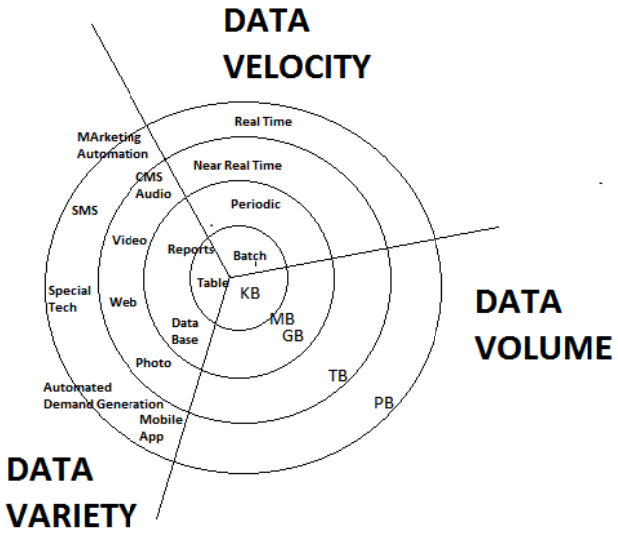
ii. **Variety** - The next aspect of Big Data is its variety. This means that the category to which Big Data belongs to is also a very essential fact that needs to be known by the data analysts. Variety refers to the various types of the data that can exist, for example, text, audio, video, and photos. This helps the people, who are closely analyzing the data and are associated with it, to effectively use the data to their advantage and thus keeping the importance of the Big Data.

iii. **Velocity** - The term 'velocity' in the context refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development. Velocity refers to the low latency, real-time speed at which the analytics need to be applied.

iv. **Variability** - This is a factor which can be a problem for those who analyses the data. This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

v. **Veracity** - The quality of the data being captured can vary greatly. Accuracy of analysis depends on the veracity of the source data.

vi. **Complexity** - Data management can become a very complex process, especially when large volumes of data come from multiple sources. These data need to be linked, connected and correlated in order to be able to grip the information that is supposed to be conveyed by these data. This situation, is therefore, termed as the 'complexity' of Big Data.



3. 5C ARCHITECTURE

Big Data Analytics for Manufacturing Applications can be based on 5C Architecture (Connection, Conversion, Cyber, Cognition and Configuration).

i. **Smart Connection**- Acquiring accurate and reliable data from machines and their components is the first step in developing a cyber-physical system application. The data might be directly measured by sensors or obtained from controller or enterprise manufacturing systems such as ERP, MES, SCM and CMM. Two important factors at this level have to be considered. First, considering various types of data, a seamless and tether-free method to manage data acquisition procedure and transferring data to the central server is required where specific protocols such as MT Connect, etc. are effectively useful. On the other hand, selecting proper sensors (type and specification) is the second important consideration for the first level.

ii. **Data-to-Information Conversion**-Meaningful information has to be inferred from the data. Currently, there are several tools and methodologies available for the data to information conversion level. In recent years, extensive focus has been applied to develop these algorithms specifically for prognostics and health management applications. By calculating health value, estimated remaining useful life, etc., the second level of CPS (Cyber Physical System) architecture brings self-awareness to machines.

iii. **Cyber**- The cyber level acts as central information hub in this architecture. Information is being pushed to it from every connected machine to form the machines network. Having massive information gathered, specific analytics has to be used to extract additional information that provides better insight over the status of individual machines among the fleet. These analytics provide machines with self-comparison ability, where the performance of a single machine can be compared with and rated among the fleet and on the other hand, similarities between machine performance and previous assets (historical information) can be measured to predict the future behavior of the machinery. In this paper we briefly introduce an efficient yet effective methodology for managing and analyzing information at cyber level.

iv. **Cognition**-Implementing CPS upon this level generates a thorough knowledge of the monitored system. Proper presentation of the acquired knowledge to expert users supports the correct decision to be taken. Since comparative information as well as individual machine status is available, decision on priority of tasks to optimize the maintaining process can be made. For this level, proper info-graphics are necessary to completely transfer acquired knowledge to the users.

v. **Configuration**-The configuration level is the feedback from cyber space to physical space and act as supervisory control to make machines self-configure and self-adaptive. This stage acts as resilience control system (RCS) to apply the corrective and preventive decisions, which has been made in cognition level, to the monitored system.

4. HADOOP SYSTEM

Hadoop includes an ecosystem of other products built over the core HDFS and Map Reduce layer to enable various types of operations on the platform.The volume of data that enterprises acquire every day is increasing exponentially. The challenge

these organizations now face is what to do with all this data and how to collect key insights from this data.

A few popular Hadoop components are as follows:

i. **Mahout**:This is an extensive library of machine learning algorithms.

ii. **Pig**: Pig is a high-level language (such as PERL) to analyze large datasets with its own language syntax for expressing data analysis programs, coupled with infrastructure for evaluating these programs.

iii. **Hive**: Hive is a data warehouse system for Hadoop that facilitates easy data summarization, ad hoc queries, and the analysis of large datasets stored in HDFS. It has its own SQL-like query language called **Hive Query Language** (**HQL**), which is used to issue query commands to Hadoop.

iv. **HBase**:**HBase** (**Hadoop Database**) is a distributed, column-oriented database. HBase uses HDFS for the underlying storage. It supports both batch style computations using Map Reduce and atomic queries (random reads).

v. **Sqoop**: Apache Sqoop is a tool designed for efficiently transferring bulk data between Hadoop and Structured Relational Databases. **Sqoop** is an abbreviation for **(SQ) L** to **Had (oop).**

Functions of Hadoop

- HDFS (Hadoop Distributed File System)
- Map Reduce

5. R SYSTEM

R is a very amazing tool that makes it a snap to run advanced statistical models on data, translate the derived models into colorful graphs and visualizations, and do a lot more functions related to data science.It is now possible to store these vast amounts of information on low cost platforms such as Hadoop.One key drawback of R, though, is that it is not very scalable. The core R engine can process and work on very limited amount of data. Using R with Hadoop will provide an elastic data analytics platform that will scale depending on the size of the dataset to be analyzed.

A few popular R components are as follows:

i. **RHIPE:**R and Hadoop Integrated Programming Environment (RHIPE) is a free and open source project. RHIPE is widely used for performing Big Data analysis with D&R analysis. D&R analysis is used to divide huge data, process it in parallel on a distributed network to produce intermediate output, and finally recombine all this intermediate output into a set.

ii. **RHDFS:** This is an R package for providing all Hadoop HDFS access to R. All distributed files can be managed with R functions.

iii. **RMR:** This is an R package for providing Hadoop Map-Reduce interfaces to R. With the help of this package, the Mapper and Reducer can easily be developed.

iv. **RHBase:**This is an R package for handling data at HBase distributed database through R.

Functions of R

- Effective programming language
- Relational database support
- Data analytics
- Data visualization
- Extension through the vast library of R packages

6. SURVEY ON BIG DATA GROWTH

Investment in big data technologies continues to extend, according to a recent survey by Gartner, Inc., which found that 73 percent of respondents have invested or plan to invest in big data in next 24 months, up from 64 percent in 2013. The survey also indicates organizations are starting to get off the fence about their big data investments - the number of organizations stating they had no plans for big data investments fell from 31 percent in 2013 to 24 percent in 2014 [18].

In 2020, over 1/3 of all data will live in or pass through the cloud. Data Production will be 44 times greater in 2020 than it was in 2009. Individuals create 70 percent of all data. Enterprise store 80 percent.

7. BIG DATA: OPPORTUNITIES AND CHALLENGES

Database technology was considered for the task, but was found to be neither well-suited nor cost-effective for those purposes. The turn of the millennium then brought further challenges as companies began to use information such as the topology of the Web and user's search histories in order to provide increasingly useful search results [5].

Web-Scale Storage: Google's technical response to the challenges of Web-Scale data management and analysis was simple, by database standards, but kicked off what has become the modern “Big Data” revolution in the systems world [6]. To handle the challenge of Web-Scale storage, the Google File System (GFS) was created [7]. GFS provides clients with the familiar OS-level byte-stream abstraction, but it does so for extremely large files whose content can span hundreds of machines in shared-nothing clusters created using inexpensive commodity hardware [5].

Data Processing: To handle the challenge of processing the data in such large files, Google pioneered its Map Reduce programming model and platform [7]. This model,

characterized by some as “parallel programming for dummies”, enabled Google's developers to process large collections of data by writing two user-defined functions, map and reduce, that the Map Reduce framework applies to the instances (map) and sorted groups of instances that share a common key (reduce) – similar to the sort of partitioned parallelism utilized in shared-nothing parallel query processing.

The Hadoop system has quickly gained traction, and it is now widely used for use cases including Web indexing, clickstream and log analysis, and certain large-scale information extraction and machine learning tasks. Soon tired of the low-level nature of the Map Reduce programming model, the Hadoop community developed a set of higher-level declarative languages for writing queries and data analysis pipelines that are compiled into Map Reduce jobs and then executed on the Hadoop Map Reduce platform.

Pig is relational-algebra-like in nature, and is reportedly used for over 60% of Yahoo!'s Map-Reduce use cases; Hive is SQL-inspired and reported to be used for over 90% of the Facebook Map Reduce use cases. Microsoft's technologies include a parallel runtime system called Dryad and two higher-level programming models, Dryad LINQ and the SQLlike SCOPE language [10], which utilizes Dryad under the covers. Interestingly, Microsoft has also recently announced that its future “Big Data” strategy includes support for Hadoop[12].

REFERENCES

[1] YouTube—YouTube Statistics Feb24, 2015lhttps://www.youtube.com/yt/press/statistics.html

[2] Facebook, Facebook Statistics, Jan27, 2015, http://www.statisticbrain.com/facebook-statistics/.

[3] Twitter, Twitter statistics,2015, http://www.statisticbrain .com/twitter-statistics/.

[4] Linked In Statistics, Oct28, 2014 http://www.statisticbrain.com/linkedin-company-profile-and-statistics/

[5] Oracle—Information Management and Big Data: A Reference Architecture, www.oracle.com/.../ info-mgmt-big-data-r..., retrieved 20/03/14.

[6] MarvinJedyk, MAKING BIG DATA, SMALL, Using distributed systems for processing, analyzing and managing large huge data sets, Software Professional's Network, Cheshire Data systems Ltd.

[7] Brad Brown, Michael Chui, and James Manikin, Are you ready for the era of big data?,McKinsey Quarterly, McKinney Global Institute, October 2011.

[8] Ghemawat, H. Gobioff, and S. Leung, “The Google File System.” in ACM Symposium on Operating Systems Principles, Lake George, NY, Oct 2003, pp. 29 – 43.

[9] J. Dean and S. Ghemawat, “Map-Reduce: Simplified data processing on large clusters,” in USENIX Symposium on Operating Systems Design and Implementation, San Francisco, CA, Dec. 2004, pp. 137–150.

[10] Seattle, Washington, USA, Proceedings / Giuseppe Santucci and Matthew Ward (eds.). IEEE Conference on Visual Analytics Science & Technology, Oct14 - 19,2012, Piscataway, NJ : IEEE, 2012, S. 173-182. - ISBN 978-1-4673-4753-2.

[11] Mark Troester—Big Data Meets Big Data Analytics,www.sas.com/resources/.../ WR46345.pdf, retrieved 10/02/14.

[12] http://developer.yahoo.com/hadoop/tutorial/pigtutorial.html

[13] http://research.microsoft.com/enus/projects/dryad/

[14] www.ibm.com/software/data/infosphere/hadoop/jaql/

[15] Windows.Azure.Storage.http://www.microsoft.com/windowsazure/features/storage/

[16] http://23.66.85.199/collateral/analyst-reports/10334-ar-promise-peril-of-big-data.pdf

[17] http://en.wikipedia.org/wiki/Big_data

[18] www.gartner.com