ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 5, May 2016

Big Data: Features, Challenges & Solutions

Chaitanya Kadam, Yasoda Thapa

Abstract -- **Big data might be a broad term for data sets so big or advanced that ancient processing applications unit may be inadequate. Challenges include analysis, capture, data correction, search, sharing, storage, transfer, visualization, querying and data privacy. The term generally refers simply to the employment of adumbrative analytics or certain completely different advanced ways to extract value from data, and often to a specific size of data set. Accuracy in huge data might cause extra assured deciding, and higher choices may result in bigger operational efficiency, value reduction and reduced risk. Big data requires exceptional technologies like A/B testing, crowdsourcing, data fusion and integration, genetic algorithms, machine learning, linguistic communication process, simulation, time series analysis and visualization. To efficiently process large quantities of data within tolerable elapsed times.**

Keywords – Big data, analytics, visualization, crowdsourcing, testing, challenges.

I. INTRODUCTION

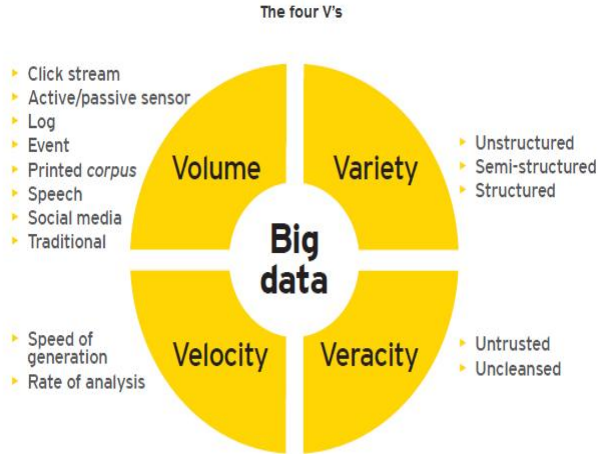
We are awash in a flood of data today. n a broad range of application areas, data is being collected at new scale. Decisions previously were based on guesswork, or on painstakingly constructed models of reality, can now be made based on the data itself. Currently drives nearly each side of our fashionable society, as well as mobile services, retail, producing, monetary services, life sciences, and physical sciences. There is broad recognition of the value of data, and products obtained through analyzing it. Scientific research has been revolutionized by Big Data. The field of Astronomy is being transformed from one where taking pictures of the sky was a large part of an astronomer's job to one where the pictures are all in a database already and the astronomer's task is to find interesting objects and phenomena in the database.

In 2010, enterprises and users stored more than 13 Exabyte of new data; this is over 50,000 times the data in the Library of Congress. The potential value of global personal location data is estimated to be $700 billion to end users, and it can result in an up to 50% decrease in product development and assembly costs, according to a recent McKinsey report [McK2011]. McKinsey predicts an equally great effect of Big Data in employment, where 140,000-190,000 workers with "deep analytical" experience will be needed in the US; furthermore, 1.5 million managers will need to become data-literate. The recent PCAST report on Networking and IT R&D [PCAST2010] known big data as a "research frontier" that may "accelerate progress across a broad vary of priorities." Even widespread news media now appreciates the worth of Big Data as evidenced by coverage in the Economist [Eco2011], the New York Times [NYT2012], and National Public Radio [NPR2011a, NPR2011b]. While the potential advantage of Big Data is real and important, and a few initial successes have already been achieved, there stay several technical challenges that must be addressed to completely notice this potential. The sheer size of the information, of course, could be a major challenge, and is that the one that's most simply recognized. However, there are others. Industry analysis companies prefer to point out that there are challenges not just in Volume, but also in Variety and Velocity.

II. BIG DATA FEATURES

Big data refers to the dynamic, giant and disparate volumes of data being created by individuals, tools and machines. It needs new, innovative and climbable technology to gather, host and analytically process the immense quantity of data gathered to derive real-time business insights that relate to customers, risk, profit, performance, productivity management and increased stockholder worth. Big data includes information garnered from social media, data from internet-enabled devices (including smartphones and tablets), machine data, video and voice recordings, and therefore the continuing preservation and logging of structured and unstructured data. it's usually characterized by the four "V's":

- *Volume:* the amount {of data| of knowledge| of information} being created is immense compared to ancient data sources
- *Variety:* data comes from completely different sources and is being created by machines also as individuals
- *Velocity:* data is being generated extraordinarily quick — a process that never stops, even while we sleep
- *Veracity:* big data is sourced from many various places, and as a result you wish to check the veracity/quality of the data.



The four V's of Big Data

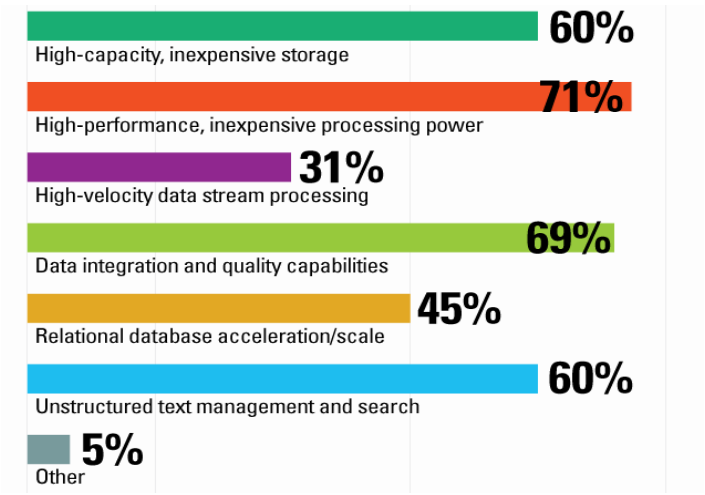


Fig: Some Other Features of Big Data

III. CHALLENGES IN BIG DATA ANALYSIS

A. *Heterogeneity and Incompleteness*

When humans consume info, an excellent deal of heterogeneousness is well tolerated. In fact, the import and richness of natural language can give valuable depth. take into account an electronic health record database design that has fields for birth date, occupation, and blood type for every patient. What can we do if one or more of those pieces of information isn't provided by a patient? Clearly, the health record continues to be placed within the database, however with the corresponding attribute values being set to NULL. Even after data cleanup and error correction, some incompleteness and a few errors in data are doubtless to stay. This incompleteness and these errors should be managed throughout data analysis. Doing this properly may be a challenge.

B. *Scale*

The word "big" is there within the very name. Managing massive and speedily increasing volumes of data has been a difficult issue for several decades. But, there's an elementary shift current now: data volume is scaling quicker than reckon resources, and central processing unit speeds area unit static. First, over the last 5 years the processor technology has created a dramatic shift -rather than processors doubling their clock cycle frequency each 18-24 months, now, thanks to power constraints, clock speeds have mostly stalled and processors area unit being engineered with increasing numbers of cores.

The second dramatic shift that's current is that the move towards cloud computing, that currently aggregates multiple disparate workloads with varied performance goals into terribly massive clusters. This level of sharing of resources on expensive and huge clusters needs new ways in which of deciding the way to run and execute processing jobs so we will meet the goals of every work cost-effectively, and to touch upon system failures, that occur a lot of ofttimes as we have a tendency to treat larger and bigger clusters.

C. *Timeliness*

The larger the info set to be processed, the longer it'll take to analyze. The design of a system that effectively deals with size is probably going additionally to end in a system that may process a given size of data set quicker. Given an oversized information set, it's typically necessary to search out components in it that meet a fixed criterion. Within the course of information analysis, this type of search is probably going to occur repeatedly. Scanning the whole information set to search out appropriate components is clearly impractical. Rather, index structures are created ahead to allow finding qualifying components quickly. The matter is that every index structure is intended to support just some categories of criteria.

D. *Privacy*

The privacy of data is another huge concern, and one that will increase within the context of big data. There is great public concern relating to the inappropriate use of private data, significantly through linking of data from multiple sources. Managing privacy is effectively both a technical and a sociological drawback, that should be addressed There are several extra challenging research issues. As an example, we have a tendency to don't understand nevertheless the way to share private data whereas limiting disclosure and ensuring enough data utility within the shared data.

E. *Human Collaboration*

Data analysis system should support input from multiple human consultants, and shared exploration of results. These multiple consultants could also be separated in space and time when it's too costly to assemble a whole team along in one room. The data system has to accept this distributed professional input, and support their collaboration. a well-liked new methodology of harnessing human ingenuity to unravel issues is thru crowd-sourcing. Wikipedia, the net reference book, is probably the most effective proverbial example of crowd-sourced knowledge.

IV. SOLUTIONS

Big data has nice potential to supply helpful data for firms which might profit the means they manage their issues. These large knowledge sets are overlarge and sophisticated for humans to effectively extract helpful data while not the help of computational tools. emerging technologies like the Hadoop framework and MapReduce supply new and exciting ways in which to process and remodel big data, outlined as complicated, unstructured, or massive amounts of data, into meaningful information.

A. Hadoop

Hadoop is a scalable, open source, fault tolerant Virtual Grid OS design for data storage and processing. It runs on artifact hardware; it uses HDFS that is fault-tolerant high information measure clustered storage design. It runs MapReduce for distributed processing and is works with structured and unstructured knowledge. For handling the velocity and heterogeneousness of data, tools like Hive, Pig and driver ar used that ar elements of Hadoop and HDFS framework. Hadoop and HDFS (Hadoop Distributed File System) by Apache is widely used for storing and managing big data.

Hadoop consists of distributed file system, data storage and analytics platforms and a layer that handles parallel computation, rate of flow (workflow) and configuration administration. HDFS runs across the nodes during a Hadoop cluster and along connects the file systems on several input and output data nodes to create them into one big file system. the current Hadoop system, consists of the Hadoop kernel, MapReduce, the Hadoop distributed file system (HDFS) and variety of connected parts like Apache Hive, HBase, Oozie, Pig and Zookeeper and these components are explained as below:

• HDFS: A extremely faults tolerant distributed file system that's accountable for storing data on the clusters.

• MapReduce: a strong parallel programming technique for distributed processing of large quantity of dataon clusters.

• HBase: A column oriented distributed NoSQL database for random read/write access.

• Pig: A high level data programing language for analyzing data of Hadoop computation.

• Hive: a data warehousing application that gives a SQL like access and relational model.

• Sqoop: A project for transferring/importing data between relational databases and Hadoop.

• Oozie: an orchestration and work flow management for dependent Hadoop jobs.

B. MapReduce

MapReduce is a programming model for processing massive data sets with a parallel, distributed algorithmic program on a cluster. Hadoop MapReduce is a programming model and software framework for writing applications that speedily method immense amounts of data in parallel on massive clusters of calculate nodes.

The MapReduce consists of 2 functions, map () and reduce (). mapper performs the tasks of filtering and sorting and reducer performs the tasks of summarizing the result. There could also be multiple reducers to parallelize the aggregations. Users can implement their own processing logic by specifying a customized map () and reduce () function. The map () function takes an input key/value pair and produces a list of intermediate key/value pairs. The MapReduce runtime system teams along all intermediate pairs based on the intermediate keys and passes them to reduce () function for producing the ultimate results. Map reduce is widely used for the Analysis of big data. Map reduce provides resolution to the mentioned problems since it supports distributed and parallel I/O scheduling. it's fault tolerant and supports scalability and its built-in processes for status and monitoring of heterogeneous and enormous datasets as in big data.

V. CONCLUSION

Through higher analysis of the massive volumes of data that are becoming available, there's the potential for creating quicker advances in several scientific disciplines and up the gain and success of the many enterprises. The challenges embrace not simply the plain problems with scale, however additionally heterogeneity, lack of structure, error-handling, privacy the least bit stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. Furthermore, these challenges would require transformative solutions, and cannot be addressed naturally by successive generation of industrial products.

VI. ACKNOWLEDGMENT

We thank our colleagues from IMCOST who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper. We thank Ms. Sheeba James and Mr. Nitin for assisting us by providing proper formats and all faculties for guiding us.

VII. REFERENCES

[1] McKinsey Global Institute (MGI), Big Data: The next frontier for innovation, competition, and productivity, Report, June, 2012.
[2] Managing Big Data. An interview with David Gorbet ODBMS Industry Watch, July 2, 2012.
[3] http://www.odbms.org/blog/2012/07/managing-big-data-an-interview-with-david-gorbet/
[4] On Big Data: Interview with Dr. Werner Vogels, CTO and VP of Amazon.com. ODBMS Industry Watch, November 2, 2011. http://www.odbms.org/blog/2011/11/on-big-data-interview-with-dr-werner-vogels-cto-and-vp-of-amazon-com/
[5] On Big Data: Interview with Shilpa Lawande, VP of Engineering at Vertica. ODBMs Industry Watch, November 16, 2011.
[6] http://dl.acm.org/citation.cfm?id=2367572
[7] Gantz, J. and E. Reinsel. 2011. "Extracting Value from Chaos", IDC's Digital Universe Study, sponsored by EMC
[8] JASON. 2008. "Data Analysis Challenges", The Mitre Corporation, McLean, VA, JSR-08-142
[9] The Economist. 2010. "Data, Data Everywhere", (online edition, February 28)

1. **Ms. Yasoda Thapa – Currently pursuing Master's in Computer Application (Third year) at ASM's Institute of Management & Computer Studies (IMCOST), Mumbai.**

2. **Mr. Chaitanya Kadam– Currently pursuing Master's in Computer Application (Third year) at ASM's Institute of Management & Computer Studies (IMCOST), Mumbai.**