ISSN: 2321-7782 (Online)
Volume 1, Issue 7, December 2013
International Journal of Advance Research in Computer Science and Management Studies
Research Paper
Available online at: www.ijarcsms.com

Big Data and HADOOP: A Big Game Changer

Umesh V. Nikam[1]
Information technology department
PRMIT & R,Badnera
Amravati - India

Anup W. Burange[2]
Information technology department
PRMIT & R,Badnera
Amravati - India

Abhishek A. Gulhane[3]
Information technology department
PRMIT & R,Badnera
Amravati - India

Abstract: ***The data management industry has matured over the last three decades, primarily based on Relational Data Base Management Systems (RDBMS) technology. Even today, RDBMS systems power a majority of backend systems for online digital media, financial systems, insurance, and healthcare, transportation, and telecommunications companies. Since the amount of data collected, and analyzed in enterprises has increased several-folds in volume, variety, and velocity of generation and consumption, organizations have started struggling with architectural limitations of traditional RDBMS architectures. As a result, a new class of systems had to be designed and implemented, giving rise to the new phenomenon of "Big Data".***

Keyword: bigdata, volume, variety, velocity, HDFS, Mapreduce, job scheduler, namenode.

I. INTRODUCTION

While there is no universally accepted definition of Big Data yet, and most of the attention in the press is devoted to the "Bigness" of Big Data, volume of data is only one factor in the requirements of modern data processing platforms. Industry analyst firm Gartner [1] defines Big Data as:

Big data is high-volume, high-velocity, and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision-making.Big data is basically vast amount of data which cannot be effectively processed, captured, and analyzed by traditional database and search tools in reasonable amount of time. Though the "big" in Big Data is subjective, it would be anywhere between few dozen terabytes to petabytes for most of the sectors. The Big Data information explosion is mainly due to the vast amounts of data generated by social media platform, data input from Omni-channels, various mobile devices, user generated data, multi-media data, and so on. Analysts term this as an expanding "Digital universe".

Big Data is usually defined by 3Vs: Volume, variety, and velocity. To put things in perspective let's examine each of these dimensions:

- Volume: IBM research finds that every day we add about 2.5 quintillion bytes (2.5 x 1018) of data; Facebook alone adds 500TB of data on daily basis; 90% of world's data is generated in last 2 years. Google processes about 1 petabyte of data every hour.
- Velocity: The rate of data growth is also astonishing. Gartner research finds that data is growing at 800% rate out of which 80% is unstructured. EMC research indicates that data increase is following Moore's law by doubling every 2 years.
- Variety: The data that is getting added is also of various types ranging from unstructured feeds, social media data,

multi-media data, sensor data etc.

Following diagram shows various technologies used in Big Data:

II. WHY BIG DATA?

Big Data is absolutely essential for the following intents:

- To spot business trends
- Determine quality of research
- To prevent diseases
- To link legal citation
- To combat crime
- To determine real time roadway communication system, where the data is created in the order of exa bytes (2^{18}).

III. WHY IT IS USED?

Areas or fields where big data are created:

- Medicine, Meteorology, Connectomics, Genomics, Complex Physics Simulation, Biological, Environment Research, and Areal Sensory System (remote sensing technologies).
- Big Science, RFID, Sensor Networks.
- Astrometry.net project keeps eye on Astrometry group via flicker for new photos of the night sky. It analyzes each image and identifies the celestial bodies such as stars, galaxies etc.

IV. DRIVERS AND OPPORTUNITIES

There is lot of drivers forcing the businesses to consider Big Data as their key business strategy. Some of them are listed below:

- Real-time prediction
- Increase operational and supply chain efficiencies
- Deep insights into customer behavior based on pattern and purchase analysis
- Information aggregation
- Better and more scientific customer segmentation for targeted marketing and product offering

Big data also provides the following opportunities:

- Improve productivity and innovation
- McKinsey predicts an increase in job opportunities ranging from 140K to 190K
- Uncover hidden patterns and rapidly respond to changing scenarios.
- Multi-channel and multi-dimensional information aggregation.

Traditional search, sort, and processing algorithms would not scale to handle the data in this range, and that too most of it being unstructured. Most of the Big Data processing technologies include machine learning algorithms, natural language processing algorithms, predictive modeling, and other artificial intelligence based techniques. Big Data is of strategic importance for many organizations. Because any new service or product will be eventually copied by competitors, but an organization can differentiate it by what it can do with the data it has.

Below diagram shows the convergence of data from various dimensions:

V. IMPACT AND APPLICATION

We will examine the impact and applications [3] of Big Data related technologies across various industry verticals and technology domains in this section.

Application across industry domains

Financial industry:

- Better financial data management
- Investment banking using aggregated information from various sources likes financial forecasting, asset pricing and portfolio management.
- More accurate pricing adjustments based on vast amount of real-time data
- Stock advises based on huge amount of stock data analysis, unstructured data like social media content etc.
- Credit worthiness analysis by analyzing huge amount of customer transaction data from various sources
- Pro-active fraudulent transaction analysis
- Regulation conformance

- Risk analytics
- Trading analytics

Retail industry:

- Better analysis of supply chain data and touch points across Omni-channel operations
- Customer segmentation based on previous transactions and profile information
- Analysis of purchase patterns and tailor made product offerings
- Unstructured data analysis from social media, multi-media to understand the tastes, preferences, and customer patterns and do sentiment analysis
- Targeted marketing based on user segmentation
- Competitor analysis

Mobility:

- Mining of customer location data, call patterns.
- Integrate with social media to provide location based services like sale offers, friend alerts, points-of-interest suggestions etc.
- Geo-location analysis :

Health care:

- Effective drug prescription [4] by analyzing all structured and unstructured medical history and records of the patient
- Avoid un-necessary prescriptions

Insurance:

- Risk analysis of customer
- Analyzing cross-sell and up-sell opportunities based on customer spending patterns
- Insurance portfolio optimization and pricing optimization

VI. WHAT IS HADOOP?

Apache Hadoop is an open-source software framework for storage and large scale processing of data-sets on clusters of commodity hardware. Hadoop is an Apache top-level project being built and used by a global community of contributors and users.[2] It is licensed under the Apache License 2.0.

The Apache Hadoop framework is composed of the following modules:

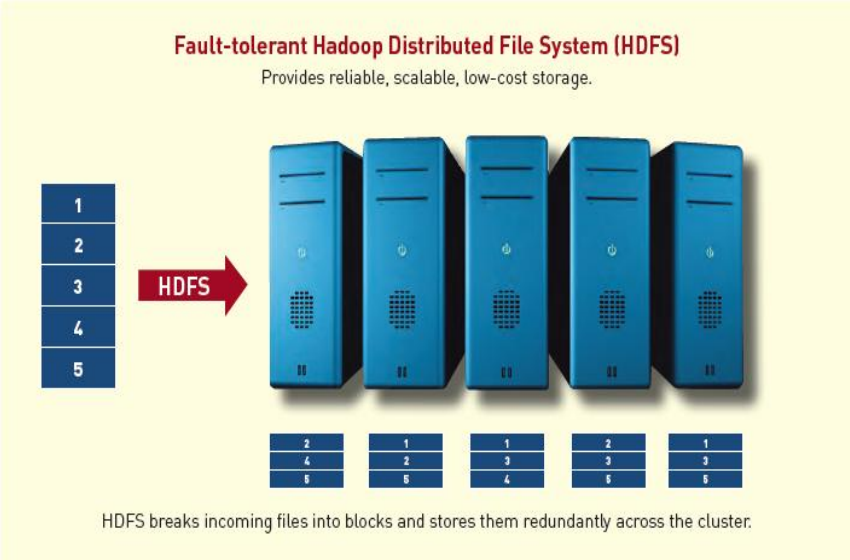
- **Hadoop Common** - contains libraries and utilities needed by other Hadoop modules
- **Hadoop Distributed File System** (HDFS) - a distributed file-system that stores data on the commodity machines, providing very high aggregate bandwidth across the cluster.
- **Hadoop YARN** - a resource-management platform responsible for managing compute resources in clusters and using them for scheduling of users' applications.
- **Hadoop MapReduce** - a programming model for large scale data processing.

All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines or

racks of machines) are common and thus should be automatically handled in software by the framework. Apache Hadoop's MapReduce and HDFS components originally derived respectively from Google's MapReduce and Google File System (GFS) papers.

1. HDFS

The file store is called the Hadoop Distributed File System, or HDFS. HDFS [2] provides scalable, fault-tolerant storage at low cost. The HDFS software detects and compensates for hardware issues, including disk problems and server failure.

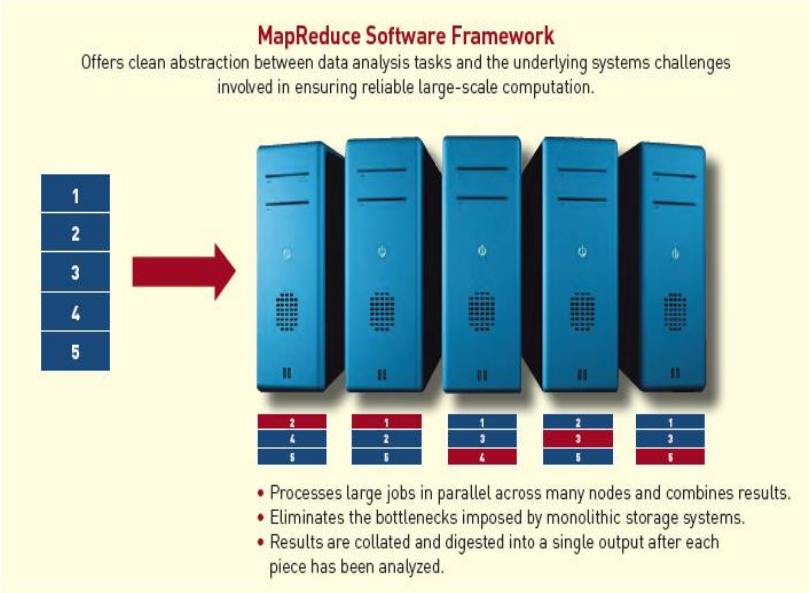


HDFS stores files across a collection of servers in a cluster. Files are decomposed into blocks, and each block is written to more than one (the number is configurable, but three is common) of the servers. This replication provides both fault-tolerance (loss of a single disk or server does not destroy a file) and performance (any given block can be read from one of several servers, improving system throughput).HDFS ensures data availability by continually monitoring the servers in a cluster and the blocks that they manage. Individual blocks include checksums. When a block is read, the checksum is verified, and if the block has been damaged it will be restored from one of its replicas. If a server or disk fails, all of the data it stored is replicated to some other node or nodes in the cluster, from the collection of replicas. As a result, HDFS runs very well on commodity hardware. It tolerates, and compensate for, failures in the cluster. As clusters get large, even very expensive fault-tolerant servers are likely to fail. Because HDFS expects failure, organizations can spend less on servers and let software compensate for hardware issues.

2. MapReduce

HDFS delivers inexpensive, reliable, and available file storage. That service alone, though, would not be enough to create the level of interest, or to drive the rate of adoption, that characterizes Hadoop over the past several years. The second major component of Hadoop is the parallel data processing system called MapReduce[5].Conceptually; MapReduce is simple.MapReduce includes a software component called the job scheduler. The job scheduler is responsible for choosing the servers that will run each user job, and for scheduling execution of multiple user jobs on a shared cluster. The job scheduler consults the NameNode for the location of all of the blocks that make up the file or files required by a job. Each of those servers is instructed to run the user's analysis code against its local block or blocks. The MapReduce processing infrastructure includes an abstraction called an input split that permits each block to be broken into individual records. There is special processing built in to reassemble records broken by block boundaries.The user code that implements a map job can be virtually anything. MapReduce allows developers to write and deploy code that runs directly on each DataNode server in the cluster. That code understands the format of the data stored in each block in the file, and can implement simple algorithms (count the number of occurrences of a single word, for example) or much more complex ones (e.g. natural language processing, pattern detection and

machine learning, feature extraction, or face recognition).At the end of the map phase of a job, results are collected and filtered by a reducer. MapReduce guarantees that data will be delivered to the reducer in sorted order, so output from all mappers is collected and passed through a shuffle and sort process. The sorted output is then passed to the reducer for processing. Results are typically written back to HDFS.Because of the replication built into HDFS, MapReduce is able to provide some other useful features. For example, if one of the servers involved in a MapReduce job is running slowly most of its peers have finished, but it is still working the job scheduler can launch another instance of that particular task on one of the other servers in the cluster that stores the file block in question. This means that overloaded or failing nodes in a cluster need not stop, or even dramatically slow down, a MapReduce job.



It is important to understand the properties that characterize good MapReduce jobs. First, algorithms must run well on the shared- nothing distributed infrastructure of Hadoop. If a processing job needs to communicate extensively among servers in the cluster, MapReduce is often a poor platform choice. At least in the map phase of any job, the best algorithms are able to examine single records, or a small number of adjacent records stored on the same DataNode, to compute a result. Some of the signature problems that characterize high- performance computing research over the last few decades weather prediction or modeling nuclear explosions, for example require extensive intra-node communication. These translate poorly to MapReduce.That said, MapReduce is a conceptually much simpler parallel processing system than older HPC systems. Existing algorithms, altered somewhat to work on data as stored on HDFS, often run very well. MapReduce is able to execute Java code, but also to use software written in other languages, including high-level languages like C or C++ and scripting languages like PHP, Python, and Perl. Speaking practically, a surprisingly large number of algorithms can be expressed as MapReduce jobs.MapReduce also excels at exhaustive processing. If an algorithm must examine every single record in a file in order to compute a result, then MapReduce is an excellent choice. Jobs run in parallel on DataNodes in a cluster. The number of DataNodes in a cluster is directly proportional to the amount of data the cluster can store. As a result, the size of a dataset affects performance of a MapReduce job much less than the complexity of the algorithm it implements.MapReduce is generally used for batch processing, not for the kind of interactive jobs that characterize many existing business intelligence applications. The key advantage of MapReduce is that it can process, literally, petabytes of data in reasonable time to answer a question. Users may have to wait minutes or even hours for a result, but can ask questions that were simply impossible to answer before MapReduce was available.

VII. WHY HADOOP?

The big web properties invented MapReduce, and built Hadoop, because they had a data problem that no existing commercial or research system could solve. The platform is now used to support an enormous variety of applications. These applications are not necessarily characterized by enormous datasets. Instead, they need the three key properties that Hadoop offers.

First, Hadoop is a single, consolidated storage platform for all kinds of data. Of course there are outstanding file and relational storage products available on the market today, and those systems will remain in use for years to come, to solve exactly the problems for which they were designed. Hadoop complements them by delivering a new repository where structured data and complex data may be combined easily.

Second, because of the economies of shared-nothing commodity servers and open source software, Hadoop provides vastly more storage at much lower cost than legacy systems. Fault-tolerant, reliable storage under HDFS costs just a few hundred dollars per terabyte today, and rides the inevitable downward price curve as server and storage vendors improve performance and reduce costs.

Third, and finally, MapReduce exploits the distributed storage architecture of HDFS to deliver scalable, reliable parallel processing services for arbitrary algorithms. Users are not limited to a small set of algorithms delivered by an RDBMS or other vendor. In Hadoop, the storage system is programmable. Users can analyze data using processors attached directly to the disks on which it resides. This combination the consolidation of all data types on a low-cost, reliable storage platform that delivers fast parallel execution of powerful analytical algorithms is new. Hadoop offers data-driven organizations ways to exploit data that they have simply never had before. MapReduce and Hadoop were created when Google and Yahoo! undertook solving an engineering problem core to their businesses: building an index of more than one billion web pages. The technology has proven invaluable there. But it has proven even more valuable in other, unexpected areas like improving page layout, advertisement selection, spell checking, map rendering, and so on. A general-purpose tool that permits analysis of all data enables new analyses that weren't previously practical or even imagined. Giving all developers easy, powerful access to all data creates a surprising multitude of business improvements.

VIII. CONCLUSION

More data usually beats better algorithm. But it is very rigid to store and analyze. However, Big Data are used for finding the customer behavior, for identifying the market trends, for increasing the innovations, for retaining the customers, for performing the operations efficiently. Flood of data coming from many sources must be handled using some non-traditional database tools. It provides more market value and systematic for the upcoming generation and Hadoop provides a new approach to handle big data.

References

1. Wikipedia, the free encyclopedia.
2. White, Tom, Hadoop: The Definitive Guide. O'Reilly Media, ISBN 978-1-4493- 3877-0.
3. An Oracle whitepaper, Jan 2012 “Oracle: Big Data for the enterprise”.
4. Scalable and Secure Sharing of Personal Health Records in Cloud Computing using Attribute-based Encryption, M. Li, S. Yu, K. Ren, and W. Lou, Sep 2010 pp 89- 106.
5. Dean, J. and Ghemawat, S., “MapReduce: Simplified Data Processing on Large Clusters.” Appeared in Proceedings of the Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December, 2004.

AUTHOR(S) PROFILE

Umesh V. Nikam, received the M.E degree in Information Technology from Prof.Ram Meghe institute of Technology & Research, Badnera-Amravati in 2011 and 2012, respectively. Currently working as an Assistant professor in Information Technology Department of Prof.Ram Meghe Institute of Technology & Research Badnera Amravati (Maharashtra).

Anup W. Burange, pursuing the M.E degree in Information Technology from Prof.Ram Meghe institute of Technology & Research, Badnera-Amravati. Currently working as an Pro-Term Lecturer in Information Technology Department of Prof.Ram Meghe Institute of Technology & Research Badnera Amravati (Maharashtra).

Abhishek A. Gulhane, received the M.E degree in Information Technology from Prof.Ram Meghe institute of Technology & Research, Badnera-Amravati in 2011 and 2012, respectively. Currently working as an Assistant professor in Information Technology Department of Prof.Ram Meghe Institute of Technology & Research Badnera Amravati (Maharashtra).