# Hadoop Performance based Big Data Analysis using Cloud Computing and Amazon Web Services

**G.Niveditha**
Assistant professor in Department of CSE
Geethanjali College of Engineering and Technology

**S.L.Anusha**
Assistant professor in Department of CSE
Geethanjali College of Engineering and Technology

***Abstract: The era of "Big Data" is upon us. From big consumer stores mining shopper data to Google using online search to predict incidence of the flu, companies and organizations are using troves of information to spot trends, combat crime, and prevent disease. Online and offline actions are being tracked, aggregated, and analyzed at dizzying rates. For example, questions like, how many calories we consumed for breakfast, how many we burned on our last run, and how long we spend using various applications on our computer, can be recorded and analyzed. We can lose weight by realizing we tend to splurge on Thursdays. We can be more efficient at work by realizing we spend time more than we thought on Facebook. Data warehousing and data mining are related terms, as is NoSQL. With data firmly in hand and with the ability given by Big Data Technologies to effectively store and analyze this data, we can find answers to these questions and work to optimize every aspect of our behavior. Amazon can know every book you ever bought or viewed by analyzing big data gathered over the years. The NSA (National Security Agency) can know every phone number you ever dialed. Facebook can and will analyze big data and tell you the birthdays of people that you did not know you knew. With the advent of many digital modalities all this data has grown to BIG data and is still on the rise. Ultimately Big Data technologies can exist to improve decision-making and to provide greater insights...faster when needed but with the downside of loss of data privacy.***

***Keywords: Big Data, ETL, Hadoop, Cloud, Web***

## 1. INTRODUCTION

Data has been a backbone of any enterprise and will do so moving forward. Storing, extracting and utilizing data has been key to many company's operations. In the past when there were no interconnected systems, data would stay and be consumed at one place. With the onset of Internet technology, ability and requirement to share and transform data has been a need. This marks invention of ETL. ETL facilitated transforming, reloading and reusing the data. Companies have had significant investment in ETL infrastructure, both data warehousing hardware and software, personnel and skills.

**BACKGROUND, MOTIVATION AND AIM:** With the advent of digital technology and smart devices, a large amount of digital data is being generated every day. Advances in digital sensors and communication technology have enormously added to this huge amount of data, capturing valuable information for enterprises, businesses. This Big data is hard to process using conventional technologies and calls for massive parallel processing. Technologies that are able to store and process exabytes, terabytes, petabytes of data without tremendously raising the data warehousing cost is a need of time. Ability to derive insights from this massive data has the potential to transform how we live, think and work. Benefits from Big data analysis range from healthcare domain to government to finance to marketing and many more [1]. Big data open source technologies have gained quite a bit of traction due to the demonstrated ability to parallelly process large amounts of data. Both parallel processing and technique of bringing computation to data has made it possible to process large datasets at high speed. These key features and ability to process vast data has been a great motivation to take a look into the architecture of the industry leading big data processing framework by Apache, Hadoop. Understand how this big data storage and analysis is achieved and experimenting with RDBMS vs Hadoop environment has proven to provide a great insight into much talked about technology.

Author of this thesis aims at understanding the dynamics involved in big data technologies mainly Hadoop, distributed data storage and analysis architecture of Hadoop, setup and explore Hadoop Cluster on Amazon Elastic Cloud. As well, conduct performance benchmarking on RDBMS and Hadoop cluster.

## 2. WHAT AND WHY BIG DATA

The amount of data generated every day in the world is exploding. The increasing volume of digital and social media and internet of things, is fueling it even further. The rate of data growth is astonishing and this data comes at a speed, with variety (not necessarily structured) and contains wealth of information that can be a key for gaining an edge in competing businesses. Ability to analyze this enormous amount of data is bringing a new era of productivity growth, innovation and consumer surplus. "Big data is the term for a collection of data sets so large and complex that it becomes difficult to process it using traditional database management tools or data processing applications. The challenges include the areas of capture, curation, storage, search, sharing, transfer, analysis, and visualization of this data" [2].

### 2.1 BIG DATA ATTRIBUTES

The three Vs - volume, velocity and variety - are commonly used to describe different aspects of big data. See Figure 1 [3]. These three attributes make it easy to define the nature of the data and the software platforms available to analyze [4].

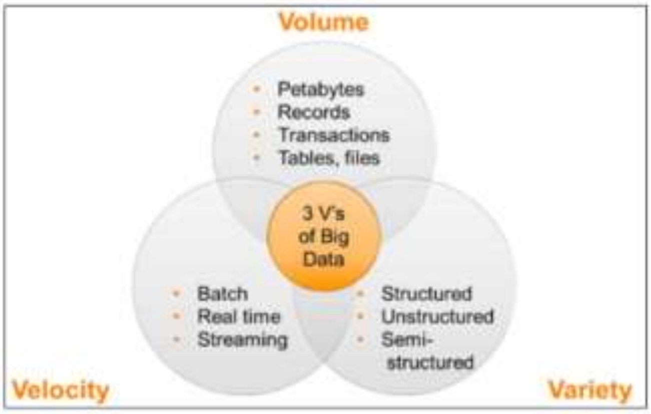

Figure 1. Three V's of big data. Source: VITRIA. The Operational Intelligence Company, 2014. http://blog.vitria.com, accessed April 2014.

***2.1.1 Volume***: Volume is the most challenging aspect of Big Data since it imposes a need for scalable storage and a distributed approach to querying. Big enterprises already have a large amount of data accumulated and archived over the years. It could be in the form of system logs, record keeping...etc. The amount of this data easily gets to the point where conventional database management systems may not be able to handle it. Data warehouse based solutions may not necessarily have the ability to process and analyze this data due to lack of parallel processing architecture. A lot can be derived from text data, locations or log files. For example, email communications patterns, consumer preferences and trends in transaction-based data, security investigations. Spatial and temporal (time-stamped) data absorb storage space quickly. Big Data technologies offer a solution to create value from this massive and previously unused/ difficult to process data.

***2.1.2 Velocity:*** Data is flowing into organizations at a large speed. Web and mobile technologies have enabled generating a data flow back to the providers. Online shopping has revolutionized consumer and provider interactions. Online retailers can now keep log of and have access to customers every interaction and can maintain the history and want to quickly utilize this information in recommending products and put the organization on a leading edge. Online marketing organizations are deriving lot of advantage with the ability to gain insights instantaneously. With the invention of the smart phone era there is even further location based data generated and its becoming important to be able to take advantage of this huge amount of data.

***2.1.3 Variety:*** All this data generated with social and digital media is rarely structured data. Unstructured text documents, video, audio data, images, financial transactions, interactions on social websites are examples of unstructured data. Conventional databases support 'large objects' (LOB's) but have their limitations if not distributed. This data is hard to fit in conventional neat relational database management structures and is not very integration friendly data and needs a lot of massaging before applications can manage it. And this leads to loss of information. If the data is lost then it's a loss that cannot be recovered. Big Data on the other hand tends to keep all the data since most of this is write once and read many times type of data. Big Data believes that there could be insights hidden in every bit of data.

## 3. HADOOP ARCHITECTURE

It's hard to omit Hadoop while talking about big data. Hadoop is the open source software platform managed by the Apache Software Foundation. It's the most widely recognized platform to efficiently and cost-effectively store and manage enormous amount of data.

### 3.1 INTRODUCTION TO HADOOP

Formal definition of Hadoop by Apache: "The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly available service on top of a cluster of computers, each of which may be prone to failures" [6]. Hadoop was initially inspired by papers published by Google, outlining its approach to handle an avalanche of data, and has since become the standard for storing, processing and analyzing hundreds of terabytes, and even petabytes of data. Hadoop framework development was started by Doug Cutting and the framework got its name from his son's elephant toy [7]. Hadoop has drawn the inspiration from Google's File System (GFS). Hadoop was spun from Nutch in 2006 to become a sub-project of Lucene and was renamed to Hadoop. Yahoo has been a key contributor to Hadoop evolution. By 2008 yahoo web search engine index was being generated by a 10,000 core Hadoop cluster. Hadoop is an open source framework by Apache, and has invented a new way of storing and processing data. Hadoop does not rely on expensive, high efficiency hardware. Instead it leverages on benefits from distributed parallel processing of huge amounts of data across commodity, low-cost servers. This infrastructure stores as well as processes the data, and can easily scale to changing needs. Hadoop is supposed to have limitless scale up ability and theoretically no data is too big to handle with distributed architecture [8]. Hadoop is designed to run on commodity hardware and can scale up or down without system interruption. It consists of three main functions: storage, processing and resource management. It is presently used by big corporations like Yahoo, eBay, LinkedIn and Facebook. Conventional data storage and analytics systems were not built keeping in mind the needs of big data. And hence no longer easily and cost-effectively support today's large data sets.

### 3.2 HADOOP ATTRIBUTES

- Fault tolerant - Fault tolerance is the ability of the system to stay functional without interruption and without losing data even if any of the system components fail [9]. One of the main goals of Hadoop is to be fault tolerant. Since hadoop cluster can use thousands of nodes running on commodity hardware, it becomes highly susceptible to failures. Hadoop achieves fault tolerance by data redundancy/ replication. And also provides ability to monitor running tasks and auto restart the task if it fails.

- Built in redundancy - Hadoop essentially duplicates data in blocks across data nodes. And for every block there is assured to be a back-up block of same data existing somewhere across the data nodes. Master node keeps track of these node and data mapping. And in case of any of the node fails, the other node where back-up data block resides, takes over making the infrastructure failsafe. A conventional RDBMS has the same concerns and uses terms like: persistence, backup and recovery. These concerns scale upwards with Big Data.

- Automatic scale up/ down - Hadoop heavily relies on distributed file system and hence it comes with a capability of easily adding or deleting the number of nodes needed in the cluster.

- Move computation to data - Any computational queries are performed where the data resides. This avoid overhead required to bring the data to the computational environment. Queries are computed parallelly and locally and combined to complete the result set.

## 4. CLOUD COMPUTING

Cloud computing is one of the most widely talked about term in the IT industry these days. Technological advances have enabled devising a solution quickly and at low cost. With cloud computing comes the ability to share computing and storage resources without having to develop infrastructure from scratch every time there is a need. The speed of delivering solution has gained significant importance in the competing technological world. And enterprises feel the pressure to quickly adapt to technologies. Cloud computing comes to an aid making infrastructures easily scalable and elastic. In simple words, cloud computing is nothing but making services or software or data available over the internet where services or software or data reside on remote machines. A client side interface is made available to access these cloud services. Cloud computing is an invention that has occurred from the need that IT organizations want to keep away from scaling infrastructure, investing money in hardware and training resources. Cloud computing usually has a front end and back end side. Front end is comprised of the client side interface and software required to access the cloud system. Whereas back end is the array of machines running different softwares depending upon the services provided by the cloud. When the time and pressure required in creating and maintaining IT infrastructures is taken off, organizations were able to focus on the core business. Also, it keeps organizations from putting valuable resources in reinventing the infrastructure every time there is a change and the organization needs to adapt to it. Infrastructure includes application servers, database servers, middleware needed to communicate between different entities. Cloud computing has allowed organizations to outsource overhead business applications like HR, payroll to third parties delivered via the cloud. See Figure [14]. Cloud computing is also meant to be flexible, elastic, reliable and secured. These are features that have led to heavy adoption of cloud computing by enterprises these days.

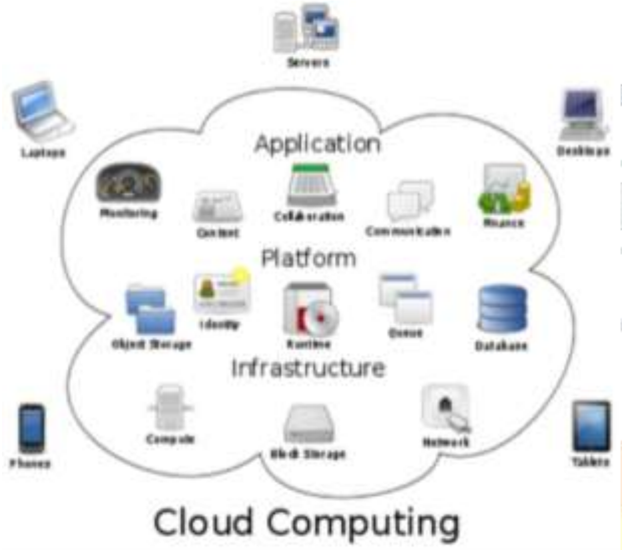

Figure 2. Cloud computing logical diagram. Source: S. Johnston. Seminar on Collaboration as a Service – Cloud Computing, 2012.

**Reasons to consider cloud computing**

- high availability
- clients are able to access system and data from anywhere
- Data is not confined to a single machine
- Reduces the need of expensive hardware on client side, and frequent hardware updates
- Keeps organizations from worrying about space and monitoring needs for data storage devices and servers
- Optimized use of servers - it can happen that server's capacity is not fully utilized by single client or single application. The server can be configured to act as multiple servers serving multiple requests. Challenges with cloud computing
- Security and privacy - sensitive data is to be stored on the machine outside of company's network
- Maintaining transparency in service delivery and billing
- Lack of standards - there is a need of technical, management and regulatory standards to provide definitions and guidelines to cloud environment.
- Responsiveness of the service
- Integration - integrating existing in-house applications to cloud services.

## 5. AMAZON WEB SERVICES

Amazon Web Services (abbreviated AWS) is a collection of remote computing services (also called web services) that is collectively known as a cloud-computing platform. Amazon provides these services over internet [15].

**5.1 WHAT IS AMAZON WEB SERVICES?** The first AWS was publicly made available in 2006 to provide online services. AWS is geographically distributed into different regions. See Figure 3. These regions have central hubs in the Eastern USA, Western USA (two locations), Brazil, Ireland, Singapore, Japan, and Australia. Each region comprises multiple smaller geographic areas called availability zones. Geographically distributing AWS is meant to avoid outages, provide back up and redundancy of the services and make them highly available.

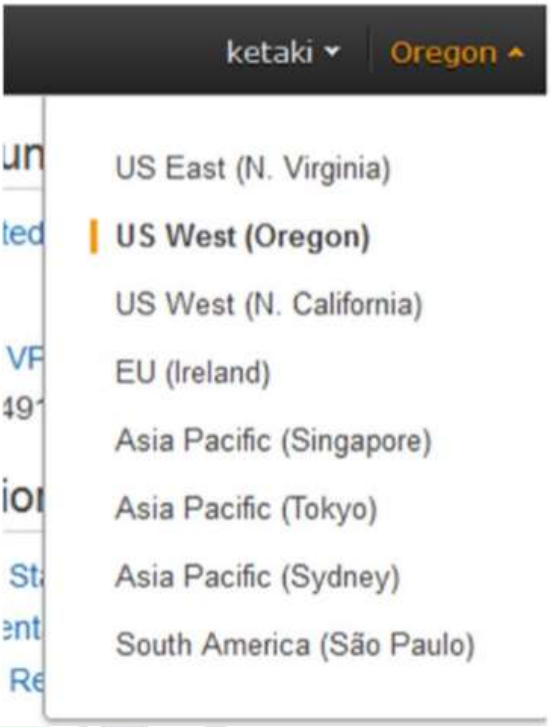

Figure 3. Screenshot of AWS regions.

Amazon Web Services (AWS) provides computing resources and services

(see Figure 3) that are very easy to access, within minutes, and meant to be cost effective at pay-as-you-go pricing. For example, for this thesis purpose, Linux machine nodes were used with a rate of $0.06 per hour and amazon has charged fees based on hours of utilization. The only cost involved is the cost when the nodes are run and are in use, without any up-front purchase costs or ongoing maintenance costs. AWS also provide easy scalability which is hard to achieve with physical in house servers. It is easy to scale up the number of nodes while experimenting on AWS.

**Benefits of using AWS that Amazon claims**

- Low cost
- No initial purchase cost
- Flexibility
- Scalability
- High availability
- Security

## 6. CONCLUSION

Big data has become highly prevalent in organization's day-to-day activities. Amount of big data and rate at which it's growing is enormous. And big data technology is sure to soon knock on the door of every enterprise, organization, and domain. RDBMS, even with multiple partitioning and parallelizing abilities fails to easily and cost-effectively scale to growing data needs. At the same time it expects data to be structured and is not so capable of storing and analyzing raw unstructured data which is common to encounter with the advent of wearable technologies, smartphones, and social networking websites. Hadoop is the most widely accepted and used open source framework to compute big data analytics in an easily scalable environment. It's a fault tolerant, reliable, highly scalable, cost-effective solution that's supports distributed parallel cluster computing on thousands of nodes and can handle petabytes of data. Two main components HDFS and MapReduce contribute to the success of Hadoop. It very well handles storing and analyzing unstructured data. Hadoop is a tried and tested solution in the production environment and well adopted by industry leading organizations like Google, Yahoo, and Facebook. Though previous versions of Hadoop did not have real time data analytics component, Apache has recently introduced Spark as a solution to real time big data analytics. Spark relies on Resilient Distributed Data and is said to provide results in split of a second. Many domains, like finance, social networking, healthcare, security, log mining are adopting big data technologies with a promise to gain insights from the ability to easily mine large amounts of data. As a future work it will be interesting to setup real time big data analytics engine and see how differently it handles data that Hadoop MapReduce, benchmark its performance against distributed batch processing architecture and understand how it helps overcome the challenges in batch processing big data analytics system.

REFERENCES

**[1].** V. Mayer-Schoönberger and K. Cukier. Big data – a revolution that will transform how we live, work, and think. Eamon Dolan/Houghton Mifflin Harcourt, Chicago, Illinois 2013.

**[2].** Wikipedia. Big data, 2014. http://en.wikipedia.org/wiki/Big_data, accessed April 2014.

**[3].** VITRIA. The Operational Intelligence Company, 2014. http://blog.vitria.com, accessed April 2014.

**[4].** E. Dumbill. What is Big Data? An Introduction to the Big Data Landscape, 2012. http://strata.oreilly.com/2012/01/what-is-big-data.html, accessed April 2014.

**[5].** M. Stonebraker, P. Brown, and D. Moore. Object-relational DBMSs, tracking the next great wave. Morgan Kauffman Publishers, Inc., San Francisco, California, 2 edition, 1998.

**[6].** Apache Hadoop. What Is Apache Hadoop?, 2014. http://hadoop.apache.org/, accessed April 2014.

**[7].** Wikipedia. Apache Hadoop, 2014. http://en.wikipedia.org/wiki/Apache_Hadoop, accessed April 2014.

**[8].** T. White. Hadoop – the definitive guide. O'Reilly Media, Inc., Sebastopol, California, 1 edition, 2009.

**[9].** V. S. Patil and P. D. Soni. Hadoop Skeleton and Fault Tolerance in Hadoop Clusters, 2011. http://salsahpc.indiana.edu/b534projects/sites/default/files/public/0_Fault%20Tolerance %20in%20Hadoop%20for%20Work%20Migration_Evans,%20Jared%20Matthew.pdf, accessed April 2014.

**[10].** Apache Hadoop. MapReduce Tutorial, 2013. https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html, accessed April 2014.

**[11].** Apache Hadoop. HDFS Architecture Guide, 2013. http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html, accessed April 2014.

**[12].** K. Kline, D. Kline, and B. Hunt. SQL in a nutshell, a desktop quick reference. O'Reilly Media, Sebastopol, California, 3 Edition, 2008.

**[13].** P. J. Sadalage, and M. Fowler. NoSQL distilled, a brief guide to the emerging world of polygot persistence. Addison-Wesley, Reading, Massachusetts, 3 edition, 2013.

**[14].** S. Johnston. Seminar on Collaboration as a Service – Cloud Computing, 2012. http://www.psirc.sg/events/seminar-on-collaboration-as-a-service-cloud-computing, accessed April 2014.

**[15].** Wikipedia, Amazon Web Services, 2014. en.wikipedia.org/wiki/Amazon_Web_Services, accessed April 2014.

**[16].** Amazon Elastic MapReduce. What is Amazon EMR?, 2009. http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-what-isemr.html accessed April 2014.

**[17].** Amazon Elastic Compute Cloud. Instances and AMIs, 2014. http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-instances-and-amis.html, accessed April 2014.

**[18].** B. Stinson. PostgreSQL – essential reference. Sams Publishing, Indianapolis, Indiana, 2001.

**[19].** GroupLens. MovieLens, 2014. http://grouplens.org/datasets/movielens/, accessed April 2014.

**[20].** R. Davies. Summary, n.d. http://files.grouplens.org/datasets/movielens/ml-10mREADME.html, accessed April 2014.

**[21].** E. Capriolo, D. Wampler, and J. Rutherglen. Programming hive. O'Reilly Media, Inc., Sebastopol, California, 2012.

**[22].** M. Barlow. Real-time big data analytics, emerging architecture. O'Reilly Media, Inc., Sebastopol, California, 2013.

**[23].** M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. Franklin, S. Shenker, and I. Stoica. Resilient Distributed Datasets, A Fault Tolerant Abstraction for In-Memory Cluster Computing, n.d. http://www.cs.berkeley.edu/~matei/talks/2012/nsdi_rdds.pdf, accessed April 2014.

**About the authors:**

**G.Niveditha** working as an Assistant professor in department of CSE at Geethanjali college of Engineering and Technology affiliated to JNTU Hyderabad. She has 3+ years teaching experience. Her Research interests includes Internet of Things (IoT) and Big Data Analytics

**S.L.Anusha** working as an Assistant professor in department of CSE at Geethanjali college of Engineering and Technology affiliated to JNTU Hyderabad. She has 2+ years teaching experience. Her Research interests include Big Data Analytics and Artificial Intelligence.