**Big Data Analysis for Data Management with AWS and Hadoop**

**A Seminar**

**Submitted in Partial Fulfillment**

**of theRequirements**

**for the award of the Degree of**

**Master of Technology**

In

**Computer Science & Engineering**

By

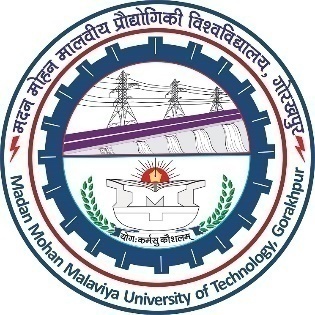
Shivani Gupta

Roll No.- 2019023117

**under the supervision of**

Mr. Rohit Kumar Tiwari

**(Assistant Professor)**

****

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MADAN MOHAN MALAVIYA UNIVERSITY OF TECHNOLOGY

**GORAKHPUR, UTTAR PRADESH**

**June, 2021**

**© M. M. M. UNIVERSITY OF TECHNOLOGY, GORAKHPUR, (U.P.) – 273010, INDIA**

**ALL RIGHTS RESERVED**

# Certificate

Certified that Shivani Gupta has carried out the research work presented in this seminar entitled “Big Data Analysis for Data Management with AWS and Hadoop” for the award of Master of Technology from Madan Mohan Malaviya University of Technology, Gorakhpur under my supervision. The report embodies result of original work and studies carried out by the student himself and the contents of the report do not form the basis for the award of any other degree to the candidate or to anybody else.

**Mr. Rohit Kumar Tiwari**

**(Assistant Professor)**

Department of Computer Science & Engineering

M. M. M. University of Technology, Gorakhpur

# Approval Sheet

This seminar entitled “Big Data Analysis for Data Management with AWS and Hadoop” by Shivani Gupta is approved for the degree of Master of Technology in Computer Science & Engineering.

**Examiner**

**Supervisor**

Mr. Rohit Kumar Tiwari

**Head of Department**

Prof. P.K. Singh

**Dean, Research & Development**

Prof.

Date:

Place: Gorakhpur

# Acknowledgement

The satisfaction that acc­­ompanies that the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible.

I express my deep sense of gratitude and indebtednessto **Mr. Rohit Kumar Tiwari**, my supervisor for his valuable advice, constant encouragement and constructive criticism during the study and during the preparation of this report.

I wish to thank my parents who have been always a source of inspiration for their never-ending support and love throughout completion of the report and I am also thankful to my friends who have helped in the successful completion of the report.

**Shivani Gupta**

# Table of Contents

[Certificate iii](#_Toc75806118)

[Approval Sheet iv](#_Toc75806119)

[Acknowledgement v](#_Toc75806120)

[Table of Contents vi](#_Toc75806121)

[List of Abbreviations vii](#_Toc75806122)

[List of Figures ix](#_Toc75806123)

[List of Tables x](#_Toc75806124)

[Abstract xi](#_Toc75806125)

[Chapter 1 1](#_Toc75806126)

[Introduction and Background 1](#_Toc75806127)

[1.1 Big Data: 1](#_Toc75806128)

[1.2 How ‘BIG’ is Big Data?: 1](#_Toc75806129)

[1.3 Amazon Web Services (AWS): 2](#_Toc75806130)

[1.3.1 Role of Cloud Computing (CC): 2](#_Toc75806131)

[1.3.2 CC comes in a picture: 3](#_Toc75806132)

[1.3.3 AWS role in Big Data Management: 3](#_Toc75806133)

[1.4 Problem Description: 4](#_Toc75806134)

[1.5 Motivation: 4](#_Toc75806135)

[CHAPTER 2 5](#_Toc75806136)

[Literature Survey 5](#_Toc75806137)

[CHAPTER 3 9](#_Toc75806138)

[Tools and Techniques 9](#_Toc75806139)

[3.1 Details of Big Data: 9](#_Toc75806140)

[3.2.2 Advantages of Big Data: 11](#_Toc75806141)

[3.3.3 Challenges associated with Big Data: 12](#_Toc75806142)

[3.2 Hadoop Framework: 12](#_Toc75806143)

[3.2.1 Hadoop Design: 13](#_Toc75806144)

[3.2.2 Architecture of Hadoop: 13](#_Toc75806145)

[3.3 Brief about HDFS: 14](#_Toc75806146)

[3.3.1 HDFS Architecture and its components: 15](#_Toc75806147)

[3.3.2 Features of HDFS: 16](#_Toc75806148)

[3.4 Running Hadoop on AWS: 17](#_Toc75806149)

[Chapter 4 18](#_Toc75806150)

[Experimental Work and Result Analysis 18](#_Toc75806151)

[5.1 System Tools and Language Used 18](#_Toc75806152)

[5.2 Launching Instances through AWS: 18](#_Toc75806153)

[5.3 Launch the Terminal: 20](#_Toc75806154)

[5.4 Implementation Work: 20](#_Toc75806155)

[Chapter 5 25](#_Toc75806156)

[Conclusion and Future Scope 25](#_Toc75806157)

[5.1 Conclusion 25](#_Toc75806158)

[5.2 Future Scope 25](#_Toc75806159)

[Bibliography 26](#_Toc75806160)

[Curriculum Vitae 28](#_Toc75806161)

# List of Abbreviations

|  |  |
| --- | --- |
| BD | Big Data |
| AWS | Amazon Web Service |
| EC2 | Elastic Compute Cloud |
| EBS | Elastic Block Store |
| HDFS | Hadoop Distributed File System |
| NN | Name Node |
| DN | Data Node |
| CN | Client Node |
| AMI | Amazon Machine Image |
| EMR | Elastic Map Reduce |
| S3 | Simple Storage Service |
| CC | Cloud Computing |
| RAM | Random Access Memory |
| GB | Giga Bytes |
| NYT | New York Times |
| HM | Human Mobility |
| FT | Fault Tolerance |
| IDE | Integrated Development Environment |

# List of Figures

Figure- 3.1 Architecture of Hadoop Ecosystem [10].

Figure- 3.2 HDFS Architecture [11].

Figure- 5.1 Details of Instance launched.

Figure- 5.2 Start and Stop procedure for launched instances.

Figure- 5.3 Linux terminal launch through AWS.

Figure- 5.4 All instances (NN, DN’s and CN).

Figure- 5.5 Master Node (NN) setup.

Figure- 5.6 DN’s setup.

Figure- 5.7 Fetching the Data (which is a long text of ‘i’).

Figure- 5.8 Tcpdump command result.

Figure- 5.9 Outcome of Creating and Attaching the volume.

Input Data

Conv1+ReLu+Dropout

Conv2+ReLu+Dropout

Conv3+ReLu+Dropout

Conv4+ReLu+Dropout

FC1

Output

# List of Tables

Table- 2.1 Surveys based on Big Data with Hadoop.

Table 5.1: Tools and System Configuration

# Abstract

Information Technology in this 21st century is reaching the sky with large-scale of data to be processed and studied to make sense of data where the traditional approach is no more efficient. Big data is a dataset displaying features of volume, velocity and variety in an OR relationship. It has no significance if it can not be exposed to strategic analysis and utilization. Big data is a problem for many of the industries because handling that much amount of data on a per-day basis become a big task for the industries to do it. Here, we are going to show you that how they are managing that huge data and storing that. There are many tools available nowadays to handle it rather than using some physical devices so, we can use this to handle with it. In this report, we have used AWS services to install the instances as per your OS (operating system) demands and after that we have used Hadoop Distributed File system (HDFS), Apache Spark and Map Reduce for doing these things without using some costly devices.

**Keywords:-** Big data, AWS, HDFS, Apache Spark, Map Reduce.

# 

# Chapter 1

# Introduction and Background

## 1.1 Big Data:

Big Data is a field that treats ways to analyse, systematically extract informations from, or otherwise deal with data sets that are too large or complex to be dealth with by traditional data processing application software.

In this technological era of large scale data. Businesses need to rethink on the modern approaches to better understand the customers to gain a competetive edge in the market. The amount of data in our world has been exploding. Companies capture ‘trillions of bytes’ of information about their customers, suppliers and operations like post a blog, written document, send messages and receiving that also consumes spaces because companies store all the information about their customers.. Millions of networked sensors are being embedded in the physical world in devices such as mobile phones and automobiles, sensing, creating, and communicating data Multimedia and individuals with smartphones and on social network sites will continue to fuel exponential growth. Big data is large pools of data that can be captured, communicated; aggregated, stored, and analyzed is now part of every sector and function of the global economy. Like other essential factors of production such as hard assets and human capital, it is increasingly the case that much of modern economic activity, innovation, and growth simply couldn’t take place without data.

## 1.2 How ‘BIG’ is Big Data?:

“Big data” refers to datasets whose size is beyond the ability of typical databasesoftware tools to capture, store, manage, and analyze. This definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data—i.e., we don’t define big data in terms of being larger than a certain number of terabytes (thousands of gigabytes). We assume that, as technology advances overtime, the size of datasets that qualify as big data will also increase. Also note that the definition can vary by sector, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry. With those caveats, big data in many sectors today will range from a few dozen terabytes to multiple petabytes(thousands of terabytes). The ability to store, aggregate, and combine data and then use the results to perform deep analyses has become ever more accessible as trends such as Moore’s Law in computing, its equivalent in digital storage, and cloud computing continue to lower costs and other technology barriers.

The means to extract insight from data are also markedly improving as software available to apply increasingly sophisticated techniques combines with growing computing horse power. Further, the ability to generate, communicate, share, and access data has been revolutionized by the increasing number of people, devices, and sensors that are now connected by digital networks. In 2010, more than 4 billion people, or 60 percent of the world’s population, were using mobile phones, and about 12 percent of those people had smartphones, whose penetration is growing at more than 20 percent a year. More than 30 million networked sensor nodes are now present in the transportation, automotive, industrial, utilities, and retail sectors.

The number of these sensors is increasing at a rate of more than 30 percent a year. There are many ways that big data can be used to create value across sectors of the global economy. Indeed, world on the cusp of a tremendous wave of innovation, productivity and growth, as well as new modes of competition and value capture all driven by big data as consumers, companies, and economic sectors exploit its potential.

## 1.3 Amazon Web Services (AWS):

AWS is basically a kind of cloud service provider. The practice of using a network of remote servers hpsted on the Internet to store, manage, and process data, rather than a local server or a personal computer is known as Cloud Computing.

### 1.3.1 Role of Cloud Computing (CC):

User always run the program on the top of OS (operating system) which is running on the top of hardware contains compute and storage unit that is RAM, CPU and Hard-Disk.

Suppose we are running with the website as program for ecommerce business and we have a server with ,minimum hardware that is 4 core CPU and 4GB RAM.

If our website goes viral, millions of clients are hitting to the site but our server dont have so much capability to maintain the traffic because of less hardware, in this case our site goes down.

### 1.3.2 CC comes in a picture:

We dont want to invest money in purchasing the upfront i.e., hardware by our own, so we bought it from the cloud resources as a per usage basis, we can pay for the upfront as pay as we go model provided by cloud, we can ask cloud do you have any service that provide us Hardware, cloud say yes we have a compute service that provide RAM and CPU and storage also that provide Hard-Disk (storage unit).

CC is like a service provider that provide services or resources to the customers and they charge us ‘as pay as we go model’.

### 1.3.3 AWS role in Big Data Management:

* AWS provides a broad and fully integrated prtfolio of the cloud computing services to help you to build, secure and deploy you big data applications with AWS, there is no hardware to produce, and no infrastructure to maintain and scale, so you can only focus on your resources on uncovering new insights with new capabilities and features added constantly. You will always be able to leverage the latest technologies without making lon-term investment commitments.
* Most Big Data technologies require large clusters of servers resulting in long provisioning and setup cycles. With AWS, you can deploy the infrastructure you need almost instantly. This means your team can be more productive, it is easier to try new things, and projects can roll out sooner.
* Big Data workloads are as the data assets they intend to analyze. A broad and deep platform means you can build virtually any ‘Big Data’ applications and support any workload reagrdless of ‘Volume’, ‘Velocity’, and ‘Variety’ of data. More than 50 services and hundreds of features added every year, AWS provides everything you need to collect, store, process, analyze, and visulaize Big Data on the cloud.
* ‘Big Data’ is sensitive data. Therefore, securing your data assets and protecting your infrastructure without losing agility is critical. AWS provides capabilities across facilities, network, software, and business processes to meet the strictest requirements. Environments are continuously audited.

## 1.4 Problem Description:

My problem statement is to find out how AWS has become this much effective for managing the large amount of data, for example:- Facebook generates more than ‘4 petabytes’ in a day so, the question how facebook deals with that huge volume of data per day.

Netflix generates 494 million GB bandwidth approximately on a per day basis. We have seen several techniques of Hadoop for solving that above without paying more cost. Here, HDFS (Hadoop Distributed File System) have used for creating ‘Master- Slave’ architecture to make a cluster of huge storage and that can store that huge amount of data. We have AWS here, because we have created or launched multiple instances through that and EC2 instance we have used for launching then instance and S3 for storage bucket. We have showed in this report that how the packets and storage are contributed to the ‘Master node’ by ‘Slave node’.

## 1.5 Motivation:

Many Organizations can use Big Data analytics systems and software to make data-driven decisions that can improve business-related outcomes. The benefits may include more effective marketing, new revenue opportunities, customer personalization and improved operational efficiency. With an effective strategy, these benefits can provide competetive advantages over rivals. Instead of using one large computer to process and store the data, Hadoop allows clustering commodity hardware together to analyze massive data sets in parallel.

Amazon EMR makes it easy to create and manage fully configured, elastic clusters of Amazon EC2 instances running Hadoop and other applications in the Hadoop ecosystems.

# CHAPTER 2

# Literature Survey

In 2004, Google published the paper that introduced MapReduce to the world. Earlyin 2005, the Nutch developers had a working MapReduce implementation in Nutch,and by the middle of that year all the major Nutch algorithms had been ported to runusing MapReduce and NDFS.

NDFS and the MapReduce implementation in Nutch were applicable beyond the realmof search, and in February 2006 they moved out of Nutch to form an independentsubproject of Lucene called Hadoop. At around the same time, Doug Cutting joinedYahoo!, which provided a dedicated team and the resources to turn Hadoop into asystem that ran at web scale (see sidebar). This was demonstrated in February 2008 when Yahoo! announced that its production search index was being generated by a10,000-core Hadoop cluster.

In January 2008, Hadoop was made its own top-level project at Apache, confirming itssuccess and its diverse, active community. By this time, Hadoop was being used by manyother companies besides Yahoo!, such as Last.fm, Facebook, and the *New York Times (NYT)*. Someapplications are covered in the case studies in Chapter 16 and on the Hadoop wiki.In one well-publicized feat, the NYTused Amazon’s EC2 compute cloudto crunch through four terabytes of scanned archives from the paper converting them to PDFs for the Web. The processing took less than 24 hours to run using 100 machines, and theproject probably wouldn’t have been embarked on without the combination of Amazon’s pay by-the-hour model (which allowed the NYT to access a large number of machines for a shortperiod) and Hadoop’s easy-to-use parallel programming model.

In April 2008, Hadoop broke a world record to become the fastest system to sort a terabyte of data. Running on a 910-node cluster, Hadoop sorted one terabyte in 209 seconds (just under 3½ minutes), beating the previous year’s winner of 297 secondsIn November of the same year, Google reported that its MapReduce implementation sorted one terabyte in 68 seconds. As the first edition of this book was going to press (May2009), it was announced that a team at Yahoo! Used Hadoop to sort one terabyte in 62 seconds.

In 2011, Gartner an information technology research and advisory company defined big data as growth challenges focus on information volume, variety and velocity [2].

In [5], Volume is defined as increasing the amount of data. Velocity is defined as the speed of data at which data flow in or out of the system. Variety is defined as various data types and sources. There are many frameworks to perform operation on big data. One of them is Apache Hadoop framework.

In [1],The grade estimation system is built on MapReduce Architecture and Hadoop based framework. The proposed architecture grading of student can be used to make predication. It can be used for analysis of various attributes of the Hadoop framework over the cloud environment. The paper has clearly deliberated the data distribution and the respective key-value pairs at each stage of Hadoop architecture.

The most popular description of big data thus far is the ‘‘3V” model, where ‘‘3V” refers to volume, variety, and velocity [7]. Volume literally means that typical big data have particularly large data volume. For example, mobile phone call records usually have 70 million data entries [8], video surveillance records can even have larger data volume in terms of data storage. Variety means that big data have diversified data sources, data structures, and potential applications. Velocity refers to the real time or quasi-real-time data updating. For instance, air quality monitoring data are often updated once or several times each day. In addition to the ‘‘3V” model, ‘‘4V” and ‘‘5V” models are emerging as researchers attempt to redefine big data. IBM promotes the conformance to veracity to explain the bias problems brought by big data and believes that the ‘‘4V” model can accurately describe big data (IBM, 2013). Several media columns argue that big data also have the features of value, variability, and visualization [9].

Big data such as mobile phone data and smart card data have incompleteness problems and other big data such as social media microblogs have biased samples. Traditional data can be used to complement such big data. Using a scientific traditional sampling survey, we can collect more detailed information of the target population including as their socio-economic characteristics, and make the collected data more representative. Many scholars are already using this approach, such as combining smart card data with travel diaries [5] and combining news report and crime incident data with social media data [6].

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Entry** | **Study** | **Big data** | **Method** | **Result** |
| Gonzalez et al. (2008) | Human Mobility (HM) | Mobile phone data | Statistical fitting | Human trajectories show a high degree of temporal and spatial regularity. |
| Roth et al. (2011) | HM | London subway ‘‘Oyster” card data | Null model | A polycentric structure is composed of large flows organized around a limited number of activity centers. |
| Krings et al. (2009) | Spatial Interaction | Mobile phone data | Gravity model | Communication intensity between two cities is proportional to the product of their sizes divided by the square of their distance. |
| Zheng et al. (2013) | Urban computing | Air monitoring data and points of interest. | A semi-supervised learning approach based on the artificial neural network and conditional random field. | The proposed method has advantage over decision tree, conditional random field, and artificial neural network. |
| Fu and Chau (2013) | Data quality | Social media (Sina Weibo) | Random sampling approach | Representative and reliable statistics on Chinese microbloggers are limited. |

**Table- 2.1 Surveys based on Big Data with Hadoop.**

# CHAPTER 3

# **Tools and Techniques**

## 3.1 Details of Big Data:

Big data typically refers to the following types of data:

1.Traditional enterprise data includes customer information from CRM systems,transactional ERP data, web store transactions, general ledger data.

2. Machine-generated /sensor data includes Call Detail Records (“CDR”), weblogs,smart meters, manufacturing sensors, equipment logs (often referred to as digitalexhaust), trading systems data.

3. Social data includes customer feedback streams, micro-blogging sites like Twitter,social media platforms like Facebook .

**3.1.1 V’s of Big Data:**

**1. Volume:** When discussing Big Data volumes, almost unimaginable sizes and unfamiliar numerical terms are required:

* Each day, the world produces 2.5 quintillion bytes of data. That is 2.3 trillion gigabytes.
* By 2020, we will have created 40 zettabytes of data, which is 43 trillion gigabytes.
* Most companies already have, on average, 100 terabytes of data stored each.
* Facebook users upload that many data daily.
* Walmart alone processes over a million transactions per hour.

**2. Velocity:** Underlying the volume numbers is an even larger trend, which is that 90 percent of extant data have been created in just the last two years. The speed at which data are generated, accumulated and analyzed is on a steep acceleration curve. As of next year, there will be 19 billion network connections globally feeding this velocity.

Although most data are warehoused before analysis, there is an increasing need for real-time processing of these enormous volumes, such as the 200 million emails, 300,000 tweets and 100 hours of Youtube videos that are passing by every minute of the day. Real-time processing reduces storage requirements while providing more responsive, accurate and profitable responses.

**3. Variety:** Another challenge of Big Data processing goes beyond the massive volumes and increasing velocities of data but also in manipulating the enormous variety of these data. Taken as a whole, these data appear as an indecipherable mass without structure. Consisting of natural language, hashtags, geo-spatial data, multimedia, sensor events and so much more, the extraction of meaning from such diversity requires ever-increasing algorithmic and computational power.

**4.** Variability: Furthermore, the intrinsic meanings and interpretations of these conglomerations of raw data depends on its context. This is especially true with natural language processing. A single word may have multiple meanings. New meanings are created and old meanings discarded over time. Interpreting connotations is, for instance, essential to gauging and responding to social media buzz. The boundless variability of Big Data therefore presents a unique decoding challenge if one is to take advantage of its full value.

**5.** **Veracity:** Understanding what Big Data is telling you is one thing. However, it is useless if the data being analyzed are inaccurate or incomplete. This situation arises when data streams originate from diverse sources presenting a variety of formats with varying signal-to-noise ratios. By the time these data arrive at a Big Data analysis stage, they may be rife with accumulated errors that are difficult to sort out. It almost goes without saying that the veracity of the final analysis is degraded without first cleaning up the data it works with.

**6.** **Visualization:** A core task for any Big Data processing system is to transform the immense scale of it into something easily comprehended and actionable. For human consumption, one of the best methods for this is converting it into graphical formats.

Spreadsheets and even three-dimensional visualizations are often not up to the task, however, due to the attributes of velocity and variety. There may be a multitude of spatial and temporal parameters and relationships between them to condense into visual forms. Solving these problems is the main impetus behind AT&T’s Nanocubes visual representation package.

**7. Value:** No one doubts that Big Data offers an enormous source of value to those who can deal with its scale and unlock the knowledge within. Not only does Big Data offer new, more effective methods of selling but also vital clues to new products to meet previously undetected market demands. Many industries utilize Big Data in the quest for cost reductions for their organizations and their customers. Those who offer the tools and machines to handle Big Data, its analysis and visualization also benefit hugely, albeit indirectly.

Although Volume, Velocity and Variety are intrinsic to Big Data itself, the other Vs of Variability, Veracity, Value and Visualzation are important attributes that reflect the gigantic complexity that Big Data presents to those who would process, analyze and benefit from it. All of them demand careful consideration, especially for enterprises not already on the Big Data bandwagon. These businesses may find that their current best practices related to data handling will require thorough revamping in order to stay ahead of the seven V’s.

### 3.2.2 Advantages of Big Data:

* **Using big data cuts your cost.**
* **Using big data increases your efficiency.**
* **Increase our pricing.**
* **We can compete with big businesses.**
* **It helps you increase sales and loyalty.**
* **Allows us to focus on local preferences.**
* **Using big data ensures you hire the right employees.**

### 3.3.3 Challenges associated with Big Data:

* **Lack of proper understanding of Big Data.**
* **Data growth issues.**
* **Confusion while Big Data tool selection.**
* **Lack of data professionals.**
* **Security of data.**
* **Integrating data from a variety of sources.**

## 3.2 Hadoop Framework:

Hadoop is an open source, Java based framework used for storing and processing big data. The data is stored on inexpensive commodity servers that run as clusters. It’s distributed file system enables concurrent processing and fault tolerance (FT). Developed by Doug Cutting and Michael J. Cafarella, Hadoop uses the MapReduce programming model for faster storage and retrieval of data from its nodes. The framework is managed by ‘Apache Software’ foundation and is licensed under the ‘Apache License-2.0’.

For years, while the processing power of application servers has been increasing manifold, databases have lagged behind due to their limiteed capacity and speed. However, today, as many applications are generating ‘Big Data’ to be processed and in this, Hadoop plays a significant role in providing a much needed makeover to the ‘Database world’.

Hadoop is a rapidly evolving ecosystem of components for implementing the Google ‘MapReduce’ algorithms in a scalable fashion on commodity hardware. Hadoop enables users to store and process ‘large volumes’ of data and analyze it in ways not previously possible with less scalable solutions or standard SQL-based approaches. As an evolving technology solution, Hadoop design considerations are new to most users and not common knowledge. As part of the Dell, Hadoop solution, Dell has developed a series of best practices and architectural considerations to use when designing and implementing Hadoop solutions. Hadoop is a highly scalable compute and storage platform.

**The project includes these modules:**

• Hadoop Distributed File System (HDFS): A distributed file system that provides high-throughput access to application data.

• Hadoop MapReduce: A YARN-based system for parallel processing of large datasets.

### 3.2.1 Hadoop Design:

This model does not include the applications and end user presentation components, but does enable those to be built in a standard way and scaled as your needs grow and your Hadoop environment is expanded. The representation is broken down into the Hadoop use cases from above: Compute, Storage, and Database workloads.

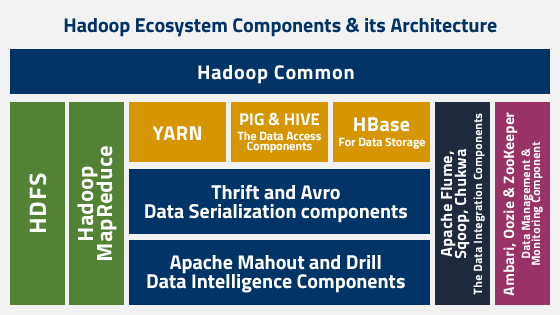
Each workload has specific characteristics for operations, deployment, architecture, and management. Although Hadoop is best known for MapReduce and its distributed file system (HDFS, renamed from NDFS), the term is also used for a family of related projects that fall under the umbrella of infrastructure for distributed computing and large-scale dataprocessing.

The entire Apache Hadoop “platform” is now commonly considered to consist of the Hadoop kernel, MapReduce and Hadoop Distributed File System (HDFS), as well as a number of related projects – including Apache Hive, Apache HBase, and others. Hadoop is written in the ‘Java programming language’ and is a top-level Apache project being built and used by a global community of contributors. Hadoop and its related projects (Hive, HBase, Zookeeper,and so on) have many contributors from across the ecosystem.

### 3.2.2 Architecture of Hadoop:

Hadoop is a framework permitting the storage of large volumes of data on node systems. The Hadoop architecture  allows parallel processing of data using several components:

* **Hadoop HDFS** to store data across slave machines.
* **Hadoop YARN** for resource management in the Hadoop cluster.
* **Hadoop MapReduce** to process data in a distributed fashion.
* **Zookeeper**  to ensure synchronization across a cluster.



**Figure- 3.1 Architecture of Hadoop Ecosystem [10].**

## 3.3 Brief about HDFS:

HDFS is a distributed, scalable, and portable file system written in Java for the Hadoop framework. Each node in a Hadoop instance typically has a single ‘NN’; a cluster of ‘DN’ form the HDFS cluster. The situation is typical because each node does not require a DN to be present. Each DN serves up blocks of data over the network using a block protocol specific to HDFS. The file system uses the TCP/IP layer for communication.

Clients use RPC to communicate between each other. HDFS stores large files (an ideal file size is a multiple of 64 MB [12]), across multiplemachines. It achieves reliability by replicating the data across multiple hosts, and hence doesnot require RAID storage on hosts. With the default replication value, 3, data is stored on three nodes: two on the same rack, and one on a different rack. Data Nodes can talk to each other to rebalance data, to move copies around, and to keep the replication of data high.

HDFS is not fully POSIX compliant, because the requirements for a POSIX file system differ from the target goals for a Hadoop application. The trade off of not having a fully POSIX compliantfile system is increased performance for data throughput. HDFS was designed to handle very large files.

HDFS has recently added high-availability capabilities, allowing the main metadata server (the namenode) to be failed over manually to a backup in the event of failure. Automatic fail-over is being developed as well. Additionally, the file system includes what is called a secondary Name-Node*,* which misleads some people into thinking that when the primary NN goes offline, the secondary namenode takes over. In fact, the secondary namenode regularly connects with the primary namenode and builds snapshots of the primary NN's directory information, which is then saved to local or remote directories. These check pointed images can be used to restart a failed primary namenode without having to replay the entire journal of file-system actions, then to edit the log to create an up-to-datedirectory structure. Because the namenode is the single point for storage and management of metadata, it can be a bottleneck for supporting a huge number of files, especially a largenumber of small files. HDFS Federation is a new addition that aims to tackle this problem to a certain extent by allowing multiple name spaces served by separate namenodes.

### 3.3.1 HDFS Architecture and its components:

There are three components of the Hadoop Distributed File System:

1. NameNode (a.k.a. masternode): Contains metadata in RAM and disk.
2. Secondary NameNode: Contains a copy of NameNode’s metadata on disk.
3. Slave Node: Contains the actual data in the form of blocks.

**Figure- 3.2 HDFS Architecture [11].**

### 3.3.2 Features of HDFS:

There are several features that make HDFS particularly useful, including:

* **Data Replication:** This is used to ensure that the data is always available and prevents data loss. For example, when a node crashes or there is a hardware failure, replicated data can be pulled from elsewhere within a cluster, so processing continues while data is recovered.
* **Fault tolerance and reliability:** HDFS' ability to replicate file blocks and store them across nodes in a large cluster ensures fault tolerance and reliability.
* **High availability:**As mentioned earlier, because of replication across notes, data is available even if the NameNode or a DataNode fails.
* **Scalability:** Because HDFS stores data on various nodes in the cluster, as requirements increase, a cluster can scale to hundreds of nodes.
* **High throughput:** Because HDFS stores data in a distributed manner, the data can be processed in parallel on a cluster of nodes.
* **Data Locality:** With HDFS, computation happens on the DataNodes where the data resides, rather than having the data move to where the computational unit is. By minimizing the distance between the data and the computing process, this approach decreases network congestion and boosts a system's overall throughput.

**3.3.3 Advantages:**

There are five main advantages to using HDFS, including:

* **Large data set storage.**
* **Fast recovery from hardware failure.**
* **Portability.**
* **Streaming data access.**
* Cost effective.

## 3.4 Running Hadoop on AWS:

Amazon EMR is a managed service that lets you process and analyze large datasets using the latest versions of Big Data processing frameworks such as Apache Hadoop, Spark, HBase, and Presto on fully customizable clusters.

* **Easy to use:** You can launch an Amazon EMR cluster in minutes. You don’t need to worry about node provisioning, cluster setup, Hadoop configuration, or cluster tuning.
* **Low cost:** Amazon EMR pricing is simple and predictable: You pay an hourly rate for every instance hour you use and you can leverage Spot Instances for greater savings.
* **Elastic:** With Amazon EMR, you can provision one, hundreds, or thousands of compute instances to process data at any scale.
* **Transient:** You can use EMRFS to run clusters on-demand based on HDFS data stored persistently in Amazon S3. As jobs finish, you can shut down a cluster and have the data saved in amazon S3. You pay only for the compute time that the cluster is running.
* **Secure:** Amazon EMR uses all common security characteristics of AWS services:
  + Identity and Access Management (IAM) roles and policies to manage permissions.
  + Encryption in-transit and at-rest to help you protect your data and meet compliance standards, such as HIPAA.
  + Security groups to control inbound and outbound network traffic to your cluster nodes.
  + AWS Cloud Trail: Audit all Amazon EMR PI calls made in your account to provide security analysis, resource change tracking, and compliance auditing.

# Chapter 4

# Experimental Work and Result Analysis

Here, we are going to explain the working procedure of this cluster, and this will help you all to create your own business model for storing and managing larger data. AWS provides differnt different Data centre regions, you can launch instances from any region and after launching the instances, we have to connect their terminal via using SSH protocol. After the terminal comes on your screen then you have to install Hadoop software and Java also in each machine and after this, we have to open Vim editor for writtimg the code which makes Master-Node (NN) and Data-Nodes to form a cluster.

## 5.1 System Tools and Language Used

**Table 5.1: Tools and System Configuration**

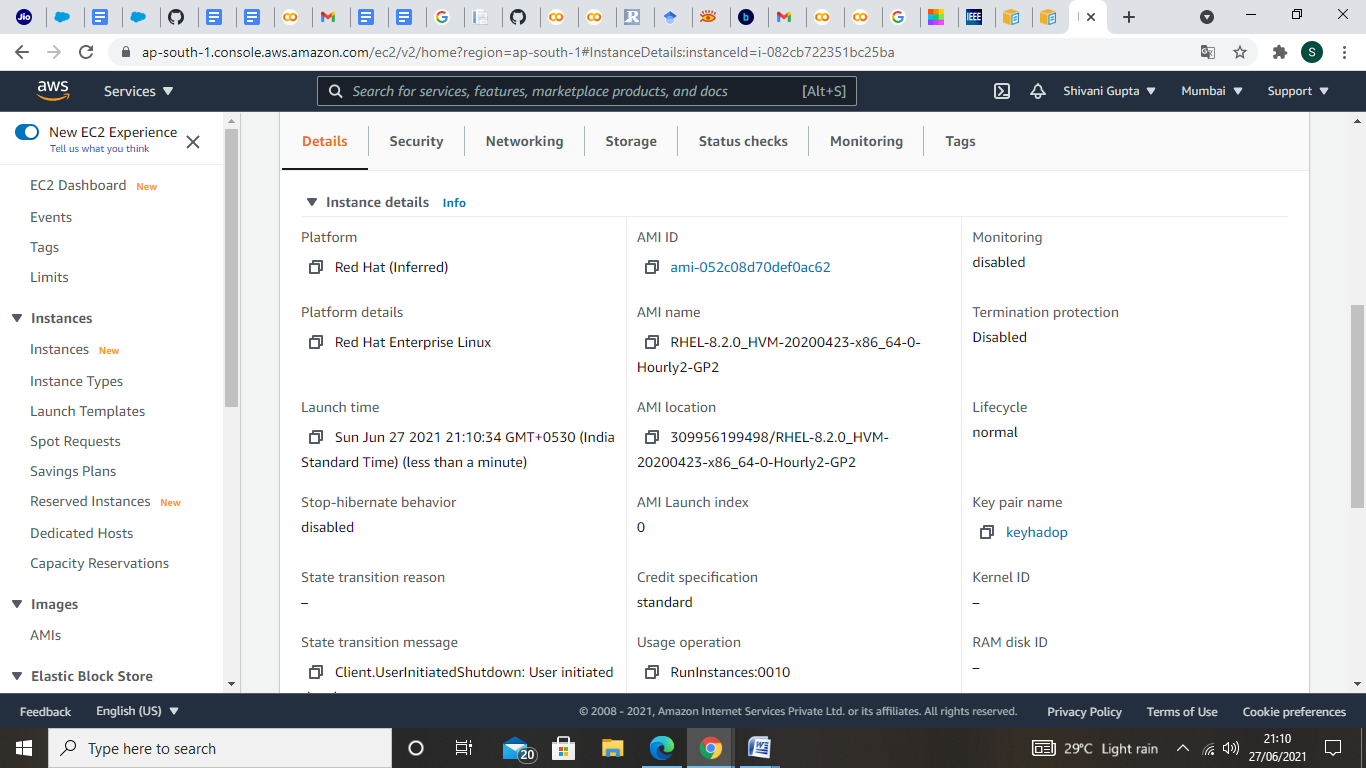
|  |  |
| --- | --- |
| **Description** | **Requirement** |
| Terminal | Linux |
| RAM | 4 GB |
| Hard Disk | 500GB |
| Processor | Intel Core i3 CPU @ 2.4 GHz |
| OS | 64-bit Linux (RedHat) |
| Language | XML |
| Service Provider | AWS |
| Software | Hadoop (HDFS) |

## 5.2 Launching Instances through AWS:

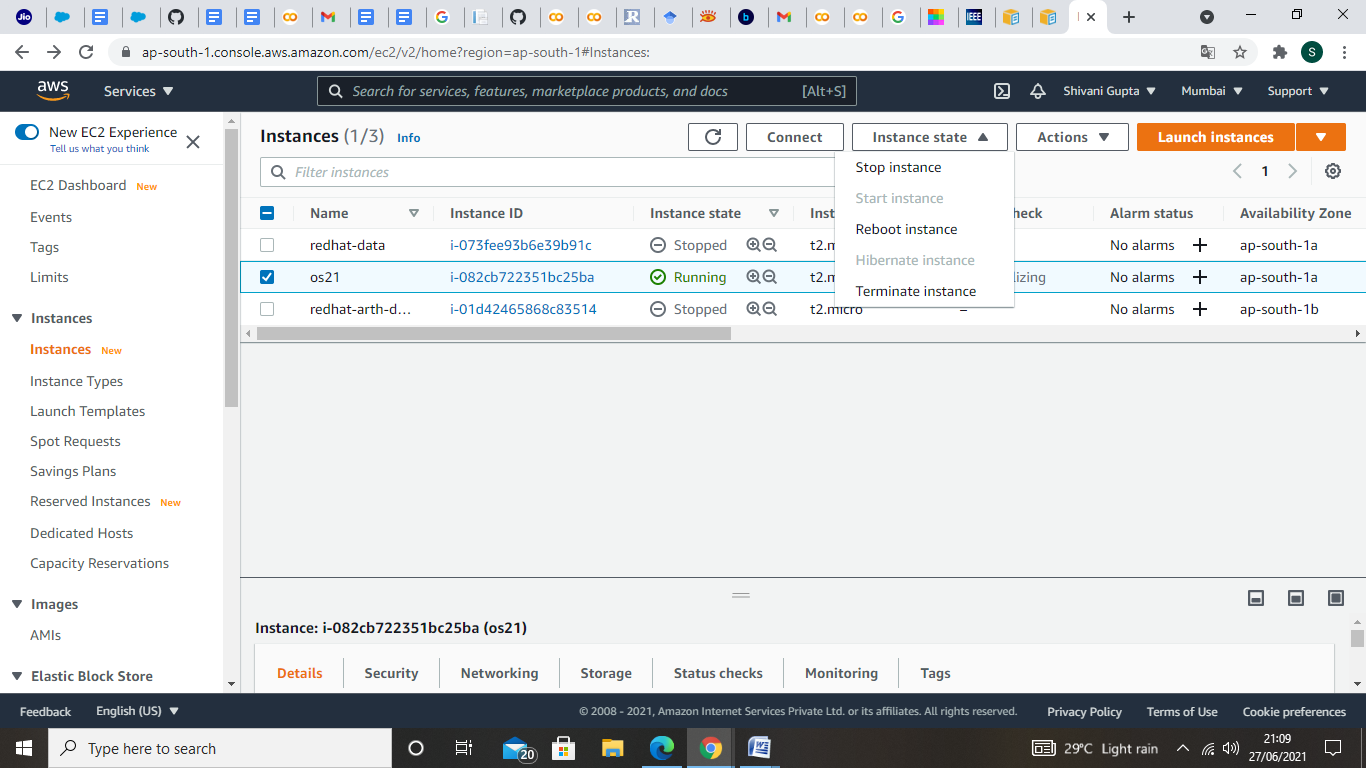
You can launch an instance using the launch instance wizard. The launch instance wizard specifies all the launch parameters required for launching an instance. Where the launch instance wizard provides a default value, you can accept the default or specify your own value. At the very least, you need to select an AMI and a key pair to launch an instance.

**Steps to Launch an Instance:**

* Choose an AMI.
* Choose an Instance type.
* Configure instance details.
* Add Storage.
* Add tags.
* Configure Security Groups.
* Review Instance launch and Select Key-Pair.

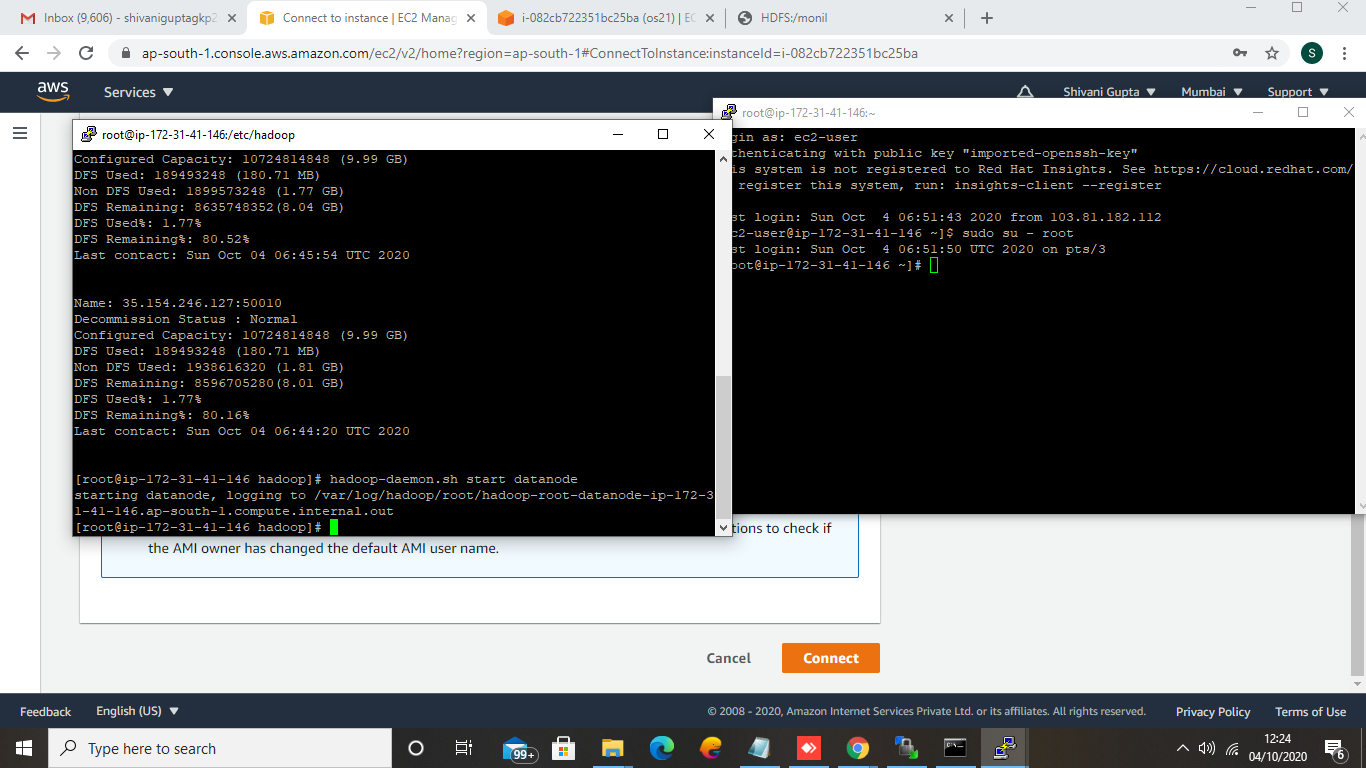


**Figure- 5.1 Details of Instance launched.**



**Figure- 5.2 Start and Stop procedure for launched instances.**

## 5.3 Launch the Terminal:



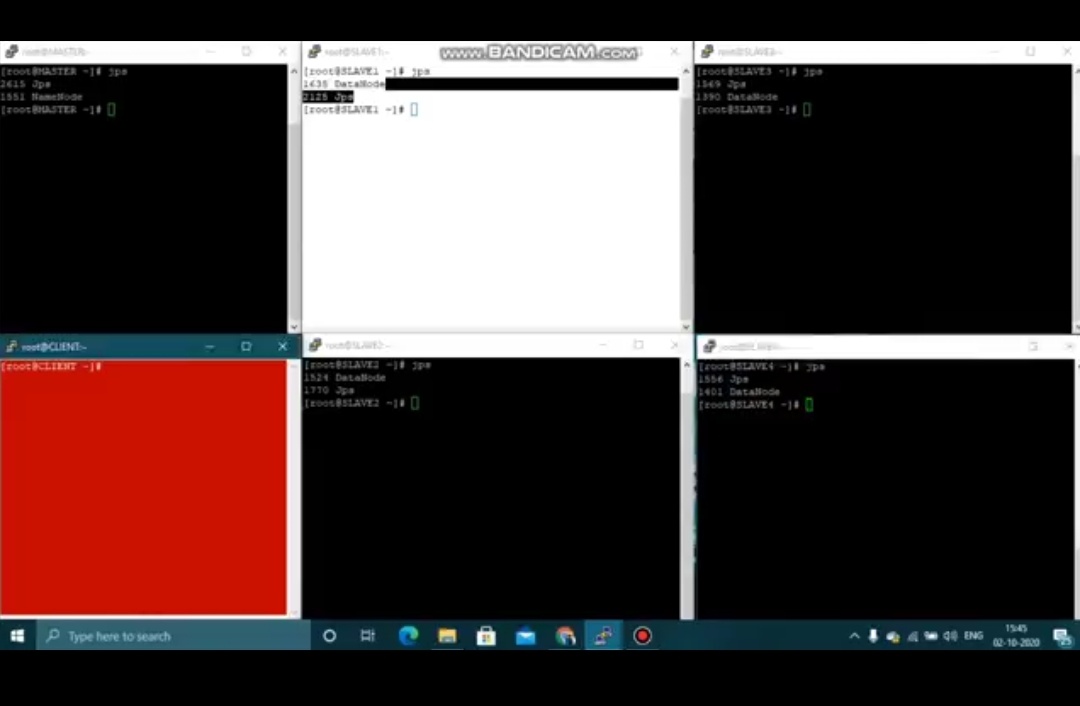
**Figure- 5.3 Linux terminal launch through AWS.**

For launching the terminal, you have to install ‘puttygen’ software for for activating the ‘Security Key Pair’ through SSH protocol and ‘WinSCP’ software for replacing the file from your windows drive to Linux terminal (generated through AWS). Above image is showing that how terminal look like and how it starts.

## 5.4 Implementation Work:

Here, we have used some commands to do this and a little bit of code also. We have mentioned below all the steps which we have followed to make cluster and how slave node contributed their storage space to the master node. How can we create an EBS volume and attached it.

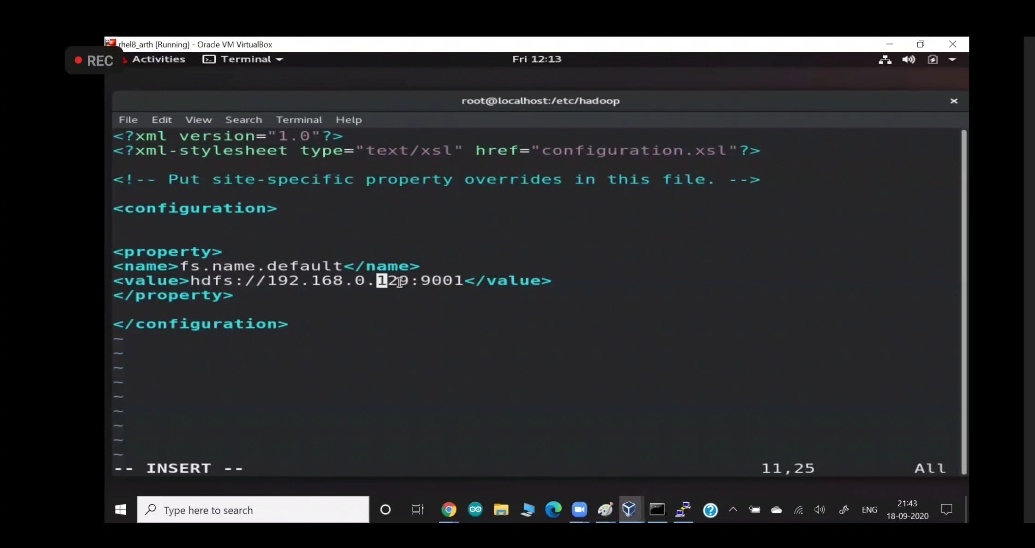
* Launch an instance as Master-Node or NN.
* Launch several instances as Slave-Node or DN.
* Launch one instance as Client Node.
* Connect all DN’s to a NN.



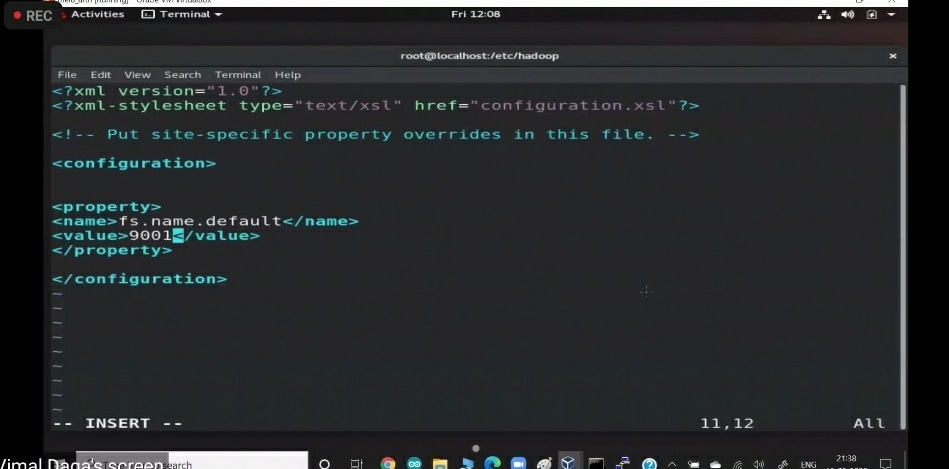
**Figure- 5.4 All Instances (NN, DN’s and CN).**

* How storage is contributed through DN’s.

We have to core-site.xml and hdfs file of master node and DN to write up the code using ‘vim’ command.

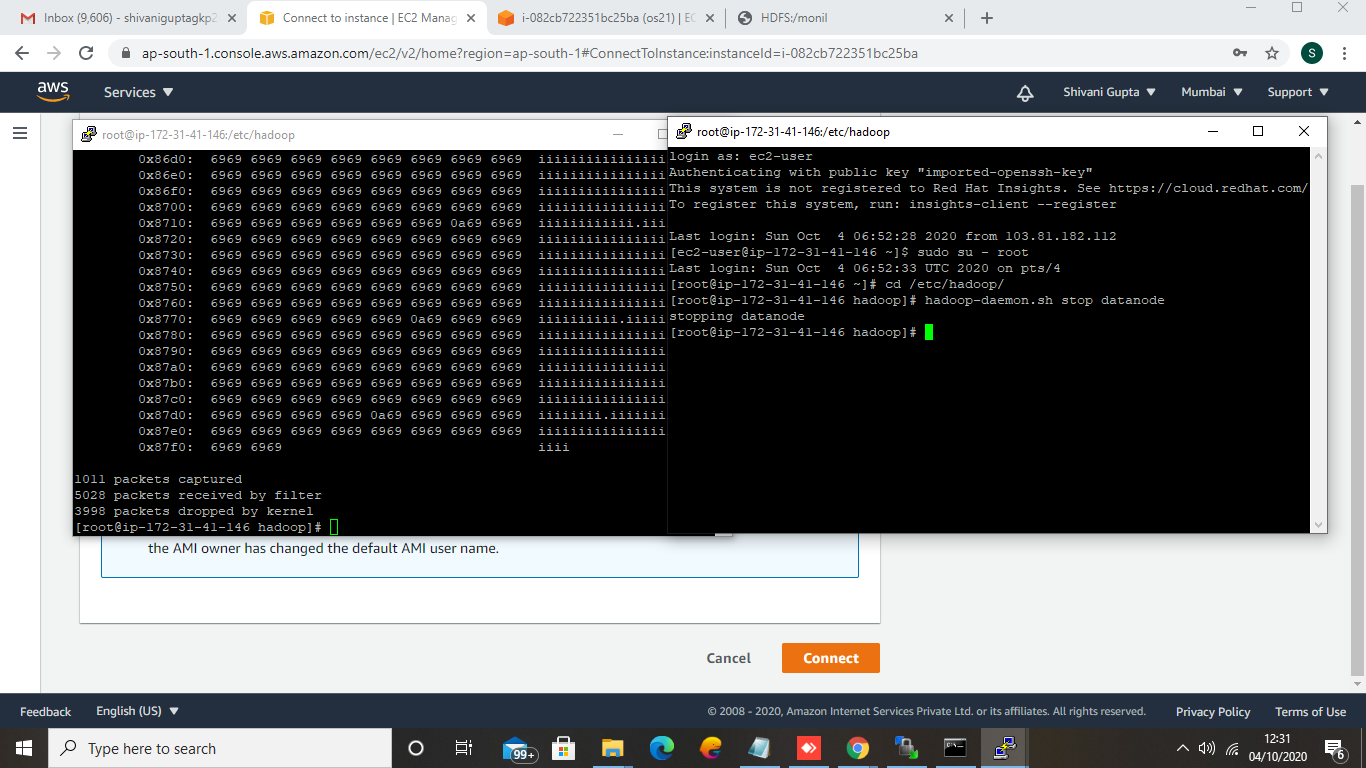


**Figure- 5.5 Master Node (NN) setup.**



**Figure- 5.6 DN setup.**

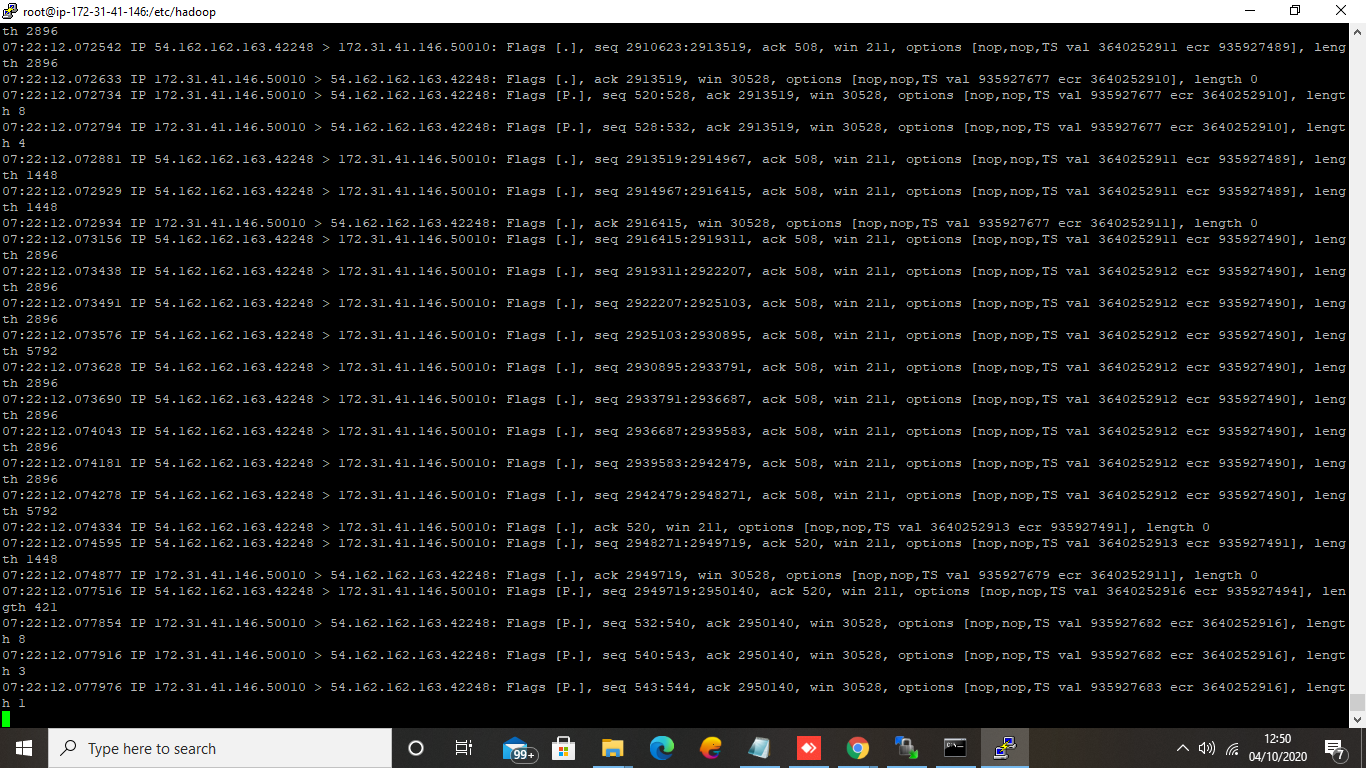
* How master node can read data from any of the slave node.



**Figure- 5.7 Fetching the Data (which is a long text of ‘i’).**

* Here, we show you all that if any of the slave node is disconnected then how master node can take data and from where it is taking.

By using ‘tcpdump’ command, we can see that the data is coming from which machine through their ‘IP’ address.



**Figure-5.8 Tcpdump command result.**

* Download the AWS CLI or work on Linux Terminal.
* Create a Key-Pair and Security Group.
* Create an EBS volume of 1GB.
* Attach the above created EBS volume to the instance.

**Steps for using AWS CLI:**

If you are having the SE account then follow these steps:

1. Login into Educate Starter Account and after that click on Account details.

2. Click on show button and copy that information containing your access key, secret key and the token.

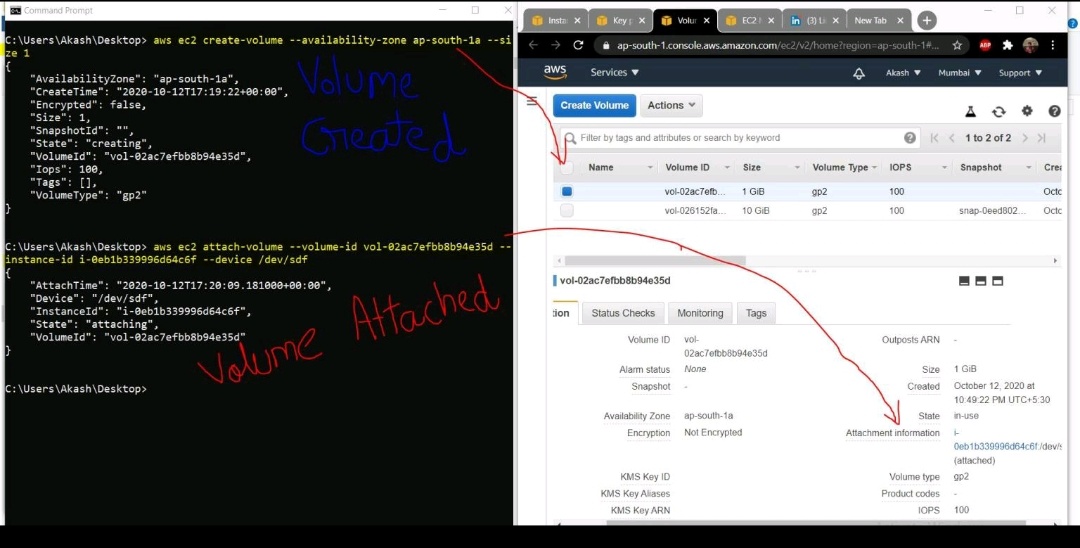
3. Open CMD and run this command:- notepad.aws/credentials.

4. This will open the notepad, remove whatever is there and paste that aws CLI information exactly.

5. Save file and exit.

**Commands:**

* To confirm wheather our AWS CLI program is insttalled or not , use commaqnd as ‘asws –version’.
* For Login, use command as ‘aws configure’.
* To create new Key-Pair, using the following command:- ‘aws ec2 create-key-pair –key-name aws\_key’.
* For creating security group then use ‘aws ec2 create-security-group –group-name MySecurity –description “:MySecurity”.
* To create volume of size 1GB of storage in this CLI, we can use command as ‘aws ec2 create-volume –volume-type gp2 –size 1 –availability-zone (name of zone)’.



**Figure- 5.9 Outcome of Creating and Attaching the volume.**

# Chapter 5

# Conclusion and Future Scope

## 5.1 Conclusion

Big data is a term for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results. The process of research into massive amounts of data to reveal hidden patterns and secret correlations named as big data analytics. These useful informations for companies or organizations with the help of gaining richer and deeper insights and getting an advantage over the competition. For this reason, big data implementations need to be analyzed and executed as accurately as possible. This work presents an overview of big data's content, scope, samples, methods, advantages and challenges and discusses privacy concern on it and also we have shown the implementation of solution Big Data problem with respect to volume and velocity.

## 5.2 Future Scope

We will try to recover some more challenges of Big Data via using AWS and Hadoop framework.

# Bibliography

**[1].** Verma, C., & Pandey, R. (2016, January). Big Data representation for grade analysis through Hadoop framework. In *2016 6th International Conference-Cloud System and Big Data Engineering (Confluence)* (pp. 312-315). IEEE.

**[2].** Warren, J., & Marz, N. (2015). *Big Data: Principles and best practices of scalable realtime data systems*. Simon and Schuster.

**[3].** Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In *2013 international conference on collaboration technologies and systems (CTS)* (pp. 42-47). IEEE.

**[4].** PETTEY, C., & GOASDUFF, L. (2012). Solving ‘big data’challenge involves more than just managing volumes of data.

**[5].** Long, Y., & Thill, J. C. (2015). Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing. *Computers, Environment and Urban Systems*, *53*, 19-35.

**[6].** Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., & Zook, M. (2013). Beyond the geotag: situating ‘big data’and leveraging the potential of the geoweb. *Cartography and geographic information science*, *40*(2), 130-139.

**[7].** Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META group research note*, *6*(70), 1.

**[8].** Gao, S., Li, L., Li, W., Janowicz, K., & Zhang, Y. (2017). Constructing gazetteers from volunteered big geo-data based on Hadoop. *Computers, Environment and Urban Systems*, *61*, 172-186.

**[9].** McNulty, E. (2014). Understanding Big Data: the seven V’s. *Dataconomy. URL: http://dataconomy. com/2014/05/seven-vs-big-data*.

**[10].**<https://www.google.com/url?sa=i&url=https%3A%2F%2Fjava2blog.com%2Fhadooparchitecture%2F&psig=AOvVaw3UzVr7lf6dvipBNBEl6rmv&ust=1624884253494000&source=images&cd=vfe&ved=0CAoQjRxqFwoTCIid6qLrt_ECFQAAAAAdAAAAABAv>.

**[11].**[**https://www.google.com/url?sa=i&url=https%3A%2F%2Fhadoop.apache.org%2Fdocs%2Fcurrent%2Fhadoop-project-dist%2Fhadoop-hdfs%2FHdfsDesign.html&psig=AOvVaw2X6wkLzWDTE8jTna2vIdq7&ust=1624885938723000&source=images&cd=vfe&ved=0CAoQjRxqFwoTCKCQmsfxt\_ECFQAAAAAdAAAAABAD**](https://www.google.com/url?sa=i&url=https%3A%2F%2Fhadoop.apache.org%2Fdocs%2Fcurrent%2Fhadoop-project-dist%2Fhadoop-hdfs%2FHdfsDesign.html&psig=AOvVaw2X6wkLzWDTE8jTna2vIdq7&ust=1624885938723000&source=images&cd=vfe&ved=0CAoQjRxqFwoTCKCQmsfxt_ECFQAAAAAdAAAAABAD)**.**

**[12].** Jahani, E., Cafarella, M. J., & Ré, C. (2011). Automatic optimization for MapReduce programs. *arXiv preprint arXiv:1104.3217*.

# Curriculum Vitae

Shivani Gupta

M. Tech. (Computer Science and Engineering)

**Education**

|  |
| --- |
| [Pursuing] M. Tech. (CSE) from **Madan Mohan Malaviya University of Technology**, Gorakhpur (IN) with an aggregate CGPA of 9.5.  [Aug, 2015 – Jun, 2019] B. Tech. (CSE) from **Indraprastha Engineering College,** Ghaziabad (IN) with 84 %. |

**Personal Information:**

**Father’s Name :** Dr. Uday Narayan Gupta

**Email :** shivaniguptagkp2@gmail.com

**Language Known -**

English Expert

Hindi Expert

**Address -**

Vill- ParsauniBujurg, Post- Tamkuhi Raj

Dist.- Kushinagar

Uttar Pradesh (India)-274407