Emerging Trends in Data Engineering : Amazon Web Services

Submitted in partial fulfillment of the requirement for the degree of
BACHELOR OF TECHNOLOGY IN BIOINFORMATICS

By

TANYA SINGH (161502)

UNDER THE GUIDANCE OF
Mrs. Princy J. Mano
Mr. Karthick Selvam

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,
WAKNAGHAT (H.P.)
June 2020

TABLE OF CONTENTS

CERTIFICATE

It is to certify that the project report titled "**Emerging Trends in Data Engineering : Amazon Web Services**" submitted by **Tanya Singh (161502)** to the Department of Biotechnology and Bioinformatics in partial fulfillment of requirement for the degree of Bachelor of Technology in Bioinformatics from Jaypee University of Information Technology has been carried out under my supervision and this work has not been submitted elsewhere for a degree.

Dr. Raj Kumar

Name of Supervisor :: Mrs. Princy J.Mano

DECLARATION

I, **Tanya Singh**, student of B.Tech (BI) hereby declare that this written submission on "**Emerging Trends in Data Engineering : Amazon Web Services**" expresses my thoughts and my ideas in my own words and I have sufficiently cited and referenced the original sources wherever others' ideas or words have been included. I also hereby declare that I have held allegiance to all the principles of academic truthfulness and integrity and I have not misrepresented or falsified or fabricated any idea/data/source in my work. I understand that any violation of the above will result in a disciplinary action by the Institute and can also evoke penal action from the sources which have not been properly cited or from whom proper permission has not been taken.

(Student Signature)

Tanya Singh (161502)

ACKNOWLEDGEMENT

I would like to express my deepest and sincere gratitude to my supervisor **Mrs. Princy J. Mano** and my trainers **Mrs. Dipti Anjarlekar** and **Mr. Karthick Selvam** for their immense help and guidance throughout the duration of my internship. Their insightful and discerning ideas helped in the successful completion of this internship. The constant support and motivation helped me steer through challenging and difficult phases of my internship.

The facilities provided by the department and my college are also acknowledgeable. I would also take this opportunity to thank my family for their perpetual encouragement throughout the internship period.

ABSTRACT

With the beginning of a new digital era and rapidly growing digital domain, data has become a new form of wealth for organizations, businesses and even countries across the globe. Data can be found in various formats, ranging from structured, numeric data of traditional databases to unstructured text documents, emails, videos, audios, stock market data and financial transactions. Emerging technology such as the Internet of Things is also one of the reasons for the streaming of data into businesses at such a fast pace. In addition to IoT, organizations collect data from other numerous sources like social media, industrial equipment, business transactions, etc.

Firstly, through this project I aspire to study about the recent changes and new emerging technologies in the sphere of Data management and Cloud. Also, my motive is to find out how cloud based technologies help in the domain of data, and how they are beneficial to the businesses.

The main objective of this project is to study about the use and benefits of Amazon Web Services.

CHAPTER 1
INTRODUCTION

As the world is becoming more and more equipped with technology, organizations have started to produce and store large quantities of data. Handling this data and then procuring meaningful information from it is challenging, but also very vital to gain an edge over the competitors. Solutions that analyze both unstructured as well as structured data are critical because they enable the company in gaining information from their privately acquired data, and also from huge quantities of data publicly available on the Internet [1]. The ability to find relation between consumer preferences' private information & products using information from blogs, product evaluations, tweets and also from the data from other social networks brings about comprehensive possibilities to companies for understanding the requirements of the customers, forecast their demands, and improvise the sustainable use of resources. This branch of analytics is popularly coined as Big Data.

Although analytics and Big Data have gained wide popularity, using them in real world applications remains a complicated and time taking task. As Yu [2] points out, it offers a great deal of value to the companies wanting to adopt it, but it simultaneously poses a significant number of challenges to attain this integrated value. A company keen to make use of analytics technology usually purchases pricey software licenses; sets up huge infrastructure for computing; and pays for the consultation of analysts who work with them to gain a better understanding of its way of doing business, assemble their data, and integrate it for analytics [3]. This collective work of the company and the analysts helps them to understand the needs of their customers, behaviors, and future prospects of demands for new product. But such an effort is usually expensive and also mostly lacks flexibility. Besides, storing and managing such large quantities of data is less viable, both technically and economically. Here's where the emergence of technologies such as '**Apache Hadoop**', '**Sqoop**', **'Hive'** come to the rescue, which have been discussed in the following sections.

Chapter 2
Structured Query Language

2.1 Overview

SQL is a computer language that is used to work with databases and is used to store, manipulate and retrieve data stored in a database. It includes creating or deleting a database, fetching as well as modifying rows of a table in a database, etc. SQL is an ANSI (American National Standards Institute) standard language, but there are many different versions of it. SQL is widely popular because it allows the user to embed within other languages using SQL modules, libraries & pre-compilers.

2.2 SQL commands

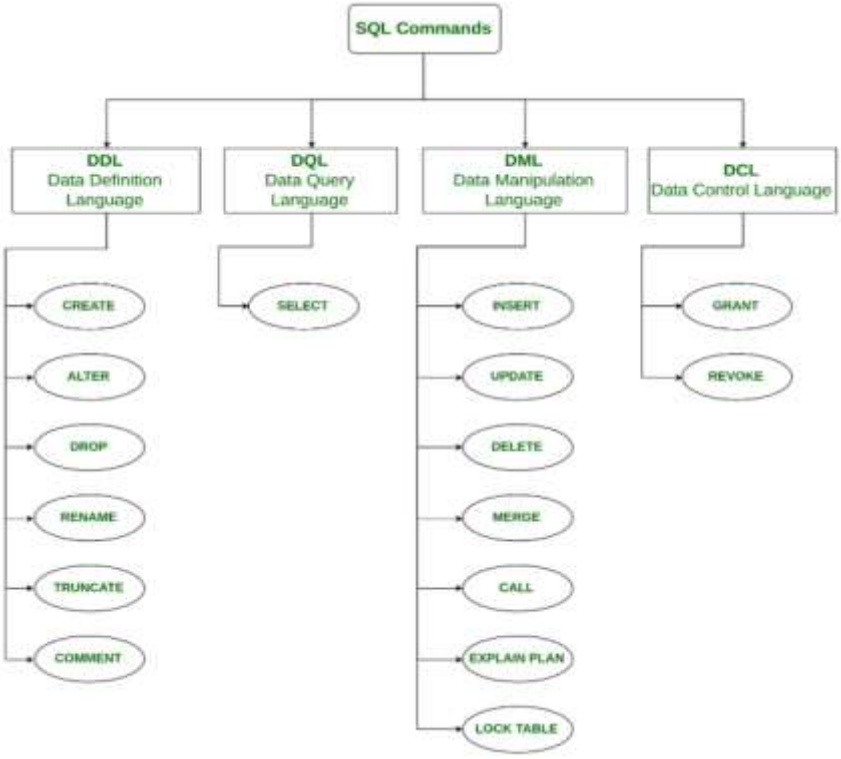

Fig 1. Types of SQL Commands

The SQL commands are divided in to the following four categories:

- **DDL:** Data Definition Language
- **DQL:** Data Query Language

- **DML:** Data Manipulation Language
- **DCL:** Data Control Language

2.2.1 Data Definition Language

DDL consists of the SQL commands which are used to define the database schema. It describes the database schema and creates and modifies the structure of objects in the database. The following commands are part of DDL:

- **Create:** Creating the database or its objects
- **Drop:** Deleting objects from the database
- **Alter:** Altering the structure of the database
- **Truncate:** Removing all the records from a table
- **Comment:** Adding comments to the data dictionary
- **Rename:** Renaming an existing object in the database

2.2.2 Data Query Language

DQL queries consist of the **SELECT** command which is used to run queries on the data composed in the schema objects. Its main objective is to get some relation according to the query run on it. **SELECT** command is used to retrieve data from the database.

2.2.3 Data Manipulation Language

DML consists of the SQL commands that manipulate the data present in the database and this includes most of the SQL statements. The following are the commands of DML:

- **Insert:** It is used to insert data in a table
- **Update:** It is used to update the data already in a table
- **Delete:** It is used to delete records of an existing table in a database

2.2.4 Data Control Language

DCL contains the commands that mainly deal with the rights, permissions and other controls of the database system. The commands included in DCL are:

- **Grant:** It gives a user the privilege to access a database
- **Revoke:** It revokes a user the privilege to access a database given to him using the 'Grant' command

Chapter 3

Apache Hadoop

The Apache Hadoop is a software framework which helps in the distributed simultaneous processing of huge data sets throughout clusters of computers making use of easy programming models. Hadoop is designed to rescale from a handful of servers to hundreds or thousands of machines with each machine providing local computation and storage. Instead of relying on hardware for delivering large-availability, the library itself is created to find out and take care of failures at the application layer, therefore realising a highly-available service upon a cluster of computers, with each of the machine equally prone to going down[4]. It is basically a combination of Processor and RAM.

It uses a collection of open-source software functionalities that help in finding solutions to the problems that involve large amounts of data(in Terabytes) and computation, using a network of many computers or systems.

Hadoop can be further bifurcated into the following:

3.1 Components of Hadoop

3.1.1 Hadoop Distributed File System (HDFS)

HDFS is a distributed file system which is developed to run on material hardware. There are many similarities between HDFS and other distributed file systems that already exist, but the differences are also significant. It is a highly fault-tolerant file system which is developed to be deployed on cheap hardware. It offers high throughput access to data of application and is appropriate for applications having large data sets[5].

Components of HDFS: NameNode and DataNodes

HDFS works in a master-slave topology. Its cluster consists of a single ‘NameNode’, a master server has the function of managing the file system namespace and regulating access of clients to the files. Besides this, there are multiple DataNodes, normally one per node in the cluster, which

has the function of managing the storage attached to the nodes on which they run. Storage of user data in files and a file system namespace is exposed by HDFS. Practically inside this system, a file is divided into one or more blocks and then stored in a set of DataNodes. Executing file system namespace operations like opening, closing, and renaming files and directories is done by the NameNode . It also decides the mapping of blocks to DataNodes. Serving read and write requests from the file system's clients is the responsibility of DataNodes. The DataNodes also create, delete, and replicate blocks when instructed by the NameNode.

The NameNode and DataNode are software pieces which are designed to run on commodity machines. These machines typically run a GNU/Linux operating system (OS). The Java language is used to build HDFS; The NameNode or the DataNode software can be run on any machine that supports Java. Usage of Java language shows that HDFS can run on a wide range of machines. In a typical deployment there is a dedicated machine which is running only the NameNode software. Every other machine of the cluster runs a single instance of the DataNode software.

3.1.2 MapReduce

Hadoop MapReduce is a software framework which simplifies and develops applications which can be used to process huge quantities of data simultaneously in-parallel on huge clusters of commodity hardware in a reliable and fault-tolerant manner[6].

The main job of MapReduce is to split-up the data-set into smaller independent parts which can then be easily processed using 'map tasks' parallely. The outputs of the map are first sorted and later fed to the 'reduce tasks' in this framework. File-system stores the input as well as output of the job. Scheduling & monitoring of tasks and re-execution of the failed tasks is done by the framework.

Ideally the compute and the storage nodes, that is, the MapReduce framework and the Hadoop Distributed File System run on the same set of nodes. Such a configuration helps in effectively scheduling tasks on the nodes already having data, resulting in very high average bandwidth across the cluster.

Components of MapReduce:

The MapReduce framework consists of a single master ‘JobTracker’ and a slave ‘TaskTracker’ for each cluster node. Scheduling a component's job is a task of the slaves, while master’s responsibility is to monitor them and re-execute the failed tasks. The master directs the slaves to execute the tasks.

3.2 Sqoop

Apache Sqoop is used for the efficient transferring of data between various data sources that may be structured, semi-structured or unstructured. An example of structured data source is Relational Database which has a well defined schema of the stored data. Examples of semi-structured data sources include Cassandra and Hbase and HDFS is an example of unstructured data source supported by Sqoop[7].

Sqoop supports progressive loading of a table, that is, a free form SQL query and saved jobs which can be run multiple times for importing updates made to a database since the last import. Populating tables in Hive or HBase can also be done using imports, whereas data from Hadoop can be put into a relational database using exports. Sqoop got its name from "SQL-to-Hadoop". It became a top-rated project of Apache in March 2012.

3.3 Hive

For providing data query and analysis, Apache Hive, a data warehouse software project was built on top of Apache Hadoop. An SQL-like interface for querying data which is stored in various databases and file systems is given by Hive, that integrates with Hadoop. Conventional queries

of SQL need to be implemented in the Java API of MapReduce for executing the queries and application of SQL over distributed data.

Hive offers the required abstraction of SQL to merge SQL-like queries with Java eliminating the need to run the queries in the Java API. Hive catalyses integration of applications based on SQL with Hadoop because a majority of applications of data warehousing work with querying languages based on SQL. Although it was initially developed by facebook, a software fork of Apache Hive is maintained by Amazon which is included in Amazon Elastic MapReduce on Amazon Web Services.

3.4 HBase

HBase is an open-source non-relational distributed database modeled after Google's Bigtable. It is written in Java. It runs on top of HDFS and is developed as part of Apache Software Foundation's Hadoop project, to provide capabilities similar to Bigtable for Hadoop, which means that it stores large quantities of sparse data in a fault-tolerant way. Tables in HBase can be accessed through the Java API, and they serve as the input as well as output for MapReduce jobs running in Hadoop. It stores key-values in a column-oriented way and is usually used because of its compatibility with Hadoop and HDFS. HBase is ideal for faster read and write operations on large datasets with high throughput and low input/output latency.

Chapter 4

Cloud Computing Basics

4.1 Virtualization

Virtualization is the capability to share the physical instance of an application or a resource among multiple organizations or users. This is achieved by logically allocating a name to all those physical resources and then serve a pointer to those physical resources based on requirement. Across an existing OS or hardware, a virtual machine is created on which other operating systems or applications can be run. This is known as Hardware Virtualization. A separate environment is provided by the virtual machine which has a logical distinction from its underlying hardware. In this arrangement, the system or the machine is the host whereas the virtual machine is the guest machine. The management of this virtual environment is done by a firmware which is known as a hypervisor.

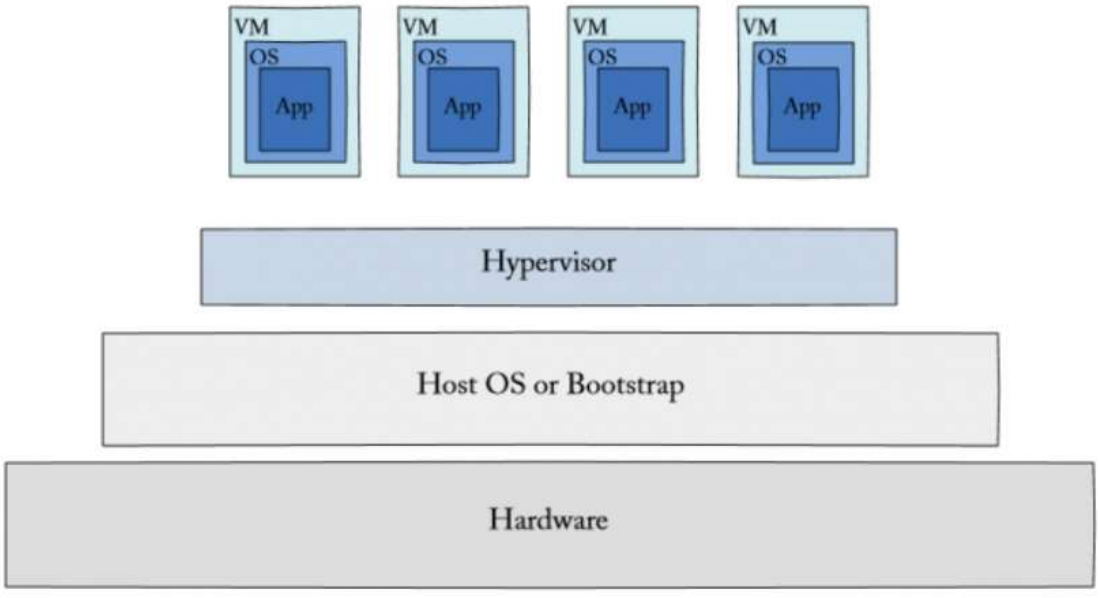

Fig 2. Cloud Virtualization Architecture

4.2 Working of Virtualization in Cloud

Virtualization plays a prominent role in cloud technology as well as its working mechanism. Ideally, in the cloud, the users share the data that is stored in the cloud similar to an application in addition to their infrastructures using virtualization. Virtualization is mostly used to provide standard versions of applications to the cloud customers and with the release of successive versions of an application; the providers can easily serve the application to the cloud and its users. This can be made possible using virtualization only.

In reality, most of the hypervisors today, use a combination of several kinds of hardware virtualization. Virtualization basically means to run multiple systems on a single machine and share all hardware resources. It helps in sharing IT resources to gain benefit in the business.

4.3 Types of Virtualization

There are two types of Virtualization based on the type of the Hypervisor. They are:

4.3.1 Type 1 Hypervisor

The hardware system on which the hypervisor runs single or multiple virtual machines is called the host machine, whereas, each virtual machine is known as a guest machine. Type 1 Hypervisor is also known as **Bare Metal Hypervisor** or **Native Hypervisor**. It directly runs on the hardware of the host machine. Besides this, it also manages the guest operating systems and controls hardware. The first hypervisors native hypervisors and they were developed by IBM. Some examples of Type 1 hypervisors are Microsoft Hyper-V, VMware fusion, Oracle VM, etc.

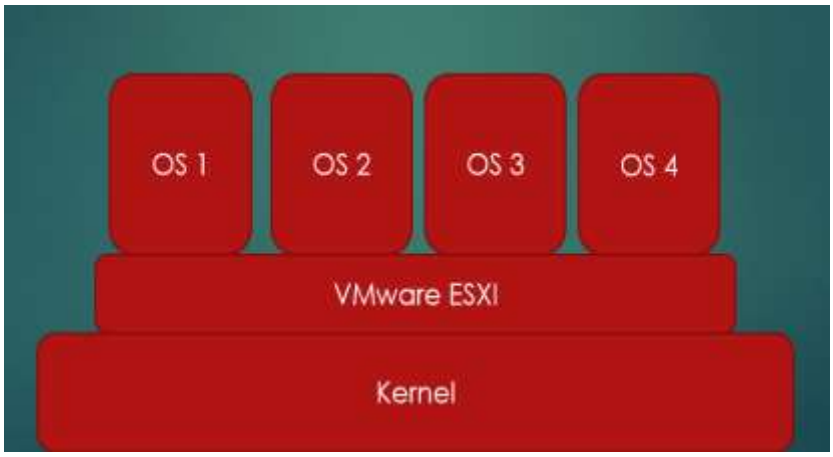

Fig 3. Type 1 Hypervisor

4.3.2 Type 2 Hypervisor

Type 2 Hypervisor is also known as a Host OS Hypervisor. This type of hypervisor run on an OS just like other computer programs, that is, there is a hypervisor on top of the OS. It provides an emulator environment to run another OS. This means that a guest operating system runs like a process on the host. VMware Workstation, VMware Player, VirtualBox, etc. are few examples.

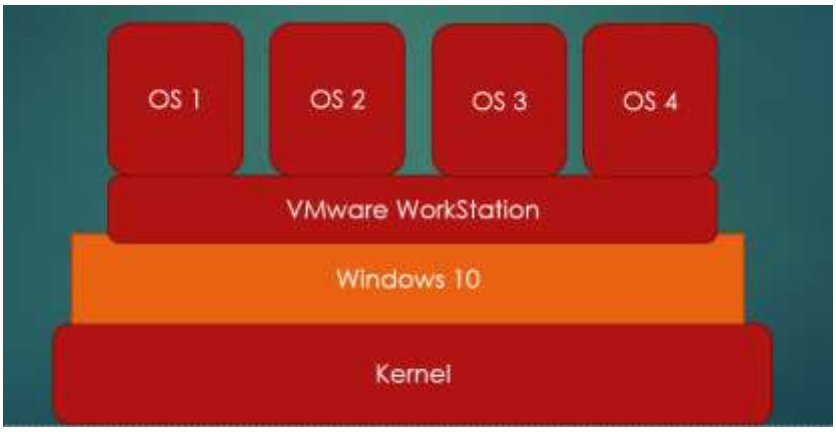

Fig 4. Type 2 Hypervisor

4.4 Cloud Computing

Cloud computing is a computing service based on the internet where huge groups of remote servers are connected in a network to allow centralized data storage and online accessibility of computing services or resources.

With the use of cloud computing, companies and organizations share and make use of computing & storage resources instead of developing, using, and improvising the infrastructure of their own.

Cloud computing model provides the following features.

- Users can acquire and release resources as per demand.
- Resources may be scaled up/down easily, as need be.
- Secured accessibility of resources over a network.
- Cloud service providers may offer a pay-as-you-go model, in which customers pay according to the type of resources and usage.

4.4.1 Types of Clouds

There are three types of clouds – Public, Private, and Hybrid cloud.

- **Public Cloud**: Arbitrator service providers build resources as well as and offer it to the users in public cloud using the Internet. Data & related security of data of customers is realised with the infrastructure owned by the service provider.
- **Private Cloud**: A private cloud offers features that are like that of the public cloud, but the resources as well as the services are either looked after by the organization or by the arbitrator for the specific user's organization. In such a cloud, the main emphasis of control is on the infrastructure which reduces the issues related to security.
- **Hybrid Cloud**: It is a mixture of private & public cloud. The choice of running on private or public cloud depends on different factors like sensitivity and application of data, certifications of industry and the required standards or regulations, etc.

4.4.2 Models of Cloud Service

There are 3 service model types in cloud viz. IaaS, PaaS, and SaaS.

- **IaaS**: IaaS is an abbreviation for Infrastructure as a Service. In such a service users are provided with the facility of provisioning processing, storage, and connectivity of network as the need be. Such a service model helps the users to create and deploy their own applications on the offered resources.
- **PaaS**: PaaS is an abbreviation for Platform as a Service. In this, the users are provided with numerous services such as databases, queues, workflow engines, e-mails, etc. These components can then be used by the customers to build their own applications. All the services, availability of resources and backup of data are managed by the service provider to help the customers focus mainly on the functionality of their application.
- **SaaS**: SaaS is an abbreviation for Software as a Service. In such a service, The customers are provided with end-user applications by the third-party providers with some administrative capability at the application level, for example the capability to deploy and manage their users. Besides,some abilities such as customizing upto some extent is possible like the customers using their choice of colors and putting up their own corporate logo, etc.

Examples of Public Cloud Providing Companies are:

- Amazon Web Services(AWS)
- Google
- Azure(Microsoft)
- Oracle
- IBM
- Dell Emc
- Digital Ocean

Chapter 5

Amazon Web Services

Amazon Web Services (AWS) is a subordinate company of Amazon which provides cloud computing platforms and APIs to individuals, companies, and governments as per demand and on a metered pay-as-you-go basis. Collectively, these web services of cloud computing offer a

set of basic technical infrastructure and distributed tools & building blocks for computing. One of the popular services is Amazon EC2 which is used to deploy a virtual cluster of computers that are always available at the user's disposal, via the Internet. The virtual computers provided are similar to real computers in most of the attributes, including the CPUs and GPUs for processing, RAM, SSD storage, OS as per choice, and pre-loaded application softwares like web servers and databases.

AWS contains more than 212 services as of now which includes popular services such as Amazon EC2 for computing and Amazon S3 for storage besides numerous other services in the domains of networking, database, analytics, services for applications, tools for development, and tools for IoT. Major other services offered are not for the end users, but they offer different functionality using APIs to the developers that can be used in different applications.

5.1 IAM

Identity and Access Management is a web service that enables the user to safely manage accessibility of AWS resources. It can be used to decide who is signed in and has the permissions for using the resources[8].

When an AWS account is created, we first start using a singular identity to sign-in which has full access to all the services as well as resources of the AWS account. Such an ID is known as the 'root user' of the account and it can be accessed by logging in using the credentials that were used for creating the account. But it is strongly advisable that the root user should not be used for the daily tasks including the administrative ones. The advisable best practice is using the root user for creating the first IAM user only and then using the root user only for the necessary tasks of account & service management.

5.2 EC2

Amazon Elastic Compute Cloud i.e. EC2 offers elastic or flexible computing capacity, as the name suggests, in the AWS cloud. Use of EC2 abolishes the need of investment in hardware, enabling one to develop & launch the applications faster. It can be used to deploy as many virtual

servers as needed, configuring the security besides networking, and also managing the storage. It enables the user to scale up/down for handling the changes of the requirements or changes in traffic inflow, decreasing the need of forecasting traffic[9].

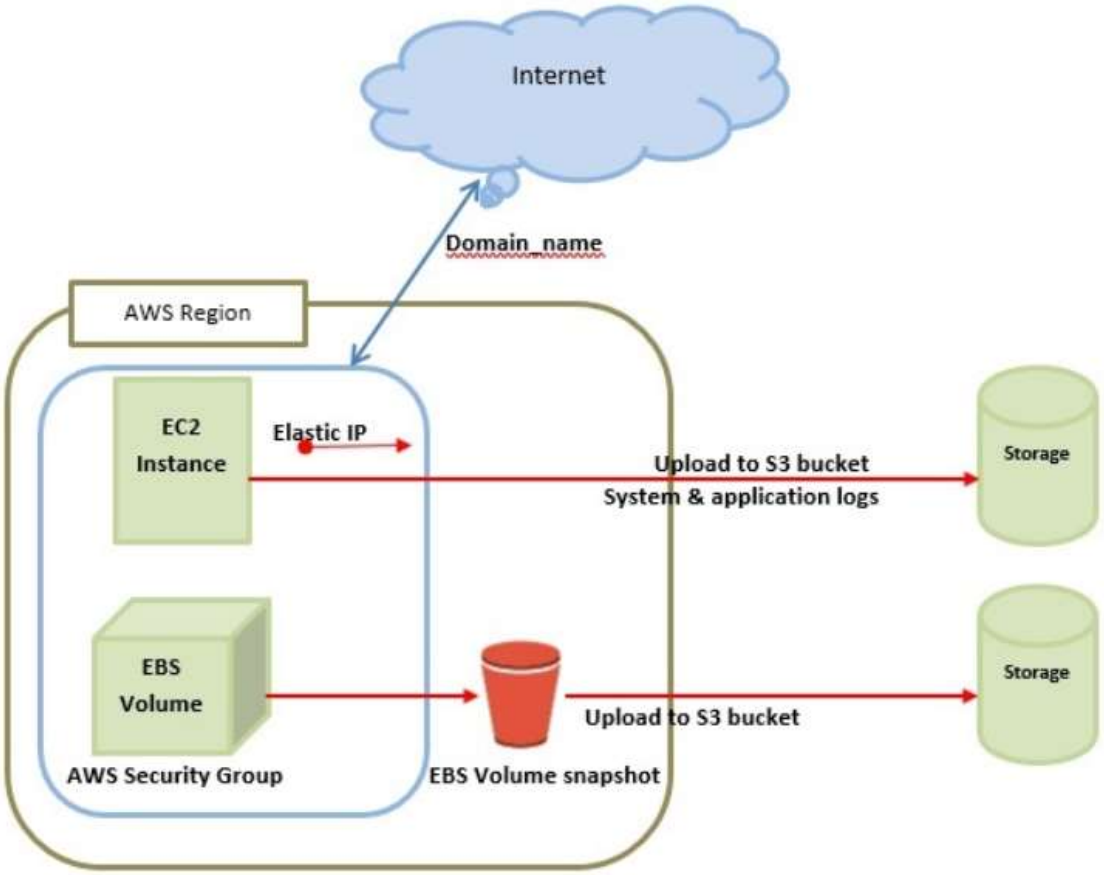

Fig 5. EC2 Architecture

As the term Elastic explains, it can extend and shrink as need be. Amazon EC2 provides the elasticity feature for computing resources such as CPU, RAM, Storage and Networking etc. in the cloud. It is an IAAS kind of service where the user takes RAM, CPU, Networking capability and Storage on rental basis and installs an OS on top of this base layer to use it.

It provides numerous features such as:

- Virtual computing environments known as instances.
- Templates for these instances which are preconfigured, known as AMI (Amazon Machine Images).

- Numerous combinations of storage, CPU & memory as well as capacity of networking of the instances, better known as the type of the instance
- Using key pairs for secure login to the instances
- A firewall that helps the user in specifying not only the protocols and ports but also the source IP address ranges which can access the instance by security groups
- Elastic IP addresses, i.e, static IP addresses of IPv4.for dynamic cloud computing
- Virtual networks can be created which are logically isolated from the rest of the AWS cloud, and which can be optionally connected the user's own network, known as virtual private clouds (VPCs)

5.3 AMI (Amazon Machine Image)

An Amazon Machine Image gives the information necessary for launching an instance. The user should state an AMI when launching an instance. A user can launch one or more instances using a single AMI whenever multiple instances of the same configuration are needed. Conversely the user can also make use of distinct AMIs for launching one or more instances having different configurations[10].

An AMI includes the following:

- Multiple EBS snapshots, or, for instance-store-backed AMIs and a template of the instance's root volume.
- Give accesses that help to decide the AWS accounts which may be able to use the AMI for launching instances.

- A map of block devices specifying the volumes to be attached to the instance whenever it is launched.

5.3.1 Using an AMI

The diagram shown below explains the lifecycle of AMI. After an AMI is created and launched, it can be used for launching new instances.

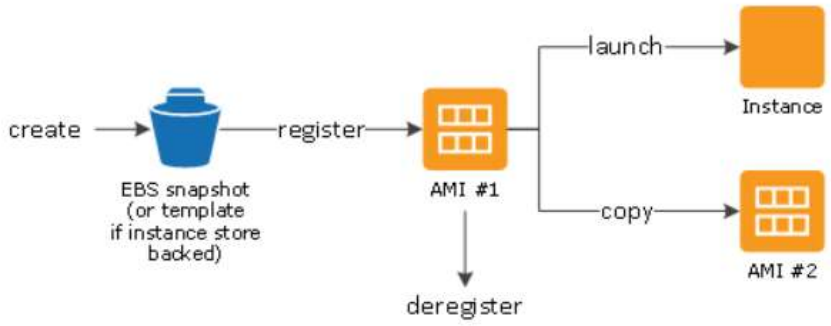

Fig 6. Lifecycle of AMI

5.4 S3

Amazon Simple Storage Service S3 is a cloud storage service. It can be used for storing and retrieving any quantity of data from anywhere on the web at any time. Such tasks can be accomplished with the use of the AWS Management Console, which is a simple and inherent web interface[11].

S3 is an object oriented storage, that is, data is stored as objects within the buckets. An object contains a file and any metadata describing the file. For storing an object in Amazon S3, the file that is needed to be stored is uploaded to the bucket in which it is to be stored. S3 is different from block & file cloud storage because here each object is given a unique ID number, and this particular ID number is used by the applications to access the object.

The S3 cloud storage service facilitates a user with the access to similar systems as used by Amazon itself for running websites of their own. It helps the users to upload/download and store feasibly any file or object which is up to 5 terabytes (TB) in size, and the ceiling for a single upload is at 5 gigabytes (GB).

5.4.1 Features of Amazon S3

99.99999999% resilience is offered by S3 to the objects that are stored in it and it also sticks to numerous security and compliance standards. A user is enabled to connect S3 to any other security and monitoring services offered by AWS such as CloudTrail and Macie.

Uploading data to S3 can be done using the APIs of S3, via the internet. It also consists of Transfer Acceleration of S3 to boost the movement over large distances, and AWS Direct Connect for a consistent and private connection between S3 and the user's data center.

5.5 Virtual Private Cloud

Amazon Virtual Private Cloud or Amazon VPC provides the ability to deploy a section of the AWS Cloud that is isolated logically, where AWS resources can be launched in a virtual network defined by the user. The user has absolute control on the environment of virtual networking, including the options to select his own range of IP addresses, create subnets, and configure the route tables as well as the network gateways. User can make use of both IPv4 as well as IPv6 in his VPC access his resources and applications securely[12].

The network configuration of an Amazon VPC can be easily customized. For instance, creation of a public-facing subnet for a web server having internet access. Backend components like databases and application servers, can also be placed in a subnet which is private-facing and

doesn’t have internet access. To control the access to EC2 instances in the subnet, numerous layers of security can be used, including security groups and network access control lists.

5.6 Amazon RedShift

Redshift is an Amazon Cloud Data Warehouse tool that has a massive parallel computing database(there’s no limitation on scaling up of the resources in terms of computing and Volume of Data). It is a column oriented database.

Amazon Redshift is a data warehousing product that forms part of AWS. It is named so to signify shifting away from Oracle, red being an allusion to Oracle, having a corporate color as red and is sometimes informally referred to as "Big Red”.

Amazon Redshift is a petabyte-scale, fully managed, cloud-based data warehouse service offered by AWS. It is an efficient solution for collection and storage of all of the user's data that enables the user to analyze the data using numerous BI tools for acquiring new useful information vital for business[13].

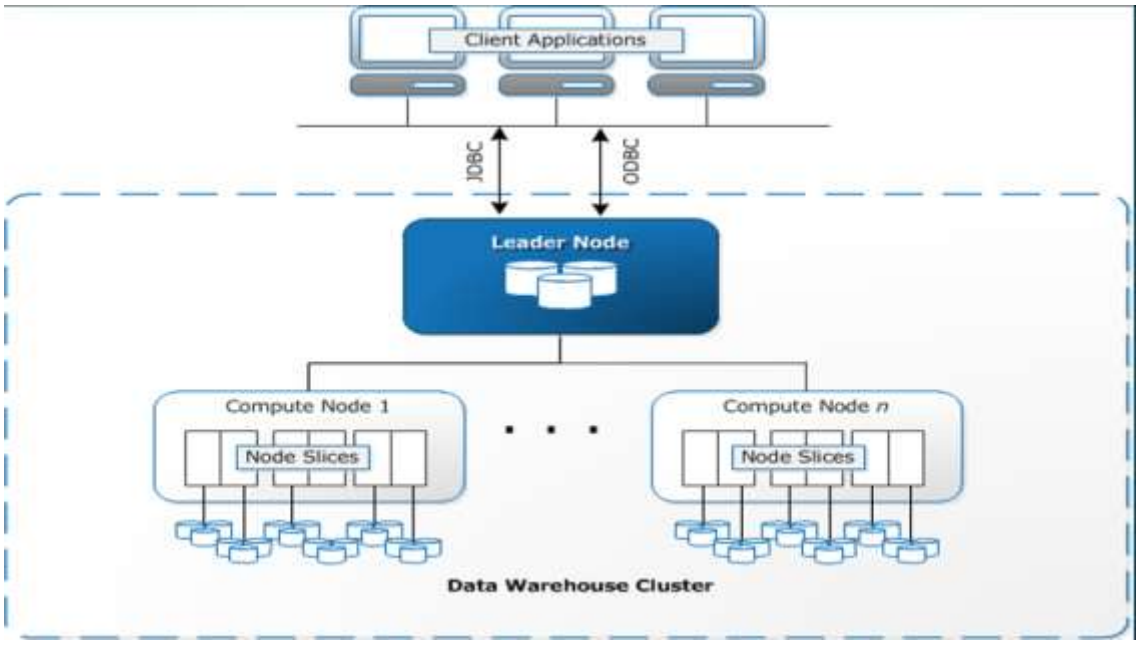

Fig 7. RedShift Architecture

5.6.1 Leader Node

The leader node looks after the process of exchange of information between the compute nodes and the client programs. It breaks down and comes up with a plan of execution to carry out operations on the database, specifically, the sequence of steps that are necessary to obtain outputs of composite queries. The code is compiled by the leader node and then distributed to the compute nodes, and then each compute node is allocated a chunk of data. The statements of SQL received by the leader node are distributed between the compute nodes whenever the tables stored in compute nodes are cited by the query. Apart from this, all the queries are executed solely on the leader node.

5.6.2 Compute Node

Compilation of code for each individual element in the execution plan is done by the leader node and then this code is assigned to the compute nodes. The compiled code is then executed by the compute nodes and the intermediary results are sent back for aggregation, to the leader node. Every compute node consists of a CPU, memory space, and disk storage attached to it depending upon the type of node. If the workload increases, storage & the computing capacity of a cluster can be expanded by adding more nodes or enhancing the type of node or both.

CHAPTER 6

PROJECT

6.1 Introduction

In this project fleet management is performed using the methodology of **ETL** which is a type of data integration that consists of the three steps namely Extract, Transform and Load. It is used to blend data from multiple sources. In this project, data is taken or extracted from a **S3 bucket**. This data is in the format of a **CSV file**. This file is then fed to a **MySQL database** running on an **Amazon EC2** instance of Ubuntu. This is the transformation step, that is, the original CSV file is converted to a MySQL database to help ease the analysis. Then this database is stored (loaded) into a data warehouse, that is, **Hive** using **Sqoop**. Hive enables **PySpark** to run SQL like queries on the data using python. Various queries are performed on this data and the output is stored in the format of a **JSON** file.

6.2 Project Architecture

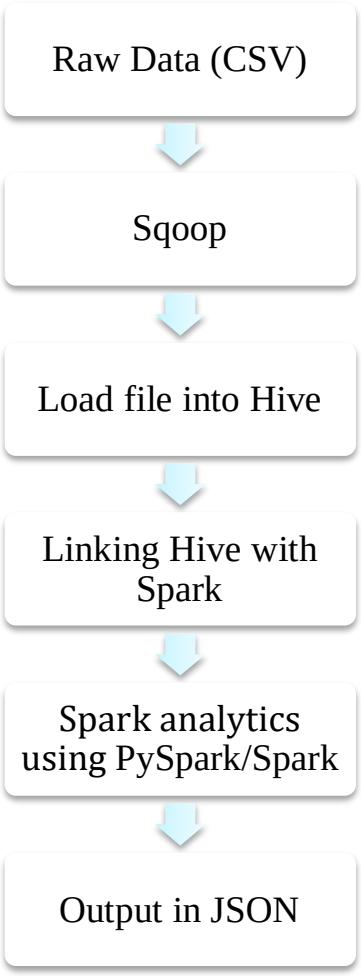

Fig.8. Project Flow

6.3 Process flow

6.3.1 Create a bucket

Created a bucket in Amazon S3 named **Blue-cognizant** to store the data file.

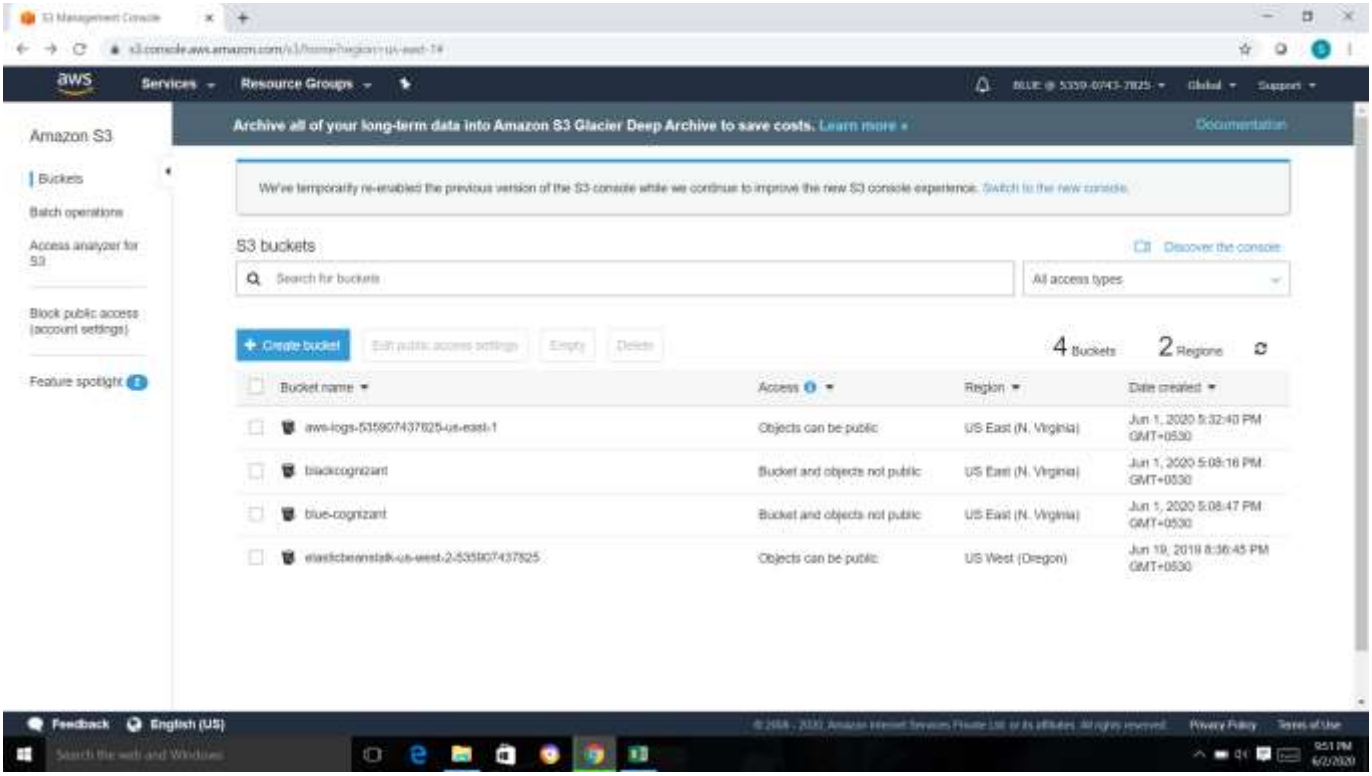

Fig 9. Amazon S3 bucket

6.3.2 Create an EC2 instance for SQL connection

Created an Ubuntu instance in Amazon EC2 instance for connecting MySQL with the data table.

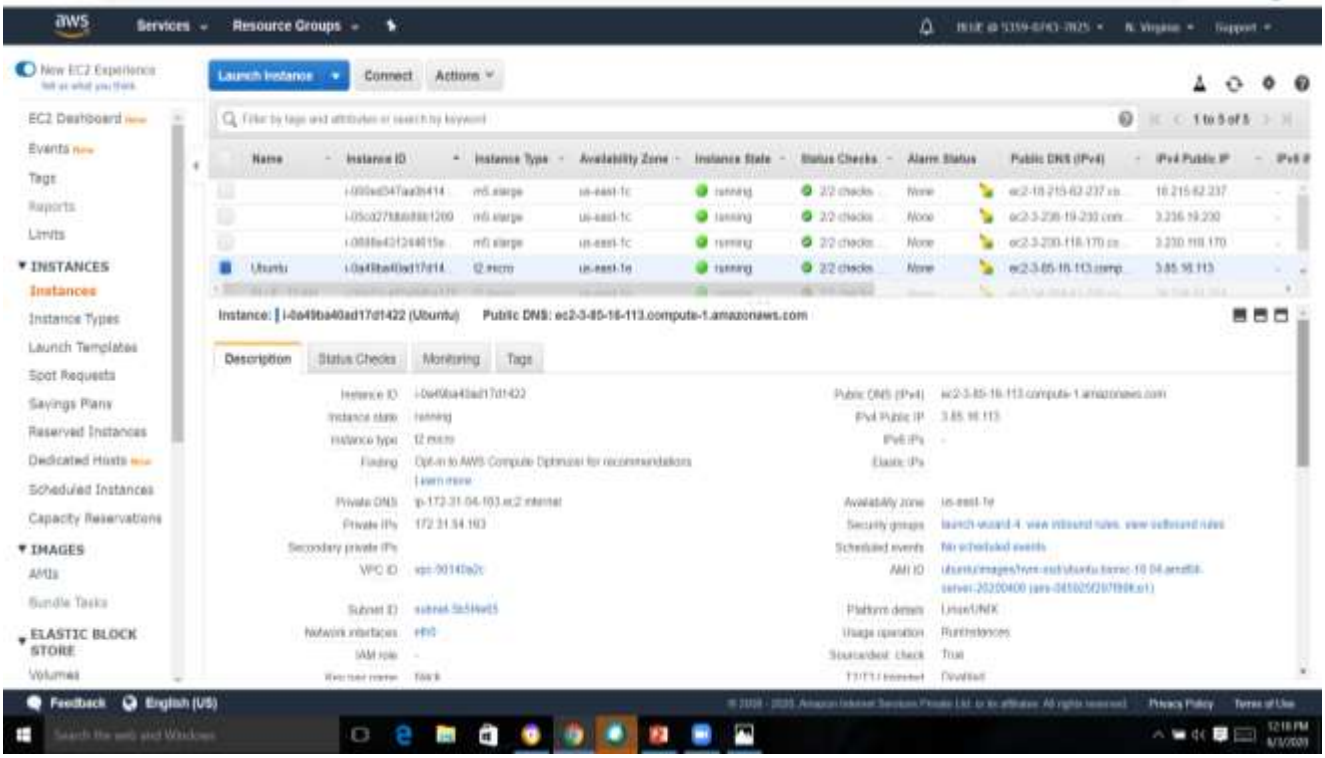

Fig 10. Amazon EC2 instance

6.3.3 Installing MySQL in the Ubuntu instance

Installed MySQL in the Ubuntu instance launched.

Granted all the privileges.

Granted Remote user access.

Set the global variable in MySQL.

Wrote the query to create a table in MySQL workbench.

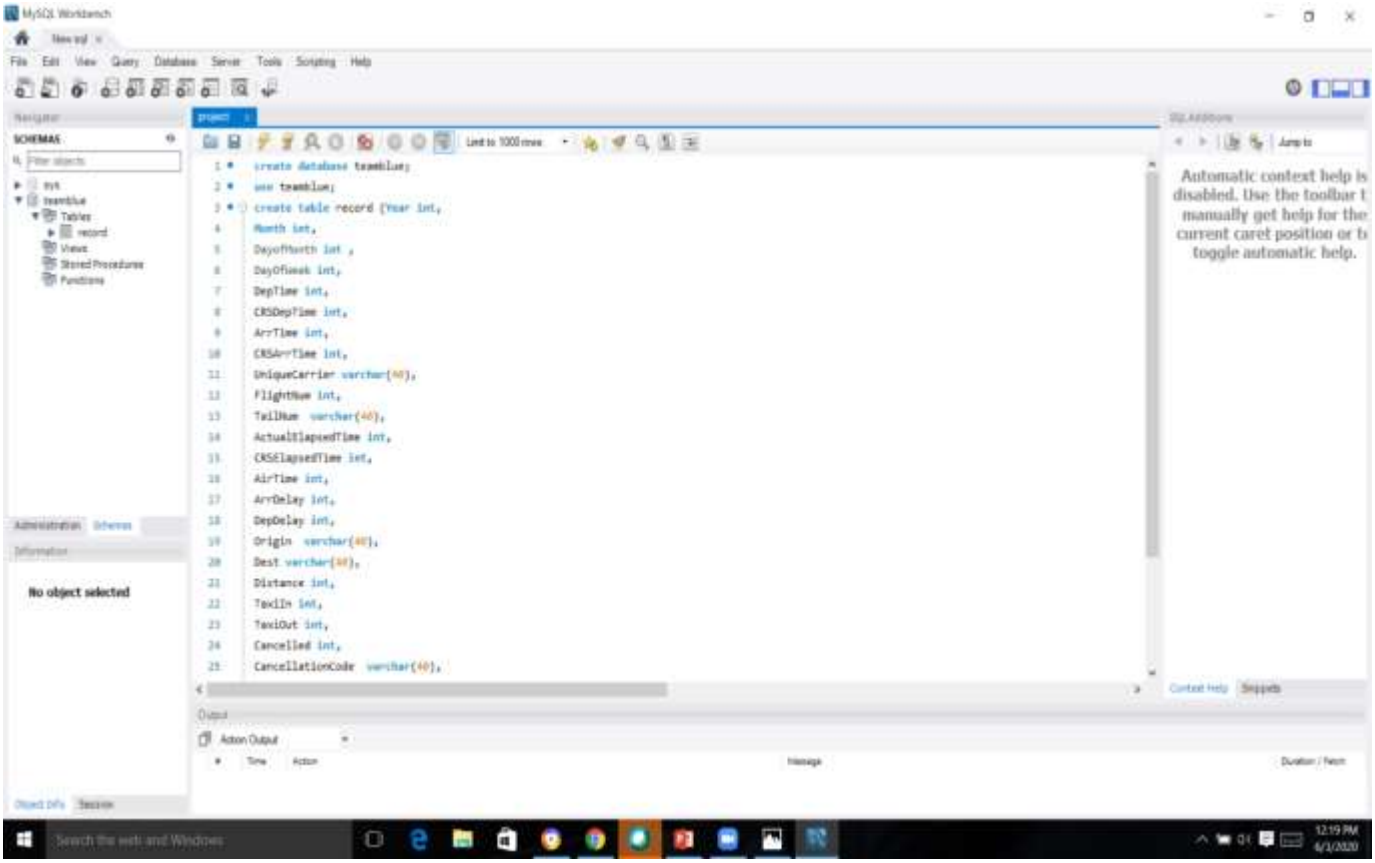

Fig 11. Table creation in MySQL workbench

6.3.4 Creating an EMR cluster

Created an EMR cluster in Amazon EMR service.

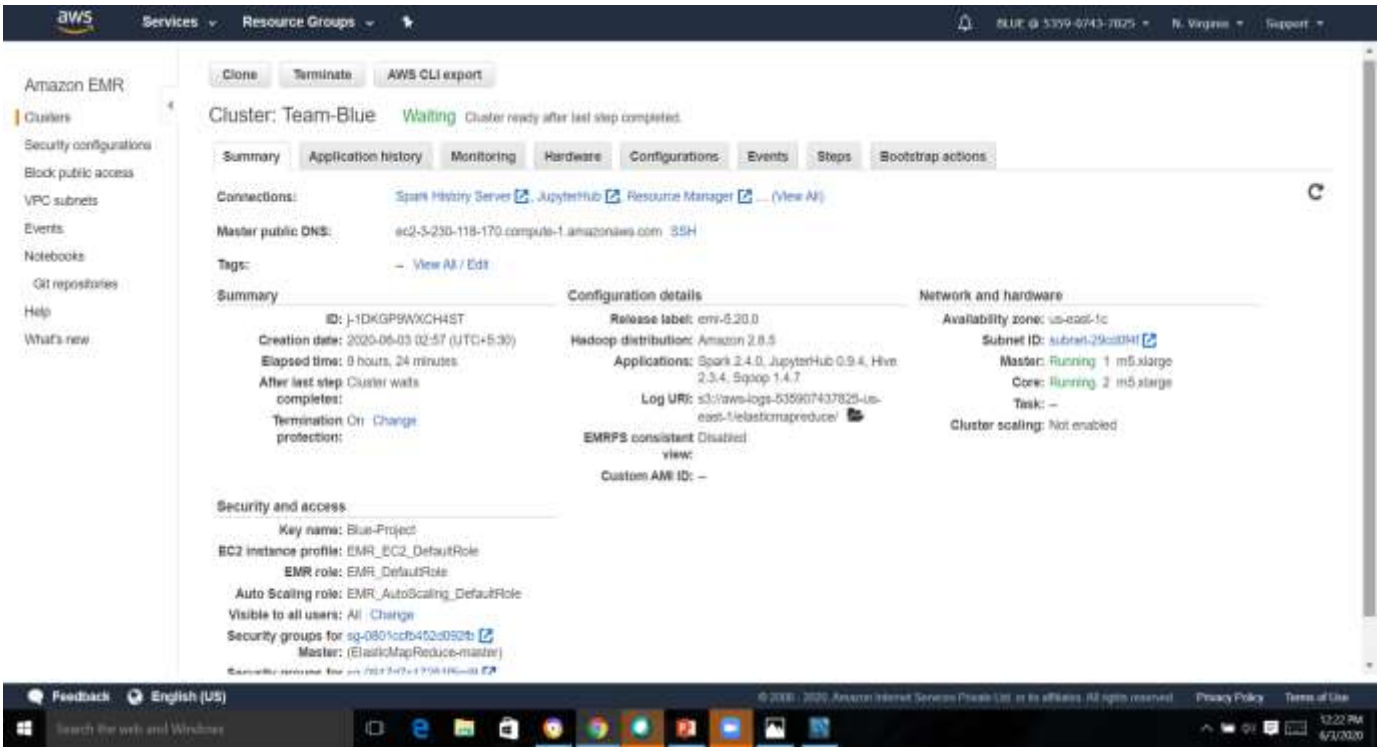

Fig 12. AWS EMR cluster

6.3.5 Connecting the database

Logged in to PuTTY using the IP address generated in EC2 instance.

Loaded the table in mysql.

Connected the database with hive using the sqoop command.

- *Sqoop import –connect jdbc:mysql://34.204.91.255/teamblue*
- *–-username blueteam -–password blue*
- *--split-by FlightNum*
- *–table record*
- *–target-dir /user/hadoop/blueteam/*
- *--hive-import*
- *–hive-table record*

Using the Jupyter link present in the Amazon EMR cluster or in PuTTY itself, ran the PySpark commands to get the final output in JSON format.

- *from pyspark.sql import HiveContext*
- *from pyspark.context import SparkContext*
- *sc = SparkContext.getOrCreate ("local", "MyApp")*
- *com = HiveContext(sc)*
- *df = com.sql(“ select * from record limit 10 ")*

6.3.6 Queries in JSON

1. Find the most Frequent tail number which is getting to the destination by maximum.

   Select Dest,TailNum,count(tailnum) as Total_count from project.record_2007 where Tail Num != 'NA' and TailNum NOT LIKE '0%' group by Dest, TailNum order by Total_coun t desc,dest

Fig 13. Output in JSON format

2. Find out the cancelled flight details for the last quarter.

   Select month,TailNum,FlightNum from project.record_2007 where Cancelled = 1 AND month In (10,11,12) group by month,FLightnum,TailNum order by month.

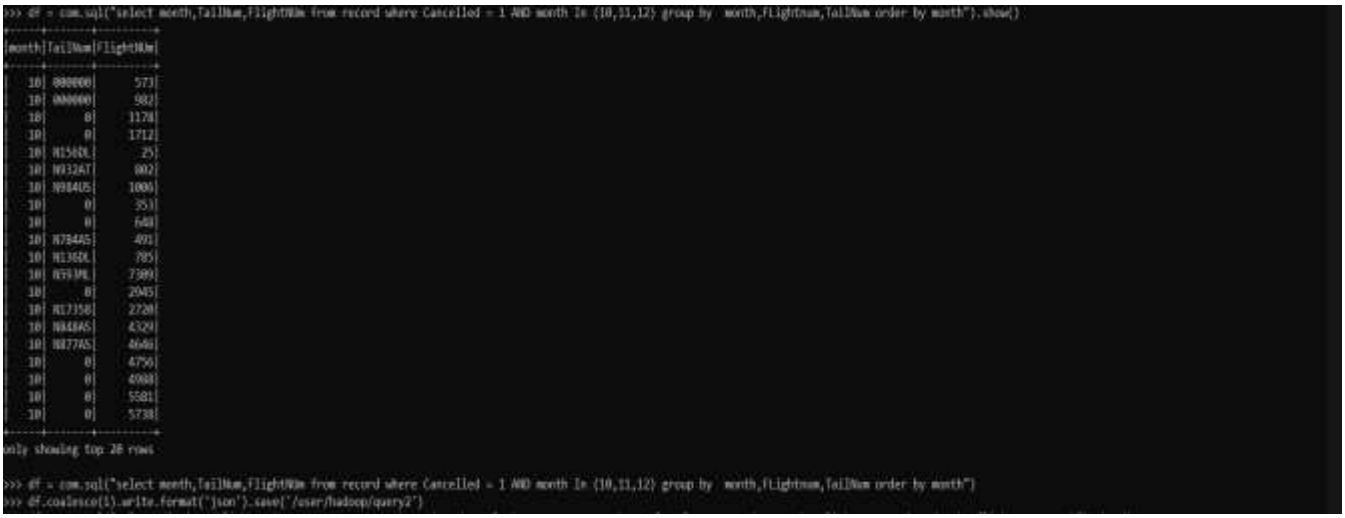

Fig 14. Output in JSON format

3. Find out the average weather delays for a particular flight per month.
   Select distinct(FlightNum) as FLIGHT_NUM,month,avg(WeatherDelay) as Average_We ather_Delay from project.record_2007 group by FlightNum,Month order by flight_num, month.

Fig 15. Output in JSON

4. Monthwise total distance travelled by each flight number.
   Select distinct FlightNum,sum(Distance) as SUM_DISTANCE,Month from project.record_2007 group by Month,FlightNum.

Fig 16. Output in JSON

5. Month-wise percentage of cancellation reason.

Fig 17. Output in JSON

6. Which flights covered maximum origin and destination by month wise.

Fig 18. Output in JSON

SUMMARY AND CONCLUSIONS

The data is currently being generated at massive amounts by the numerous social activities and it has never been so big, and the speed of this data generation process is also ever increasing. This trend of Big Data is now being considered by organizations as an instrument to gain a competitive advantage over others. If they are able to make sensible inferences out of the information contained in the data before others do, then they will be able to captivate more customers, gain more from each customer, optimise their operation, and decrease their costs.

But, analytics of Big Data remains a very exigent and time consuming task that needs costly software, huge infrastructure for computation, and huge efforts. This is where Cloud computing plays a helpful role in combating such issues, by providing resources according to the demand and with costs relative to the actual usage. In addition to this, it enables the software infrastructure to be levelled up and down rapidly as need be, adapting the system to the changing demand.

In this report, we discussed the key stages of analytics workflows and surveyed the nuances of every stage. We studied about various key services offered by Amazon Web Services and put some services to actual use in our ETL project. To conclude, the field of data analysis is ever changing and the emergence of cloud computing technologies and service providers such as AWS are pushing the limits in this field of technology, enabling easy and cost effective analysis and benefitting numerous organizations to attract more customers.

REFERENCES

[1] F. Schomm, F. Stahl, G. Vossen, Marketplaces for data: An initial survey, SIGMOD Record 42 (1) (2013) 15–26.

[2] P.S. Yu, On mining big data, in: J. Wang, H. Xiong, Y. Ishikawa, J. Xu, J. Zhou (Eds.), Web-AgeInformation Management, in: Lecture Notes in Computer Science, vol. 7923, Springer-Verlag, Berlin, Heidelberg, 2013, p. XIV.

[3] X. Sun, B. Gao, Y. Zhang, W. An, H. Cao, C. Guo, W. Sun, Towards delivering analytical solutions in cloud: Business models and technical challenges, in: Proceedings of the IEEE 8th International Conference on e-Business Engineering (ICEBE 2011), IEEE Computer Society, Washington, USA, 2011, pp. 347–351.

[4] Hadoop.apache.org. 2020. *Apache Hadoop*. [online]
Available at: <http://hadoop.apache.org/> [Accessed 16 June 2020].

[5] Hadoop.apache.org.2020.*HDFS Architecture Guide*. [online]
Available at: <https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html#Introduction>
[Accessed 16 June 2020].

[6] Hadoop.apache.org. 2020. [online] Available at:Hadoop.apache.org. 2020. [online] Available at: <https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html#Purpose>
[Accessed 16 June 2020].

[7] Sqoop.apache.org. 2020. *Apache Sqoop Documentation — Apache Sqoop Documentation*. [online]
Available at: <https://sqoop.apache.org/docs/1.99.7/index.html> [Accessed 16 June 2020].

[8] Docs.aws.amazon.com. 2020. *What Is IAM? - AWS Identity And Access Management*. [online] Available at: <https://docs.aws.amazon.com/IAM/latest/UserGuide/introduction.html>
[Accessed 16 June 2020].

[9] Docs.aws.amazon.com. 2020. *What Is Amazon EC2? - Amazon Elastic Compute Cloud*. [online] Available at: <https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts.html>
[Accessed 16 June 2020].

[10] Docs.aws.amazon.com. 2020. *Amazon Machine Images (AMI) - Amazon Elastic Compute Cloud*. [online]
Available at: <https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/AMIs.html>
[Accessed 16 June 2020].

[11] Docs.aws.amazon.com. 2020. *Getting Started With Amazon Simple Storage Service- Amazon Simple storage service.* [online]
Available at: <https://docs.aws.amazon.com/AmazonS3/latest/gsg/GetStartedWithS3.html>
[Accessed 16 June 2020].

[12] Docs.aws.amazon.com. 2020. *What Is Amazon VPC? - Amazon Virtual Private Cloud*. [online]
Available at: <https://docs.aws.amazon.com/vpc/latest/userguide/what-is-amazon-vpc.html>
[Accessed 16 June 2020].

[13] Docs.aws.amazon.com. 2020. *Amazon Redshift Management Overview - Amazon Redshift*. [online]
Available at: <https://docs.aws.amazon.com/redshift/latest/mgmt/overview.html>
[Accessed 16 June 2020].

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

PLAGIARISM VERIFICATION REPORT

Date: ....19/07/2020.....

Type of Document (Tick): **PhD** | **M.Tech Dissertation/** | **B.Tech Project** ✔ | **Pape**

Name: Tanya Singh **Department:** Bioinformatics **Enrolment No** 161502

Contact No. 9418651713 **E-mail.** tanyasingh3@hotmail.com

Name of the Supervisor: Dr. Raj Kumar

Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters):

EMERGING TRENDS IN DATA ENGINEERING: AMAZON WEB SERVICES

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

Complete Thesis/Report Pages Detail:

- Total No. of Pages = 44
- Total No. of Preliminary pages = 7
- Total No. of pages accommodate bibliography/references = 2

(Signature of Student)

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found **Similarity Index** at..... (%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

(Signature of Guide/Supervisor)  **Signature of HOD**

FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

<table>
<tr><td>Copy Received on</td><td rowspan="4">• All Preliminary Pages
• Bibliography/Im a ges/Quotes
• 14 Words String</td><td rowspan="2">Similarity Index (%)
7 %</td><td>Word Counts</td><td></td></tr>
<tr><td>Report Generated on</td><td rowspan="2">Character Counts</td><td rowspan="2"></td></tr>
<tr><td rowspan="2"></td><td rowspan="2">Submission ID</td></tr>
<tr><td>Total Pages Scanned</td><td></td></tr>
<tr><td></td><td></td><td></td><td>File Size</td><td></td></tr>
</table>

Checked by
Name & Signature **Librarian**

.....

Please send your complete thesis/report in (PDF) with Title Page, Abstract and Chapters in (Word File) through the supervisor at plagcheck.juit@gmail.com