A Major Project Synopsis on

**Job Vacancy Analysis in New York City**

Submitted to Manipal University, Jaipur

Towards the partial fulfillment for the Award of the Degree of

**BACHELORS OF TECHNOLOGY**

In Computers Science and Engineering

# 2016-2020

By

# Rohan Chopra

# (169105152)



Under the guidance of

# Mr. Rohit Verma

**Department of Computer Science and Engineering School of Computing and Information Technology Manipal University Jaipur**

**Jaipur, Rajasthan**

**Introduction**

Today, two mainstream technologies are the center of concern in IT – Big Data and Cloud Computing. Fundamentally different, Big data is all about dealing with the massive scale of data whereas Cloud computing is about infrastructure. However, the simplification offered by Big data and Cloud technology is the main reason for their huge enterprise adoption.

Big data deals with massive structured, semi-structured or unstructured data to store and process it for data analysis purpose. In this project, our aim is to analyze the job vacancies in New York City. We have been provided with a dataset that announces job vacancies in various department of civil services in different locations of New York. The dataset is so large and complex that it cannot be dealt by using traditional data-processing application software. Thus, certain operations will be performed on data source to extract, load and transform data which can offer greater statistical power, easy to visualize and analyze yielding high value returns.

Moreover, the focus of this project is not only to manipulate big data but also bring in the use of Cloud Computing. It refers to the on-demand delivery of IT resources and applications via the internet. With cloud computing instead of having to design and build our own data centers, we access a data center and all of its resources, via the internet – allowing us to scale up or down based on our actual needs. Cloud computing offers services to the users on a pay-as-you-go model.

Through hardware virtualization, cloud computing provides the option of storing significant amounts of data with the help of scalability, fault tolerance and availability. This allows Big Data to be available, scalable and fault tolerant through cloud computing. Instances can be created on cloud to operate on Big Data and user will be charged only when the instance is in operation. The greatest benefit of Cloud Computing is that it provides a great deal of benefit to the organizations bring on the use of cloud technology.

**Motivation**

The motivation for doing this project was primarily an interest in undertaking a challenging project in an interesting area of research. The opportunity to learn about a new area of computing not covered in lectures was appealing.

Big data involves manipulating petabytes (and perhaps soon, exabytes and zettabytes) of data, and the cloud’s scalable environment makes it possible to deploy data-intensive applications that power business analytics. The cloud also simplifies connectivity and collaboration within an organization, which gives more employees access to relevant analytics and streamlines data sharing.

This area is possibly an area that I might study at postgraduate level.

**Statement of Problem**

The goal of this project is to gain knowledge of big data and cloud computing both of which are valuable in its own and how to harness big data in cloud. Its objectives are:

* To compare different Jobs in different Cities from the given data by performing ETL operations.
* Helps the candidates with different choices of jobs in different cities based on their various criteria.
* Help students to identify in which there is higher chances to get a job.
* To bring in the use of Cloud Computing by using AWS Web Services
* Performing Queries on Amazon Athena and storing results generated in Amazon S3 Bucket
* Learning about visualizing data on Amazon Quicksight.

**Methodology**

We need to develop the job vacancy analysis by building the functionality incrementally in each of the applications such as MySQL, Hive, Spark using ETL(Extract, Transform, Load) tools such as SQOOP.

Project Flow:

**Results generated using**

**Different tools**

**Data**

**Warehouse**

**Extract, Transform and Load**

**Heterogeneous Sources i.e. Operational Data (Flat Files)**

Steps to be followed to achieve the project objectives are:

1. Load data into relational database such as MySQL
2. Creating table in Hive using SerDe(Serializer/Deserializer) properties
3. Transfer data into data warehouse such as HDFS(Hadoop File System) using SQOOP(ETL Tool)
4. Clean the data by handling NULL/Blank Values using HiveQL(Hive Query Language)
5. Running queries on SPARK for better performance
6. Storing data in Amazon S3
7. Performing queries on Amazon Athena

**Facilities Required**

**Minimum System Requirement:**

1. Intel Pentium 90 or higher (P166 recommended)
2. Microsoft Windows 2010 or above
3. Memory: 8GB of RAM or more
4. Internet Explorer 10.0 or higher
5. MySQL
6. HIVE
7. BEELINE
8. SQOOP (ETL Tool)
9. Apache SPARK

**Bibliography and References**

1. Capgemini Material and Resources Only.
2. AWS Training
3. Plurasight AWS Training