

New York Job Vacancy Analysis
A PROJECT REPORT
Submitted in partial fulfilment of the
requirement for the award of the degree
of
BACHELOR OF TECHNOLOGY (B.Tech)
in
Computer Science Engineering
by
Romil Nagar
169105154



**MANIPAL UNIVERSITY
JAIPUR**

Department of Computer Science & Engineering,
School of Computing and IT,
MANIPAL UNIVERSITY JAIPUR
JAIPUR-303007
RAJASTHAN, INDIA

May/2020

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
MANIPAL UNIVERSITY JAIPUR, JAIPUR – 303 007 (RAJASTHAN), INDIA

CERTIFICATE

This is to certify that the project titled **New York Job Vacancy Analysis** is a record of the bonafide work done by **Romil Nagar** (169105154) submitted in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology (B.Tech) in Computer Science and Engineering of Manipal University Jaipur, during the academic year 2019-20.

Prof. Rohit Verma

*Project Guide, Dept. of Computer Science and
Engineering Manipal University Jaipur*

Dr. Sandeep Joshi

*HOD, Dept. of Computer Science and
Engineering Manipal University Jaipur*

(On company letterhead)

Date:28/04/2020

CERTIFICATE

This is to certify that the project entitled **NEW YORK JOB VACANCY ANALYSIS** was carried out by **ROMIL NAGAR** (169105154) at **CAPGEMINI, PUNE** under my guidance during **Starting Date (23\01\2020), 2020** to **Ending Date (27\04\2020)**.

Ms. Radhika MohanRaj
Senior Analyst,
Capgemini, Pune

ACKNOWLEDGMENTS

I would like to express my special thanks of gratitude to Capgemini who gave me the golden opportunity to learn this great technology and do this wonderful project on the Big Data topic, I came to know about so many versatile aspects and I am really thankful to them. I take this opportunity to thank all those magnanimous persons who rendered their full services to my work.

It's with lot of happiness we are expressing gratitude to our guide **Mr. Rohit Verma** B.Tech, Assistant professor, Computer Science and Engineering, for her timely and kind help, guidance and for providing us with most essential materials required for the completion of the completion of this project. We are very thankful to him for his indomitable guidance. This inspiration up to the last moment had made things possible in a nice manner.

Finally, I thank each and every one who helped to complete my project work with their cordial support.

ABSTRACT

In Job Vacancy scenario the main strategic goal is to find the available job openings for the un- employed public in different agencies and department and find the relevant job for them among different job categories. During the last years Big Data has become the buzzword across various industries, but it is difficult to know exactly what Big Data can do to improve business value and which Big Data applications marketers should consider to invest both their time and money in. Our goal in this paper is to show a comprehensive list of data-driven use cases and their value, which are deployed by successful marketing teams today. Each use case is chosen for its relevancy to the job sector and is backed by case studies from real organizations.

To find the results to the chosen use cases the main tasks were divided into three stages

1. Finding Data
2. Cleaning and Manipulating Data
3. Analyzing the Data
4. Visualizing the Data

Data can't be used directly from the files as the data was having many missing values as well as the data was not in standard format that can be directly analyzed by spark. So in cleaning stage data was transformed and cleaned. After that in analysis stage query are run to find the results to the cases pre-defined. After that output are properly displayed so that it can be presented and can be understood by the bank. Query analysis on the data helps us to draw insights from it so that it can be beneficial for Descriptive analysis as well as predictions. The results from these cases can be used to find the vacancies in different sectors. These case studies also display the preferred skills as well as job description which can be used to fetch the required skills for a particular job.

Tools which were used in different stages are

1. **Cleaning Data:** - Hadoop is used to clean the data.
2. **Data Ingestion:** - Sqoop was used to ingest the data.
3. **Analyzing Stage:** - Hadoop and Apache Spark were used to run queries to the test cases.
4. **Visualizing Stage:** - Tableau is used to visualize the data.

In this project, we started learning different technologies required to build the project like python, java, OracleSQL, Unix, Hadoop, Spark, etc. After that we used all those technologies to build a project which was based on job vacancy analysis. The whole big data project consists of analyzing the data, handling null values in it, using data ingestion tools to import it into Hadoop.

Moreover, we took the data and build a database in MySQL and tried to handle the null values in MySQL using case statements. After that it was imported into hive using a data ingestion tool called Sqoop. In hive we did the query analysis to draw useful information and creating a solution for our problem statement. At last we used tableau to generate graphs and charts for visualizing the data.



Capgemini Technology Services India Limited
(Formerly known as IGATE Global Solutions Limited)
No.158-162P & 165-170P, EPIP Phase II, Whitefield,
Bengaluru – 560066, Karnataka India.
T: +91-80-4104-0000 | F: +91-80-4125-9090
www.in.capgemini.com

Capgemini/HR/HM/AG/TA/SJ

May 27, 2020

TO WHOMSOEVER IT MAY CONCERN

This is to certify **Mr. Romil Nagar** has successfully completed his internship in "**Business Intelligence**" from **January 23, 2020** to **March 31, 2020**

He interned under the guidance of "**Ms. Radhika Mohanraj**".

We wish him all the best for his future endeavors.

Yours truly,
For **Capgemini Technology Services India Limited**

Signature Not Verified
Digitally signed by HRUSHIKESH
MANGALAMPALLI
Date: 2020.05.27 12:08:03 +05:30

*This is a digitally signed document and does not require any signatures on it.

Regd. Off.:No.14, Rajiv Gandhi Infotech Park, Hinjawadi Phase III,
MIDC – Sez, Village Man, Taluka Mulshi, Pune – 411057, Maharashtra, India.
Tel: +91.20.6699 1000 | Fax:+91.20.6699 5050 | CIN: U85110PN1993PLC145950

LIST OF FIGURES

Figure No	Figure Title	Page No
1.	Report Organization	4
2.	MySQL Logo	9
3.	Sqoop Logo	10
4.	Hadoop Logo	11
5.	Apache Spark Logo	12
6.	Tableau Logo	13
7.	Map Reduce Structure	15
8.	Spark RDD Lineage Graph	16
9.	Project Flow	17
10.	Project Structure	17
11.	Table Schema	18
12.	Dataset Capture	19
13.	Dataset '@' Separated	20
14.	Dataset ' ' Separated	20
15.	Schema Creation	21
16.	Query 1 Output	23
17.	Most In Demand Job Vacancies	24
18.	Query 2 Output	25
19.	No. of External and Internal jobs	26
20.	Query 3 Output	27
21.	No. of part time and full time job vacancies	28
22.	Query 4 Output	29
23.	Query 5 Output	30
24.	Query 6 Output	31
25.	February and March job postings	32
26.	April and May job postings	32
27.	May and June Job Posting	33
28.	Query 7 Output	34
29.	Query 8 Output	35
30.	Time taken in job application	36
31.	Query 9 Output	37

Contents		
		Page No
Acknowledgement		iv
Abstract		v
List Of Figures		vi
Chapter 1	INTRODUCTION	1
1.1	Overview	1
1.2	Motivation	2
1.3	Project Statement	3
1.4	Organization of Report	4
1.5	Minimum System Requirements	4
Chapter 2	BACKGROUND OVERVIEW	5
2.1	Conceptual Overview (<i>Concepts/ Theory used</i>)	5
2.2	Technologies Involved	9
Chapter 3	METHODOLOGY	14
3.1	System Architecture	14
3.2	Algorithm and Techniques	15
3.3	Overall Structure	17
Chapter 4	IMPLEMENTATION	18
4.1	Input Stage	18
4.2	Cleaning Stage	20
4.3	Modeling Stage	21
4.4	Retrieving Stage	22
Chapter 5	RESULTS AND ANALYSIS	23
5.1	Case Study 1	23
5.1	Case Study 2	25
5.1	Case Study 3	27
5.1	Case Study 4	29
5.1	Case Study 5	30
5.1	Case Study 6	31
5.1	Case Study 7	34
5.1	Case Study 8	35
5.1	Case Study 9	37
Last Chapter	CONCLUSIONS & FUTURE SCOPE	38
6.1	Conclusions	38
6.2	Future Scope of Work	38
REFERENCES		39

1. INTRODUCTION

1.1 Overview

According to the study by IDC, the worldwide revenue for big data and business analytics solutions is expected to reach \$260 billion by 2022. This year, the projected numbers will hit \$166 billion, up 11.7% compared to 2018. It comes as no surprise that banking is one of the business domains that makes the highest investment in big data and BA technologies.

Big Data refers to all the data that is being generated across the globe at an unprecedented rate. This data could be either structured or unstructured. Today's business enterprises owe a huge part of their success to an economy that is firmly knowledge-oriented. Data drives the modern organizations of the world and hence making sense of this data and unraveling the various patterns and revealing unseen connections within the vast sea of data becomes critical and a hugely rewarding endeavor indeed. There is a need to convert Big Data into Business Intelligence that enterprises can readily deploy. Better data leads to better decision making and an improved way to strategize for organizations regardless of their size, geography, market share, customer segmentation and such other categorizations. Hadoop is the platform of choice for working with extremely large volumes of data.

Some of the use cases of big data are:

1. **Personalized customer experience:** According to Oracle, 84% of the surveyed executives agree that customers are looking for a more individualized, tailored experience. Your data can give you valuable insights into user behavior and help you optimize your customer experience accordingly. For example, by having a complete customer profile and exhaustive data on product engagement at hand, you can predict and prevent churn.
2. **User segmentation and targeting:** McKinsey finds that using data to make better decisions can save up to 15-20% of your marketing budget. Taking into account that banks spend on average 8% of their overall budgets on marketing, tapping into big data sounds like a great opportunity to not only save, but generate additional revenue through highly targeted marketing strategies.
3. **Business process optimization and automation:** JP Morgan Chase & Co. is one of the automation pioneers in the banking services industry. The company currently employs several artificial intelligence and machine learning programs to optimize some of their processes, including algorithmic trading and commercial-loan agreements interpretation.

1.2 Motivation

The computing revolution that began more than 2 decades ago has led to large amounts of digital data being amassed by corporations. Advances in digital sensors; proliferation of communication systems, especially mobile platforms and devices; massive scale logging of system events; and rapid movement toward paperless organizations have led to a massive collection of data resources within organizations. And the increasing dependence of businesses on technology ensures that the data will continue to grow at an even faster rate.

Having the ability to drill down through huge amounts of data to find the really powerful metrics and nuggets of information is the core purpose of big data and is also a key tool for employees to be able to track their progress. The success of ‘increasing website visitors’ or ‘increasing revenue’ comes from the success of the elements that make up the larger goals, which are in turn made from smaller metrics. Being able to accurately track these at a macro level gives genuine goals that can be achieved daily, creating almost a gamification model and further increasing motivation to meet long term, larger goals.

1.3 Project Statement

There are many useful cases that can be created using this data which can be used by the companies to segment the customers as well as find their trends. These cases can also help to optimize their resources according to the customer's behavior. So here are some of the important cases

The aim of the project is to find results of the following cases using big data technologies

1. Most in demand job vacancies in each Agency
2. Number of external and internal jobs vacancy in each Agency.
3. Determine number of part time and full time jobs vacancies in different job categories.
4. Display different jobs with salary ranges
5. Agency and Job Category having Highest paying jobs
6. Month wise number of job posted in all agencies.
7. Most popular preferred skills per job category
8. Days taken in job application for each work unit
9. Number of job positions in each work location

1.4 Organization of Report

The report is divided into three main parts

1. **Methodology:** This covers the flow chosen to achieve the desired result. This section covers the technologies used in different sections such as technologies used in data cleaning, data moving, analysis of the data. It also covers the details of the software used and version used during creation of this project.
1. **Implementation:** This section covers the algorithms and approaches used in this project. This section consists of the details of the data moment from cleaning to data manipulation to analysis. It also contains the techniques used in different tasks like how cleaning is done? Why there is need to clean and manipulate the data? How data is modeled to run the appropriate queries to find the result of the cases.
2. **Result and Analysis:** This section contains the final output of the project and performing the descriptive analysis on the project to draw insights, useful information and creating a solution for our problem statement. It also contains the bar graph and diagrams for the visualization of output of the cases mentioned in the aim of project.

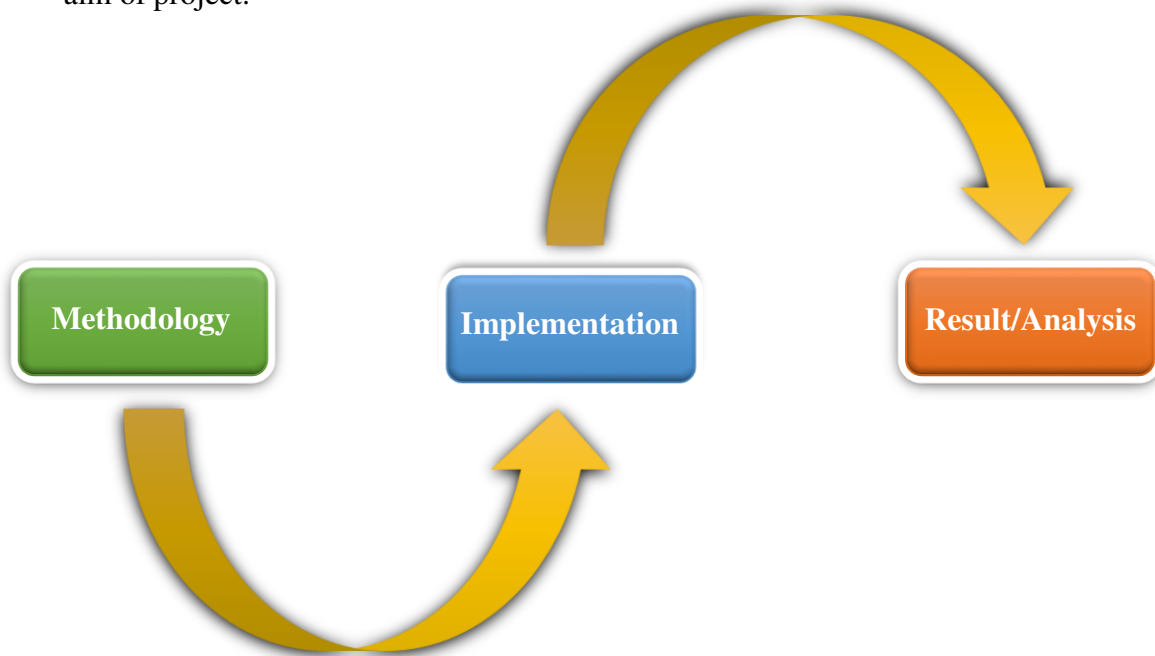


Figure 1: Report Organizations

1.5 Minimum System Requirements

- Intel Pentium 90 or higher (P166 recommended)
- Microsoft Windows 2010 or above
- Memory: 8GB of RAM (4GB or more recommended)
- Internet Explorer 10.0 or higher
- Oracle 11g / MySQL/ HIVE/Beeline
- SQOOP (ETL Tool)

2. BACKGROUND OVERVIEW

2.1 Conceptual Overview

BIG DATA is a blanket term for the non-traditional strategies and technologies needed to gather, organize, process, and analyze insights from large datasets. While the problem of working with data that exceeds the computing power or storage of a single computer is not new, the pervasiveness, scale and value of this type of computing has greatly expanded in recent years.

What Is Big Data?

An exact definition of "big data" is difficult to nail down because projects, vendors, practitioners and business professionals use it quite differently. With that in mind, generally speaking, big data is:

1. Large datasets
2. The category of computing strategies and technologies that are used to handle large datasets

In 2001, Gartner's Doug Laney first presented what became known as the "three Vs of big data" to describe some of the characteristics that make big data different from other data processing:

Volume

These datasets can be orders of magnitude larger than traditional datasets, which demands more thought at each stage of the processing and storage life cycle.

Velocity

Another way in which big data differs significantly from other data systems is the speed that information moves through the system. Data is frequently flowing into the system from multiple sources and is often expected to be processed in real time to gain insights and update the current understanding of the system.

Variety

Data can be ingested from internal systems like application and server logs, from social media feeds and other external APIs, from physical device sensors, and from other providers. Big data seeks to handle potentially useful data regardless of where it's coming from by consolidating all information into a single system.

The formats and types of media can vary significantly as well. Rich media like images, video files and audio recordings are ingested alongside text files, structured logs, etc. While more traditional data processing systems might expect data to enter the pipeline already labeled, formatted, and organized, big data systems usually accept and store data closer to its raw state. Ideally, any transformations or changes to the raw data will happen in memory at the time of processing.

What Does a Big Data Life Cycle Look Like?

So how is 'data' actually processed when dealing with a big data system? While approaches to implementation differ, there are some commonalities in the strategies and software that we can talk about generally. While the steps presented below might not be true in all cases, they are widely used.

The general categories of activities involved with big data processing are:

1. Ingesting data into the system
2. Persisting the data in storage
3. Computing and Analyzing data
4. Visualizing the results

Before we look at these four workflow categories in detail, we will take a moment to talk about **clustered computing**, an important strategy employed by most big data solutions. Setting up a computing cluster is often the foundation for technology used in each of the life cycle stages.

Ingesting Data into the System

Data ingestion is the process of taking raw data and adding it to the system. The complexity of this operation depends heavily on the format and quality of the data sources and how far the data is from the desired state prior to processing. Technologies like **Apache Sqoop** can take existing data from relational databases and add it to a big data system. Similarly, **Apache Flume** and **Apache Chukwa** are projects designed to aggregate and import application and server logs.

Persisting the Data in Storage

The ingestion processes typically hand the data off to the components that manage storage, so that it can be reliably persisted to disk. While this seems like it would be a simple operation, the volume of incoming data, the requirements for availability, and the distributed computing layer make more complex storage systems necessary. This usually means leveraging a distributed file system for raw data storage. Solutions like **Apache Hadoop's HDFS** file system allow large quantities of data to be written across multiple nodes in the cluster. This ensures that the data can be accessed by compute resources, can be loaded into the clusters RAM for in-memory operations, and can gracefully handle component failures. Data can also be imported into other distributed systems for more structured access. Distributed databases, especially NoSQL databases, are well-suited for this role because they are often designed with the same fault tolerant considerations and can handle heterogeneous data.

Computing and Analyzing Data

Once the data is available, the system can begin processing the data to surface actual information. The computation layer is perhaps the most diverse part of the system as the requirements and best approach can vary significantly depending on what type of insights desired. Data is often processed repeatedly, either iteratively by a single tool or by using a number of tools to surface different types of insights.

While batch processing is a good fit for certain types of data and computation, other workloads require more **real time processing**. Real time processing demands that information be processed and made ready immediately and requires the system to react as new information becomes available. One way of achieving this is **stream processing**, which operates on a continuous stream of data composed of individual items.

Another common characteristic of real-time processors is in-memory computing, which works with representations of the data in the clusters memory to avoid having to write back to disk.

Apache Storm, Apache Flink, and Apache Spark provide different ways of achieving real-time or near real-time processing. There are trade-offs with each of these technologies, which can affect which approach is best for any individual problem. In general, real-time processing is best suited for analyzing smaller chunks of data that are changing or being added to the system rapidly.

Visualizing the results

Due to the type of information being processed in Big Data system, recognizing trends or changes in data over time is often more important than the values themselves. Visualizing data is one of the most useful way to spot trends and make sense of a large number of data points.

One popular way of visualizing data is with Tableau Studio:

Tableau is greatly used because data can be analyzed very quickly with it. Also, visualizations are generated as dashboards and worksheets. Tableau allows one to create dashboards that provide actionable insights and drives the business forward. Tableau products always operate in virtualized environments when they are configured with the proper underlying operating system and hardware. Tableau is used to explore data with limitless visual analytics.

Tableau software is used to translate queries into visualization. It is also used for managing metadata. Tableau software imports data of all sizes and ranges. For a non-technical user, Tableau is a life saver as it offers the facility to create ‘no-code’ data queries.

Big Data Glossary

Some of key Big data concepts are as follows that are used in this project:

1. **Big data:** Big data is an umbrella term for datasets that cannot reasonably be handled by traditional computers or tools due to their volume, velocity, and variety. This term is also typically applied to technologies and strategies to work with this type of data.
2. **Cluster computing:** Clustered computing is the practice of pooling the resources of multiple machines and managing their collective capabilities to complete tasks. Computer clusters require a cluster management layer which handles communication between the individual nodes and coordinates work assignment.
3. **Data lake:** Data lake is a term for a large repository of collected data in a relatively raw state. This is frequently used to refer to the data collected in a big data system which might be unstructured and frequently changing. This differs in spirit to data warehouses (defined below).
4. **Data mining:** Data mining is a broad term for the practice of trying to find patterns in large sets of data. It is the process of trying to refine a mass of data into a more understandable and cohesive set of information.
5. **Data warehouse:** Data warehouses are large, ordered repositories of data that can be used for analysis and reporting. In contrast to a data lake, a data warehouse is composed of data that has been cleaned, integrated with other sources, and is generally well ordered. Data warehouses are often spoken about in relation to big data, but typically are components of more conventional systems.
6. **ETL:** ETL stands for extract, transform, and load. It refers to the process of taking raw data and preparing it for the system's use. This is traditionally a process associated with data warehouses, but characteristics of this process are also found in the ingestion pipelines of big data systems.
7. **Hadoop:** Hadoop is an Apache project that was the early open source success in big data. It consists of a distributed file system called HDFS, with a cluster management and resource scheduler on top called YARN (Yet Another Resource Negotiator). Batch processing capabilities are provided by the MapReduce computation engine. Other computational and analysis systems can be run alongside MapReduce in modern Hadoop deployments.
8. **In-memory computing:** In memory computing is a strategy that involves moving the working datasets entirely within a cluster's collective memory. Intermediate calculations are not written to disk and are instead held in memory. This gives in-memory computing systems like Apache Spark a huge advantage in speed over i/o bound systems like Hadoop's MapReduce.
9. **Map reduce (big data algorithm):** Map reduce (the big data algorithm, not Hadoop's MapReduce computation engine) is an algorithm for scheduling work on a computing cluster. The process involves splitting the problem set up (mapping it to different nodes) and computing over them to produce intermediate results, shuffling the results to align like sets, and then reducing the results by outputting a single value for each set.

2.2 Technologies Involved

MySQL

The MySQL™ software delivers a very fast, multithreaded, multi-user, and robust SQL (Structured Query Language) database server. MySQL Server is intended for mission critical, heavy-load production systems as well as for embedding into mass-deployed software.

A database is a structured collection of data. It may be anything from a simple shopping list to a picture gallery or the vast amounts of information in a corporate network. To add, access, and process data stored in a computer database, you need a database management system such as MySQL Server. Since computers are very good at handling large amounts of data, database management systems play a central role in computing, as standalone utilities, or as parts of other applications.

MySQL Server can run comfortably on a desktop or laptop, alongside your other applications, web servers, and so on, requiring little or no attention. If you dedicate an entire machine to MySQL, you can adjust the settings to take advantage of all the memory, CPU power, and I/O capacity available. MySQL can also scale up to clusters of machines, networked together.

MySQL Server was originally developed to handle large databases much faster than existing solutions and has been successfully used in highly demanding production environments for several years. Although under constant development, MySQL Server today offers a rich and useful set of functions. Its connectivity, speed, and security make MySQL Server highly suited for accessing databases on the Internet.

The MySQL Database Software is a client/server system that consists of a multithreaded SQL server that supports different back ends, several different client programs and libraries, administrative tools, and a wide range of application programming interfaces (APIs).



Figure 2: MySQL Logo

Sqoop

Apache Sqoop, short for “SQL to Hadoop,” was created to perform bidirectional data transfer between Hadoop and almost any external structured data store. Taking advantage of MapReduce, Hadoop’s execution engine, Sqoop performs the transfers in a parallel manner.

Sqoop supports the Linux operating system, and there are several installation options. One option is the source tarball that is provided with every release. This tarball contains only the source code of the project. You can’t use it directly and will need to first compile the sources into binary executables. For your convenience, the Sqoop community provides a binary tarball for each major supported version of Hadoop along with the source tarball.

As Sqoop is not a cluster service, you do not need to install it on all the nodes in your cluster. Having the installation available on one single machine is sufficient. As a Hadoop application, Sqoop requires that the Hadoop libraries and configurations be available on the machine.

A significant strength of Sqoop is its ability to work with all major and minor database systems and enterprise data warehouses. To abstract the different behavior of each system, Sqoop introduced the concept of connectors: all database-specific operations are delegated from core Sqoop to the specialized connectors. Sqoop itself bundles many such connectors; you do not need to download anything extra in order to run Sqoop.

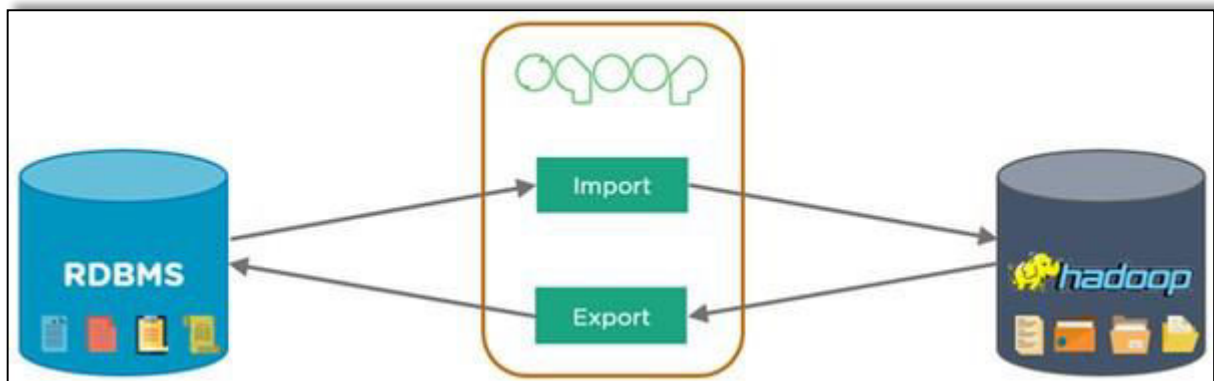


Figure 3: Sqoop Logo

Hadoop

Apache Hadoop is a collection of Open source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model. Originally designed for computer clusters built from commodity hardware still the common use it has also found use on clusters of higher-end hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common occurrences and should be automatically handled by the framework.

The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part which is a MapReduce programming model. Hadoop splits files into large blocks and distributes them across nodes in a cluster. It then transfers packaged code into nodes to process the data in parallel. This approach takes advantage of data locality, where nodes manipulate the data they have access to, this allows the dataset to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are distributed via high-speed networking.



Figure 4: Hadoop Logo

Apache Spark

Apache Spark is an open-source distributed general—purpose cluster-computing framework. Spark provides an interface for programming entire clusters with implicit data parallelism and faulttolerance.

Apache Spark has as its architectural foundation the resilient distributed dataset (RDD), a read-only multiset of data items distributed over a cluster of machines, that is maintained in a fault-tolerant way. The Dataframe API was released as an abstraction on top of the RDD, followed by the Dataset API. In Spark 1.x, the RDD was the primary application programming interface (API), but as of Spark 2.x use of the Dataset API is encouraged even though the RDD API is not deprecated. The RDD technology still underlies the Dataset API.

Spark and its RDDs were developed in 2012 in response to limitations in the MapReduce cluster computing paradigm, which forces a particular linear dataflow structure on distributed programs: MapReduce programs read input data from disk, map a function across the data, reduce the results of the map, and store reduction results on disk. Spark's RDDs function as a working set for distributed programs that offers a (deliberately) restricted form of distributed shared memory.

Spark facilitates the implementation of both iterative algorithms, which visit their data set multiple times in a loop, and interactive/exploratory data analysis, i.e., the repeated database style querying of data. The latency of such applications may be reduced by several orders of magnitude compared to Apache Hadoop MapReduce implementation. Among the class of iterative algorithms are the training algorithms for machine learning systems, which formed the initial impetus for developing Apache Spark.



Figure 5: Apache Spark Logo

Tableau

Tableau is a data visualization, exploration, and analysis tool. It allows you to visualize your data in new and varied ways that enhance your analysis. Sometimes, it tends to bring out the kid in you, making you excited and giddy about the pending eureka moments.

Tableau allows you to connect and mash up your data to see it in different forms and possibilities, such as overlaying weather data onto sales data, or Twitter feed trends combined with survey data. This helps you understand your data, uncover insights, or at least helps you ask the next questions.

Tableau's product line includes desktop design and analysis tools for creating and consuming data. For larger deployments, Tableau Server permits information consumers to access reports in a secure environment without the need to load software. Reports are consumed in Tableau Server via a web- browser. Tableau Server also enables reports to be consumed on iOS or Android tablet computers. Tableau Public is a free tool that facilitates sharing public data on the web via blogs or webpages. For those that want a hosted solution, Tableau Public Premium is a fee-based service that uses the same technology as Tableau Public in a private consumption environment.

Tableau Desktop is the design tool for creating visual analytics and dashboards. There are two versions: Personal Edition and Professional Edition. Professional Edition is more popular because it connects to a wider variety of data sources than Personal Edition. Less common data sources can be accessed via the Open Database Connectivity (ODBC) standard.



Figure 6: Tableau Logo

3. METHODOLOGY

3.1 System Architecture

1. General Structure

1.1 Cleaning

1.1.1 Generating two files consisting of '|' and '@' as delimiters to separate fields.

1.1.2 Replace the null values with 'NA' in fields.

1.2 Creation of schema in MySQL and uploading the data into it

1.3 Ingestion of data in hive

1.4 Query Creation and Execution

1.5 Saving the results

1.6 Running the queries in spark

1.7 Visualization

1.7.1 Creation of Bar charts and diagram.

2. Frameworks

2.1 MySQL

2.2 Sqoop

2.3 Apache Hadoop

2.4 Beeline

2.5 Apache Spark

2.6 Tableau Software

3. Working Environment

3.1 MySQL

3.2 Hive shell

3.3 Spark shell

3.4 Tableau work space

3.2 Algorithms & Techniques

1. Map Reduce

- 1.1. Map reduce is a Framework for processing parallelizable problems across large datasets using a large number of computer (nodes), collectively referred to as a cluster (if all nodes are on the same local network and used similar hardware) or a grid (if the nodes are shared across geographically and administratively distributed Systems, and use more heterogeneous hardware). Processing can occur on data stored either in a file system (unstructured) or in a database (structured). MapReduce can take advantage of the locality of data, processing it near the place it is stored in order to minimize communication overhead.
- 1.2. A MapReduce framework (or system) is usually composed of three operations (or steps):
- 1.3. **Map:** each worker node applies the map function to the local data, and writes the output to temporary storage. A master node ensures that only one copy of redundant input data is processed.
- 1.4. **Shuffle:** worker nodes redistribute data based on the output keys (produced by the map function), such that all data belonging to one key is located on the same worker node
- 1.5. **Reduce:** worker nodes now process each group of output data, per key, in parallel.
- 1.6. MapReduce allows for distributed processing of the map and reduction operations. Maps can be performed in parallel, provided that each mapping operation is independent of the others: in practice, this is limited by the number of independent data sources and/or the number of CPUs near each source. Similarly, a set of 'reducers' can perform the reduction phase, provided that all outputs of the map operation that share the same key are presented to the same reducer at the same time, or that the reduction function is associative. While this process can often appear inefficient compared to algorithms that are more sequential (because multiple instances of the reduction process must be run), MapReduce can be applied to significantly larger datasets than a single "commodity" server, can handle a large server farm can use MapReduce to sort a petabyte of data in only a few hours. The parallelism also offers some possibility of recovering from a partial failure of servers or storage during the operation: if one mapper or reducer fails, the work can be rescheduled assuming the input data are still available.

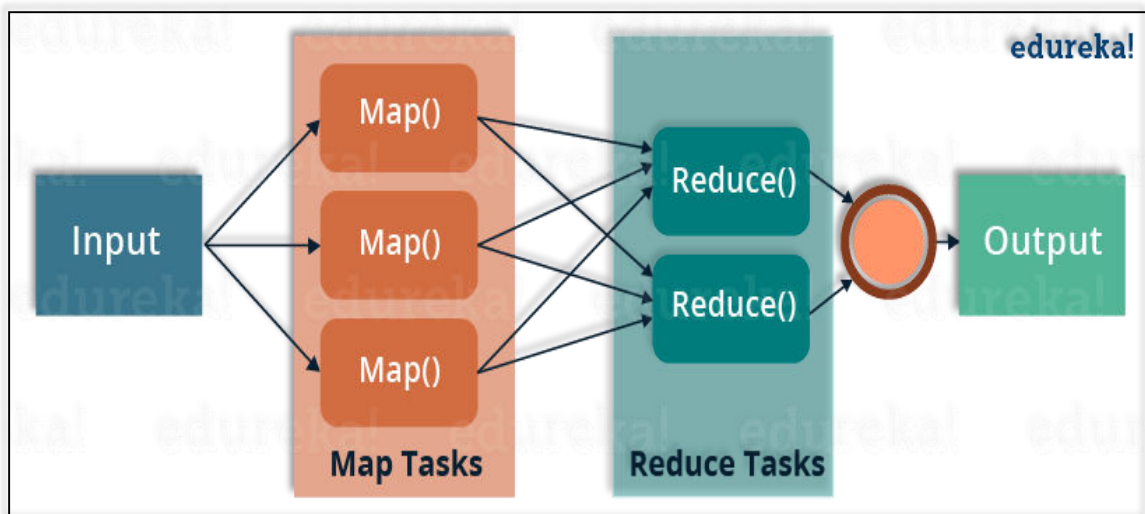


Figure 7: MapReduce Structure

2. Resilient Distributed Datasets

- 2.1 Resilient Distributed Datasets (RDD) is a fundamental data structure of Spark. It is an immutable distributed collection of objects. Each dataset in RDD is divided into logical partitions, which may be computed on different nodes of the cluster. RDDs can contain any type of Python, Java, or Scala objects, including user-defined classes.
- 2.2 Formally, an RDD is a read-only, partitioned collection of records. RDDs can be created through deterministic operations on either data on stable storage or other RDDs. RDD is a fault-tolerant collection of elements that can be operated on in parallel.
- 2.3 There are two ways to create RDDs parallelizing an existing collection in your driver program or referencing a dataset in an external storage system, such as a shared file system, HDFS, HBase, or any data source offering a Hadoop Input Format.
- 2.4 Spark makes use of the concept of RDD to achieve faster and efficient MapReduce operations.
- 2.5 Since RDDs are created over a set of transformations, it logs those transformations, rather than actual data. The graph of transformations to produce one RDD is called a Lineage Graph.
- 2.6 Spark RDD Lineage Graph

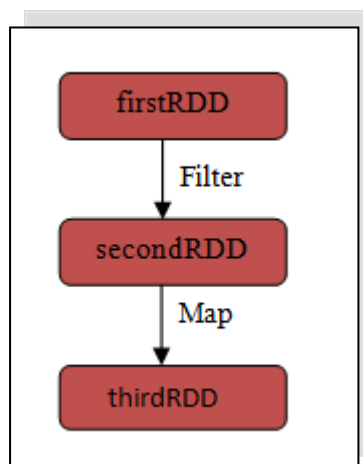


Figure 8: Spark RDD Lineage Graph

- 2.7 In case we lose some partition of RDD, we can replay the transformation on that partition in lineage to achieve the same computation, rather than doing data replication across multiple no. This characteristic is the biggest benefit of RDD because it saves a lot of efforts in data management and replication and thus achieves faster computations.

3.3 Overall Structure

- Database project and table nyc_job was created in MySQL
- Data was loaded in MySQL from the given dataset
- Database job vacancy and table nyc_job was created in hive
- Data was imported from MySQL into Hadoop using SQOOP
- Hive was connected using Beeline CLI for better GUI
- To reduce effort, we directly created a table in hive using SerDe properties and load data into that table.
- To save time, we used Spark as an alternative for MapReduce job.
- Link hive data warehouse to spark
- Load file into RDD and create DataFrame.
- Perform queries in spark using sqlContext.sql () which would reflect in hive table.
- Visualizing the results and plotting bar charts and trend charts.

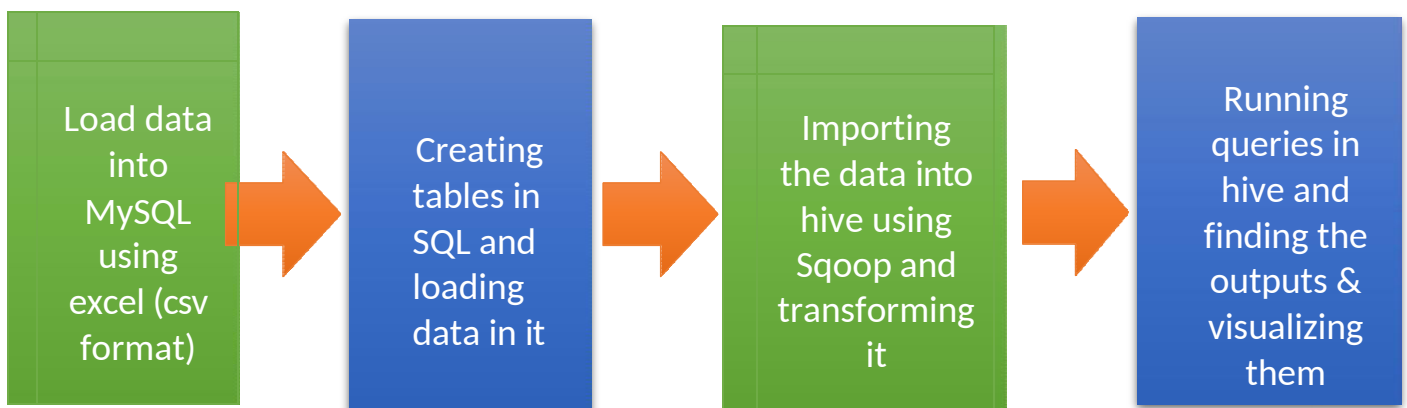


Figure 9: Project Flow

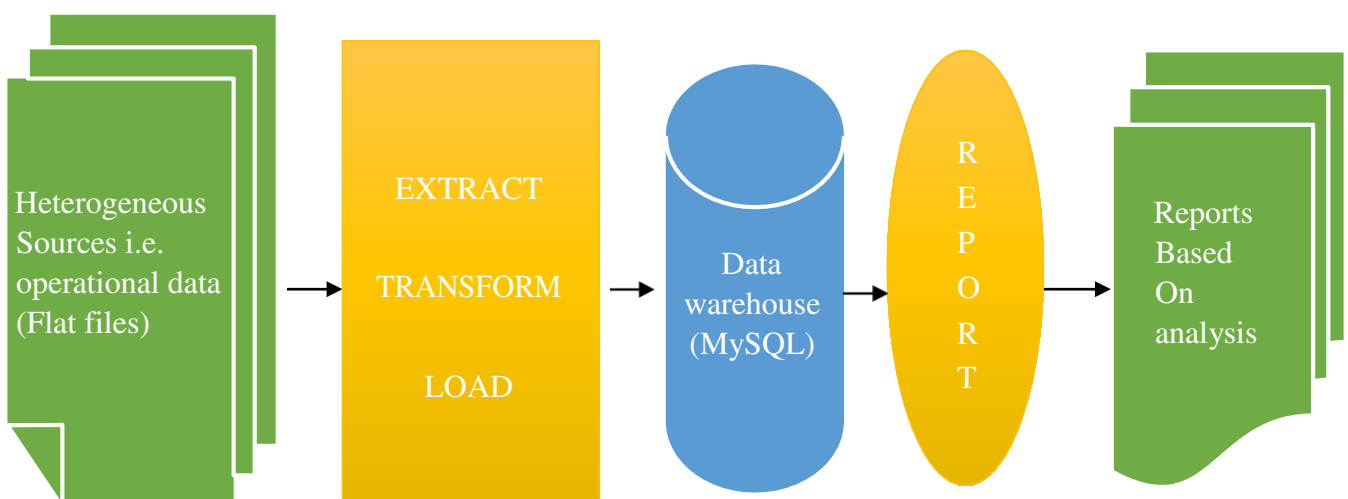
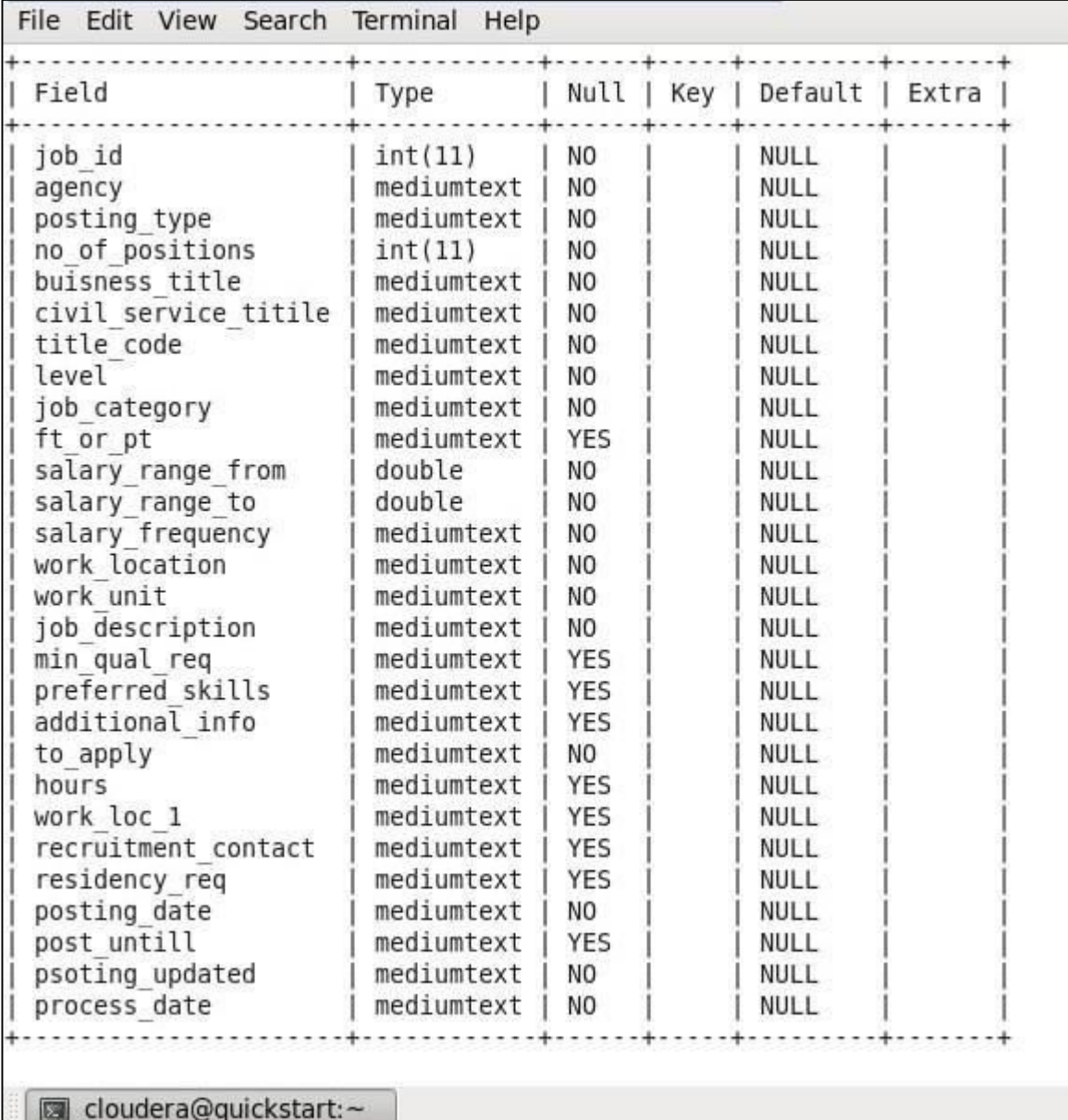


Figure 10: Project Structure

4. IMPLEMENTATION

4.1 Input Stage

- The project has a single table schema approach.
- It consists of 28 fields.
- The fields of tables are job_id, agency, posting_type, no_of_positions, buisness_title, civil_service_titile, title_code, level, job_category, ft_or_pt, salary_range_from, salary_range_to, salary_frequency, work_location, work_unit, job_description, min_qual_req, preferred_skills, additional_info, to_apply, hours, work_loc_1, recruitment_contact, residency_req, posting_date, post_unfill, psoting_updated, process_date.



Field	Type	Null	Key	Default	Extra
job_id	int(11)	NO		NULL	
agency	mediumtext	NO		NULL	
posting_type	mediumtext	NO		NULL	
no_of_positions	int(11)	NO		NULL	
buisness_title	mediumtext	NO		NULL	
civil_service_titile	mediumtext	NO		NULL	
title_code	mediumtext	NO		NULL	
level	mediumtext	NO		NULL	
job_category	mediumtext	NO		NULL	
ft_or_pt	mediumtext	YES		NULL	
salary_range_from	double	NO		NULL	
salary_range_to	double	NO		NULL	
salary_frequency	mediumtext	NO		NULL	
work_location	mediumtext	NO		NULL	
work_unit	mediumtext	NO		NULL	
job_description	mediumtext	NO		NULL	
min_qual_req	mediumtext	YES		NULL	
preferred_skills	mediumtext	YES		NULL	
additional_info	mediumtext	YES		NULL	
to_apply	mediumtext	NO		NULL	
hours	mediumtext	YES		NULL	
work_loc_1	mediumtext	YES		NULL	
recruitment_contact	mediumtext	YES		NULL	
residency_req	mediumtext	YES		NULL	
posting_date	mediumtext	NO		NULL	
post_unfill	mediumtext	YES		NULL	
psoting_updated	mediumtext	NO		NULL	
process_date	mediumtext	NO		NULL	

Figure 11: Table Schema

Dimension of table and its columns:

1. nyc_job

• job_id	int
• agency	string
• posting_type	string
• no_of_positions	int
• buisness_title	string
• civil_service_title	string
• title_code	string
• level	string
• job_category	string
• ft_or_pt	string
• salary_range_from	double
• salary_range_to	double
• salary_frequency	string
• work_location	string
• work_unit	string
• job_description	string
• min_qual_req	string
• preferred_skills	string
• additional_info	string
• to_apply	string
• hours	string
• work_loc_1	string
• recruitment_contact	string
• residency_req	string
• posting_date	string
• post_until	string
• psoting_updated	string
• process_date	string

And data consist of comma separated values as shown in the figure below:

```
Job ID,Agency,Posting Type,# Of Positions,Business Title,Civil Service Title,Title Code  
No,Level,Job Category,Full-Time/Part-Time indicator,Salary Range From,Salary Range  
To,Salary Frequency,Work Location,Division/Work Unit,Job Description,Minimum Qual  
Requirements,Preferred Skills,Additional Information,To Apply,Hours/Shift,Work  
Location 1,Recruitment Contact,Residency Requirement,Posting Date,Post Until,Posting  
Updated,Process Date  
382638,DEPARTMENT OF PROBATION,External,28,Summer College Intern,SUMMER  
COLLEGE INTERN,10234,0,"Public Safety, Inspections, &  
Enforcement",P,15,17.5,Hourly,"33 Beaver St, New York Ny",Admin Executive  
Offices,"POSITION TITLE Summer Intern/Youth Worker The New York City  
Department of Probation (DOP) contributes to safer communities by supervising people  
on probation and fostering opportunities for them to move out of the criminal justice  
system and into meaningful education, employment, health services, family engagement and  
community participation. We are located at 15 offices in every borough across the City  
and provide three core services – juvenile intake, pre-sentence investigations, and  
probation supervision. In summary, DOP ensures that people who enter our system are  
supervised according to their risk level and receive the support and services they need to  
abide by the law and be an asset to their communities. Juvenile Operations serves young  
people who have been arrested and are between the ages of 7 and 15 at the time of the
```

Figure 12: Dataset Capture

4.2 Cleaning Stage

1. Generating two files consisting of '|' and '@' as delimiters to separate fields: Data consist of some columns in which the fields were enclosed within " " character.
2. Hence, it was not being mapped with the schema therefore we used '|' and '@'(for Spark) as delimiter to create a new file.
3. Replace the null values with 'NA' in fields: Data contained many fields which were empty and so the respective fields were replaced with a string 'na'.

Similarly, all files were passed to obtain the data as shown in the figure:

```
382638@DEPARTMENT OF PROBATION@External@28@Summer College Intern@SUMMER COLLEGE INTERN@10234@0@Public Safety, Inspections, & Enforcement@P@15@17.5@Hourly@33 Beaver St, New York Ny@Admin Executive Offices@"POSITION TITLE Summer Intern/Youth Worker The New York City Department of Probation (DOP) contributes to safer communities by supervising people on probation and fostering opportunities for them to move out of the criminal justice system and into meaningful education, employment, health services, family engagement and community participation. We are located at 15 offices in every borough across the City and provide three core services â€" juvenile intake, pre-sentence investigations, and probation supervision. In summary, DOP ensures that people who enter our system are supervised according to their risk level and receive the support and services they need to abide by the law and be an asset to their communities. Juvenile Operations serves young people who have been arrested and are between the ages of 7 and 15 at the time of the alleged offense. A young person's disposition may include Probation supervision, which offers him or her chance to demonstrate an ability to function in the community, in part, by making positive behavioral changes and developing better decision-making skills in order to avoid further delinquent activity. â€¢ Assist coordinator with the design, development and implementation of programming and curricula for leadership through arts and culture, health and wellness, and recreational activities. â€¢ Implement
```

Figure 13: Dataset '@' Separated

```
382638|DEPARTMENT OF PROBATION|External|28|Summer College Intern|SUMMER COLLEGE INTERN|10234|0|Public Safety, Inspections, & Enforcement|P|15|17.5|Hourly|33 Beaver St, New York Ny|Admin Executive Offices|"POSITION TITLE Summer Intern/Youth Worker The New York City Department of Probation (DOP) contributes to safer communities by supervising people on probation and fostering opportunities for them to move out of the criminal justice system and into meaningful education, employment, health services, family engagement and community participation. We are located at 15 offices in every borough across the City and provide three core services â€" juvenile intake, pre-sentence investigations, and probation supervision. In summary, DOP ensures that people who enter our system are supervised according to their risk level and receive the support and services they need to abide by the law and be an asset to their communities. Juvenile Operations serves young people who have been arrested and are between the ages of 7 and 15 at the time of the alleged offense. A young person's disposition may include Probation supervision, which offers him or her chance to demonstrate an ability to function in the community, in part, by making positive behavioral changes and developing better decision-making skills in order to avoid further delinquent activity. â€¢ Assist coordinator with the design, development and implementation of programming and curricula for leadership through arts and culture, health and wellness, and recreational activities. â€¢ Implement daily activities
```

Figure 14: Dataset '|' Separated

4.3 Modeling Stage

To query the data, it has to model into a class with fixed data types so that data can be retrieved using SQL queries. Spark classes have the limit of 23 types so it can't be used to model the data with fields greater than 23 which is what happens in the first two files. As they contained more than 23 categories. So to model this I have used a different approach.

Which is as follows:

1. Schema is created for each file by defining column name with data type

```
scala> import org.apache.spark.sql.types.{StructType,
StructField,IntegerType,DoubleType,StringType};
scala> import org.apache.spark.sql.Row;
scala> val schema = StructType(fields.split(",").map(fieldName => StructField
(fieldName, StringType, true)))
```

Figure 15: Schema Creation

2. Each file rdd is created using the built function `parallize()` function which takes input the array of the data and converts it into rdd.
3. Now using inbuilt spark function `createDataFrame ()` adds are converted to data frames on which queries are run.
4. Single data frame is created.
5. `CreateDataFrame ()` takes to input the rdd and the schema which is used to map the data.

4.4 Retrieving Stage

Following are the queries written to extract the required data for the following questions:

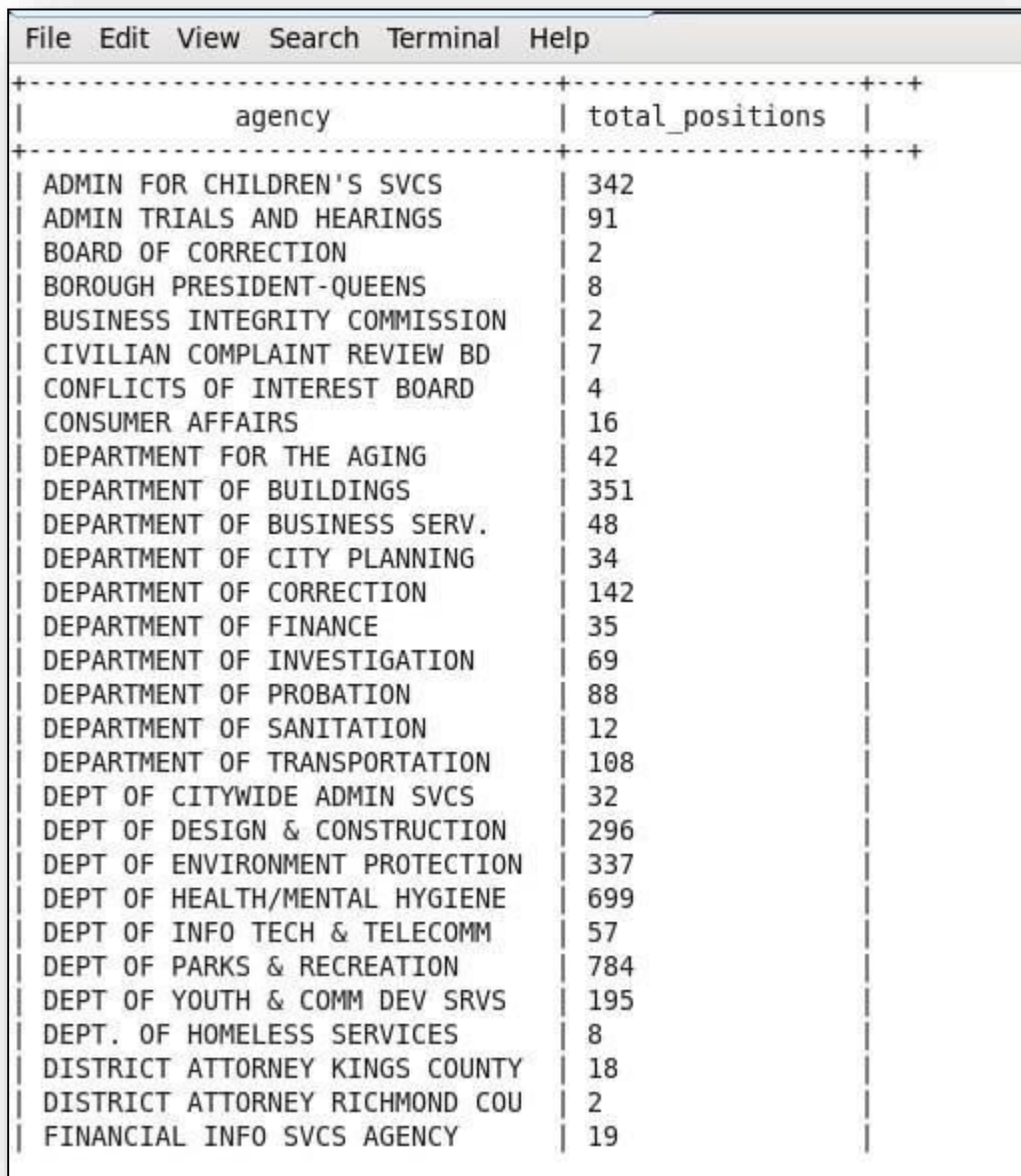
1. **Most in demand job vacancies in each Agency**
select agency,sum(no_of_positions) as Total_Positions from nyc_job group by agency;
2. **Number of external and internal jobs vacancy in each Agency.**
select agency, posting_type, sum(no_of_positions) as TOTAL_VACANT_JOBS
from partition_posting_type where posting_type='External' OR
posting_type='Internal' group by agency,posting_type;
3. **Determine number of part time and full time jobs vacancies in different job categories.** select job_category,ft_or_pt,sum(no_of_positions) as
JOB_VACANCIES from nyc_job group by job_category,ft_or_pt;
4. **Display different jobs with salary range**
select distinct title_code,job_category,(salary_range_to-salary_range_from) as
salary_range from nyc_job order by title_code;
5. **Agency and Job Category having Highest paying jobs**
select z.agency,z.job_category,z.wages from (select
agency,job_category, case when salary_frequency in ('Hourly') then
salary_range_to*9*269 when salary_frequency in ('Daily') then
salary_range_to*269
else salary_range_to
end as wages from nyc_job)z where z.wages in
(select max(u.wages) from (select agency,job_category,
case when salary_frequency in ('Hourly') then
salary_range_to*9*269 when salary_frequency in ('Daily') then
salary_range_to*269
else salary_range_to
end as wages from nyc_job)u);
6. **Month wise number of job posted in all agencies.**
select agency,month(posting_date),sum(no_of_positions) as job_posted from
nyc_job group by month(posting_date),agency;
7. **Most popular preferred skills per job category**
select distinct(job_category) as JOB_CATEGORY,preferred_skills as
SKILLS_REQUIRED from nyc_job;
8. **Days taken in job application for each work unit**
Select distinct (work_unit), datediff(to_date(process_date), to_date(posting_date)) as
NUMBER_OF_HOURS from nyc_job;
9. **Number of job positions in each work location**
select work_location,sum(no_of_positions) as TOTAL_JOB_POSITIONS from
nyc_job group by work_location order by TOTAL_JOB_POSITIONS desc;

5. RESULT AND DISCUSSION

Reports that can be generated:

5.1 Most in demand job vacancies in each Agency.

select agency,sum(no_of_positions) as Total_Positions from nyc_job group by agency;



The image shows a screenshot of a database query result window. The window has a menu bar with 'File', 'Edit', 'View', 'Search', 'Terminal', and 'Help'. Below the menu bar is a table with two columns: 'agency' and 'total_positions'. The table contains 28 rows of data, listing various NYC agencies and their corresponding total positions. The data is as follows:

agency	total_positions
ADMIN FOR CHILDREN'S SVCS	342
ADMIN TRIALS AND HEARINGS	91
BOARD OF CORRECTION	2
BOROUGH PRESIDENT-QUEENS	8
BUSINESS INTEGRITY COMMISSION	2
CIVILIAN COMPLAINT REVIEW BD	7
CONFLICTS OF INTEREST BOARD	4
CONSUMER AFFAIRS	16
DEPARTMENT FOR THE AGING	42
DEPARTMENT OF BUILDINGS	351
DEPARTMENT OF BUSINESS SERV.	48
DEPARTMENT OF CITY PLANNING	34
DEPARTMENT OF CORRECTION	142
DEPARTMENT OF FINANCE	35
DEPARTMENT OF INVESTIGATION	69
DEPARTMENT OF PROBATION	88
DEPARTMENT OF SANITATION	12
DEPARTMENT OF TRANSPORTATION	108
DEPT OF CITYWIDE ADMIN SVCS	32
DEPT OF DESIGN & CONSTRUCTION	296
DEPT OF ENVIRONMENT PROTECTION	337
DEPT OF HEALTH/MENTAL HYGIENE	699
DEPT OF INFO TECH & TELECOMM	57
DEPT OF PARKS & RECREATION	784
DEPT OF YOUTH & COMM DEV SRVS	195
DEPT. OF HOMELESS SERVICES	8
DISTRICT ATTORNEY KINGS COUNTY	18
DISTRICT ATTORNEY RICHMOND COU	2
FINANCIAL INFO SVCS AGENCY	19

Figure 16: Query 1 output

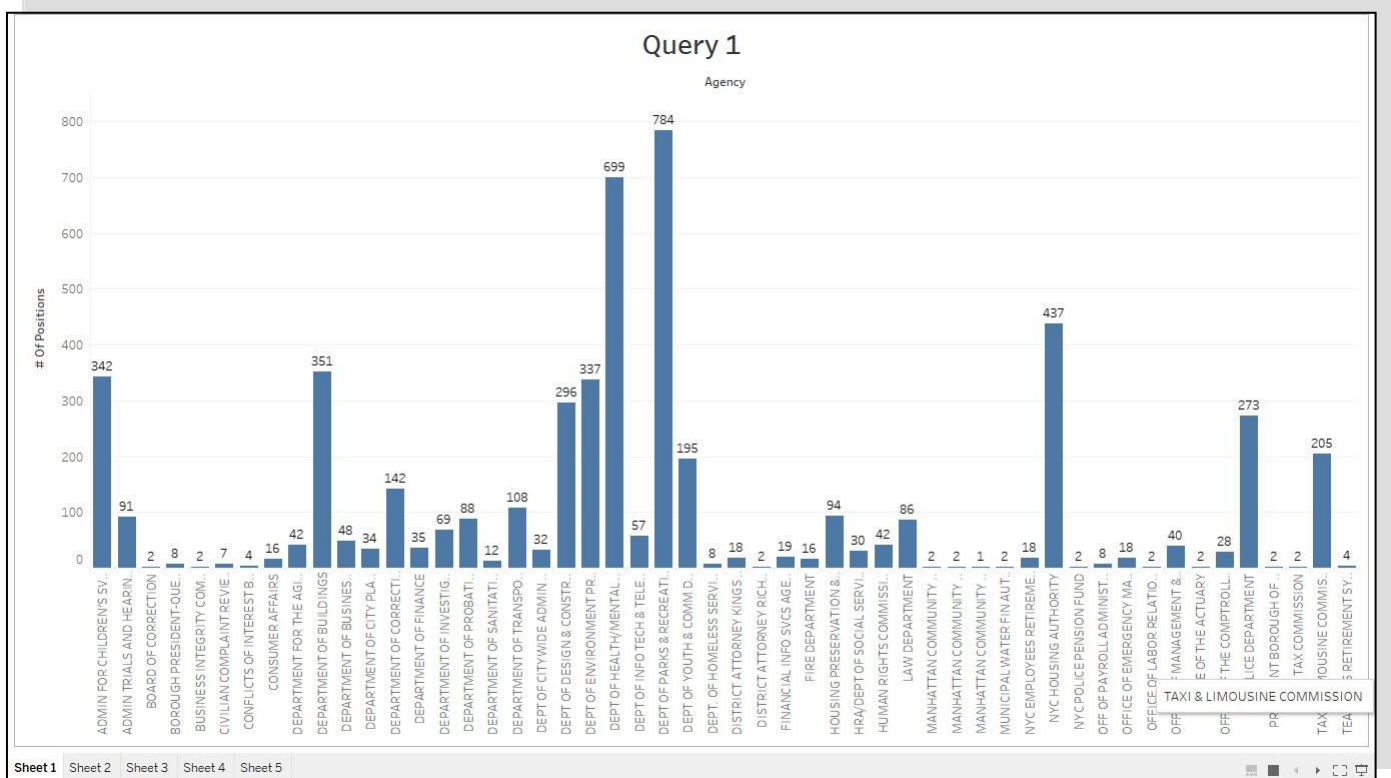
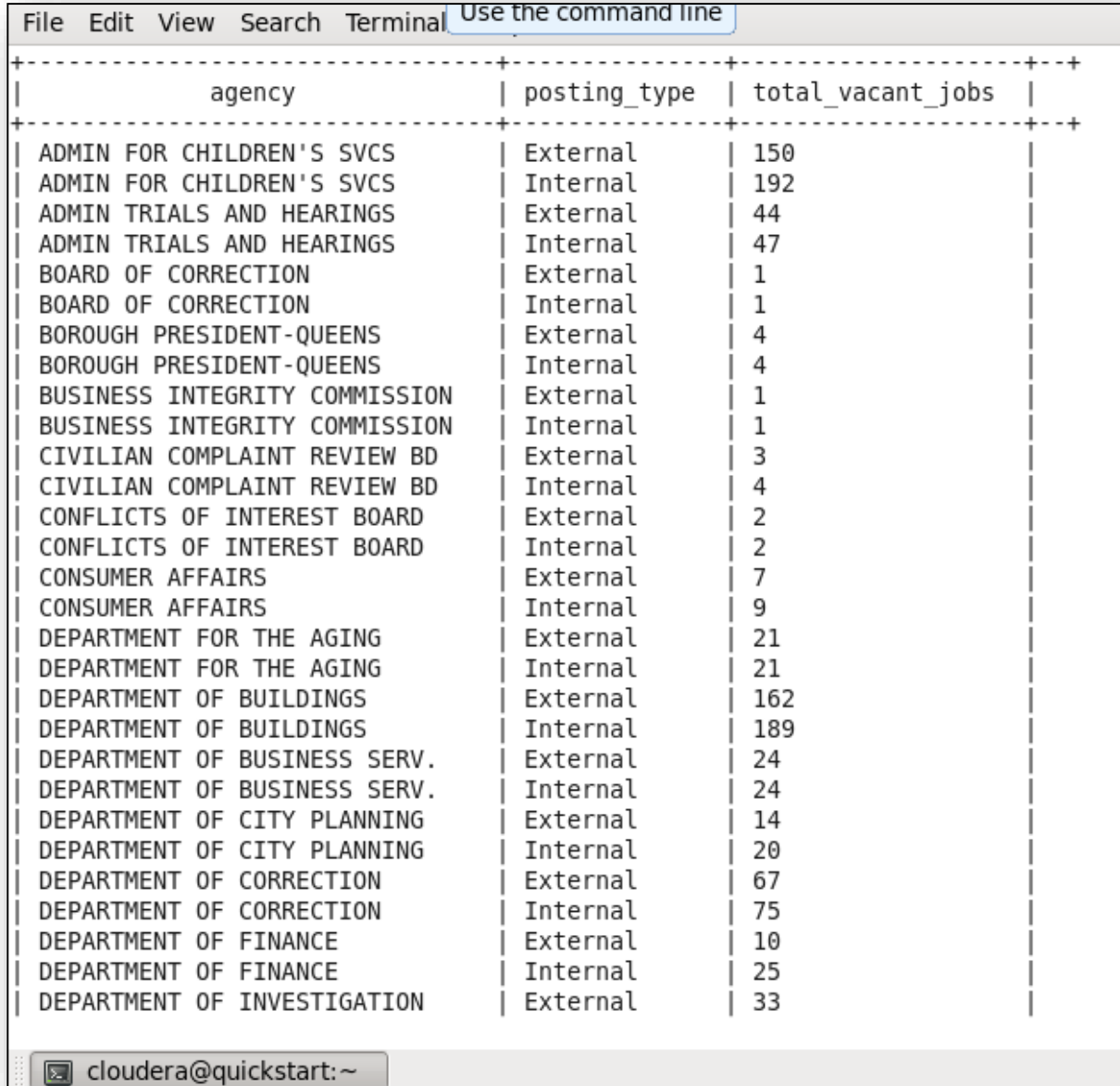


Figure 17: Most in demand job vacancies

5.2 Number of external and internal jobs vacancy in each Agency.

```
select agency, posting_type, sum(no_of_positions) as TOTAL_VACANT_JOBS
from partition_posting_type where posting_type='External' OR
posting_type='Internal' group by agency,posting_type;
```



agency	posting_type	total_vacant_jobs
ADMIN FOR CHILDREN'S SVCS	External	150
ADMIN FOR CHILDREN'S SVCS	Internal	192
ADMIN TRIALS AND HEARINGS	External	44
ADMIN TRIALS AND HEARINGS	Internal	47
BOARD OF CORRECTION	External	1
BOARD OF CORRECTION	Internal	1
BOROUGH PRESIDENT-QUEENS	External	4
BOROUGH PRESIDENT-QUEENS	Internal	4
BUSINESS INTEGRITY COMMISSION	External	1
BUSINESS INTEGRITY COMMISSION	Internal	1
CIVILIAN COMPLAINT REVIEW BD	External	3
CIVILIAN COMPLAINT REVIEW BD	Internal	4
CONFLICTS OF INTEREST BOARD	External	2
CONFLICTS OF INTEREST BOARD	Internal	2
CONSUMER AFFAIRS	External	7
CONSUMER AFFAIRS	Internal	9
DEPARTMENT FOR THE AGING	External	21
DEPARTMENT FOR THE AGING	Internal	21
DEPARTMENT OF BUILDINGS	External	162
DEPARTMENT OF BUILDINGS	Internal	189
DEPARTMENT OF BUSINESS SERV.	External	24
DEPARTMENT OF BUSINESS SERV.	Internal	24
DEPARTMENT OF CITY PLANNING	External	14
DEPARTMENT OF CITY PLANNING	Internal	20
DEPARTMENT OF CORRECTION	External	67
DEPARTMENT OF CORRECTION	Internal	75
DEPARTMENT OF FINANCE	External	10
DEPARTMENT OF FINANCE	Internal	25
DEPARTMENT OF INVESTIGATION	External	33

Figure 18: Query 2 Output

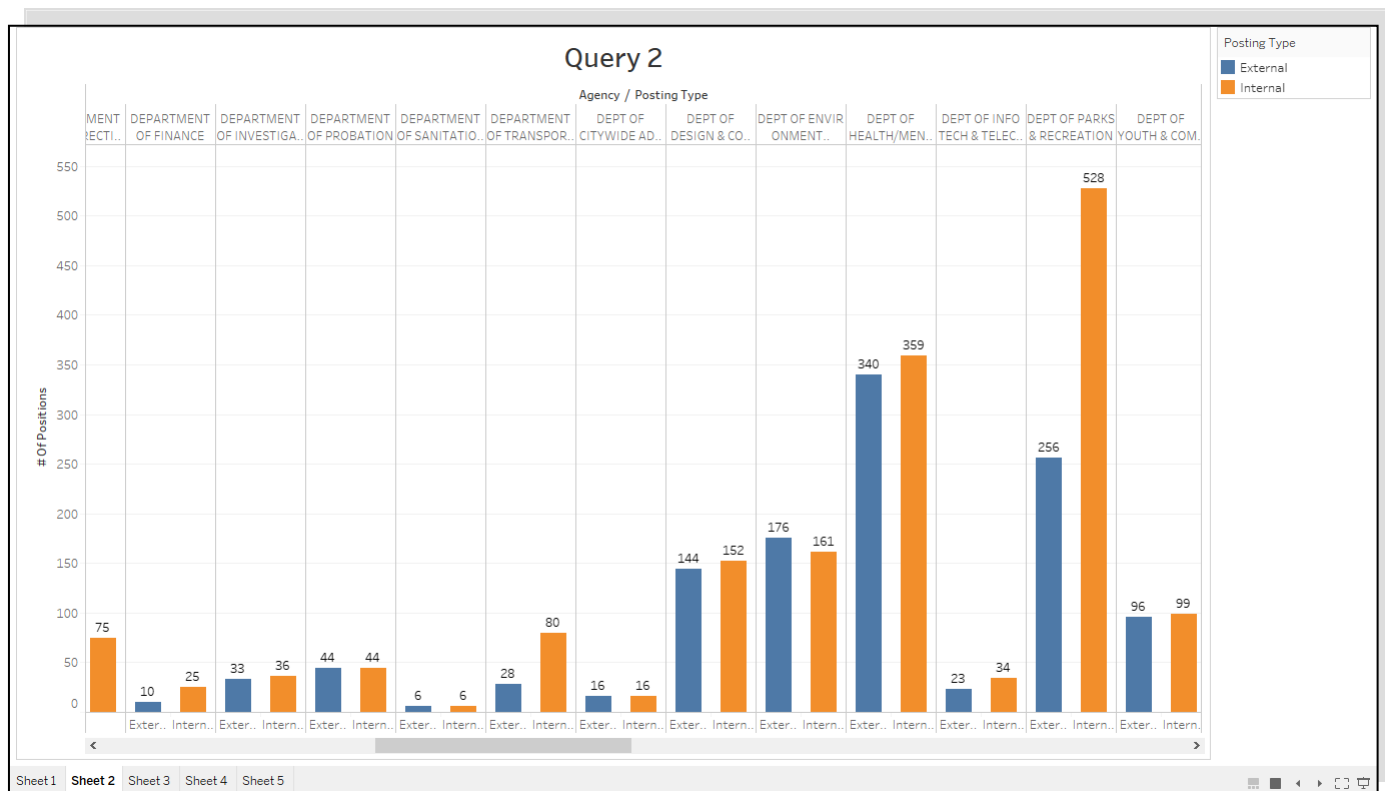


Figure 19: No of internal and external jobs

5.3 Determine number of part time and full time jobs vacancies in different job categories.

```
select job_category,ft_or_pt,sum(no_of_positions) as JOB_VACANCIES from
nyc_job group by job_category,ft_or_pt;
```

File Edit View Search Terminal Help			
job_category	ft_or_pt	job_vacancies	
Health Technology, Data & Innovation	F	6	
Health Technology, Data & Innovation	P	2	
Legal Affairs	F	168	
Legal Affairs	NA	9	
Legal Affairs	P	70	
Legal Affairs Policy, Research & Analysis	F	48	
Legal Affairs Public Safety, Inspections, & Enforcement	F	51	
Policy, Research & Analysis	F	55	
Policy, Research & Analysis	NA	14	
Policy, Research & Analysis	P	4	
Policy, Research & Analysis Public Safety, Inspections, & Enforcement	F		40
Policy, Research & Analysis Social Services	F	24	
Public Safety, Inspections, & Enforcement	F	555	
Public Safety, Inspections, & Enforcement	NA	180	
Public Safety, Inspections, & Enforcement	P	76	
Public Safety, Inspections, & Enforcement Social Services	F	301	
Social Services	F	113	
Technology, Data & Innovation	F	120	
Technology, Data & Innovation	NA	14	
Technology, Data & Innovation Policy, Research & Analysis	F	7	
Technology, Data & Innovation Public Safety, Inspections, & Enforcement	F		5
Technology, Data & Innovation Social Services	F	2	

122 rows selected (289.38 seconds)

Figure 20: Query 3 output

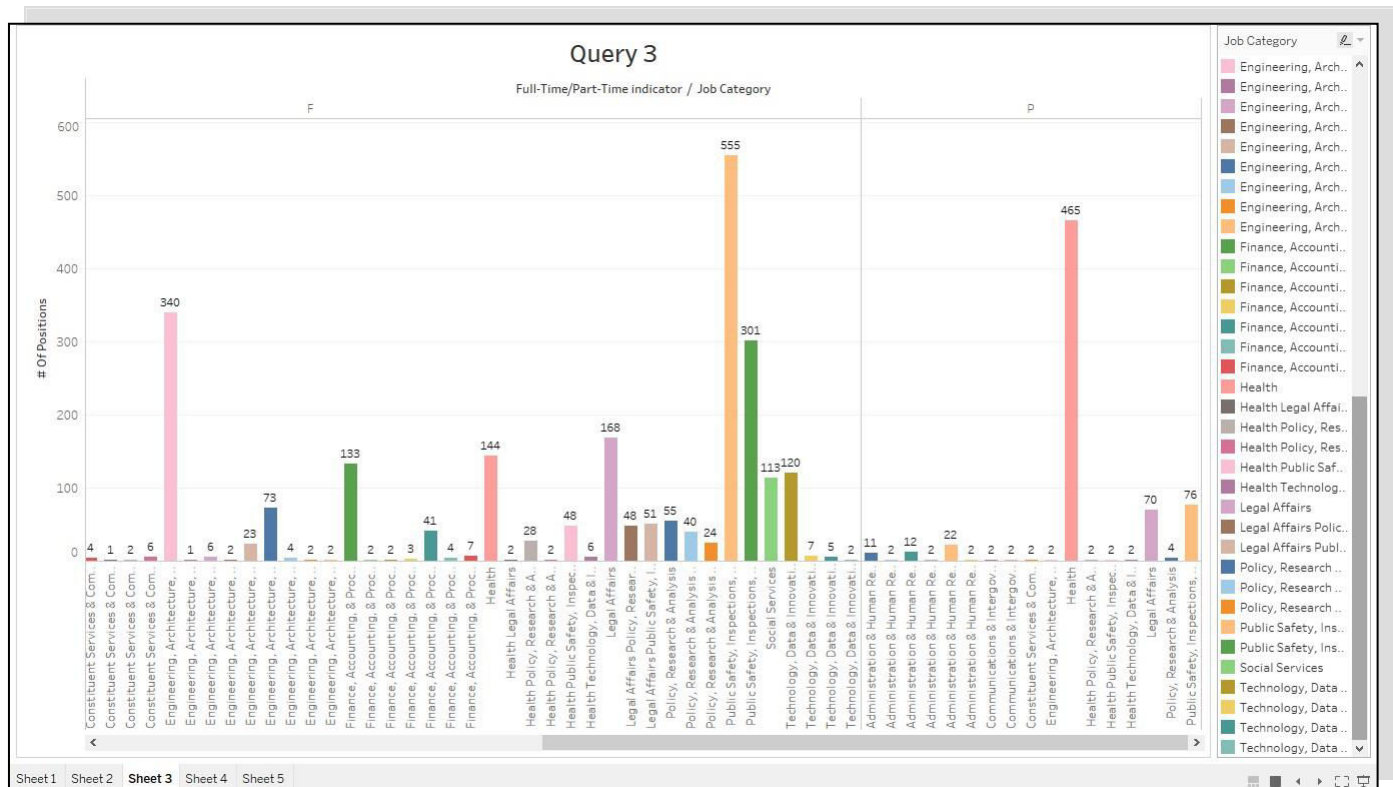


Figure 21: No. of part-time and full-time job vacancies

5.4 Display different jobs with salary ranges (0 indicates that salary is fixed and there is no increment).

```
select distinct title_code,job_category,(salary_range_to-salary_range_from) as
salary_range from nyc_job order by title_code;
```

File	Edit	View	Search	Terminal	Help
title_code	job_category	salary_range			
6766	Technology, Data & Innovation	2085.0			
6766	Technology, Data & Innovation	4000.0			
6766	Technology, Data & Innovation	10000.0			
6776	Health	0.0			
6798	Technology, Data & Innovation	25000.0			
6798	Technology, Data & Innovation	20000.0			
6798	Technology, Data & Innovation	45000.0			
6798	Technology, Data & Innovation	65000.0			
6798	Technology, Data & Innovation	75000.0			
6801	Health	19056.0			
70316	Building Operations & Maintenance	11579.0			
70810	Public Safety, Inspections, & Enforcement	8985.0			
80112	Building Operations & Maintenance	7955.0			
80184	Engineering, Architecture, & Planning	10000.0			
80201	Administration & Human Resources Social Services	22591.0			
80293	Policy, Research & Analysis	35000.0			
80305	Building Operations & Maintenance	24575.0			
80609	Administration & Human Resources	20316.0			
81303	Building Operations & Maintenance	12900.0			
81350	Building Operations & Maintenance	23630.0			
81361	Constituent Services & Community Programs Building Operations & Maintenance	5764.0			
81805	Health	7550.32			
81815	Health	2.5600000000000023			
81815	Health	4.130000000000003			
82011	Building Operations & Maintenance	24564.0			
82107	Health Public Safety, Inspections, & Enforcement	11940.0			
82976	Finance, Accounting, & Procurement	15000.0			
8297A	Finance, Accounting, & Procurement	15000.0			
8297A	Finance, Accounting, & Procurement	82207.0			

Figure 22: Query 4 output

5.5 Agency and Job Category having Highest paying jobs.

```
select z.agency,z.job_category,z.wages from (select
agency,job_category, case when salary_frequency in ('Hourly') then
salary_range_to*9*269 when salary_frequency in ('Daily') then
salary_range_to*269
else salary_range_to
end as wages from nyc_job)z where z.wages in
(select max(u.wages) from (select agency,job_category,
case when salary_frequency in ('Hourly') then
salary_range_to*9*269 when salary_frequency in ('Daily') then
salary_range_to*269
else salary_range_to
end as wages from nyc_job)u)
```

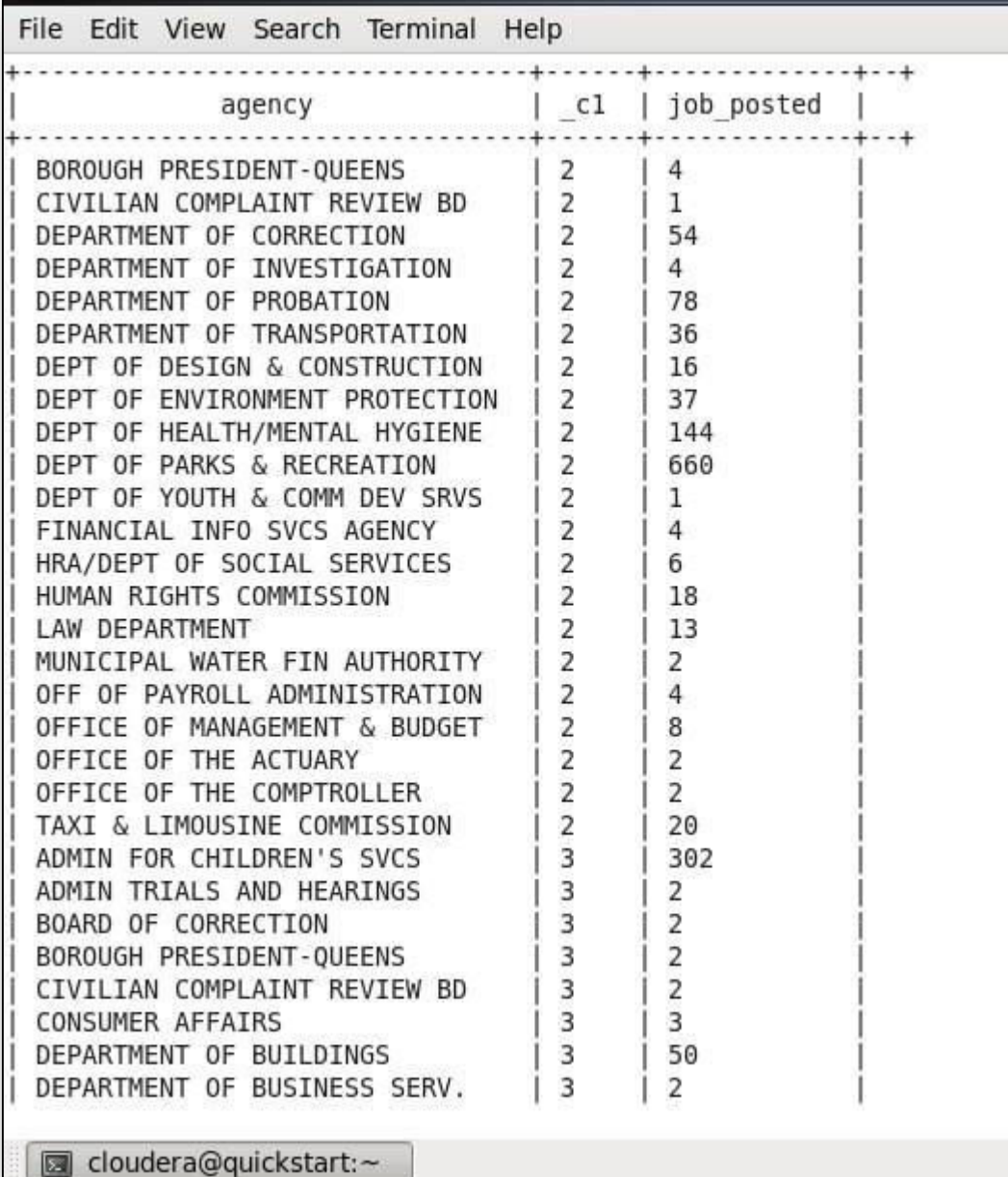
z.agency	z.job_category	z.wages
NYC HOUSING AUTHORITY	Building Operations & Maintenance	234402.0
NYC HOUSING AUTHORITY	Building Operations & Maintenance	234402.0

2 rows selected (794.63 seconds)

Figure 23: Query 5 output

5.6 Month wise number of job posted in all agencies.

```
select agency,month(posting_date),sum(no_of_positions) as job_posted from  
nyc_job group by month(posting_date),agency;
```



agency	_c1	job_posted
BOROUGH PRESIDENT-QUEENS	2	4
CIVILIAN COMPLAINT REVIEW BD	2	1
DEPARTMENT OF CORRECTION	2	54
DEPARTMENT OF INVESTIGATION	2	4
DEPARTMENT OF PROBATION	2	78
DEPARTMENT OF TRANSPORTATION	2	36
DEPT OF DESIGN & CONSTRUCTION	2	16
DEPT OF ENVIRONMENT PROTECTION	2	37
DEPT OF HEALTH/MENTAL HYGIENE	2	144
DEPT OF PARKS & RECREATION	2	660
DEPT OF YOUTH & COMM DEV SRVS	2	1
FINANCIAL INFO SVCS AGENCY	2	4
HRA/DEPT OF SOCIAL SERVICES	2	6
HUMAN RIGHTS COMMISSION	2	18
LAW DEPARTMENT	2	13
MUNICIPAL WATER FIN AUTHORITY	2	2
OFF OF PAYROLL ADMINISTRATION	2	4
OFFICE OF MANAGEMENT & BUDGET	2	8
OFFICE OF THE ACTUARY	2	2
OFFICE OF THE COMPTROLLER	2	2
TAXI & LIMOUSINE COMMISSION	2	20
ADMIN FOR CHILDREN'S SVCS	3	302
ADMIN TRIALS AND HEARINGS	3	2
BOARD OF CORRECTION	3	2
BOROUGH PRESIDENT-QUEENS	3	2
CIVILIAN COMPLAINT REVIEW BD	3	2
CONSUMER AFFAIRS	3	3
DEPARTMENT OF BUILDINGS	3	50
DEPARTMENT OF BUSINESS SERV.	3	2

Figure 24: Query 6 output

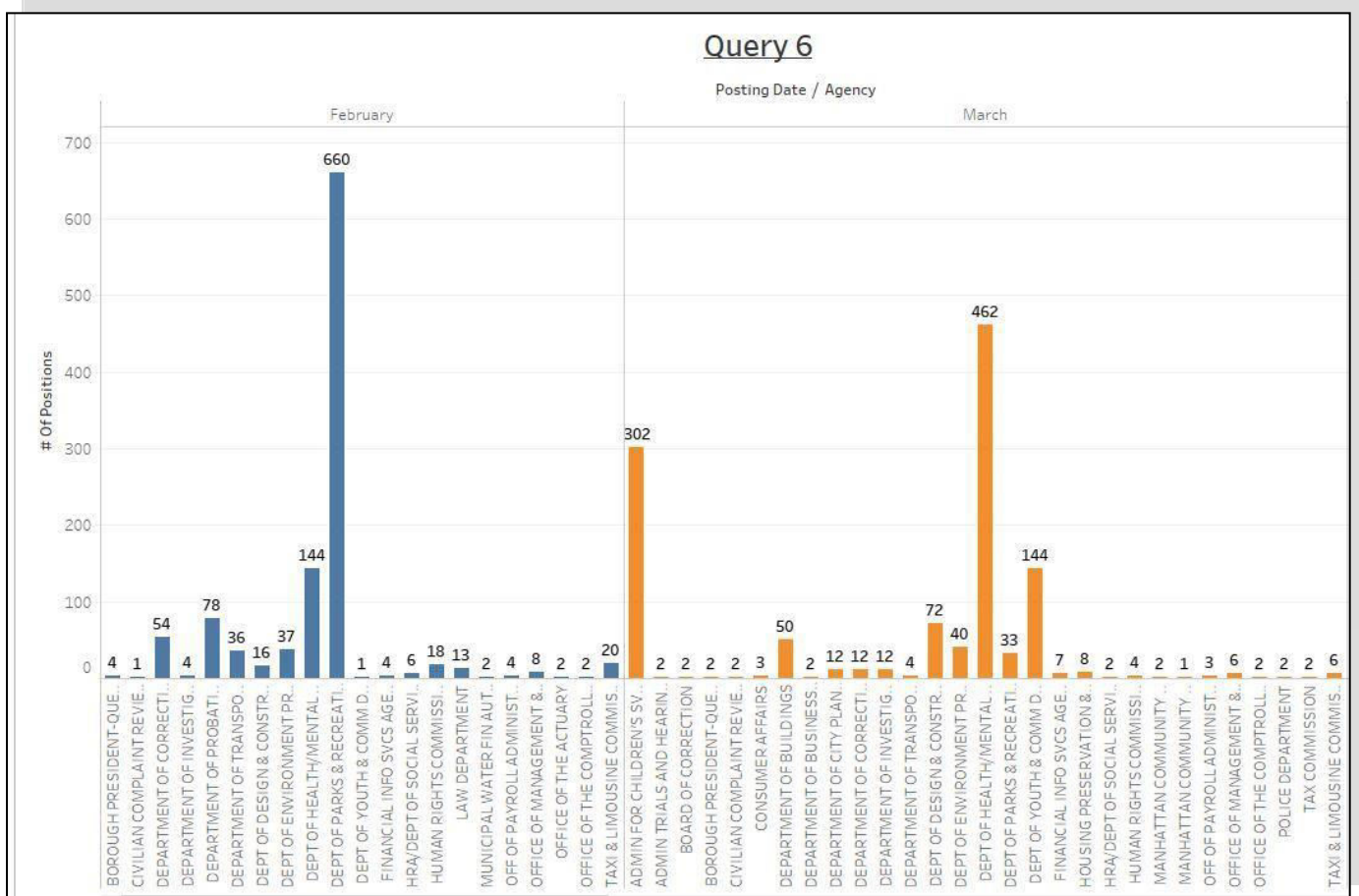


Figure 25: February and March job postings

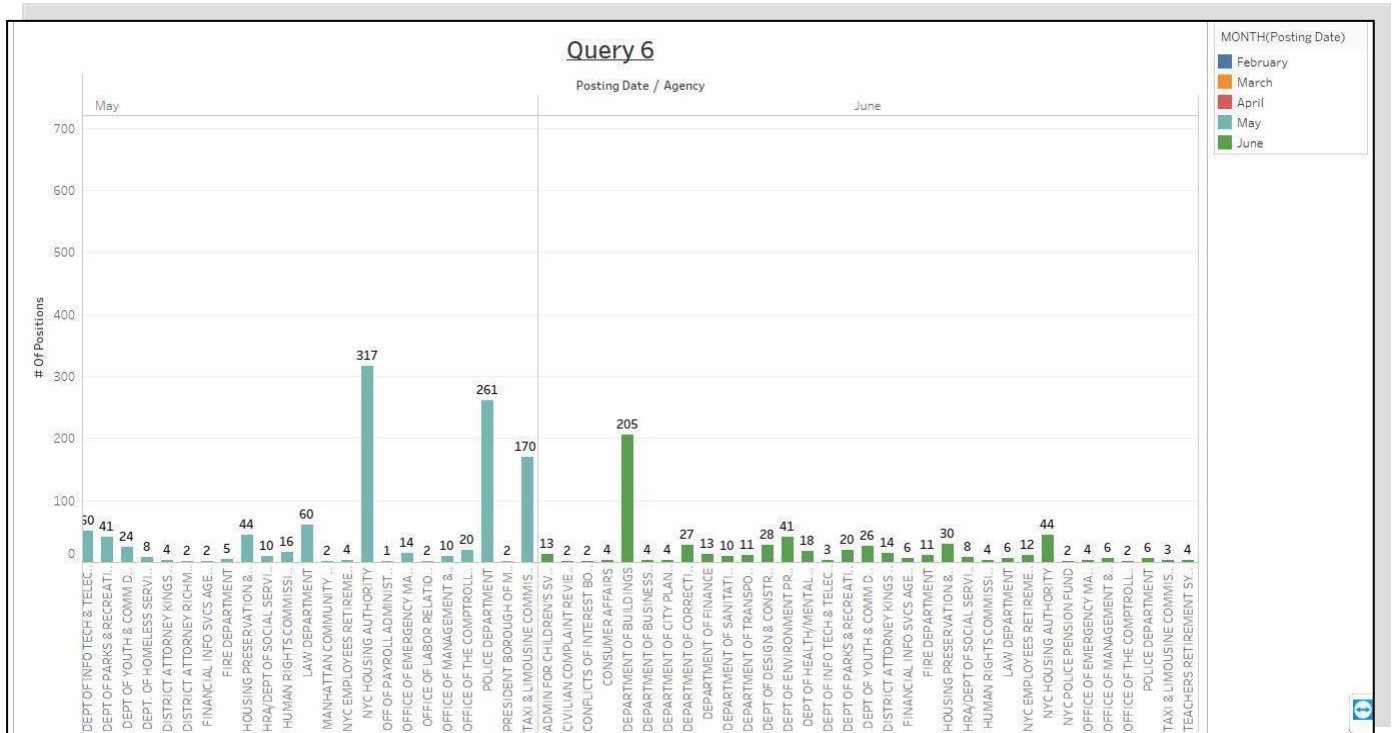


Figure 26: April and May job postings

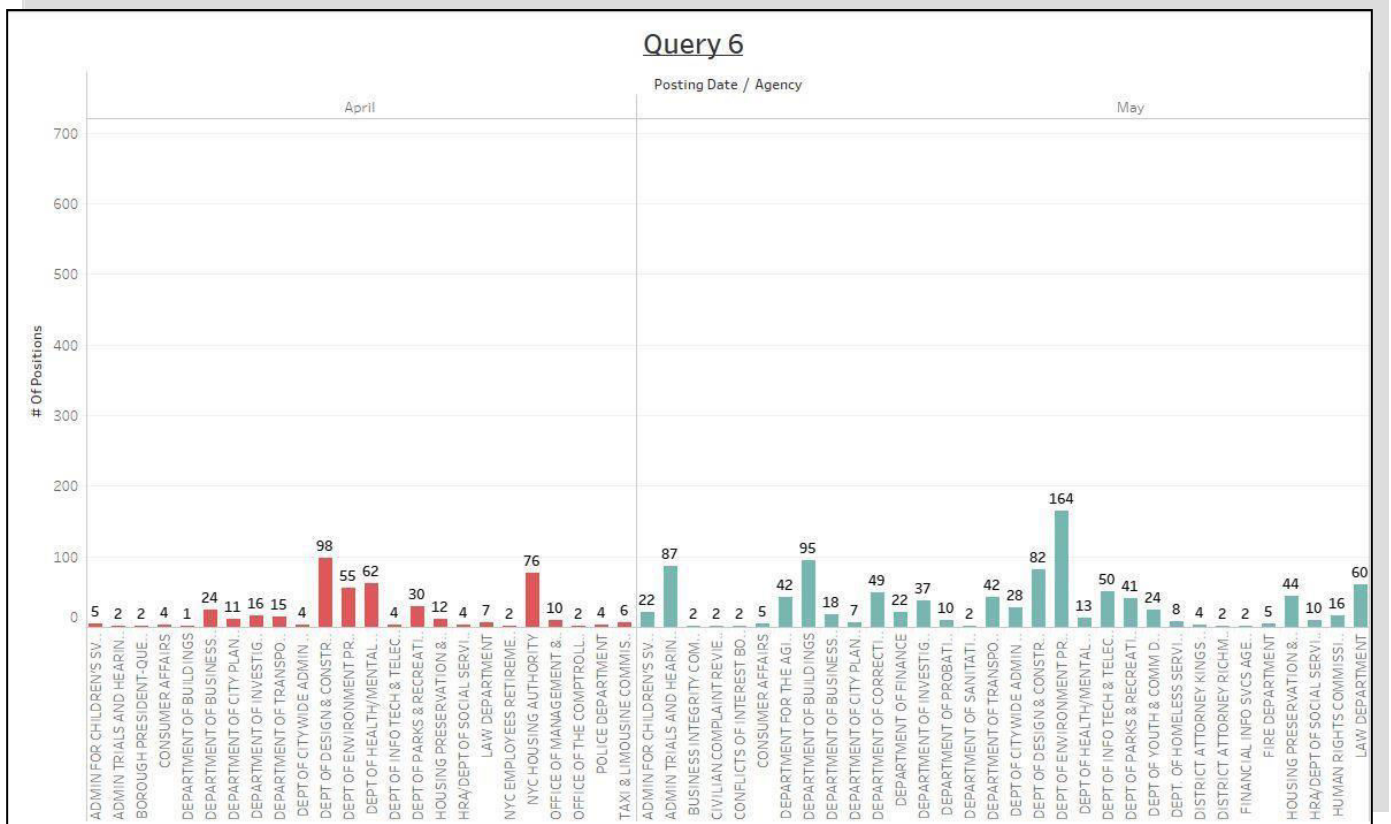


Figure 27: May and June job postings

5.7 Most popular preferred skills per job category

select distinct(job_category) as JOB_CATEGORY,preferred_skills as
SKILLS_REQUIRED from nyc_job;

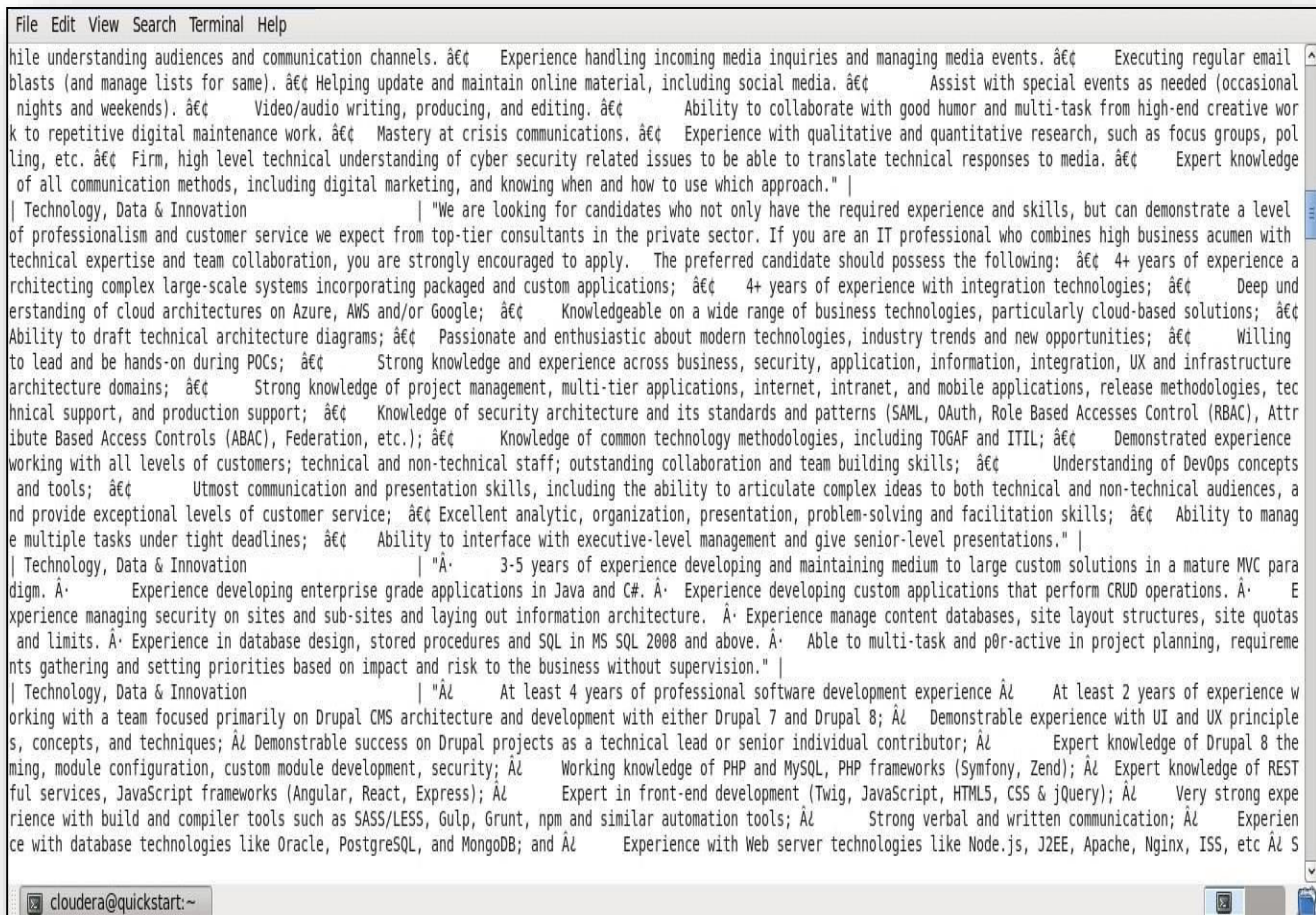
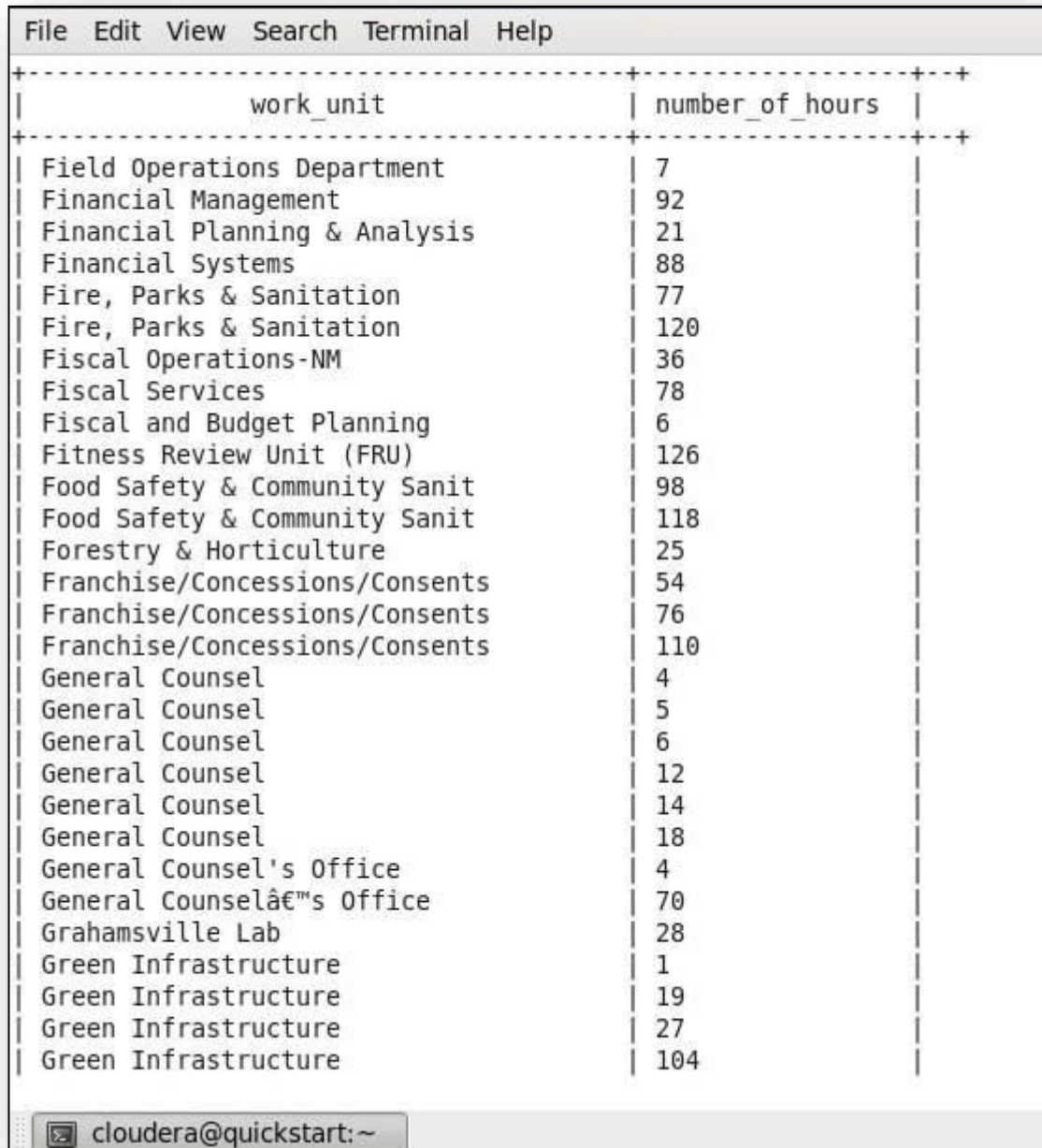


Figure 28: Query 7 output

5.8 Days taken in job application for each work unit

Select distinct (work_unit), datediff(to_date(process_date), to_date(posting_date)) as
NUMBER_OF_HOURS from nyc_job;



work_unit	number_of_hours
Field Operations Department	7
Financial Management	92
Financial Planning & Analysis	21
Financial Systems	88
Fire, Parks & Sanitation	77
Fire, Parks & Sanitation	120
Fiscal Operations-NM	36
Fiscal Services	78
Fiscal and Budget Planning	6
Fitness Review Unit (FRU)	126
Food Safety & Community Sanit	98
Food Safety & Community Sanit	118
Forestry & Horticulture	25
Franchise/Concessions/Consents	54
Franchise/Concessions/Consents	76
Franchise/Concessions/Consents	110
General Counsel	4
General Counsel	5
General Counsel	6
General Counsel	12
General Counsel	14
General Counsel	18
General Counsel's Office	4
General Counsel's Office	70
Grahamsville Lab	28
Green Infrastructure	1
Green Infrastructure	19
Green Infrastructure	27
Green Infrastructure	104

Figure 29: Query 8 output

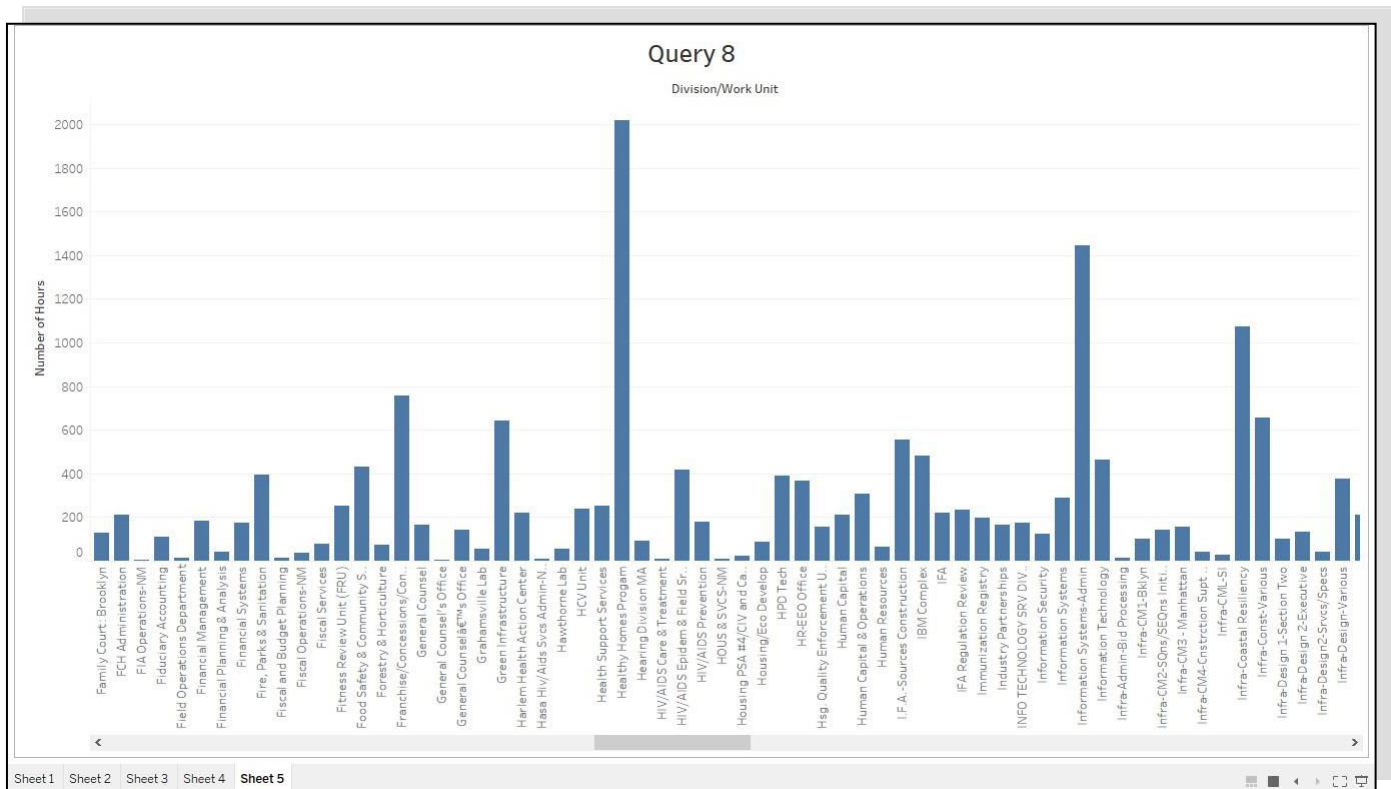
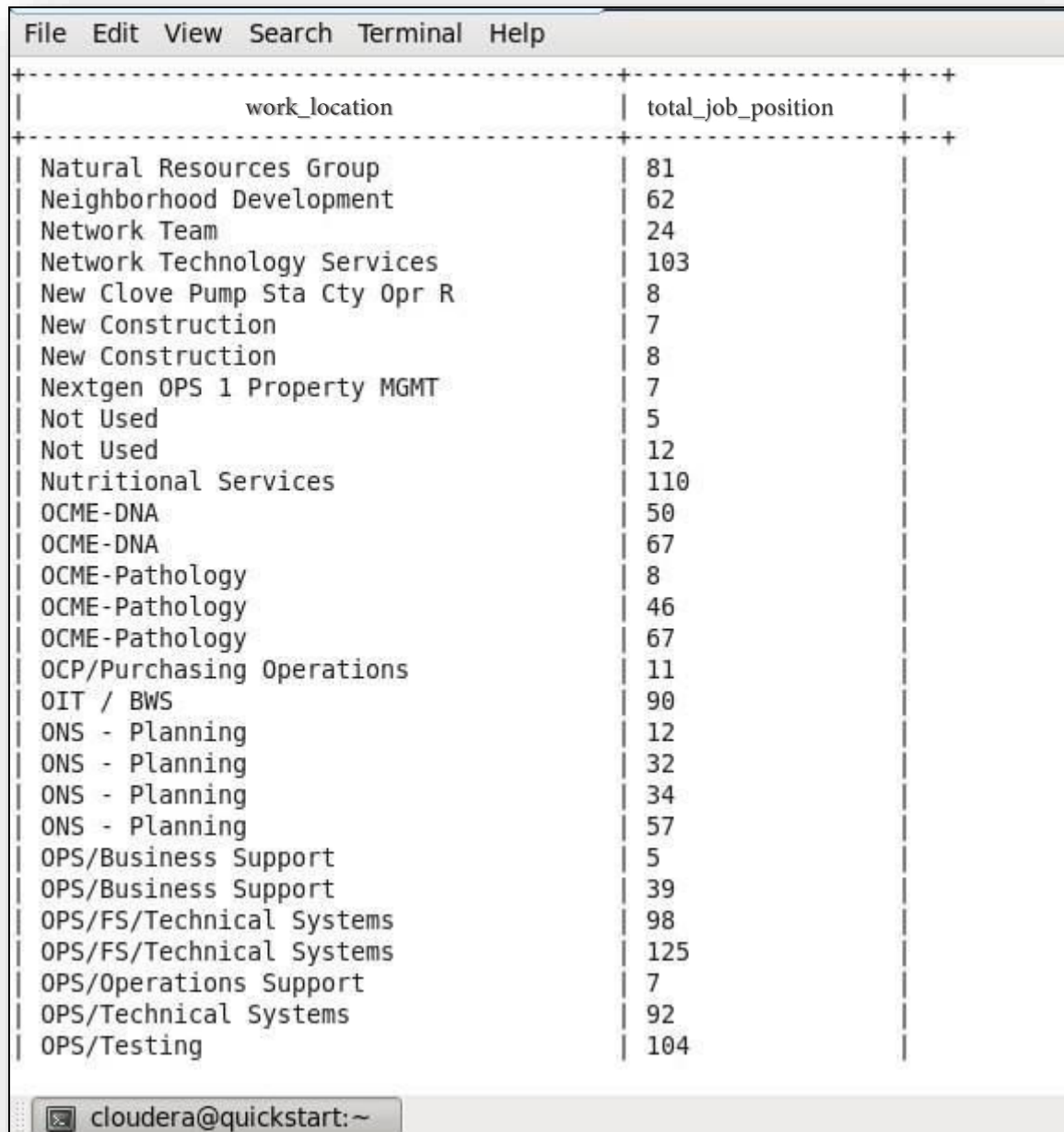


Figure 30: Time taken in job application

5.9 Number of job positions in each work location

```
select work_location,sum(no_of_positions) as TOTAL_JOB_POSITIONS from  
nyc_job group by work_location order by TOTAL_JOB_POSITIONS desc;
```



work_location	total_job_position
Natural Resources Group	81
Neighborhood Development	62
Network Team	24
Network Technology Services	103
New Clove Pump Sta Cty Opr R	8
New Construction	7
New Construction	8
Nextgen OPS 1 Property MGMT	7
Not Used	5
Not Used	12
Nutritional Services	110
OCME-DNA	50
OCME-DNA	67
OCME-Pathology	8
OCME-Pathology	46
OCME-Pathology	67
OCP/Purchasing Operations	11
OIT / BWS	90
ONS - Planning	12
ONS - Planning	32
ONS - Planning	34
ONS - Planning	57
OPS/Business Support	5
OPS/Business Support	39
OPS/FS/Technical Systems	98
OPS/FS/Technical Systems	125
OPS/Operations Support	7
OPS/Technical Systems	92
OPS/Testing	104

Figure 31: Query 9 output

6. CONCLUSION & FUTURE SCOPE OF WORK

6.1 Conclusion

From this project we can conclude that with right tools like Big Data tools we can infer some great things from a big dataset which cannot be processed using old technologies. These big data technologies not only give the power to find some reference we can integrate with machine learning technologies but also to find some depth analytics that cannot be inferred using simple query methods.

6.2 Future Scope

We can improve the model by undergoing following steps:

- We can use machine learning methods to help job seekers find vacancies in any particular organization.
- Use better questionnaire to find some more important trends.
- Better modelling of the data.

REFERENCES

Reference / Hand Books

- [1] MapReduce Design Patterns by Adam Shook, Donald Miner. O'Reilly Media, 2012.
- [2] Capgemini Material and Resources. <https://capgemini.ontidwit.com/>
- [3] Mining the Social Web by Matthew A. Russell. O'Reilly Media, 2013.
- [4] The tutorial point website: https://www.tutorialspoint.com/hadoop/hadoop_big_data_overview.htm
- [5] Apache Hadoop 2.7.2 documentation <https://hadoop.apache.org/docs/r2.7.2/>
- [6] Spark <https://spark.apache.org/docs/1.6.0/>
- [7] Mining of Massive Datasets by Anand Rajaraman, Jure Leskovec, and Jeffrey D. Ullman. University of Stanford, 2014.
- [8] Getting Data Right by Michael Stonebraker, Tom Davenport, James Markarian et al., O'Reilly, 2015 (work in progress).
- [9] Coursera <https://www.coursera.org/learn/gcp-fundamentals>

Biometric Analysis Using Google Cloud
A PROJECT REPORT
Submitted in partial fulfilment of the
requirement for the award of the degree
of
BACHELOR OF TECHNOLOGY (B.Tech)
in
Computer Science & Engineering

by

Romil Nagar
169105154



MANIPAL UNIVERSITY
JAIPUR

Department of Computer Science & Engineering,
School of Computing and IT,
MANIPAL UNIVERSITY JAIPUR
JAIPUR-303007
RAJASTHAN, INDIA

May/2020

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
MANIPAL UNIVERSITY JAIPUR, JAIPUR – 303 007 (RAJASTHAN), INDIA

Date: 30/06/2020

CERTIFICATE

This is to certify that the project titled **Biometric Analysis using Google Cloud** is a record of the bonafide work done by **Romil Nagar** (169105154) submitted in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology (B.Tech) in Computer Science and Engineering of Manipal University Jaipur, during the academic year 2019-20.

Prof. Rohit Verma

*Project Guide, Dept. of Computer Science and
Engineering Manipal University Jaipur*

Dr. Sandeep Joshi

*HOD, Dept. of Computer Science and
Engineering Manipal University Jaipur*

(On company letterhead)

Date:30/06/2020

CERTIFICATE

This is to certify that the project entitled **BIOMETRIC ANALYSIS USING GOOGLE CLOUD** was carried out by **ROMIL NAGAR** (169105154) at **MANIPAL UNIVERSITY JAIPUR** under my guidance during **Starting Date (30\04\2020)** to **Ending Date (30\06\2020)**.

Prof. Rohit Verma

Assistant Professor,
Manipal University, Jaipur

ACKNOWLEDGMENTS

I would like to express my special thanks of gratitude to Manipal University Jaipur who gave me the golden opportunity to learn this great technology and do this wonderful project on the Biometric Analysis using Google Cloud topic, I came to know about so many versatile aspects and I am really thankful to them. I take this opportunity to thank all those magnanimous persons who rendered their full services to my work.

It's with lot of happiness we are expressing gratitude to our guide **Mr. Rohit Verma** B.Tech, Assistant professor, Computer Science and Engineering, for her timely and kind help, guidance and for providing us with most essential materials required for the completion of the completion of this project. We are very thankful to him for his indomitable guidance. This inspiration up to the last moment had made things possible in a nice manner.

Finally, I thank each and every one who helped to complete my project work with their cordial support.

ABSTRACT

Human gait identification has become an active area of research due to increased security requirements. Human gait identification is a potential new tool for identifying individuals beyond traditional methods. The emergence of motion capture techniques provided a chance of high accuracy in identification because completely recorded gait information can be recorded compared with security cameras. The aim of this research was to build a practical method of gait identification and investigate the individual characteristics of gait on cloud.



CERTIFICATE

I hereby certify that the Project Work entitled "Biometrics Analysis using Google Cloud ", which is being submitted by Mr. Romil Nagar (169105154) in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology In Computer Science and Engineering submitted, Manipal University Jaipur, is an authentic record of his own work carried out under the supervision of Prof. Anubha Parashar in duration from 30 April 2020 to 30 June 2020.

The matter presented in this thesis work has not been submitted for the award of any other degree of this or any other university.

Project Guide

A handwritten signature in blue ink, which appears to read "Anubha", is positioned above the printed name.

Anubha Parashar
Assistant Professor
Department of Computer Science and Engineering
School of Computing and Information Technology
Manipal University Jaipur, India

LIST OF FIGURES

Figure No	Figure Title	Page No
1.	The Gait with Mood. left: Normal walk; right: Tired walk	4
2.	The Retargeting Process	5
3.	Image sequence-background subtraction-image binarisation and normalization	6
4.	Stick figure in a motion capture system with 40 markers	6
5.	Structure of Neural Network	9
6.	Classification between Cat and Dog	10
7.	Google Drive	12
8.	Access to Google Drive	13
9.	Google Drive	14
10.	TensorFlow	16
11.	Keras	17
12.	Architecture of our Model.	19
13.	Hyper Parameter Setting for GEINet v3	19
14.	Hyper Parameter Setting for GEINet v4	20
15.	Hyper Parameter Setting for GEINet v6	20
16.	Covariates in the Dataset.	21
17.	Dataset Description.	21
18.	CMC Curve for V3	24
19.	CMC Curve for V4	24
20.	CMC Curve for V6	25
21.	ROC Curve for V3	26
22.	ROC Curve for V4	26
23.	ROC Curve for V6	27
24.	Training Process Plot for V3	28
25.	Training Process Plot for V4	29
26.	Training Process Plot for V6	30
27.	Comparison Table	31

Contents		
		Page No
Acknowledgement		iv
Abstract		v
List Of Figures		vi
Chapter 1	INTRODUCTION AND LITERATURE REVIEW	1
1.1	Overview	1
1.2	Motivation	2
1.3	Project Statement	3
1.4	Literature Review	4
1.5	Data Recording Technique	6
1.6	Data Processing Methods	7
Chapter 2	BACKGROUND OVERVIEW	9
2.1	Conceptual Overview (<i>Concepts/ Theory used</i>)	9
2.2	Technologies Involved	12
Chapter 3	METHODOLOGY	18
3.1	Proposed System	18
3.2	Architecture of the System	19
3.3	Dataset Collection	21
3.4	Feature Extraction	22
Chapter 4	RESULTS AND ANALYSIS	24
4.1	CMC Curve	24
4.2	ROC Curve	26
4.3	Training Process Plot	28
Last Chapter	CONCLUSIONS & FUTURE SCOPE	31
6.1	Conclusions	31
6.2	Future Scope of Work	31
REFERENCES		32

INTRODUCTION AND LITERATURE REVIEW

1.1 Overview

Human gait analysis, the systematic study of human walking, has been developed from early descriptive studies to newer studies involving mathematical analysis and modelling and has become an important part of human motion analysis. Gait analysis has been applied in many areas, including biomechanical, psychological, and security disciplines.

The goal of researchers is to analyze a walker's status based on their gait, such as gender, age, and health (Mather & Murdoch 1994; Nigg et al. 1994; Powers & Perry 1997; Grabiner et al. 2001; Troje 2002; Zhang et al. 2009; Menant et al. 2009a). Furthermore, researchers aim to identify individuals (Foster et al. 2003; Wang et al. 2003; Han & Bhanu 2005; Sarkar et al. 2005; Chellappa et al. 2007). Recently, the use of soft biometrics for recognition has been studied. In (Wang et al. 2005), a video analysis framework using soft biometrics signatures such as skin tone and clothing color was used for airport security surveillance. In (Moustakas et al. 2010), an efficient framework combining soft biometrics such as "height" and "stride length" with gait features was proposed.

With the development of gait recording techniques, research methods were also advanced. The question that was first proposed in the 1970s, 'Can people recognize their friends or family by gait' has been developed to 'Can we identify a particular person by gait' (Cutting & Kozlowski 1977; Foster et al. 2003; Han & Bhanu 2006; Moustakas et al. 2010). Intuitively, we know that individual gaits are different and include some personal information. Can gait features were used like a 'biometric signature' to identify individuals, similar to the use of DNA or handwriting? This question inspired the research aims of this thesis.

In this research, a novel approach for identifying individuals was proposed based on 3D motion capture data. A novel gait feature set was proposed and evaluated. It was investigated the different influences on gait features from gait phases. A novel gait phases definition was proposed. The similarity and dissimilarity between the left and right sides of the body in gait were investigated. Besides, the relationship between gait and attractiveness was analyzed and a predictable model for gait attractiveness was built.

1.2 Motivation

Gait is defined as “a manner of walking”. We extend our definition of gait to include both the appearance and the dynamics of human walking motion. Johansson had shown in the 1970’s that observers could recognize walking subject’s familiar to them by just watching video sequences of lights affixed to joints of the walker. The effective representation of gait is a key issue. Currently, there are several successful representation models such as appearance-based models, stochastic statistical models articulated biomechanics models in which a set of parameters describes the gait, and other parameter-based models. Several of these models can be combined to further improve the representation of gait. this paper, a state-of-the-art deep convolutional model (VGG-D) which is consisting of 16 convolutional and pooling layers was first trained in a fully supervised setting. The convolutional model just uses very small convolution filters and was the winner of ILSVRC-2014. We then used the responses from the pre-trained CNN model using *CASIA-B* dataset (classification annotations only) as generic features for gait representation generation from which I got motivated and decided to choose this research paper for analysis work.

1.3 Problem Statement

- We want to build a network, which can transform different kind of binary silhouettes, whether it's composed of 1 frames or 10 frames of Gait silhouettes or different start frame, direct to complete GEI, which has almost one kind of shape for each subject.
- Build the different types of models and compare their accuracies.
- Use pre-trained models and compare their accuracy with our algorithms.
- Extraction of best features through Convolutional neural networks.
- Build a framework which will be able to perform the preprocessing as well as the model prediction on CASIA-B dataset

1.4 Literature Review

Gait is defined as “a particular way or manner of moving on foot”. Gait analysis was a purely academic discipline in the beginning. Over time, it has been transformed into a useful tool in the diverse fields of physiology, clinical medicine and security. Psychology research (Johansson 1973) would seem to suggest that humans can recognize movement patterns merely from the temporal component. Since the first complete description of the gait cycle given by the Weber brothers in Germany in 1836 (Weber & Weber 1836), gait studies have revealed gait to be related to anatomy, physiology, and biomechanics. From the 1970s to the 1990s, many types of gait analysis research focused on different targets. These studies attempted to reveal the relationship between gait patterns and gender, age, health, wealth, and so on (Mather & Murdoch 1994; Schmitt & Atzwanger 1995; Cho et al. 2004; Boston & Sharpe 2005; Chiu & Wang 2007; Bennett et al. 2008; Røislien et al. 2009; Menant et al. 2009b; Bockemuhl et al. 2010). The biomechanical analysis of gait has been successfully applied in human clinical gait analysis (Whittle 1996).

With the development of motion capture techniques in the last two decades, new research areas have attracted interest. Motion capture techniques provide 3D motion data by motion capture system, whereas videos or cameras only provide 2D image data (3D is abbreviation of three dimensions, 2D is abbreviation of two dimensions). Based on 3D gait data, many medical studies sought to investigate the difference between healthy individuals and patients with specific diseases (Powers & Perry 1997; Rosengren et al. 2009; Zhang et al. 2009). Human identification research also received greater attention with the advent of motion capture techniques, particularly in the field of security (Ma et al. 2006; Shan et al. 2008). Other research has sought to recognize human action, such as walking, jumping, and running. However, gait analysis based on image data has continued since the 1970s (Cutting & Kozlowski 1977; Barton & Lees 1997; Collins et al. 2002; Keren 2003; Jokisch et al. 2006; Bodor et al. 2009).

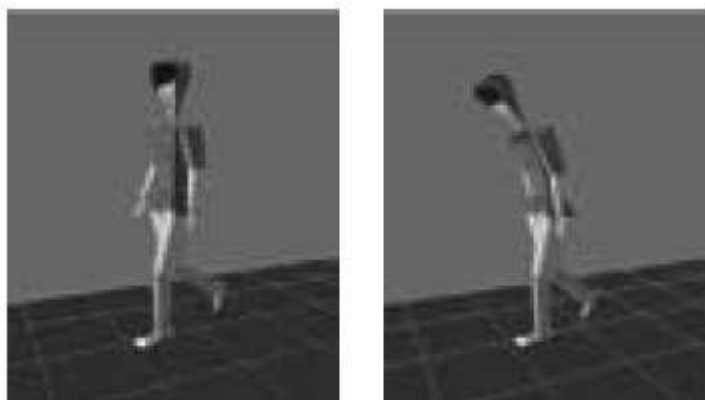


Figure 1: The Gait with Mood. left: Normal walk; right: Tired walk (Munetoshi et al. 1995)

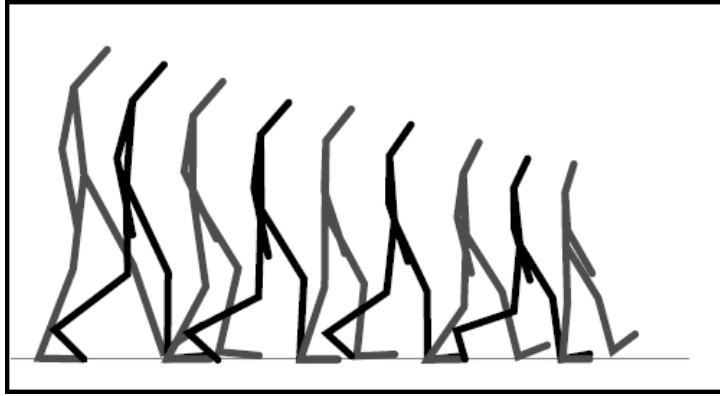


Figure 2: The retargeting process, which adapts the motion as the character morphs to 60% of its original size (Gleicher 1998).

1.5 Data Recording Technique

For gait analysis, it is common for high-speed video data to be collected and analyzed on a frame-by-frame basis. There have been many studies based on video data. In those studies, the silhouettes of walkers were identified from the background. A silhouette is the image of a person, an object or scene consisting of the outline and a basically featureless interior, with the silhouetted object usually black (Figure 3).



Figure 3: Image sequence-background subtraction-image binarisation and normalization (Lam et al. 2007).

With the advent of motion capture technology, the recording of 3D gait data has become another common technique. Data are collected using infrared cameras that track the motion of markers that are placed on the crucial points on body segments. The markers' X, Y, Z coordinates can be recorded. A real-time model of gait will be captured in the motion capture system. Motion capture systems are currently represented by two main groups, optical systems and non-optical systems. Optical systems require the subject to wear a form-fitting suit with markers that reflect light back to the camera's lens to obtain the markers' 3D positions. Optical systems use multiple cameras to capture the markers exact positions. The more cameras used, the higher the accuracy of the recorded data will be. An optical system usually contains 7-13 cameras.

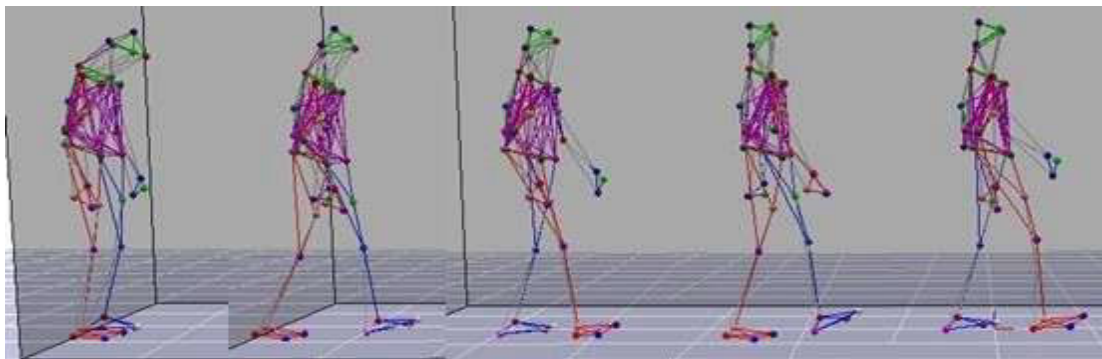


Figure 4: Stick figure in a motion capture system with 40 markers

Recorded data from a motion capture system unquestionably contain more detailed data than that from a video camera. The advantage of video camera data is that the database sample could contain hundreds of people because security cameras provide data easily. For example, gait recognition research based on video images used 114 subjects in (Foster et al. 2003), 126 subjects in (Moustakas et al. 2010), and 80 subjects in (Zhang & Troje 2005). 3D motion capture database samples normally contain only dozens of subjects because more experiment conditions are required. For example, 20 subjects were used in (Rosengren et al. 2009), 37 subjects were used in (Kennedy et al. 2009), 16 subjects were used in (Allet et al. 2008), and 36 subjects were used in (Menant et al. 2009a). Although many important studies have been based on video data, research based on 3D data has attracted more attention in the last two decades, particularly in the fields of medicine and security.

1.6 Data Processing Methods

In general, a complete analysis approach for gait data includes the representation method for gait, the analysis method, and the distinguishing method if portions of the research involve recognition or identification. The distinguishing method refers to the method used to recognize or identify gait. Representation and analysis methods differ depending on the dimension of the data.

1.6.1 Representation Methods

Basically, there are two major approaches in gait analysis: feature-based approaches and model-based approaches (Wang & Singh 2003; Boulgouris et al. 2005). In feature-based approaches, the extracted features are used to represent gait. Silhouettes were commonly used for 2D data (Foster et al. 2003; Wang et al. 2003; Boulgourisa et al. 2006; Barnich & Van Droogenbroeck 2009). In 3D databases, the gait representation was first determined by how many markers were used on the subjects. Those markers were attached at joints. Usually, the number of markers varied from 10 to 40. Some researchers used indicators of joints: degrees of freedom (DOF) (Bockemuhl et al. 2010), joint rotation (Bruijn et al. 2008), joint angles (Ormoneit et al. 2005) and so on. In model-based approaches, mathematical tools such as Fourier expansion and singular value decomposition (SVD) were used to represent gait (Troje 2002; Cunado et al. 2003; Ormoneit et al. 2005). In (Cunado et al. 1999), thigh motion was modelled as a pendulum for representation.

1.6.2 Analysis Methods

A. Feature-based approach

The main component of the feature-based approach is the extraction of gait features from gait. It is a common method in gait analysis. Techniques such as Fourier transforms, motion energy images, Eigen space transformation, principal component analysis, and canonical space transformation are often used to reduce data dimensionality and generate features for gait analysis.

In research based on video image, silhouettes are the primary features (Foster et al. 2003; Wang et al. 2003; Boulgourisa et al. 2006; Barnich & Van Droogenbroeck 2009). In (Boulgouris & Chi 2007), body component-wise in silhouettes was used. Feature images or templates in silhouette were used in (Masoud & Papanikolopoulos 2003; Lam et al. 2007). Torso length, upper-arm length, lower-arm length, thigh length, calf length, and foot length were used in (Han & Bhanu 2005). Other methods include gait energy images, proposed in 2006 (Han & Bhanu 2006), and composite energy features: clusters of energy filters, to identify gait (Dosil et al. 2008). Clothes, footwear, walking surface, emotional state, and walking speed can also be features (Boulgouris et al. 2005).

Initially, these analyses related to view point. Subsequently, researchers gradually proposed methods that are view point independent (Zhang & Troje 2005; Bodor et al. 2009). Based on motion capture data, more features about body segments are chosen. Hip-knee angles were used as features for gait recognition (Barton & Lees 1997; Cunado et al. 2003). Hip flexion in swing and lower limb joint angles have been studied previously (Vrieling et al. 2008). Arm movement has been received more attention recently. Swinging arm regions were used for gait phase detection (Wang et al. 2009). Arm motion was used in human motion recognition (Ganesh & Bajcsy 2008). The features used in previous research show that limbs and hips are important in gait recognition based on 3D gait data. Furthermore, joint motion trajectories were used to extract

gait features as a signature via wavelet (Lakany 2008). Silhouettes are always used in side-view, and curve spread as an efficient descriptor of front-view gait is used in recognition (Soriano et al. 2004).

3D data captured by motion capture systems can produce more accurate identification results because more information is recorded. Many more potential features can be chosen: hip flexion in swing, lower limb joint angles in (Vrieling et al. 2008); velocity in (Kressig et al. 2004; Bennett et al. 2008; Menant et al. 2009a); pelvic rotation and thorax in (Bruijn et al. 2008); arm swing in (Ford et al. 2007); hip-knee angles in (Barton & Lees 1997; Cunado et al. 2003); and motion trajectory in (Wu & Li 2009).

B. Gait Signature for Identification via Feature-based Methods

A novel set of gait features purely extracted from gait as gait features were first proposed to represent gait. The features were then analyzed to determine if they could represent personal gait. Then, CNN were used to extract gait signatures for identification based on a normalized gait cycle. A softmax was used to identify subjects, and identification results were compared based on different methods.

The data are derived from a gait cycle normalized by linear interpolation. Many previous studies did not incorporate this step. The advantage of normalized gait is that subjects have the same gait cycle, same initial gait pose, and same frame numbers in one gait cycle, which improves the accuracy of individual gait identification.

BACKGROUND OVERVIEW

2.1 Conceptual Overview

Deep learning is a smaller element. It is a more specific application of AI, or more precisely a subset of machine learning and one of its algorithms. Right now DL is the most popular ML-algorithm that refers to a technique for creating an AI-powered layered neural network, much like a simplified replica of the human brain. DL gives computers the ability to solve more complex problems than other ML algorithms.

An example — during a game of chess, a neural network is trained predominantly.

What is a neural network?

Neural networks, like biological ones, are composed of neurons. In machines, the neurons are virtual basically bits of code running statistical regressions. String enough of these virtual neurons together and you get a virtual neural network.

A neural network is a learning system, that is, it operates not only on the basis of given algorithms but also based on our own experience. A neuron in deep learning can be thought of as a “black box” with many inlets and one outlet. At the input, the neuron receives signals and forms an output based on them

We know from biology that our ability to learn is based on the unique properties of a brain of 80 billion neurons. The collective work of these cells now allows you to understand what I am telling now. And neural network algorithms are trying to build a model of this process, although implausible biologically, but inspired by the laws of nature.

In these programs, calculations are made by a network consisting of separate elements that process and transmit information to each other. In the process of spreading through the network, information changes we call this process learning.

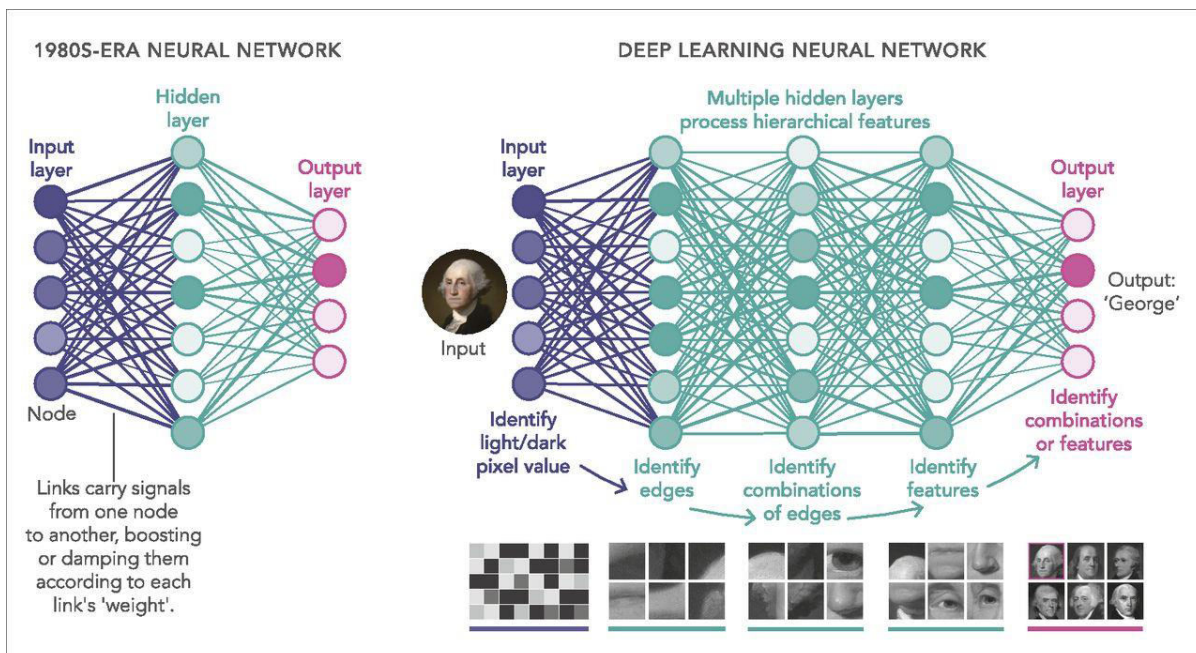


Figure 5: Structure of Neural Network

If you search for deep learning in Wikipedia, you will notice this super hot word is far away from being new. Actually this technology was introduced to the machine learning community by Rina Dechter in 1986, and to artificial neural networks by Igor Aizenberg and colleagues in 2000, in the context of Boolean threshold neurons.

So yes, the roots of deep learning are decades old, but the term “deep learning” nor this approach was so popular before 2012. To the greatest extent, major breakthroughs occur after research and hard work of three scientists that is now well known as fathers of the Deep Learning:

Yoshua Bengio, Geoffrey Hinton and Yann LeCun — fathers of the Deep Learning

The architecture of a neural network can consist of many layers — information processing is divided into many stages. This is where the “deep” came from, by the way. In general, there are three types of layers of neurons in a neural network:

- **Input Layer:** Input variables, sometimes called the visible layer.
- **Hidden Layers:** Layers of nodes between the input and output layers. There may be one or more of these layers.
- **Output Layer:** A layer of nodes that produce the output variables.

For example, it is necessary for the computer to recognize the cat in the photo. We collect data — millions of photos of cats — and give (feed) this data to the algorithm.

There can be a lot of layers, but imagine that to solve the problem, you need only 4. Each input of the first layer of neurons receives an incoming pixel of the picture.

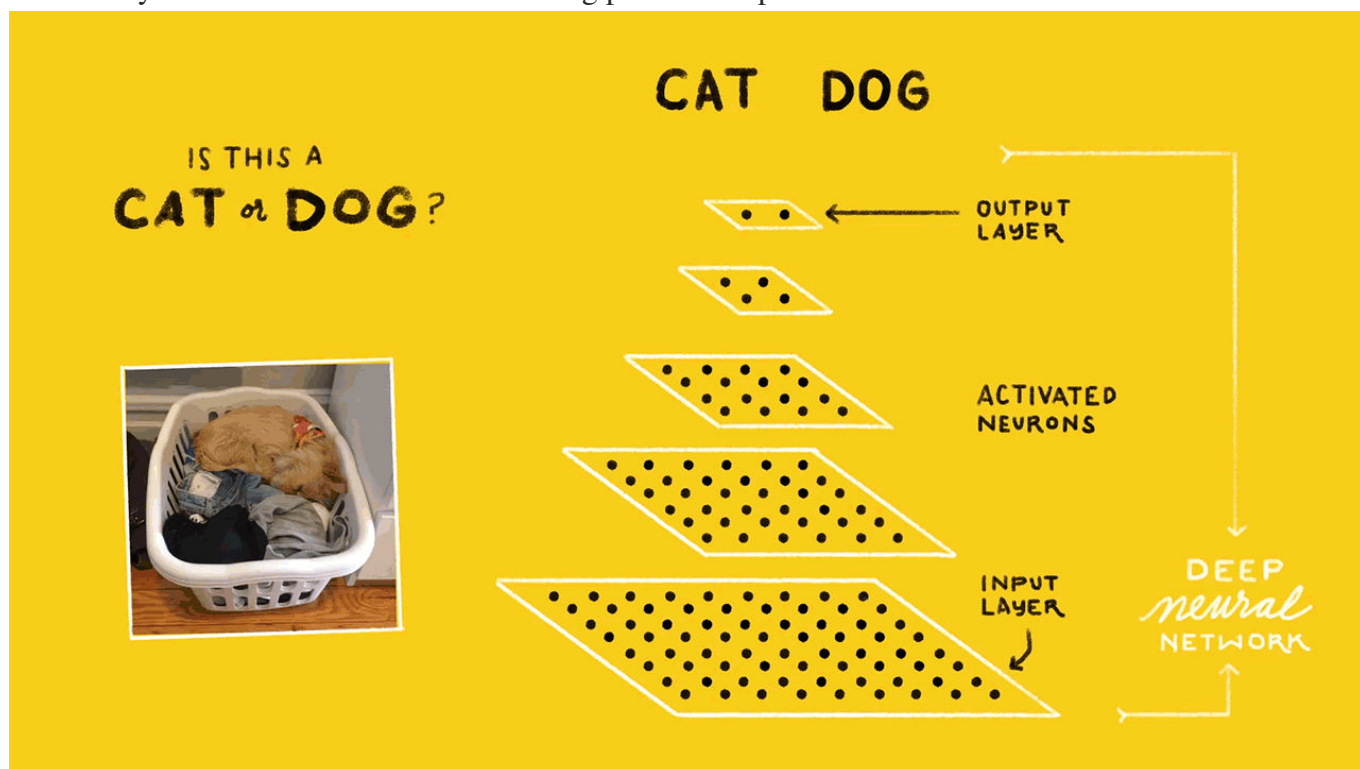


Figure 6: Classification between Cat and Dog

In this case, each next layer combines information obtained at previous levels:

- 1 — the first layer of neurons can only recognize lines, points, and circles. When it understands where these objects are in the photo, it passes the information to the next layer.
- 2 — based on this data, the algorithm determines that the next layer will be able to distinguish between triangles and squares, for example, to understand where are the kitten's ears.
- 3 — when the third layer finds out about this, he is already able to understand where the kitten's head is in the image, and where is the body.
- 4 — combining all the information received, the 4th layer of neurons understands that in front of it is an image of a kitten. So, the training was successful.

But how this magic works on?

Each connection between neurons is associated with a weight. This weight dictates the importance of the input value. The initial weights are set randomly. Predicting the cat on the photo, cat's features are one of the heavier factors. Hence, the cat's features neuron connections will have a big weight.

Each neuron has an Activation Function. These functions are hard to understand without mathematical reasoning. Simply put, one of its purposes is to “standardize” the output from the neuron. Once a set of input data has passed through all the layers of the neural network, it returns the output data through the output layer.

How to train a neural network?

So, we already know how a neural network works. Now we need to understand how the results are calculated. The main difficulty is to assign the right value for each connection in the neural network, and that's why it needs to be trained.

The neural network processes all the data and sets the value to each neuron until it comes to the right conclusions about each element on which the data is collected. At the end of this stage, the value of each element becomes constant, and the neural network can more accurately give predictions.

Traditional machine learning vs deep learning

While traditional machine learning algorithms are linear, deep learning algorithms are stacked in a hierarchy of increasing complexity and abstraction. To understand it better, imagine a toddler whose first word is a dog. The toddler learns what a dog is (and is not) by pointing to objects and saying the word dog. The parent says, “Yes, it is a dog,” or, “No, it is not a dog.”

As the toddler continues to point to objects, he becomes more aware of the features that all dogs possess. What the toddler does, without knowing it, is to clarify a complex abstraction (the concept of dog) by building a hierarchy in which each level of abstraction is created with the knowledge that was gained from the preceding layer of the hierarchy.

Computer programs that use deep learning go through much the same process. Each algorithm in the hierarchy applies a nonlinear transformation on its input and uses what it learns to create a statistical model as output. Iterations continue until the output has reached an acceptable level of accuracy. The number of processing layers through which data must pass is what inspired the label deep.

2.2 Technologies Involved

Google Drive

Google Drive lets you store your company's data in one place and safely share your documents internally and outside your organization with access restrictions which improves collaboration and productivity while securing your organization's data. Google Drive allows easy accessibility of your files and folders from any internet enabled device while putting you in control of how your teams share them. Depending on the SKU you take up, G Suite Basic has 30GB cloud storage shared across both Gmail and Drive while G Suite Business and G Suite Enterprise edition both have unlimited cloud storage.

What can you do with Drive?

1.Upload and store files

You can store any file in Drive: pictures, drawings, videos, and more. You only need to store a file in Drive on one device, and it will automatically be available on all your other devices.

Store files on your desktop

Use less of your PC/Mac disk space & stream directly from the cloud Drive File Stream gives you access to files directly from your computer, without impacting all of your disk space. Spend less time waiting for files to sync and more time being productive.

If you want to work on files from your desktop, install Drive File Stream. All your Drive files appear and can be streamed on demand, so they don't take up all your storage space on your computer. (If you decide later to uninstall Drive File Stream, your Drive files won't be affected. They can still be accessed from Drive on the web.) Drive File Stream is only available if your G Suite administrator has turned it on for your organization or team.

Upload files from your phone or tablet

You can also use the Drive app to store files on your Android or iOS device. (If you decide later to uninstall the app, your Drive files won't be affected and can still be accessed from Drive on the web.)

Access your files

Drive simplifies your work by making the latest version of your file available automatically across the web and all your devices. After you store your files in Drive, you can get to them on any computer, smartphone, or tablet. When you change or delete a file stored in one location, Drive makes the same change everywhere else, so you don't have to.



Figure 7: Google Drive

Google Colab

Whether you are a student interested in exploring Machine Learning but struggling to conduct simulations on enormous datasets, or an expert playing with ML desperate for extra computational power, Google Colab is the perfect solution for you. Google Colab or “the Colaboratory” is a free cloud service hosted by Google to encourage Machine Learning and Artificial Intelligence research, where often the barrier to learning and success is the requirement of tremendous computational power.

Benefits of Colab

Besides being easy to use (which I’ll describe later), the Colab is fairly flexible in its configuration and does much of the heavy lifting for you.

- Python 2.7 and Python 3.6 support
- Free GPU acceleration
- Pre-installed libraries: All major Python libraries like TensorFlow, Scikit-learn, Matplotlib among many others are pre-installed and ready to be imported.
- Built on top of Jupyter Notebook
- Collaboration feature (works with a team just like Google Docs): Google Colab allows developers to use and share Jupyter notebook among each other without having to download, install, or run anything other than a browser.
- Supports bash commands
- Google Colab notebooks are stored on the drive

Setting Free GPU

It is so simple to alter default hardware (CPU to GPU or vice versa); just follow Edit > Notebook settings **or** Runtime>Change runtime type and select GPU **as** Hardware accelerator.

Running or Importing .py Files with Google Colab

Run these codes first in order to install the necessary libraries and perform authorization.



```
from google.colab import drive
drive.mount('/content/drive/')
```

... Go to this URL in a browser: <https://accounts.google.com/o/oauth2/auth?c>

Enter your authorization code:

Figure 8: Access to Google Drive

Click the link, **copy** verification code and **paste** it to text box. After you can access your drive through colab.

Python

What Is Python?

- Interpreted **high-level object-oriented dynamically-typed scripting** language.
- As a result, **run time errors** are usually encountered.

Why Python?

- Python is the most popular language due to the fact that it's easier to code and understand it.
- Python is an object-oriented programming language and can be used to write functional code too.
- It is a suitable language that bridges the gaps between business and developers.
- Subsequently, it takes less time to bring a Python program to market compared to other languages such as C#/Java.
- Additionally, there are a large number of python machine learning and analytical packages.
- A large number of communities and books are available to support Python developers.
- Nearly all types of applications, ranging from forecasting analytical to UI, can be implemented in Python.
- There is no need to declare variable types. Thus it is quicker to implement a Python application.

Why Not Python?

- Python is slower than C++, C#, Java. This is due to the lack of Just In Time optimisers in Python.
- Python syntactical white-space constraint makes it slightly difficult to implement for new coders.
- Python does not offer advanced statistical features as R does.
- Python is not suitable for low-level systems and hardware interaction.

How Does Python Work?

This image illustrates how python runs on our machines:

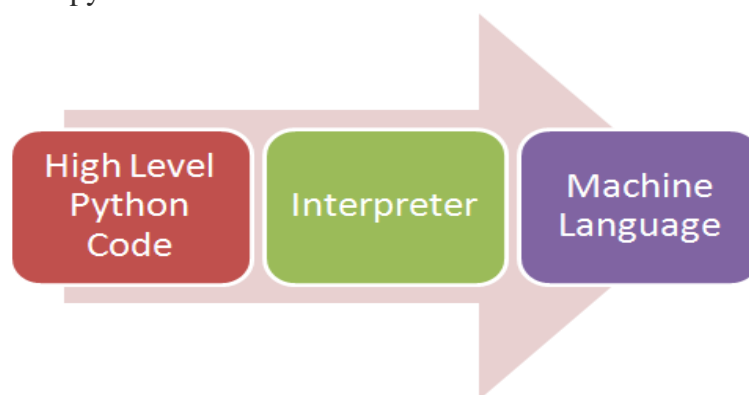


Figure 9: Working of Python

The key here is the Interpreter that is responsible for translating high-level Python language to low-level machine language.

The way Python works is as follows:

1. A Python virtual machine is created where the packages (libraries) are installed. Think of a virtual machine as a container.
2. The python code is then written in .py files
3. CPython compiles the Python code to bytecode. This bytecode is for the Python virtual machine.
4. When you want to execute the bytecode then the code will be interpreted at runtime. The code will then be translated from the bytecode into the machine code. The bytecode is not dependent on the machine on which you are running the code. This makes Python machine-independent.

Tensorflow

TensorFlow is a computational framework for building machine learning models. It is the second generation system from Google Brain headed by Jeff Dean. Launched in early 2017, it has disrupted the ML world by bringing in numerous capabilities from scalability to building production ready models.

The Framework

TensorFlow provides a variety of different tool kits that allow you to write code at your preferred level of abstraction. For instance, you can write code in the Core TensorFlow (C++) and call that method from Python code. You can also define the architecture on which your code should run (CPU, GPU etc.). In the above hierarchy, the lowest level in which you can write your code is C++ or Python. These two levels allow you to write numerical programs to solve mathematical operations and equations. Although this is not highly recommended for building Machine Learning models, it offers a wide range of math libraries that ease your tasks. The next level in which you can write your code is using the TF specific abstract methods which are highly optimised for model components. For example, using the `tf.layers` method abstract you can play with the layers of a neural net. You can build a model and evaluate the model performance using the `tf.metrics` method. The most widely used level is the `tf.estimator` API, which allows you to build (train and predict) production ready models with easy. The estimator API is insanely easy to use and well optimised. Although it offers less flexibility, it has all that is needed to train and test your model. Let's see an application of the estimator API to build a classifier using just three lines of code.

Mostly TensorFlow is used as a backend framework whose modules are called through Keras API. Typically, TensorFlow is used to solve complex problems like Image Classification, Object Recognition, Sound Recognition, etc. In this article, we have learnt about the structure and components of TensorFlow. In the next article, we shall dive into Machine Learning and build our first Linear Regression model using TensorFlow.



Figure 10: TensorFlow

Keras

Keras is a deep learning framework for Python that provides a convenient way to define and train almost any kind of deep learning model. Keras is a high-level neural networks API, written in Python which is capable of running on top of **Tensorflow**, **Theano** and **CNTK**. It was developed for enabling fast experimentation.

Keras doesn't handle low-level operations such as tensor manipulations and differentiation. Instead, it relies on a specialized, well-optimized tensor library to do so which serves as the backend engine of Keras. We can use several backend engines for keras, and currently three existing backend implementations are the Tensorflow backend, the Theano backend, and the Microsoft Cognitive Toolkit (CNTK) backend.

Keras has the following features:

- Allows for easy and fast prototyping
- Run seamlessly on CPU and GPU
- Supports both convolutional networks (for computer vision) and recurrent networks (for sequence and time-series), as well as the combination of two.
- It supports arbitrary network architectures: multi-input or multi-output models, layer sharing, model sharing and so on. This means Keras is appropriate for building deep learning models, from generative adversarial networks to a neural Turing machine.

The typical Keras workflows looks like:

- Define your training data: input tensor and target tensor
- Define a network of layers (or model) that maps input to our targets.
- Configure the learning process by choosing a loss function, an optimizer, and some metrics to monitor.
- Iterate your training data by calling the *fit ()* method of your model.



Figure 11: Keras

METHODOLOGY

3.1 Proposed System

The proposed system detects the walking pattern of different person under different lightning conditions. The need for effective and efficient gait recognition system cannot be overemphasized. This is because gait recognition can be used in a number of different scenarios. One example would be to analyze the video stream from surveillance cameras. If

an individual walks by the camera who's gait has been previously recorded and they are a known threat, then the

system will recognize them and the appropriate authorities can be automatically alerted and the person can be apprehended before they are allowed to become a threat. The threat can be detected from a distance, creating a time buffer for authorities to take action. This system has a large amount of potential application domains, such as airports, banks and general high security area. A gait recognition system was developed to provide a means for identifying humans based on their gait pattern. This system was described and implemented on Intel Core i5, CPU 2.4 GHz, 12GB RAM using Python programming language, The results of the developed system are satisfactory . This system will go a long way in assisting both security agents and different organizations like banks cub threats, aid identification and investigation.

3.2 Architecture of the System

Architecture contains Two sequential triplets of convolution, pooling, and normalization layers, and two subsequent fully connected layers, which outputs a set of similarities to individual training subjects given a GEI as an input.

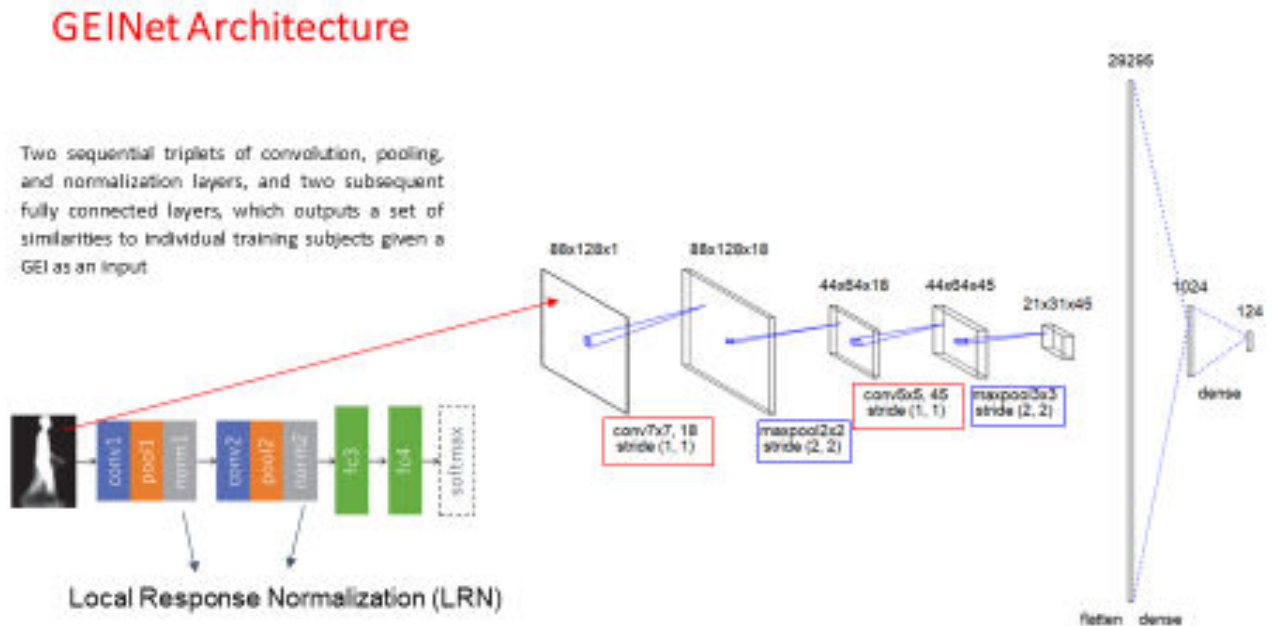


Figure 12: Architecture of our Model.

GEINet v3 contains two convolutional and pooling layers and two dense layers with a flatten layer.

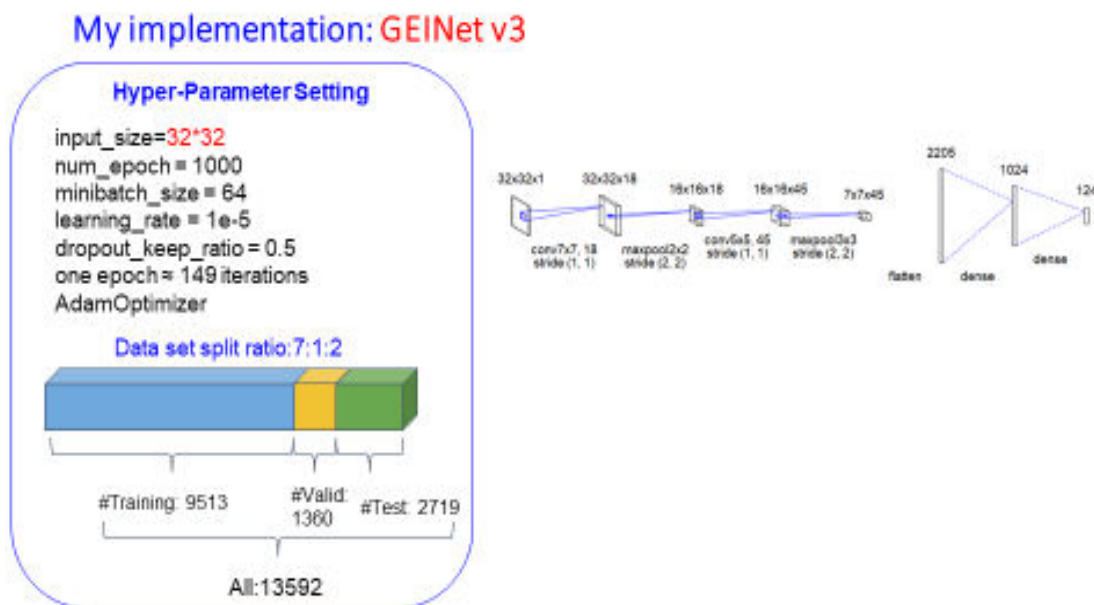


Figure 13: Hyper Parameter Setting for GEINet v3

GEINet v4 contains two convolutional and pooling layers and two dense layers with a flatten layer and two layers for learning rate.

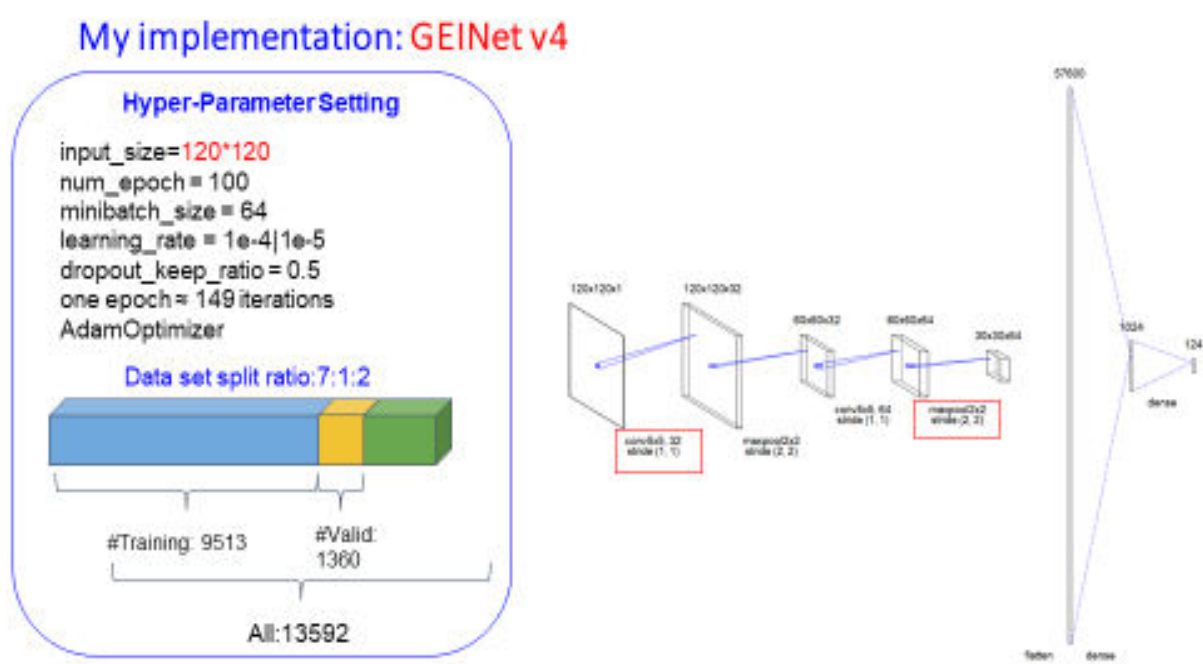


Figure 14: Hyper Parameter Setting for GEINet v4

GEINet v6 contains two convolutional and pooling layers and two dense layers with a flatten layer.

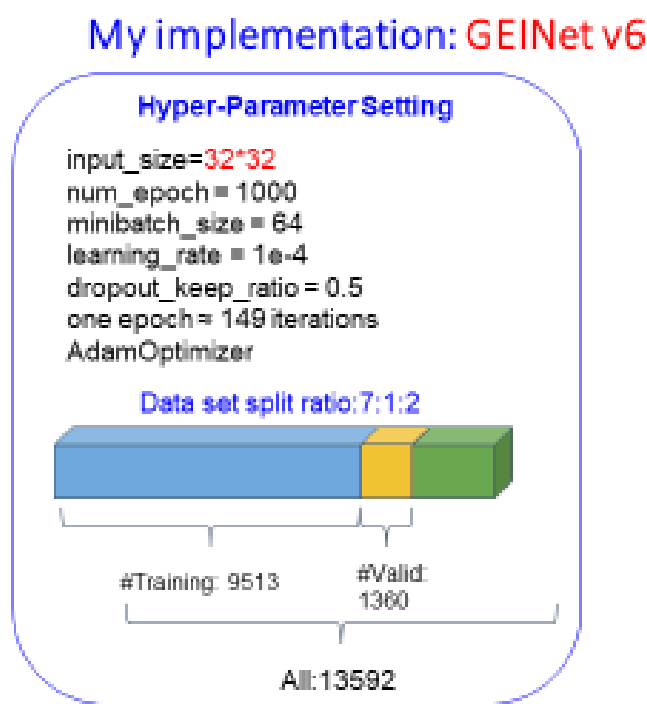


Figure 15: Hyper Parameter Setting for GEINet v6

3.3 Dataset Collection

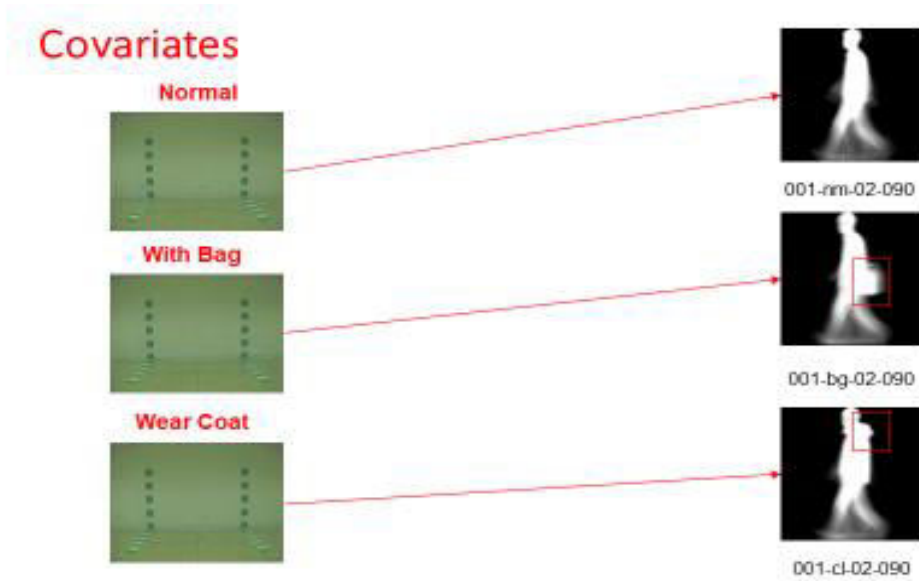


Figure 16: Covariates in the Dataset.

- The CASIA Gait database was used to evaluate the effectiveness of the proposed feature selection algorithms. It is an indoor gait database consisting of 124 subjects captured from 11 different views simultaneously starting from 0° to 180° with an increment of 18° .
- The database has 10 walking sequences for each individual consisting of 6 normal walking sequences (Set A), 2 carrying-bag sequences (Set B) and 2 wearing-coat sequences (Set C).
- The total number of sequences in the database is 13640. We used the first 4 sequences of each individual in Set A as the training set (Set A1) and the rest as the test set including the rest sequences in Set A (Set A2), Set B and Set C.
- The original image size of the database is 320×240 . After size normalization, the size of the GEIs became 128×88 (i.e. the original feature space has a dimensionality of 11264).

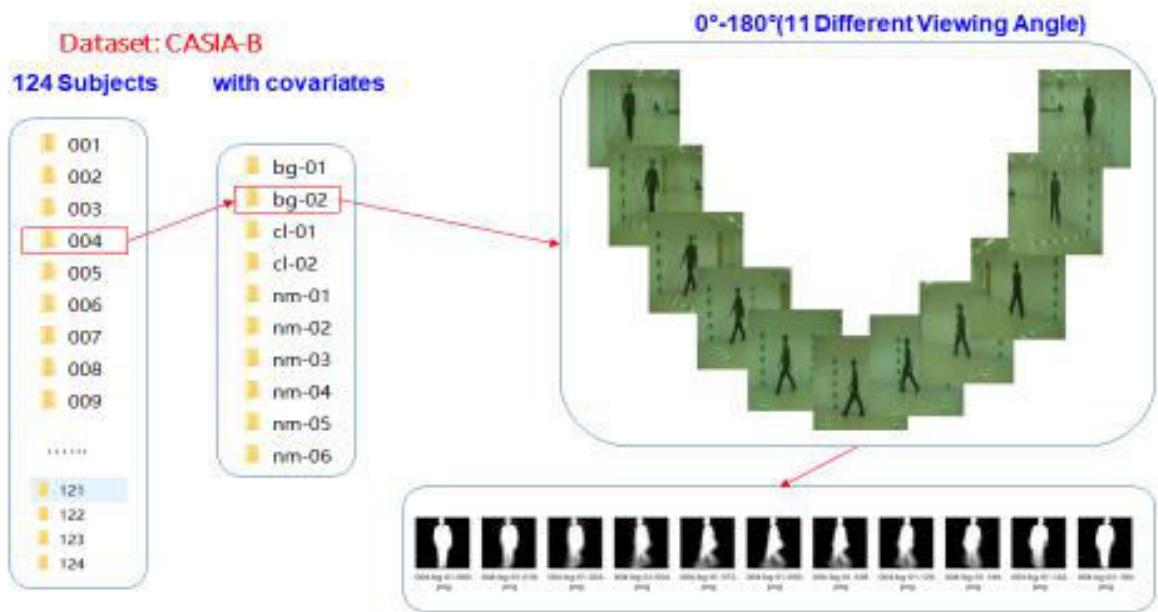


Figure 17: Dataset Description.

3.4 Feature Extraction

Our proposed system takes the input Gait energy images(GEI) and extracts the edge features from the image using neural networks which automatically detect the edge features. **Edge detection** includes a variety of mathematical methods that aim at identifying points in a digital image at which the image brightness changes sharply or, more formally, has discontinuities. The points at which image brightness changes sharply are typically organized into a set of curved line segments termed *edges*. The same problem of finding discontinuities in one-dimensional signals is known as step detection and the problem of finding signal discontinuities over time is known as change detection. Edge detection is a fundamental tool in image processing, machine vision and computer vision, particularly in the areas of feature detection and feature extraction. The purpose of detecting sharp changes in image brightness is to capture important events and changes in properties of the world. It can be shown that under rather general assumptions for an image formation model, discontinuities in image brightness are likely to correspond to:

- discontinuities in depth,
- discontinuities in surface orientation,
- changes in material properties and
- variations in scene illumination.

In the ideal case, the result of applying an edge detector to an image may lead to a set of connected curves that indicate the boundaries of objects, the boundaries of surface markings as well as curves that correspond to discontinuities in surface orientation. Thus, applying an edge detection algorithm to an image may significantly reduce the amount of data to be processed and may therefore filter out information that may be regarded as less relevant, while preserving the important structural properties of an image. If the edge detection step is successful, the subsequent task of interpreting the information contents in the original image may therefore be substantially simplified. However, it is not always possible to obtain such ideal edges from real life images of moderate complexity.

Set Description

The set is divided into three parts i.e. the training set, testing set and validation set. These sets are taken in a ratio of 7:2:1.

Detailed Design Methodologies

The major steps involved consisted of data collection, data acquisition, classifier design, training and testing. The CNN was implemented using Python 3.7. Training was performed on a laptop with a 12 GB of memory.

For data collection the dataset taken was CASIA-B. It consists of 124 different subjects with silhouette image sequences present at different angles like 0,18,36,54,72,90,108,126,144,162 and 180 degrees.

At first we extracted all the silhouettes of all the subjects and set them in a loop. Then we clipped all the sequences and calculated the Region of Interest Points (ROI). After that we calculated the

gait period cycle which is the time period or sequence of events or movements during locomotion in which one foot contacts the ground to when that same foot again contacts the ground, and involves propulsion of the center of gravity in the direction of motion. After having the number of images involved in the Gait cycle we combined all of those images and averaged out them to calculate the Gait energy images of every subject at every angle. Here we are dealing with normal images which does not include any bag, coat etc. After calculating the GEI we then set them in the directories accordingly.

We extracted the GEI and made a program which will store all the energy images in a `X_train` variable which will then dumped or saved in a pkl (pickle) extension file. Subject no. corresponding to it are stored in a `Y_train` variable and stored as well. We made two pkl files. One contains the GEI of size 32×32 and the other one containing size 120×120 .

We now make the three different architectures that are V3, V4 and V6 , V3 and V6 architecture requires an input image size of 32×32 whereas V4 requires input image size equal to 120×120 . Now we passed all the data through Convolutional layers which will filter out the features and we trained them. V3 and V6 were trained on 1000 epochs whereas V4 was trained on 100 epochs and validated it on 10 percent of the partitioned data. All the results were then plotted and accuracy was compared.

RESULTS

4.1 CMC Curve

- Each probe biometric sample is compared against all gallery samples
- The resulting scores are sorted and ranked •Determine the rank at which a true match occurs
- True Positive Identification Rate (TPIR): Probability of observing the correct identity within the top K ranks
- CMC Curve: Plots TPIR against ranks

V3 CMC Curve

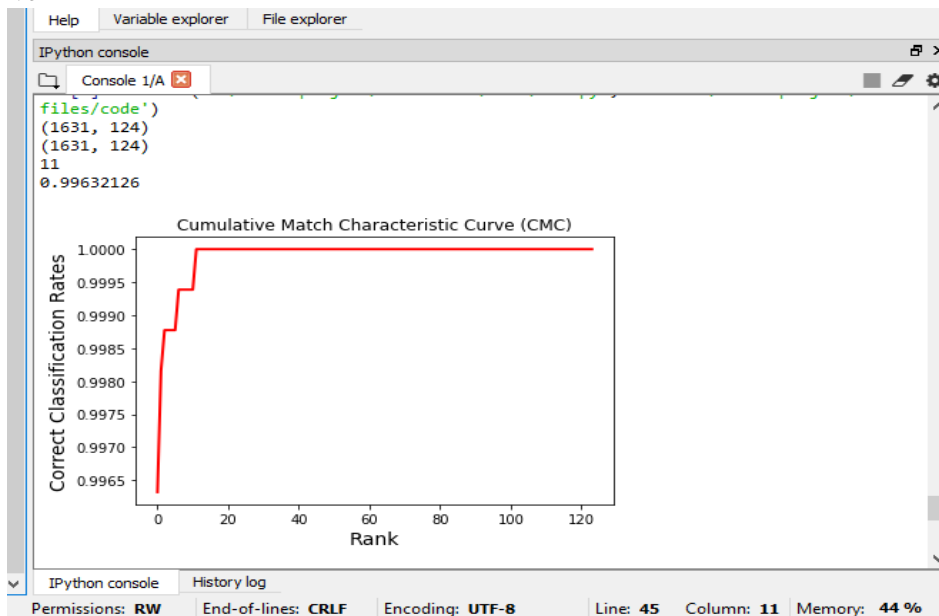


Figure 18: CMC Curve for V3

V4_CMC_CURVE

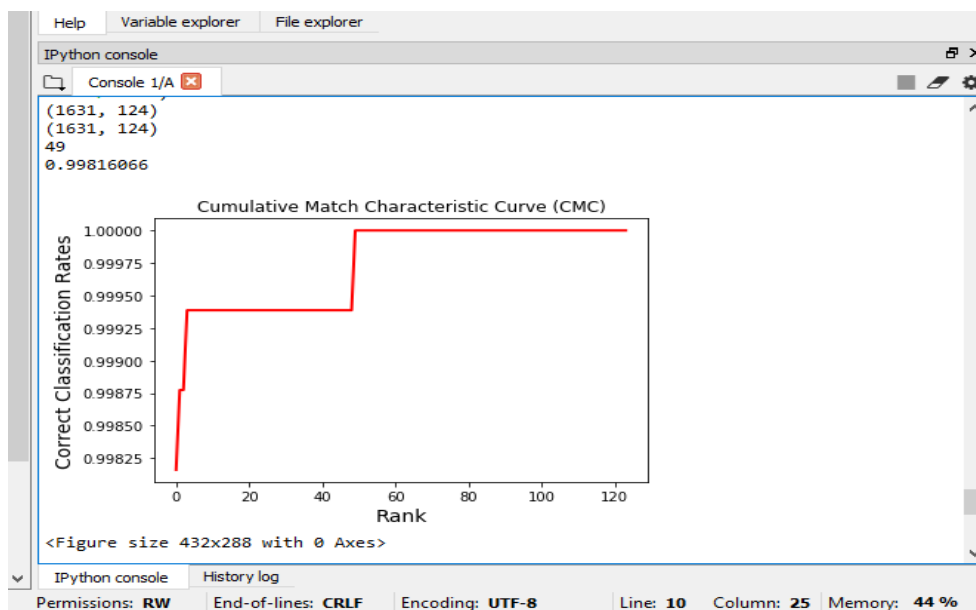


Figure 19: CMC Curve for V4

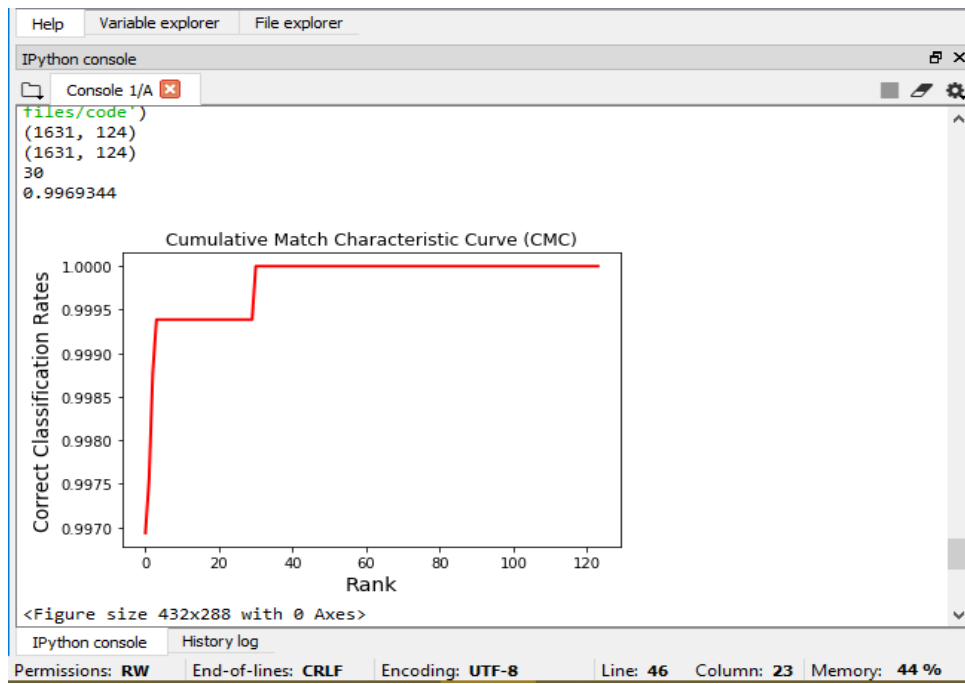


Figure 20: CMC Curve for V6

4.2 ROC Curve

- Biometrics samples are compared against each other
- Genuine and impostor scores are generated
- False Match Rate (FMR) and False Non-match Rate (FNMR) are computed at multiple thresholds
- ROC Curve: True Match Rate versus False Match Rate
- ROC Curve: Aggregate Statistics

V3

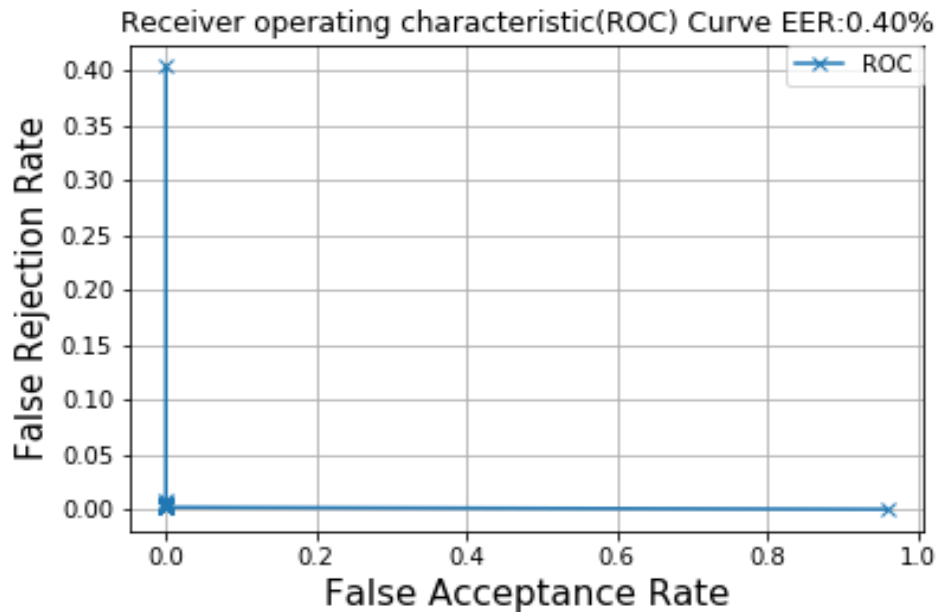


Figure 21: ROC Curve for V3

V4_LRN

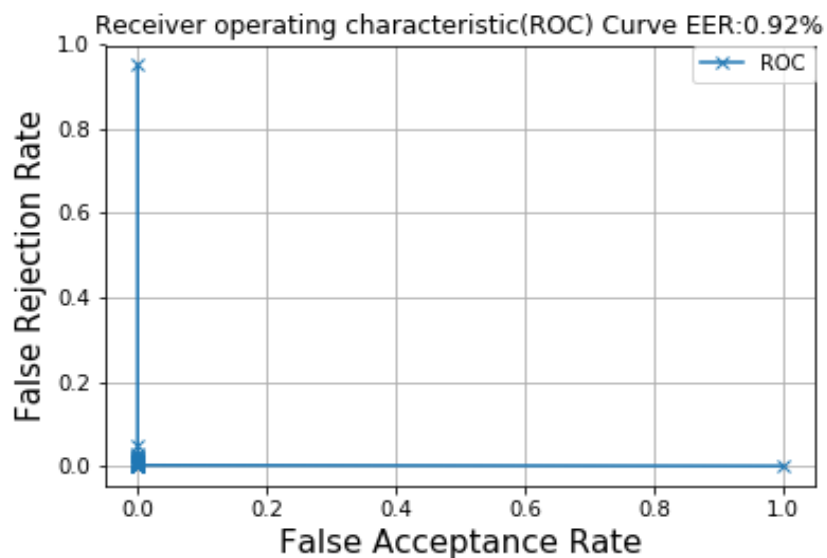


Figure 22: ROC Curve for V4

V6

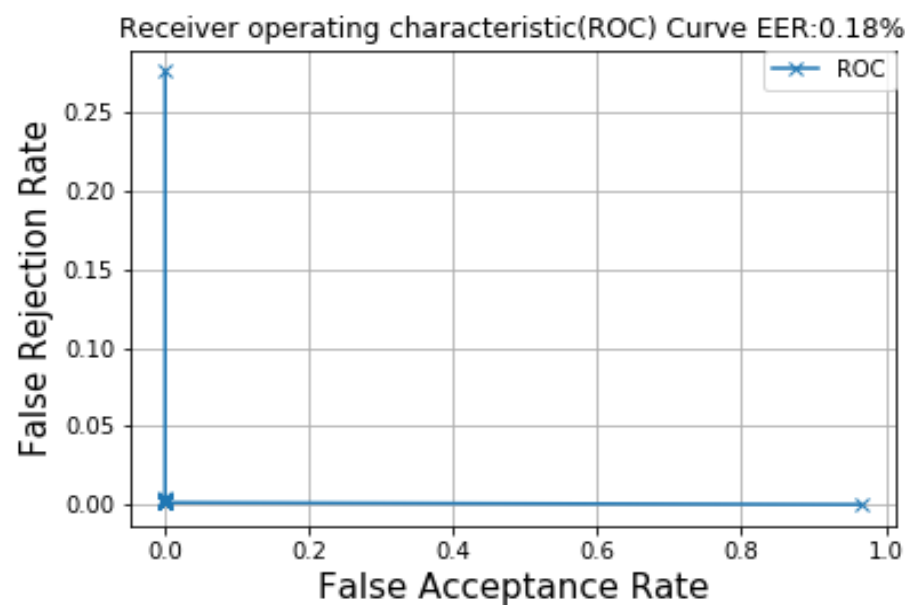


Figure 23: ROC Curve for V6

4.3 Training Process Plot

The concept of the Learning Curve basically states that there is less and less learning as more repetitive steps are taken. The Boston Consulting Group conducted some empirical studies and below are the conclusions from that study:

- The time required to perform a task decreases as the task is repeated.
- The amount of improvement decreases as more units are produced, and
- The rate of improvement has sufficient consistency to allow its use as a prediction tool.

V3

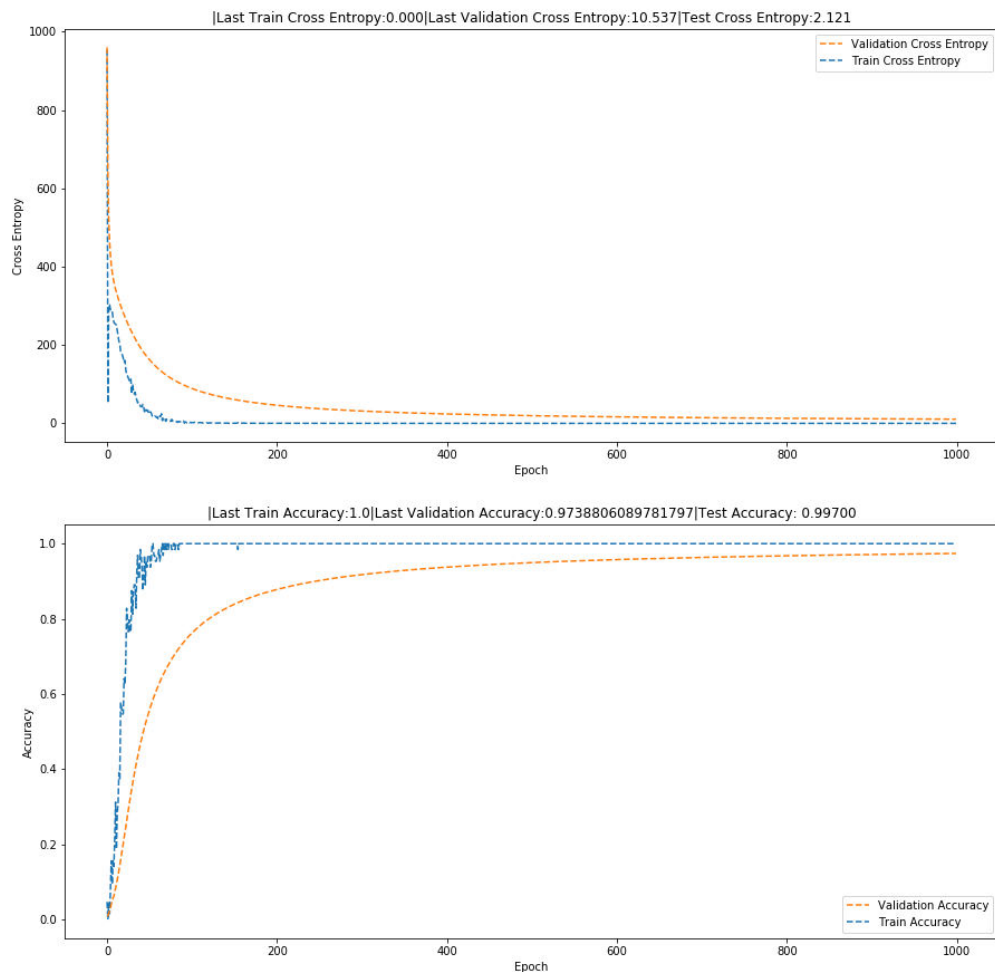


Figure 24: Training Process Plot for V3

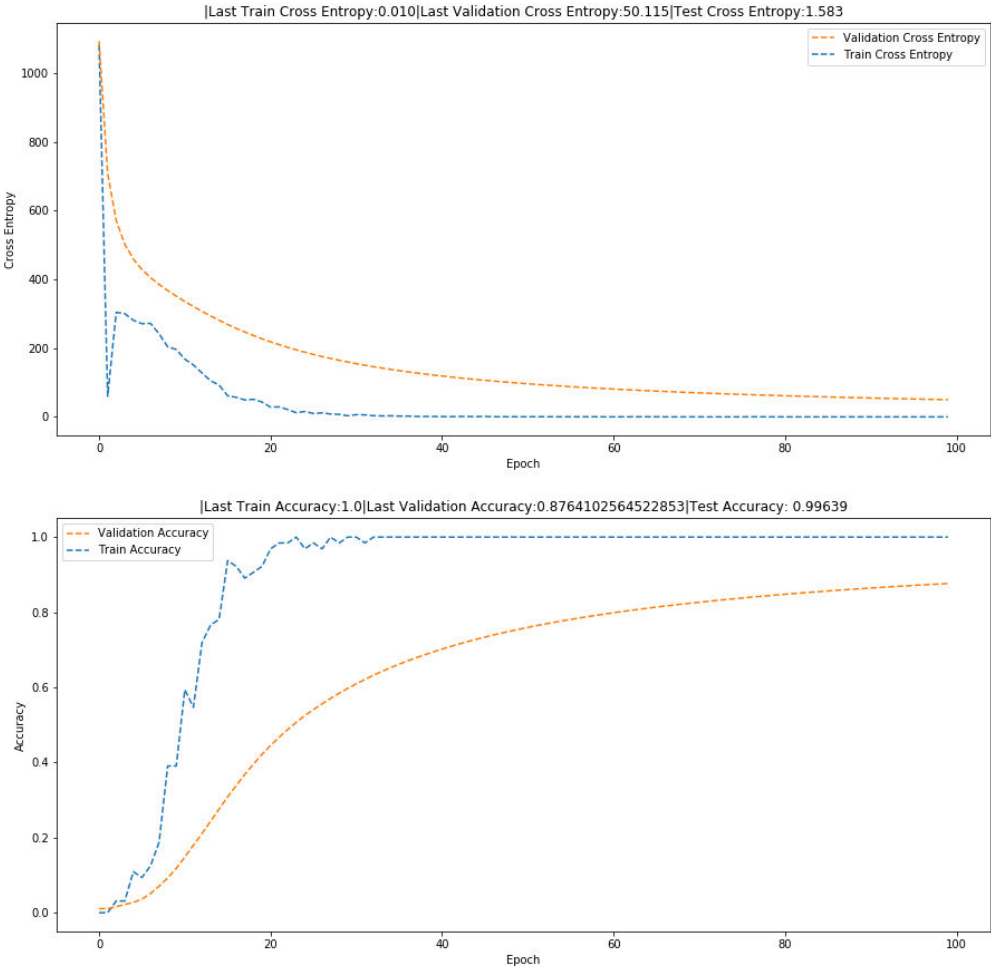


Figure 25: Training Process Plot for V4

V6

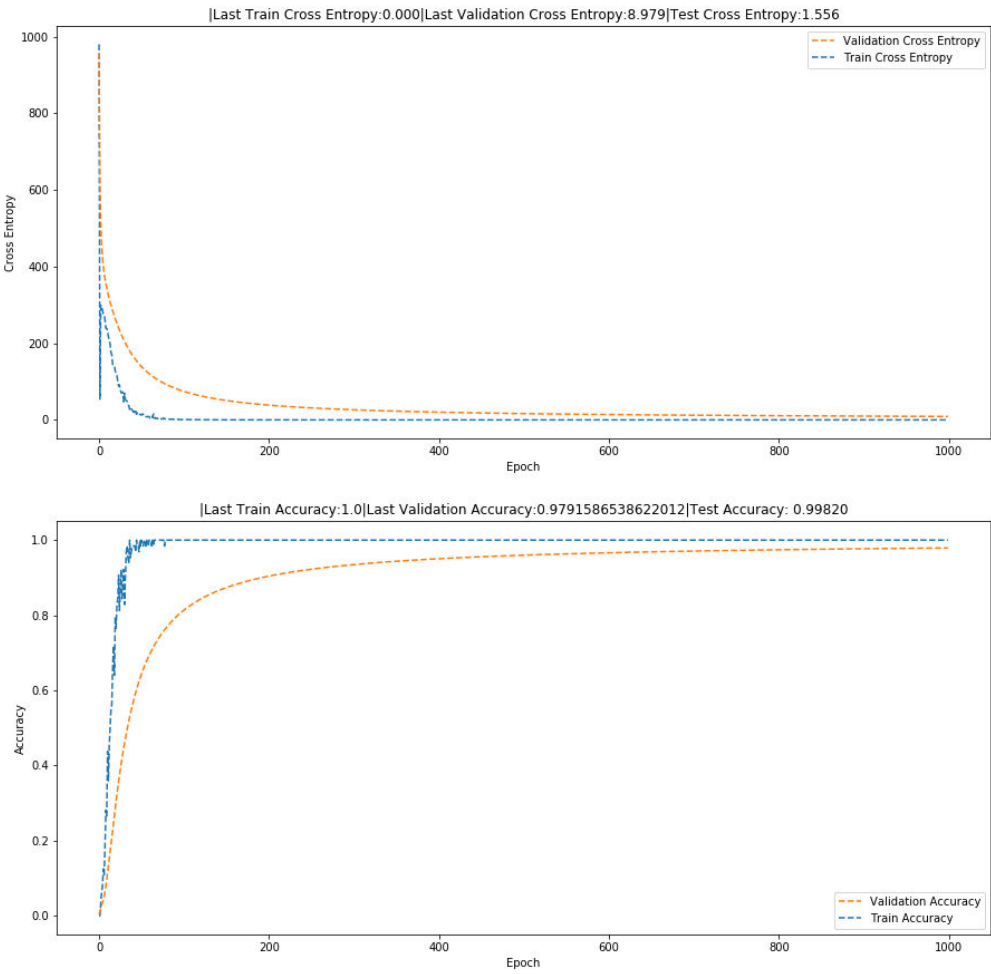


Figure 26: Training Process Plot for V6

CONCLUSION AND FUTURE SCOPE

The main work encompassed gait identification and gait analysis of individuals. First, a novel approach for identifying individuals was proposed. This identification method achieved very high accuracy in a data sample with very similar subjects. Then, a solution was provided to the question: which of these features should be extracted to represent gait and why. A novel gait cycle and its corresponding phases were defined. The influence of gait features from the gait phases was investigated.

In addition, the relationship between gait and attractiveness was analyzed and a predictable model for gait attractiveness was built. The similarity and dissimilarity between the left and right sides of the body in gait were investigated. This research showed the most asymmetric body parts in each subjects' in gait and revealed the common similarities and asymmetric appearances in gait among different subjects.

Input Size	Learning Rate	Model	EER(%)	Rank1 ACC
32*32	1e-5	V3	0.40	99.63%
32*32	1e-4 1e-5	V6	0.18	99.69%
120*120	1e-4 1e-5	v4+LRN	0.97	99.81%

Figure 27: Comparison Table

We conclude that the accuracy at image size of 32*32 is best achieved at 99.69%. and when the image size was 120*120 it was 99.81.

Future Scope

Prominent future research domain demands on the effective fusion of gait with other biometrics so as to accomplish a comprehensive surveillance system in environments that are characterized by “special interest”, high traffic, much transient in nature such as airports. Encompassing the GRS as a part, in launching a comprehensive multi model Automated Human Recognition System is to be experimented in depth and to be optimistically realized in the near future.

REFERENCES

1. Murray, M.P., Drought, A.B., Kory, R.C.: Walking patterns of normal men. *J Bone Joint Surg Am* 46(2), 335{360 (1964)
2. Cutting, J.E., Kozlowski, L.T.: Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the psychonomic society* 9(5), 353{356 (1977)
3. Hossain, E., Chetty, G.: Multimodal feature learning for gait biometric based human identity recognition. In: Lee M., Hirose A., Hou ZG., Kil R.M. (eds.) *ICONIP 2013. LNCS*, vol. 8227, pp. 721{728. Springer, Berlin, Heidelberg (2013)
4. Alotaibi, M., Mahmood, A.: Improved gait recognition based on specialized deep convolutional neural networks. In: *2015 IEEE Applied Imagery Pattern Recognition Workshop*. pp. 1{7. IEEE Press, New York (2015)
5. Wolf, T., Babaee, M., Rigoll, G.: 2016 Multi-view gait recognition using 3d convolutional neural networks. In: *IEEE International Conference on Image Processing*. pp. 4165{4169. IEEE Press, New York (2016)
6. Shiraga, K., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: Geinet: View-invariant gait recognition using a convolutional neural network. In: *Biometrics(ICB), 2016 International Conference on*. pp. 1{8. IEEE Press, New York (2016)
7. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. _sherfaces:Recognition using class speci_c linear projection. *IEEE Transactions on pattern analysis and machine intelligence* 19(7), 711{720 (1997)
8. Mansur, A., Makihara, Y., Muramatsu, D., Yagi, Y.: Cross-view gait recognition using view-dependent discriminative analysis. In: *2014 IEEE International Joint Conference on Biometrics (IJCB)*. pp. 1{8. IEEE Press, New York (2014)