MAJOR PROJECT ON:DATA PROFILING AND INSIGHTS

Major Project Submitted in Fulfillment of the Requirements

for the degree of

Master of Computer Application By

Samhita Keshri
30701017021
Anisha Mitra
30701017045

Under the guidance of

Vishal Kumar Yadav

Third Eye Data Analytics Services India Pvt Ltd.
8th Floor, Plot, 37/2, GN Block, Sector V, Bidhannagar, Kolkata,
West Bengal 700091
2020

THIRDEYE

Third Eye Data Analytics Services India Pvt Ltd.
8th Floor, Plot, 37/2, GN Block, Sector V, Bidhannagar, Kolkata, West Bengal 700091

CERTIFICATE OF RECOMMENDATION

This is to certify that Samhita Keshri, Anisha Mitra have completed their project work titled “Major project on: Data Profiling And Insights “ under the direct supervision and guidance of Mr.Vishal Kumar Yadav. We are satisfied with their work,which is being presented for the fulfillment of the degree of Master of Computer Application (MCA), Maulana Abul Kalam Azad University of Technology, Kolkata 700064.

Vishal kr yadav

____

(Data Scientist and Delivery Lead)
Vishal Kumar Yadav
Date:30.04.2020

TECHNO MAIN,SALTLAKE

FACULTY OF MCA DEPARTMENT

CERTIFICATE OF APPROVAL

The foregoing Major project is hereby approved as a creditable study of Master of Computer Application (MCA) and presented in a manner satisfactory to warrant its acceptance as a pre-requisite to the degree for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or any statement made, opinion expressed or conclusion therein but approve this Minor project only for the purpose for which it is submitted.

Signature of Examiners

GOAL OF THE MINOR PROJECT

The project “Data Profiling And Insights” is aimed to promote live streaming of data using Kafka and zookeeper . Whenever any file comes in a specified folder the data is consumed and then produced to a dataframe using spark.the dataframes are then written to csv files . The csv files are then profiled and analysed to be cleaned. After that the cleaned dataframes are written to csv files. Thereafter meaningful insights are obtained from the cleaned csv files which are plotted using highcharts.

CONTENT

ACKNOWLEDGEMENT

We hereby take the opportunity to thank Mr Vishal Kumar Yadav,Data Scientist and Delivery Head,Third Eye Data Analytics India Pvt Ltd. for allowing us towork on data profiling and for providing us with all the necessary facilities to make our project work and of worth.

Samhita Keshri

Samhita Keshri

Anisha Mitra

Anisha Mitra

INTRODUCTION

Big Data is also **data** but with a **huge size**. Big Data is a term used to describe a collection of data that is huge in volume and yet growing exponentially with time. In short such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.

In simple terms, Kafka is a messaging system that is designed to be fast, scalable, and durable. It is an open-source stream processing platform.It aims at providing a high-throughput, low-latency platform for handling real-time data feeds.

Apache describes Kafka as a distributed streaming platform that lets us:

1. Publish and subscribe to streams of records.
2. Store streams of records in a fault-tolerant way.
3. Process streams of records as they occur.

Apache Spark is a lightning-fast cluster computing technology, designed for fast computation. It is based on Hadoop MapReduce and it extends the MapReduce model to efficiently use it for more types of computations, which includes interactive queries and stream processing. The main feature of Spark is its in-memory cluster computing that increases the processing speed of an application.Spark is designed to cover a wide range of workloads such as batch applications, iterative algorithms, interactive queries and streaming.Apart from supporting all these workload in a respective system, it reduces the management burden of maintaining separate tools.

BACKGROUND

SOFTWARE SPECIFICATION:

Apache Spark

Apache Kafka

Python

linux (O.S)

HARDWARE SPECIFICATION:

CPU: unless ssl or compression are required no fast cpu is needed.LZ4 codec is required

RAM: minimum 6GB of RAM is required.with heavy production upto 32GB is required.

DISK: Kafka thrives when using multiple drives in a RAID setup. SSDs don't deliver much of an advantage due to Kafka's sequential disk I/O paradigm, and NAS should not be used.

NETWORK AND FILESYSTEM: XFS is recommended.

BACKGROUND

KAFKA:

Kafka is generally used for two broad classes of applications:

- Building real-time streaming data pipelines that reliably get data between systems or applications
- Building real-time streaming applications that transform or react to the streams of data

To understand how Kafka does these things, let's dive in and explore Kafka's capabilities from the bottom up.

First a few concepts:

- Kafka is run as a cluster on one or more servers that can span multiple datacenters.
- The Kafka cluster stores streams of *records* in categories called *topics*.
- Each record consists of a key, a value, and a timestamp.

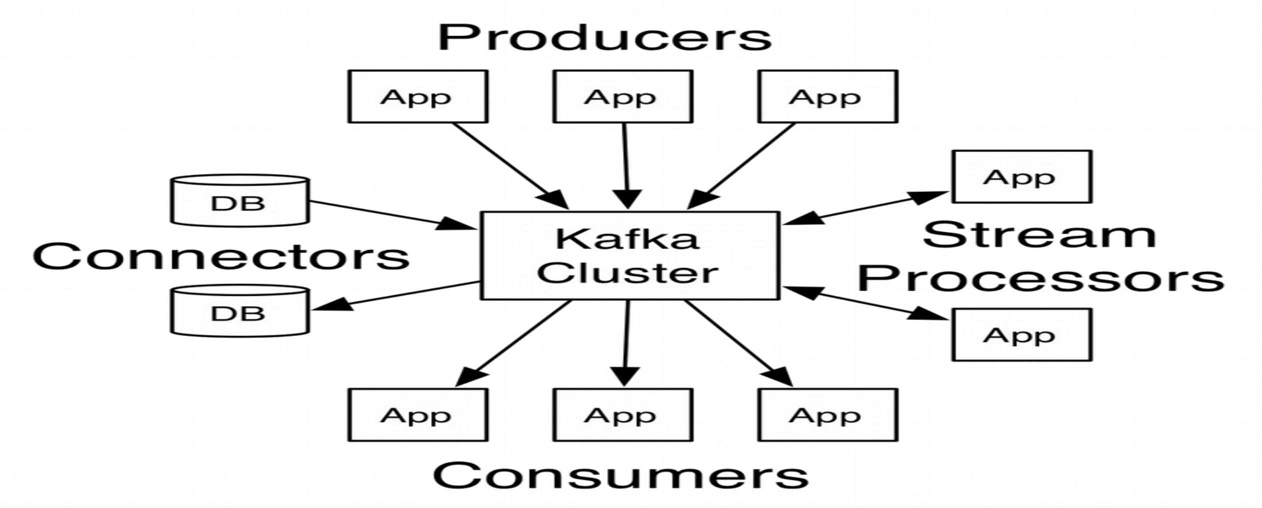
Kafka has five core APIs:

- The Producer API allows an application to publish a stream of records to one or more Kafka topics.
- The Consumer API allows an application to subscribe to one or more topics and process the stream of records produced to them.
- The Streams API allows an application to act as a *stream processor*, consuming an input stream from one or more topics and producing an output stream to one or more output topics, effectively transforming the input streams to output streams.
- The Connector API allows building and running reusable producers or consumers that connect Kafka topics to existing applications or data

systems. For example, a connector to a relational database might capture every change to a table.

- The Admin API allows managing and inspecting topics, brokers and other Kafka objects.

In Kafka the communication between the clients and the servers is done with a simple, high-performance, language agnostic TCP protocol. This protocol is versioned and maintains backwards compatibility with older versions. We provide a Java client for Kafka, but clients are available in many languages.



SPARK:

Spark is an Apache project advertised as “lightning fast cluster computing”. It has a thriving open-source community and is the most active Apache project at the moment.

Spark provides a faster and more general data processing platform. Spark lets you run programs up to 100x faster in memory, or 10x faster on disk, than Hadoop. Last year, Spark took over Hadoop by completing the 100 TB

Daytona GraySort contest 3x faster on one tenth the number of machines and it also became the fastest open source engine for sorting a petabyte.

Spark also makes it possible to write code more quickly as you have over 80 high-level operators at your disposal. To demonstrate this, let’s have a look at the “Hello World!” of BigData: the Word Count example. Written in Java for MapReduce it has around 50 lines of code, whereas in Spark (and Scala) you can do it as simply as this:

```
sparkContext.textFile("hdfs://...")
          .flatMap(line => line.split(" "))
          .map(word => (word, 1)).reduceByKey(_ + _)
          .saveAsTextFile("hdfs://...")
```

Another important aspect when learning how to use Apache Spark is the interactive shell (REPL) which it provides out-of-the box. Using REPL, one can test the outcome of each line of code without first needing to code and execute the entire job. The path to working code is thus much shorter and ad-hoc data analysis is made possible.

DATA PROFILING:

Data profiling is the process of examining the data available from an existing information source (e.g. a database or a file) and collecting statistics or informative summaries about that data. The purpose of these statistics may be to:

1. Find out whether existing data can be easily used for other purposes
2. Improve the ability to search data by tagging it with keywords, descriptions, or assigning it to a category
3. Assess data quality, including whether the data conforms to particular standards or patterns
4. Assess the risk involved in integrating data in new applications, including the challenges of joins

5. Discover metadata of the source database, including value patterns and distributions, key candidates, foreign-key candidates, and functional dependencies
6. Assess whether known metadata accurately describes the actual values in the source database
7. Understanding data challenges early in any data intensive project, so that late project surprises are avoided. Finding data problems late in the project can lead to delays and cost overruns.
8. Have an enterprise view of all data, for uses such as master data management, where key data is needed .

DATA CLEANING:

Data cleansing or **data cleaning** is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting.

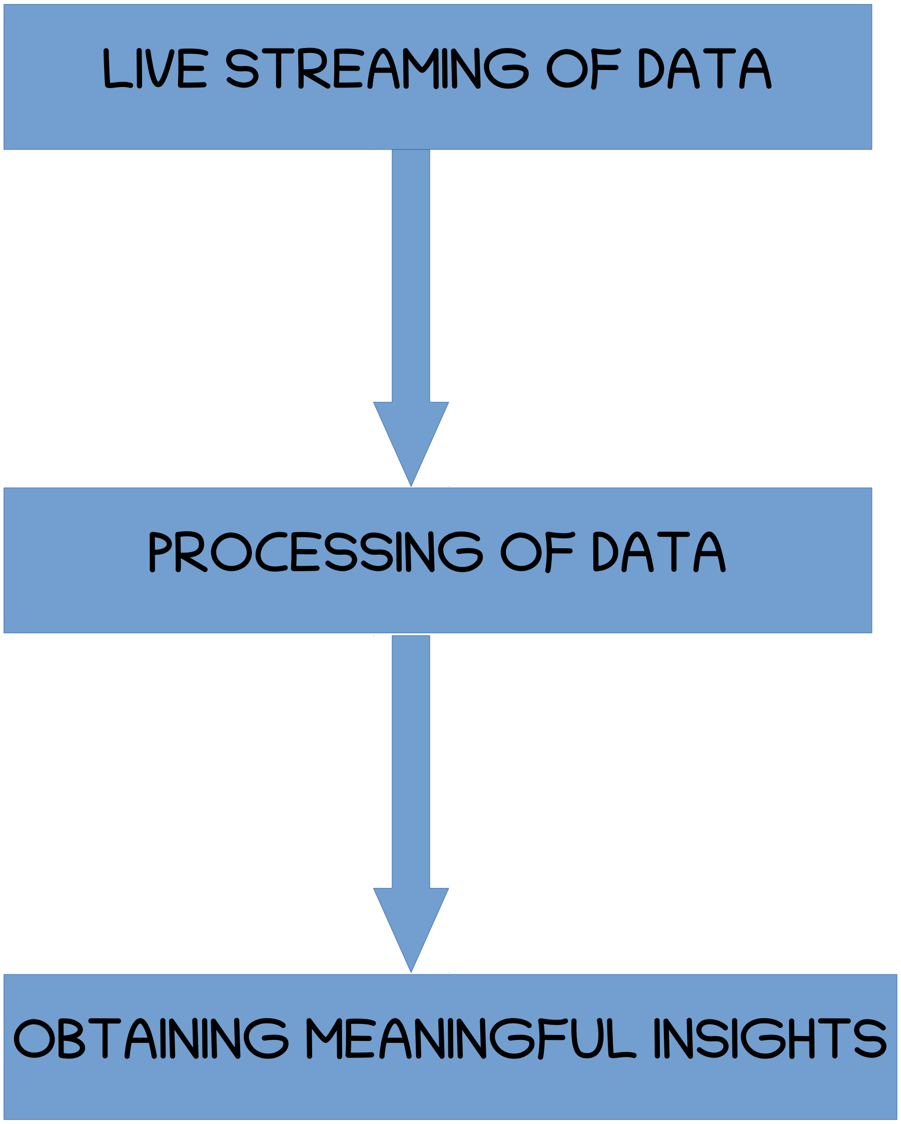
After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleaning differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data.

The actual process of data cleansing may involve removing typographical errors or validating and correcting values against a known list of entities. The validation may be strict (such as rejecting any address that does not have a valid postal code) or fuzzy (such as correcting records that partially match existing, known records). Some data cleansing solutions will clean data by cross-checking with a validated data set. A common data cleansing practice is data enhancement, where data is made more complete by adding related information. For example, appending addresses with any phone numbers related to that address. Data cleansing may also involve harmonization (or normalization) of data, which is the process of bringing together data of "varying file formats, naming conventions, and columns", and transforming it into one cohesive data set.

CSV FILES:A Comma Separated Values (CSV) file is a plain text file that contains a list of data. These files are often used for exchanging data between different applications. For example, databases and contact managers often support CSV files.These files may sometimes be called Character Separated Values or Comma Delimited files. They mostly use the comma character to separate (or delimit) data, but sometimes use other characters, like semicolons.

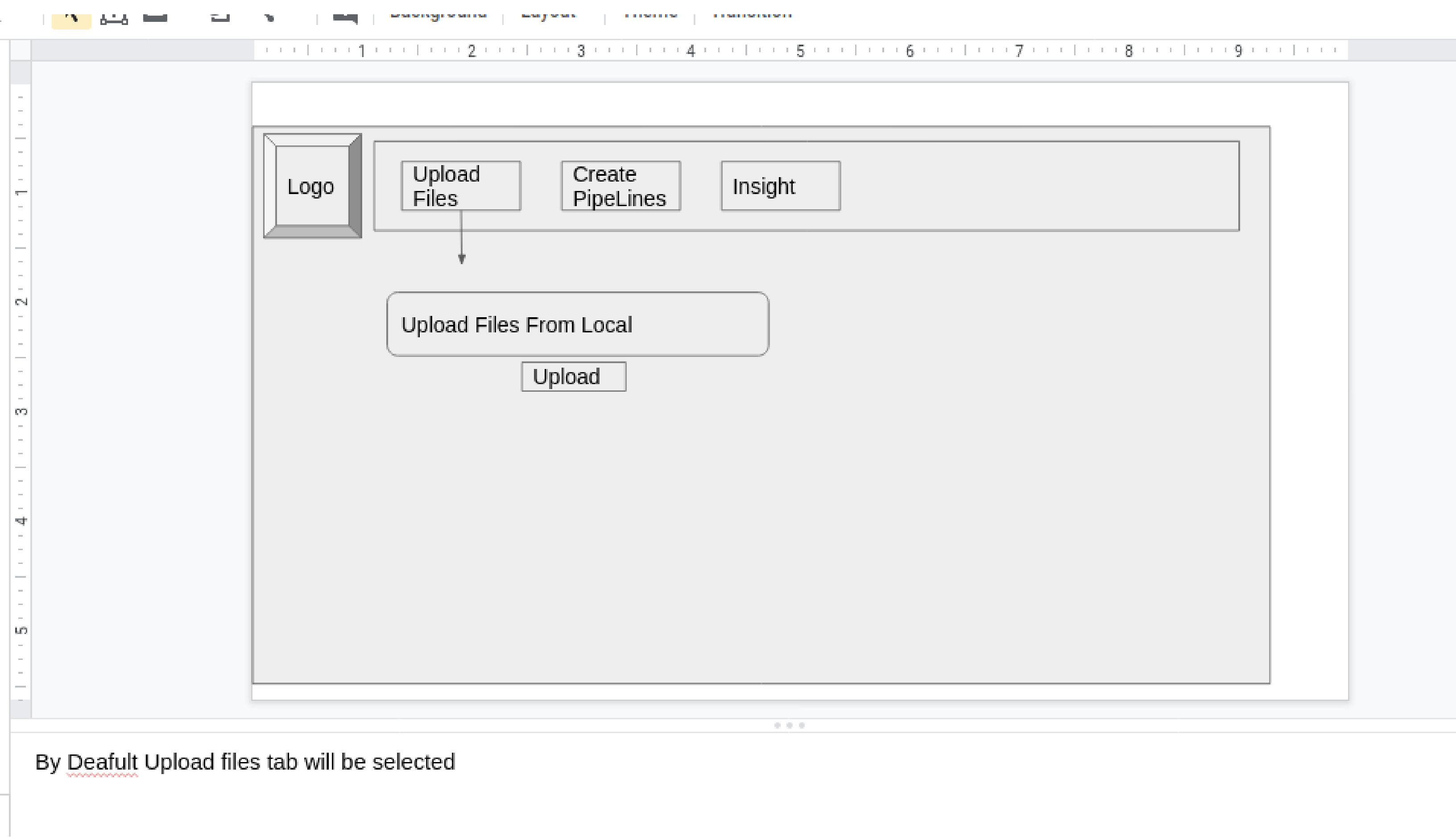
HIGHCHARTS: **Highcharts** is a charting library written in pure **JavaScript**, offering an easy way of adding interactive charts to your web site or web application.

FLOW DIAGRAM



LIVE STREAMING OF DATA

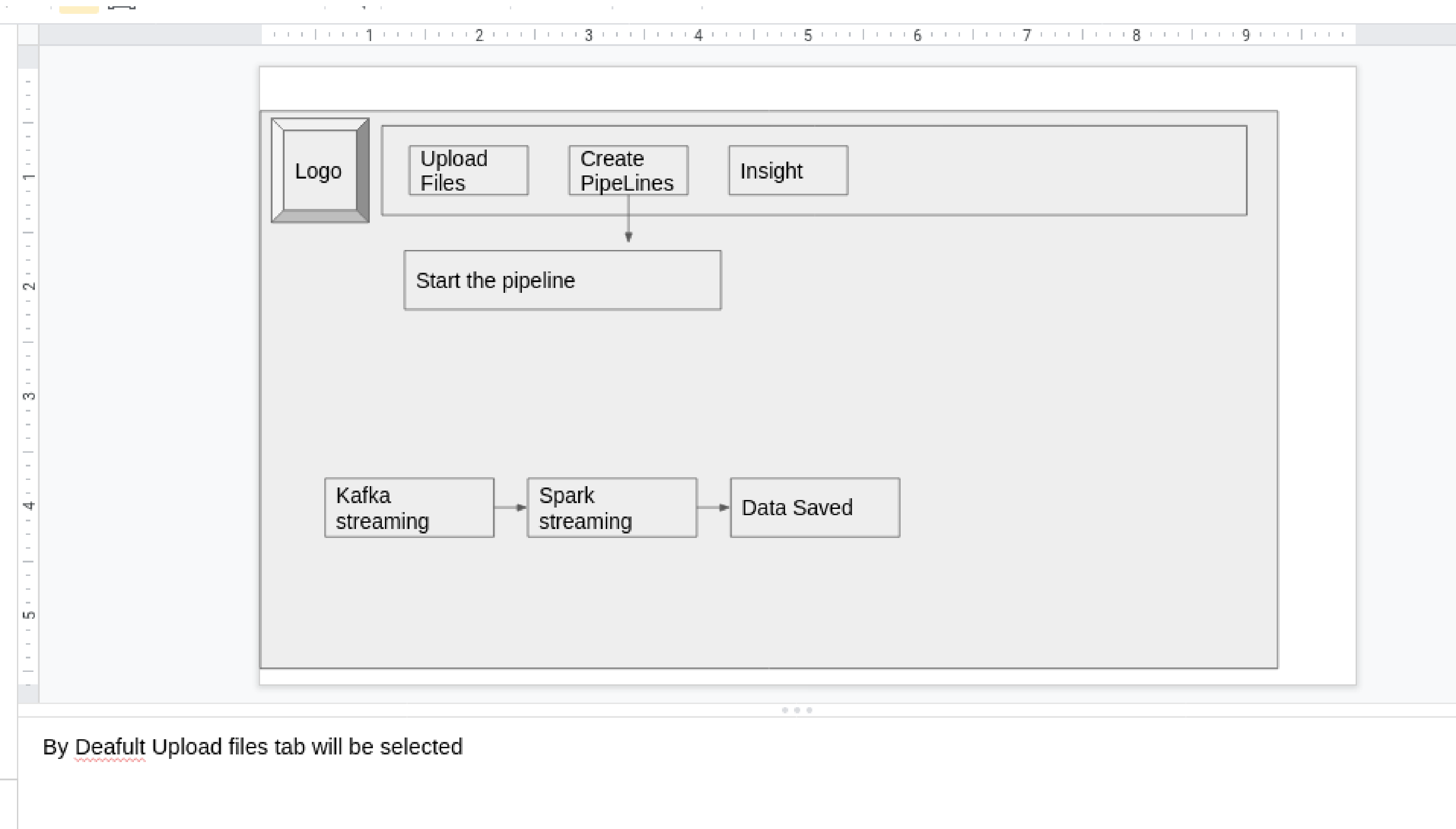
There are two event watchers . Each has different purpose.the first event watcher which is called watchdog1 resembles the role of kafka,wich is live streaming of data.in this case there is a specified location,(in this case it is a folder.)the event watcher that is watchdig1 listens to the folder.whenever there is any file in the folder,the event watcher sends the path of the file to the producer program of the kafka.the file is read line by line and simultaneously the contents are send to the consumer program using kafka.thus the live streaming is concluded by the principle that whenever a file is present in a specified location its contents are read .



By Deafult Upload files tab will be selected

PROCESSING OF DATA

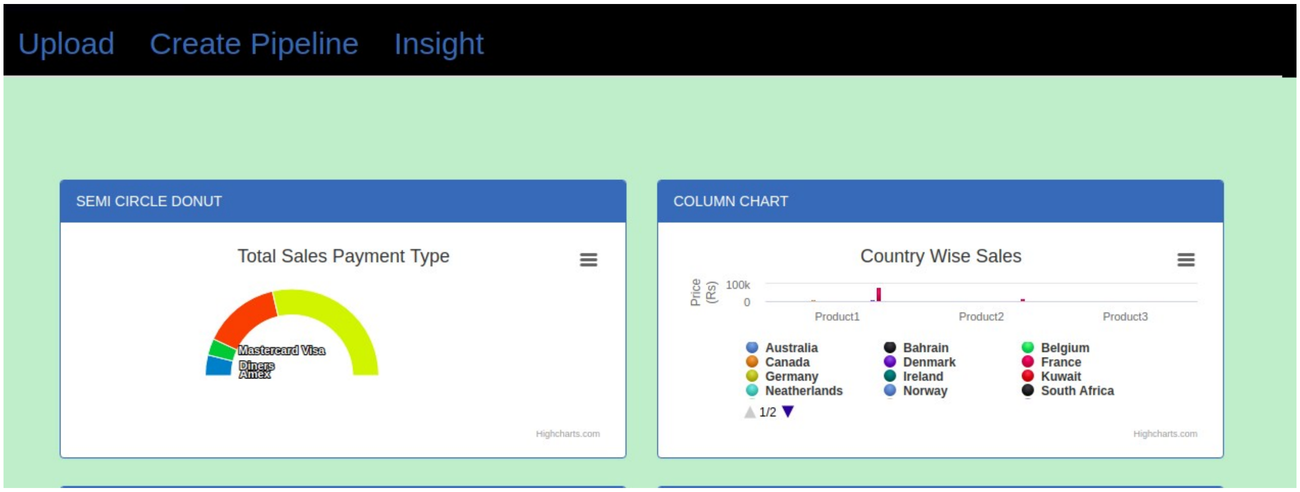
Since the application is based on implementation of big data,the processing is carried out using the processor for big data , that is spark .the file which is read line by line using the consumer of kafka in the previous step,is now stored in a predefined dataframe using spark .the data which is now in dataframes are logically equivalent to two dimentional arrays.but these tables are full of uncleaned data. these data are written to csv files .the csv files are equivalent to excel sheets.since these files are not in a proper manner to be analysed to gain insights,they are firstly examined for varoius inconsistancies .this examination is called data profiling. after profiling,the inconsistancies are found out .then the files are cleaned to remove the inconsistancies .the cleaned csv are stored in dataframes.

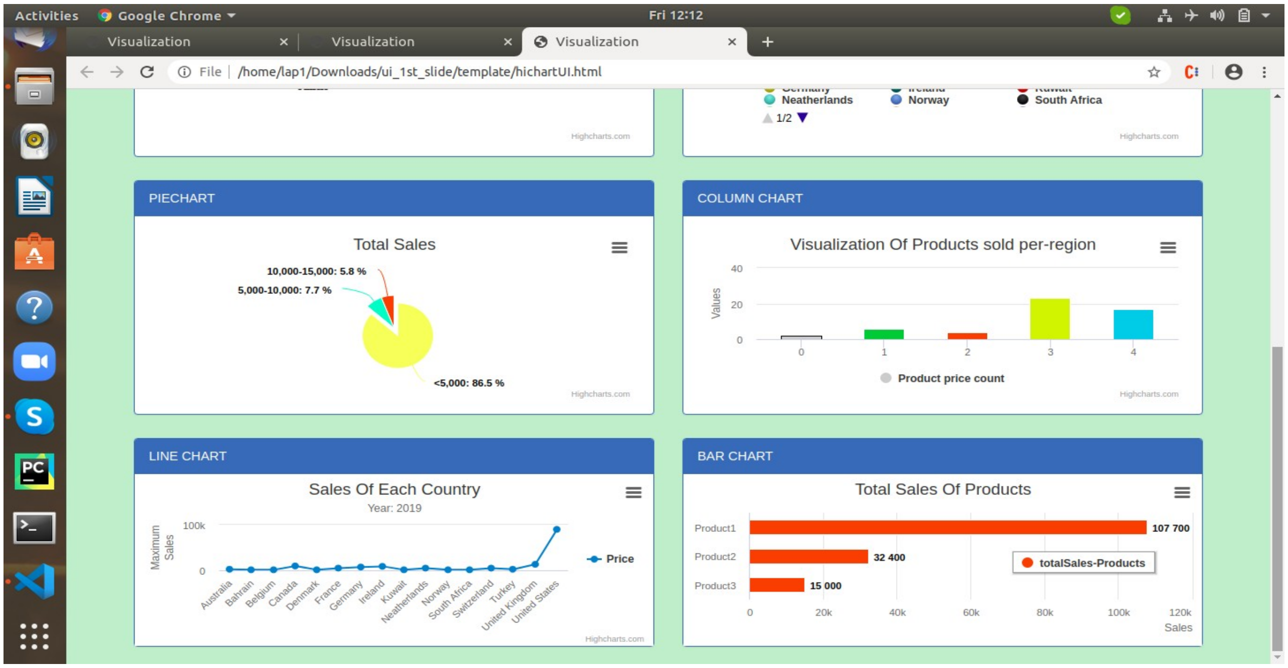


By Deafult Upload files tab will be selected

OBTAINING MEANINGFUL INSIGHTS

The cleaned csv files in the previous step are then stored in separate dataframes .in this palce the second event watcher that is watchdog2 comes into play.the watchdod2 calls the data generator API. This API stores the cleaned csv files to formatted dataframes. these dataframes are used to obtain meaningful insights and plot them using highcharts.

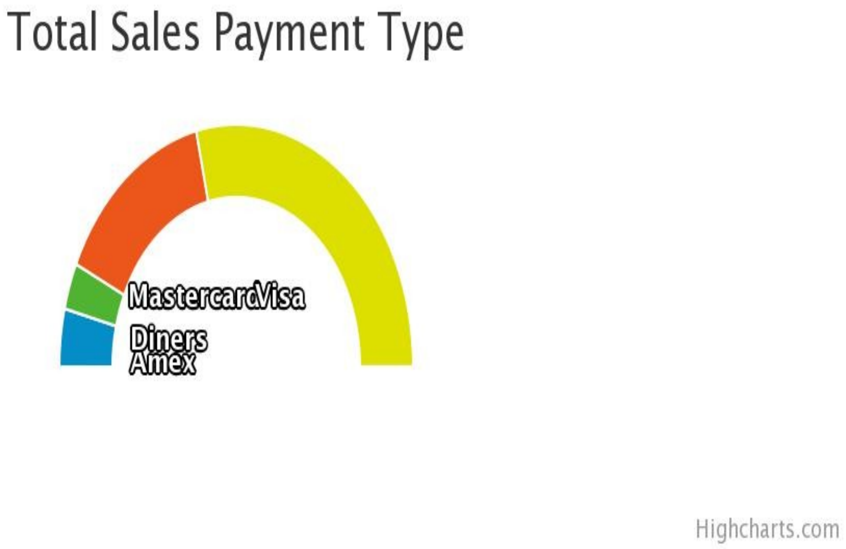




LIST OF CSV FILES

totsales_paymettype_country.csv

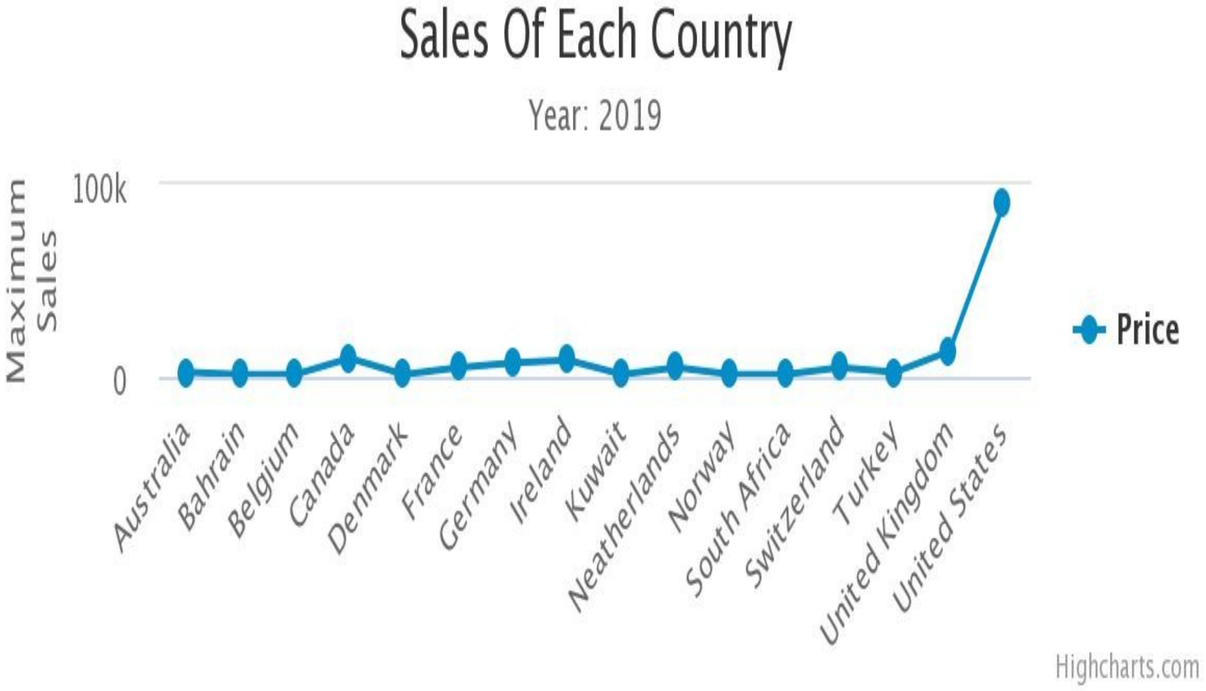
Paymenttype	Country	Price



Here paymenttype is the type of credit or debit card used.the country column has names of countries and price is the total amount on each type of card in each country.from this file the total sales of each type of credit cards are detemined.

totsales_countrytype.csv

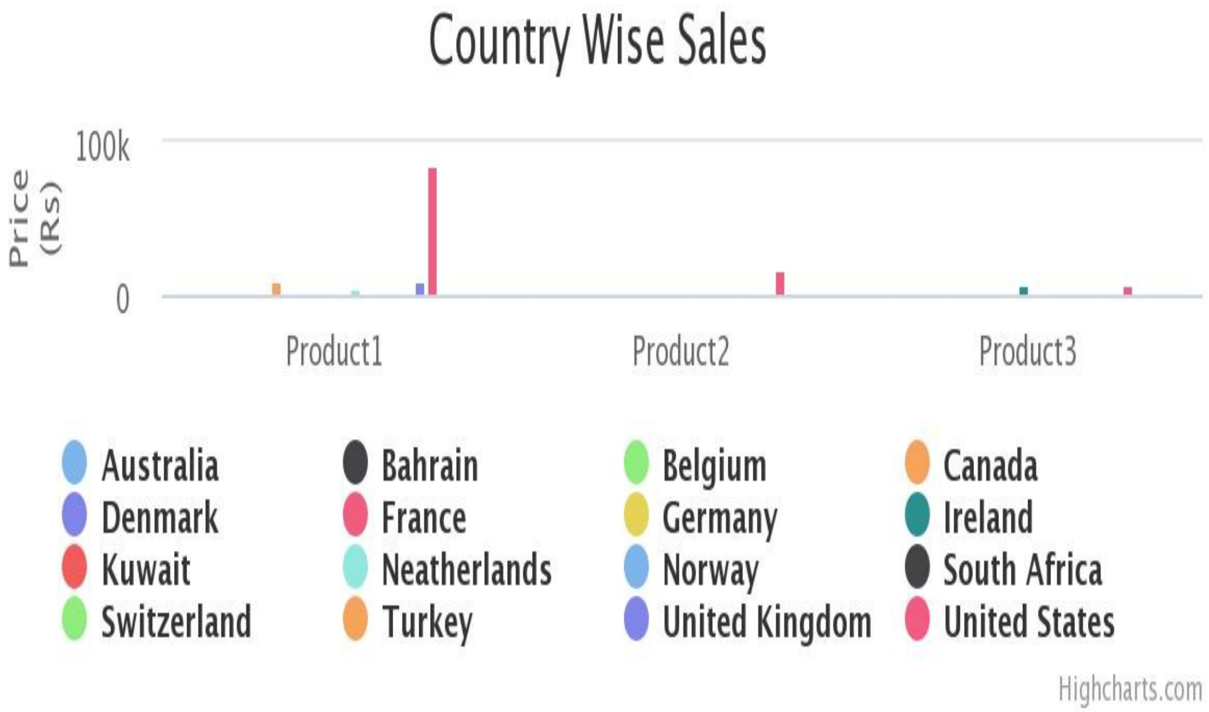
Country	Price



In this file names of different countries and the total sales price for the countries are given.the chart shows total sales in each countries.this is a line chart with countries plotted along x axis and price plotted along y axis.

totsales_producttype_country.csv

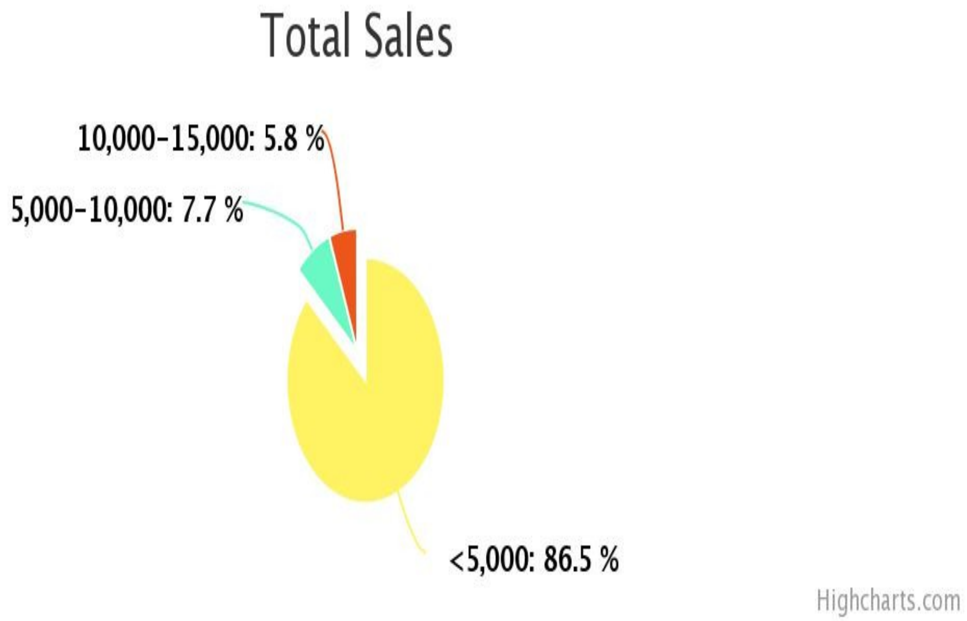
Product	Country	Price



There are three types of products which are grouped and plloted to show the country wise sales of different products.

totsales_country_state.csv

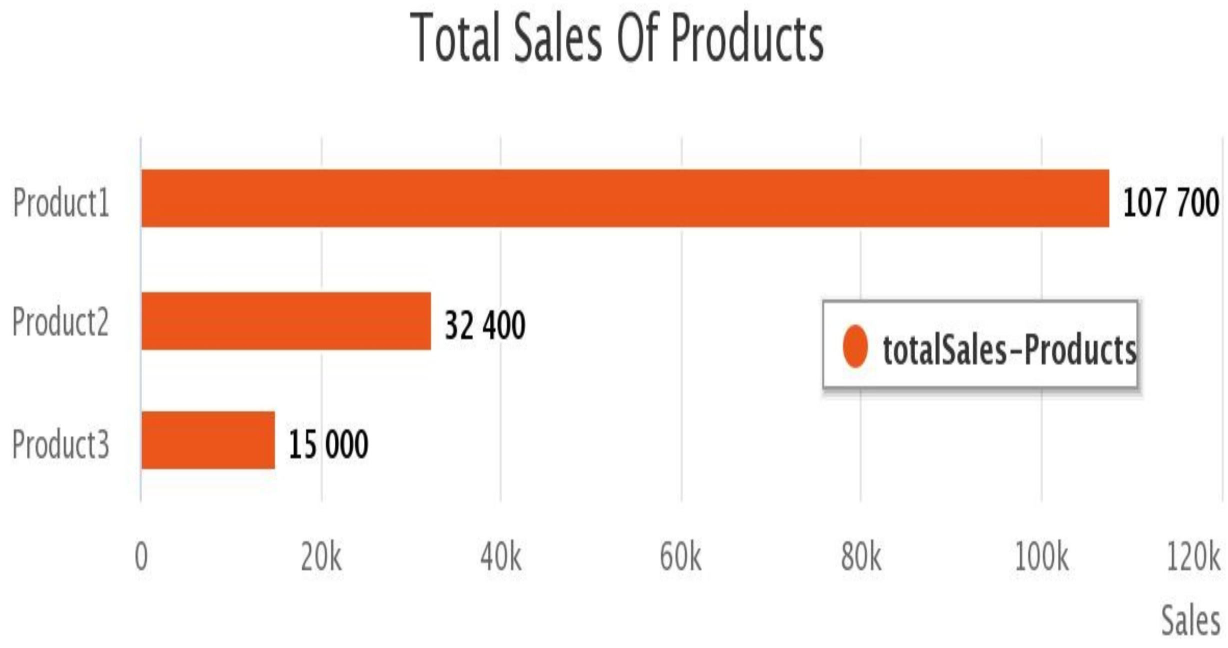
Country	State	Price



Here the total sales is catagorized price wise and is shown using pie charts. the country column has names of the countries. The state has names of the states of the countries and price is the sales in different states of the countries.

totsales_producttype.csv

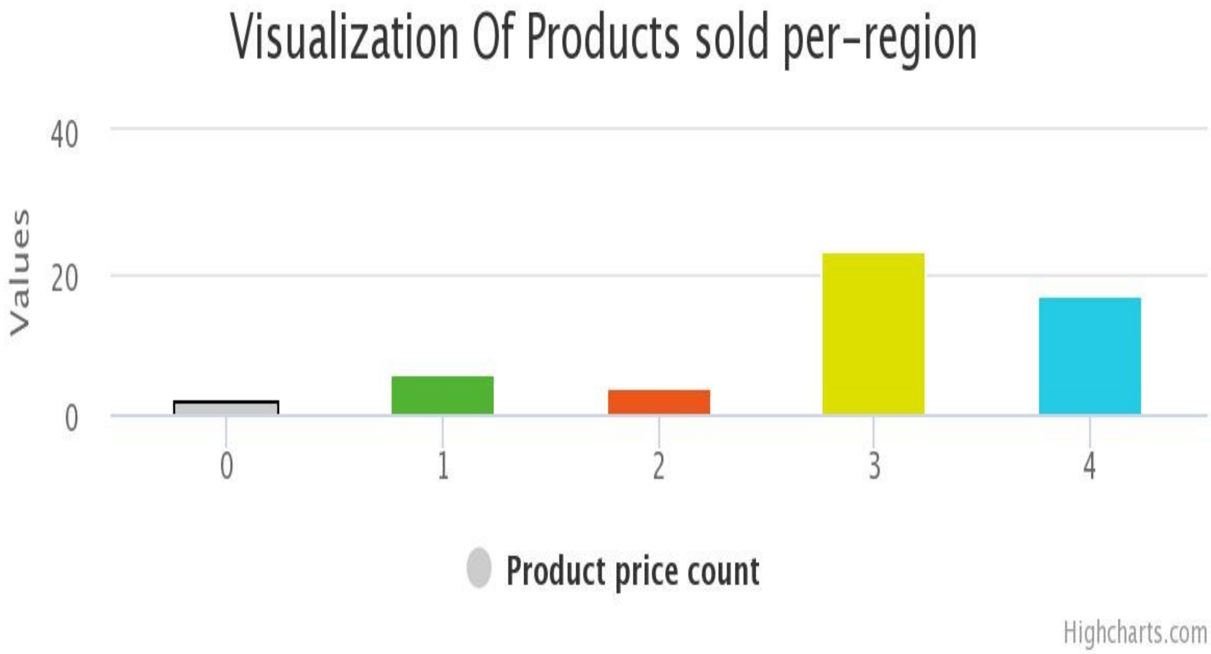
Product	Price



The price of different products are shown in the bar chart.along x axis the the price in thouands are plotted and along y axis the products are plotted.which infers that product1 has highest sales,then product2 holds its position.product3 has lowest sales record.

totsales_country_state.csv

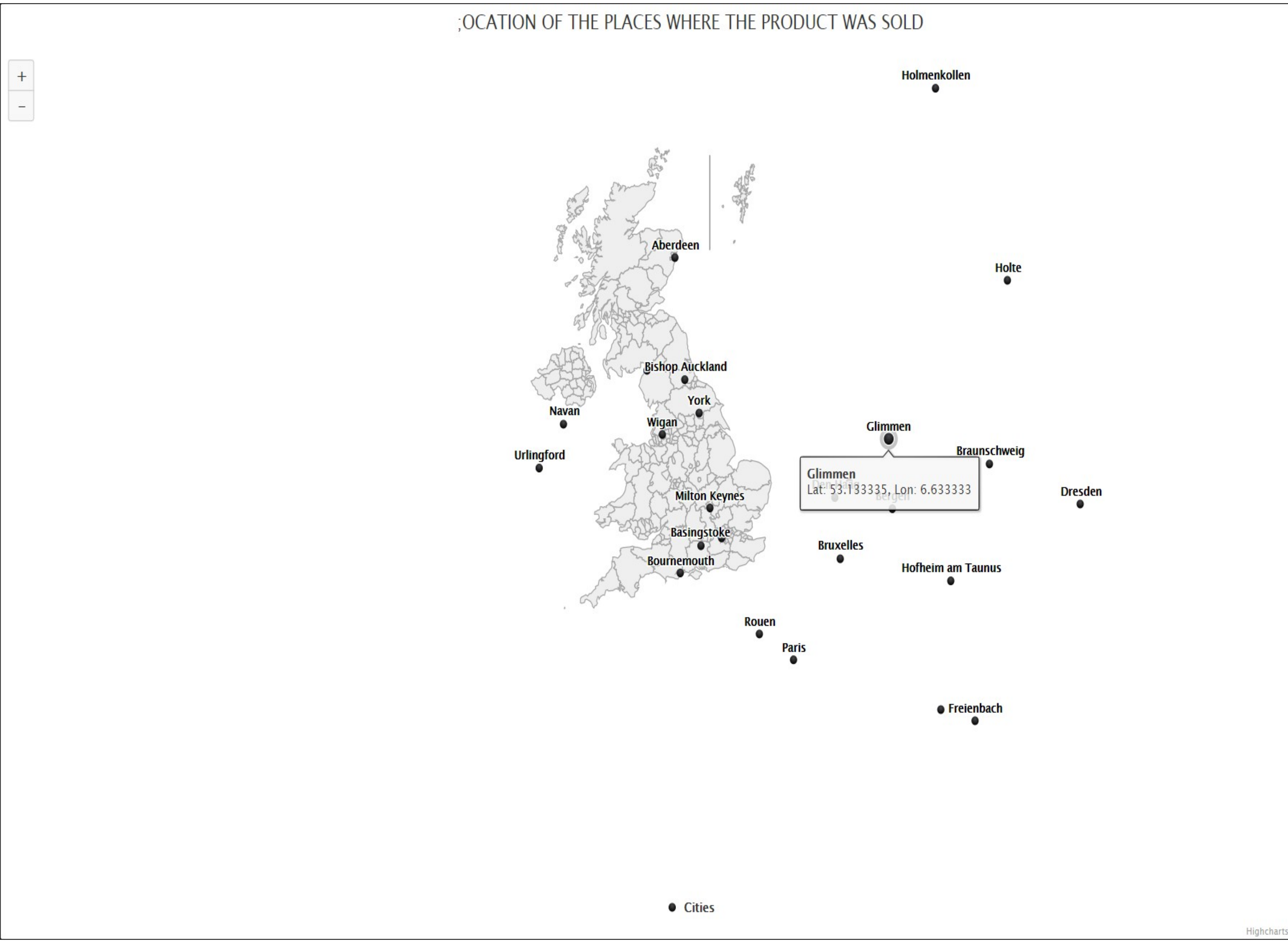
Country	State	Price



Here country wise prices are shown.along x axis there are five different countries.along y axis there are prices plotted.the countries are grouped to form five different groups.

Outlier.csv

Transaction Date	Product	Price	Payment_type	Nmae	City	State	Country	Acount_created	Last_login	Ltitude	Longitude	Q1	Q3	Outlier



CONCLUSION

The main aim of the project is to implement applications of big data ,spark and kafka.the kafka is mainly reaponsible for implementing live streaming of data.and spark is responsible for processing of the live data that is coming from kafka.here different files are present which is derived from the main file called outlier.csv.with this the live streaming takes place and processing is done to produce five other files.these five files are cleaned to obtain meaningful insights.these insights are plotted using javascript library called highcharts.