**Abstract**

The main idea of the project to build ETL pipelines, after that we are doing analysis of that data as like aggregation and join report.

In UberMedia project Data we got normally from these locations:

* Places visited/ Points of interest/ Common evening locations captured via mobile device signals. For Example- visit to an auto dealership, a home improvement retailer, jewelry shops, Malls etc.

This project we are done in two part: -

1)Raw Layer

2)Business Layer

Analysis on the first Row Layer and after that we are applying analysis on business layer.

We are making the jars using tools are sbt and maven.

After making jars using in the spark submit and give the path of the location which location of data are process and after processing which location data are sending.

we are using AWS S3 for storage purpose and making folder for which data are good records and data are bad records all files are storage in s3 bucket.

After this data are processed and sending the good records for future purpose.

After that we are submitted that data for future analysis.

**Introduction**

Writing Spark Jobs for Data Consumption from S3 buckets and transform it as per the requirement and then ingesting it into HDFS raw zone.

Built an Enterprise Data Lake by doing various aggregations as per the business requirement from HDFS raw zone data to provide a single abstracted view of business data to our client’s business units for advanced analytics.

The objective of UberMedia project is get insight of customers based on their geolocation data or through their visit using their device id(Hashed format). Data which we got contains latitude, longitude, Mobile device id in hashed format, polygon id and UNIX time.

So we get customer analysis like:

Total number of visit customers

Max/max visit time of customers.

In this project we are using the uber data that means any person which are booked an uber drive the details of that person are in the uber data cluster.

In this detail the polygon\_id, latitude and longitude are using and analyze for this location what actually this location and which category this is considering.

**Project Description (features)**

Data we got normally from these locations:

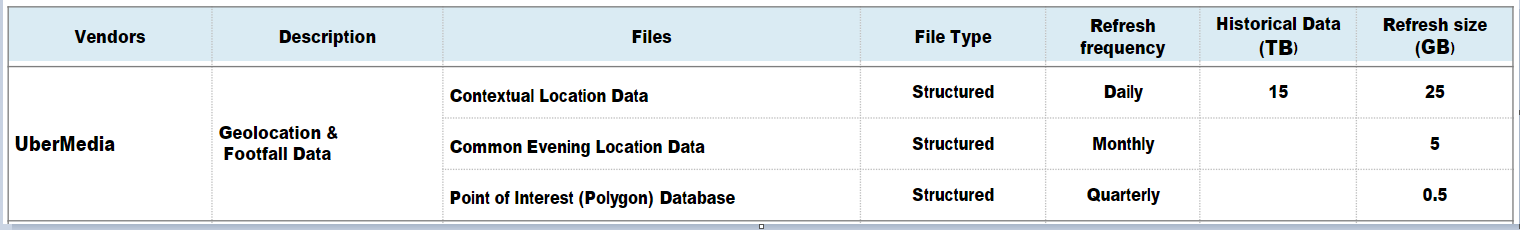
* Places visited/ Points of interest/ Common evening locations captured via mobile device signals. For Example- visit to an auto dealership, a home improvement retailer, jewelry shops, Malls etc.

Figure:-1(UberMedia Geolocation Data)

In ubermedia we got 3 types of data files

Contextual location data:

* CLD data is in structured format and occur daily.
* Which contains: device\_id, unixtime, latitude, longitude, polygon\_id.

Common Evening Location data:

* + Its same as CLD but its data of common location and occur monthly
  + Which contains: device\_id, country, last seen, latitude, longitude,
  + state, zip code, census.

Point of Interest data:

* It contains polygon data which occur quarterly Which contains: Polygon ID, Polygon Name, Full Address, Address1, Address2, City, State, Zip, Zip4, Polygon Category.
* These data we got in AWS S3 bucket and we read this data using spark then apply cleaning, transformation and then save PII information to S3 and Non PII information to HDFS/Hive.

For this process we require two layers: Raw layer and Business Layer

1) Raw Layer steps:

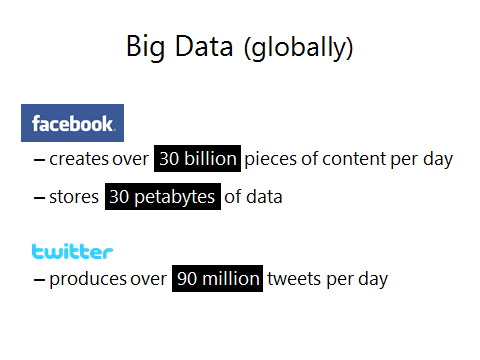
* Integration between S3 and spark.
* Reading data from S3 using spark sql
* Pre validation on data
* Header validation
* Custom filtration
* Data transformation and segregating good and bad records
* Write good records and Non PII to Hive/HDFS
* PII Records to S3
* Bad records in S3
* Post validation
* Audit
* Logs

2) Business Layer steps involved:

* Integration between hive and spark.
* Pre validation on data
* data from hive with latest file date
* Perform Aggregation, timestamp transformation, data analysis.
* Post validation
* Audit

**Big Data  
 What it is and why it matters: -**

* In today modern life are using big data is everywhere.
* Lots of data are being collected and warehoused from the system
* Web of data, electronic commerce
* Purchased by any authority or the grocery store
* Bank /credit card transaction system
* Social sites network system.



**IBM Definition**

Big data is created from everywhere every time in whole world. Every process and each social media process are creating it. Systems mobile devices are creating them

**What is Big Data?**

* + This data counted from everywhere: devices and sensors helpful in collective weather forecasting info, digital images and video formats, bought transaction data, and mobile phones tracking signals to name some. That is called the big data.

**Competitive advantage **

Data is being generated from the new resources in the modern advantage

**Decision making**  

Decision making is move from the lite data flow.

**Value of data **

As the value of data continues grow in the newest industry .

**Vs of big data: -**

**Volume**

In the real world huge amount of data are generated day by day so we very high volume of data will be generated.

**Velocity**: -

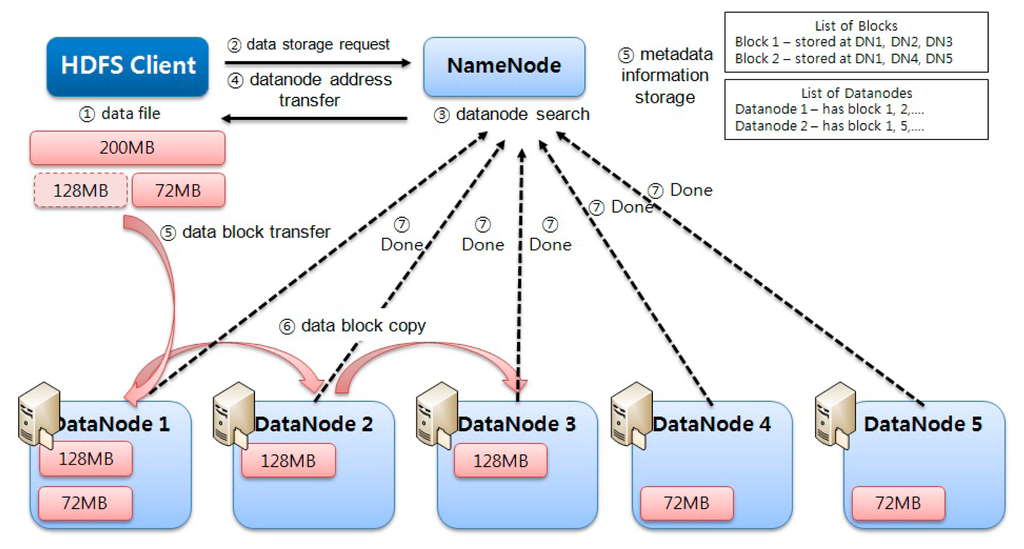
The varies type of data will be produced in the day by day so we have seen various velocity of data are coming. The velocity of data will be coming.

**What to do with these data?**

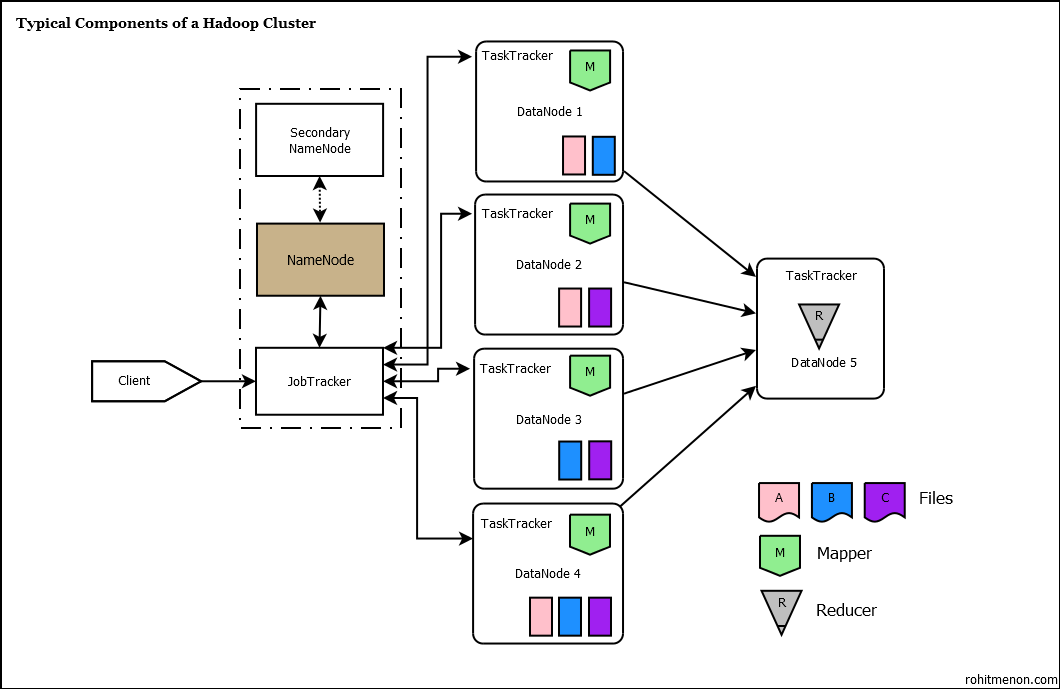
* Aggregation and Statistics
  + Data warehouse and OLAP
* Indexing, Searching, and Querying
  + Keyword based search
  + Pattern matching (XML/RDF)

(High level architecture of Hadoop)

**How Block placement Happens.**

****

**Components of Hadoop Cluster: -**

****

**Software Requirements: -**

Tools/Software Required: -

* **Big data Technologies:** Hadoop, Spark, AWS (additional)
* **Hadoop Components:** HDFS, Hive
* **Cloud Components:** S3(additional)
* **Languages:**  Unix/Linux, Scala/Python
* **Database:** Hive
* **Build Tools:** SBT, Maven
* **IDE:** Eclipse, IntelliJ
* **Operating Systems:** MS Windows, Unix/Linux
* **Other tool:**  MobaXterm, VMware, mapr

This project we are completed using some tools of big data.

This project we are completed in two ways -In the local server or the cloud server of aws(Amazon Web Server).We are using the local server for best performance of our project because in the cloud server some time are network error are occurring so that's why we are using the local VMware to execute of our project.

**System Design**

Spark Submit: -

spark-submit --class com.analytic.ubermedia.raw.ds\_um\_driver /home/mapr/jar/pipelines-0.0.1-SNAPSHOT.jar UM\_DAILY\_CONTEXTUAL\_LOCATION ubermedia/raw/cld/input/01JUN2019 /user/mapr/ubermedia/raw/cld/output\_non\_pii /user/mapr/ubermedia/raw/cld/error /user/mapr/ubermedia/raw/cld/archive /user/mapr/ubermedia/raw/cld/audit /user/mapr/ubermedia/raw/cld/output\_pii /user/mapr/ubermedia/raw/cld/logs

spark-submit --class com.analytic.ubermedia.raw.ds\_um\_driver /home/mapr/jar/pipelines-0.0.1-SNAPSHOT.jar UM\_POI /user/mapr/ubermedia/raw/poi/input/201906 uber.poi /user/mapr/ubermedia/raw/poi/error /user/mapr/ubermedia/raw/poi/archive /user/mapr/ubermedia/raw/poi/audit /user/mapr/ubermedia/raw/poi/pii\_table /user/mapr/ubermedia/raw/poi/logs

spark-submit --class com.analytic.um.bssn.ds\_um\_driver /home/mapr/jar/pipelines-0.0.1-SNAPSHOT.jar UM\_DAILY\_CONTEXTUAL\_LOCATION uber.poi uber.non\_pii uber.business /user/mapr/ubermedia/business/logs

**Table Create: -**

create external table non\_pii(

accenture\_device\_id string,

lat double,

lon double,

polygon\_id string,

unixtime string,

batch\_id string,

data\_collection\_date string,

input\_file\_name string,

event\_datetime string,

file\_date date

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t'

STORED AS TEXTFILE

LOCATION '/user/mapr/ubermedia/raw/cld/output\_non\_pii';

create table poi

(

PolygonID string,PolygonName string,FullAddress string,Address1 string,Address2 string,City string,State string,Zip string,Zip4 string,PolygonCategory array<string>,input\_file\_name string,run\_date string,batch\_id string

)

partitioned by (file\_date string)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t'

STORED AS TEXTFILE;

create table uber.business

(

accenture\_device\_id string,lat\_lon string, mintime string,maxtime string,lat double,lon double,tzoneid string,mintime\_local string,maxtime\_local string,Polygon\_id string,PolygonCategory string,input\_file\_name string,batch\_id string,load\_date string

)

partitioned by (file\_date string,data\_collection\_date string)

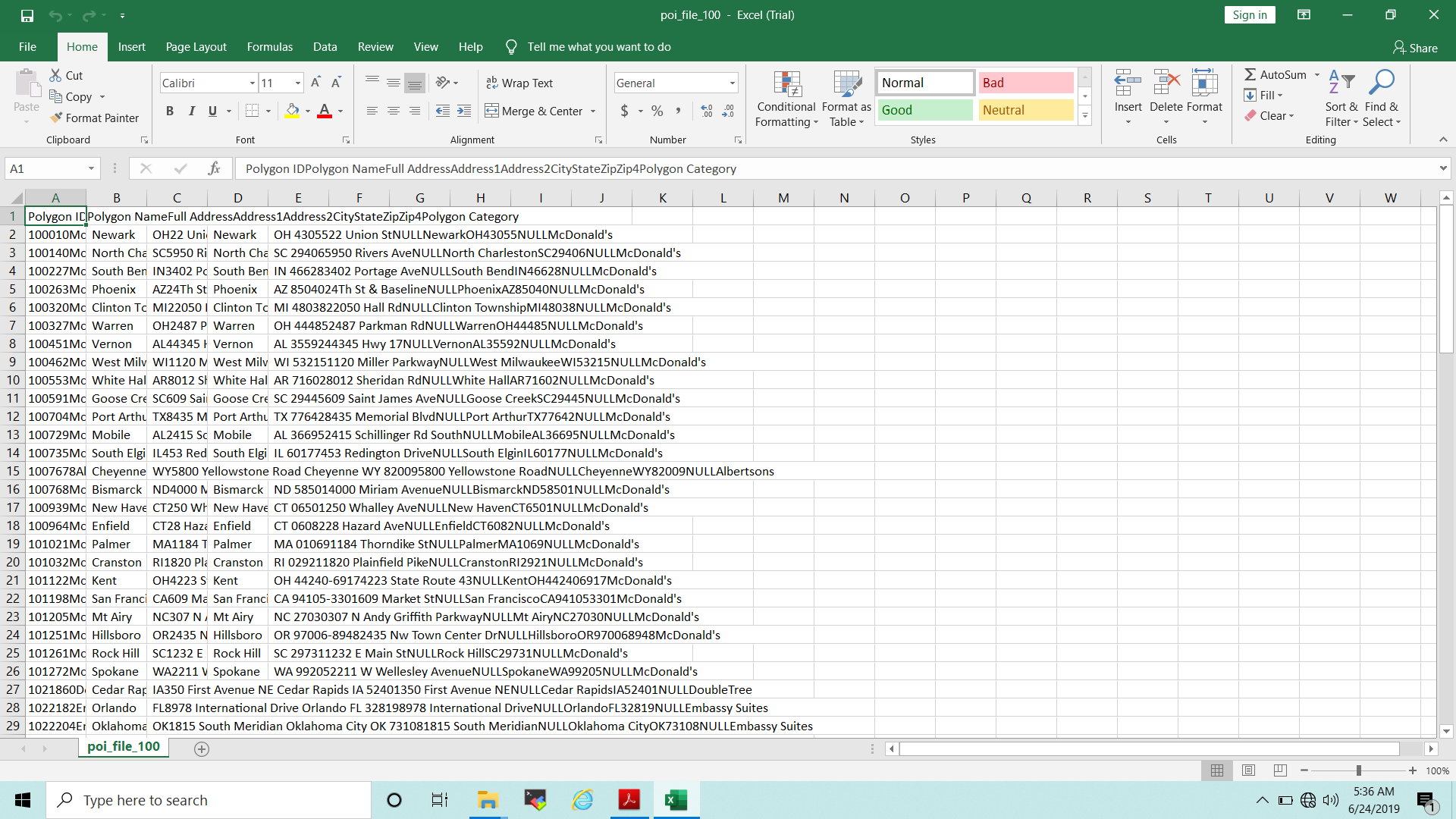
ROW FORMAT DELIMITED

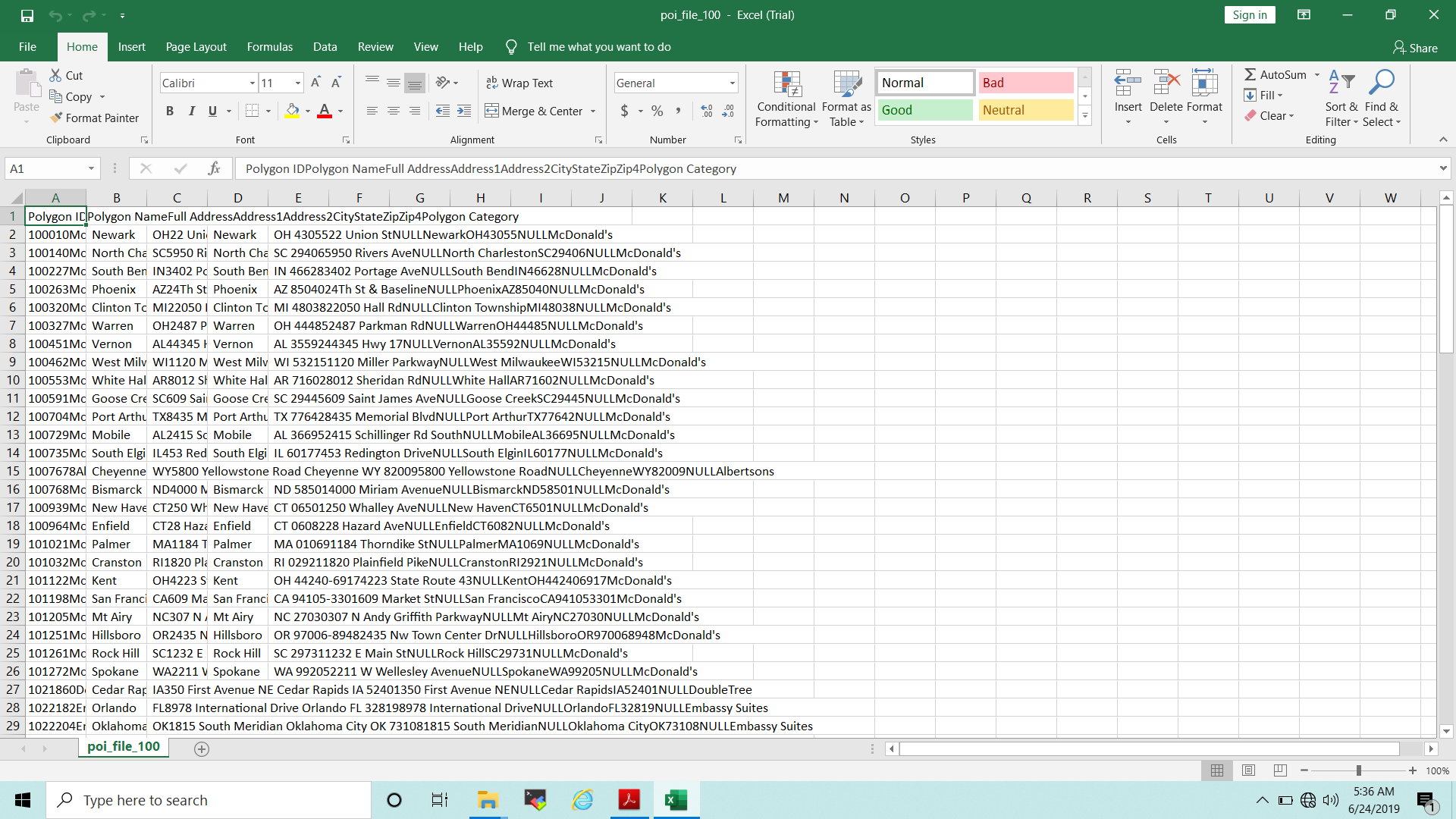
FIELDS TERMINATED BY '\t'

STORED AS TEXTFILE;

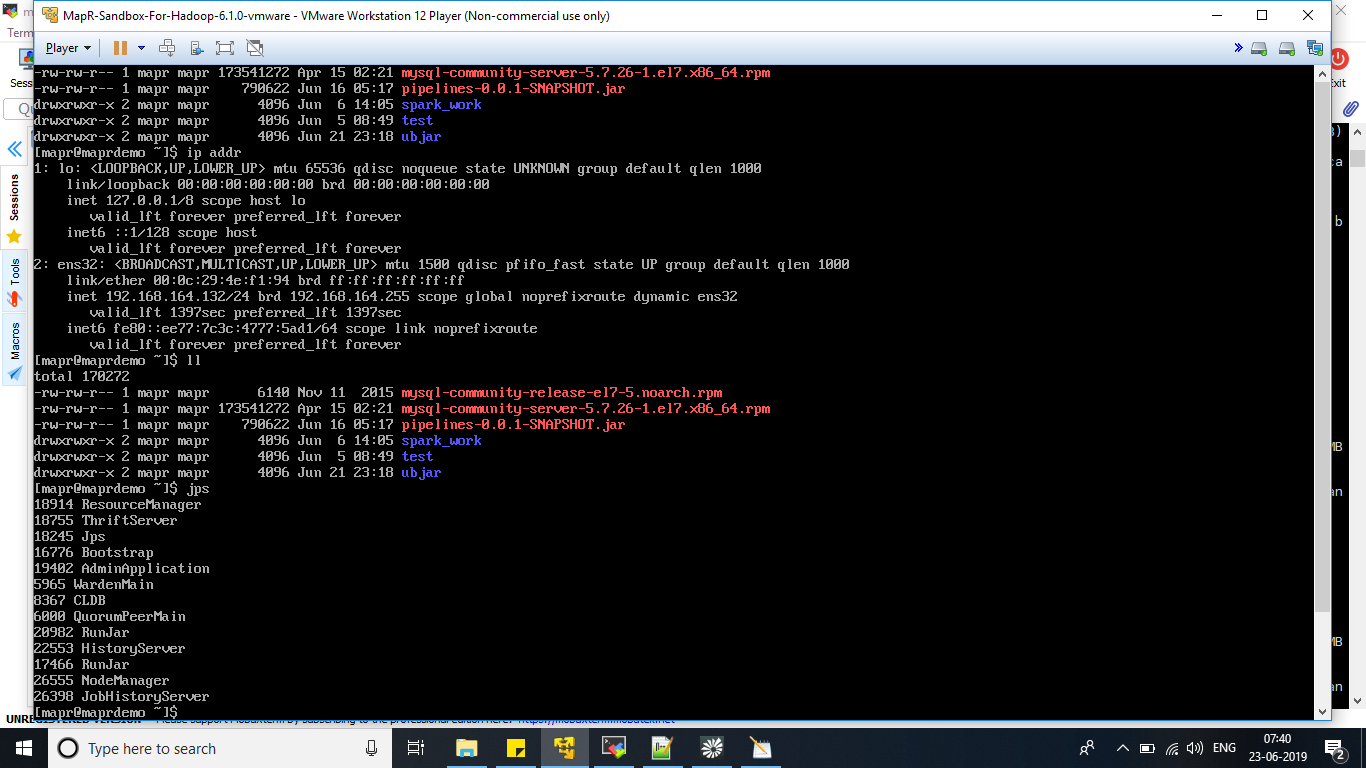
**Implementation Result**

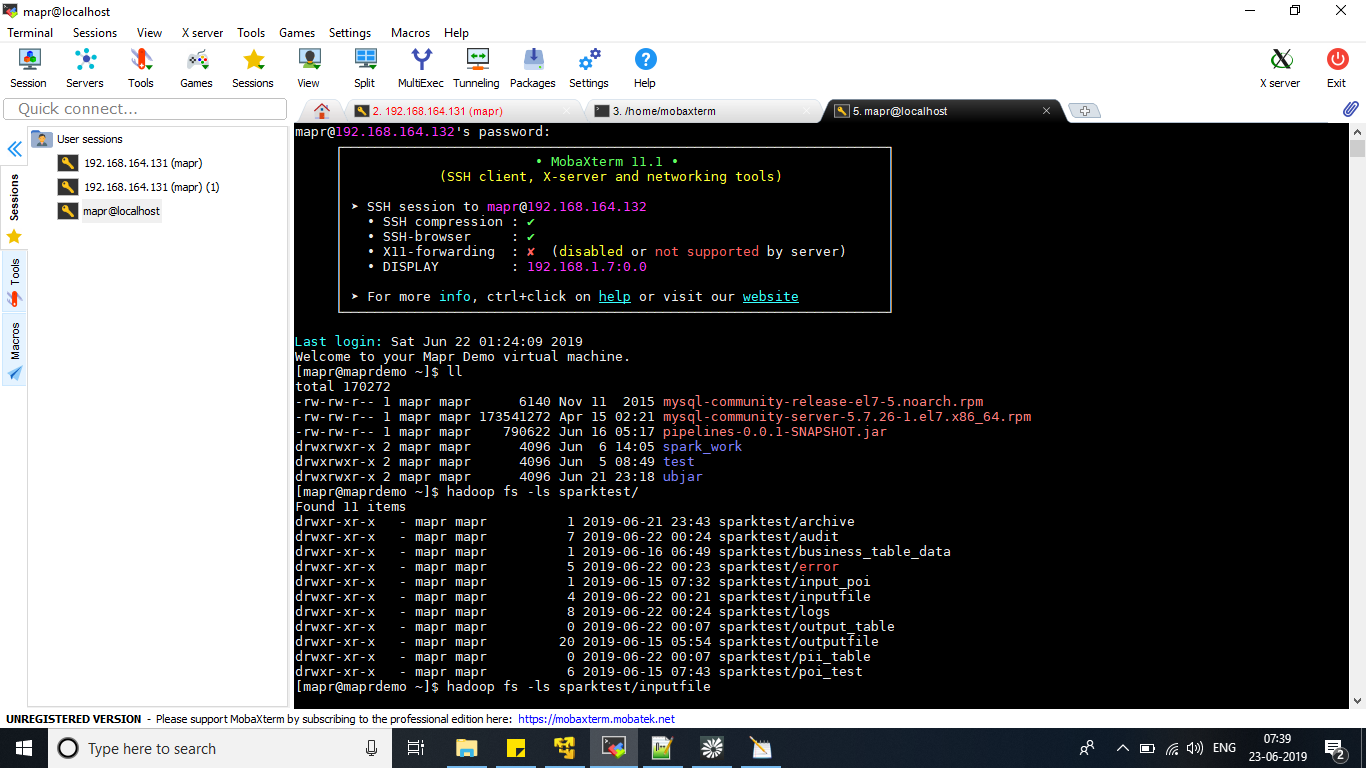
**Dataset: -**

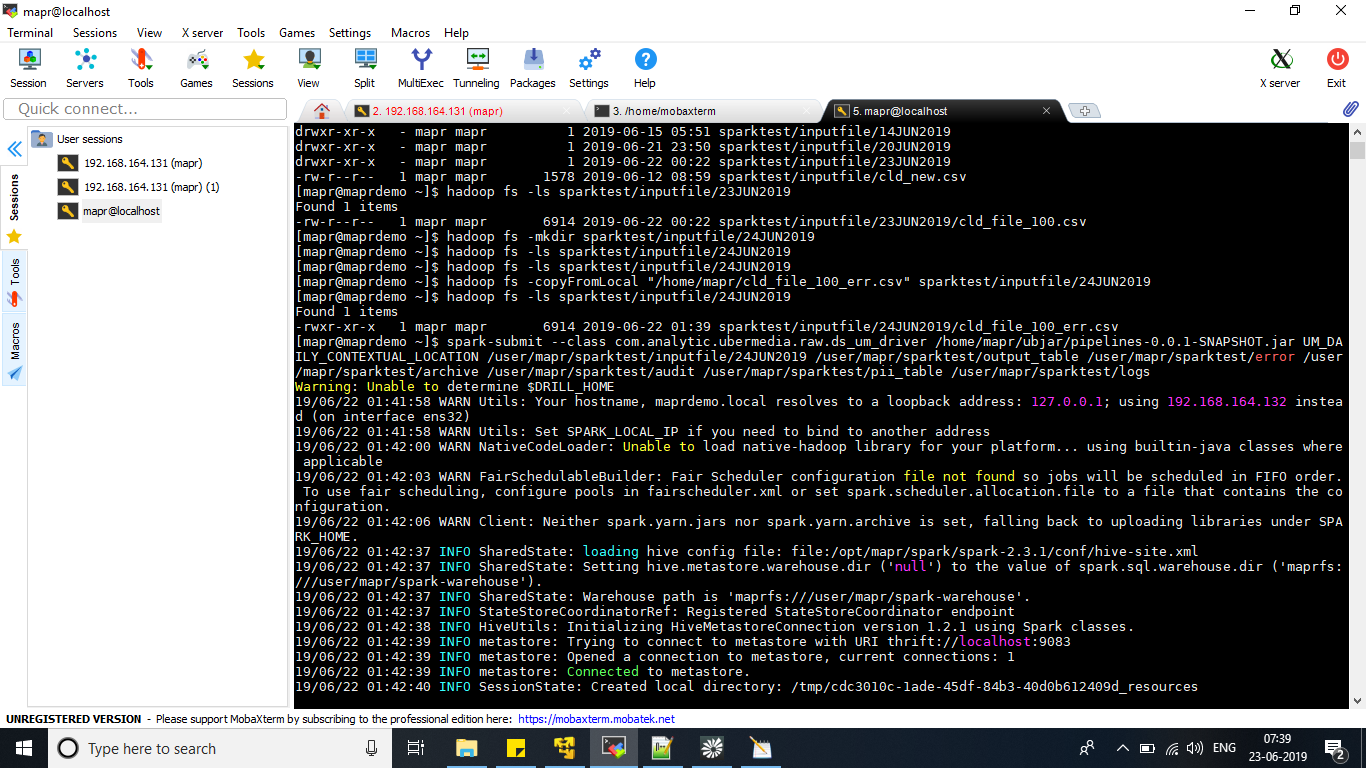


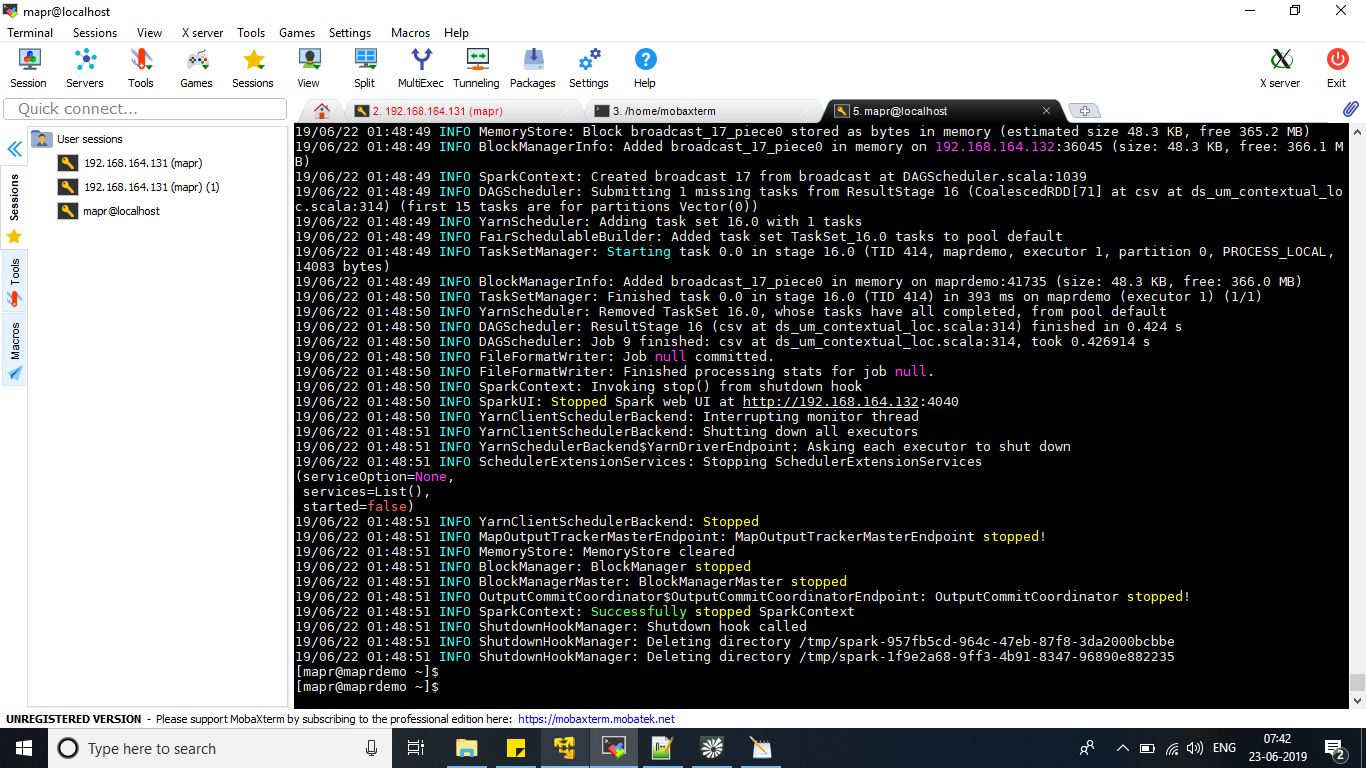


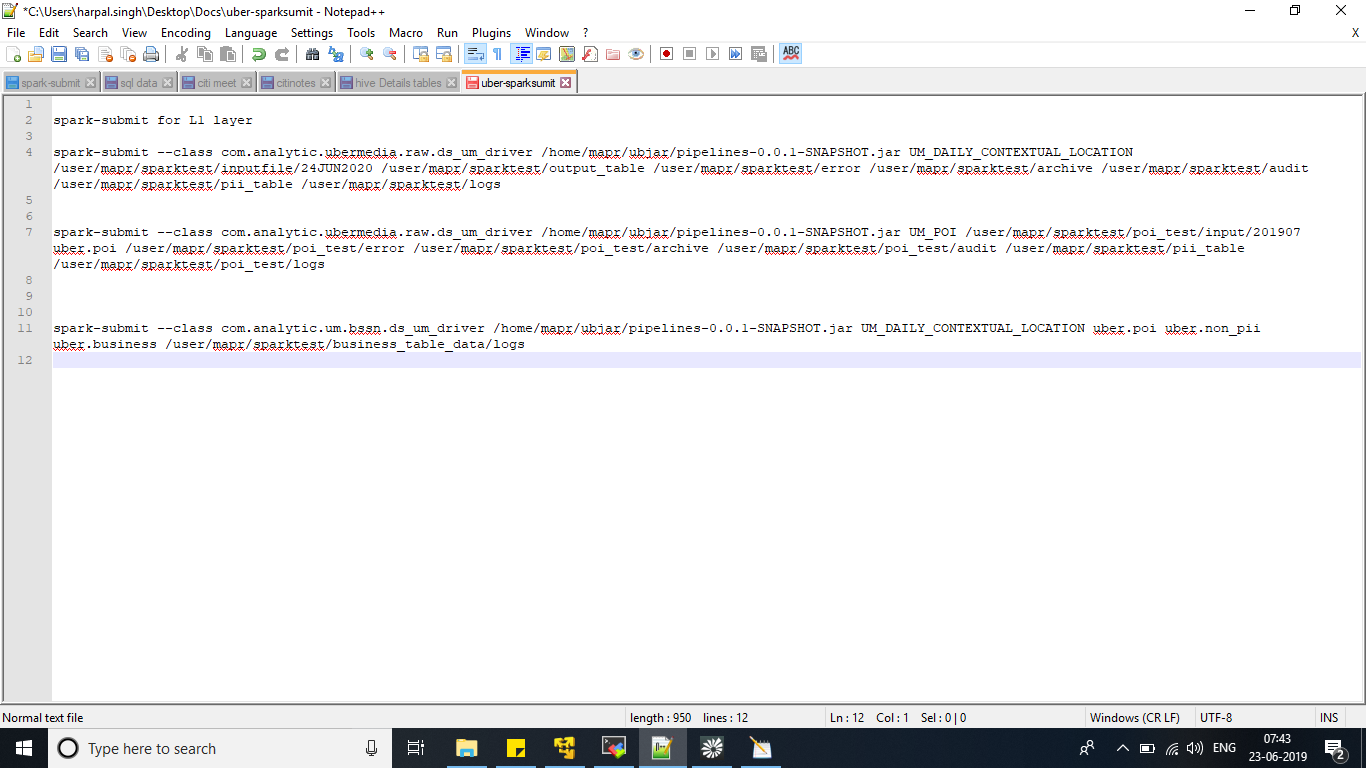
**Result**

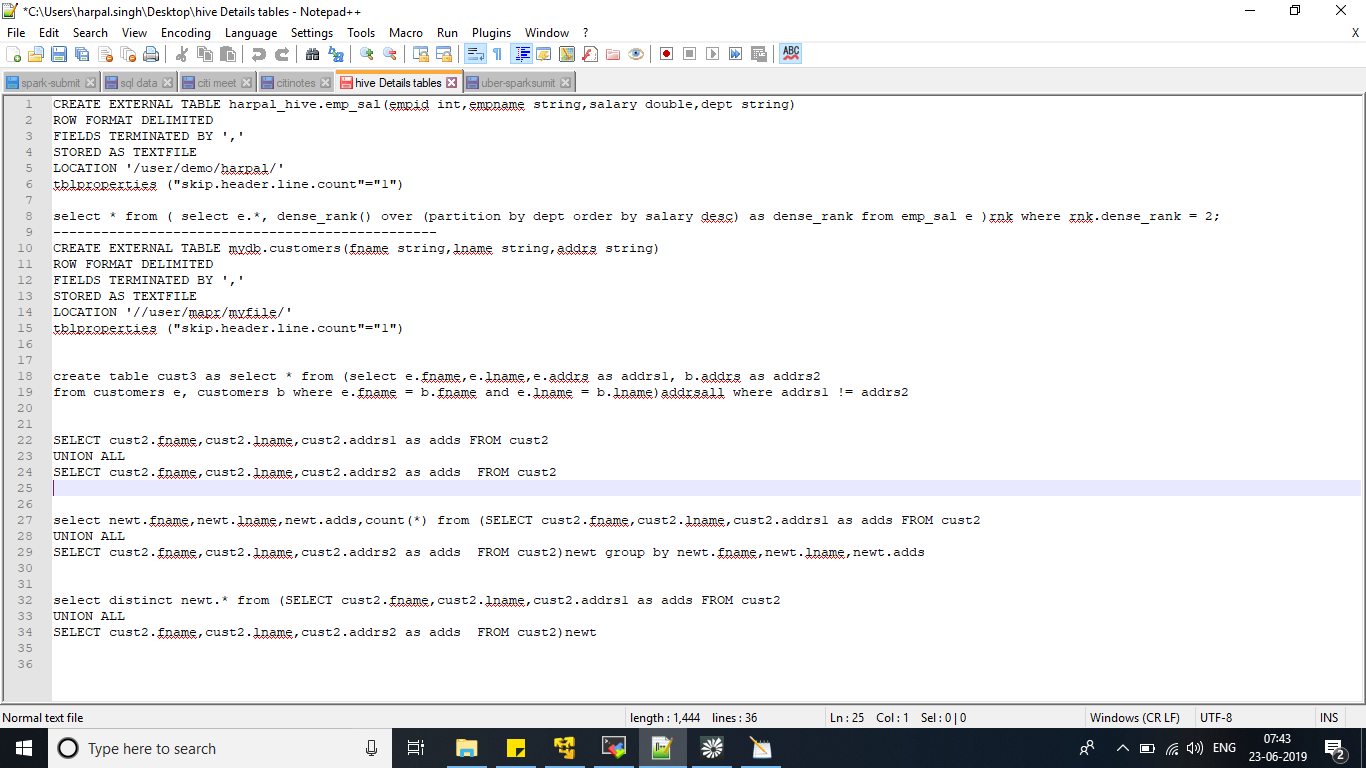


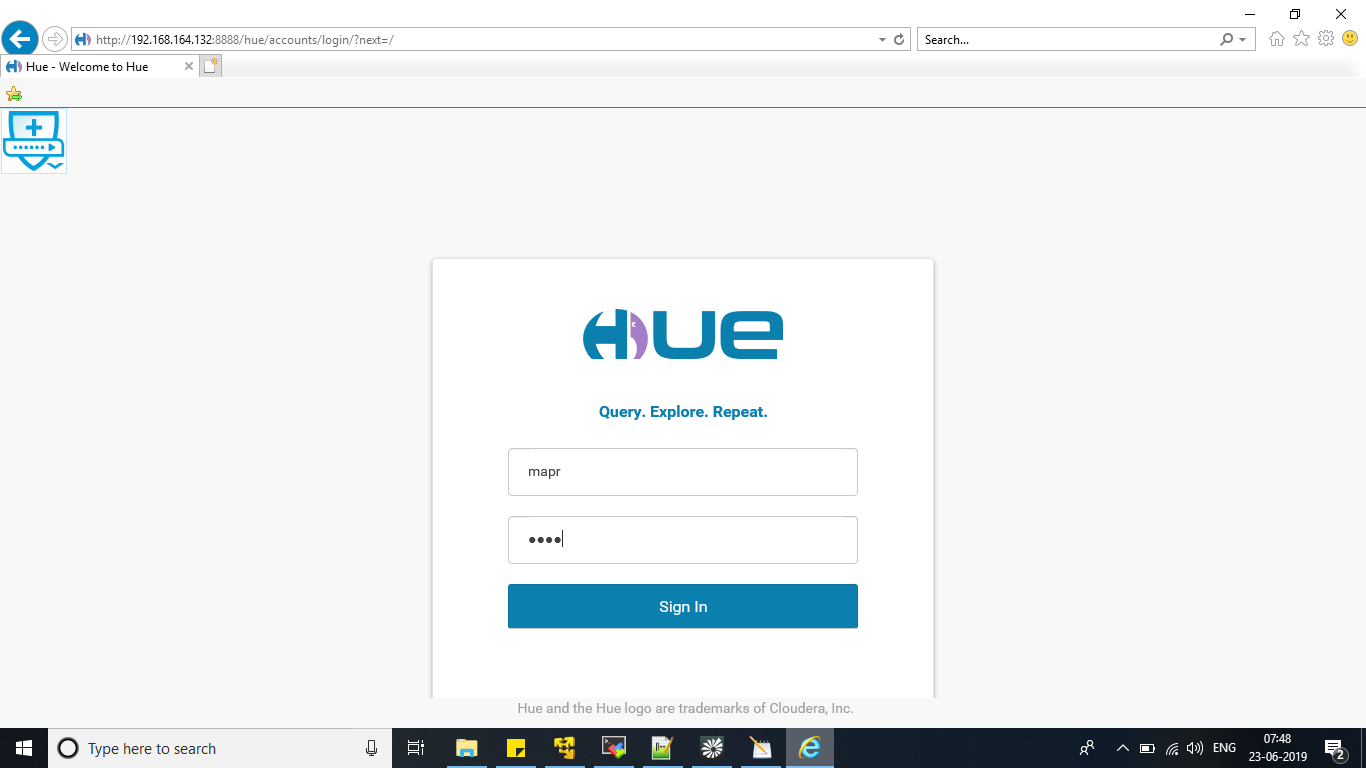


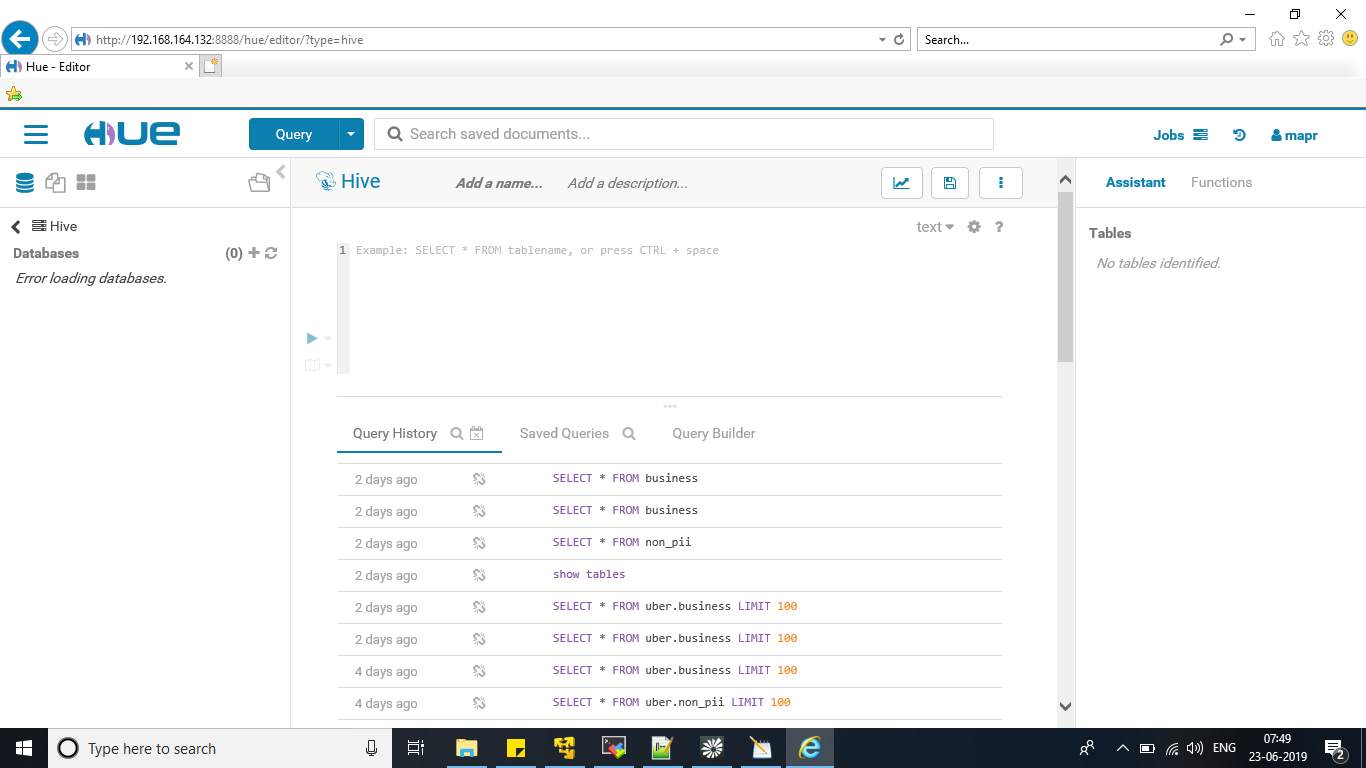


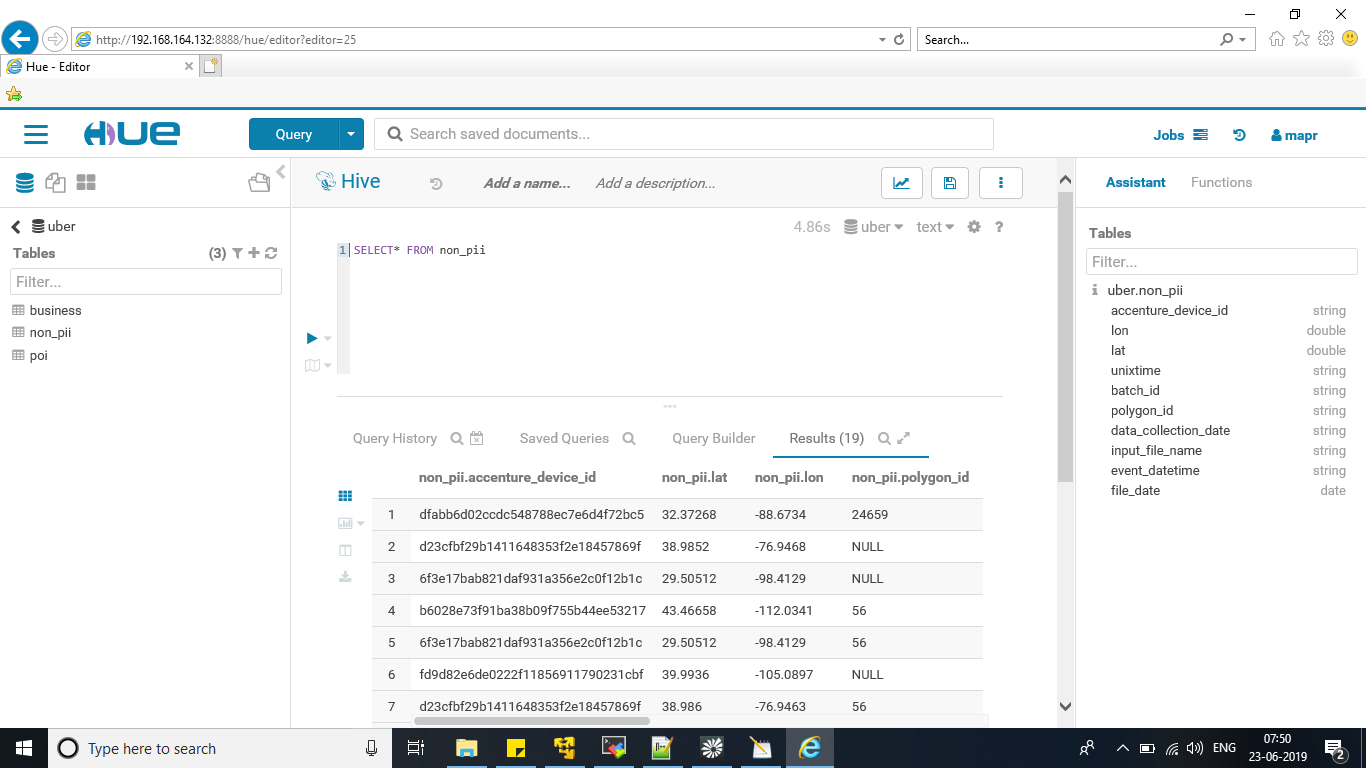


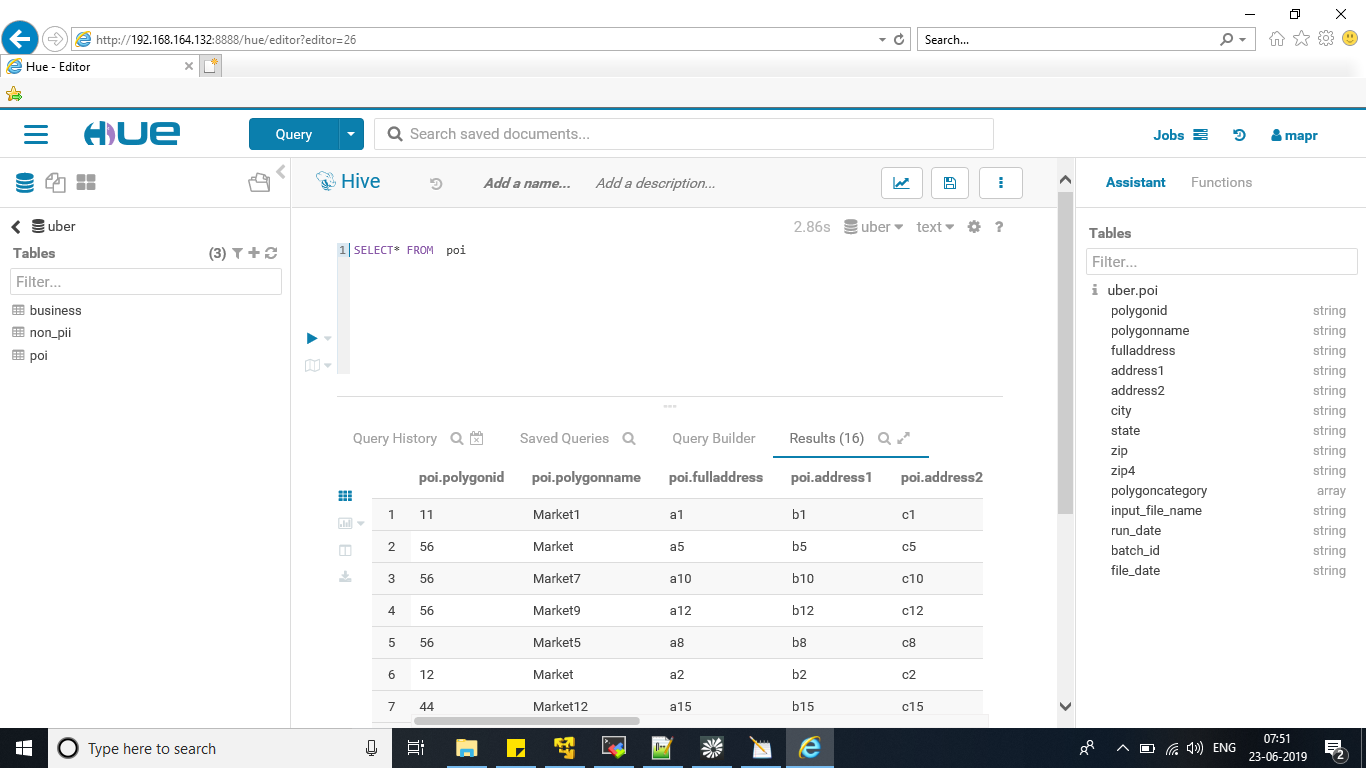


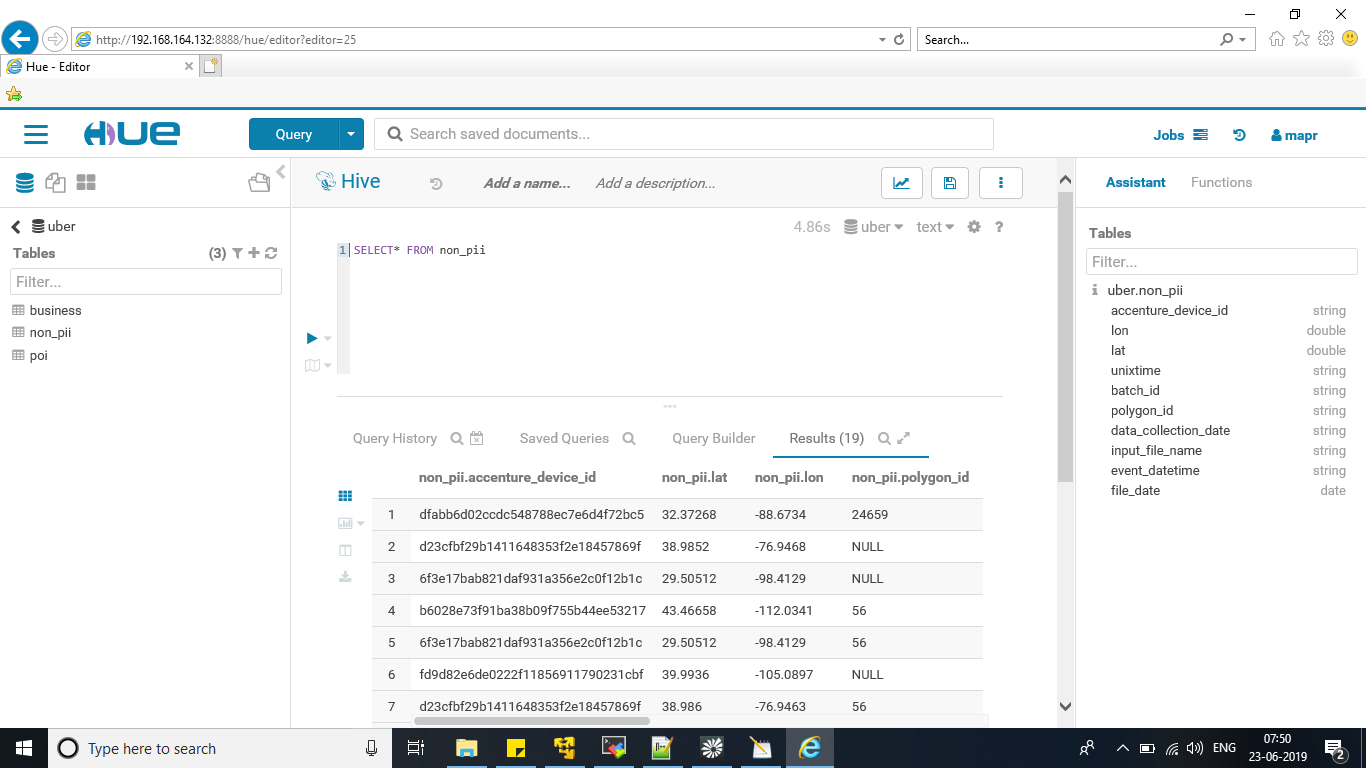


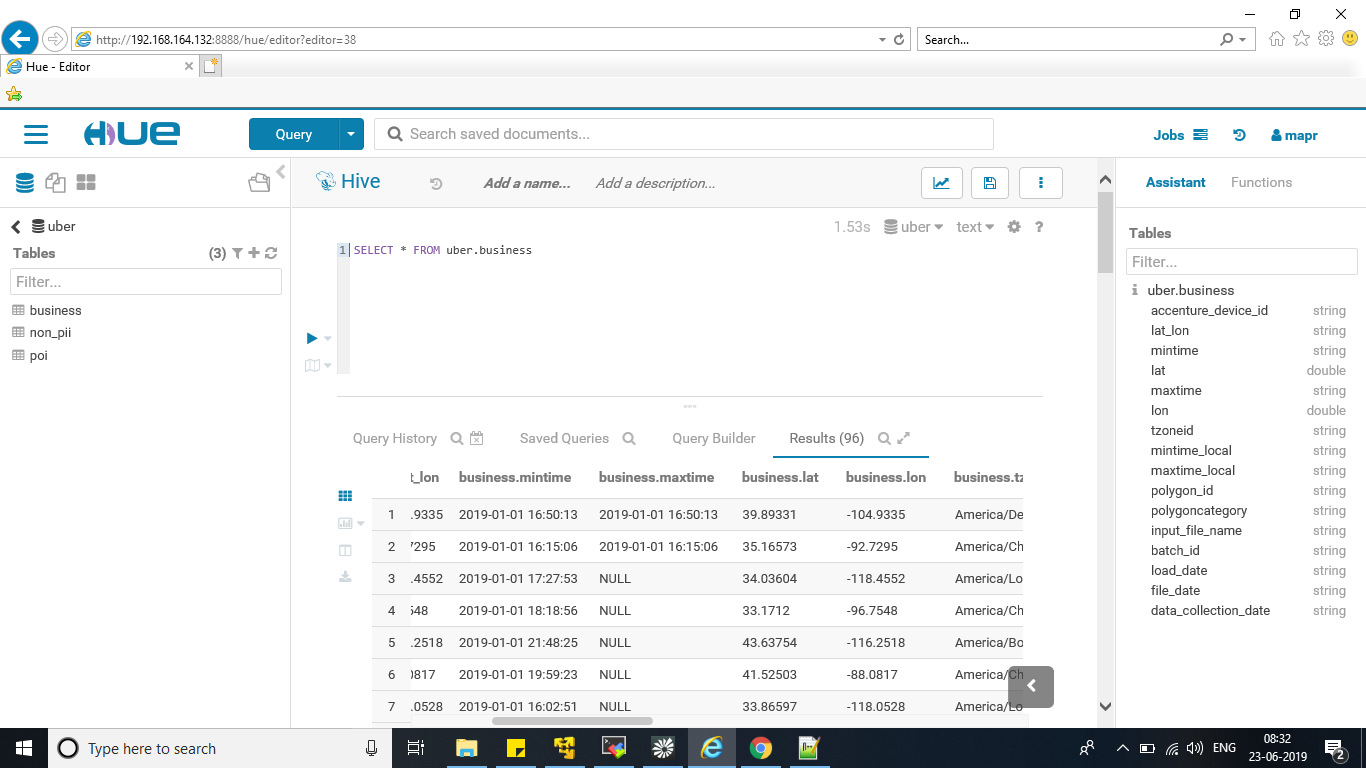


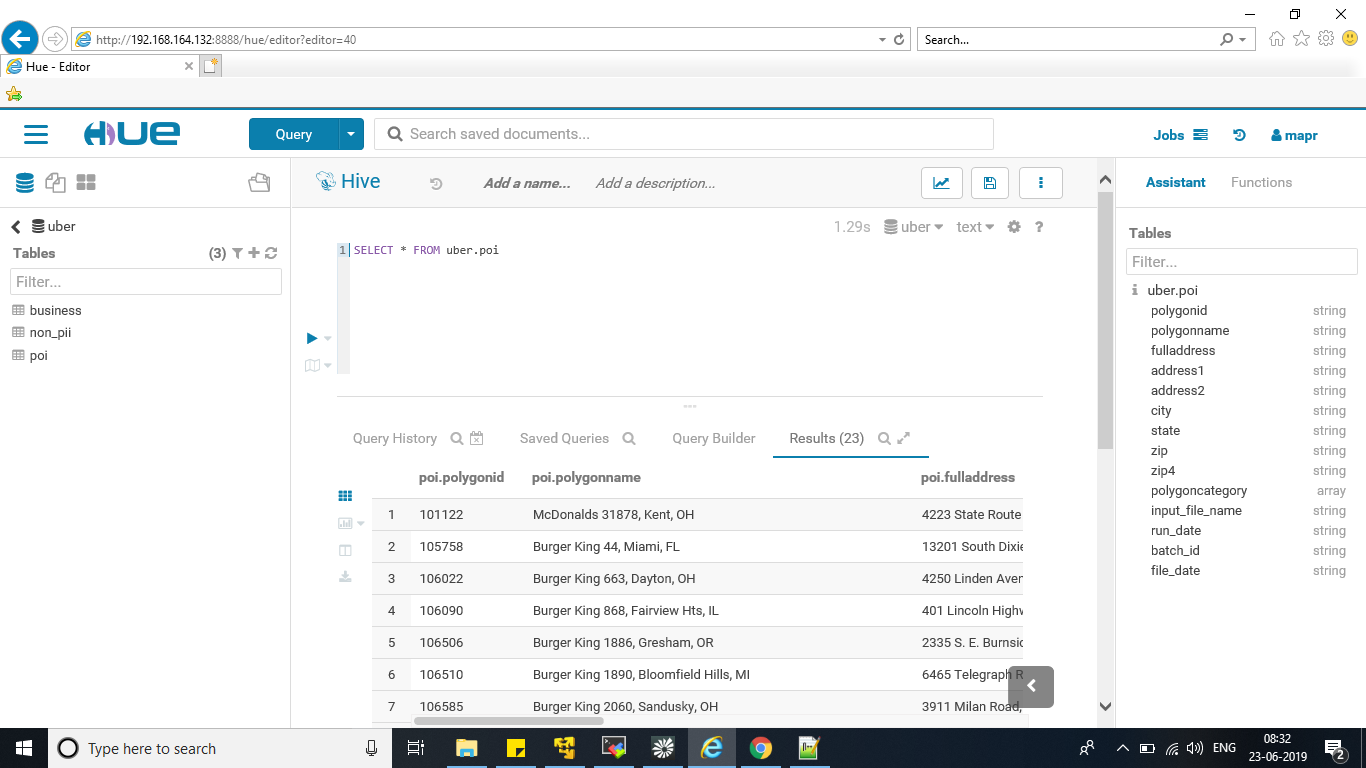


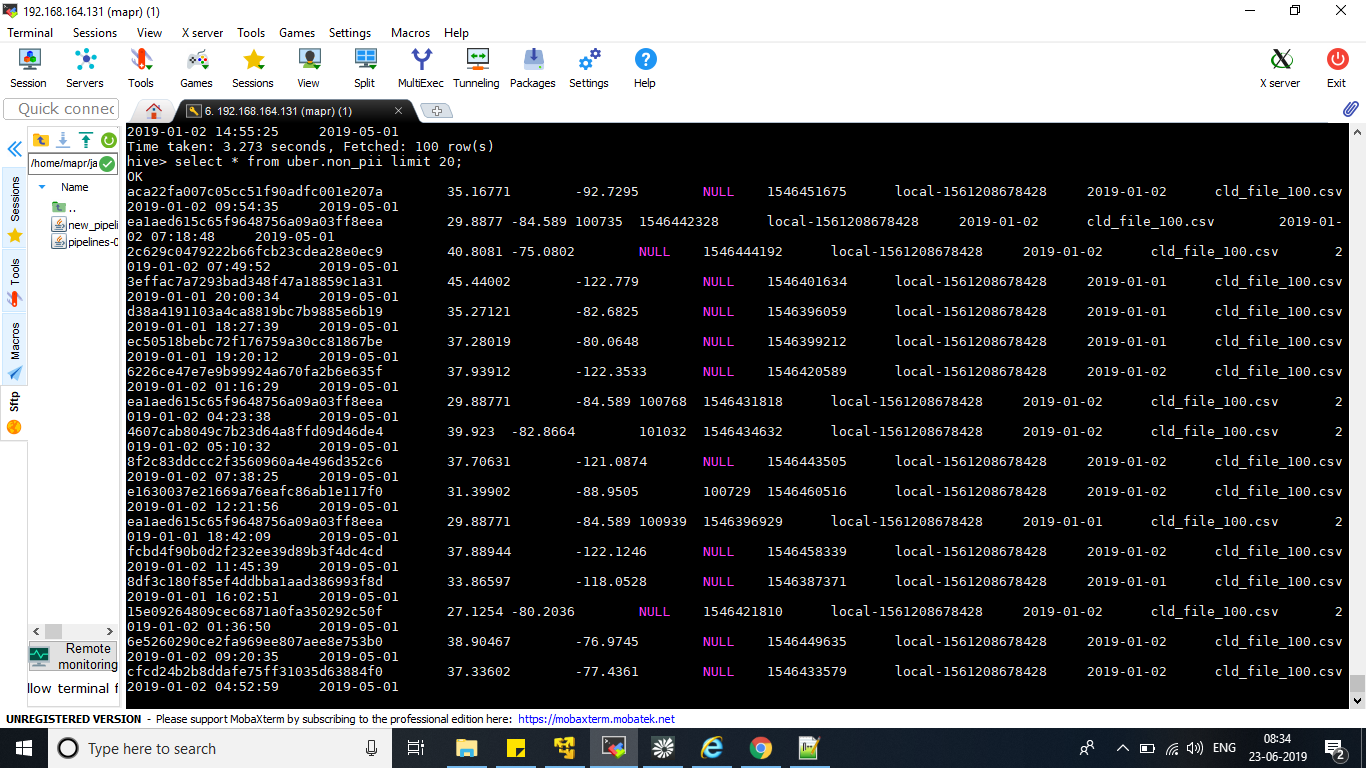


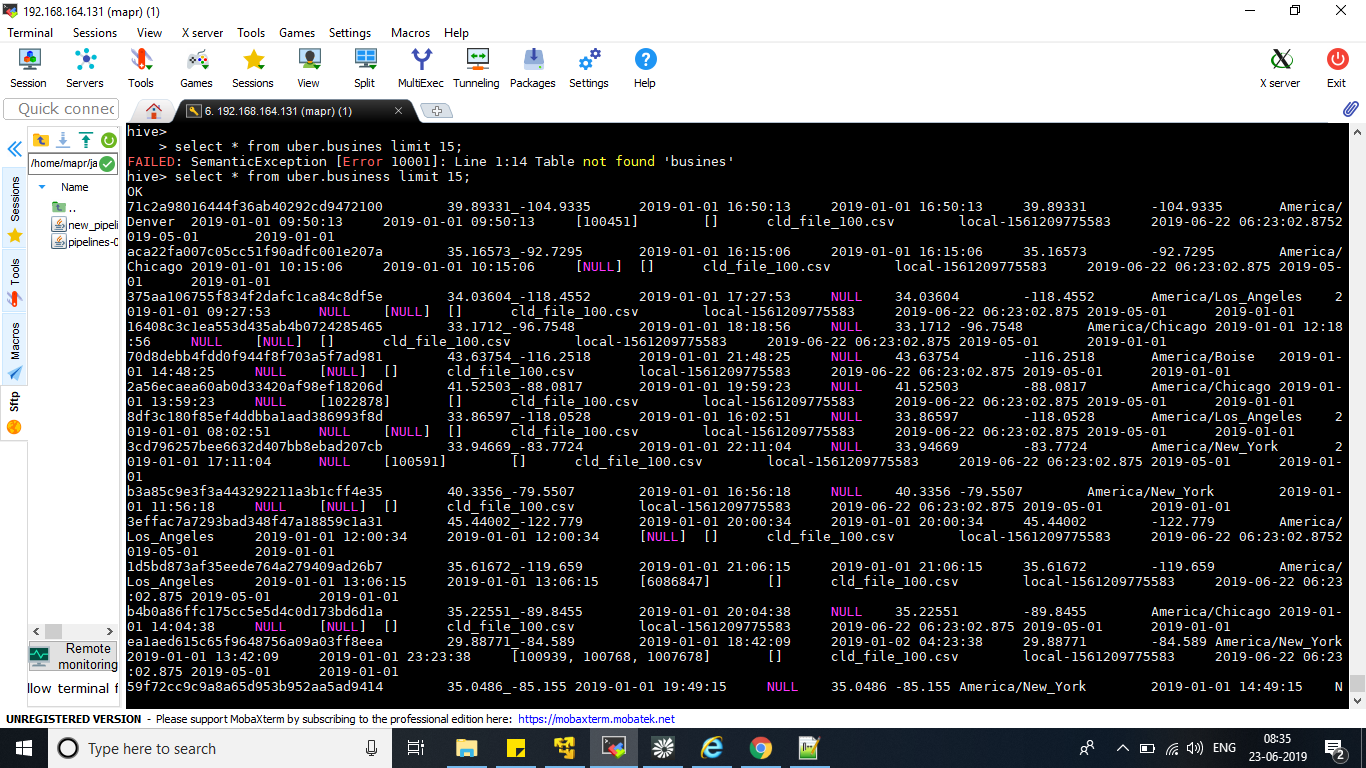


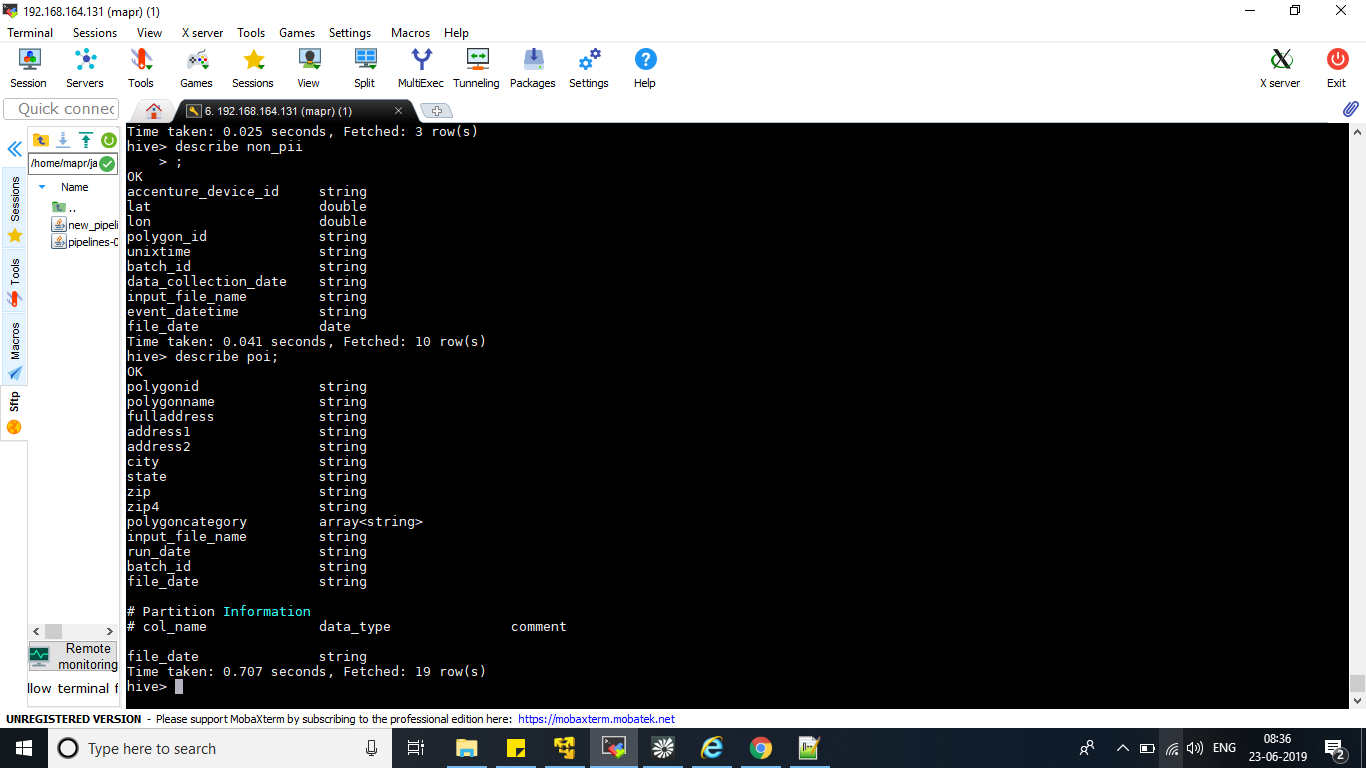


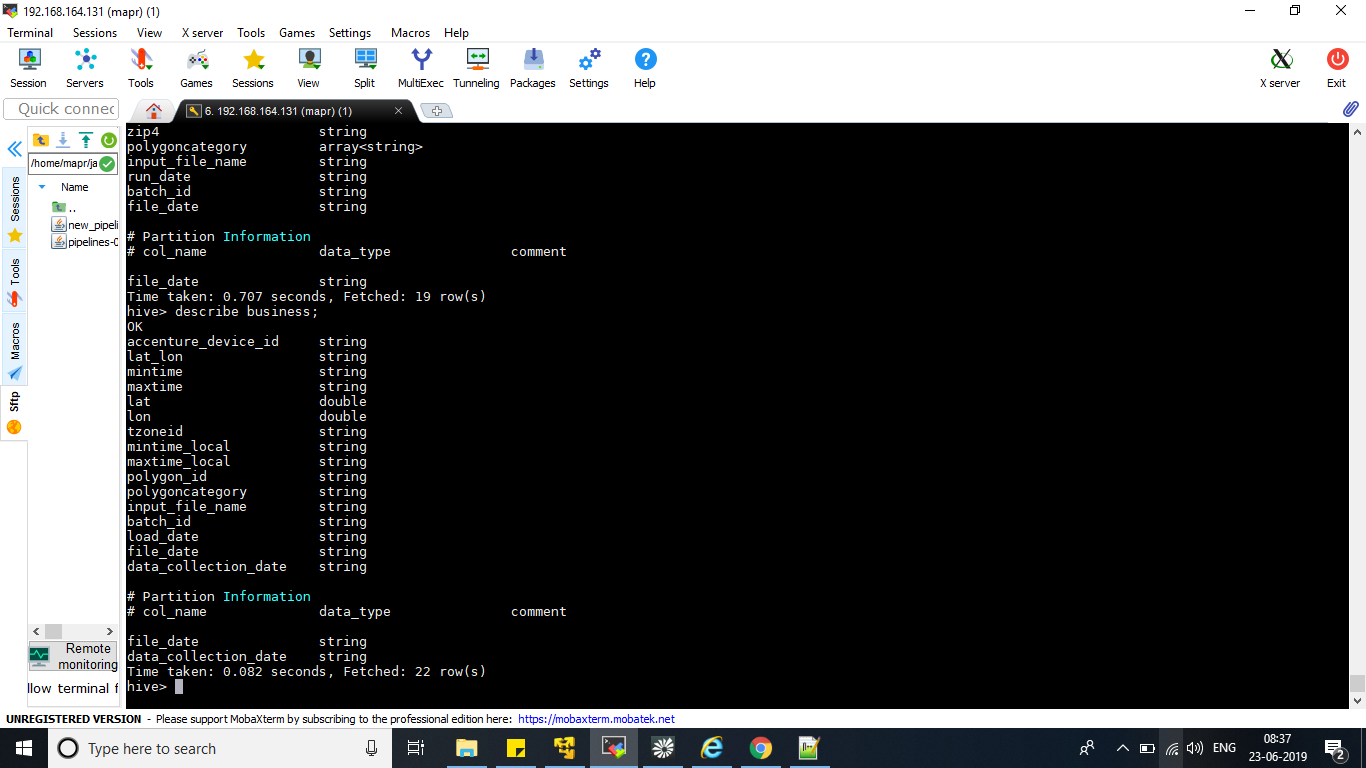












**Conclusion and Future scope of the Project**

Data analysis is a process through we are processing data, cleaning and modified data using tools.

After all those processes completion final result show in the form of good records, using of good data we are making strategies and get desired business results.

This analyzed data we are using for market research

for example, Banking, HealthCare, Government and Education.

Data analysis provide both of them speed and accuracy for business analysis.

Business have realized the importance of utilizing big data to maximize the profit.

All they know this is very important to business purpose and in the sector of healthcare.

As a result, we can say that big data analysis is more important part becoming in the future to making the big economy of the country.

At last if we can say that big data is big buzz in the market, all e-commerce company and social sites like Facebook, Instagram, WhatsApp are using big data technology Hadoop.