Table of Content

1. Introduction

1.1 **What is Big Data?** Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. Challenges include capture, storage, analysis, search, sharing, transfer, visualization, querying, updating and information privacy. The term "big data" often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set.

1.1.1 Volume, Velocity & Variety have been the conventional distinguishing factors of Big Data. This triad of factors are explained as follows:-

Volume - Indicates the amount of data to be processed.

Velocity - The fast rate at which data is received and (perhaps) acted upon.

Variety - The types of data available. Broadly classified as structured or unstructured and include text, audio, video, streams etc.

1.1.2 The evolution of Big Data technologies has brought to fore certain other factors such as veracity, variability, visualization and value which characterize Big Data. These are broadly explained as follows:-

Veracity - Veracity refers to the quality of data in terms of its trustworthiness/ reliability and accuracy. Since data is derived from many different sources, it is important to assess its quality so as to allow ease in linking, matching, cleaning and transforming the data across systems.

Variability - Data flows are unpredictable and change often. These could vary based on event triggers or seasonal factors.

Visualization - Capability to visualize data for analytic purposes.

Value - Perhaps, in the current scenario where there is a deluge of available data, value is the most important aspect. It primarily implies to the Merit and associated advantages that the analysis of this data offers in providing Business Solutions. Thus necessitating its storage and analysis.

1.2 **Hadoop**.

1.2.1 Hadoop is the Apache open source software framework for reliable, scalable, distributed computing Big Data. It hides the underlying system details and complexities from user and is developed in Java.

1.2.2 Hadoop consists of three sub projects, namely:-

- MapReduce
- Hadoop Distributed File System (aka. HDFS)
- Hadoop Common

1.2.3 Hadoop has a large ecosystem with both open-source & proprietary Hadoop-related projects which include Hive, pig, sqoop, Hbase etc.

1.3 **Sqoop**.

1.3.1 Sqoop is a tool designed to transfer data between Hadoop and relational database servers. It is used to import data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases.

1.4 **Hive**

4.1 Hive is a data warehouse infrastructure tool to process structured data in Hadoop(used for structure and semi structured data analysis and processing). It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.

1.5 **Pig**

1.5.1 Apache Pig is a high-level data flow platform for executing MapReduce programs of Hadoop. The language used for Pig is Pig Latin. The Pig scripts get internally converted to Map Reduce jobs and get executed on data stored in HDFS. Apart from that, Pig can also execute its job in Apache Tez or Apache Spark.

1.5.2 Pig can handle any type of data, i.e., structured, semi-structured or unstructured and stores the corresponding results into Hadoop Data File System. Every task which can be achieved using Pig can also be achieved using java used in MapReduce.

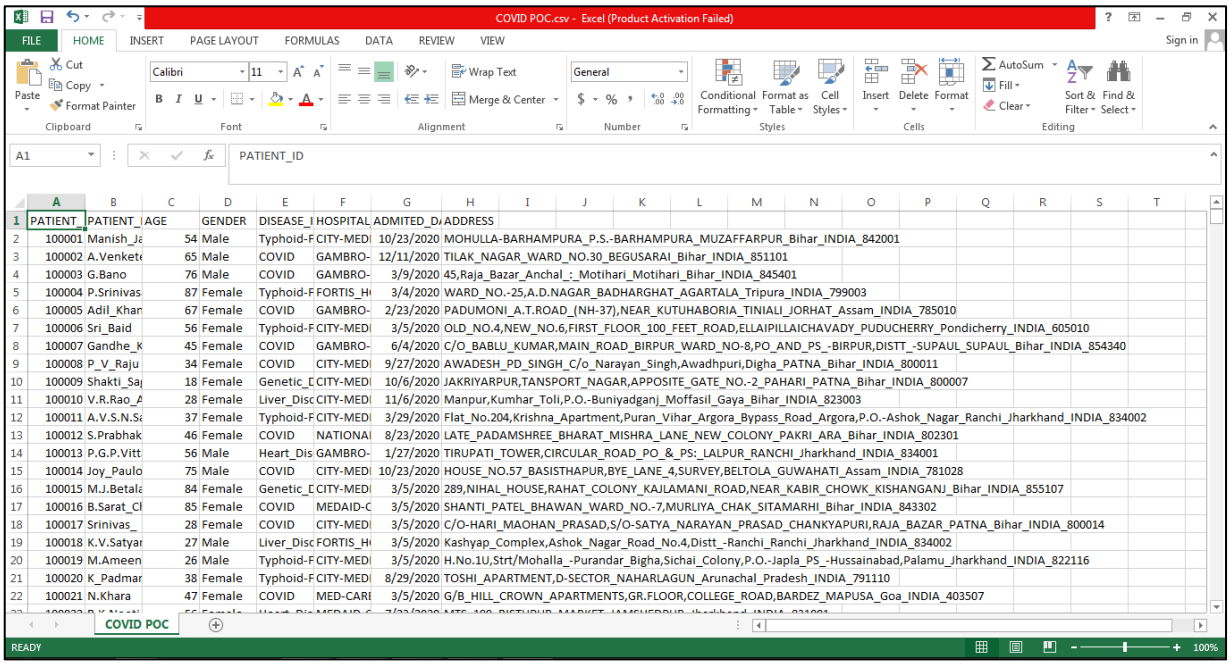
2. **Dataset**

2.1 **Brief about the Dataset -** The dataset comprises of 3000 randomly collected patient health records, in a comma separated file format (csv), from different parts of the country. It includes information about the various types of diseases contracted by different people including COVID infections. It primarily gives a sample perspective of the country wise COVID infections alongwith other diseases for which individuals were admitted in hospitals. The dataset has been cleaned and prepared for the tasks to be performed in hadoop.

2.2 **Purpose -** The dataset is being used to demonstrate the usage of various aspects of the Hadoop Software Framework/ Ecosystem.

2.3 **Dataset Attributes** - These are as follows:-

Attribute	Description	Data Type
PATIENT_ID	Unique number given to each patient	Int
PATIENT_NAME	Name of the individual	Str
AGE	Age of the individual	Int
GENDER	Male/ Female	Str
DISEASE_INFO	Single term name of the disease for which individual admitted	Str
HOSPITAL_NAME	Name of the Hospital to which individual admitted	Str
ADMITED_DATE	Date of admission	Date Time
ADDRESS	Address of the Individual	Str



PATIENT_	PATIENT_	AGE	GENDER	DISEASE_I	HOSPITAL	ADMITED_D	ADDRESS
100001	Manish_Ja	54	Male	Typhoid-F	CITY-MEDI	10/23/2020	MOHULLA-BARHAMPURA_P.S.-BARHAMPURA_MUZAFFARPUR_Bihar_INDIA_842001
100002	A.Venkete	65	Male	COVID	GAMBRO-	12/11/2020	TILAK_NAGAR_WARD_NO.30_BEGUSARAI_Bihar_INDIA_851101
100003	G.Bano	76	Male	COVID	GAMBRO-	3/9/2020	45,Raja_Bazar_Anchal_:_Motihari_Motihari_Bihar_INDIA_845401
100004	P.Srinivas	87	Female	Typhoid-F	FORTIS_H	3/4/2020	WARD_NO.-25,A.D.NAGAR_BADHARGHAT_AGARTALA_Tripura_INDIA_799003
100005	Adil_Khan	67	Female	COVID	GAMBRO-	2/23/2020	PADUMONI_A.T.ROAD_(NH-37),NEAR_KUTUHABORIA_TINIALI_JORHAT_Assam_INDIA_785010
100006	Sri_Baid	56	Female	Typhoid-F	CITY-MEDI	3/5/2020	OLD_NO.4,NEW_NO.6,FIRST_FLOOR_100_FEET_ROAD,ELLAIPILLAICHAVADY_PUDUCHERRY_Pondicherry_INDIA_605010
100007	Gandhe_K	45	Female	COVID	GAMBRO-	6/4/2020	C/O_BABLU_KUMAR,MAIN_ROAD_BIRPUR_WARD_NO-8,PO_AND_PS_-BIRPUR,DISTT_-SUPAUL_SUPAUL_Bihar_INDIA_854340
100008	P_V_Raju	34	Female	COVID	CITY-MEDI	9/27/2020	AWADESH_PD_SINGH_C/o_Narayan_Singh,Awadhpuri,Digha_PATNA_Bihar_INDIA_800011
100009	Shakti_Sa	18	Female	Genetic_D	CITY-MEDI	10/6/2020	JAKRIYARPUR,TANSPORT_NAGAR,APPOSITE_GATE_NO.-2_PAHARI_PATNA_Bihar_INDIA_800007
100010	V.R.Rao_A	28	Female	Liver_Disc	CITY-MEDI	11/6/2020	Manpur,Kumhar_Toli,P.O.-Buniyadganj_Moffasil_Gaya_Bihar_INDIA_823003
100011	A.V.S.N.Sa	37	Female	Typhoid-F	CITY-MEDI	3/29/2020	Flat_No.204,Krishna_Apartment,Puran_Vihar_Argora_Bypass_Road_Argora,P.O.-Ashok_Nagar_Ranchi_Jharkhand_INDIA_834002
100012	S.Prabhak	46	Female	COVID	NATIONAL	8/23/2020	LATE_PADAMSHREE_BHARAT_MISHRA_LANE_NEW_COLONY_PAKRI_ARA_Bihar_INDIA_802301
100013	P.G.P.Vitt	56	Male	Heart_Dis	GAMBRO-	1/27/2020	TIRUPATI_TOWER,CIRCULAR_ROAD_PO_&_PS:_LALPUR_RANCHI_Jharkhand_INDIA_834001
100014	Joy_Paulo	75	Male	COVID	CITY-MEDI	10/23/2020	HOUSE_NO.57_BASISTHAPUR,BYE_LANE_4,SURVEY,BELTOLA_GUWAHATI_Assam_INDIA_781028
100015	M.J.Betala	84	Female	Genetic_D	CITY-MEDI	3/5/2020	289,NIHAL_HOUSE,RAHAT_COLONY_KAJLAMANI_ROAD,NEAR_KABIR_CHOWK_KISHANGANJ_Bihar_INDIA_855107
100016	B.Sarat_Cl	85	Female	COVID	MEDAID-C	3/5/2020	SHANTI_PATEL_BHAWAN_WARD_NO.-7,MURLIYA_CHAK_SITAMARHI_Bihar_INDIA_843302
100017	Srinivas_	28	Female	COVID	CITY-MEDI	3/5/2020	C/O-HARI_MAOHAN_PRASAD,S/O-SATYA_NARAYAN_PRASAD_CHANKYAPURI,RAJA_BAZAR_PATNA_Bihar_INDIA_800014
100018	K.V.Satyar	27	Male	Liver_Disc	FORTIS_H	3/5/2020	Kashyap_Complex,Ashok_Nagar_Road_No.4,Distt_-Ranchi_Ranchi_Jharkhand_INDIA_834002
100019	M.Ameen	26	Male	Typhoid-F	CITY-MEDI	3/5/2020	H.No.1U,Strt/Mohalla_-Purandar_Bigha,Sichai_Colony,P.O.-Japla_PS_-Hussainabad,Palamu_Jharkhand_INDIA_822116
100020	K_Padmar	38	Female	Typhoid-F	CITY-MEDI	8/29/2020	TOSHI_APARTMENT,D-SECTOR_NAHARLAGUN_Arunachal_Pradesh_INDIA_791110
100021	N.Khara	47	Female	COVID	MED-CARE	3/5/2020	G/B_HILL_CROWN_APARTMENTS,GR.FLOOR,COLLEGE_ROAD,BARDEZ_MAPUSA_Goa_INDIA_403507

Fig 1. Screenshot of Dataset

2.4 **Objectives** - The dataset is being used to derive the following information/queries:-

2.3.1 Select top 10 records from the table.

2.3.2 Select all COVID cases recorded.

2.3.3 Find records with age > 35 && Gender is ‘Male’ & Disease_Info is COVID.

2.3.4 Find records with age >= 50 && Gender is ‘Male’ & Disease_Info is COVID.

2.3.5 Average age of patients with COVID in the sample.

2.3.6 Use PIG Script to filter the Map Reduce Output in the following manner:-
- Provide Age wise data
- Provide age and disease relationship (using ‘foreach’ and ‘generate’ in Pig)
- Isolate the occurrence of COVID in males.

3. Execution

3.1 Hadoop Distributed File System (HDFS)

3.1.1 Moving data into Hadoop Distributed File System (HDFS) -

Step 1. Make a directory to store data. (Dir - ‘pocinput’ created)

hadoop fs –mkdir /user/cloudera/pocinput

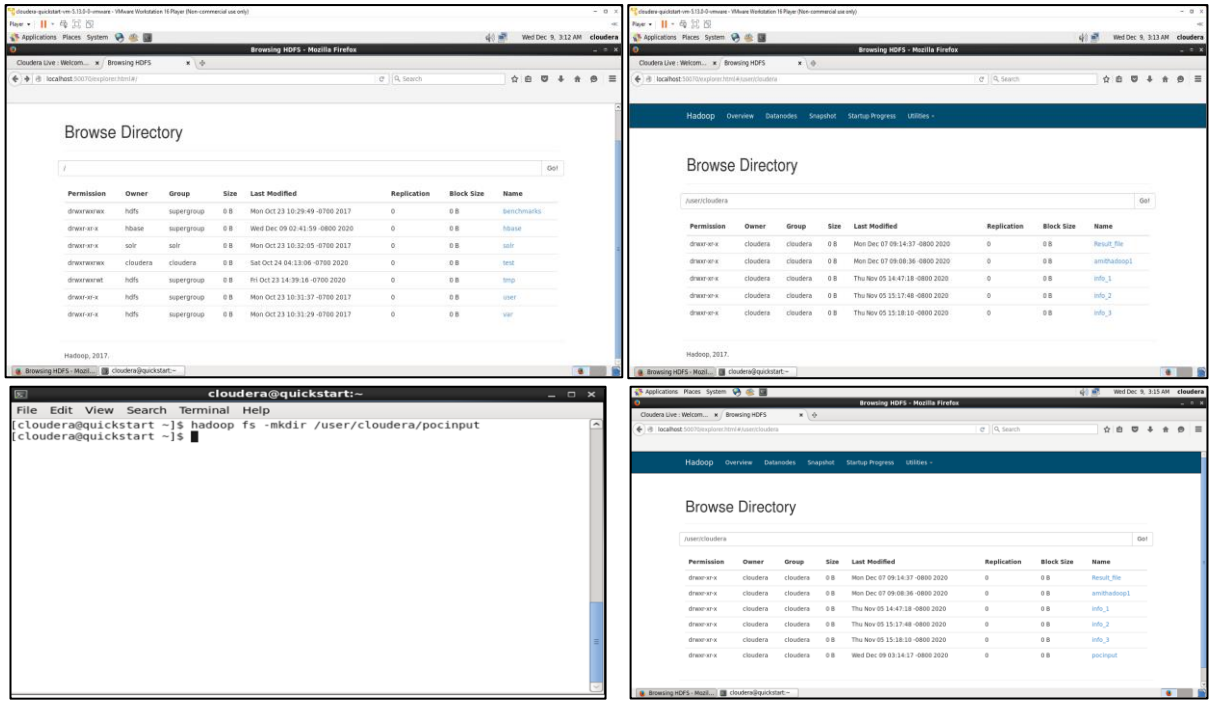


Fig 2. Screenshots of 3.1.1 (Step 1)

Step 2. Move the dataset into the ‘/pocinput’ directory

hadoop fs –put /home/cloudera/Desktop/covid_poc.csv /user/cloudera/pocinput

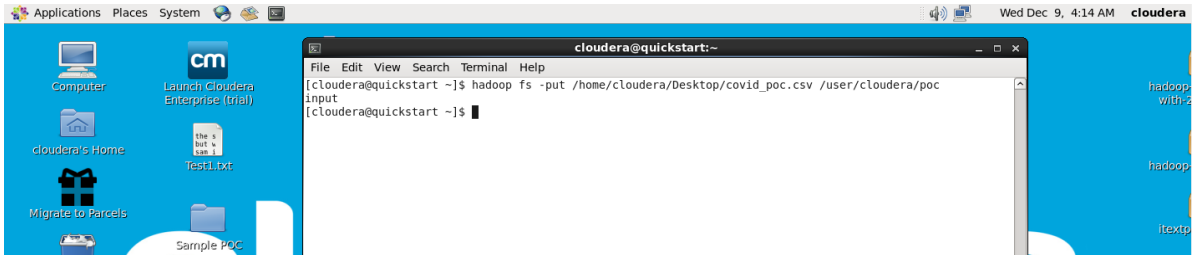
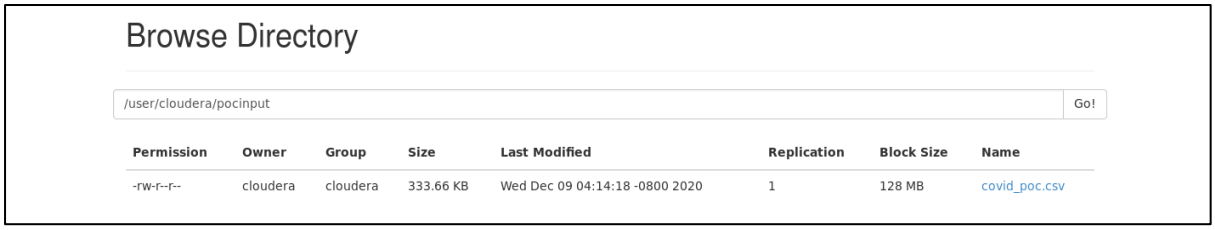


Fig 3a. Screenshot of 3.1.1 (Step 2)



Browse Directory

/user/cloudera/pocinput Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	cloudera	cloudera	333.66 KB	Wed Dec 09 04:14:18 -0800 2020	1	128 MB	covid_poc.csv

Fig 3b. Screenshot of 3.1.1 (Step 2) – Dataset loaded

3.1.2 For queries in hive we will load the data in HDFS and for queries in Pig we will use the data from local file system.

3.2 <u>Sqoop</u>

3.2.1 Loading Data into Sqoop environments.

Step 1. Create the table in sql inorder to insert data into it from sqoop.

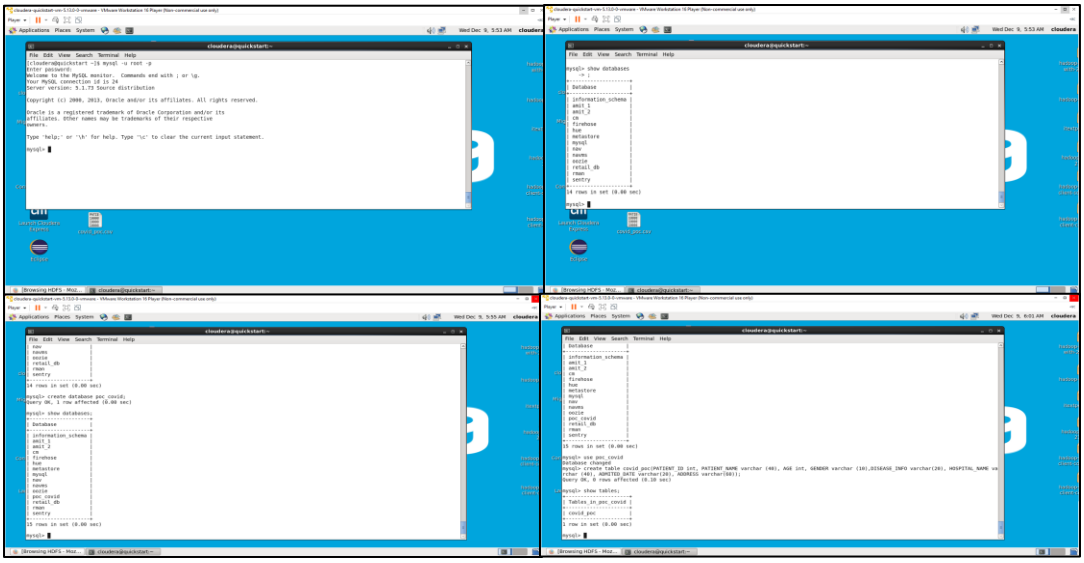


Fig 4. Screenshot of 3.2.1 (Step 1)

Step 2. Write 'sqoop' command on the terminal. This will open the shell.

Step 3. Create the table in sql and insert data into it from sqoop.

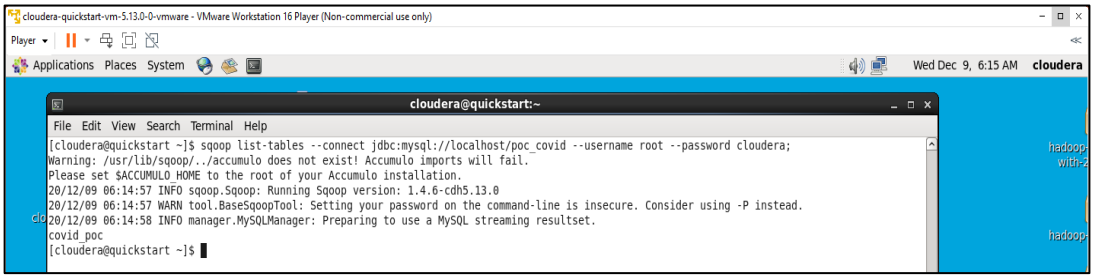


Fig 5a. Screenshot of 3.2.1 (Step 3)

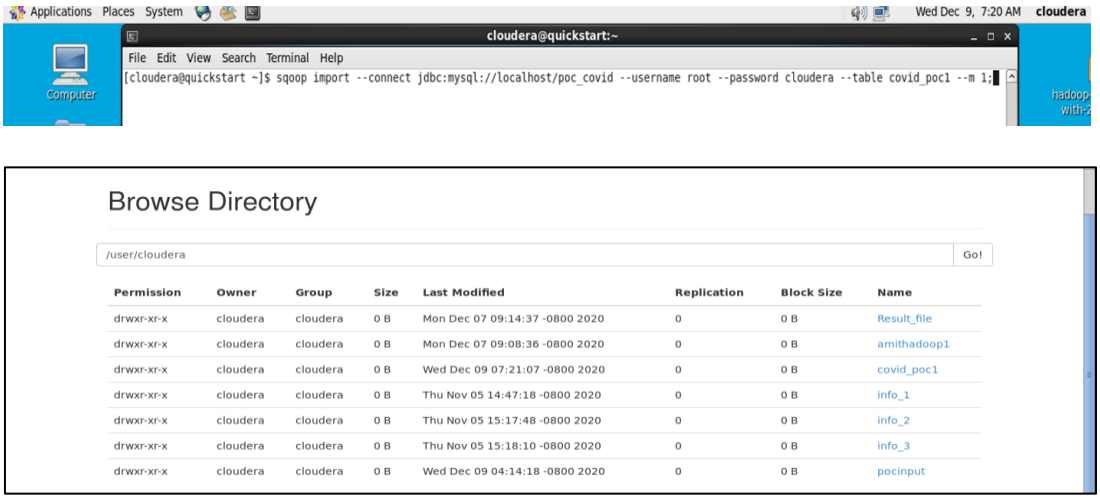


Fig 5b. Screenshot of 3.2.1 (Step 3)

3.3 **Hive**

3.3.1 **Load Data into hive environments.**

Step 1. Open the terminal.

Step 2. Write the 'hive' command on the terminal. This will open the hive shell.

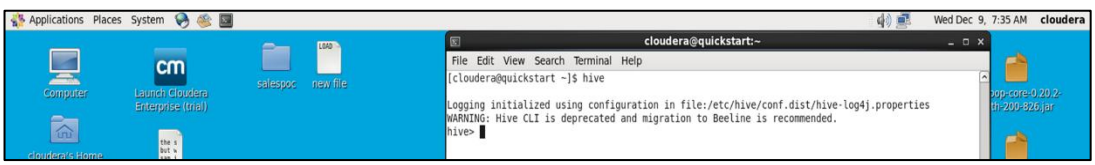


Fig 6. Screenshot of 3.3.1 (Step 1 & 2)

Step 3. Create the table in hive and load data into it from HDFS for data analysis.

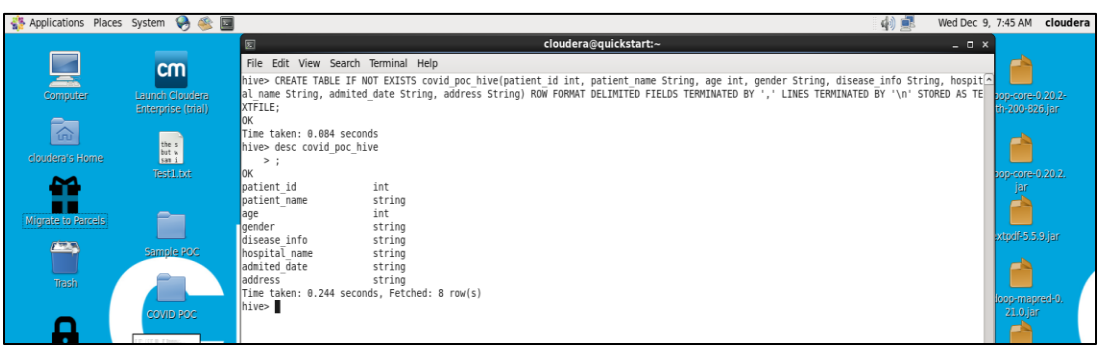


Fig 7. Screenshot of 3.3.1 (Step 3)

3.3.2 **Answers to Queries** as set out in the objective are as follows:-

Query 2.3.1. Select top 10 records from the table.

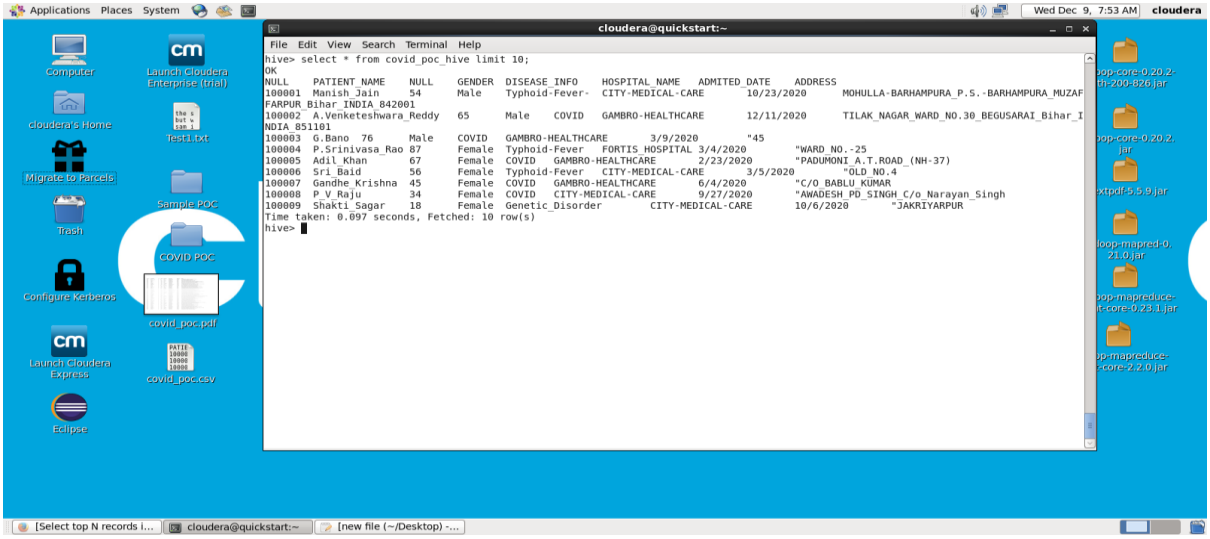


Fig 8. Screenshot of Query 2.3.1

Query 2.3.2 Select all COVID cases recorded.

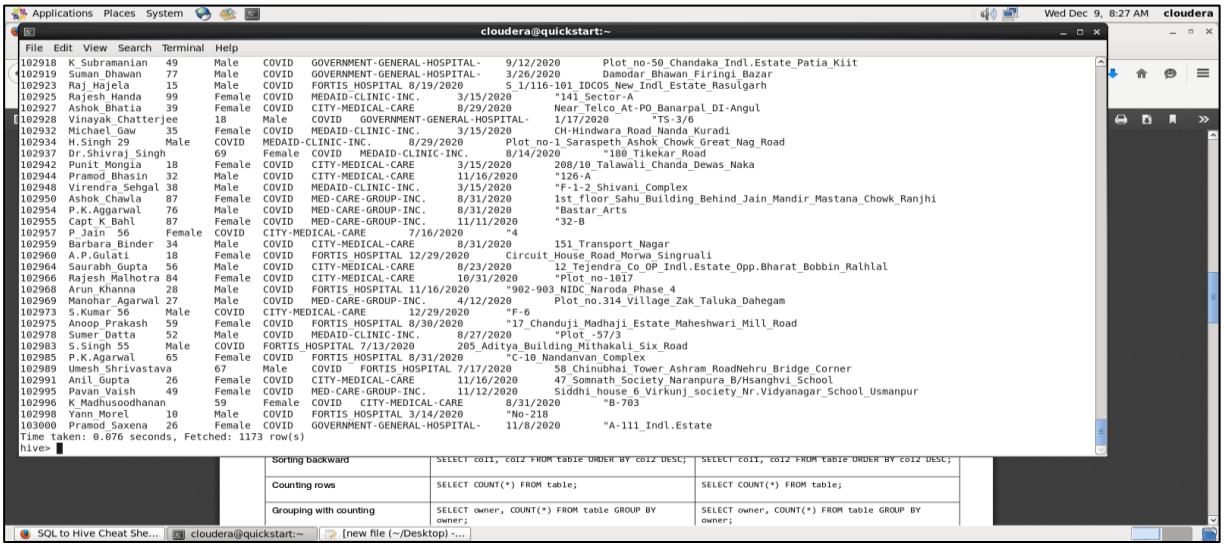


Fig 9. Screenshot of Query 2.3.2

Query 2.3.3 Find records with age > 35 && Gender is ‘Male’ & Disease_Info is COVID (Query changed to age = 35 for showing results)

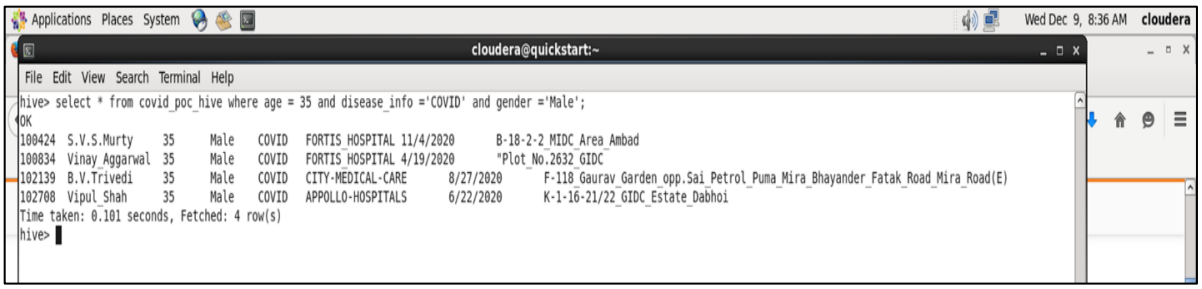


Fig 10. Screenshot of Query 2.3.3

Query 2.3.4 Find records with age >= 50 && Gender is ‘Male’ & Disease_Info is COVID.

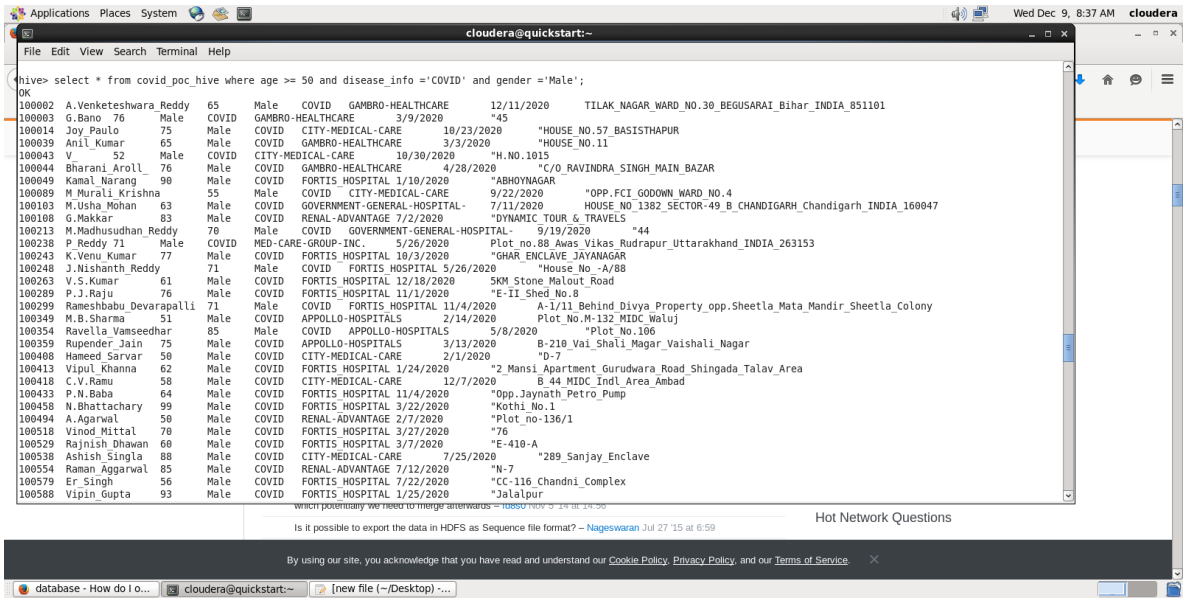


Fig 11. Screenshot of Query 2.3.4

3.4 <u>**Pig**</u>

3.4.1 Load Data into Pig environments.

Step 1. Open the terminal.
Step 2. Since working in local file system, hence type, Pig Command (pig –x local) on the terminal. This will open the grunt shell in the local mode.

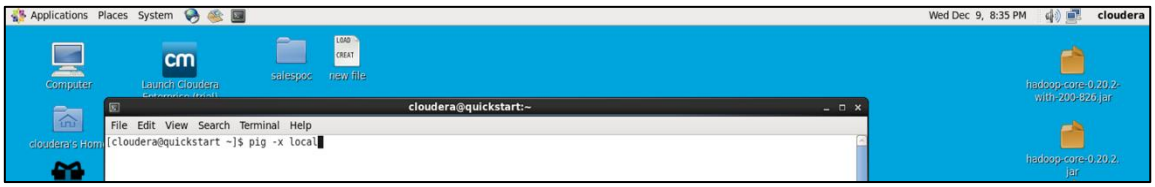
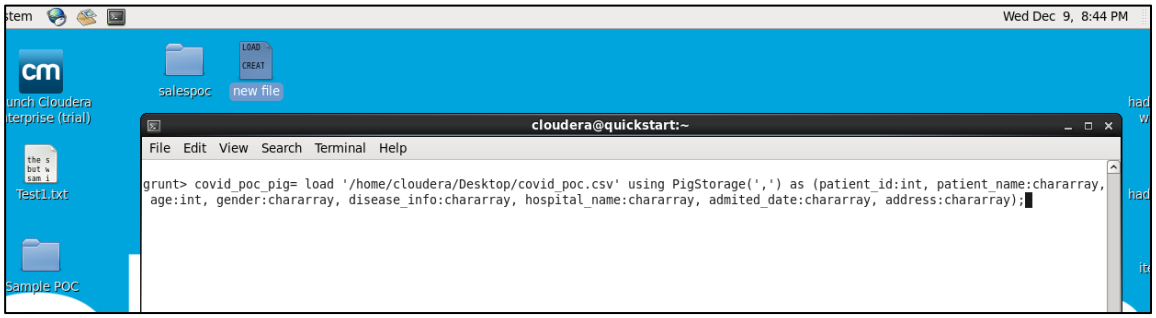


Fig 12. Running Pig

Step 3. Create a bag (table) and load data into it from local file system.



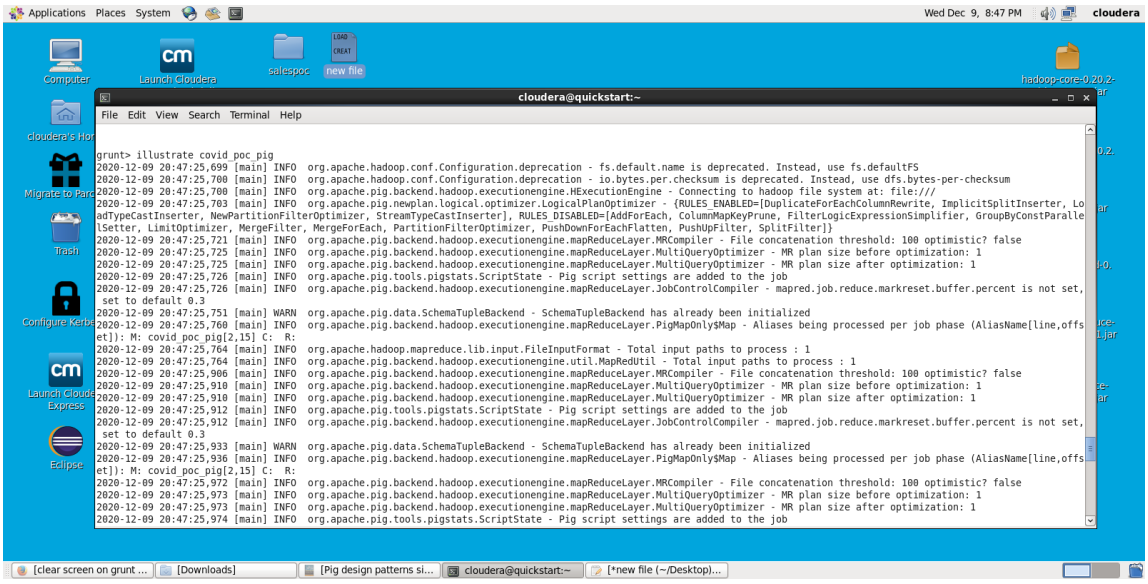


Fig 13. Creating &describing Table in Pig

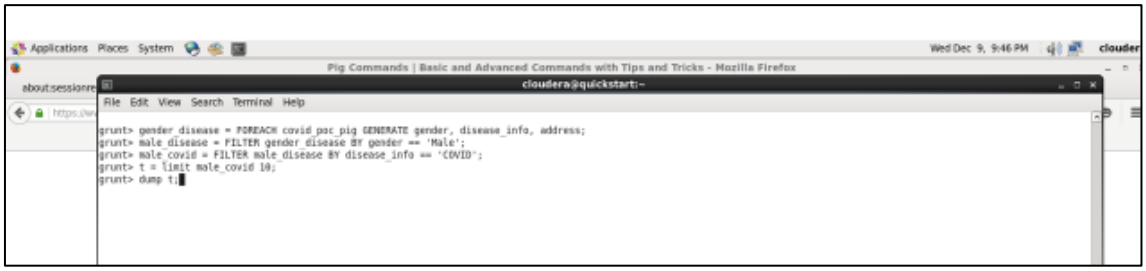
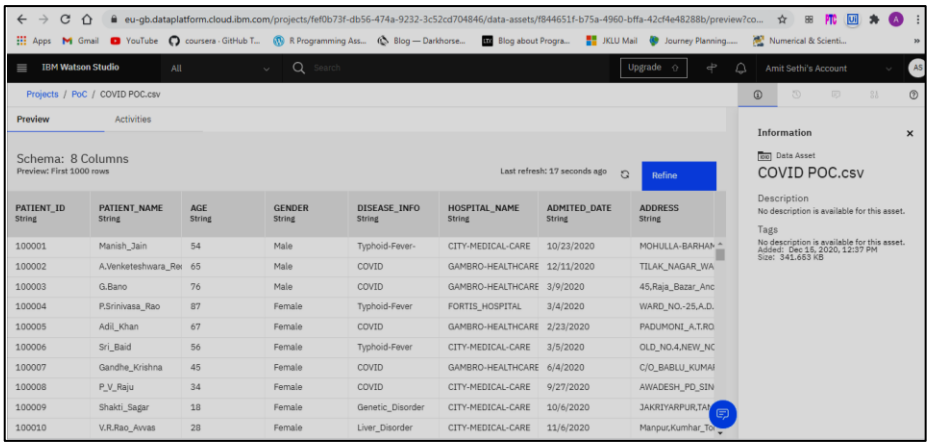


Fig 14. 'Foreach' Query in Pig

3.5 **IBM Watson**

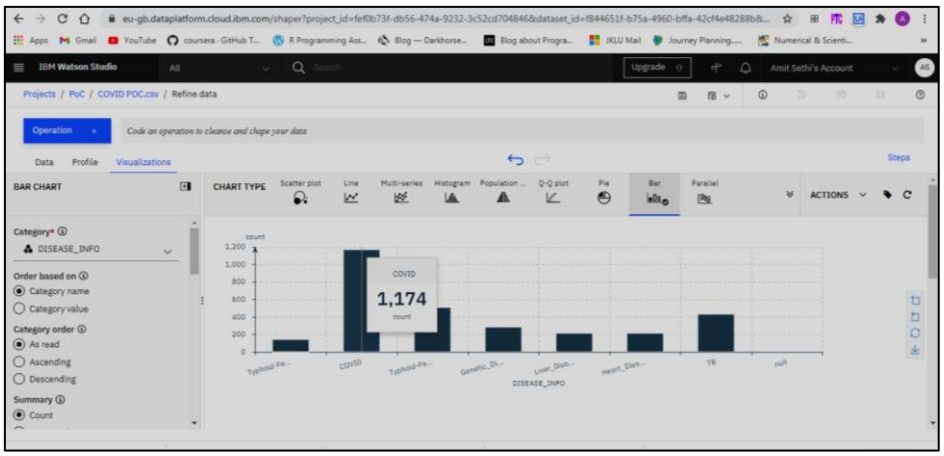
3.5.1 Inorder to utilize IBM Watson, a Lite account was created on https://cloud.ibm.com. The dataset was uploaded and some of the queries were answered utilizing the visualization module of IBM Watson.

3.5.2 Select top 10 records from the table. Shown in IBM Watson.

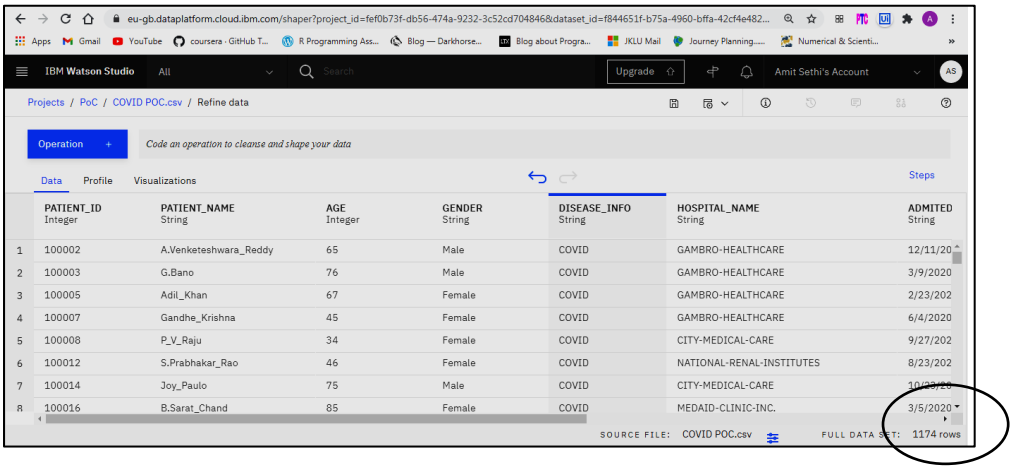


Top 10 Records

3.5.3 Select all COVID cases recorded.

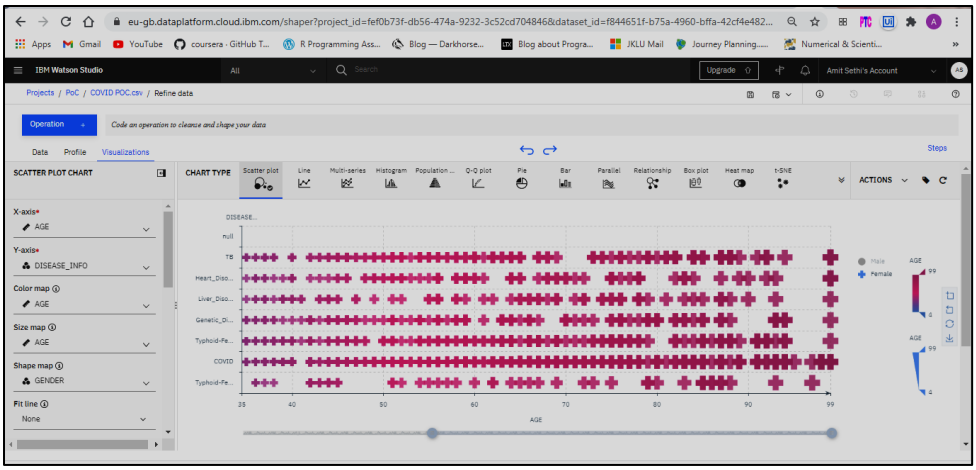


Bar Chart Depicting total COVID cases

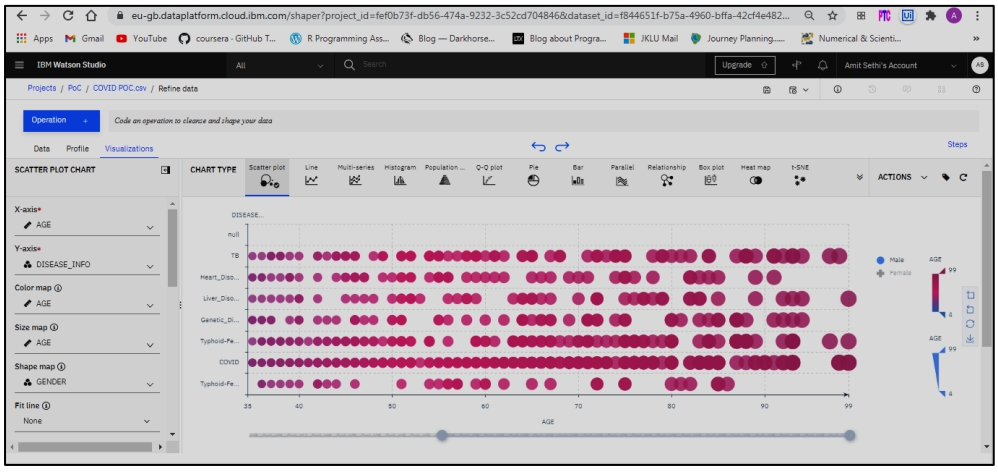


Query giving 1174 records which are COVID cases

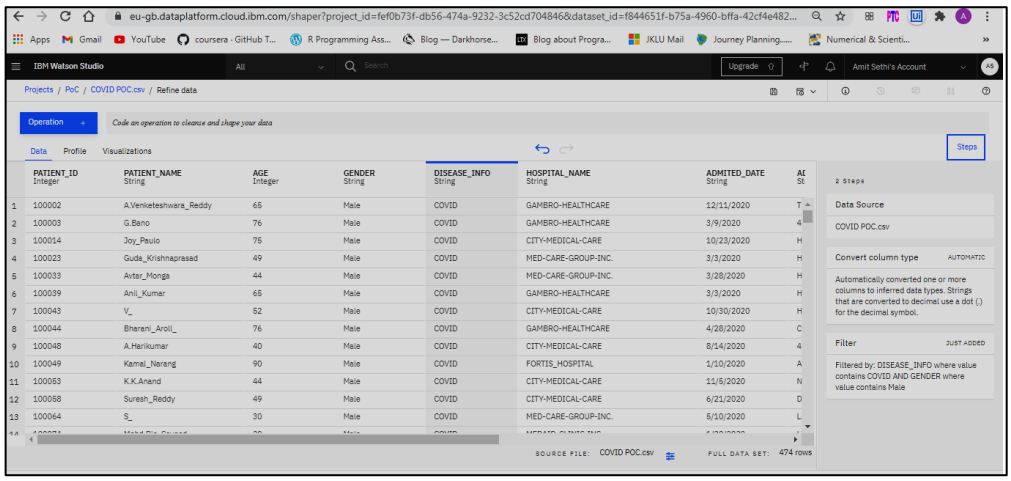
3.5.4 Find records with age > 35 && Gender is ‘Male’ & Disease_Info is COVID.



Female COVID individuals with age greater than 35

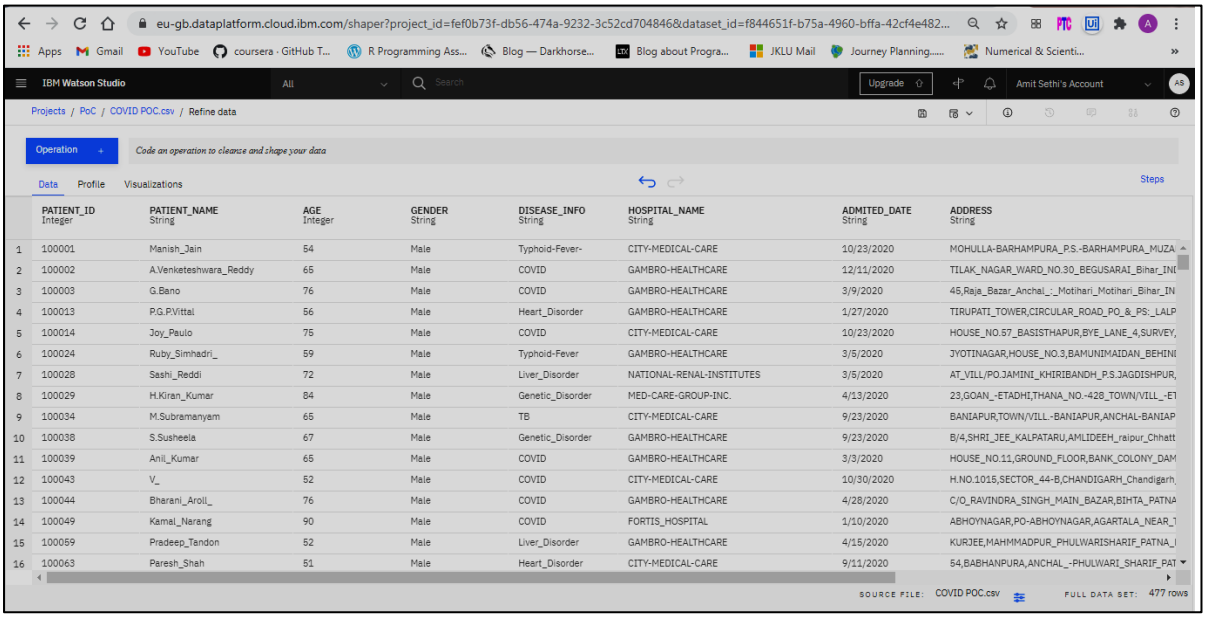


Male COVID individuals with age greater than 35



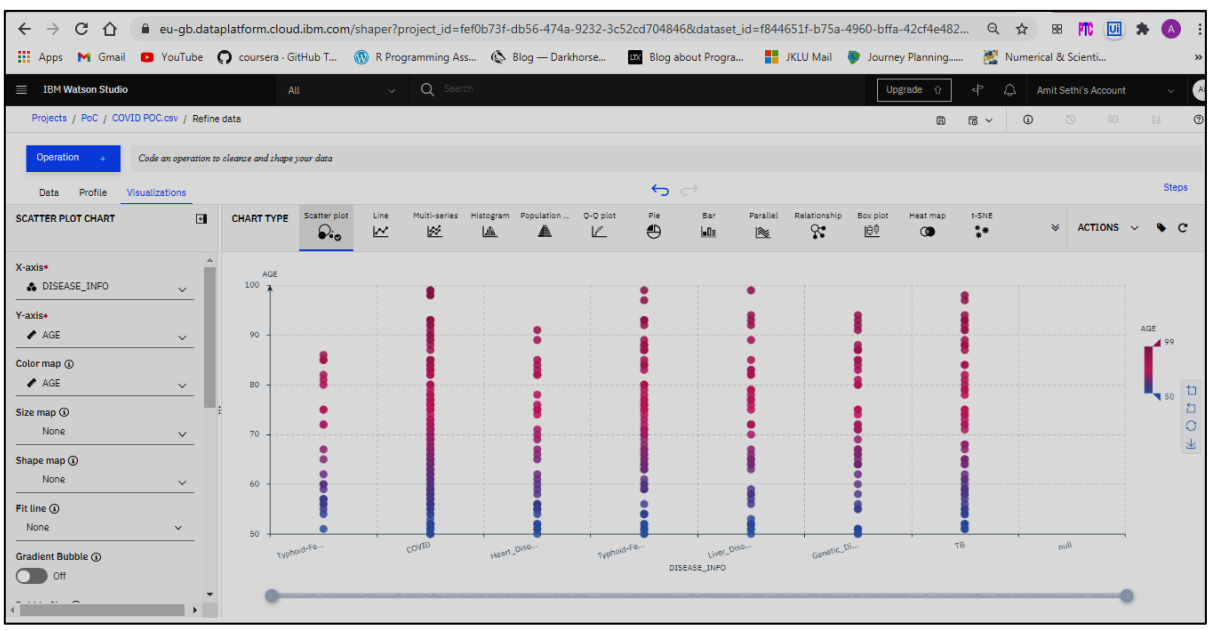
Filter Record Output for Male COVID with age greater than 35

3.5.5 Find records with age >= 50 && Gender is ‘Male’ & Disease_Info is COVID.



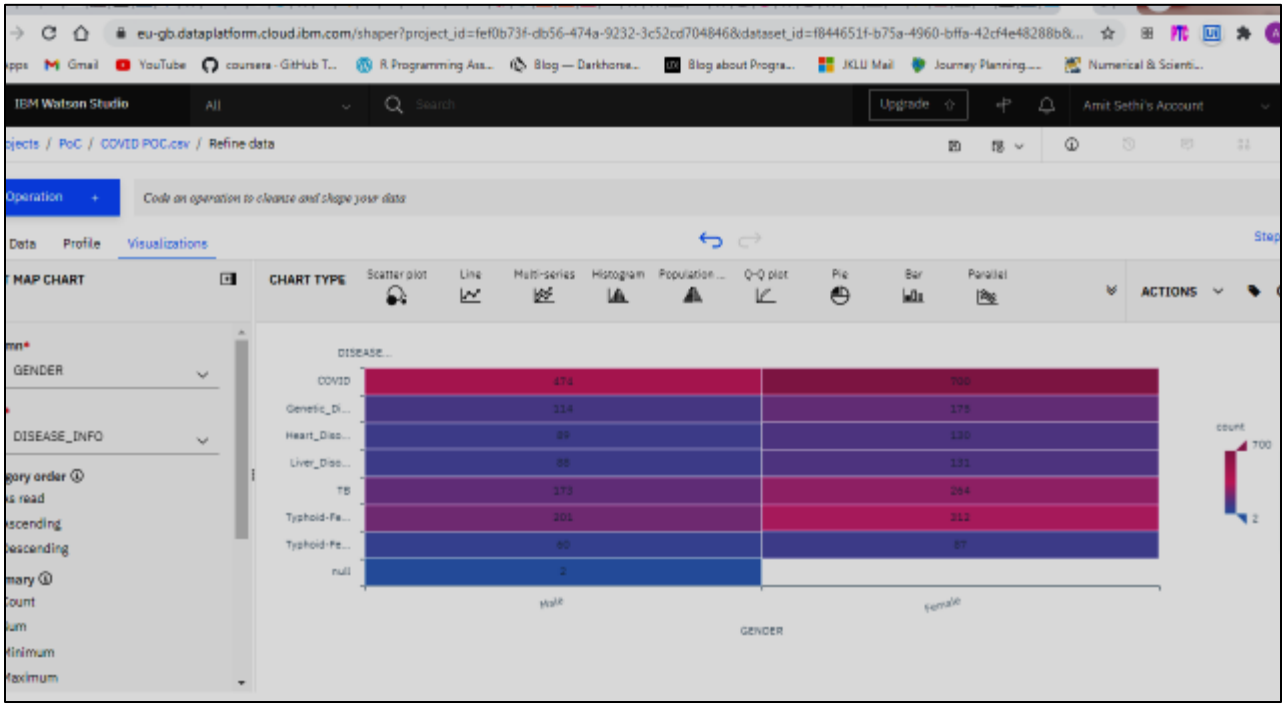
	PATIENT_ID Integer	PATIENT_NAME String	AGE Integer	GENDER String	DISEASE_INFO String	HOSPITAL_NAME String	ADMITED_DATE String	ADDRESS String
1	100001	Manish_Jain	54	Male	Typhoid-Fever-	CITY-MEDICAL-CARE	10/23/2020	MOHULLA-BARHAMPURA_P.S.-BARHAMPURA_MUZA
2	100002	A.Venketeshwara_Reddy	65	Male	COVID	GAMBRO-HEALTHCARE	12/11/2020	TILAK_NAGAR_WARD_NO.30_BEGUSARAI_Bihar_IN
3	100003	G.Bano	76	Male	COVID	GAMBRO-HEALTHCARE	3/9/2020	45,Raja_Bazar_Anchal_:_Motihari_Motihari_Bihar_IN
4	100013	P.G.P.Vittal	56	Male	Heart_Disorder	GAMBRO-HEALTHCARE	1/27/2020	TIRUPATI_TOWER,CIRCULAR_ROAD_PO_&_PS:_LALP
5	100014	Joy_Paulo	75	Male	COVID	CITY-MEDICAL-CARE	10/23/2020	HOUSE_NO.57_BASISTHAPUR,BYE_LANE_4,SURVEY,
6	100024	Ruby_Simhadri_	59	Male	Typhoid-Fever	GAMBRO-HEALTHCARE	3/5/2020	JYOTINAGAR,HOUSE_NO.3,BAMUNIMAIDAN_BEHINI
7	100028	Sashi_Reddi	72	Male	Liver_Disorder	NATIONAL-RENAL-INSTITUTES	3/5/2020	AT_VILL/PO.JAMINI_KHIRIBANDH_P.S.JAGDISHPUR,
8	100029	H.Kiran_Kumar	84	Male	Genetic_Disorder	MED-CARE-GROUP-INC.	4/13/2020	23,GOAN_-ETADHI,THANA_NO.-428_TOWN/VILL_-ET
9	100034	M.Subramanyam	65	Male	TB	CITY-MEDICAL-CARE	9/23/2020	BANIAPUR,TOWN/VILL.-BANIAPUR,ANCHAL-BANIAP
10	100038	S.Susheela	67	Male	Genetic_Disorder	GAMBRO-HEALTHCARE	9/23/2020	B/4,SHRI_JEE_KALPATARU,AMLIDEEH_raipur_Chhatt
11	100039	Anil_Kumar	65	Male	COVID	GAMBRO-HEALTHCARE	3/3/2020	HOUSE_NO.11,GROUND_FLOOR,BANK_COLONY_DAM
12	100043	V_	52	Male	COVID	CITY-MEDICAL-CARE	10/30/2020	H.NO.1015,SECTOR_44-B,CHANDIGARH_Chandigarh,
13	100044	Bharani_Aroll_	76	Male	COVID	GAMBRO-HEALTHCARE	4/28/2020	C/O_RAVINDRA_SINGH_MAIN_BAZAR,BIHTA_PATNA
14	100049	Kamal_Narang	90	Male	COVID	FORTIS_HOSPITAL	1/10/2020	ABHOYNAGAR,PO-ABHOYNAGAR,AGARTALA_NEAR_T
15	100059	Pradeep_Tandon	52	Male	Liver_Disorder	GAMBRO-HEALTHCARE	4/15/2020	KURJEE,MAHMMADPUR_PHULWARISHARIF_PATNA_I
16	100063	Paresh_Shah	51	Male	Heart_Disorder	CITY-MEDICAL-CARE	9/11/2020	54,BABHANPURA,ANCHAL_-PHULWARI_SHARIF_PAT

Male age greater than 50 and suffering from COVID

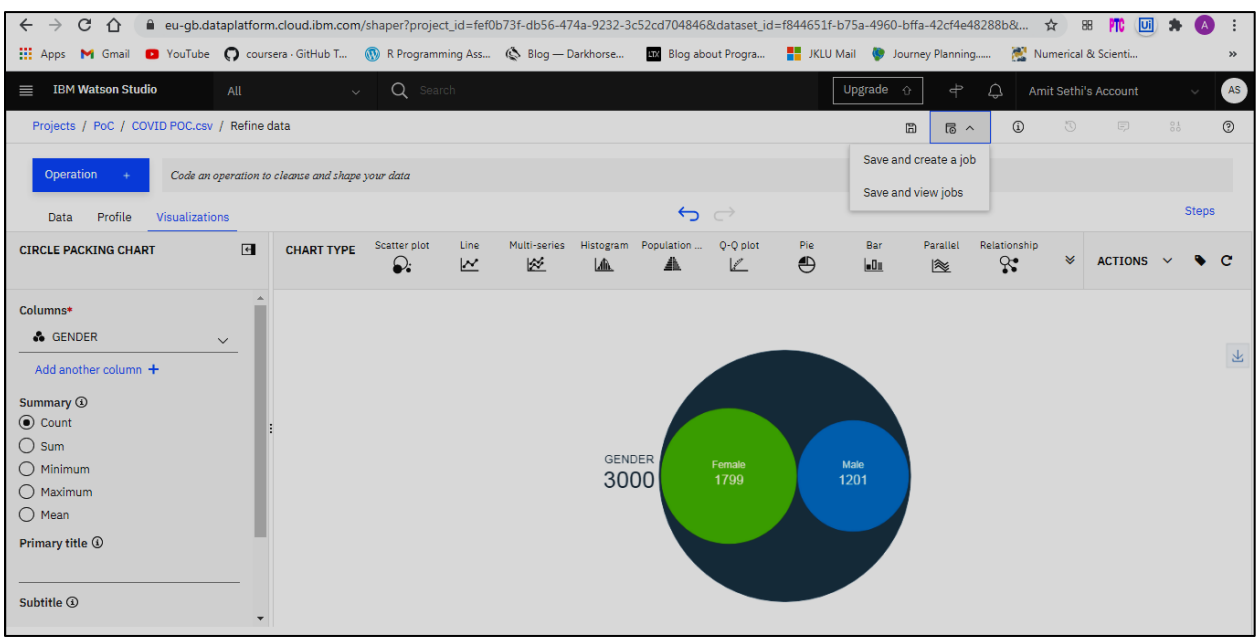


Male age Greater than 50 and sufferin from COVID as well as other diseases

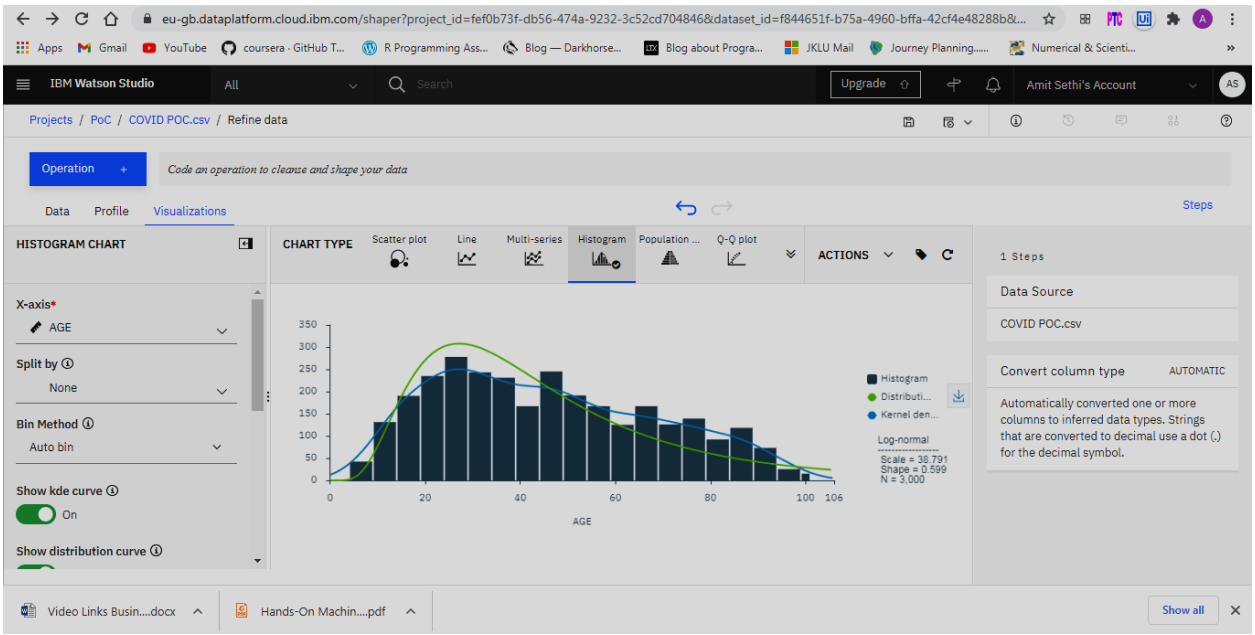
3.5.6 Other visualizations obtained from IB Watson



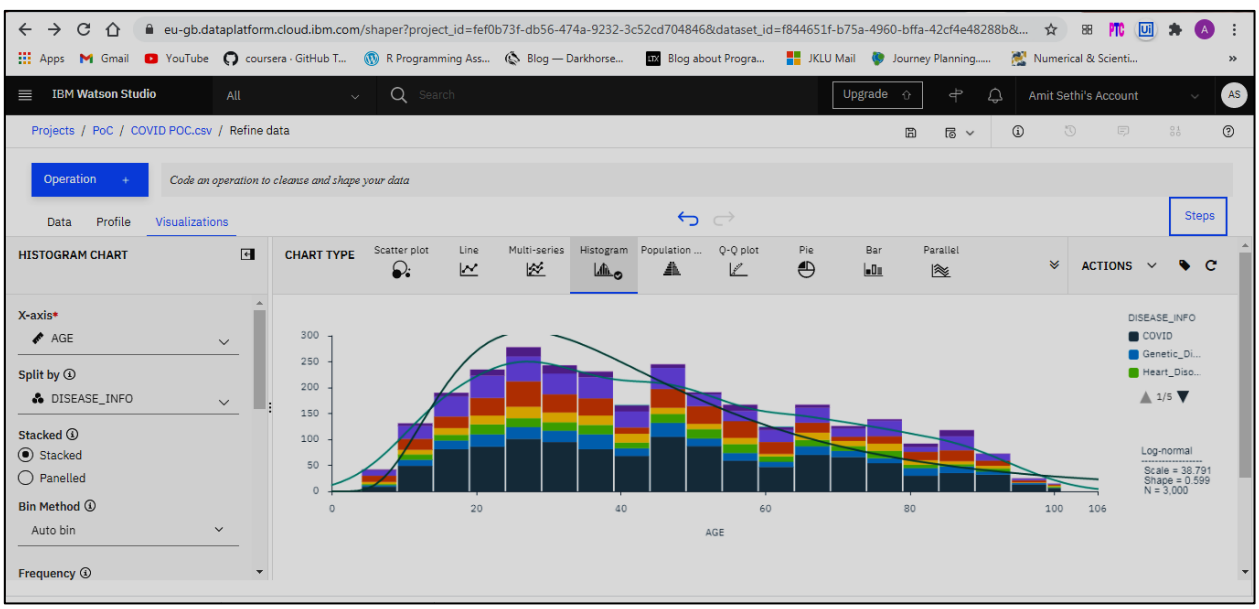
Heat Map of Male and Female gender distributed with respect to diseases. (COVID is on top)



Gender wise Distribution of Dataset



Age wise Distribution of Dataset



Age and Disease wise distribution of Dataset