Assignment #2 (Team work)

Personal Ethics & Academic Integrity Statement

Student name: Rahul Raj Sharma Student ID: 300111840

Student Name: Harman Bhutani Student ID: 300144160

By typing in my name and student ID on this form and submitting it electronically, I am attesting to the fact that I have reviewed not only my own work, but the work of my team member, in its entirety.

I attest to the fact that my own work in this project adheres to the fraud policies as outlined in the Academic Regulations in the University's Undergraduate Studies Calendar. I further attest that I have knowledge of and have respected the "Beware of Plagiarism" brochure found on the Telfer School of Management's doc-depot site. To the best of my knowledge, I also believe that each of my group colleagues has also met the aforementioned requirements and regulations. I understand that if my group assignment is submitted without a completed copy of this Personal Work Statement from each group member, it will be interpreted by the school that the missing student(s) name is confirmation of non-participation of the aforementioned student(s) in the required work.

We, by typing in our names and student IDs on this form and submitting it electronically,

- warrant that the work submitted herein is our own group members' work and not the work of others
- acknowledge that we have read and understood the University Regulations on Academic Misconduct
- acknowledge that it is a breach of University Regulations to give or receive unauthorized and/or unacknowledged assistance on a graded piece of work

Part A - For this part of the assignment please answer questions to the best explanation possible.

Note: You may consult with resources on the Internet to research and answer these the questions. However, please do not copy and paste your answers directly from a google search! The instructor is looking for your understanding of the topic in your own words.

[5 Pts for each question]

Question 1: What is RDD? Briefly explain Transformations and Actions in the context of RDDs. Give 5 examples each for transformations and actions functions with a short description. (1 line description only for the examples)

Answer: RDD : Resilient Distributed Datasets is a fundamental data structure of Spark which is an immutable distribution collection of objects which computes on the different node of the cluster. Each dataset in RDD is divided into logical partitions, which may be computed on different nodes of the cluster.

Transformation : Transformation is a operation which will transform RDD data from one form to another and when we apply this operation to any RDD, we will get the transformed data in new RDD. Transformation are of two types first one is Narrow and another one is Wide. Below are the operations performed in transformation:

1. **map(func) -** The map function iterates over every line in RDD and split into new RDD
2. **flatMap() -** With the help of **flatMap()** function, to each input element, we have many elements in an output RDD
3. **union(dataset) -** With the **union()** function, we get the elements of both the RDD in new RDD.
4. **distinct() -** It returns a new dataset that contains the **distinct** elements of the source dataset. It is helpful to remove duplicate data.
5. **coalesce() -** To avoid full shuffling of data we use coalesce() function. In **coalesce()** we use existing partition so that less data is shuffled.

Action : Transformations create RDDs from each other, but when we want to work with the actual dataset, at that point action is performed. When the action is triggered after the result, new RDD is not formed. Actions are Spark RDD operations that give non-RDD values. Below are the operations that are performed under action:

1. **collect() -** The action **collect()** is the common and simplest operation that returns our entire RDDs content to driver program.
2. **take(n) -** The action **take(n)** returns n number of elements from RDD.
3. **countByVlaue() -** The **countByValue()** returns, many times each element occur in RDD.
4. **reduce() -** The **reduce()** function takes the two elements as input from the RDD and then produces the output of the same type as that of the input elements.

5. **Aggregate() -** It gives us the flexibility to get data type different from the input type. The **aggregate()** takes two functions to get the final result.

Question 2: What is the Spark stack? Briefly explain about any 4 libraries that constitute the Spark ecosystem? (3-4 lines each, no diagrams required)

Answer: Spark is a general purpose cluster computing framework that empowers its central engine to exploit certain higher-level components. It is operationable with Apache Hadoop, in the sense that it can read and write data to and from HDFS and can also integrate the data with other storage systems that are supporated by Hadoop API. Below are the libraries that constitute the spark ecosystem :

1. **Spark Core :** In a way the Spark center is close to an operating system's kernel. It is the engine of general operation, which is both quick and sensitive of fault. Upon this main engine is designed the whole Spark ecosystem. It is designed primarily to do job preparation, task delivery and job control through worker nodes. It is also responsible for memory allocation, interaction with various unidentical storage systems and various other operations.
2. **Spark SQL :** This Spark framework is designed for querying, evaluating and running organized system operations. It is a very critical component of the whole Spark stack due to the fact that most operational data are organized, while unstructured data are increasingly that. Acting as a distributed database engine, it makes requests from Hadoop Hive to run up to 100 times faster without any change
3. **Mlib :** MLlib is the built-in machine learning library in the Spark stack. This was introduced in Spark 0.8. The aim is to render machine learning fast and scalable. Developers will use Spark SQL, Spark Streaming and GraphX effortlessly in their favorite programming language, whether it's Java, Python or Scala. MLlib includes the functions required to conduct different statistical research such as associations, regression, hypothesis checking, etc.
4. **SparkR :** The SparkR project started to combine R's mathematical analytics and machine learning capabilities with Spark's scalability. This discussed R's weakness, which was their capacity to store as many data as a single machine's memory suit. R programs will now scale over SparkR in a distributed environment.

Question 3: You have been appointed as a Data Consultant by a company that would like to take advantage of Big data analytics. You have been asked to provide a brief comparative analysis of Hadoop and Spark big data processing tools and present a summary report to the executive committee. (Maximum 1 page) You are free to make assumptions about the requirements of the company for big data analytics. Remember your assumptions will set the grounds of your preferred tool. (Refer to presentation slides on Spark for requirements assumption)

Answer: Bhutani Corporation is the leading Computer, Laptop and Accessory retailer in the country. The company has recently moved from in-store to the e-commerce platform. As a business, a year of product sales has been successfully completed both in the physical and online E-commerece store.

Harman Bhutani, the company's CEO needs an analysis of last year's business revenues, because the company is facing pressure and consumers have become more competitive–because they want business processes to be quicker, quicker, and product quality to be superior and low compared to their rivals.

Company wants to learn more about consumer desires, expectations, the desire to buy for any viable business. With the rise in data sources, Harman intended to take a step away from conventional resources. Business made use of common resources like Access and Excel.As a Data Consulatant , suggestion to company is to migrate to more advanced data analytics tools such as Apache Hadoop, Apache Spark and many more.

Hadoop is the oldest and most mature platform, but it would always suggest incorporating Spark in our setting as Apache Spark is the contemporary big data processing engine that provides rapid solutions for any vulnerabilities compared to Hadoop. It can be used effectively in identifying specific trends that are helpful to the company perspective

General difference between Spark and Hadoop

The Basis Of Comparison Between Hadoop vs Spark	**Hadoop**	**Spark**
Category	Basic Data processing engine	Data analytics engine

Usage	Batch processing with a huge volume of data	Process real-time data, from real-time events like Twitter, Facebook
Latency	High latency computing	Low latency computing
Data	Process data in batch mode	Can process interactively
Ease of use	Hadoop's MapReduce model is complex, need to handle low-level APIs	Easier to use, abstraction enables a user to process data using high-level operators
Scheduler	External job scheduler is required	In-memory computation, no external scheduler required
Security	Highly secure	Less secure as compare to Hadoop
Cost	Less costly since MapReduce model provide a cheaper strategy	Costlier than Hadoop since it has an in-memory solution

Productive Implication of Spark in ABC environment

Commonly, retail sales data include the latest product details as it is continually updated. In our company, for example, the mega deals will be updated both at the sites and in the retail stores. To uplift the company, cashback deals will be suggested in order to encourage buyers to shop at certain retail locations, as a result of which they will get huge profits. Thus, Apache Spark aims to deeply process data in real-time.

Because Apache Spark offers an application programming interface based on a data structure known as the robust distributed dataset (RDD), sales data objects would be spread around a computer cluster, which will be managed in a fault-tolerant manner.

Via Apache Spark, streaming data can be obtained in this way, which will make it easier for us to sort them into different categories, which will lead to improved business skills. These collected data items can also be used in predictive analytics such as assessing particular customer interest in various demographic and geographic aspects.

Spark is an open source and highly flexible to understand the success of industry and can actually meet the demands of ever-growing data scale.

Question 4: Briefly explain the following in the context of Apache Spark. (2 Sentences maximum)

Answer:

a. **Apache Spark :** Apache Spark is a data processing system that can execute computational tasks easily on very broad data sets, and can also spread data processing tasks through many devices, either individually or in combination with other distributed computing resources. Such two attributes are essential to the environments of big data and machine learning that involve a huge computational power organising to crunch across broad data stores.

b. **Spark Context :** SparkContext is the main entry point for spark functionality. This represents the connection to a Spark cluster, and can be used to create RDDs, accumulator and broadcast variables on that cluster. Only one SparkCluster may be active per JVM.

c. **Master & Slave :** The machine on which the spark standalone cluster managers runs is called the master node. The resources are allocated by master. It uses the "Workers" running throughout the cluster for the creation of Executors for the "Driver".
Slave node or the worker node are spawned during the life of a spark application. The execute work on the data within the node and report back to the master node.

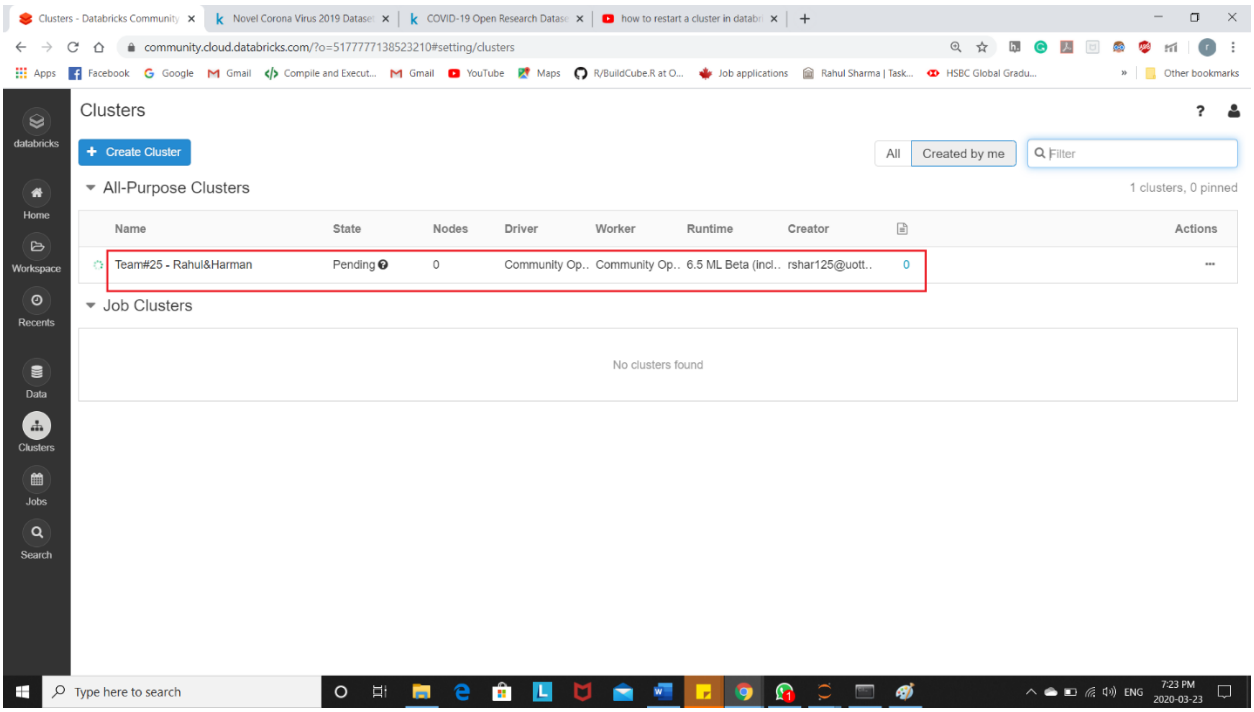
d. **Fault Tolerance :** Fault refers to failure, thus fault tolerance in Apache Spark is the potential to work and restore damage after a failure.If we want our device to be tolerant to error, it should be redundant, because we need a redundant portion to get the missing information. Redundant data is used to restore the defective data.

e. **DAG :** Directed Acyclic Graph is a finite direct graph with no directed cycles. There are finite number of vertices and edges. Where each edge directed from one vertex to another. DAG operation can do better global optimization than other systems like MapReduce.

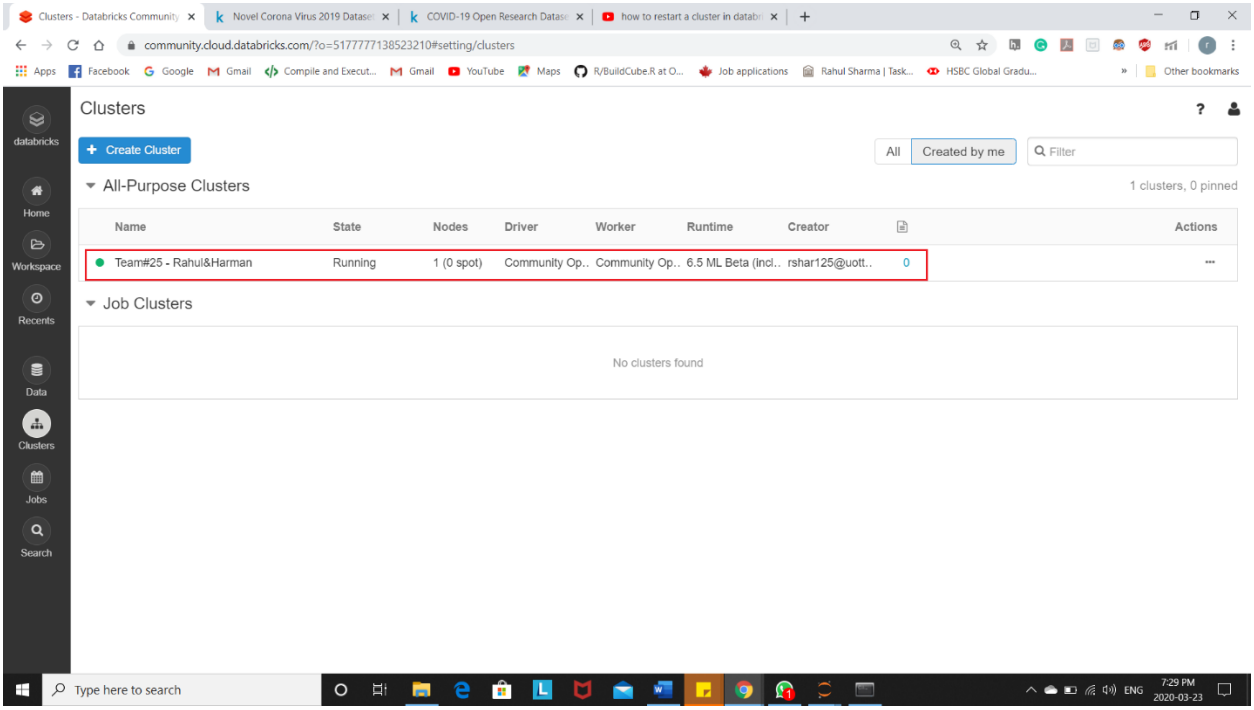
Part B – For this part of the assignment please refer to the manual posted on the course website called *Spark_Experiment_on_Databricks.doc*. Create an account on Databricks® community edition and provide snapshots of your experiment as part of the answers to the questions (5) – (8). You are free to choose any two data files. One must have normal text content (up to 50 lines) and the second data file must be a data file (in excel or CSV format, with up to 1000 records/rows). Please submit a copy of your data files along with this assignment through BrightSpace.

Question 5 [2 pts]: Create a Cluster in the databricks community edition, name the cluster in the format Team# and provide the following **SCREENSHOTS** with the cluster name clearly visible:

a) Cluster pending state – Below screenshot shows the cluster is in pending state

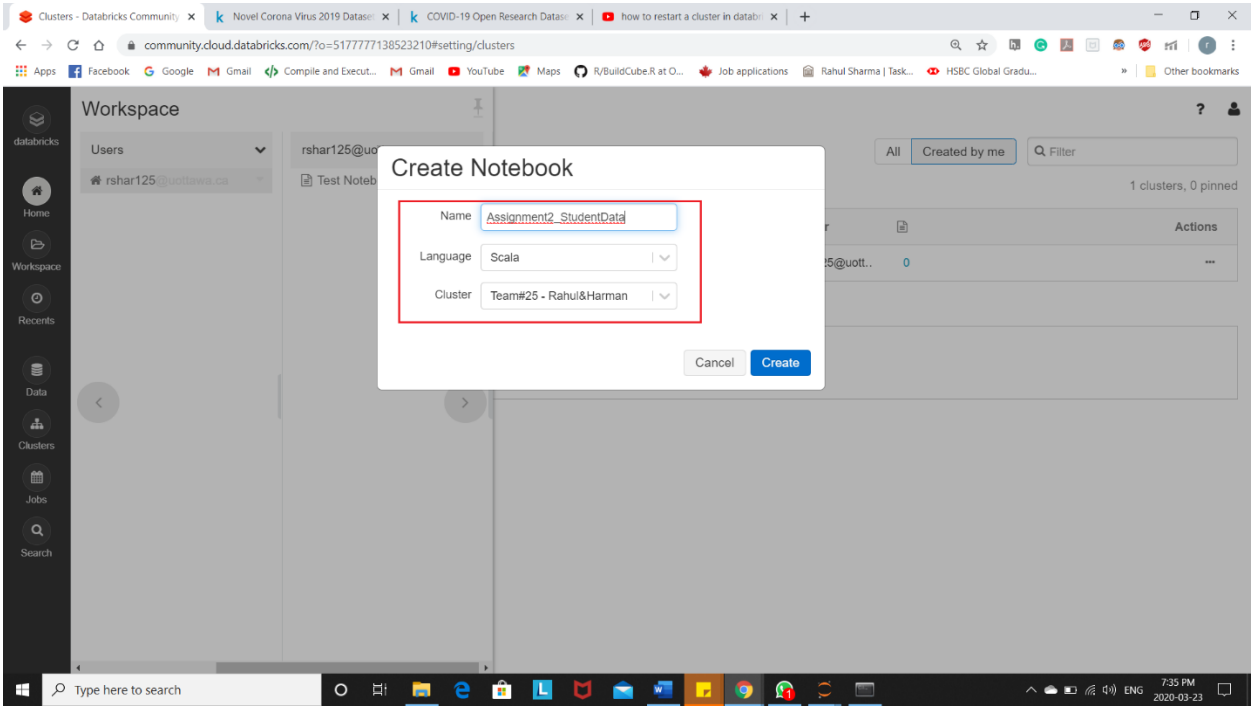


b) Cluster running state – Below screenshot shows cluster is in running state

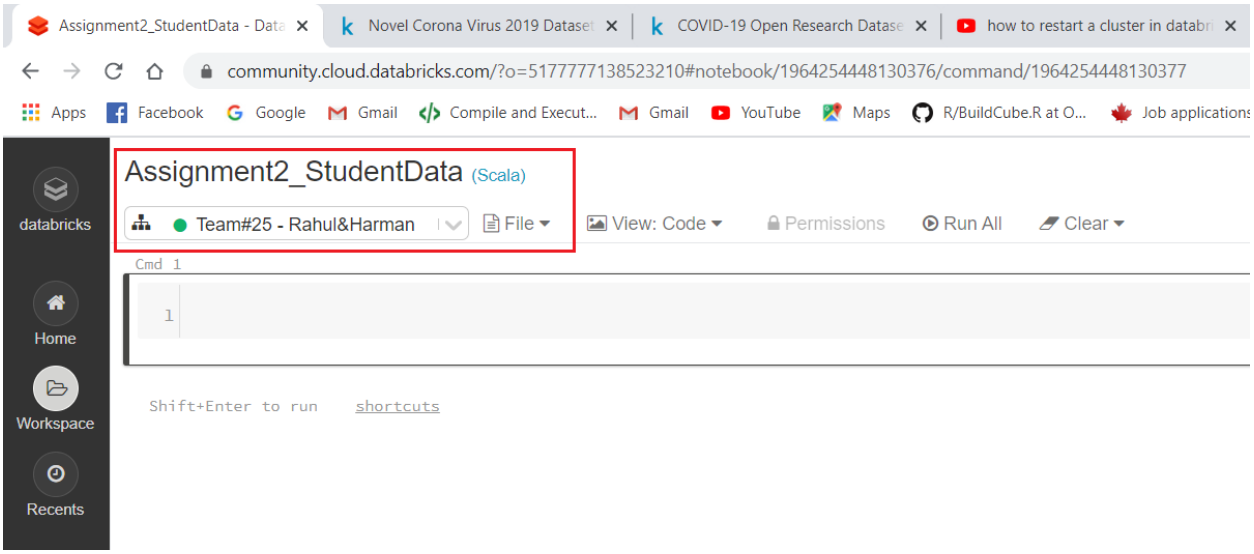


Question 6 [6 pts]: Create a notebook in Scala using your active cluster, upload the text file (.txt format) selected for your team, and copy the data source file link. Provide **SCREENSHOTS** of the following:

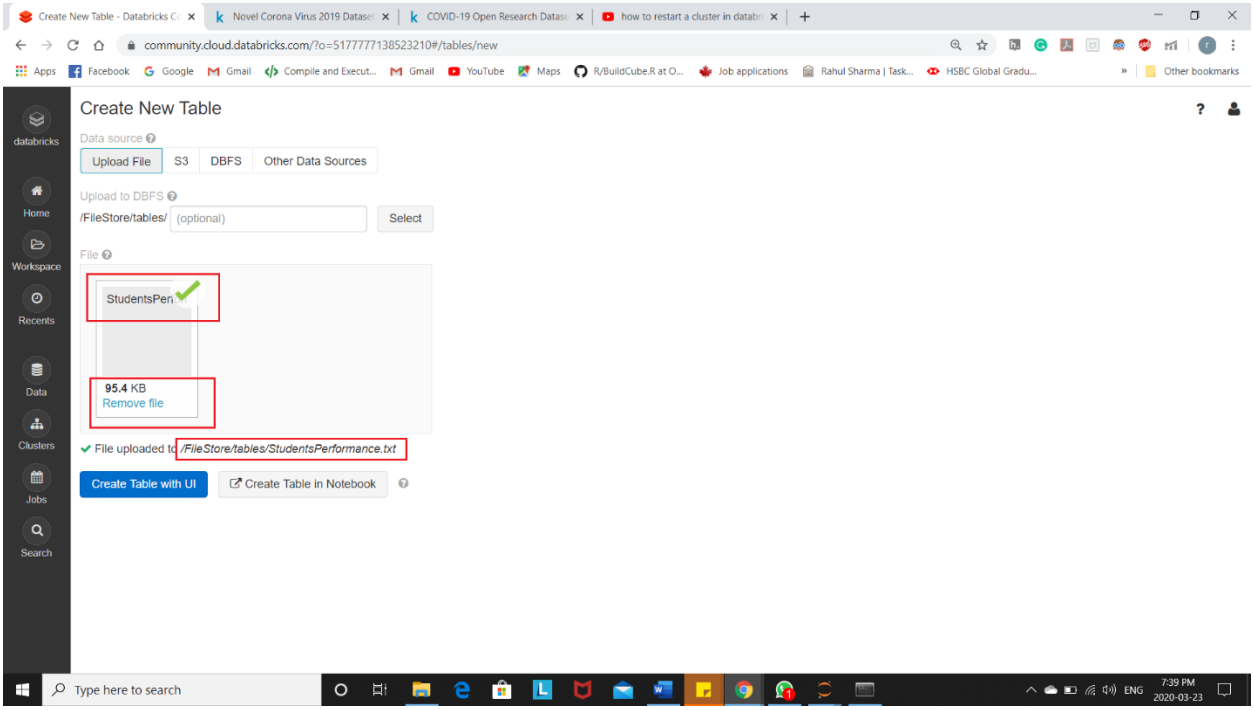
a) Your new Scala notebook – Below screenshot shows the creation on new notebook elements which we have set like name of the notebook, language which we use and the cluster name associated with the notebook.



Below screenshot shows the notebook is ready for the coding

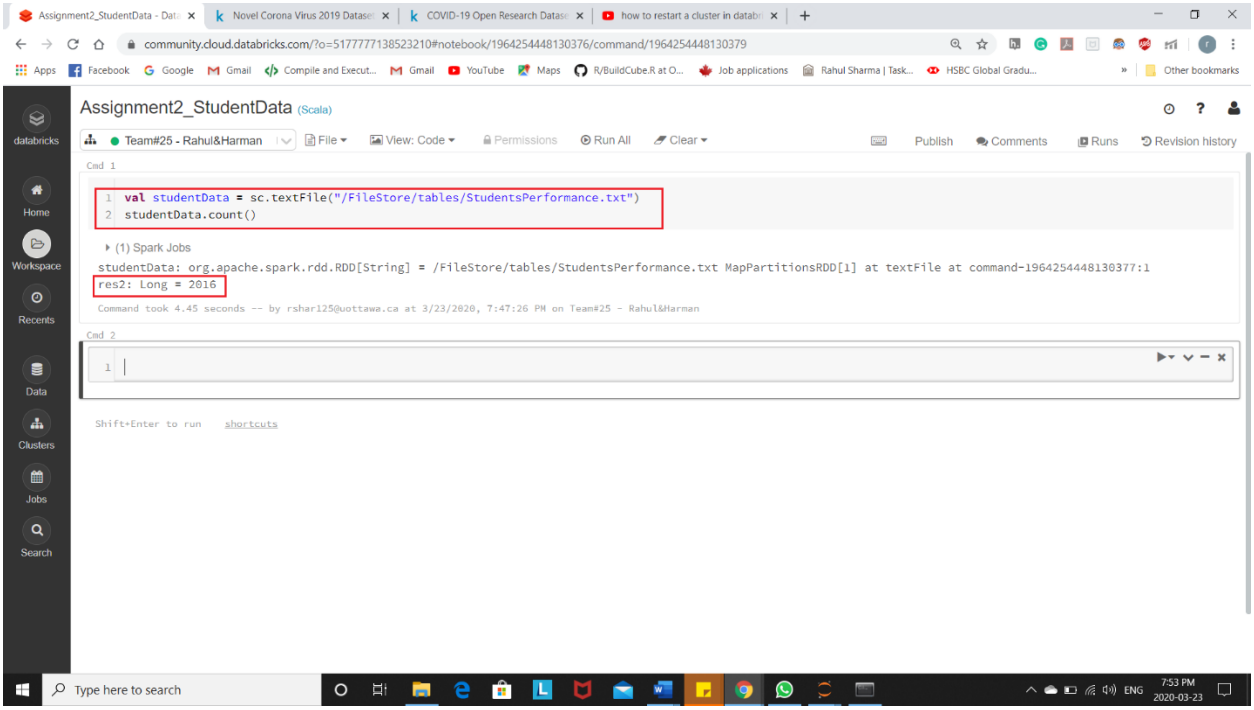


b) Uploaded data file in databricks showing the file name, size, and the databricks file link – Below screenshot shows the file is successfully uploaded along with the name of the file, size of the file and the link of the file where the file is uploaded

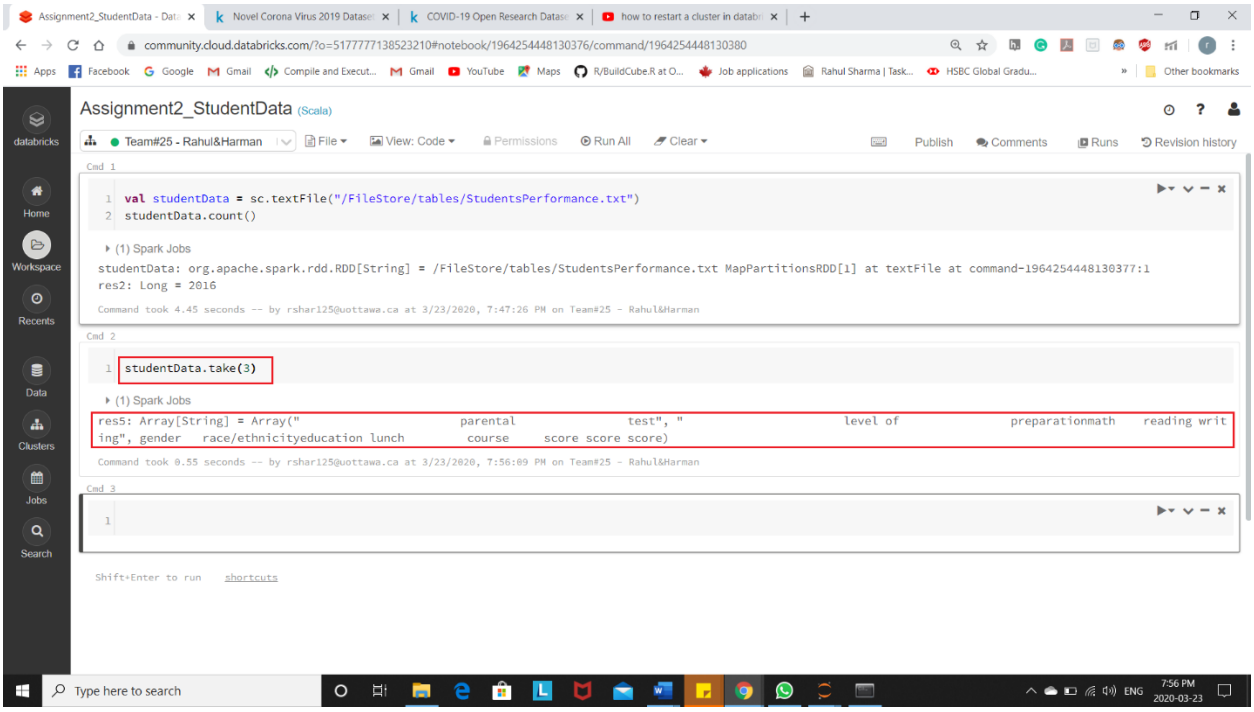


For questions (c) – (f), please provide a **SCREENSHOTS** of the code with the result from your Scala notebook:

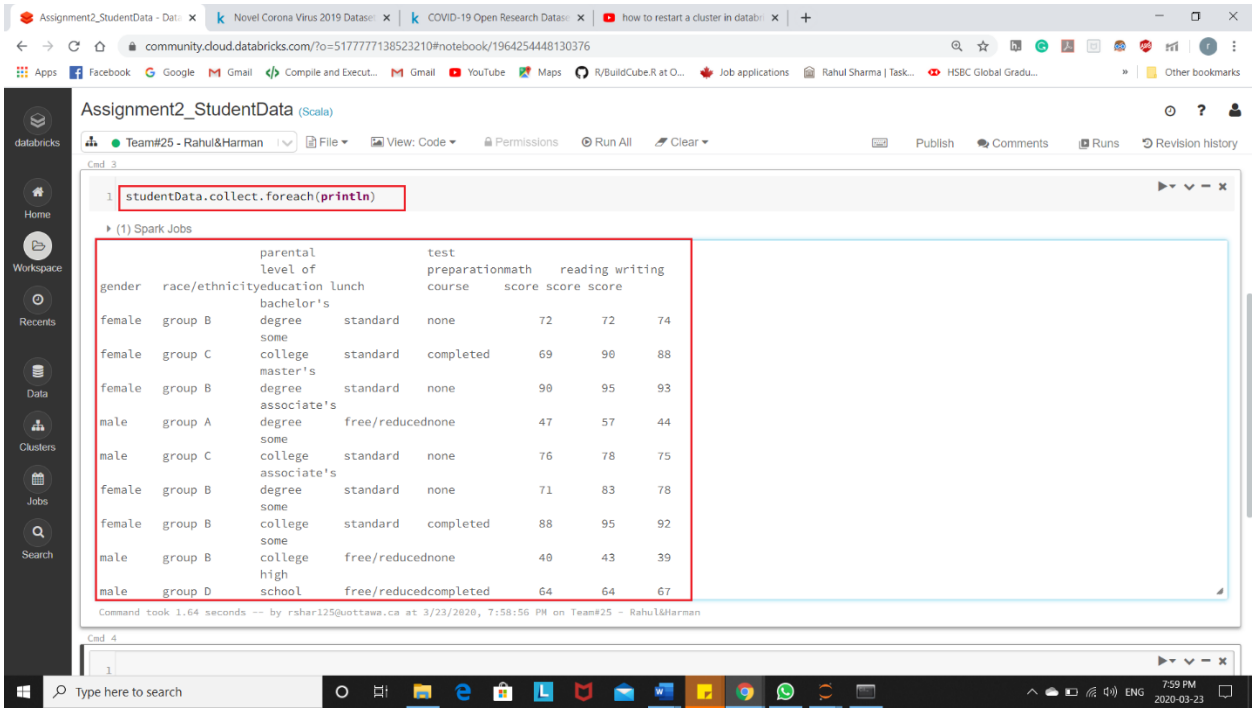
c) Count the number of lines of text in your data file – Below screenshot shows the file is successfully uploaded and the no. of lines is equal to 2016



d) Retrieve the first 3 rows of the data file – Beloq screenshot shows the first 3 line of the file

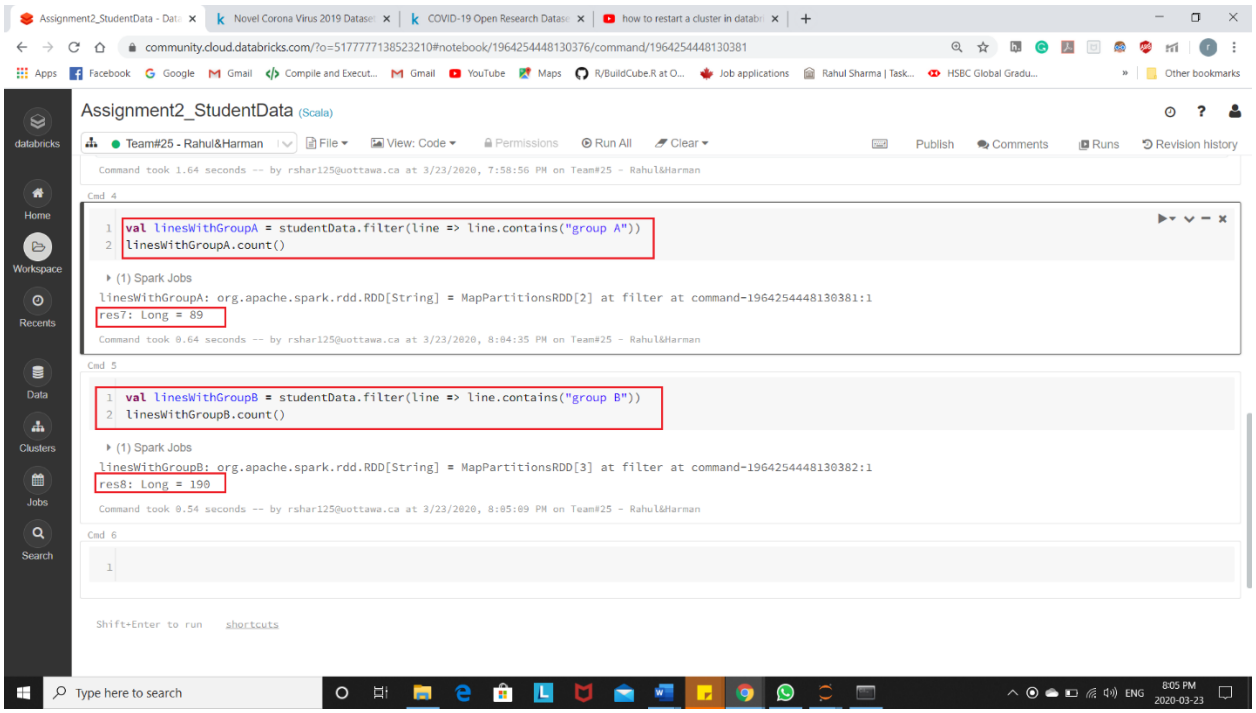


e) Print of all the lines in your data file (please provide a screenshot of only the lines that are visible on screen at one time) – Below screenshot shows the data of the file



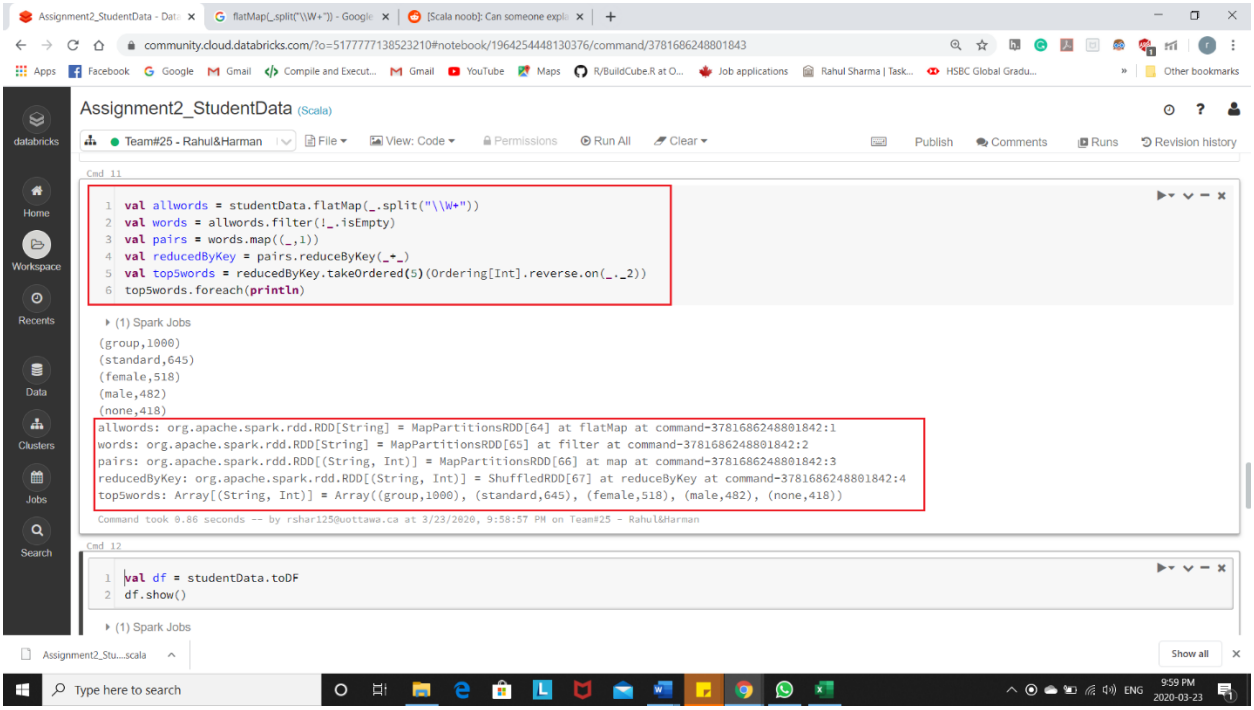
f) Pick any two words that exist in your data file, filter the data file for lines with these words, and count the total occurrence of the words.

Below screenshot shows the total no. of occurences for the word ”group A” which is 89 highlighted first and second highlighted shows the no. of occurences for word “group B” which is 190. We have choosen these 2 words to find out the category in data for these 2 groups



Question 7 [6 pts]: Use any 4 Scala transformation functions in a pipeline format to extract some analytic information from your data file (from question 7). Please provide:

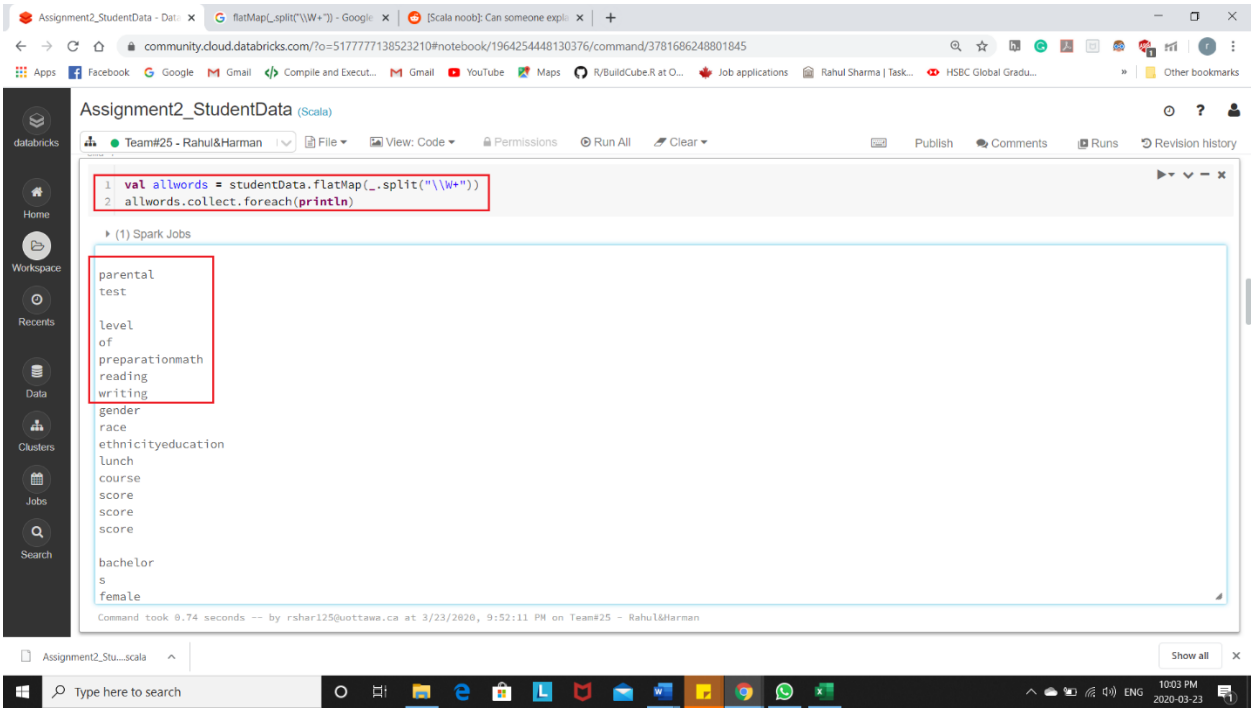
a) A screenshot of the notebook with code showing the 4 transformation functions in a pipeline – Below screenshot shows the 4 transformation which we have used like flatMap(), filter(), map(), pairs()



b) Provide a brief explanation of the transformations

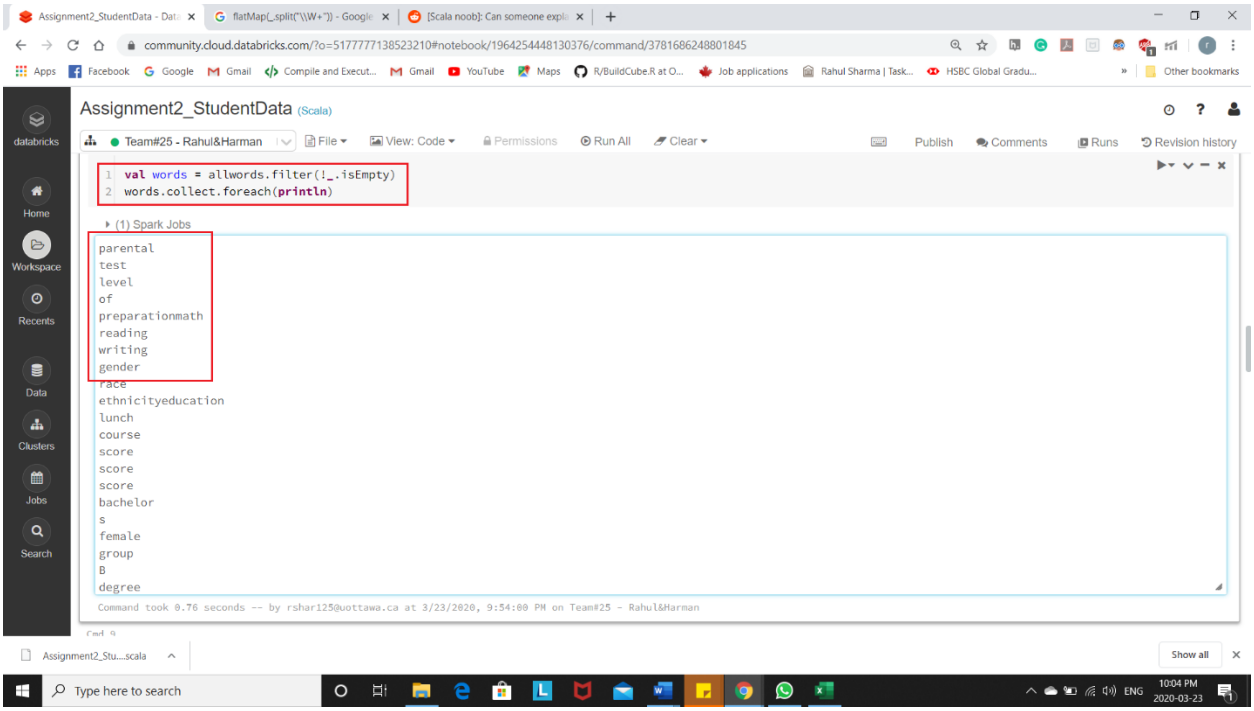
Transformation 1

First transformation is done on flatMap() which will divide the whole file into single words and it will easy for us to read all the data by words



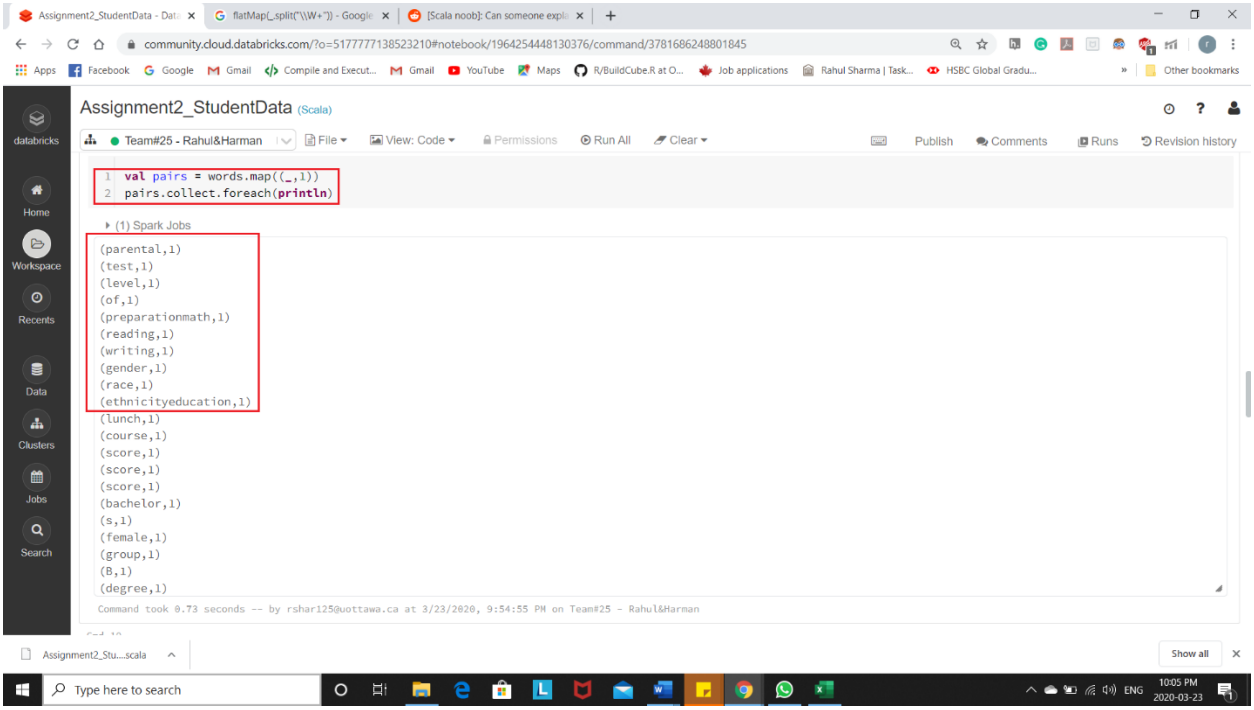
Transformation 2

We have used filter() in second transformation in order to filter out the empty lines from the file



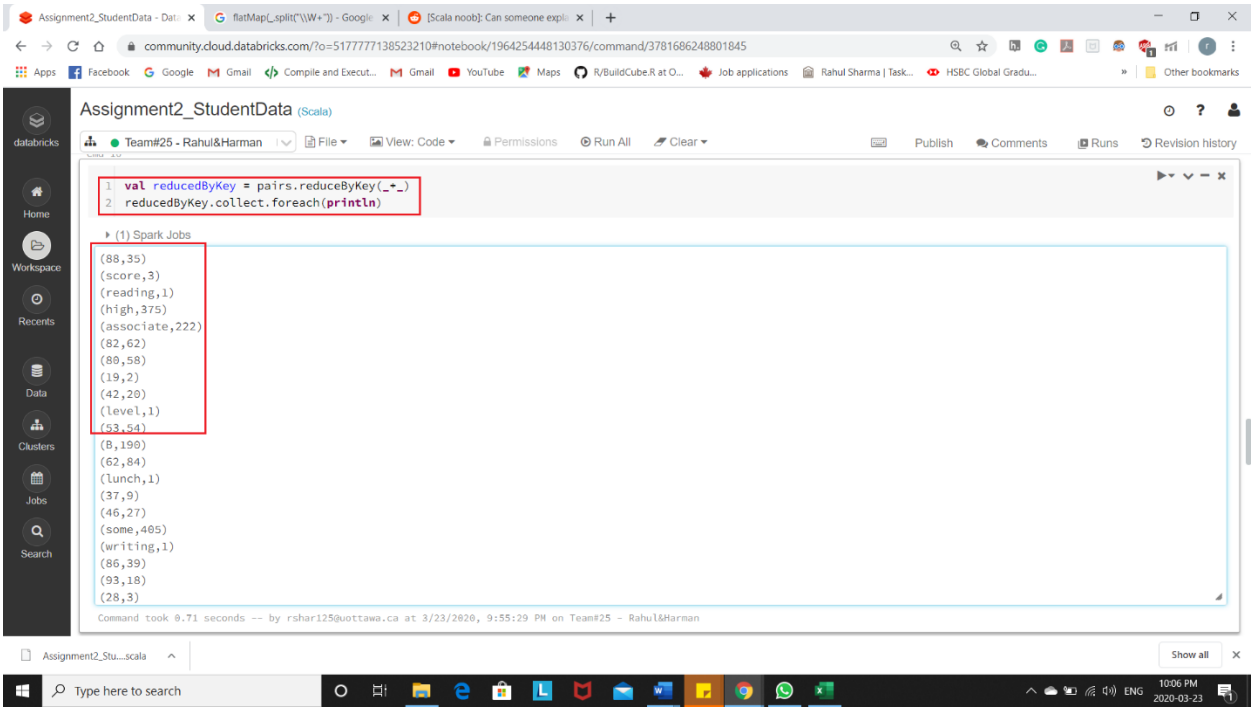
Transformation 3

In third transformation we have used map in order to pair all the word in key value pair like for each occurrence just write it as word and no. of occurrence as 1



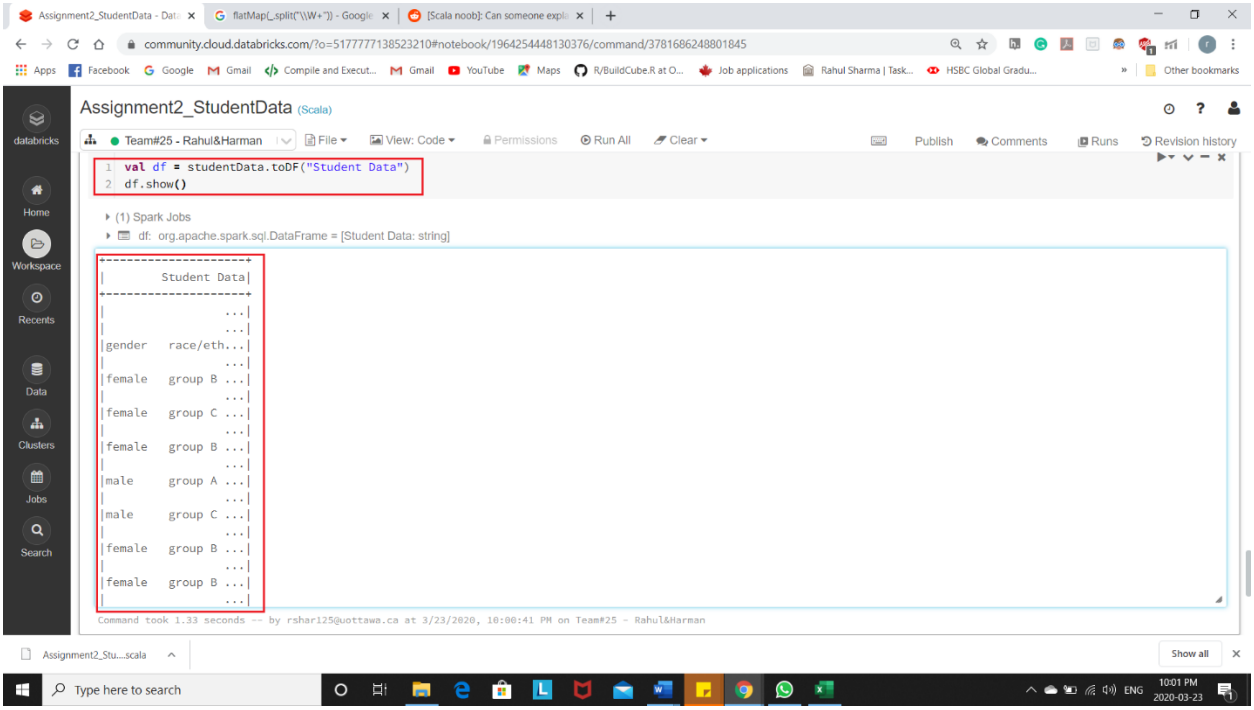
Transformation 4

In transformation 4 we have used pair() in order to pair the same words and to find out the total no. of occurrences for each word



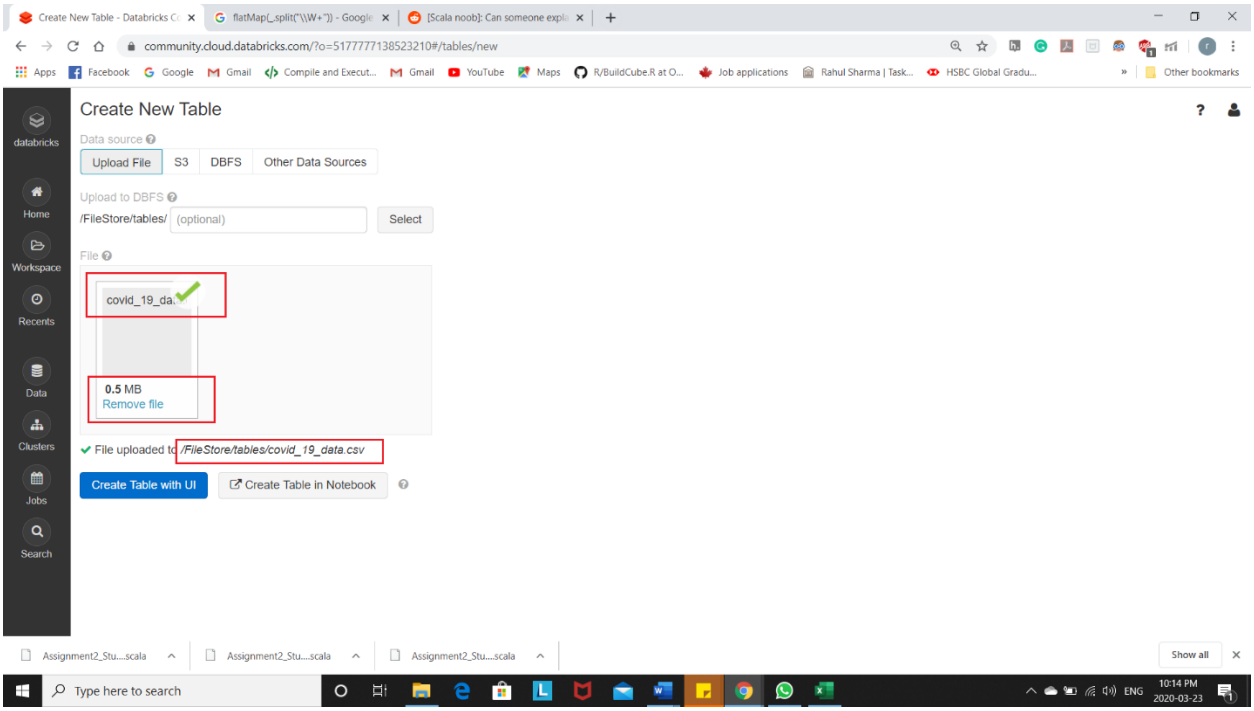
c) A screenshot of the transformation data frame result

Below screenshot shows the transformed data after all the above 4 transformation is successfull

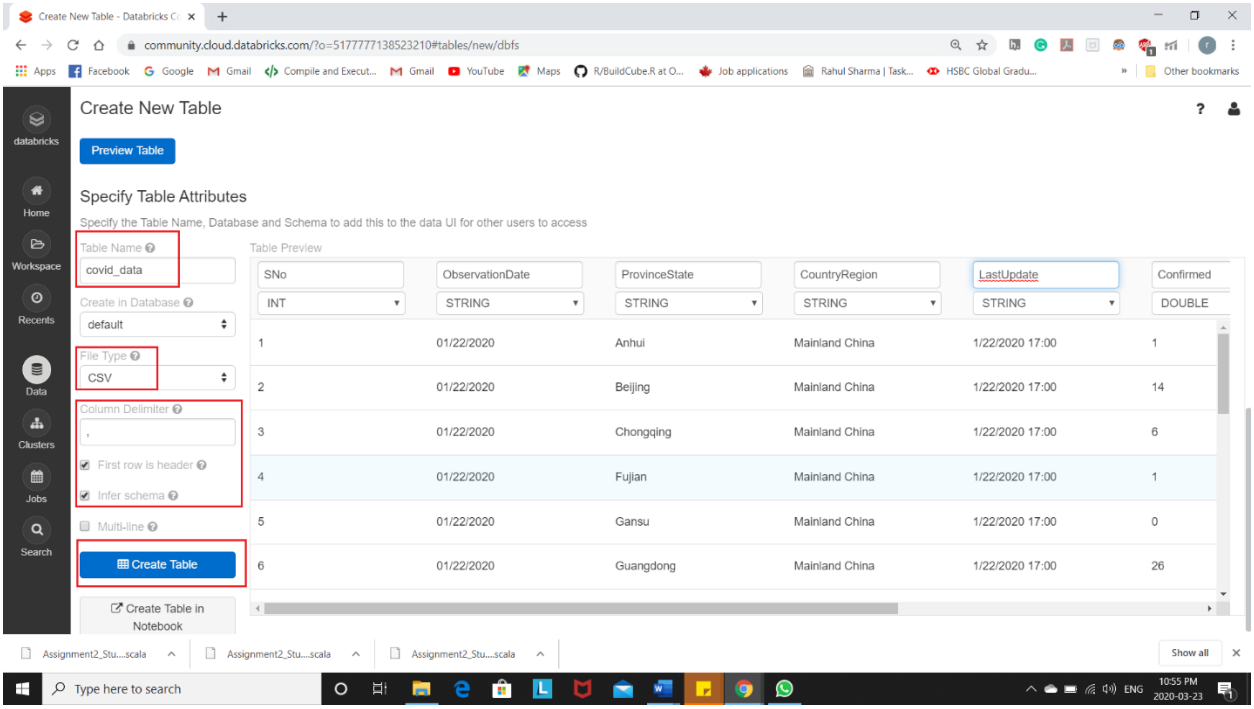


Question 8 [6 pts]: Please upload the selected csv file for your team into databricks and create a table. Create a new python notebook and display the data frame showing the first 30 rows using the Spark sql context. Using the same notebook, query your data for any 3 columns (including at least one summary data) from your data file and display the result data frame showing those columns. Convert the result into a dashboard of your choice. Please provide snapshots of the following activities:

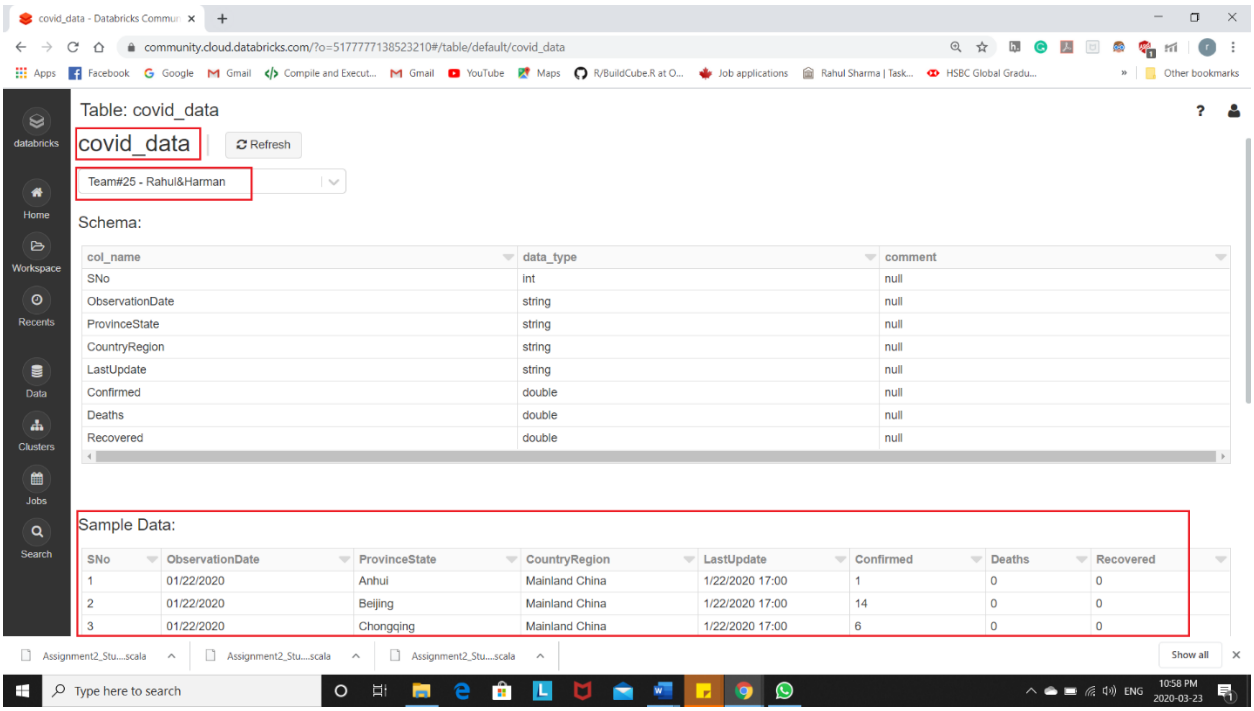
a) Data import confirmation page – Below screenshot shows the file is successfully uploaded in the cluster. Highlighted are the file name, size and link of the file



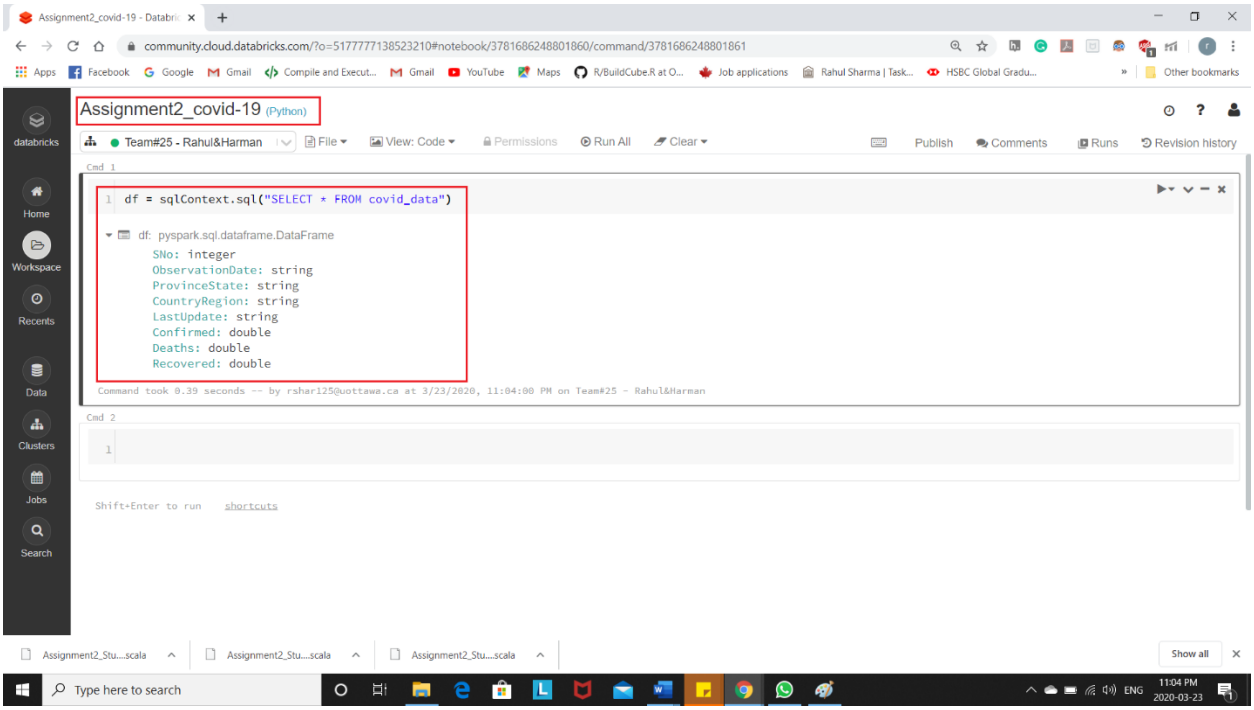
b) Created table from imported data – Below screenshot shows the formation of the table with the uploaded data. Highlighted are the table name, file type and column deliminators



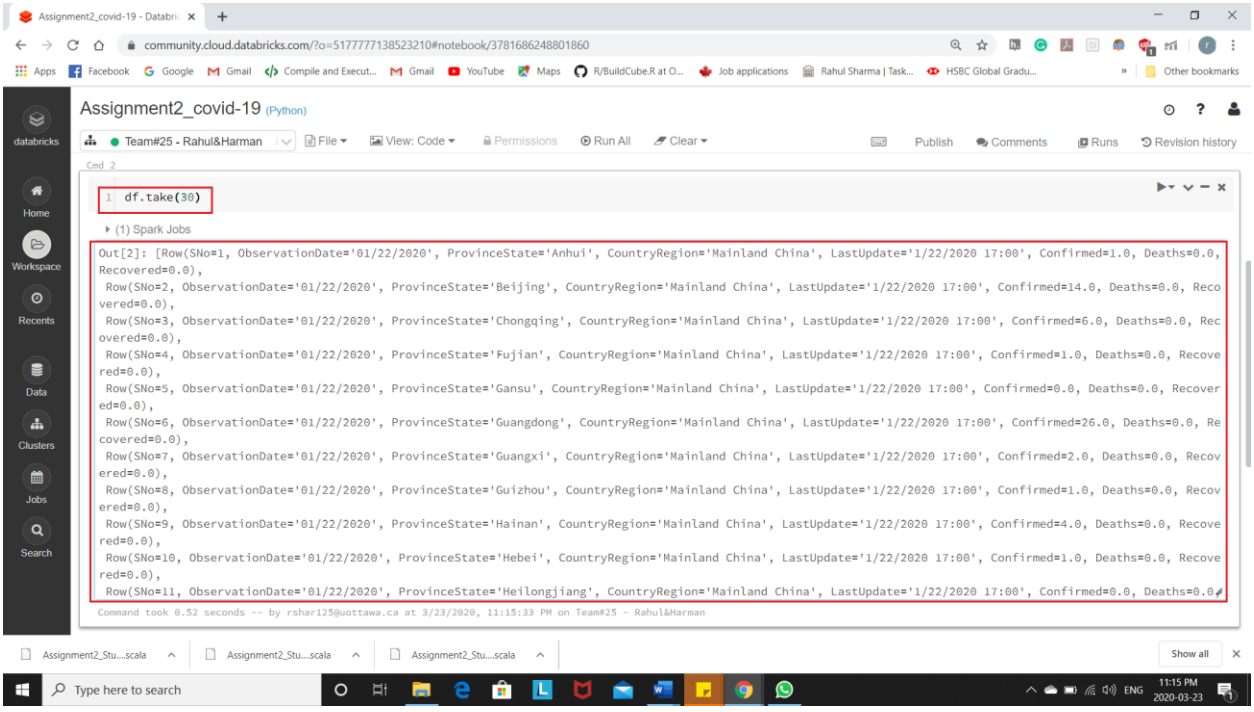
Below screenshot shows the table is created successfully



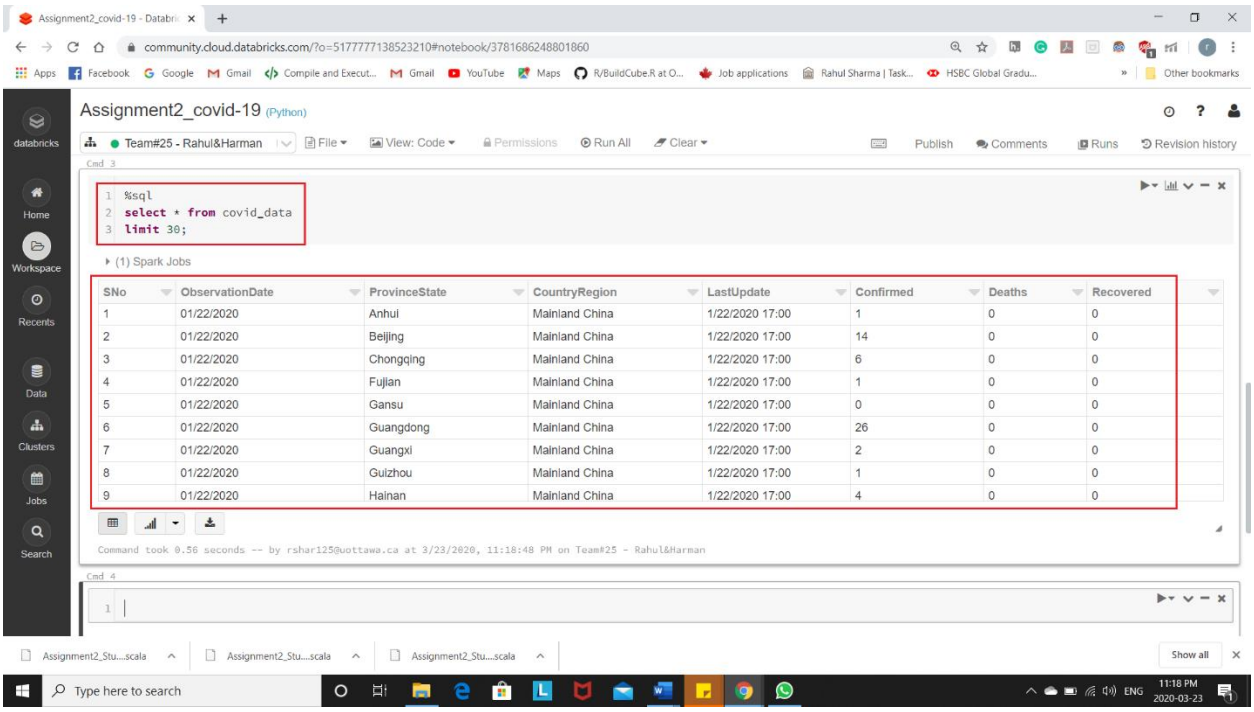
c) Your python notebook with code to create a data frame using the Spark sql context – Below screenshot shows the python notebook is created successfully and the data frame is successfully created with the data of the table



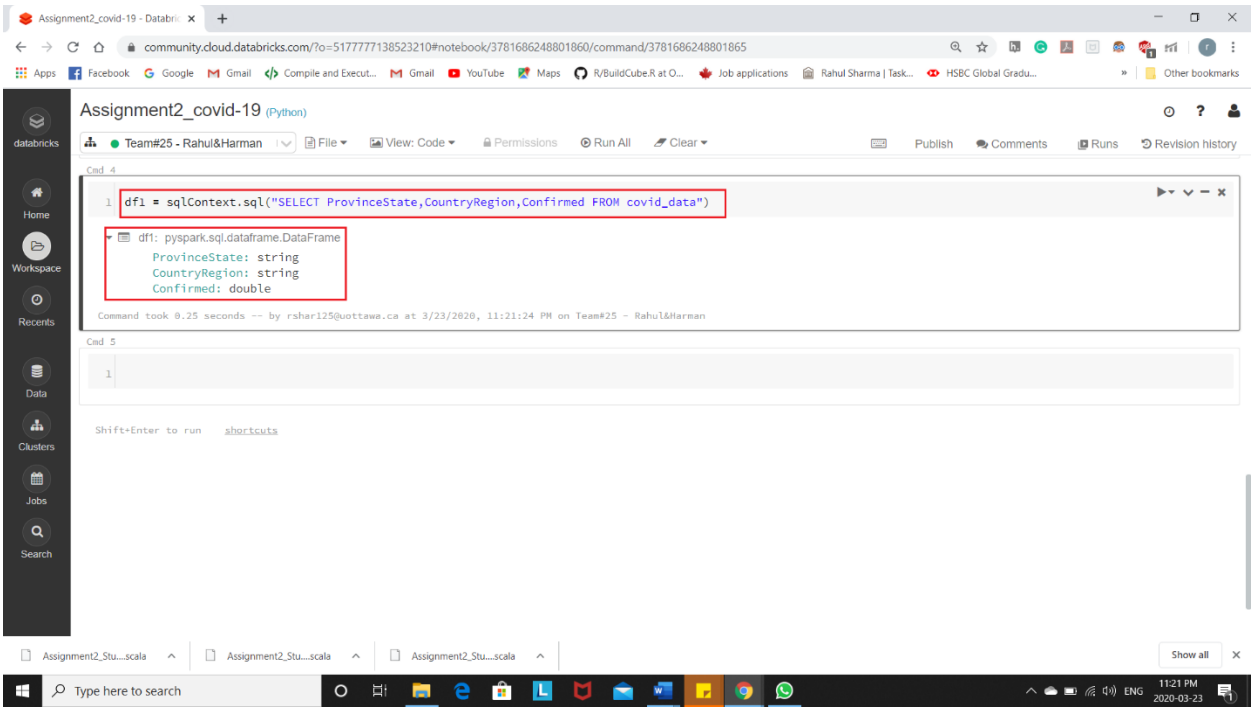
d) Data frame showing the first 30 rows of your data file – Below screenshot shows the first 30 rows of the data which we have uploaded as a table



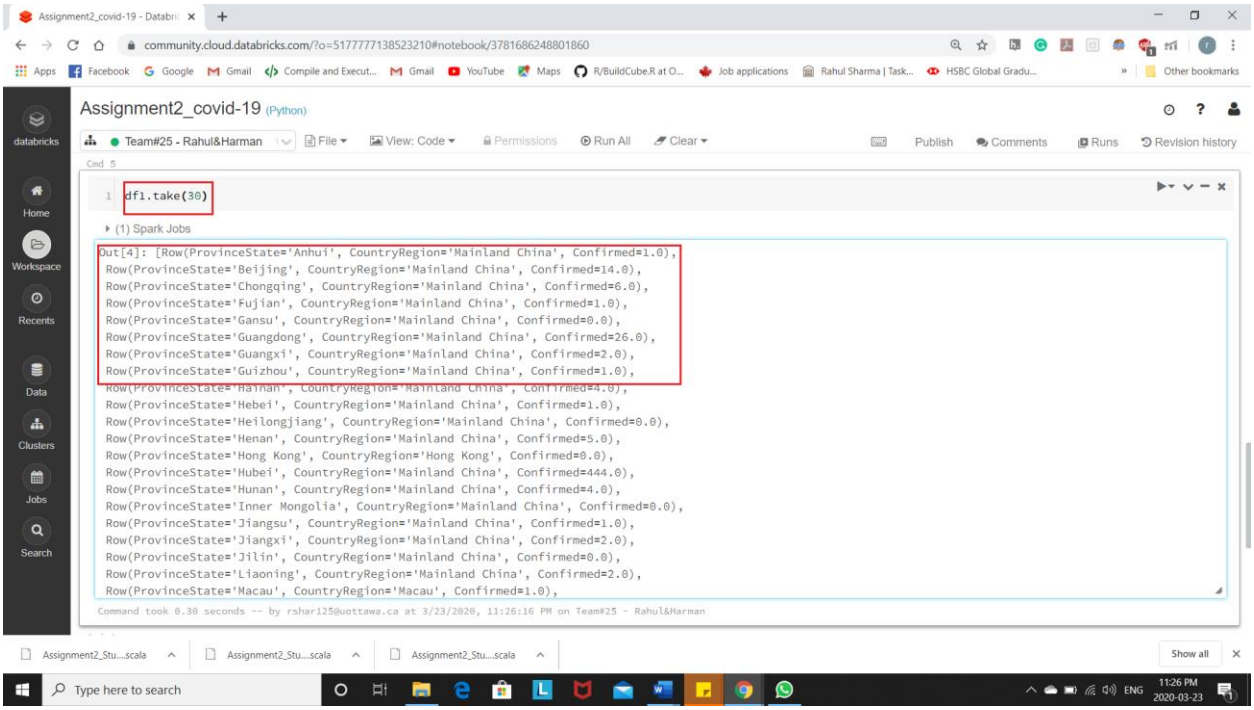
Below screenshot shows the same 30 rows which we have retrieved using sql commands



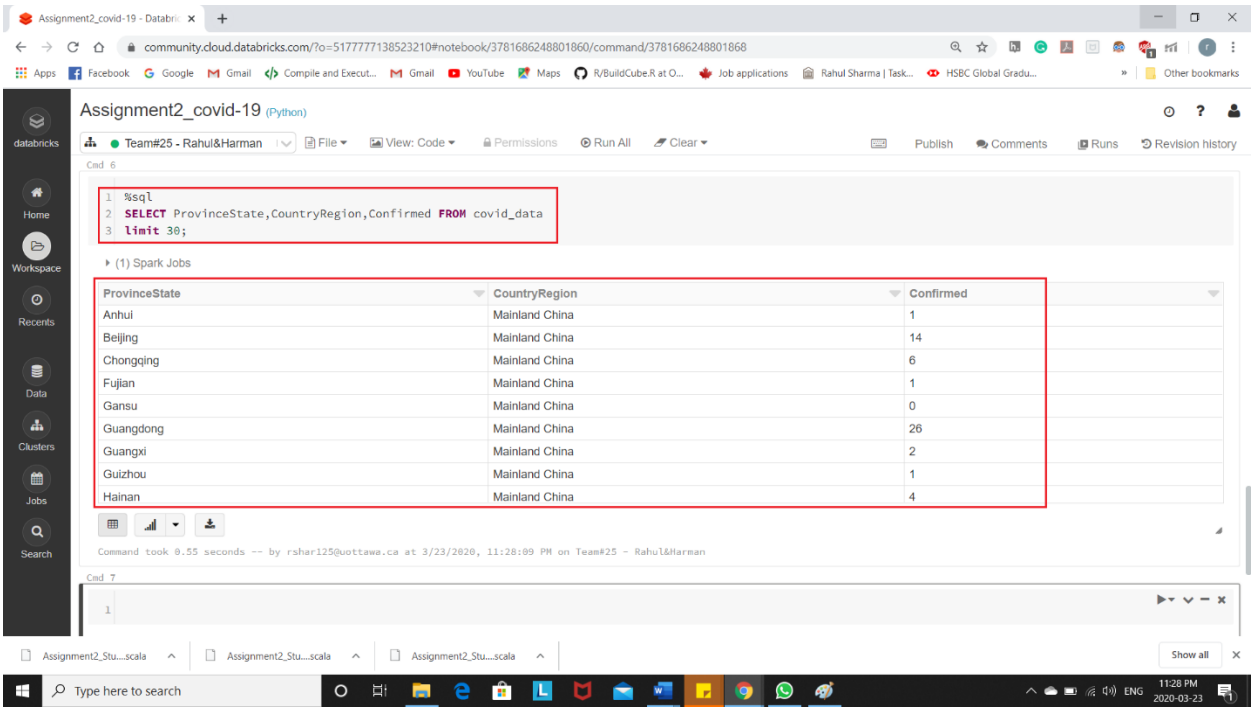
e) Code to create a data frame showing the 3 columns from your data file – Below is the screenshot shows the 3 columns selected to do the analysis of the data



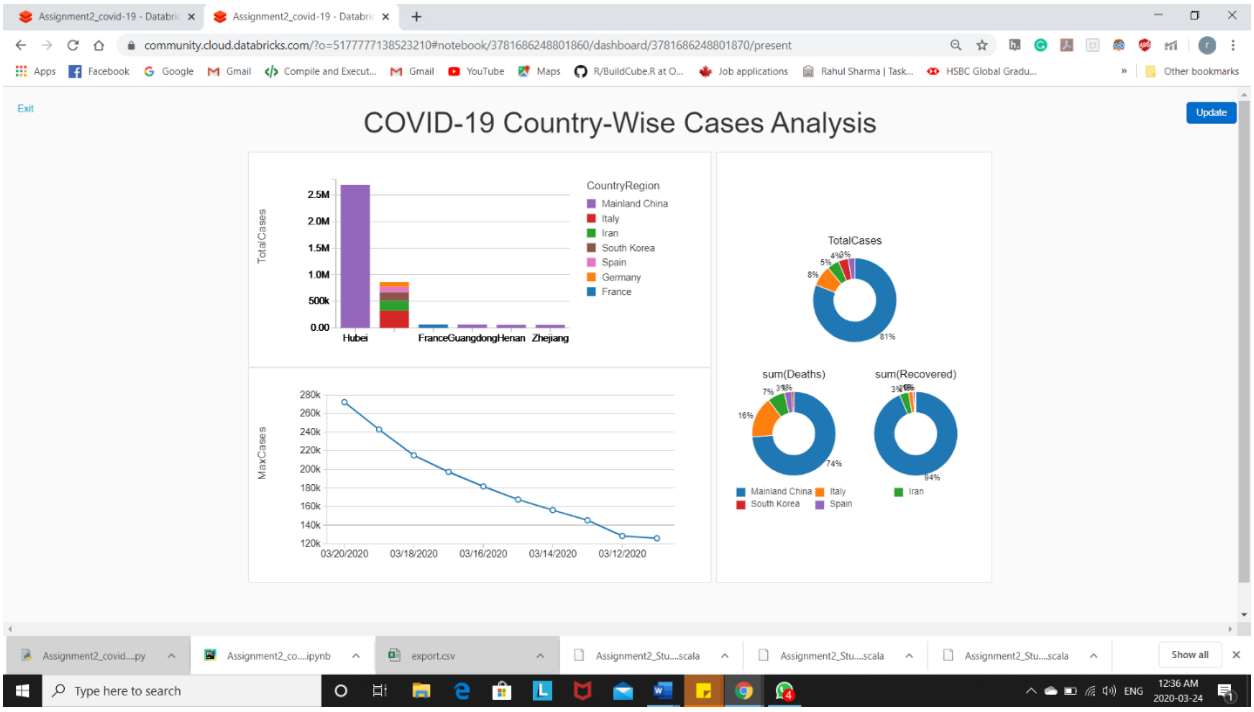
Below screenshot shows the data associated with the 3 columns which we have selected for the analysis using data frame



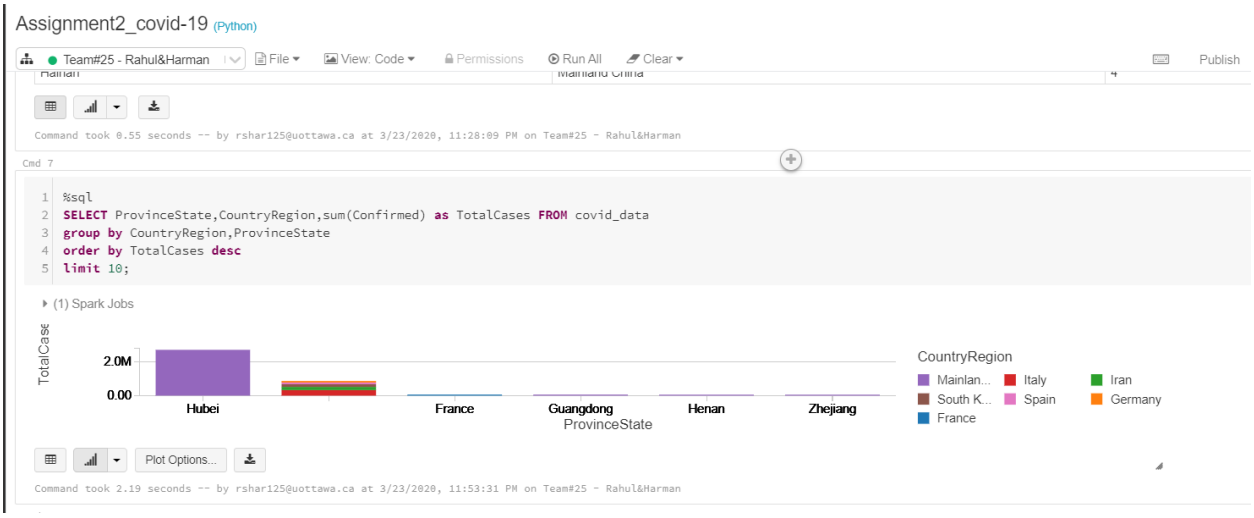
Below is the screenshot of the data associated with the 3 columns using sql commands



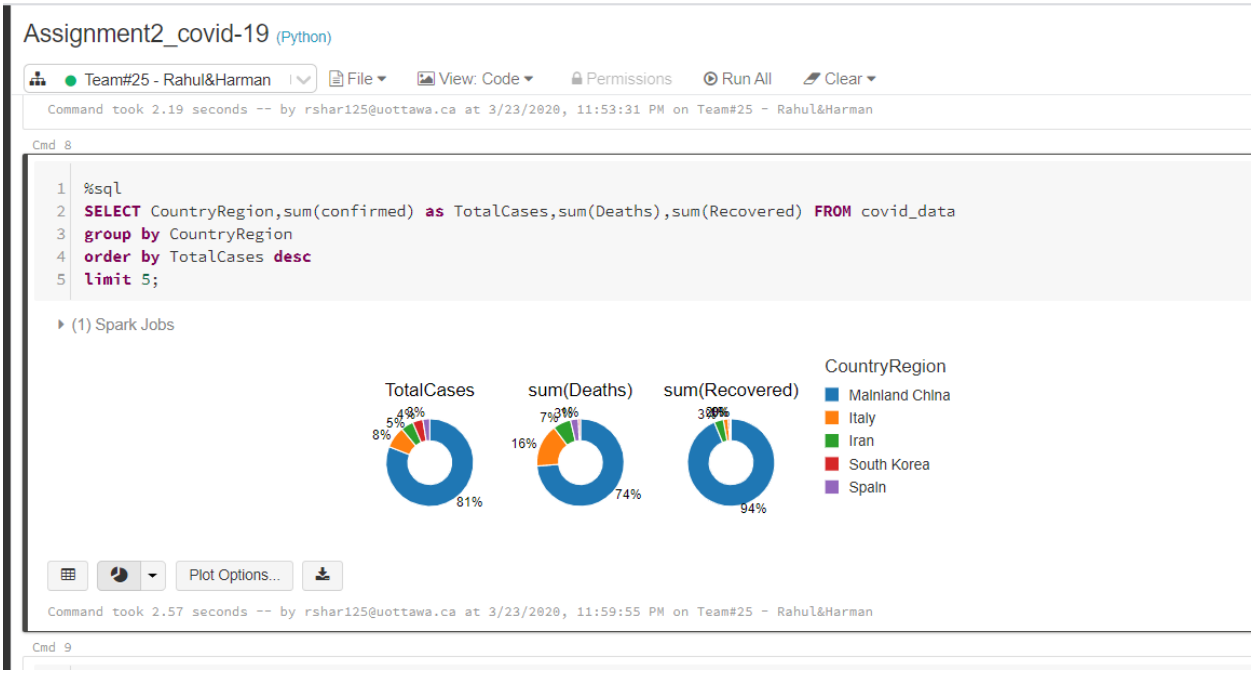
f) Screenshot of a dashboard using the results from (e) – Below is the screenshot of the dashboard created with the 3 graphs



Below screenshot shows the first graph of the dashboard which basically shows the top-10 country and province which has total maximum cases of the covid-19



Below screenshot shows the top-5 countries with total no. of cases along with the cases which they have recovered and the total death of there particular region due to covid-19



Below screenshot shows the affect of covid -19 daya by day