Aakash Sarap

Northeastern University **Assignment 02**

ALY6110 Data Management and Big Data

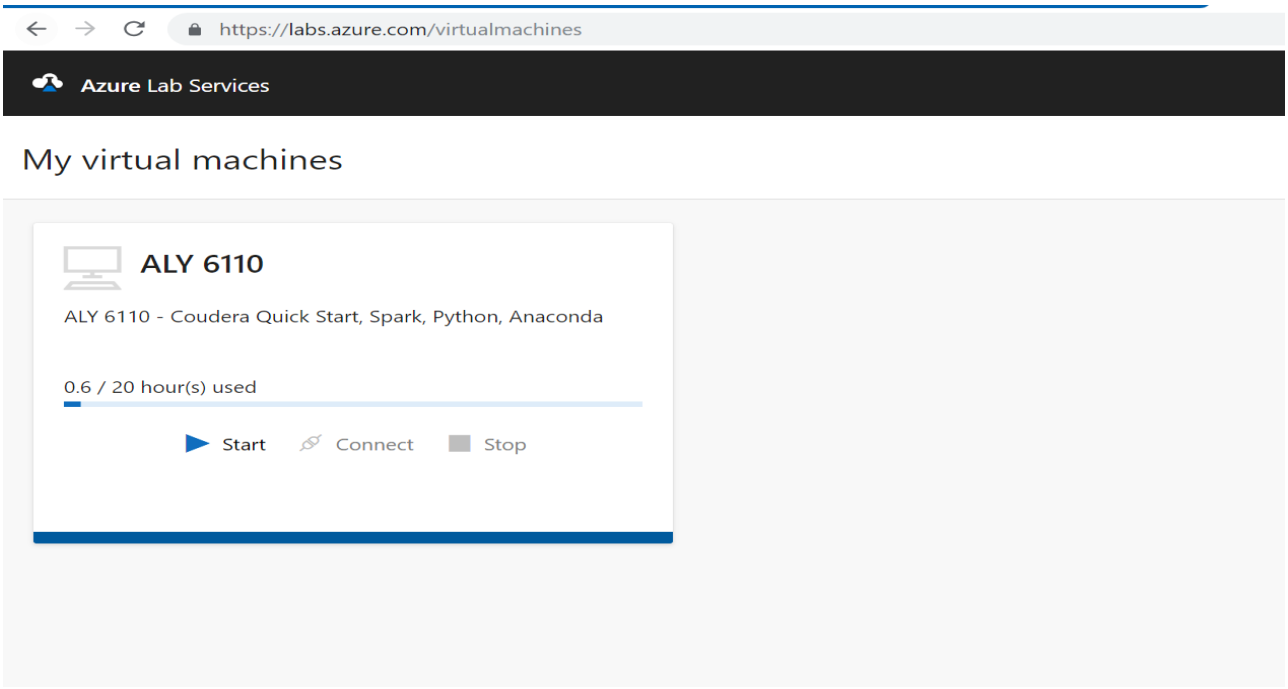
https://northeastern.blackboard.com

Problem 1:

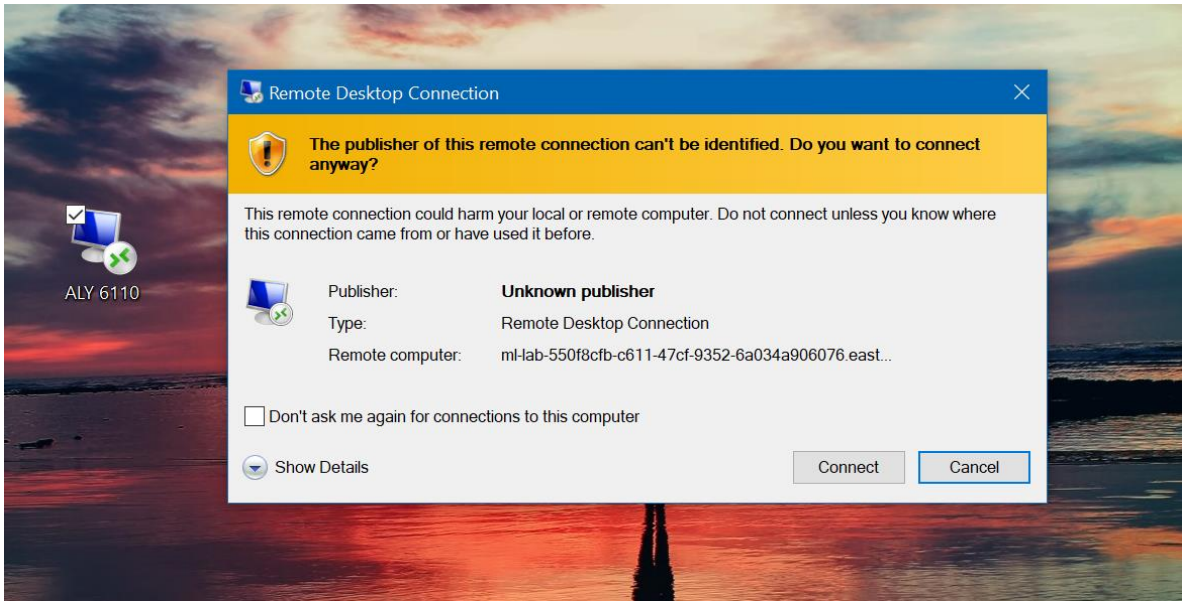
Start up the Microsoft Azure Lab – Virtual Machine and get all the Hadoop Services running.

Note*: I am using the Microsoft Azure Lab for Class ALY 6110 for running Cloudera and Hadoop Services which already has the Hyper-V Manager and the Cloudera installed in it.*

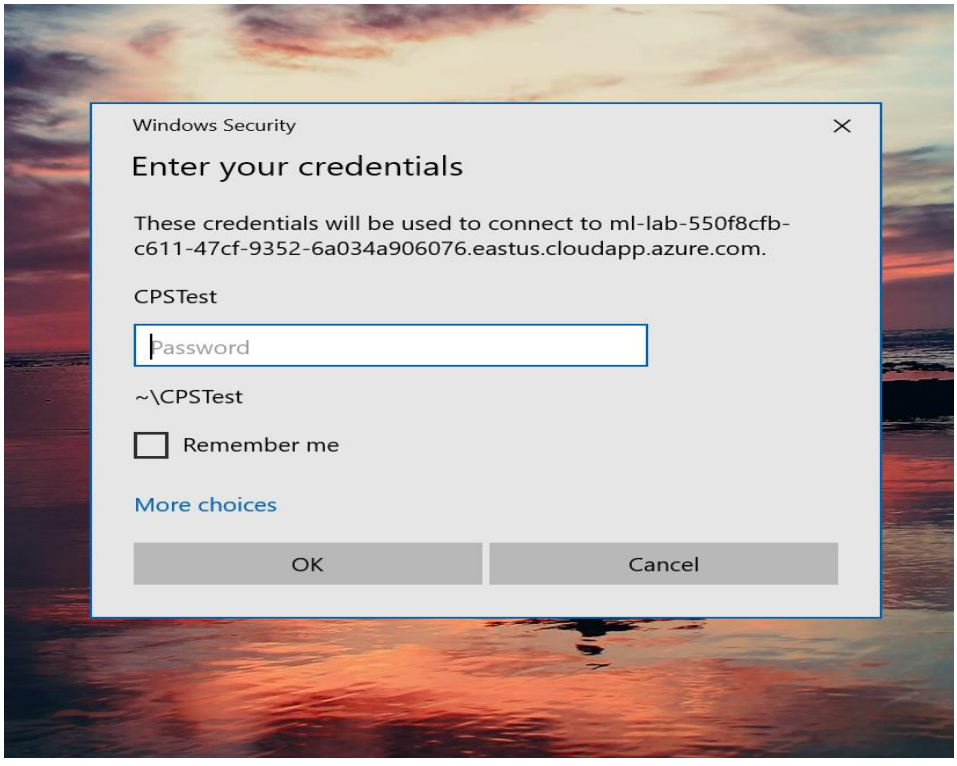
1. As I am using the Microsoft Azure Lab Machine ALY 6110 for running Cloudera and Hadoop Services, we first visit the URL https://labs.azure.com/virtualmachines.



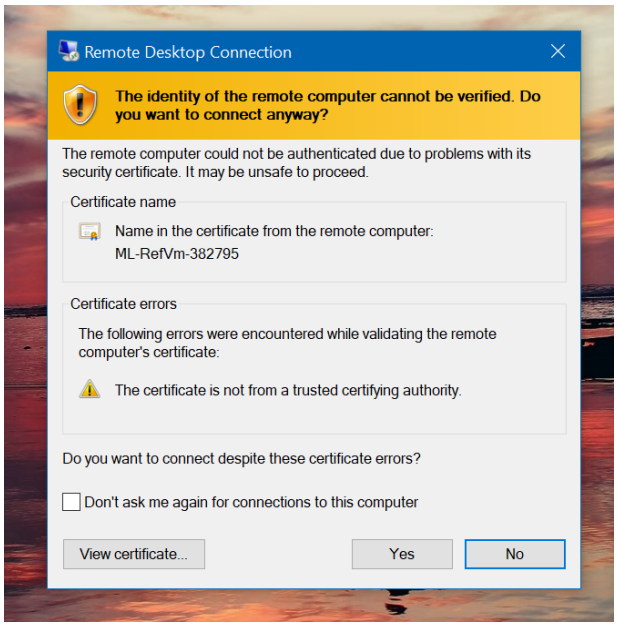
2. Now click on “Start” to start the virtual machine. After we click on Start, a remote desktop client will get installed. I have downloaded it on the Desktop. Now double click on the Remote Desktop Client named as “ALY 6110”. You will see the window as shown below. Click on “Connect”.



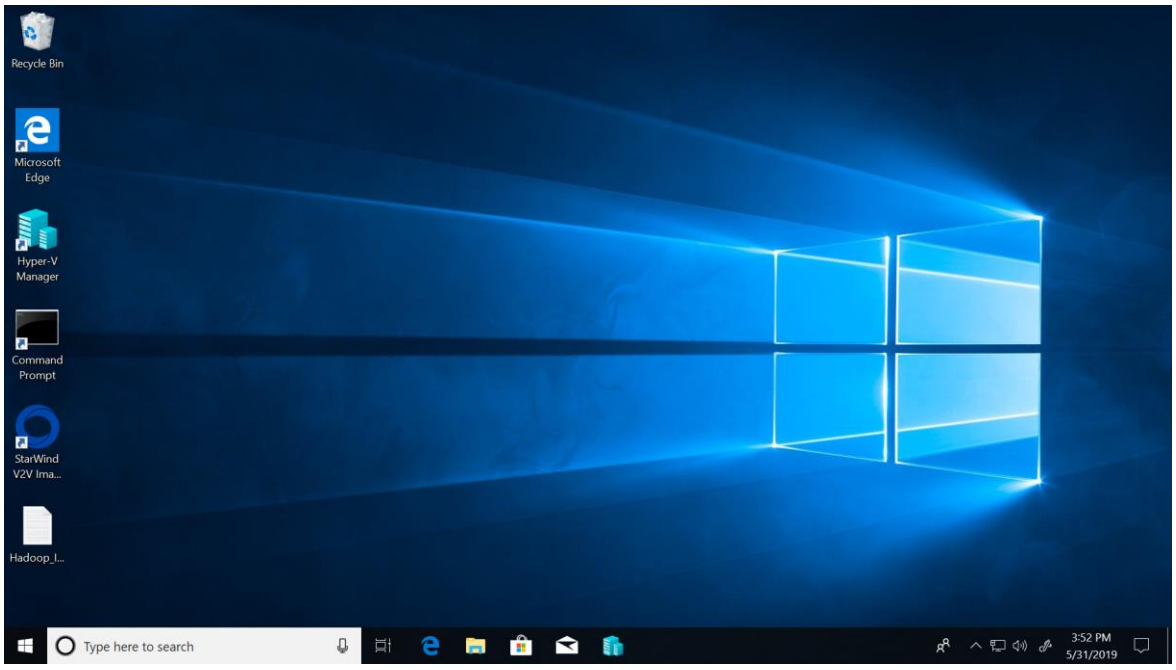
3. It will then ask for the password for starting the virtual machine. Enter the password and click "OK".



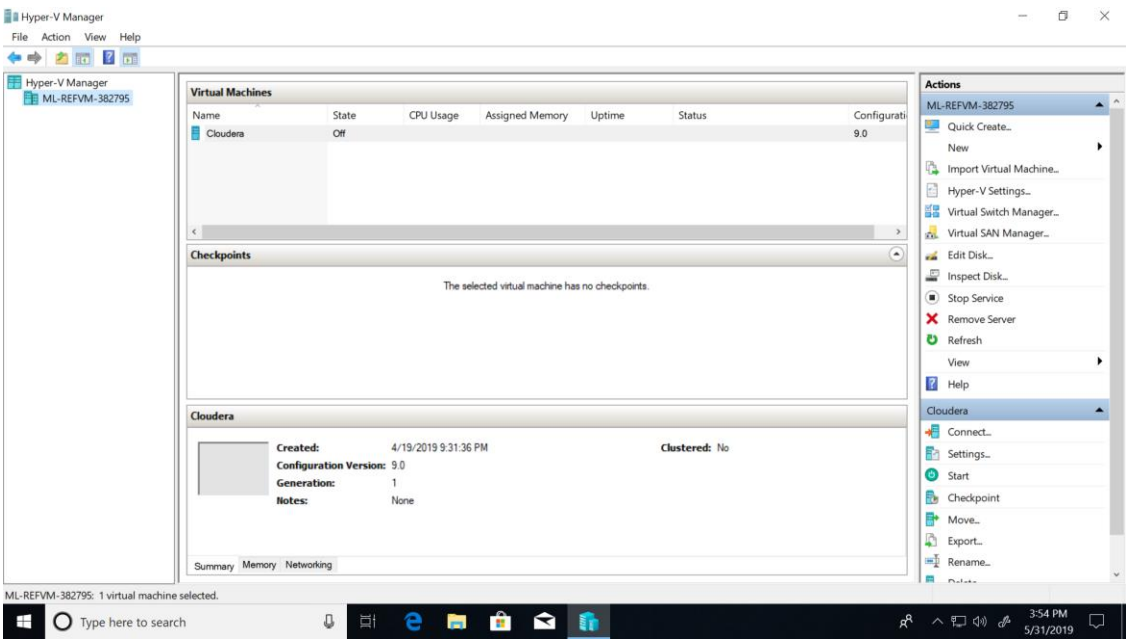
4. It will then give a prompt as shown below because of the certificate errors. But we will be connecting despite these certificate errors and click on “Yes”.



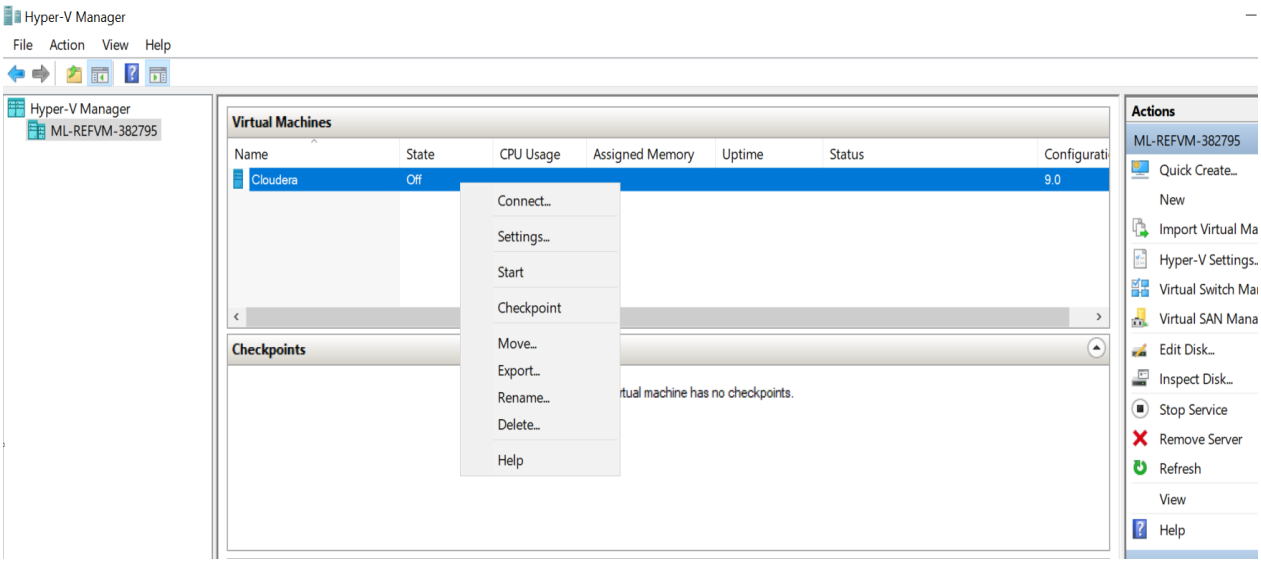
5. Our Virtual Machine will get started and looks like the one below in the screenshot.



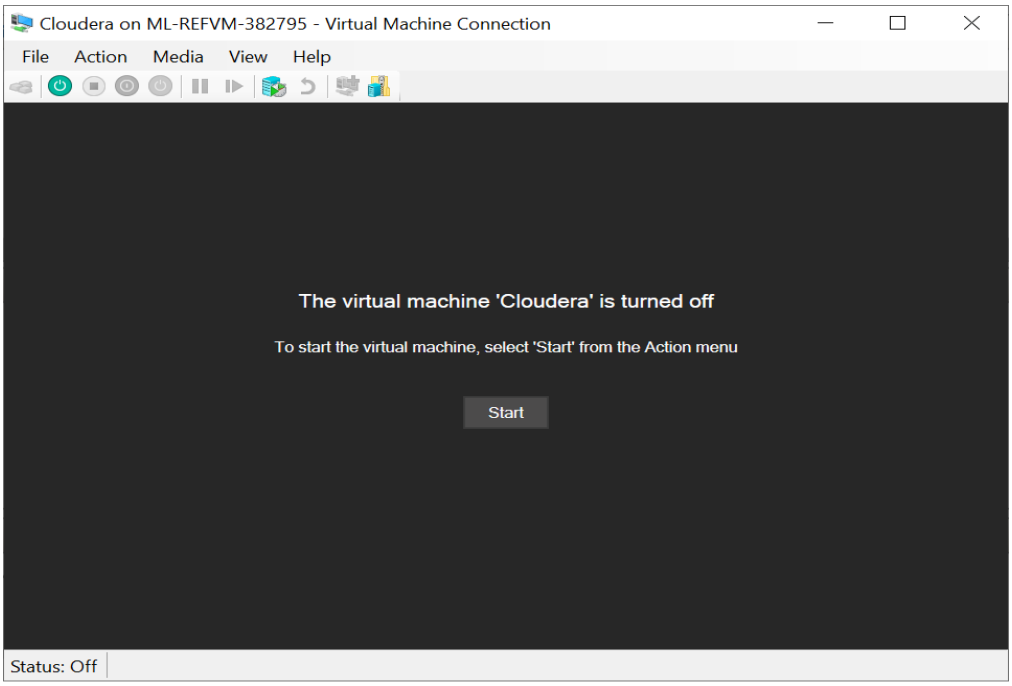
6. Open Hyper-V Manager present on the desktop. Refer the below screenshot. We can see that under Virtual Machines, status of Cloudera is "Off".



7. Right click on Cloudera and select "Connect" as shown below.



8. We can see that a new window with Cloudera VM comes up but it can be seen at the bottom left that the status is still Off.



9. Click on "Start". The Cloudera VM will now get started as shown below.

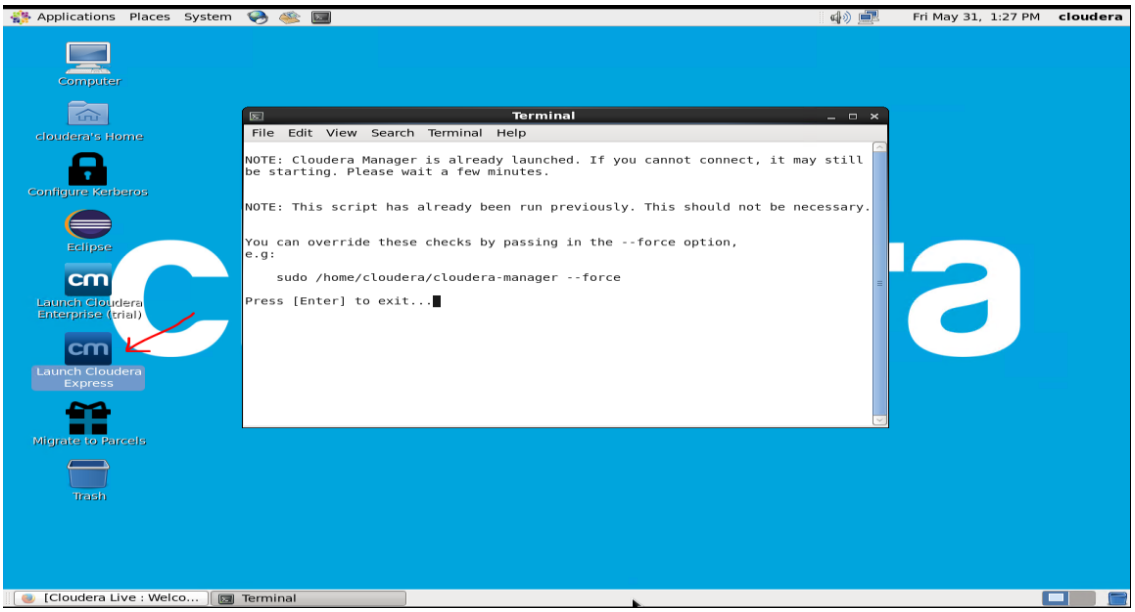
Aakash Sarap

Northeastern University Assignment 02

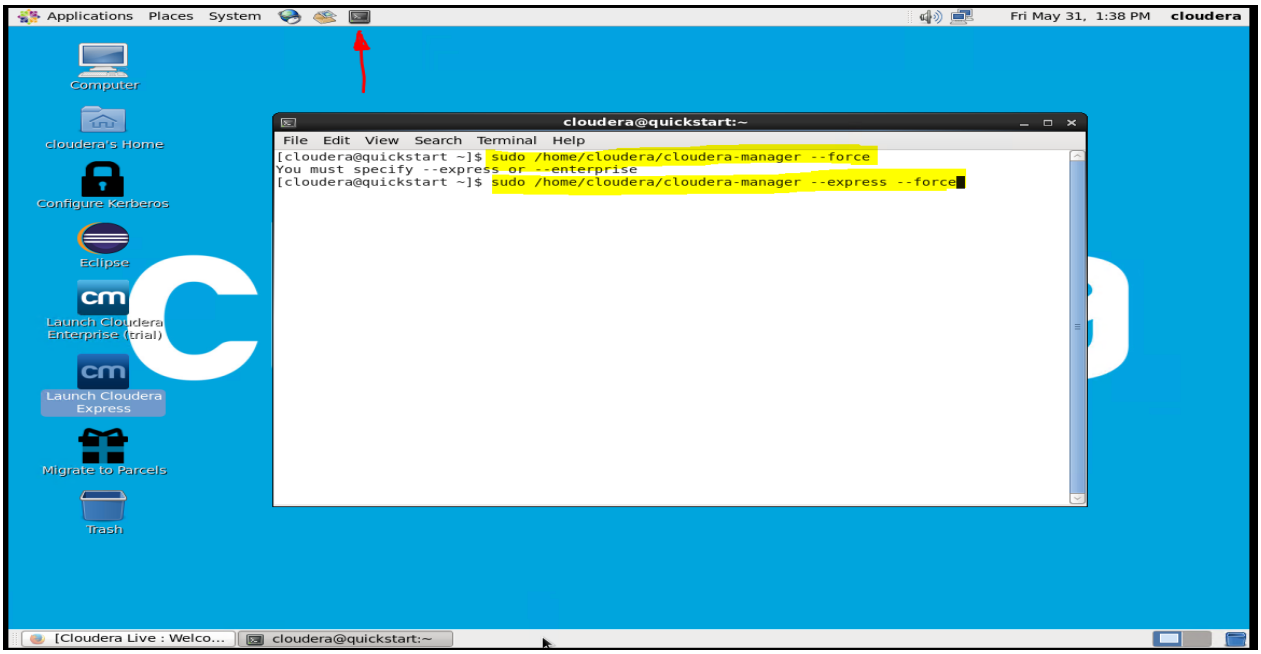
ALY6110 Data Management and Big Data

https://northeastern.blackboard.com

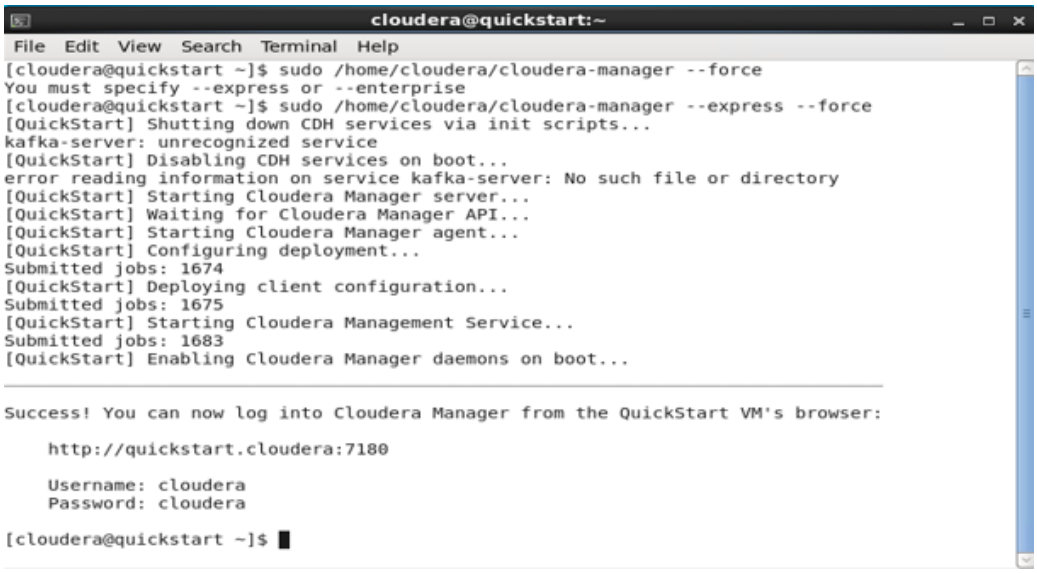
10. After the Cloudera gets started, open the "Launch Cloudera Express" as shown in the below screenshot. A new terminal window gets started.



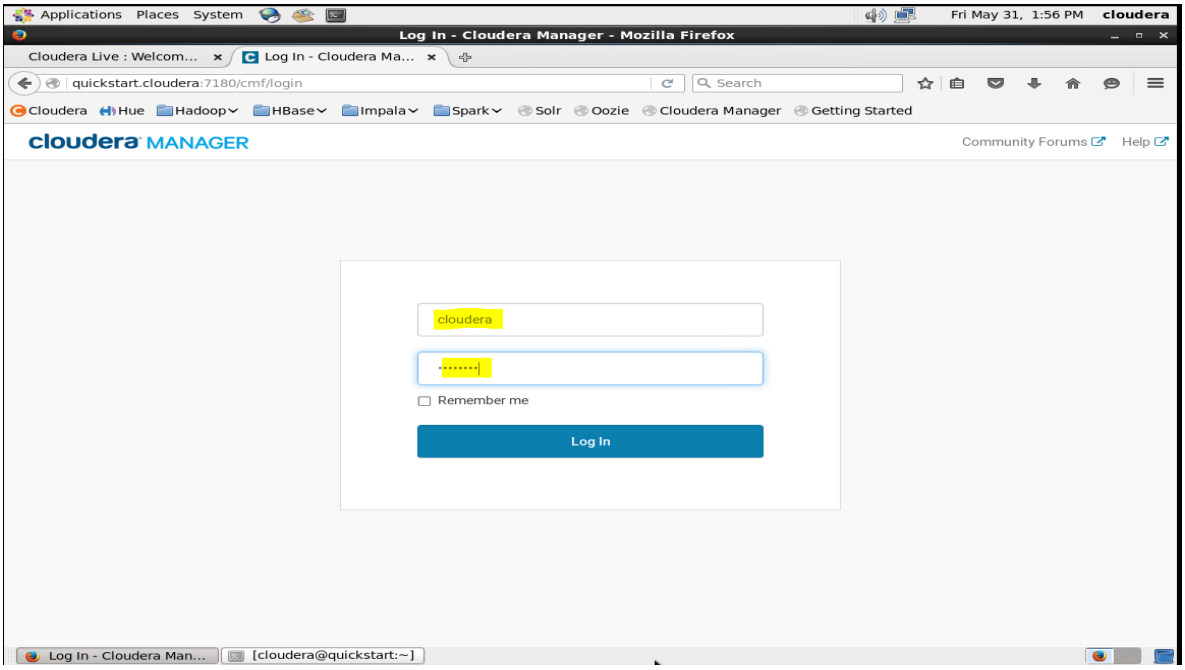
11. Copy the path *"**sudo /home/cloudera/cloudera-manager - - force**"* from the above terminal and paste it in the new terminal. Now we must specify *"**- - express**"* and hit enter. Refer the below screenshot.



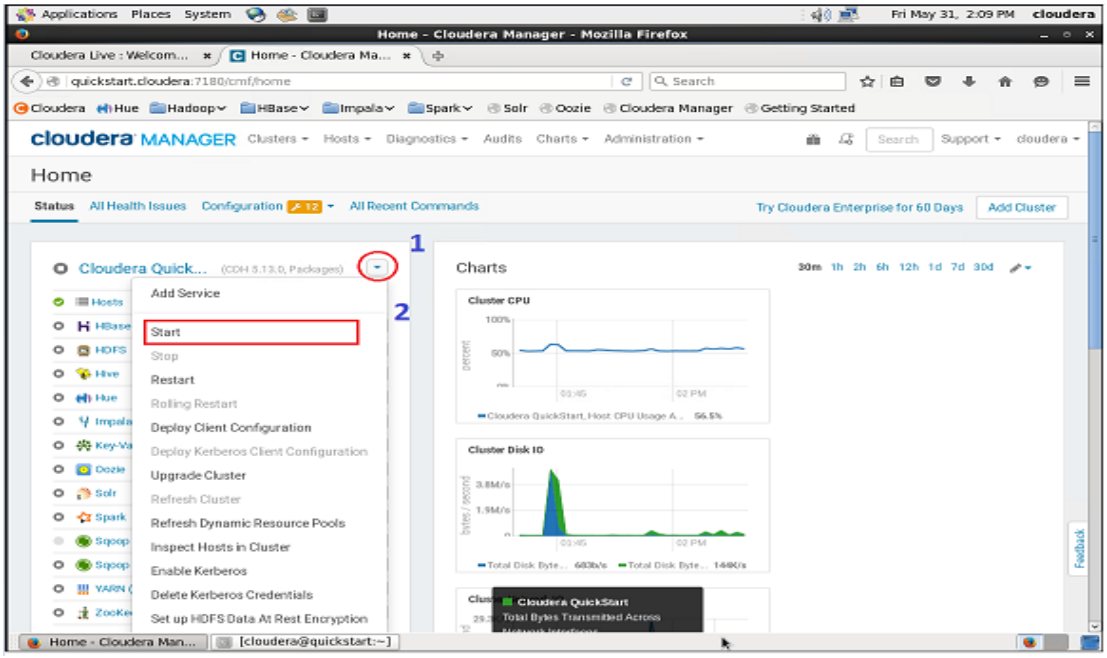
12. The terminal will now provide us with the username and password for the Cloudera Manager as shown below.



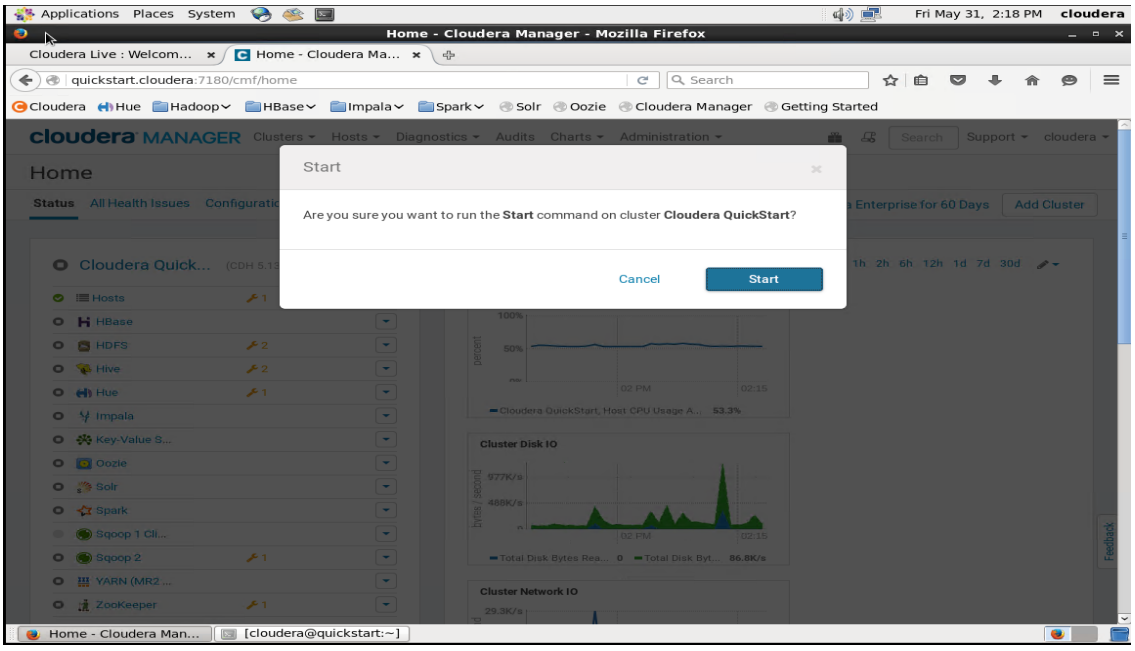
13. Open the browser in the Cloudera and go to Cloudera Manager. Use the credentials provided by the terminal to login to Cloudera Manager.



14. Go to the down arrow besides “Cloudera Quickstart” and click on “Start”.



15. We can then again see a start prompt. Click on “Start” as shown below. It makes all the services to run.



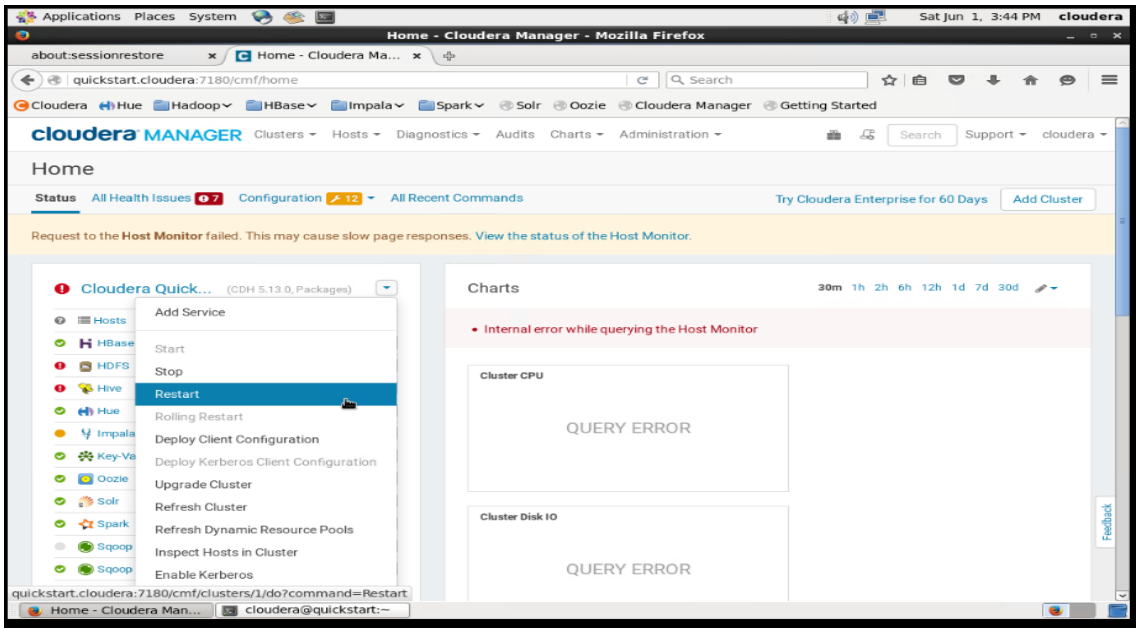
Aakash Sarap

Northeastern University Assignment 02

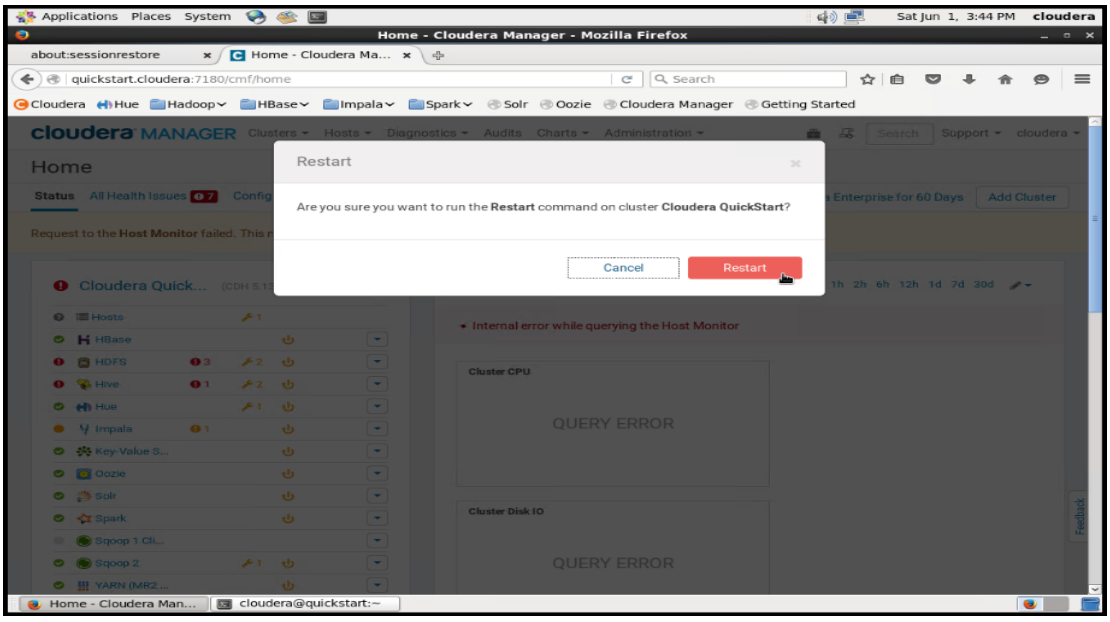
ALY6110 Data Management and Big Data

https://northeastern.blackboard.com

16. Due to inactivity, I had been logged out of my Azure Lab, so I again followed the same steps and logged on to Cloudera Manager. Now, instead of Start, I can see "Restart" under Cloudera Quickstart. Refer the below screenshot.



17. Confirm that you want to restart all the services for Cloudera Quickstart.



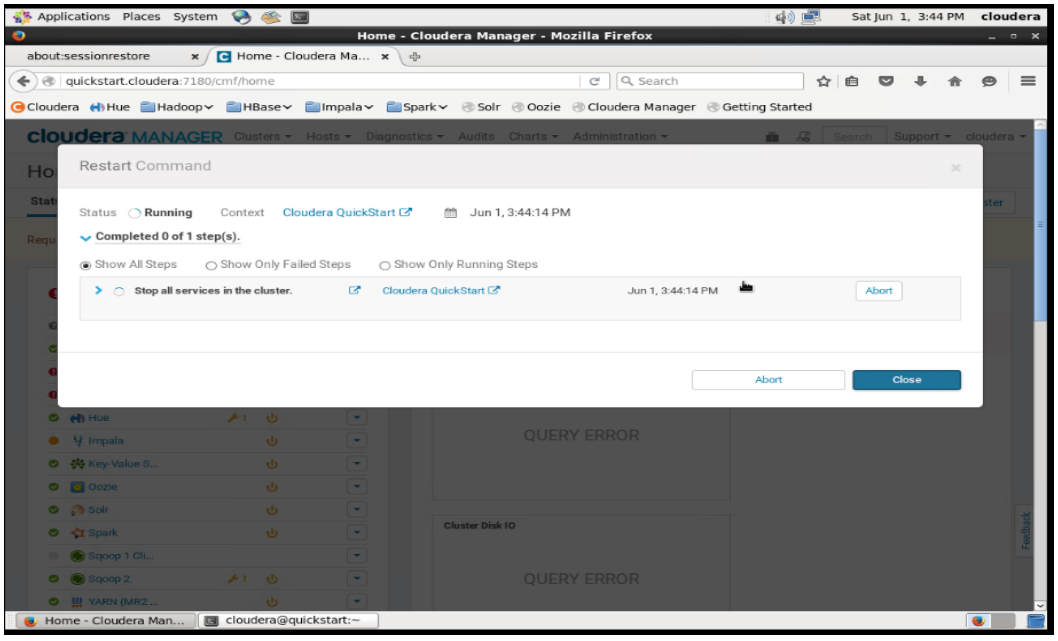
Aakash Sarap

Northeastern University Assignment 02

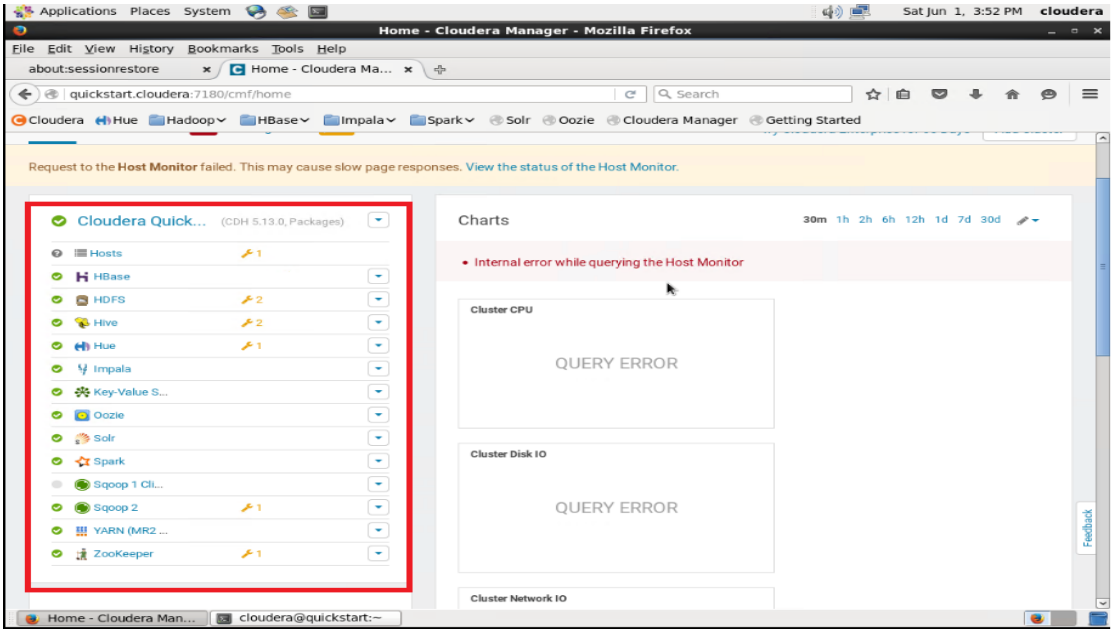
ALY6110 Data Management and Big Data

https://northeastern.blackboard.com

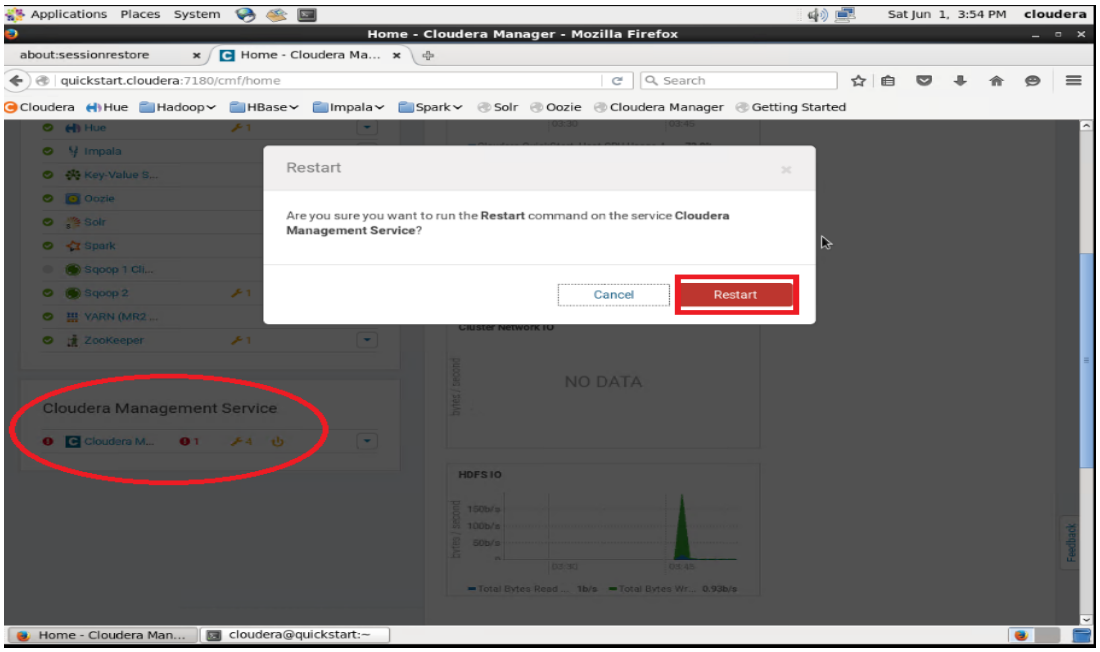
18. See the below screenshot to see how all the services are getting started.



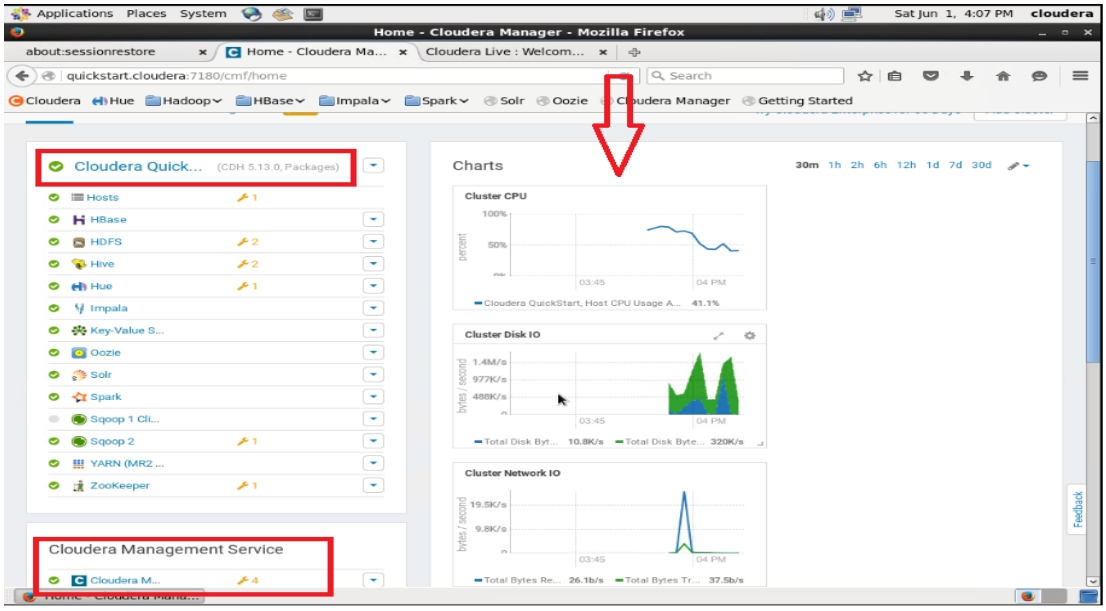
19. As you can see below, there is a green tick against the service that is successfully running.



20. Similarly, we an “Restart” the services under Cloudera Management Service as shown below.



21. The below screenshot shows that all our services are now running. We can also see the charts showing the Cluster CPU, Disk IO & Network IO.



Aakash Sarap

Northeastern University Assignment 02

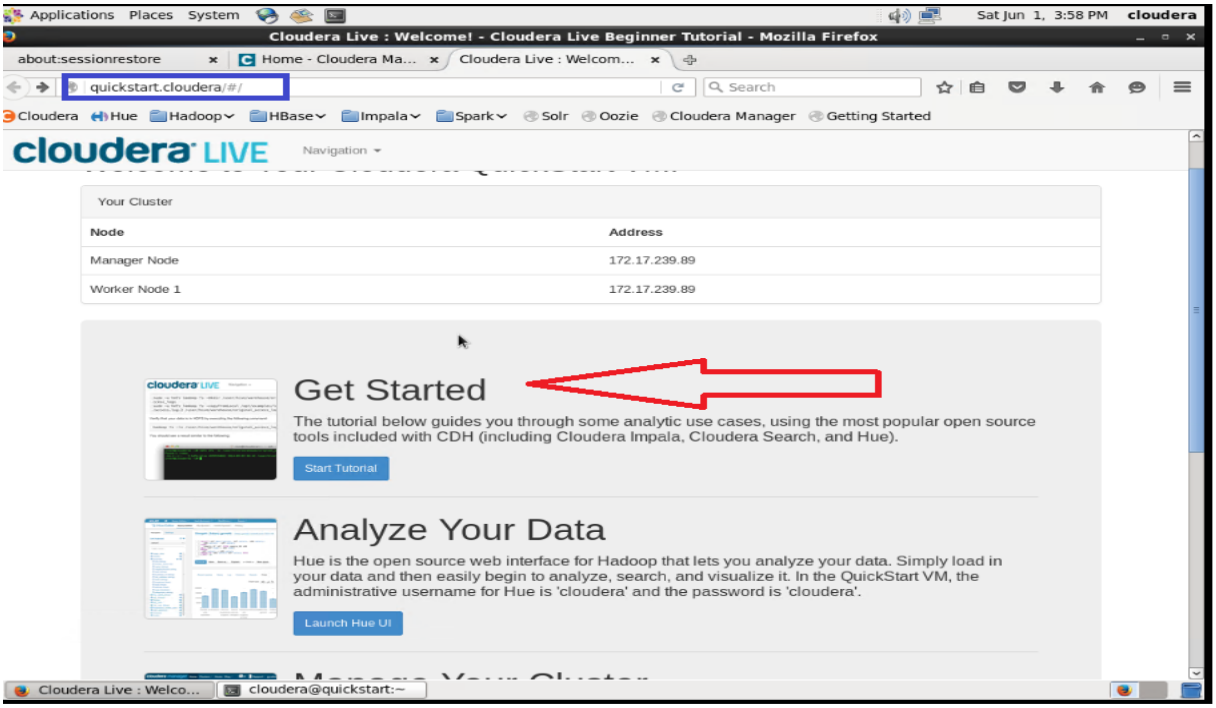
ALY6110 Data Management and Big Data

https://northeastern.blackboard.com

Problem 2:

a. **Use this page as a guide: https://www.cloudera.com/developers/get-started-with-hadoop-tutorial/setup.html**
b. **Complete the "The "Getting Started with Hadoop" Tutorial at https://www.cloudera.com/developers/get-started-with-hadoop-tutorial/setup.html.**

Go to Cloudera Live through the URL "quickstart.cloudera/#/" to start the tutorials "Get Started" by clicking on "Start Tutorial".



Below are the steps which were followed to complete the above tutorial.

The "Getting Started with Hadoop" Tutorial

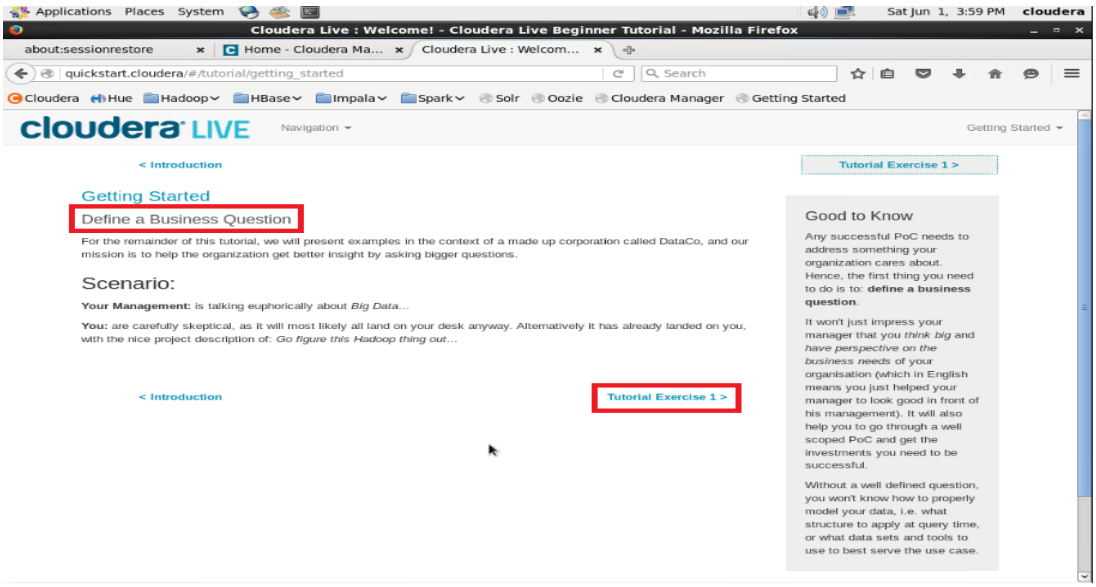
Defining our Business Question: The first step in data analysis is to define a business question. Below screenshot gives a detailed description of the scenario for Tutorial 1.

Aakash Sarap

Northeastern University Assignment 02

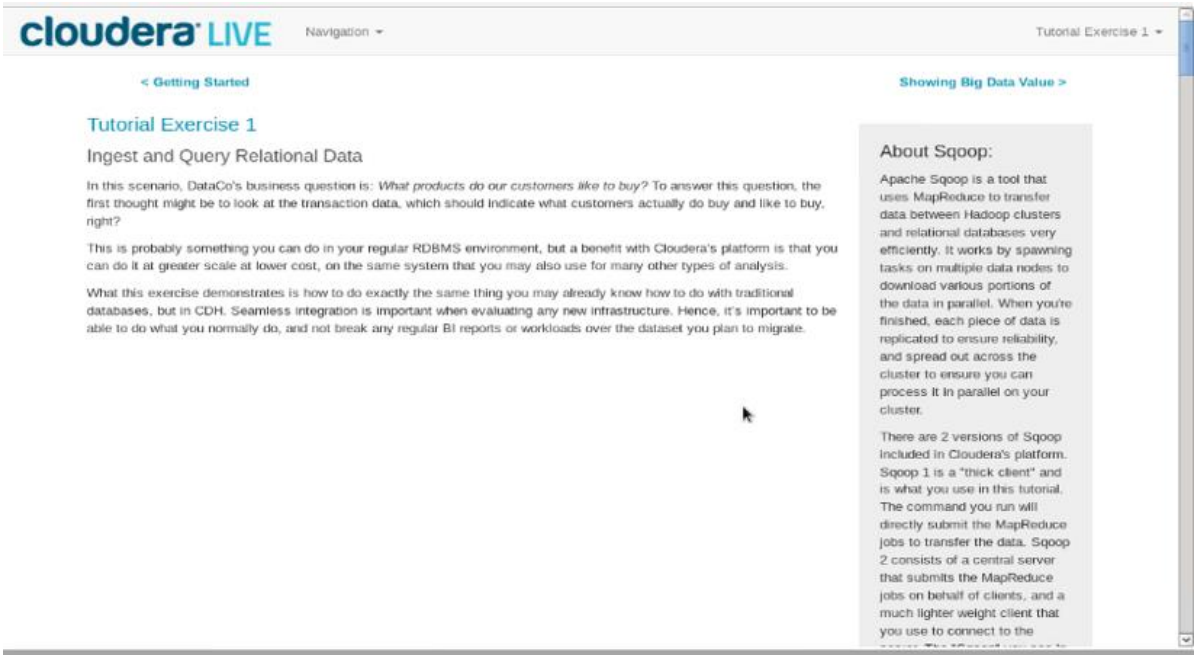
ALY6110 Data Management and Big Data

https://northeastern.blackboard.com



Tutorial Exercise 1: Ingest and Query Relational Data

Business Question for this Scenario: ***What products do the customers like to buy?***



Ingest and Query Relational Data

Aakash Sarap

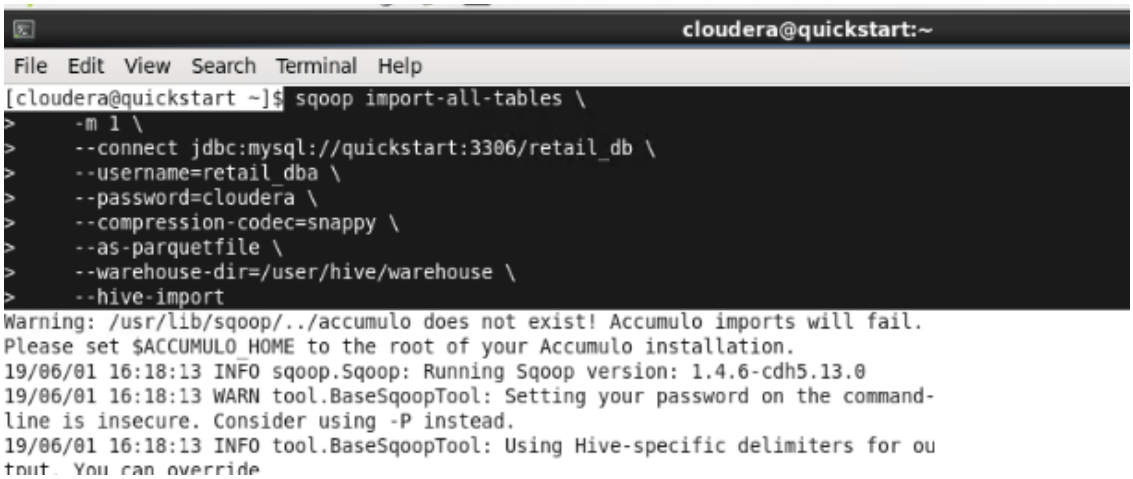
Northeastern University Assignment 02

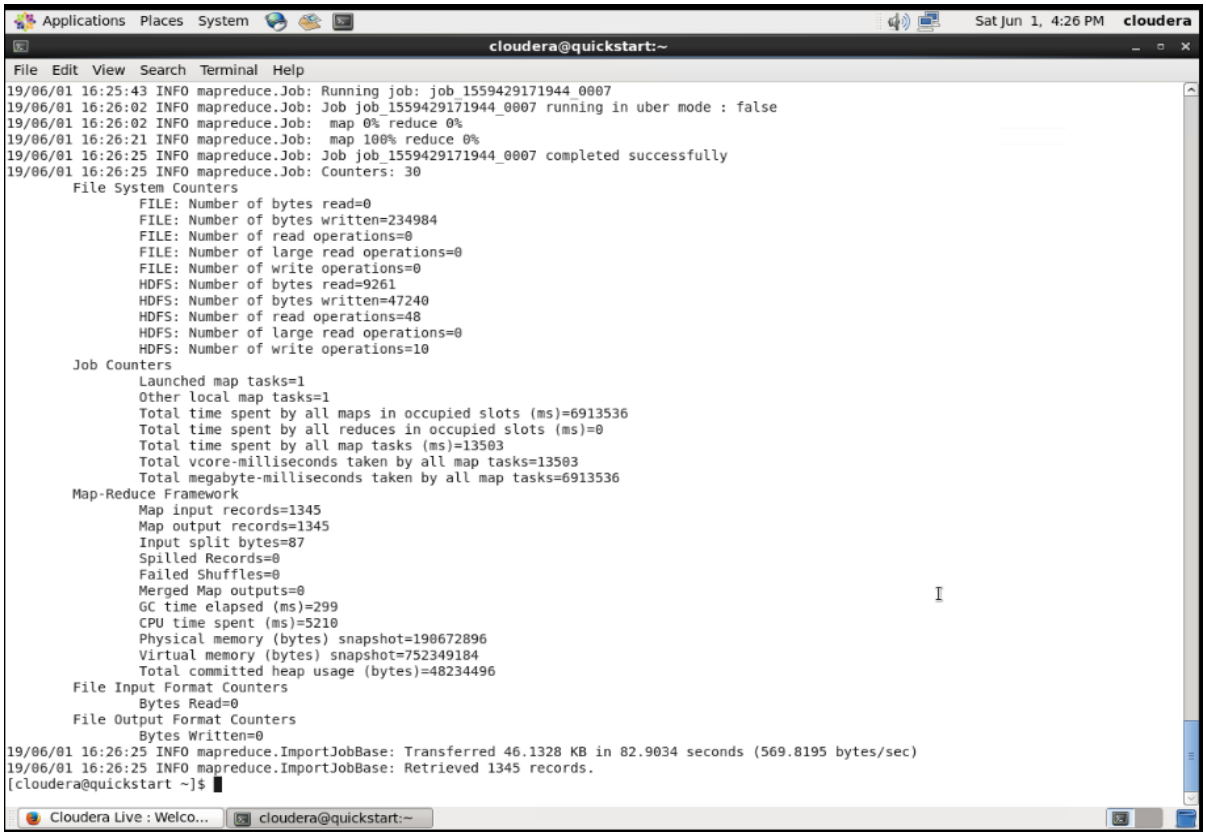
ALY6110 Data Management and Big Data

https://northeastern.blackboard.com

In this tutorial, we would be using Apache Sqoop to transfer data between Hadoop clusters and relational database. It works on parallel processing.

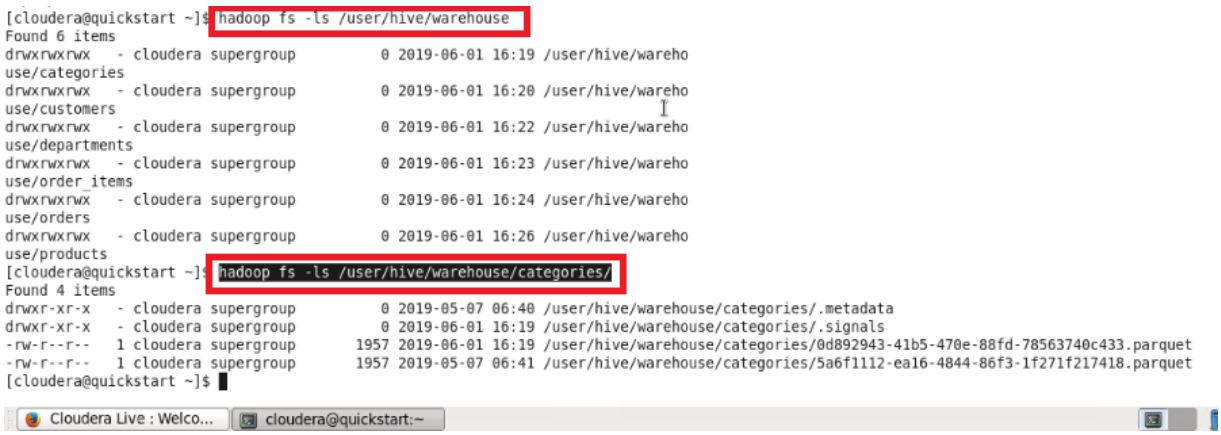
We have opened the terminal and written the command as shown in the below image.



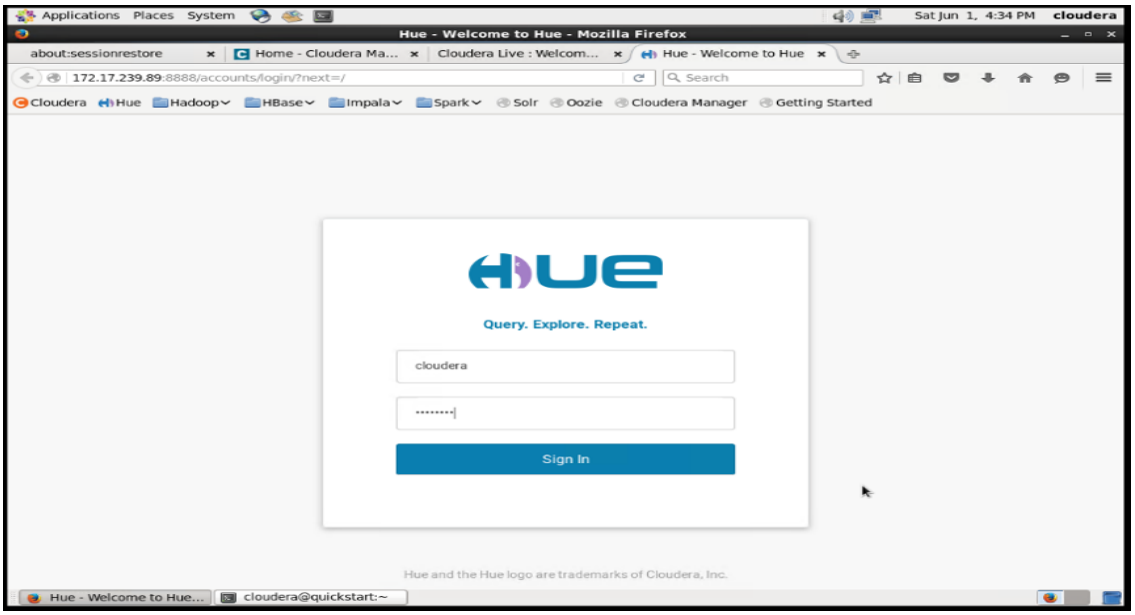


In the above image, we can see how Sqoop uses MapReduce jobs to pull the data from our MySQL database and write the data to HDFS in parallel, distributed across the cluster in Apache Parquet format.

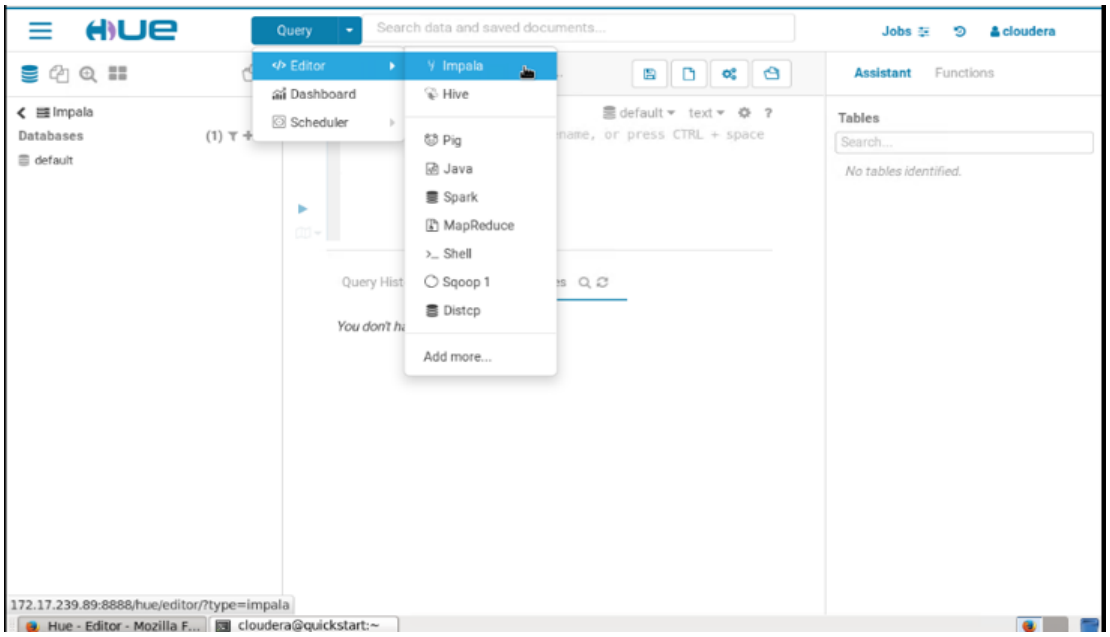
Verification: We will now confirm if our data files exist in HDFS using the commands as shown in the screenshot. We can see the directories and files inside HDFS.



Using Hue: Hue is a web-based interface for many of the tools in CDH and we would be using Impala query editor in Hue to query out tables. Open the Hue webpage from the browser in Cloudera. Enter the Username and Password for Hue in VM is the *"Cloudera"*.

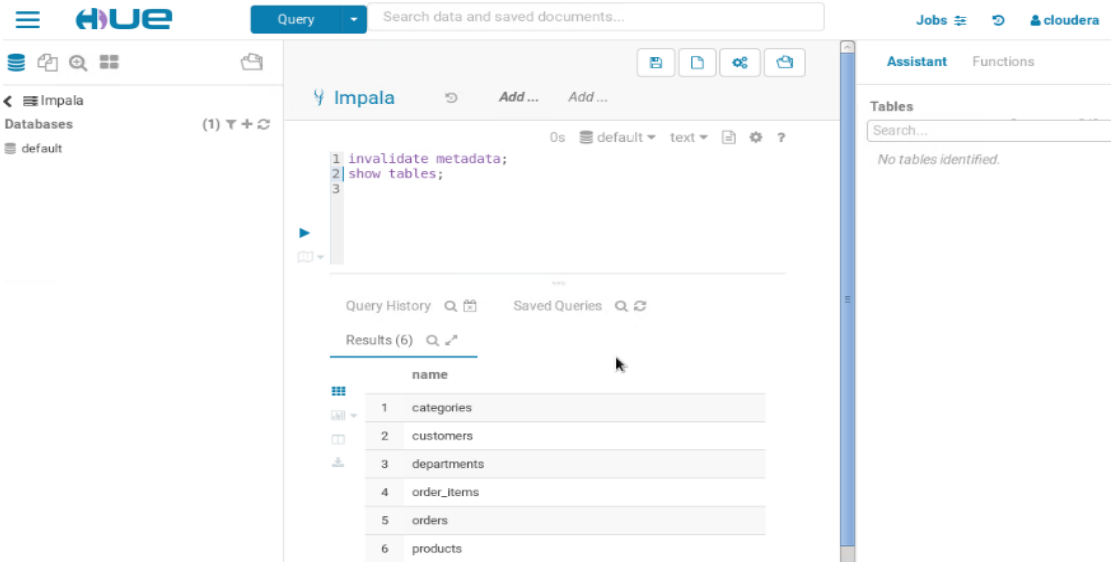


Once, we go inside Hue, go to ***Query > Editor > Impala***. This will open the Impala Query Editor for us. Refer the below image for better understanding.



Impala Query Editor in Hue

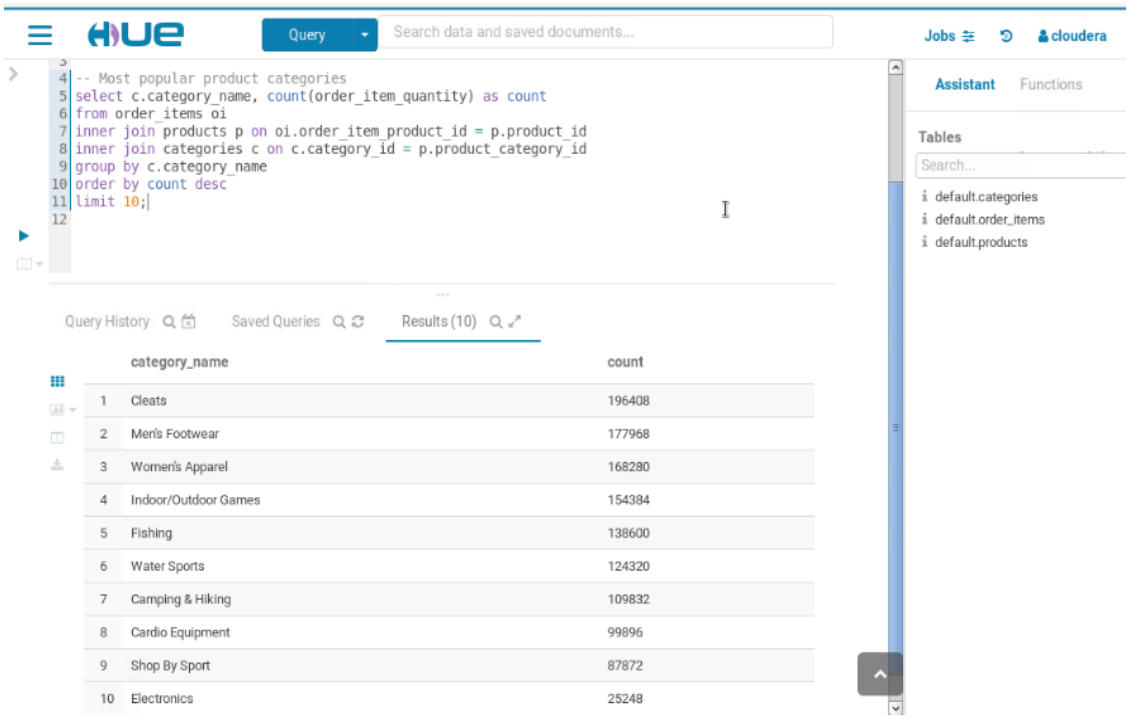
We first need to inform Impala that its metadata is out of data and it is done by executing the query *"invalidate metadata"*. Then we use the query *"show tables"* to have a look at the tables as shown below.



The query executes successfully to give us the list of six tables.

Now, we will focus on our business question. We have written the SQL queries to calculate total revenue per product and for displaying the top 10 revenue generating products as shown below.

```
-- Most popular product categories
select c.category_name, count(order_item_quantity) as count
from order_items oi
inner join products p on oi.order_item_product_id = p.product_id
inner join categories c on c.category_id = p.product_category_id
group by c.category_name
order by count desc
limit 10;
```



	category_name	count
1	Cleats	196408
2	Men's Footwear	177968
3	Women's Apparel	168280
4	Indoor/Outdoor Games	154384
5	Fishing	138600
6	Water Sports	124320
7	Camping & Hiking	109832
8	Cardio Equipment	99896
9	Shop By Sport	87872
10	Electronics	25248

Most popular product categories

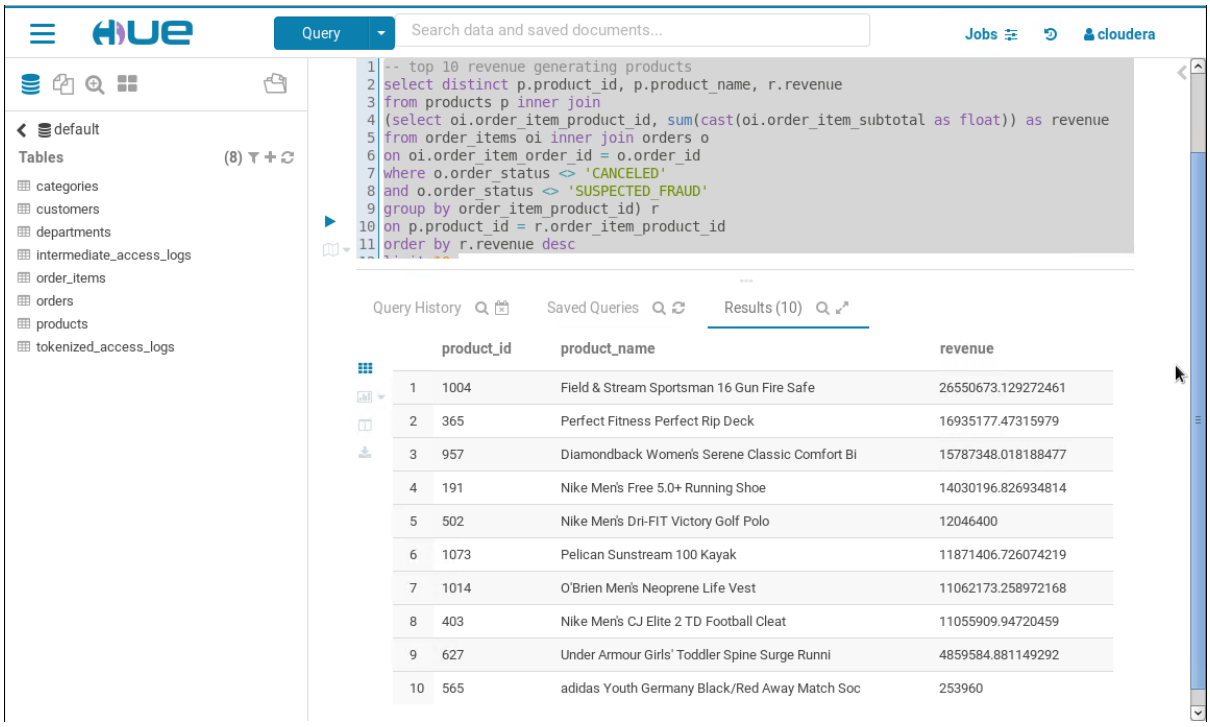
Aakash Sarap

Northeastern University Assignment 02

ALY6110 Data Management and Big Data

https://northeastern.blackboard.com

```
-- top 10 revenue generating products
select distinct p.product_id, p.product_name, r.revenue
from products p inner join
(select oi.order_item_product_id, sum(cast(oi.order_item_subtotal as float)) as revenue
from order_items oi inner join orders o
on oi.order_item_order_id = o.order_id
where o.order_status <> 'CANCELED'
and o.order_status <> 'SUSPECTED_FRAUD'
group by order_item_product_id) r
on p.product_id = r.order_item_product_id
order by r.revenue desc
limit 10;
```



	product_id	product_name	revenue
1	1004	Field & Stream Sportsman 16 Gun Fire Safe	26550673.129272461
2	365	Perfect Fitness Perfect Rip Deck	16935177.47315979
3	957	Diamondback Women's Serene Classic Comfort Bi	15787348.018188477
4	191	Nike Men's Free 5.0+ Running Shoe	14030196.826934814
5	502	Nike Men's Dri-FIT Victory Golf Polo	12046400
6	1073	Pelican Sunstream 100 Kayak	11871406.726074219
7	1014	O'Brien Men's Neoprene Life Vest	11062173.258972168
8	403	Nike Men's CJ Elite 2 TD Football Cleat	11055909.94720459
9	627	Under Armour Girls' Toddler Spine Surge Runni	4859584.881149292
10	565	adidas Youth Germany Black/Red Away Match Soc	253960

Top 10 revenue generating products

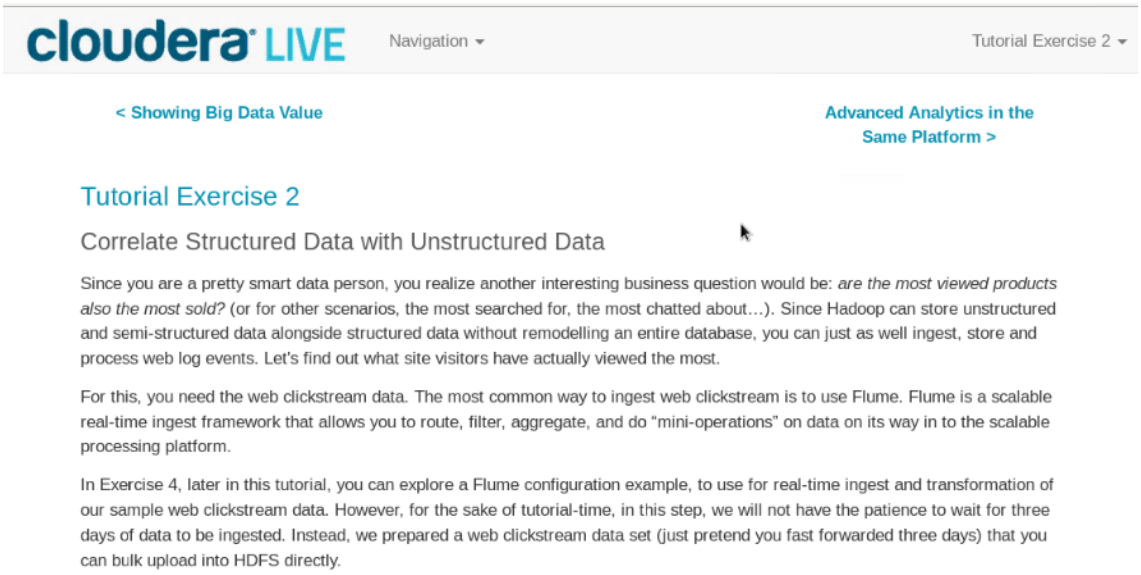
From the results, we can easily identify the most popular products as well as the top 10 revenue generating products. For e.g., Field & Stream Sportsman 16 Gun Fire Safe is the product that is generating the maximum revenue for the company.

Conclusion

Through this tutorial we learned how to Sqoop structured data into HDFS, transforming of the data into Parquet file format and creating hive tables. We also learned how to use SQL like queries in Hadoop environment with the help of Impala query editor.

Tutorial Exercise 2: Correlate Structured Data with Unstructured Data

In this tutorial, we are using *Apache Flume* to ingest web clickstream data. Flume is a scalable real-time ingest framework that allows us to route, filter, aggregate, and do "mini-operations" on data.



Bulk Upload Data

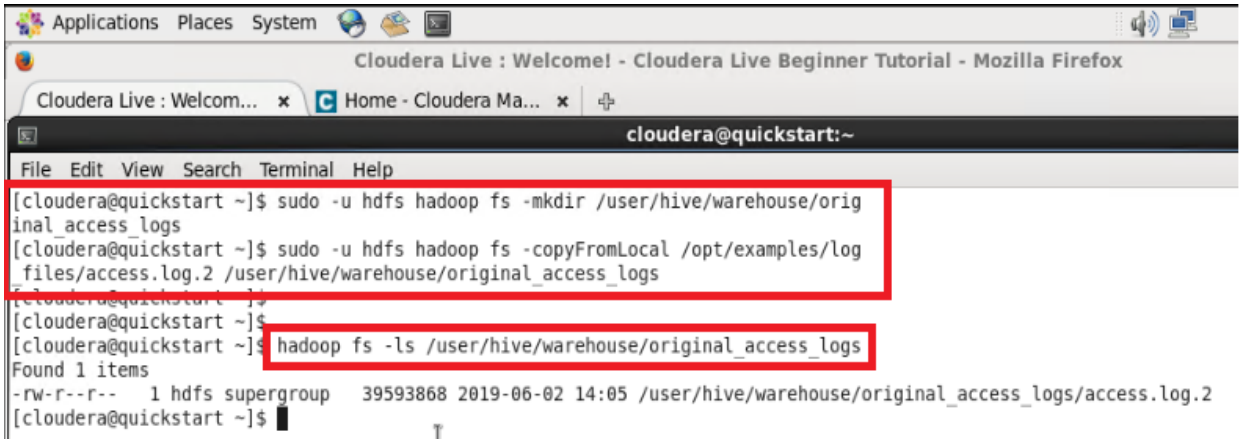
There are some pre-loaded sample access log data into */opt/examples/log_data/access.log.2*. We have moved data from local file system to HDFS with the help of the commands as shown in the below image.

Aakash Sarap

Northeastern University Assignment 02

ALY6110 Data Management and Big Data

https://northeastern.blackboard.com



Now we must build a table in Hive and query the data using Apache Impala and Hue. The first step in this would be to parse the logs into individual fields using a regular expression using SerDes. The second step would be transferring the data from this table to the one which doesn't need any SerDe.

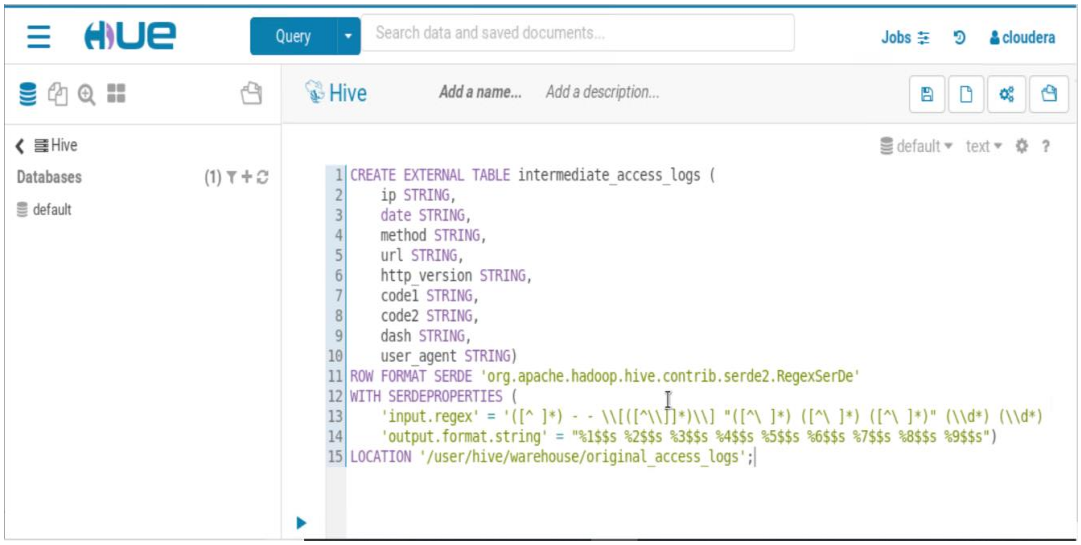
Let's go to Hive query editor in Hue to execute the queries shown in below image.

Aakash Sarap

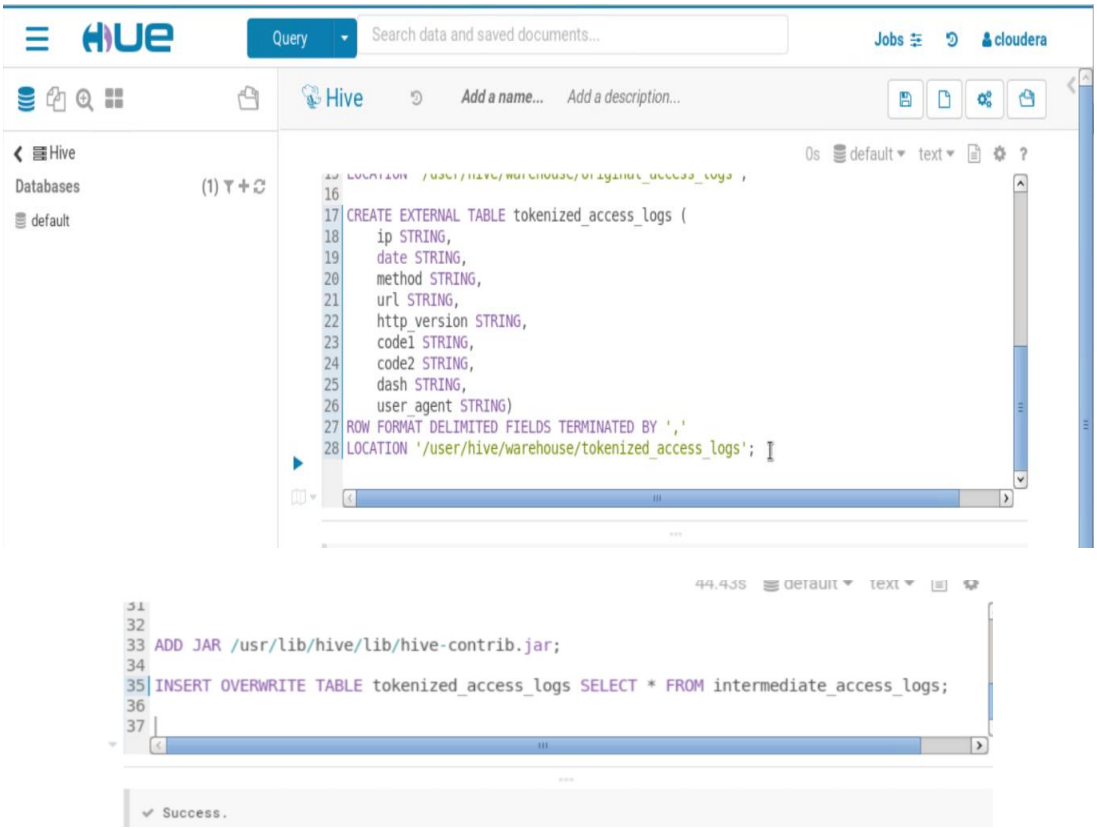
Northeastern University Assignment 02

ALY6110 Data Management and Big Data

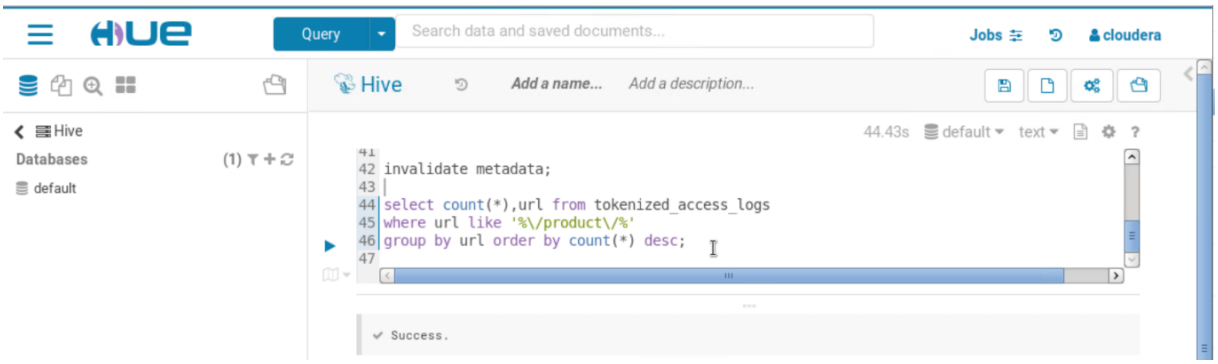
https://northeastern.blackboard.com

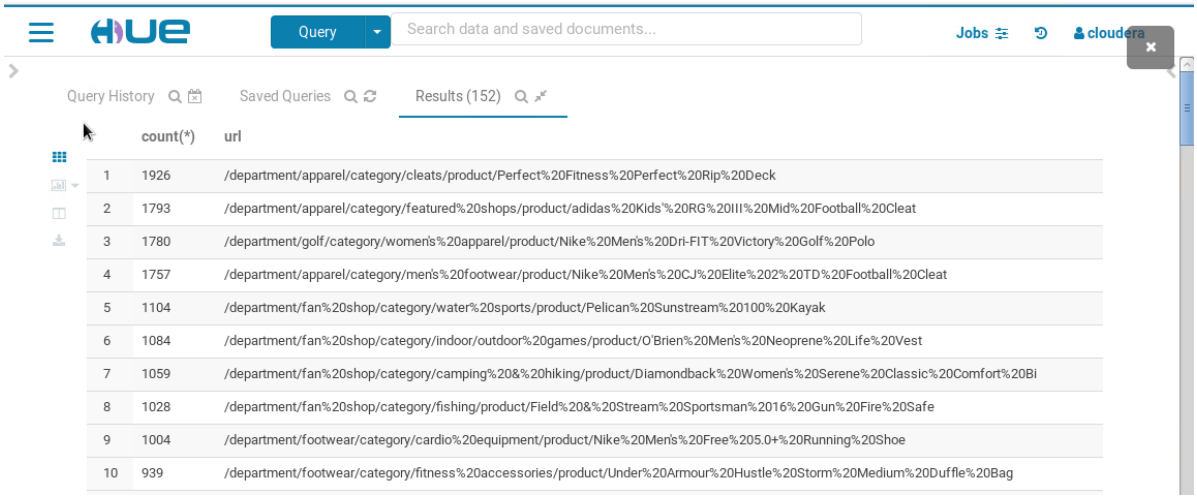


The second query which is shown in the below image takes some time to run. It works in the same way as Sqoop import works using a MapReduce job.



Now, we need to go to Impala query editor and enter the queries as shown in the below image.





	count(*)	url
1	1926	/department/apparel/category/cleats/product/Perfect%20Fitness%20Perfect%20Rip%20Deck
2	1793	/department/apparel/category/featured%20shops/product/adidas%20Kids'%20RG%20III%20Mid%20Football%20Cleat
3	1780	/department/golf/category/women's%20apparel/product/Nike%20Men's%20Dri-FIT%20Victory%20Golf%20Polo
4	1757	/department/apparel/category/men's%20footwear/product/Nike%20Men's%20CJ%20Elite%202%20TD%20Football%20Cleat
5	1104	/department/fan%20shop/category/water%20sports/product/Pelican%20Sunstream%20100%20Kayak
6	1084	/department/fan%20shop/category/indoor/outdoor%20games/product/O'Brien%20Men's%20Neoprene%20Life%20Vest
7	1059	/department/fan%20shop/category/camping%20&%20hiking/product/Diamondback%20Women's%20Serene%20Classic%20Comfort%20Bi
8	1028	/department/fan%20shop/category/fishing/product/Field%20&%20Stream%20Sportsman%2016%20Gun%20Fire%20Safe
9	1004	/department/footwear/category/cardio%20equipment/product/Nike%20Men's%20Free%205.0+%20Running%20Shoe
10	939	/department/footwear/category/fitness%20accessories/product/Under%20Armour%20Hustle%20Storm%20Medium%20Duffle%20Bag

Output: 152 rows returned

After carefully observing the results, we found that the list of most sold products is displayed from the previous tutorial steps, but there is one product which was missing in the previous list.

In this example of DataCo from Tutorial Exercise 1, the reason for that missing product could be assumed as, on the view page where a lot of visitors stopped and checked that product but because of a typo in the price of that item. And after the typo was fixed and the displayed price was corrected, the sales for that product increased rapidly.

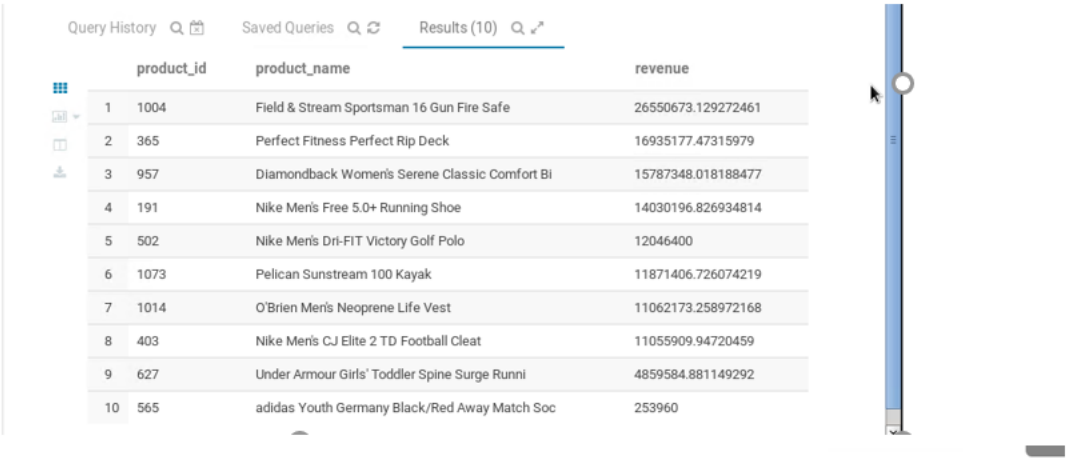
Refer the below screenshot from the tutorial link for this missing product details.

Aakash Sarap

Northeastern University Assignment 02

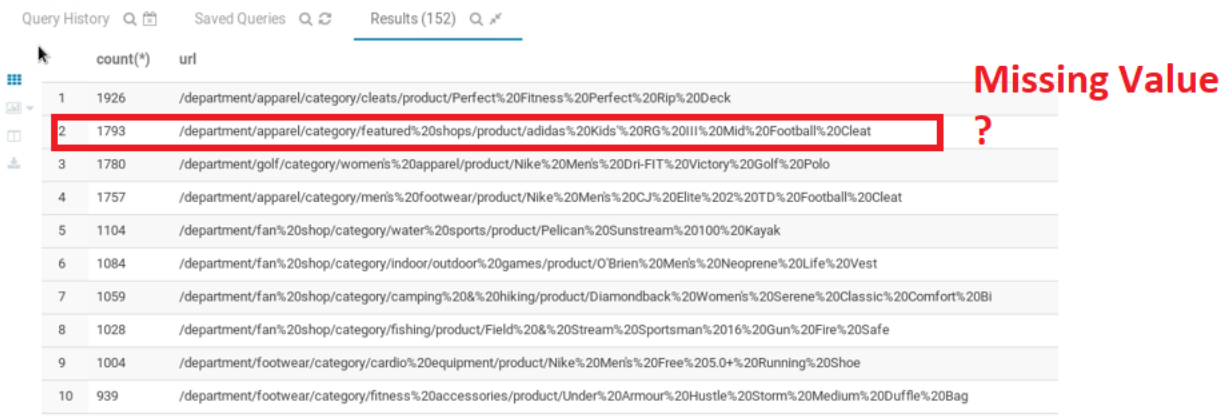
ALY6110 Data Management and Big Data

https://northeastern.blackboard.com



Query History | Saved Queries | Results (10)

	product_id	product_name	revenue
1	1004	Field & Stream Sportsman 16 Gun Fire Safe	26550673.129272461
2	365	Perfect Fitness Perfect Rip Deck	16935177.47315979
3	957	Diamondback Women's Serene Classic Comfort Bi	15787348.018188477
4	191	Nike Men's Free 5.0+ Running Shoe	14030196.826934814
5	502	Nike Men's Dri-FIT Victory Golf Polo	12046400
6	1073	Pelican Sunstream 100 Kayak	11871406.726074219
7	1014	O'Brien Men's Neoprene Life Vest	11062173.258972168
8	403	Nike Men's CJ Elite 2 TD Football Cleat	11055909.94720459
9	627	Under Armour Girls' Toddler Spine Surge Runni	4859584.881149292
10	565	adidas Youth Germany Black/Red Away Match Soc	253960



Query History | Saved Queries | Results (152)

	count(*)	url
1	1926	/department/apparel/category/cleats/product/Perfect%20Fitness%20Perfect%20Rip%20Deck
2	1793	/department/apparel/category/featured%20shops/product/adidas%20Kids'%20RG%20III%20Mid%20Football%20Cleat
3	1780	/department/golf/category/women's%20apparel/product/Nike%20Men's%20Dri-FIT%20Victory%20Golf%20Polo
4	1757	/department/apparel/category/men's%20footwear/product/Nike%20Men's%20CJ%20Elite%202%20TD%20Football%20Cleat
5	1104	/department/fan%20shop/category/water%20sports/product/Pelican%20Sunstream%20100%20Kayak
6	1084	/department/fan%20shop/category/indoor/outdoor%20games/product/O'Brien%20Men's%20Neoprene%20Life%20Vest
7	1059	/department/fan%20shop/category/camping%20&%20hiking/product/Diamondback%20Women's%20Serene%20Classic%20Comfort%20Bi
8	1028	/department/fan%20shop/category/fishing/product/Field%20&%20Stream%20Sportsman%2016%20Gun%20Fire%20Safe
9	1004	/department/footwear/category/cardio%20equipment/product/Nike%20Men's%20Free%205.0+%20Running%20Shoe
10	939	/department/footwear/category/fitness%20accessories/product/Under%20Armour%20Hustle%20Storm%20Medium%20Duffle%20Bag

Missing Value ?

Conclusion

Through this tutorial we got to know that if we correlate two datasets for the same business questions adds value to any business and as the analysis was being on the same platform, it makes life easy for both the big data analyst as well as for the business.

Tutorial Exercise 3: Relationship Strength Analytics using Spark

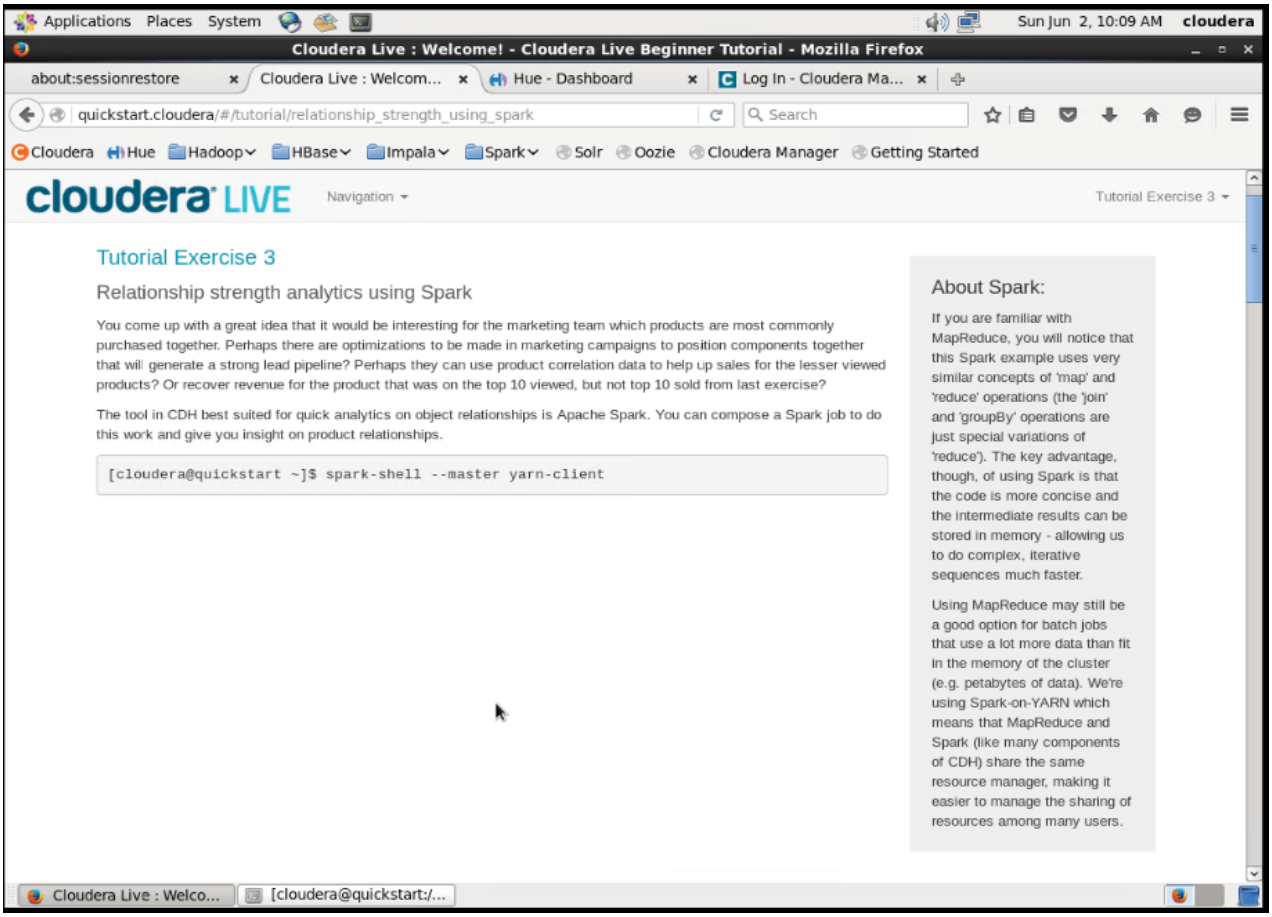
In this tutorial, we are using Apache Spark which uses the similar concepts of MapReduce operations, but it has the ability to perform in-memory operations and thus gives concise results very fast.

Aakash Sarap

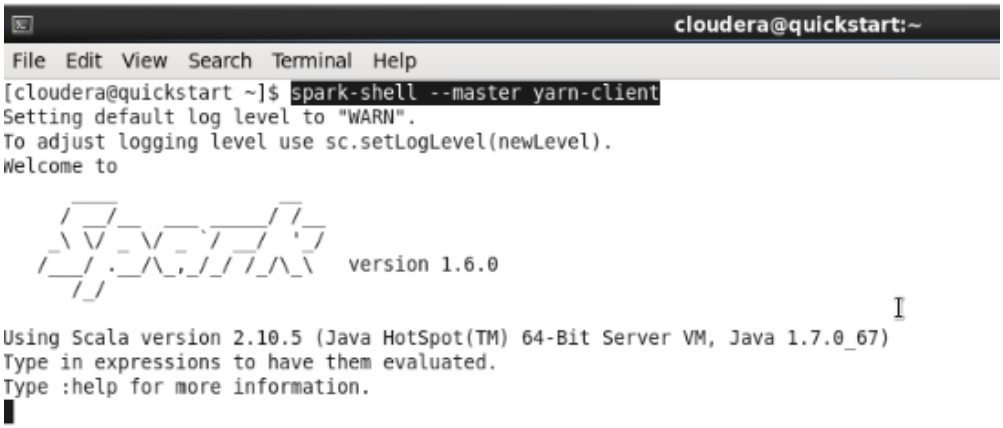
Northeastern University Assignment 02

ALY6110 Data Management and Big Data

https://northeastern.blackboard.com



To install Spark on Cloudera, we write the command shown in the below screenshot.



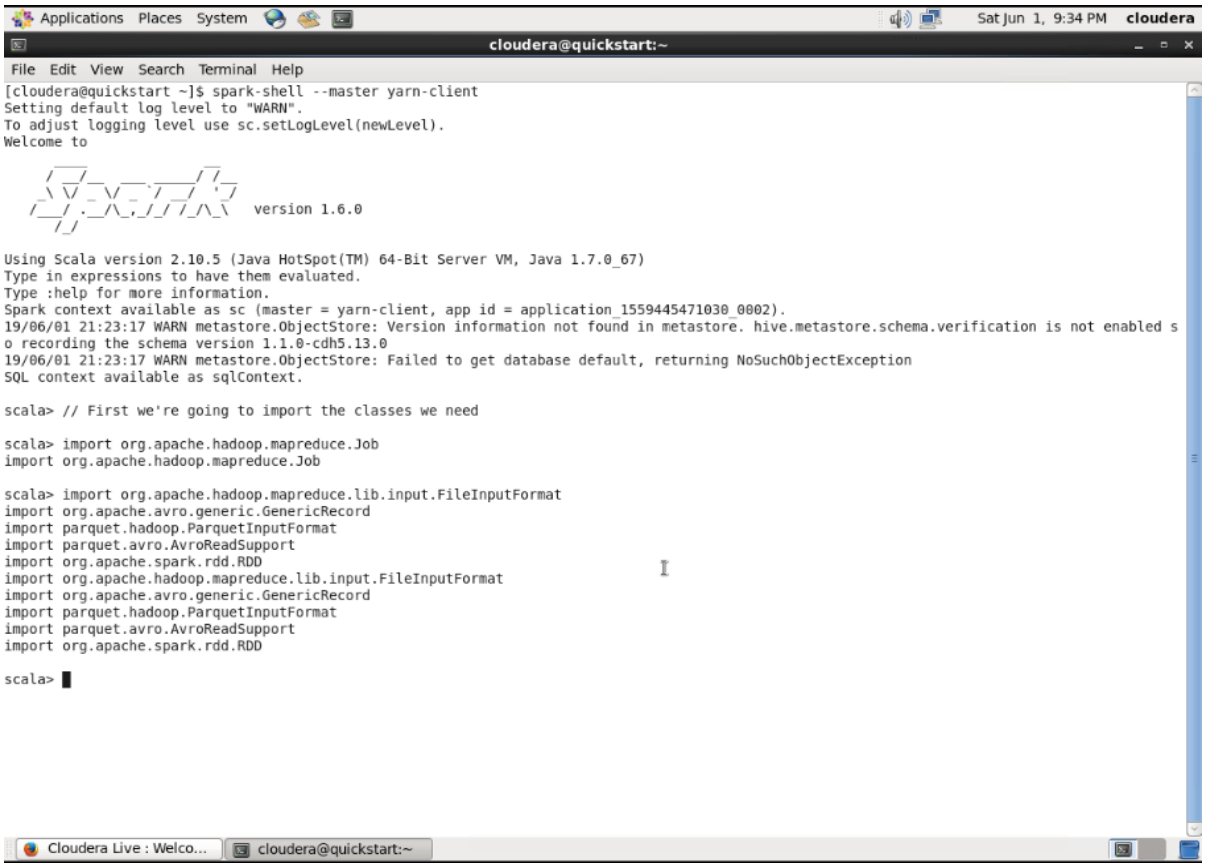
Aakash Sarap

Northeastern University Assignment 02

ALY6110 Data Management and Big Data

https://northeastern.blackboard.com

The below screenshot shows that the Spark has been successfully installed as it shows Scala at the terminal prompt.



As the Scala prompt appeared, we will write the code as shown in the below image to import the classes we need.

```
scala> // First we're going to import the classes we need

scala> import org.apache.hadoop.mapreduce.Job
import org.apache.hadoop.mapreduce.Job

scala> import org.apache.hadoop.mapreduce.lib.input.FileInputFormat
import org.apache.avro.generic.GenericRecord
import parquet.hadoop.ParquetInputFormat
import parquet.avro.AvroReadSupport
import org.apache.spark.rdd.RDD
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat
import org.apache.avro.generic.GenericRecord
import parquet.hadoop.ParquetInputFormat
import parquet.avro.AvroReadSupport
import org.apache.spark.rdd.RDD
```

After this, we will be creating RDD's for the files we imported from MySQL with Sqoop. RDD's are the low-level API for Spark's data structures for working with distributed datasets.

```
scala> // Then we create RDD's for 2 of the files we imported from MySQL with Sqoop
// RDD's are Spark's data structures for working with distributed datasets

scala> def rddFromParquetHdfsFile(path: String): RDD[GenericRecord] = {
    val job = new Job()
    FileInputFormat.setInputPaths(job, path)
    ParquetInputFormat.setReadSupportClass(job,
        classOf[AvroReadSupport[GenericRecord]])
    return sc.newAPIHadoopRDD(job.getConfiguration,
        classOf[ParquetInputFormat[GenericRecord]],
        classOf[Void],
        classOf[GenericRecord]).map(x => x._2)
}
warning: there were 1 deprecation warning(s); re-run with -deprecation for details
rddFromParquetHdfsFile: (path: String)org.apache.spark.rdd.RDD[org.apache.avro.generic.GenericRecord]
```

Now, we need to extract the fields of interest from the tables *order_items* and *products*. After that, we try to get the list of every product, its name, and quantity grouped by order.

```
scala> // Next, we extract the fields from order_items and products that we care about
// and get a list of every product, its name and quantity, grouped by order

scala> val orders = order_items.map { x => (
    x.get("order_item_product_id"),
    (x.get("order_item_order_id"), x.get("order_item_quantity")))
}.join(
  products.map { x => (
    x.get("product_id"),
    (x.get("product_name")))
  }
).map(x => (
    scala.Int.unbox(x._2._1._1), // order_id
    (
        scala.Int.unbox(x._2._1._2), // quantity
        x._2._2.toString // product_name
    )
)).groupByKey()
orders: org.apache.spark.rdd.RDD[(Int, Iterable[(Int, String)])] = ShuffledRDD[10] at groupByKey at <console>:55
```

Now, we will count how many times each combination of products appears together in an order and then we sort them and only take 10 most frequent combinations.

Aakash Sarap

Northeastern University Assignment 02

ALY6110 Data Management and Big Data

https://northeastern.blackboard.com

```
scala> val orders = order_items.map { x => (
    x.get("order_item_product_id"),
    (x.get("order_item_order_id"), x.get("order_item_quantity")))
}.join(
  products.map { x => (
    x.get("product_id"),
    (x.get("product_name")))
  }
).map(x => (
    scala.Int.unbox(x._2._1._1), // order_id
    (
        scala.Int.unbox(x._2._1._2), // quantity
        x._2._2.toString // product_name
    )
)).groupByKey()
orders: org.apache.spark.rdd.RDD[(Int, Iterable[(Int, String)])] = ShuffledRDD[10] at groupByKey at <console>:55

scala> val cooccurrences = orders.map(order =>
  (
    order._1,
    order._2.toList.combinations(2).map(order_pair =>
        (
            if (order_pair(0)._2 < order_pair(1)._2)
                (order_pair(0)._2, order_pair(1)._2)
            else
                (order_pair(1)._2, order_pair(0)._2),
            order_pair(0)._1 * order_pair(1)._1
        )
    )
  )
)
val combos = cooccurrences.flatMap(x => x._2).reduceByKey((a, b) => a + b)
val mostCommon = combos.map(x => (x._2, x._1)).sortByKey(false).take(10)
cooccurrences: org.apache.spark.rdd.RDD[(Int, Iterator[((String, String), Int)])] = MapPartitionsRDD[11] at map at <console>:43
combos: org.apache.spark.rdd.RDD[((String, String), Int)] = ShuffledRDD[13] at reduceByKey at <console>:57
mostCommon: Array[(Int, (String, String))] = Array((290678,(Perfect Fitness Perfect Rip Deck,Perfect Fitness Perfect Rip Deck)), (246633,(Nik
e Men's Dri-FIT Victory Golf Polo,Nike Men's Dri-FIT Victory Golf Polo)), (225067,(O'Brien Men's Neoprene Life Vest,O'Brien Men's Neoprene Li
fe Vest)), (140254,(Nike Men's Free 5.0+ Running Shoe,Nike Men's Free 5.0+ Running Shoe)), (119904,(Under Armour Girls' Toddler Spine Surge R
unni,Under Armour Girls' Toddler Spine Surge Runni)), (67876,(Nike Men's Dri-FIT Victory Golf Polo,Perfect Fitness Perfect Rip Deck)), (629..
```

Now, we will print results, each output in one line and then exit the Spark shell.

```
scala> // We print our results, 1 per line, and exit the Spark shell

scala> println(mostCommon.deep.mkString("\n"))
(290678,(Perfect Fitness Perfect Rip Deck,Perfect Fitness Perfect Rip Deck))
(246633,(Nike Men's Dri-FIT Victory Golf Polo,Nike Men's Dri-FIT Victory Golf Polo))
(225067,(O'Brien Men's Neoprene Life Vest,O'Brien Men's Neoprene Life Vest))
(140254,(Nike Men's Free 5.0+ Running Shoe,Nike Men's Free 5.0+ Running Shoe))
(119904,(Under Armour Girls' Toddler Spine Surge Runni,Under Armour Girls' Toddler Spine Surge Runni))
(67876,(Nike Men's Dri-FIT Victory Golf Polo,Perfect Fitness Perfect Rip Deck))
(62924,(O'Brien Men's Neoprene Life Vest,Perfect Fitness Perfect Rip Deck))
(54399,(Nike Men's Dri-FIT Victory Golf Polo,O'Brien Men's Neoprene Life Vest))
(39656,(Nike Men's Free 5.0+ Running Shoe,Perfect Fitness Perfect Rip Deck))
(35092,(Perfect Fitness Perfect Rip Deck,Under Armour Girls' Toddler Spine Surge Runni))

scala> exit
warning: there were 1 deprecation warning(s); re-run with -deprecation for details
Jun 1, 2019 9:44:18 PM INFO: parquet.hadoop.ParquetInputFormat: Total input paths to process : 2
Jun 1, 2019 9:44:18 PM INFO: parquet.hadoop.ParquetInputFormat: Total input paths to process : 2
[cloudera@quickstart ~]$
```

Aakash Sarap

Northeastern University Assignment 02

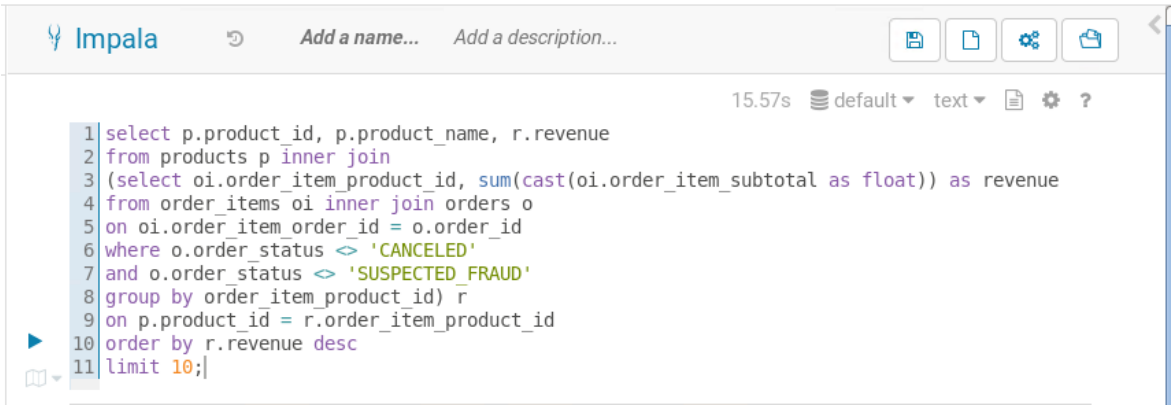
ALY6110 Data Management and Big Data

https://northeastern.blackboard.com

Problem 3:

Answer these questions:

A. What is the 5th most revenue generating product?



Query History | Saved Queries | Results (10)

	product_id	product_name	revenue
1	1004	Field & Stream Sportsman 16 Gun Fire Safe	26550673.129272461
2	365	Perfect Fitness Perfect Rip Deck	16935177.47315979
3	957	Diamondback Women's Serene Classic Comfort Bi	15787348.018188477
4	191	Nike Men's Free 5.0+ Running Shoe	14030196.826934814
5	502	Nike Men's Dri-FIT Victory Golf Polo	12046400
6	1073	Pelican Sunstream 100 Kayak	11871406.726074219
7	1014	O'Brien Men's Neoprene Life Vest	11062173.258972168
8	403	Nike Men's CJ Elite 2 TD Football Cleat	11055909.94720459
9	627	Under Armour Girls' Toddler Spine Surge Runni	4859584.881149292
10	565	adidas Youth Germany Black/Red Away Match Soc	253960

After running the above query in Impala Query Editor, we got the top 10 most revenue generating products. From the output, we can see that ***5th most Revenue Generating Product is Nike Men's Dri-Fit Victory Golf Polo.***

Aakash Sarap

Northeastern University Assignment 02

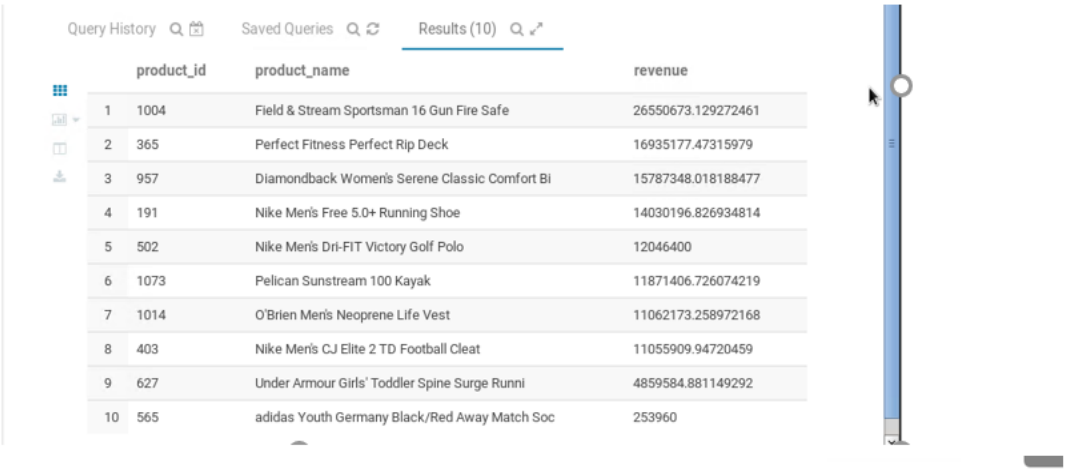
ALY6110 Data Management and Big Data

https://northeastern.blackboard.com

B. How much revenue does the Nike men's dry fit polo earn?

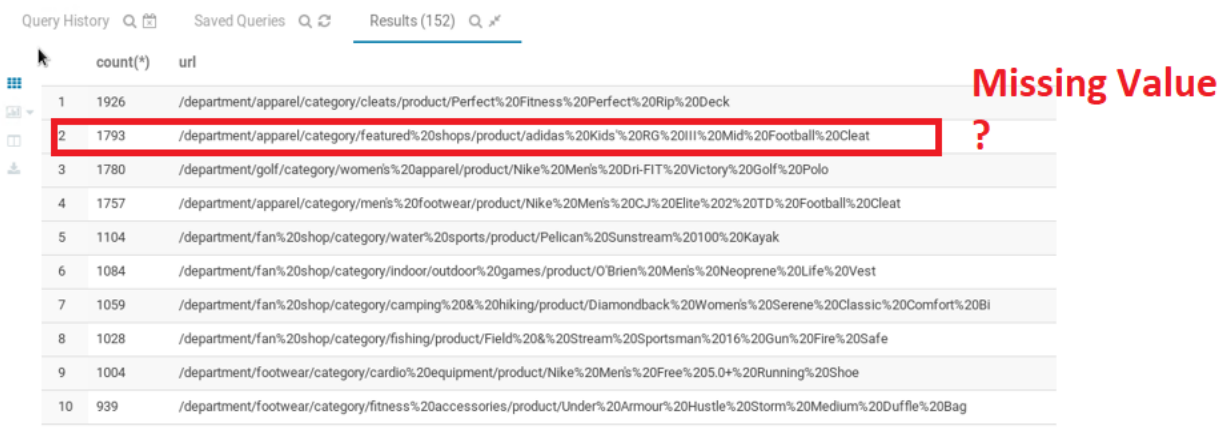
From the above output screenshot, we can see that ***Nike Men's Dry-Fit Polo earns the revenue of $12046400***

C. There is one product that did not show up in the previous result. It seems to be viewed a lot, but never purchased. Why?



Query History | Saved Queries | Results (10)

	product_id	product_name	revenue
1	1004	Field & Stream Sportsman 16 Gun Fire Safe	26550673.129272461
2	365	Perfect Fitness Perfect Rip Deck	16935177.47315979
3	957	Diamondback Women's Serene Classic Comfort Bi	15787348.018188477
4	191	Nike Men's Free 5.0+ Running Shoe	14030196.826934814
5	502	Nike Men's Dri-FIT Victory Golf Polo	12046400
6	1073	Pelican Sunstream 100 Kayak	11871406.726074219
7	1014	O'Brien Men's Neoprene Life Vest	11062173.258972168
8	403	Nike Men's CJ Elite 2 TD Football Cleat	11055909.94720459
9	627	Under Armour Girls' Toddler Spine Surge Runni	4859584.881149292
10	565	adidas Youth Germany Black/Red Away Match Soc	253960



Query History | Saved Queries | Results (152)

	count(*)	url
1	1926	/department/apparel/category/cleats/product/Perfect%20Fitness%20Perfect%20Rip%20Deck
2	1793	/department/apparel/category/featured%20shops/product/adidas%20Kids'%20RG%20III%20Mid%20Football%20Cleat
3	1780	/department/golf/category/women's%20apparel/product/Nike%20Men's%20Dri-FIT%20Victory%20Golf%20Polo
4	1757	/department/apparel/category/men's%20footwear/product/Nike%20Men's%20CJ%20Elite%202%20TD%20Football%20Cleat
5	1104	/department/fan%20shop/category/water%20sports/product/Pelican%20Sunstream%20100%20Kayak
6	1084	/department/fan%20shop/category/indoor/outdoor%20games/product/O'Brien%20Men's%20Neoprene%20Life%20Vest
7	1059	/department/fan%20shop/category/camping%20&%20hiking/product/Diamondback%20Women's%20Serene%20Classic%20Comfort%20Bi
8	1028	/department/fan%20shop/category/fishing/product/Field%20&%20Stream%20Sportsman%2016%20Gun%20Fire%20Safe
9	1004	/department/footwear/category/cardio%20equipment/product/Nike%20Men's%20Free%205.0+%20Running%20Shoe
10	939	/department/footwear/category/fitness%20accessories/product/Under%20Armour%20Hustle%20Storm%20Medium%20Duffle%20Bag

Missing Value ?

After carefully observing the results, we found that the list of most sold products is displayed from the previous tutorial steps, but there is one product which was missing in the previous list and that product is ***Adidas Kids III Mid Football Cleat.*** In this example of DataCo from

Aakash Sarap

Northeastern University Assignment 02

ALY6110 Data Management and Big Data

https://northeastern.blackboard.com

Tutorial Exercise 1, the reason for that missing product could be assumed as, on the view page where a lot of visitors stopped and checked that product but because of a typo in the price of that item. And after the typo was fixed and the displayed price was corrected, the sales for that product increased rapidly.

References

Cloudera. (2019). *The "Getting Started With Hadoop" tutorial, setup | Cloudera*. [online] Available at https://www.cloudera.com/developers/get-started-with-hadoop-tutorial/setup.html [Accessed 3 Jun. 2019].