DATA MANAGEMENT AND BIG DATA  
  
  
ALY6110, WINTER 2021  
MODULE 2 ASSIGNMENT

NAME: SHIVANI ADSAR

CLASS NUMBER & CLASS NAME: ALY6110 & DATA MANAGEMENT AND BIG DATA

CRN NUMBER: 22824

ASSIGNMENT NAME: WEEK 2 ASSIGNMENT 1

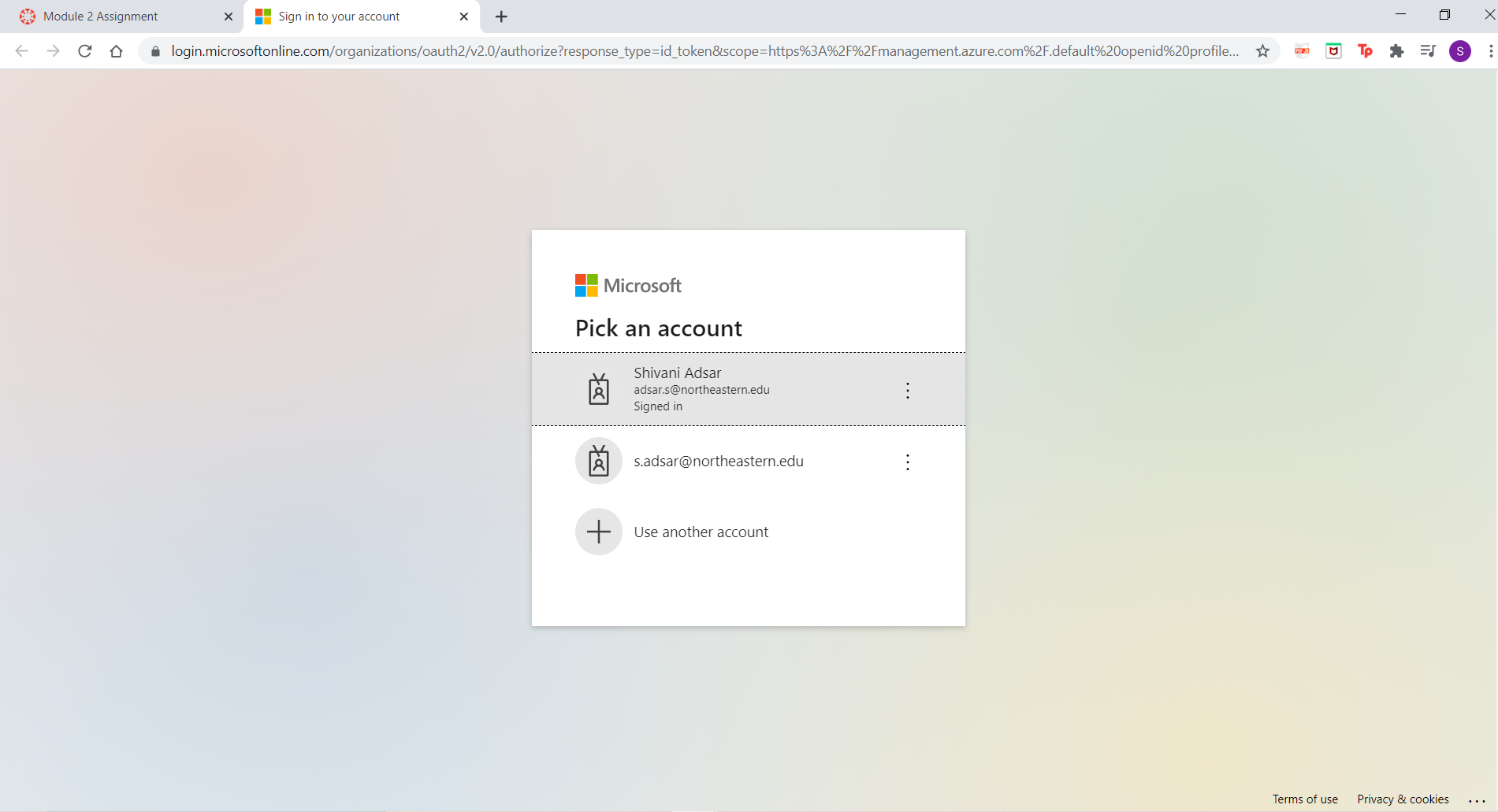
CONTACT: adsar.s@northeastern.edu

**Summary**

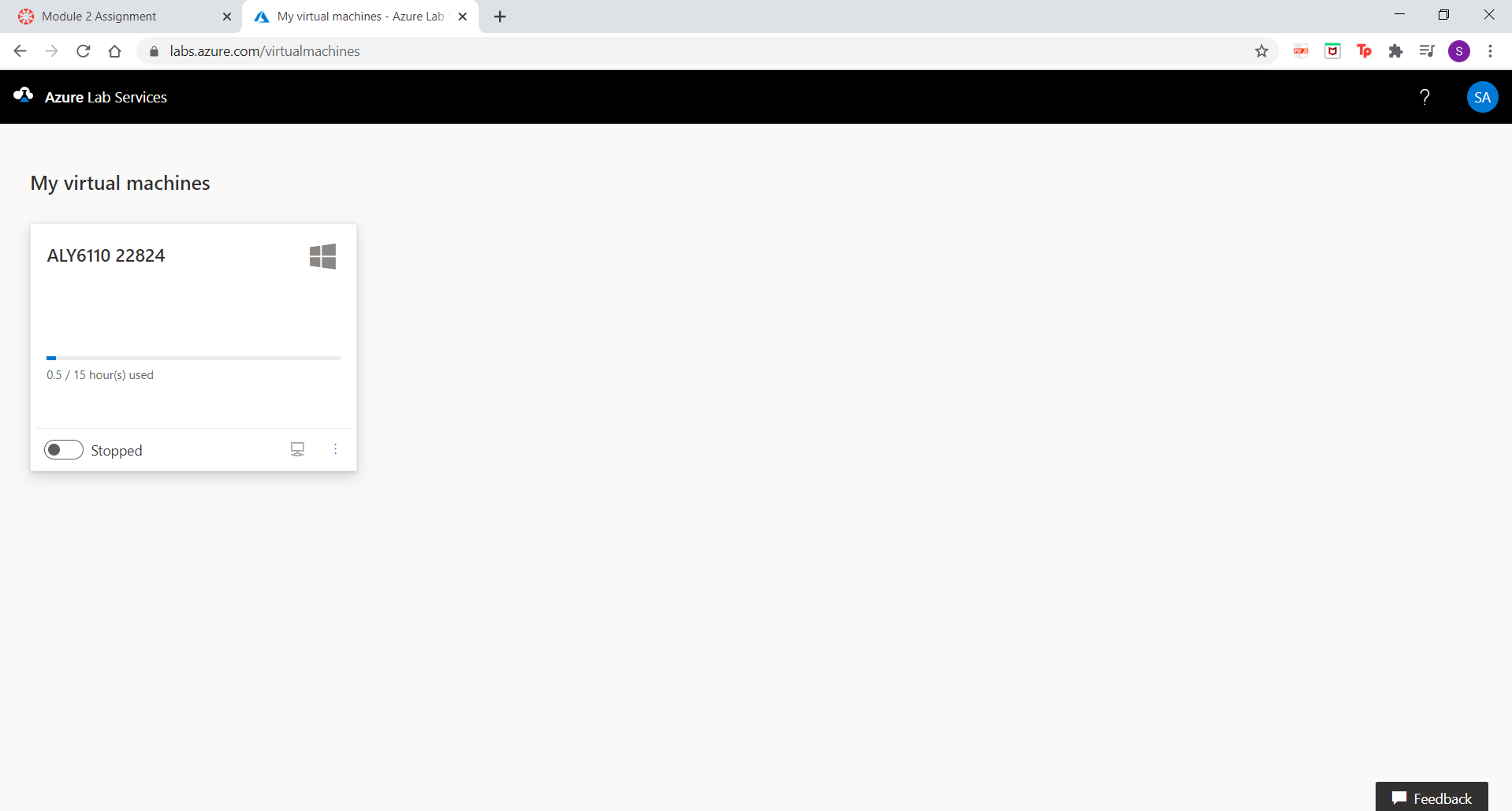
The assignment provides practical experience on Azure virtual machine and then, connecting to the Cloudera virtual machine. Moreover, the assignment provides exposure on Hadoop and querying on Impala using Hue on Cloudera. The Cloudera virtual machine helps in managing the data effectively for Hadoop platform. Nowadays, huge amounts of data is being generated in companies which needs efficient management, this can be performed using Cloudera framework as it is highly scalable. This assignment educated us in registering with the Microsoft Azure and then establishing a remote desktop connection. This ensured connectivity with Cloudera virtual machine, which helped in initiating connections with other Hadoop services. These services were tracked and utilized un their healthy state. Further, we logged into Hue and worked on the SQL queries using the Impala query editor. The Apache Impala works on parallel processing over SQL engine using Hadoop. This has helped us in understanding ways to perform analysis on big data using Cloudera virtual machine.

**Part 1**

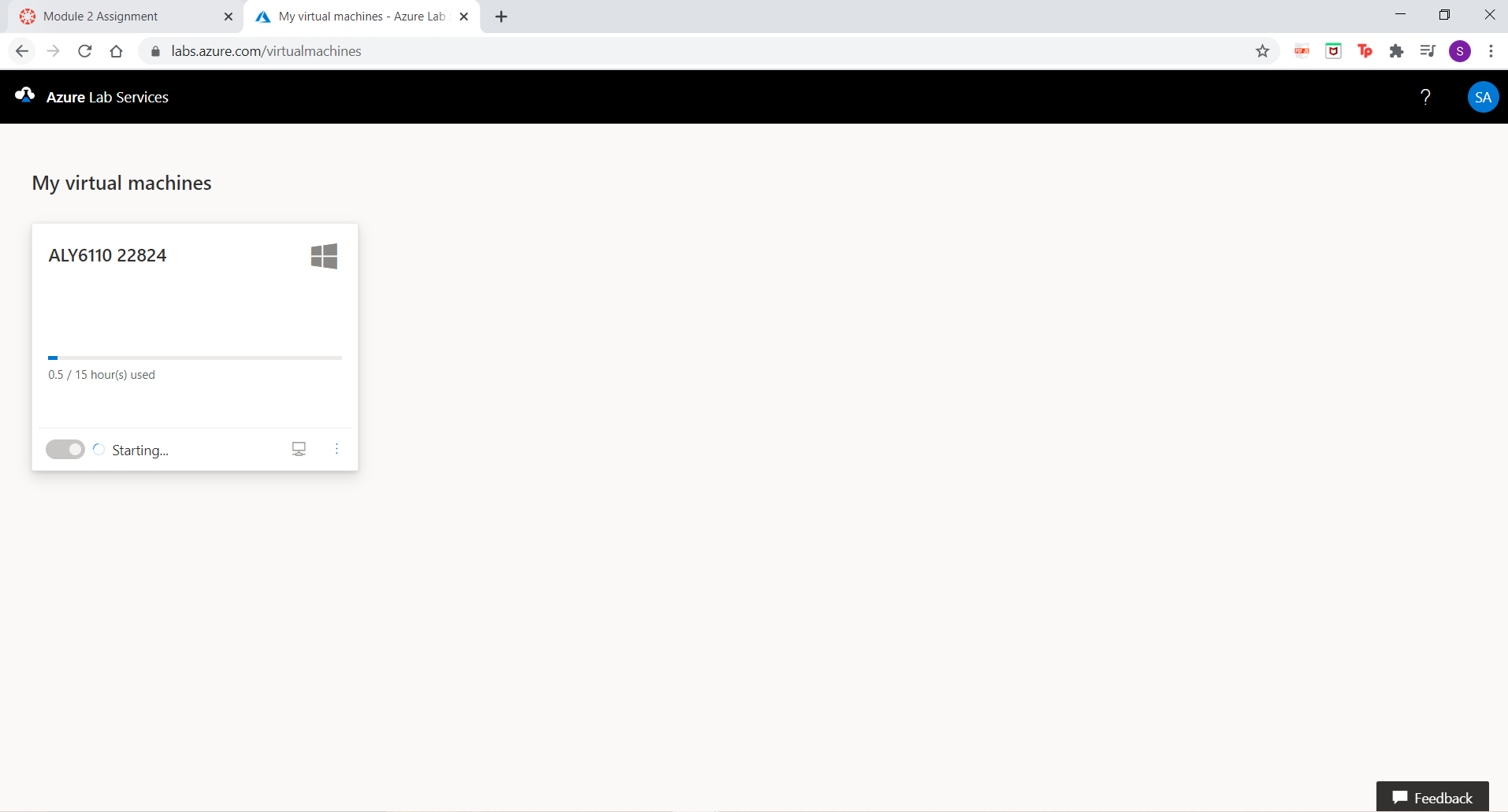
* We used the Microsoft environment through the link <https://labs.azure.com/register/r2ltwp63> for registration with north eastern credentials.



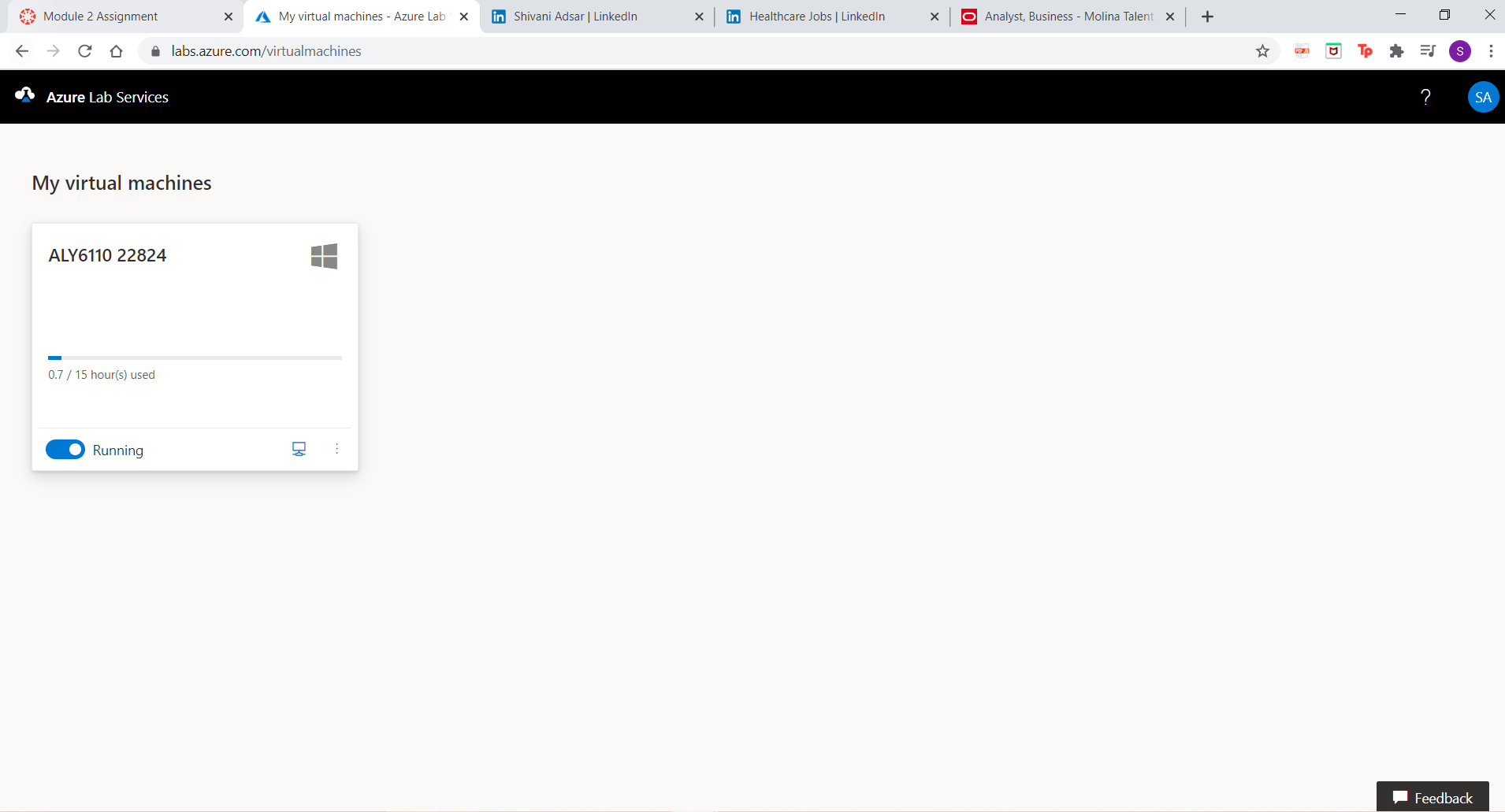
* We used the Azure Lab service environment, to access the virtual machine. This allows us to view the number utilized and then start or stop the virtual machine as required.



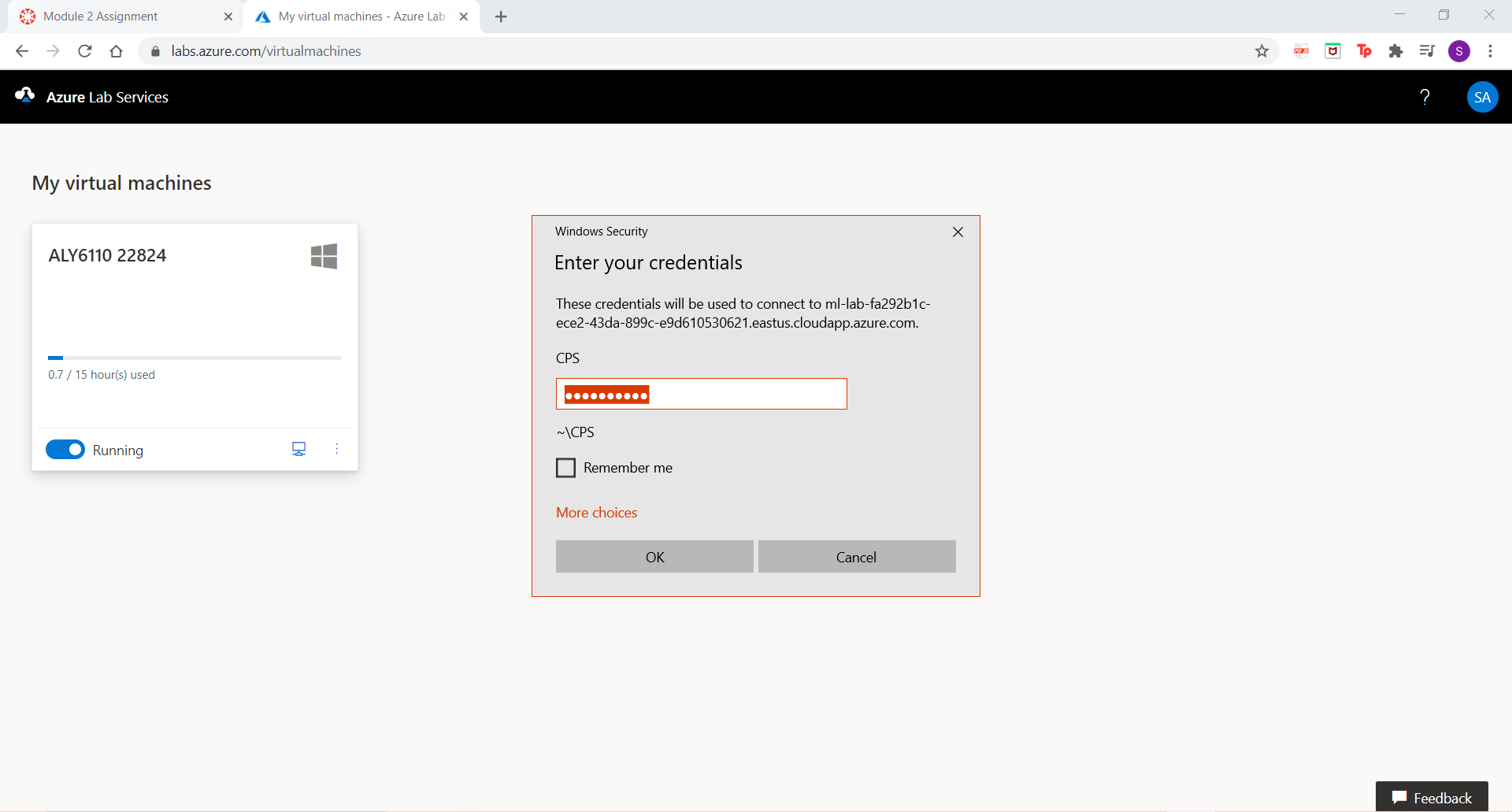
* The below screenshot shows the starting of the virtual machine after clicking the toggle bar.



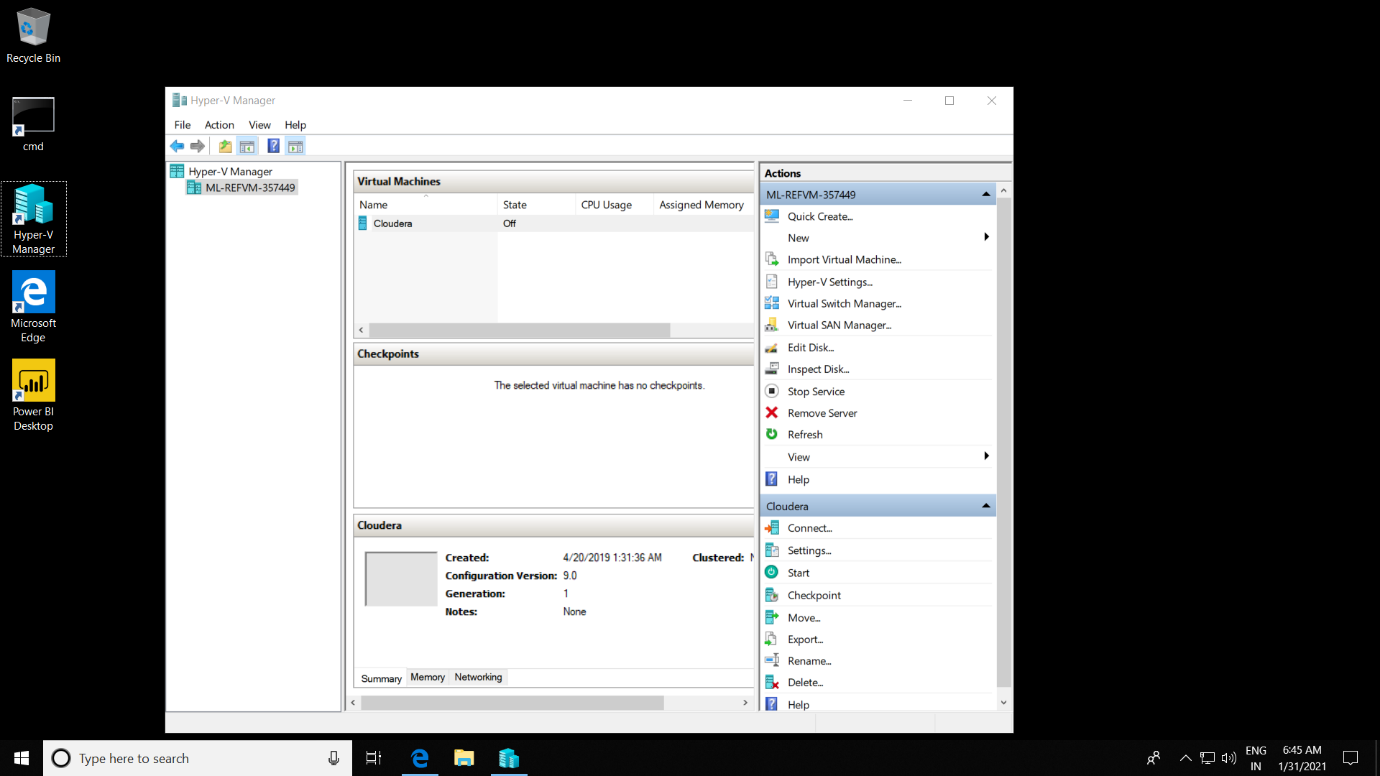
* After starting the virtual machine, the status will change from starting to running, and then, clicking on the monitor icon will download a RDP file which will help in connecting with the lab virtual machine. This RDP file will establish a remote desktop connection.



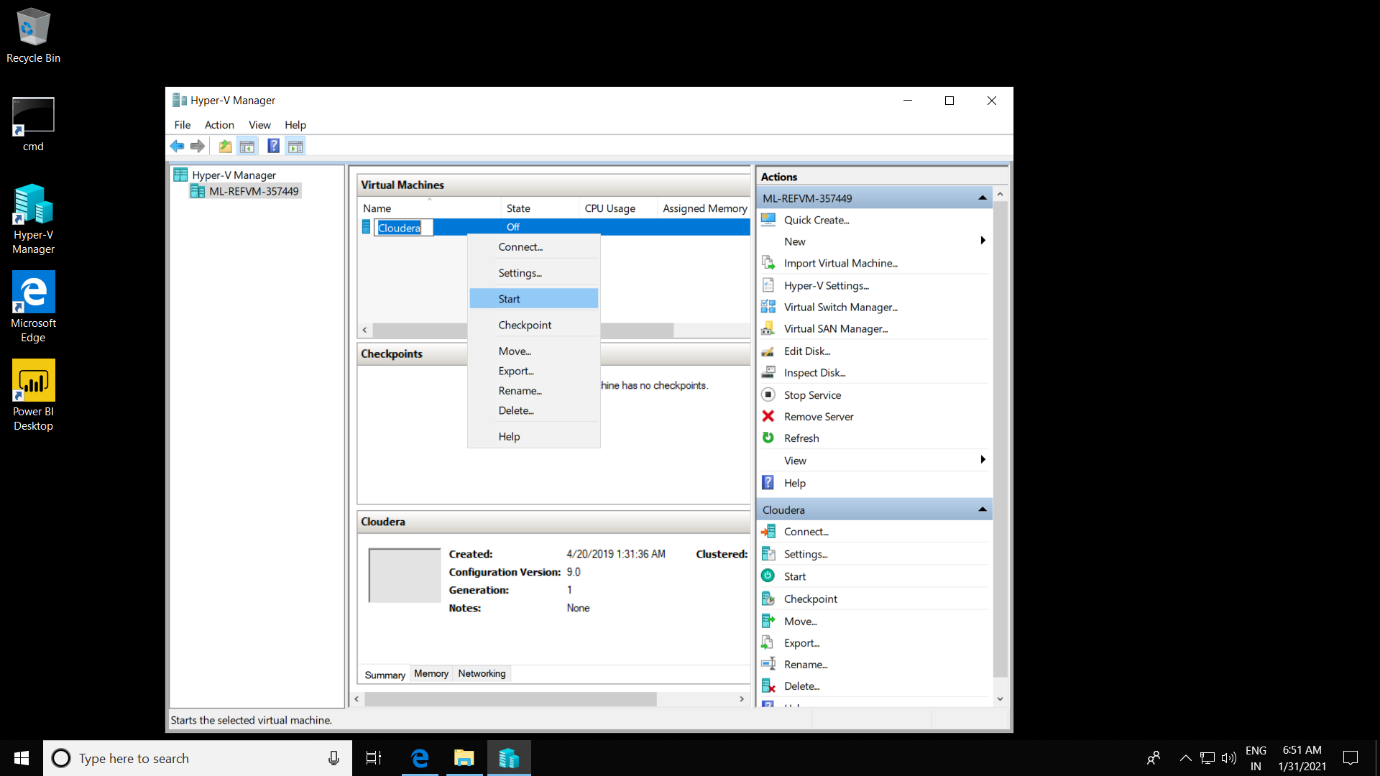
* When we run the RDP file, a dialog box will pop up for our credentials, and we would be using the password “ClassTest2” for “CPS” username.



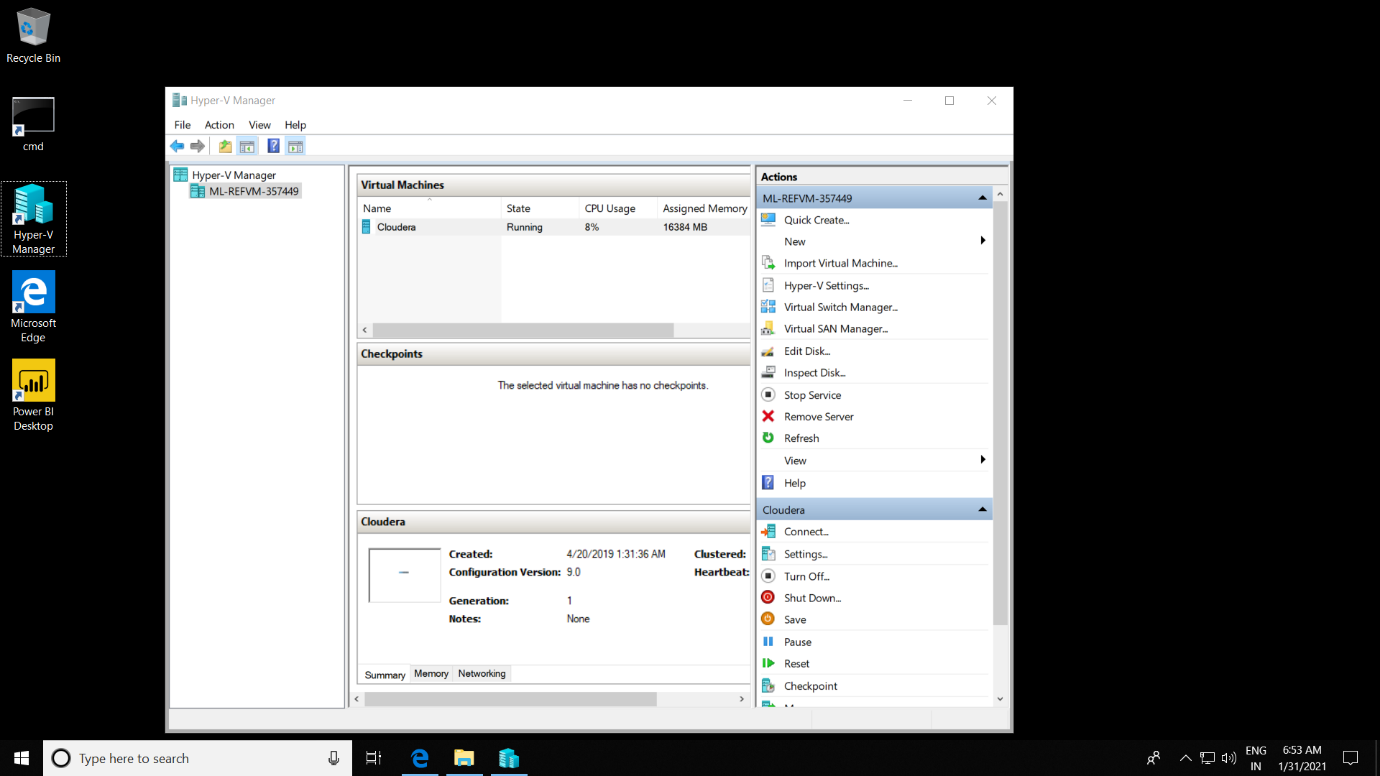
* This will open a virtual machine, where we click on the “Hyper-V Manager” which allows us to run a virtual environment. This gives access to Cloudera, which allows us to perform data analytics on cloud.



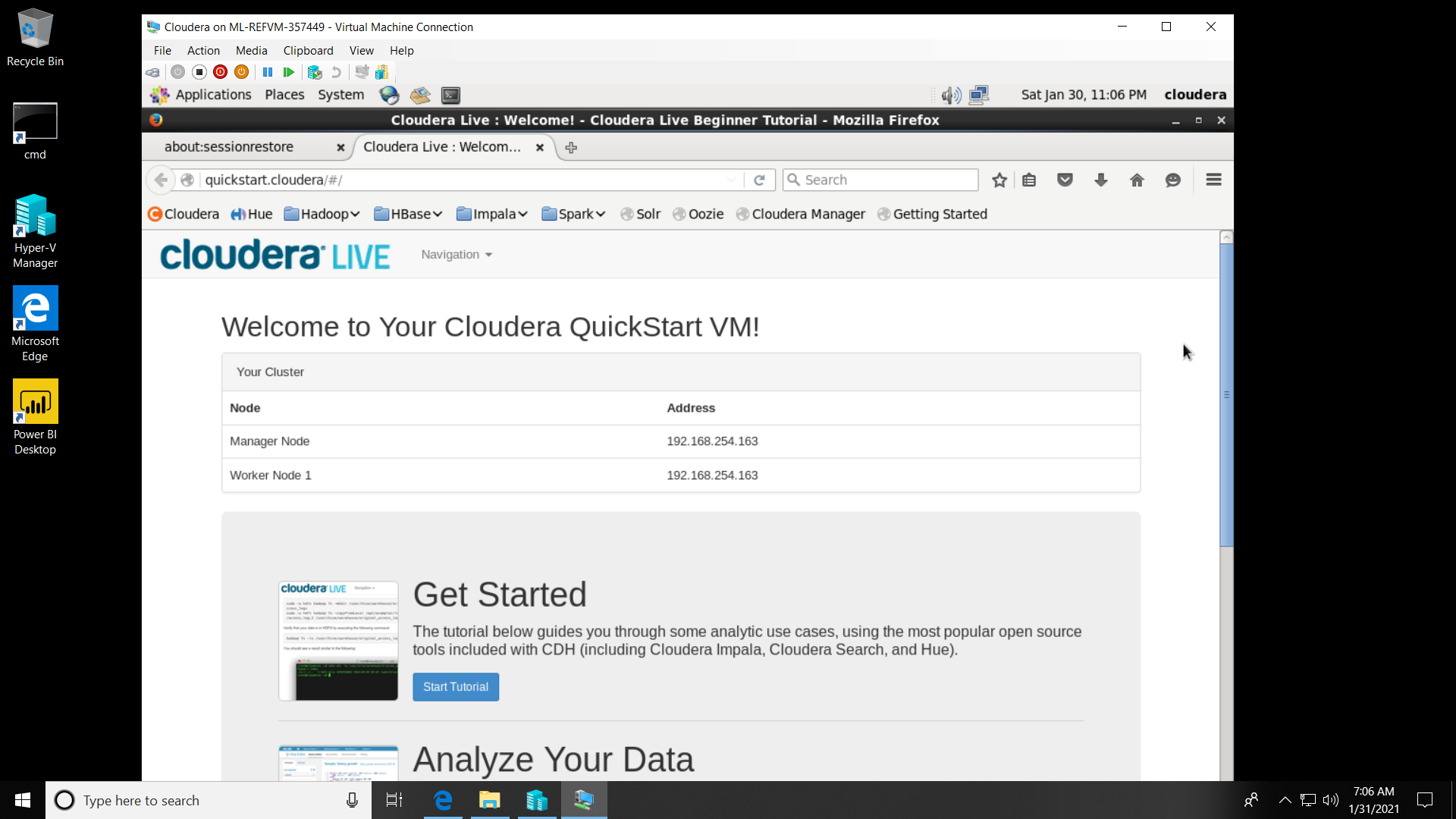
* Initially, the status for Cloudera is set to “Off”, which will require starting and therefore we select “Start” option. Further, we will select the “Connect” option for connecting with the Cloudera platform.



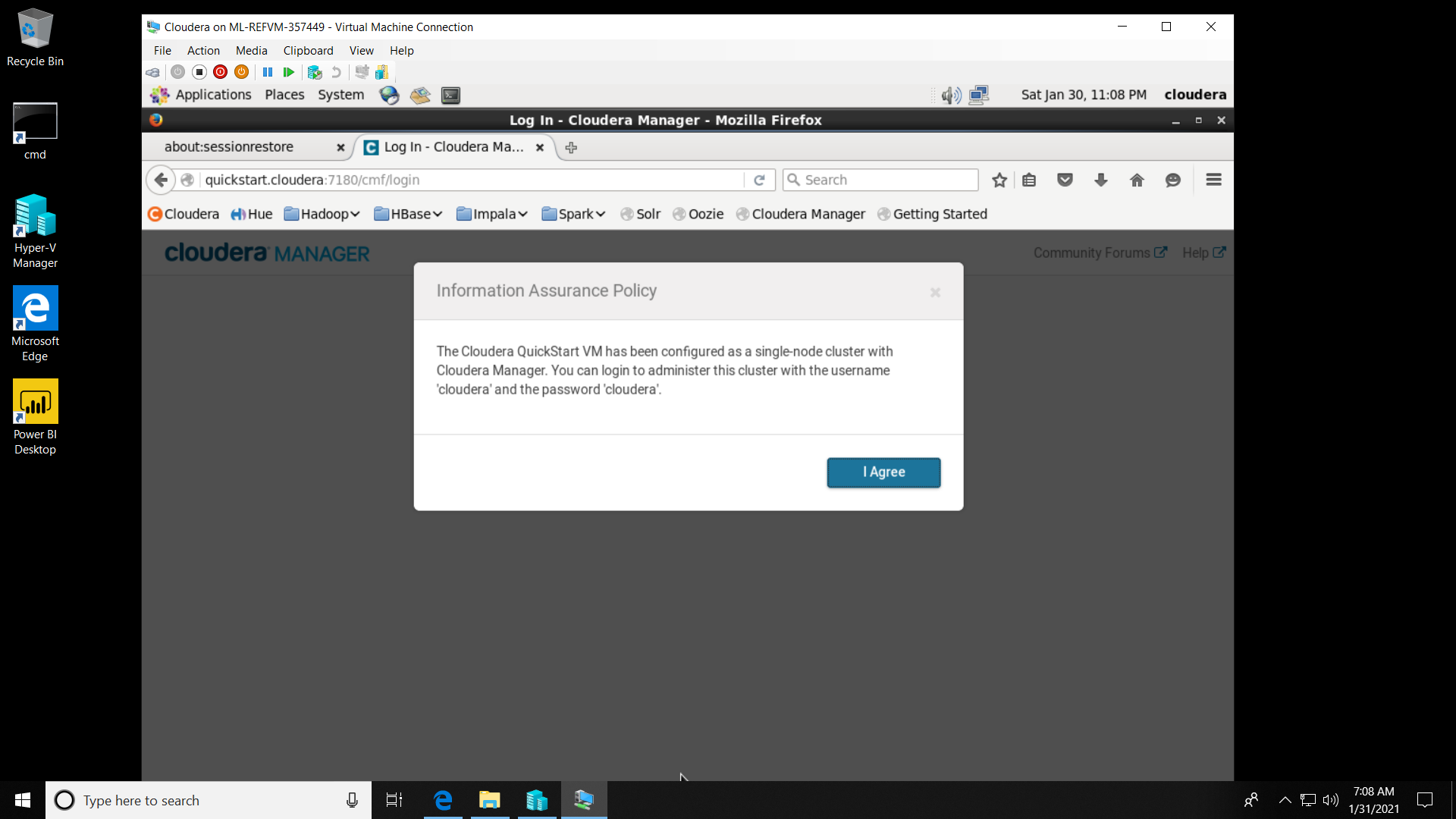
* After starting the Cloudera virtual machine, the status will change from “Off” to “Running” and the Cloudera Virtual Machine connection will open up.



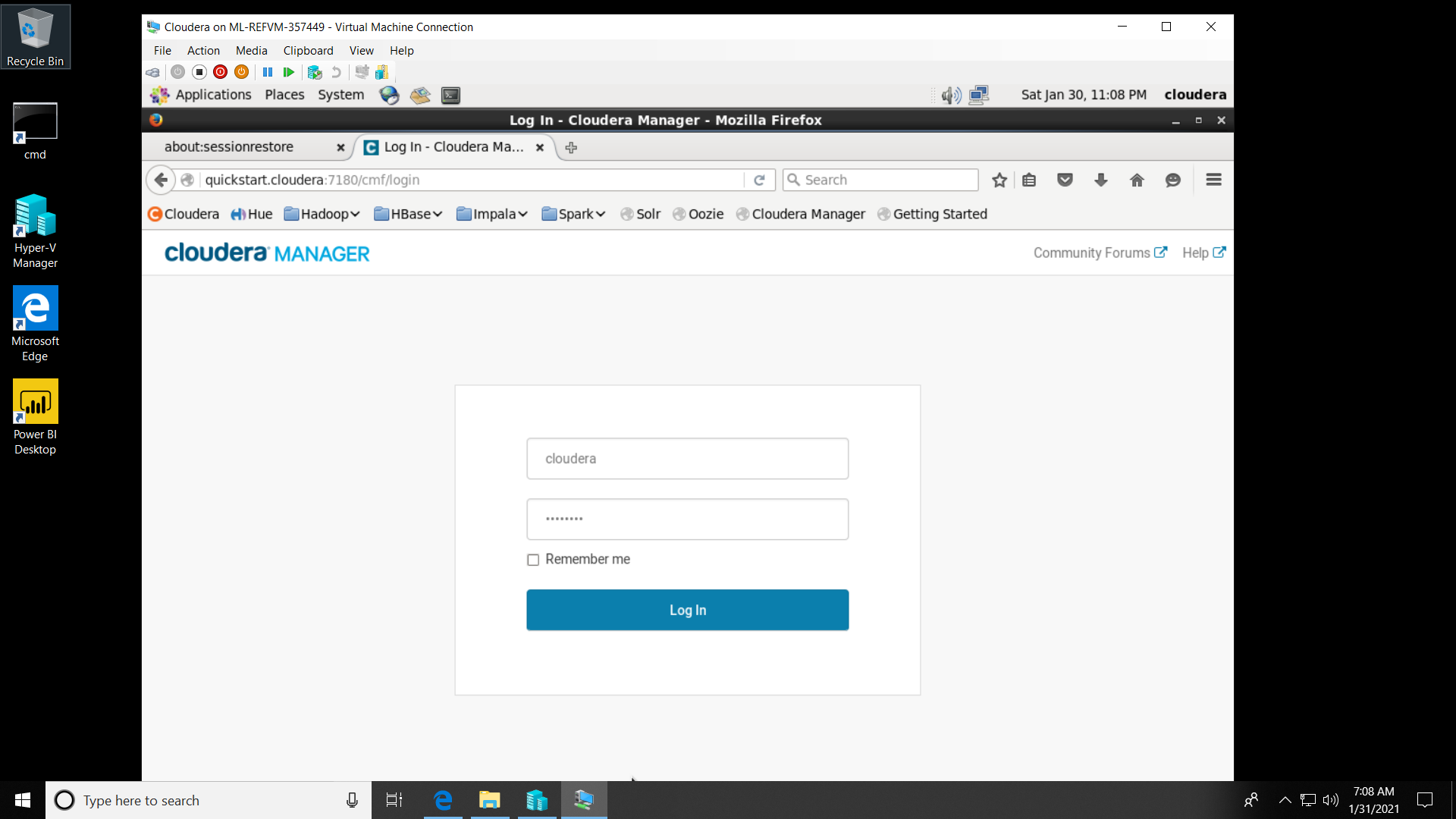
* The Cloudera virtual connection will help us in logging into the application.



* When we log into the Cloudera application, a dialog box will open up to agree the required Information Assurance Policy.

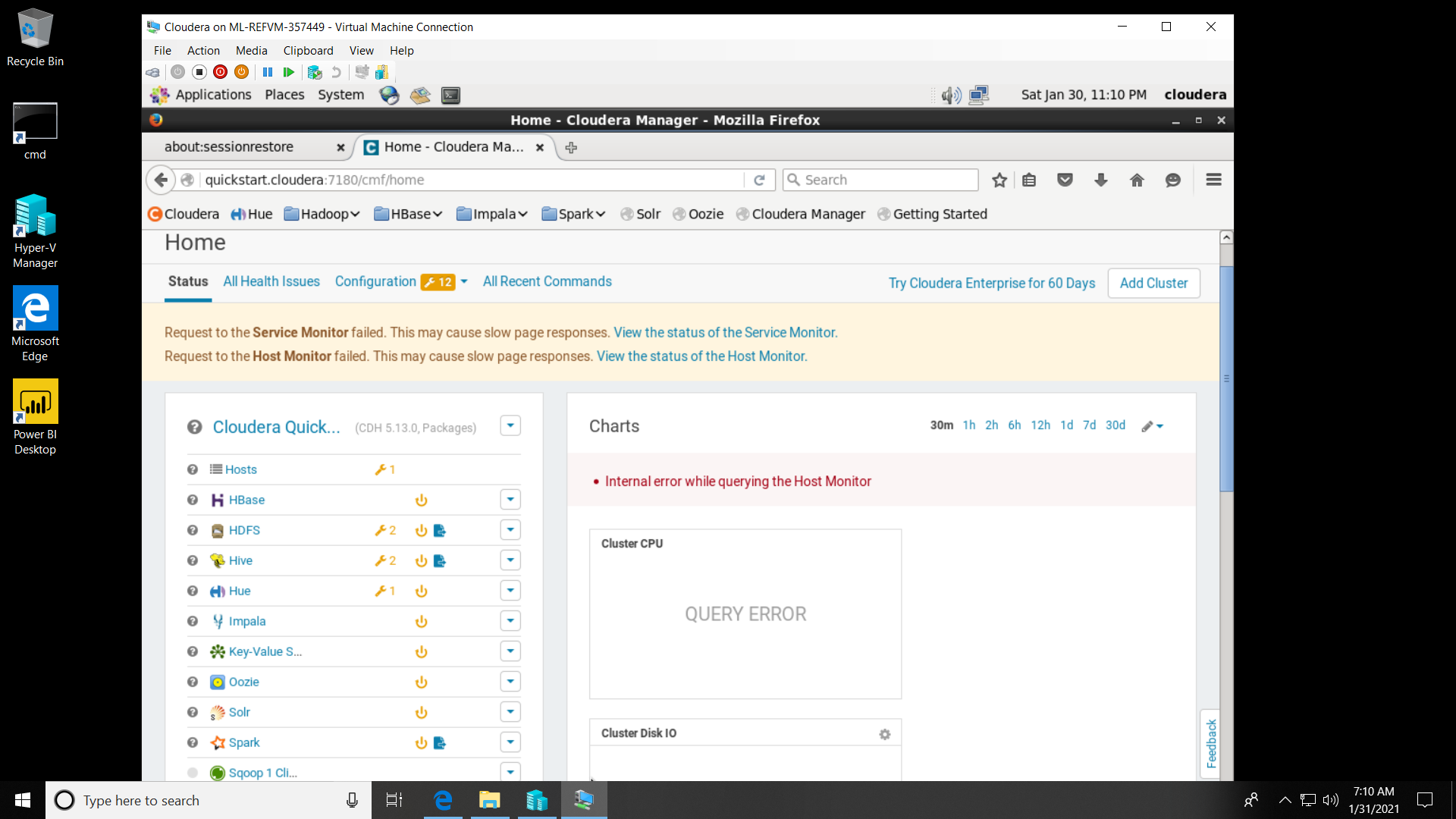


* After agreeing to the privacy policy, we will be prompted to add the required credentials to access the application. Therefore, we have used Cloudera as the username and password.

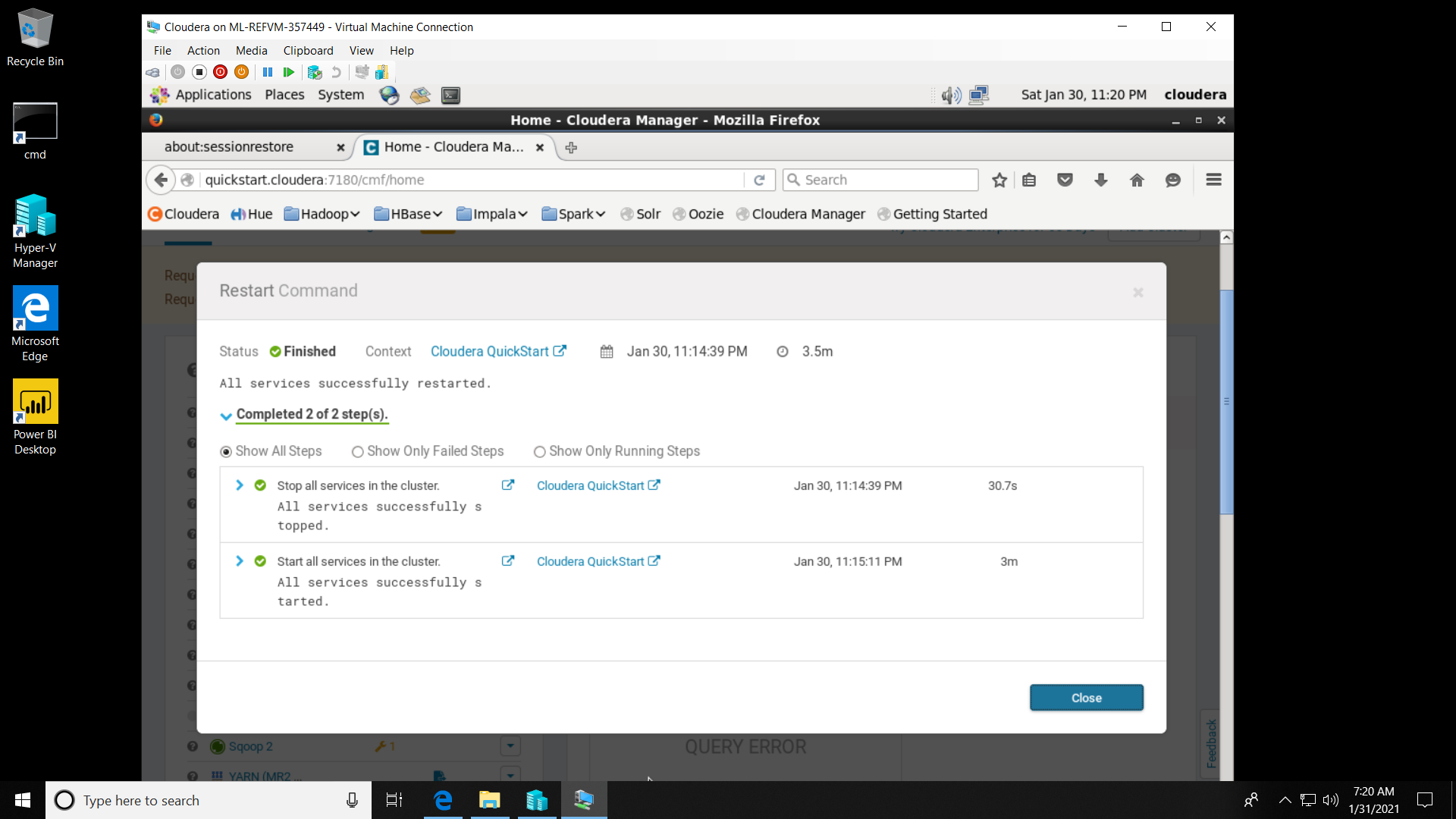


**Part 2**

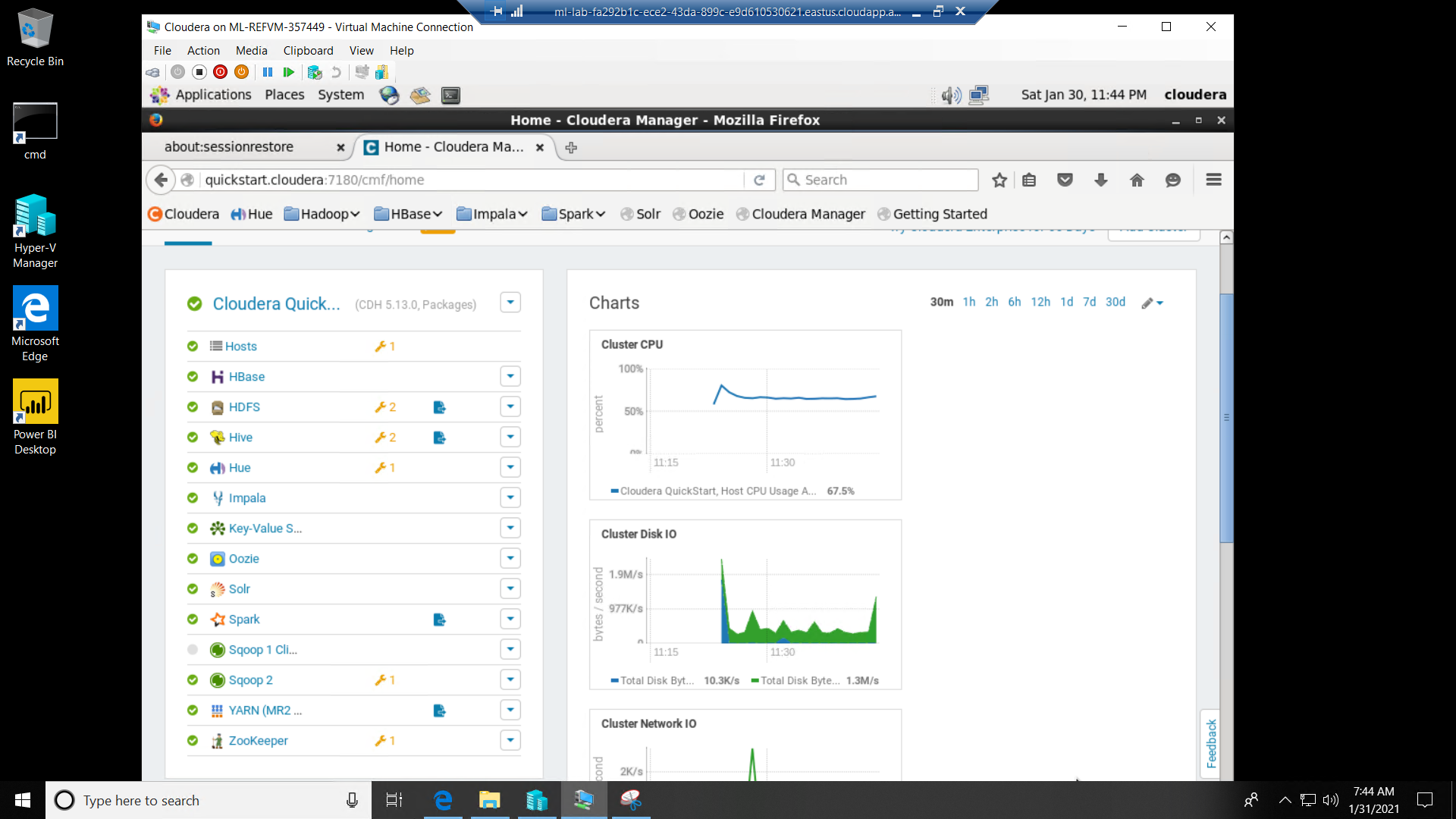
* When we log into Cloudera, and access the Cloudera Manager, it shows the health status of Hadoop services. These services show an internal error initially, therefore, we need to restart the services to re-initiate them.



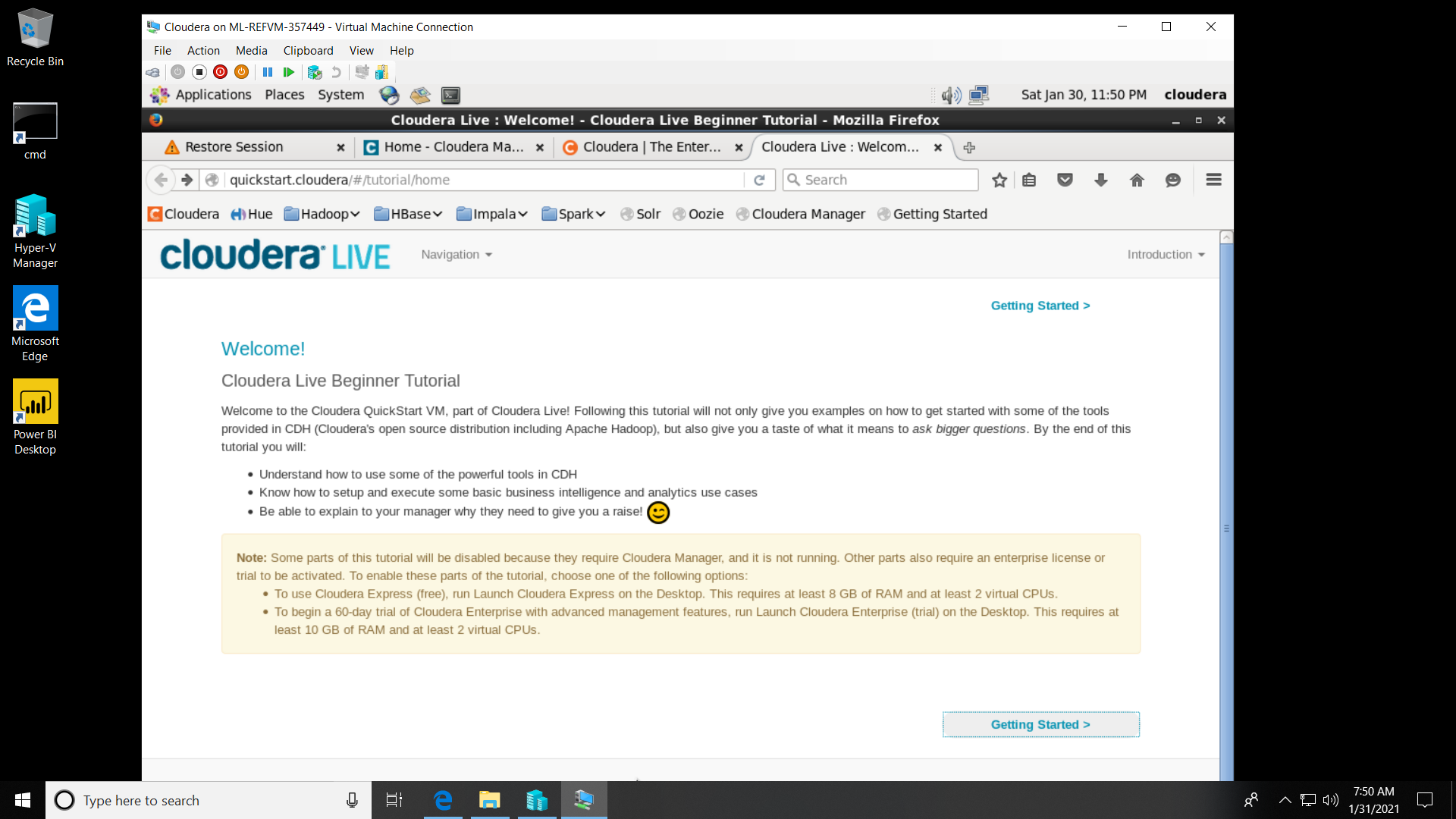
* The screenshot shows that Hadoop services have been restarted.



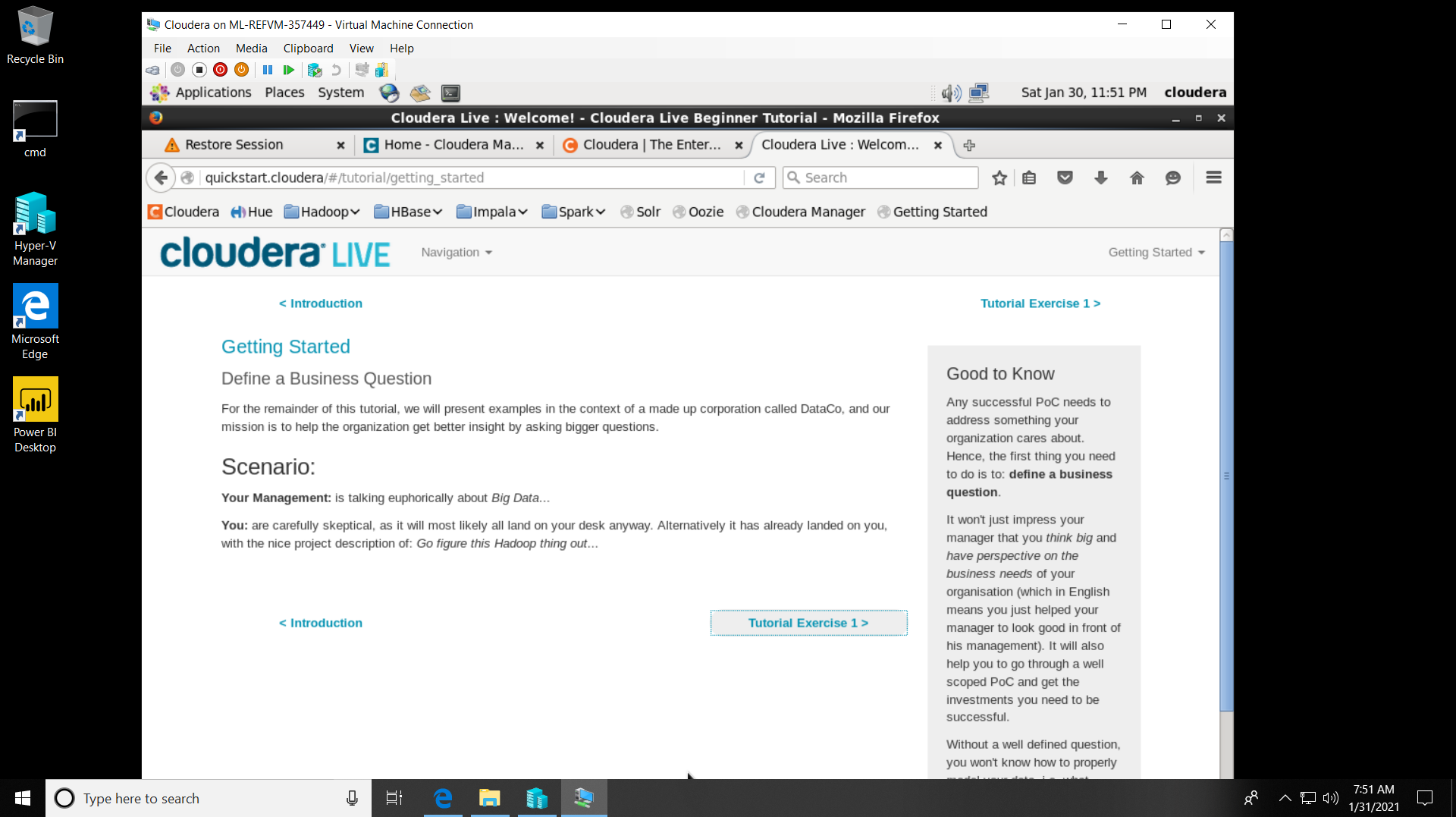
* The below graph shows the CPU and storage utilisation of various Hadoop services like Hive, Impala, Hue, HDFS, Spark etc., after restarting.



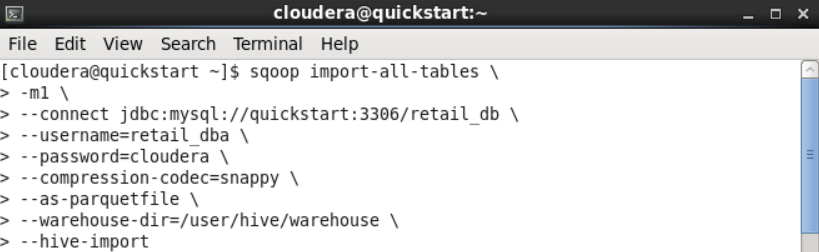
* In order to check the working of Hadoop services, we visit the Cloudera  <http://quickstart.cloudera/#/tutorial/home> website in the virtual machine. The below screenshot shows the Cloudera tutorial.



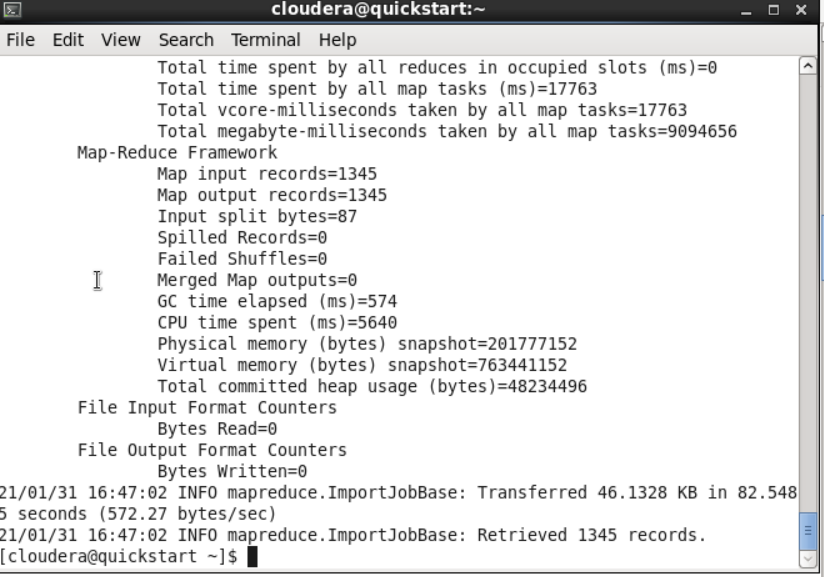
* As we click on “Getting Started”, we will view the next page for the tutorial, and then we would proceed to the Tutorial Exercise 1.



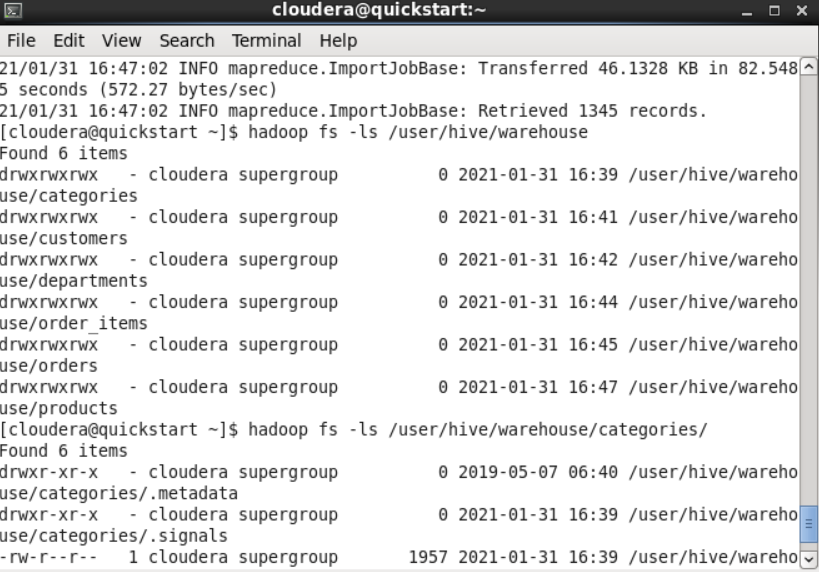
* In order to check if the Hadoop services are working, we will use some commands on the Cloudera terminal which is present in the top right and then open the sqoop job.



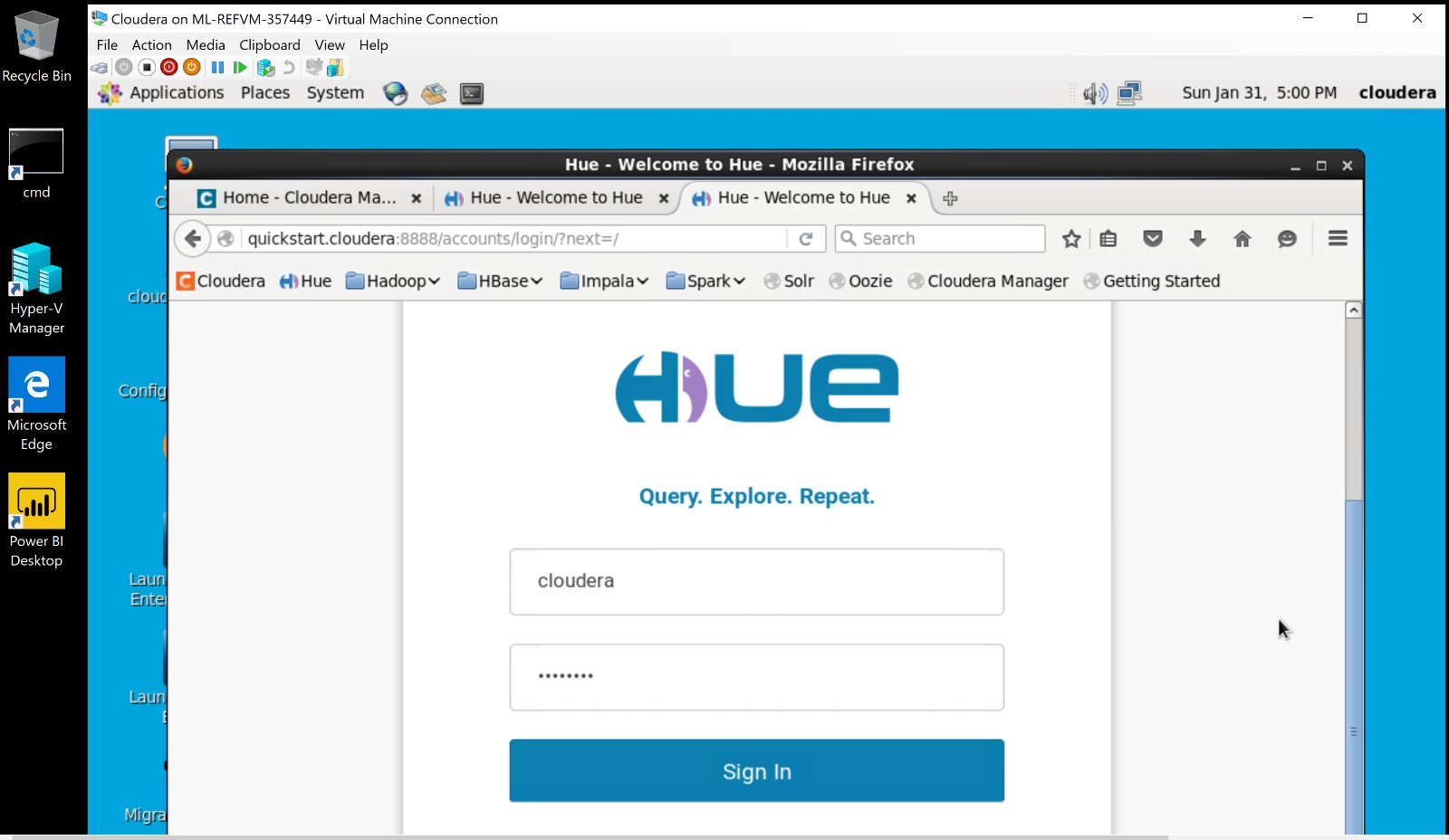
* This will start the MapReduce requests which will fetch data from the SQL server and produce the data to the HDFS.



* Further, we use the below commands to check if the data was loaded onto the HDFS file system.



The sqoop has created the tables which can be queried at the backend, and we are using the Impala tool present in Hue for querying purposes. When we launch Hue, we use the username and password as “cloudera” .

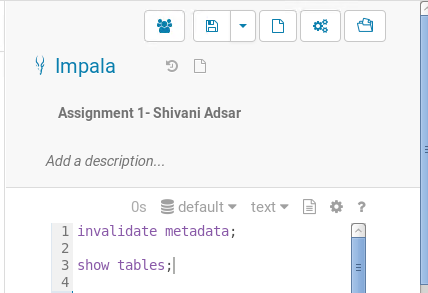
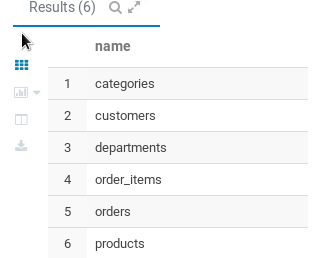


**Part 3**

As we navigate into the Hue, we will be able to select the Impala query editor which has helped in performing queries on the data. We have invalidated all the initial metadata present in impala and displayed the tables.

Code Explanation

invalidate metadata; -- *This will invalidate the pre-existing data*  
show tables; -- *To display tables in the database*

Query Output

**The 3rd most popular product category**

Code Explanation

select c.category\_name, count(order\_item\_quantity) as count -- *This creates objects and selects variable names, count function will count the items*

from order\_items oi -- *This selects the table name*

inner join products p on oi.order\_item\_product\_id = p.product\_id

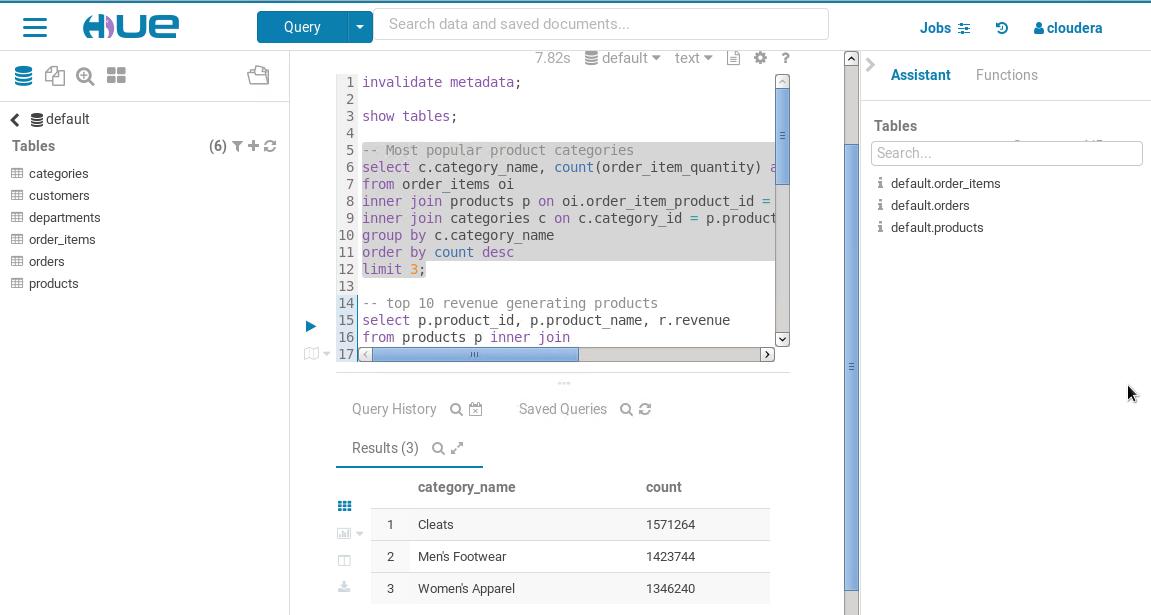
--*The inner join is used to join the order\_item\_product\_id in order\_items table with the product\_id in the products table*

inner join categories c on category\_id = p.product\_category\_id -- *Since we need to display the category name, we have used inner join on category\_id in the categories table with product\_category\_id in the products table*

group by c.category\_name -- *This groups results by the category name*

order by count desc -- *This will display the count in the descending order*

limit 3; -- *To limit our results to three items*



As seen in the above query and its output, the 3rd most popular product category is Women’s Apparel and the count for the category is 1346240.

**Revenue earned through Nike Men's Free 5.0+ running shoe**

The below query and its output shows the revenue earned through Nike Men's Free 5.0+ running shoe. It seems that the revenue earned was 56120787.3077 . We have used the inner join for joining the “products” and “order\_items” tables. The inner join has been applied on the “order\_item\_order\_id” and “order\_id” fields present in the “order\_items” table. In order to retrieve the revenue for the Nike Men’s shoe, we have applied a filter on the “product\_name” .

Code Explanation

select distinct(p.product\_id), p.product\_name, r.revenue -- *This creates objects for every table and selects distinct product\_id, product\_name and revenue.*

from products p inner join --*This fetches data from the “products” table using inner join*

(select oi.order\_item\_product\_id, sum(cast(oi.order\_item\_subtotal as float)) as revenue --*This will select order\_item\_product\_id, and convert order\_item\_subtotal to float and display their sum as a revenue field.*

from order\_items oi inner join orders o --*This inner join will join the order\_item\_order\_id* *in order\_items table with order\_id in*

on oi.order\_item\_order\_id = o.order\_id *the orders table*

where o.order\_status <> ‘CANCELED’ --*This filters results which do not have order\_status as “Canceled”*

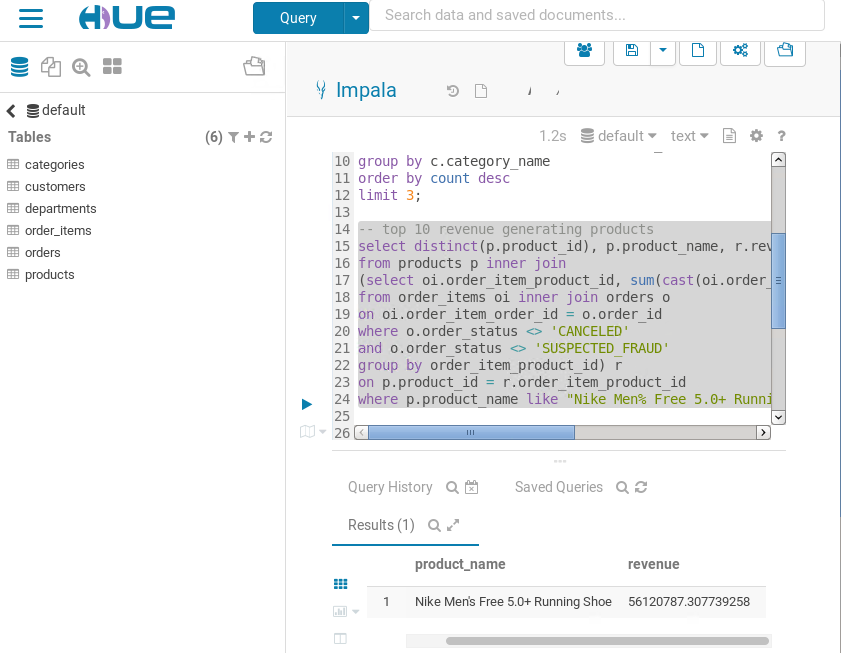
and o.order\_status <> ‘SUSPECTED\_FRAUD’ --*This filters results which do not have order\_status as “Suspected Fraud”*

group by order\_item\_product\_id) r *--This will group the results by order\_item\_product\_id*

on p.product\_id = r.order\_item\_product\_id *-- Then , joins the results using product\_id on order\_item\_product\_id*

where p.product\_name like “Nike Men's Free 5.0+ running shoe” *--This will filter the results for the specific product name*

limit 10; -- *This will display the first 10 records*



**Product name with the least revenue and the revenue of the products**

Since, we require all the products with least revenue, we have ordered the revenue by ascending order. Also, we have applied the distinct function on the “product\_id” to display unique values.

It seems like “Bushnell Pro X7 Jolt Slope Rangefinder” has the least revenue, with a revenue of 76798.718 .

Code Explanation

select distinct(p.product\_id), p.product\_name, r.revenue -- *This creates objects for every table and selects distinct product\_id, product\_name and revenue.*

from products p inner join --*This fetches data from the “products” table using inner join*

(select oi.order\_item\_product\_id, sum(cast(oi.order\_item\_subtotal as float)) as revenue --*This will select order\_item\_product\_id, and convert order\_item\_subtotal to float and display their sum as a revenue field.*

from order\_items oi inner join orders o --*This inner join will join the order\_item\_order\_id* *in order\_items table with order\_id in*

on oi.order\_item\_order\_id = o.order\_id *the orders table*

where o.order\_status <> ‘CANCELED’ --*This filters results which do not have order\_status as “Canceled”*

and o.order\_status <> ‘SUSPECTED\_FRAUD’ --*This filters results which do not have order\_status as “Suspected Fraud”*

group by order\_item\_product\_id) r *--This will group the results by order\_item\_product\_id*

on p.product\_id = r.order\_item\_product\_id *-- Then , joins the results using product\_id on order\_item\_product\_id*

order by r.revenue asc *-- This will order the results by the revenue in the ascending order*



**Analysis**

**Pros**

The Cloudera is a very flexible and well data integrated platform for managing huge volumes of data. It helps us in managing Apache Hadoop services, in analyzing the data while keeping data integrity intact. Moreover, the platform provides some of the crucial Hadoop tools including:

Cloudera Impala, which is an parallel processing interactive platform for running SQL queries seamlessly.

Cloudera Manager- it is a web interface which provides accessibility to monitor, track and repair issues using APIs and further maintain health information of the Hadoop services. Moreover, we can track our deployed data.

The Cloudera Virtual Machine is very beneficial as it allows us to run multiple operating systems at the same time.

**Cons**

The Cloudera Manager is a great platform to manage our data on Hadoop. However, there are some disadvantages:

The installation process is time consuming and sophisticated. Moreover, the Impala editor has a very slow functioning. Also, on restarting the virtual machine, there have been connectivity issues which causes delay in processing. Therefore, every time we restart the virtual machine, we need to reinitiate and repair the Hadoop services for maintaining their good health status within the platform.

In addition, it is cumbersome to perform the copy-paste of some data between the virtual machine and our operating system.

**Conclusion**

We have learnt about Microsoft Azure, which helped in understanding about the usability of the platform in performing big data analytics on large data. The Azure virtual machine platform is very flexible, cloud analytics platform which enables us to implement cloud analytics, without purchasing the licensed version. The virtual machine allows us to create virtual machines through the images, and then host services on these virtual machines. The platform allows us to maintain a backup of our data in case the system hardware fails. Since, we have used the Cloudera platform using Microsoft Azure, it can be noted that, Azure was able to host the Cloudera platform within its virtual machine thereby, enabling us to access Hadoop services within Cloudera such as Hive, Oozie, MapReduce, Impala etc. Moreover, using Cloudera, we were able to access Hue which is a web interface to Hadoop, and enabled us to implement SQL queries on the impala editor. The virtualization and cloud technology has helped businesses prosper and companies would continue to use these hosting services in future. The Microsoft Azure is a very scalable platform to explore cloud analytics, data warehousing, data integration and Security. The feedback for the virtual machine platform would be that it lacks computational power and has slow performance. Also, some of the findings that can be analyzed through the SQL implementation, is that, “Bushnell Pro X7 Jolt Slope Rangefinder” has the least revenue. In addition, most popular product category amongst the customers, is Cleats, followed by Men’s footwear and Women’s Apparel. The scope for such big data platforms like Microsoft azure is tremendous in the software industry as huge amounts of data is being generated every day.

**References**

[1] Everything you ever wanted to know about Microsoft Azure. (2020, February 21). Retrieved February 03, 2021, from <https://www.nigelfrank.com/blog/everything-you-ever-wanted-to-know-about-microsoft-azure/>

[2] Microsoft Azure explained: What it is and how to use It: Ccb tecnology. (2020, June 18). Retrieved February 03, 2021, from https://ccbtechnology.com/what-microsoft-azure-is-and-why-it-matters/