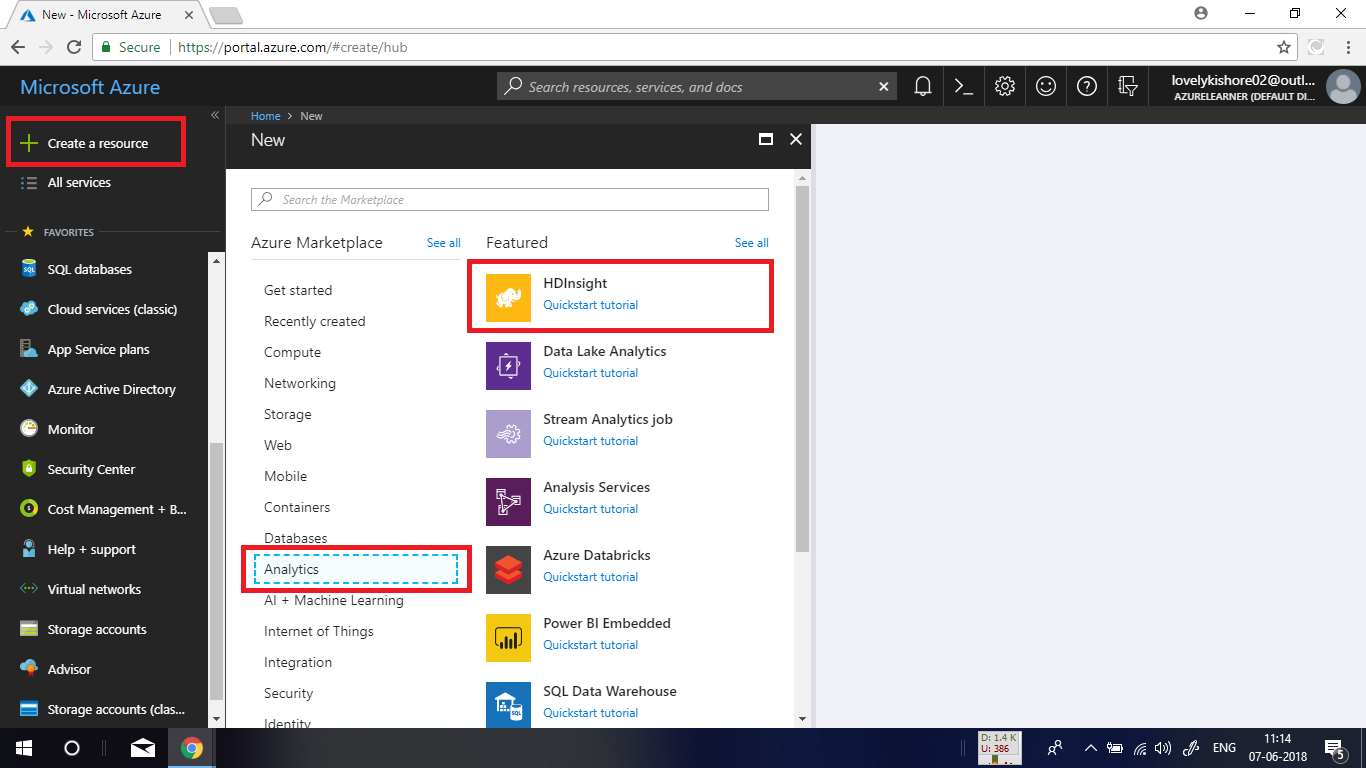
# **Creating Spark Cluster in HDInsight using Azure Portal**

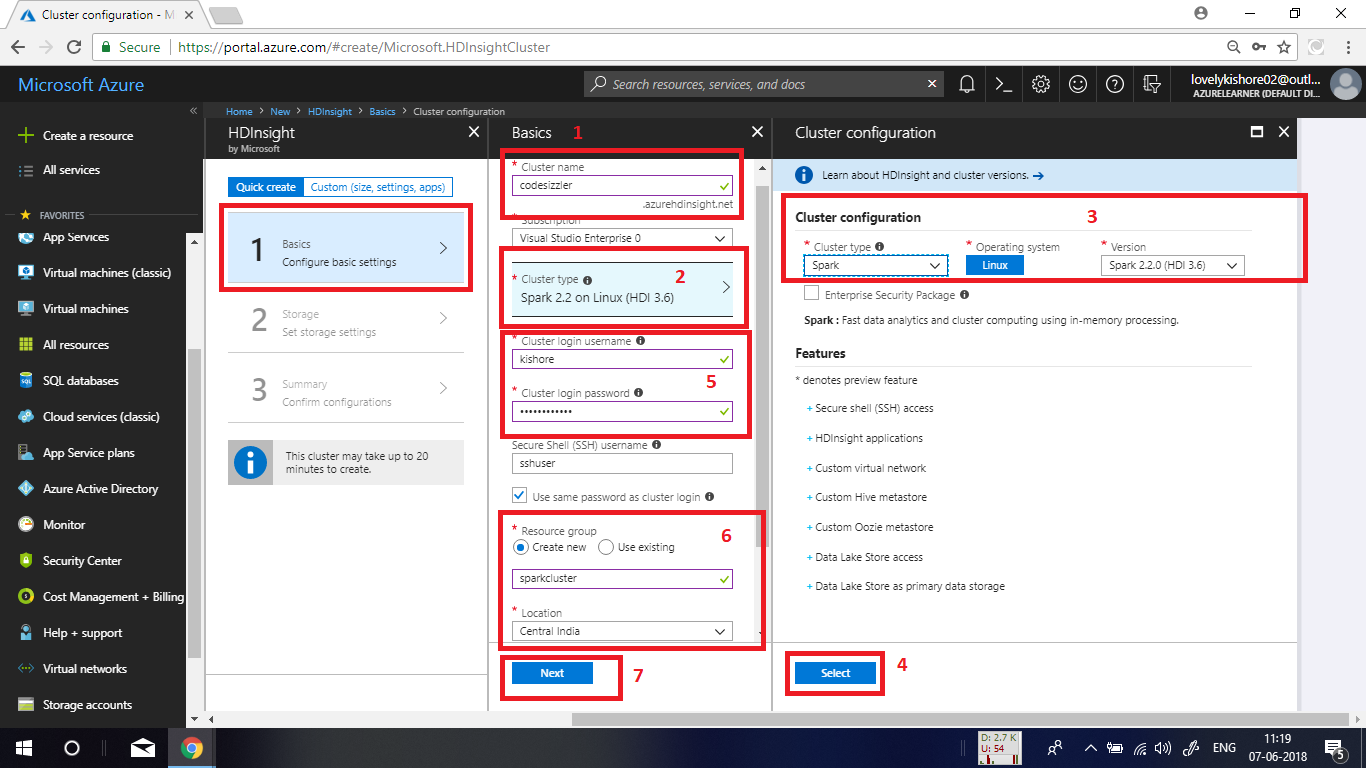
This demo compromises the method of creating Spark Cluster in HDInsight using Azure Portal and deploying Jupyter notebook in it.

**Creating HDInsight and Configuring Spark Cluster:**

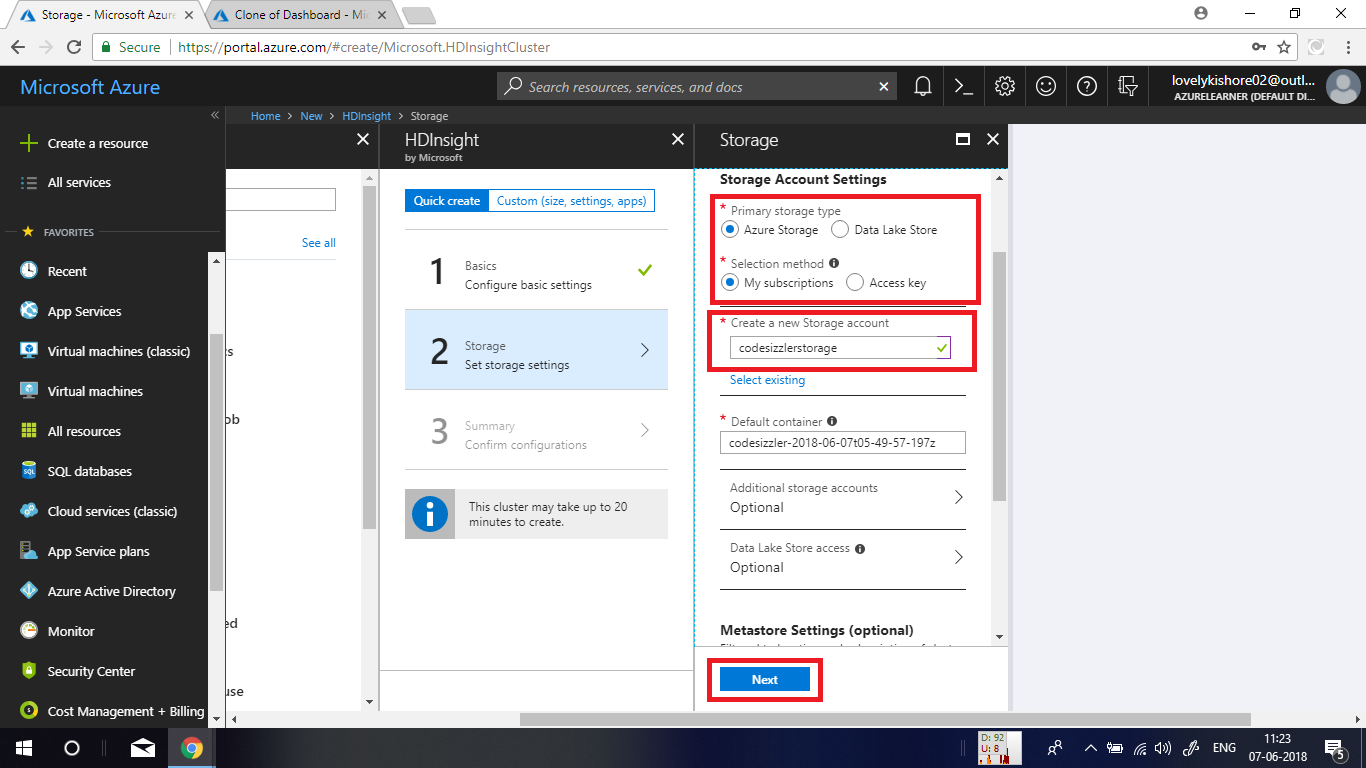
Go to **+ Create Resource -> Analytics -> HDInsight**.



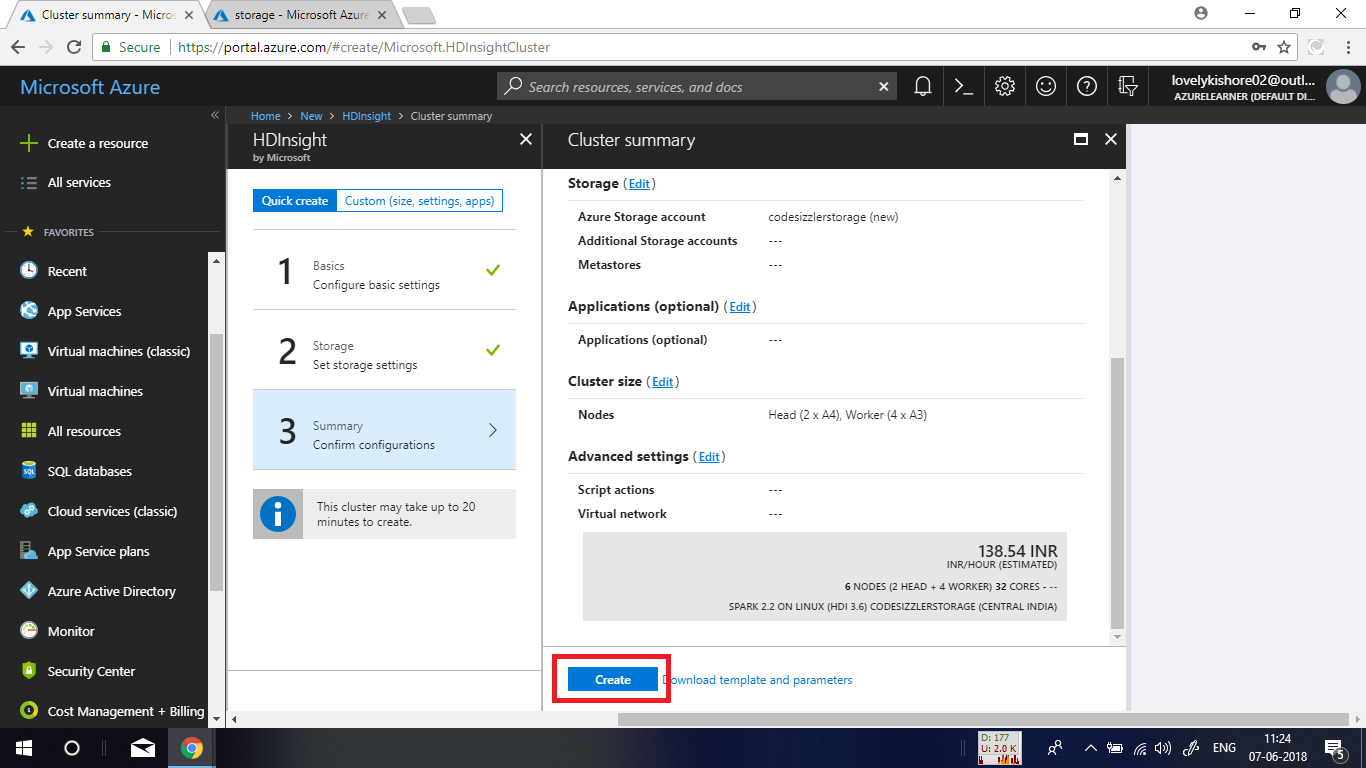
In the basic settings blade, give a name for **cluster.** In the cluster type, choose **Spark** from the drop-down list and select latest version of it and click on **Select** button. Give an **Username** and **Password.** Choose a resource group and click on **Next.**



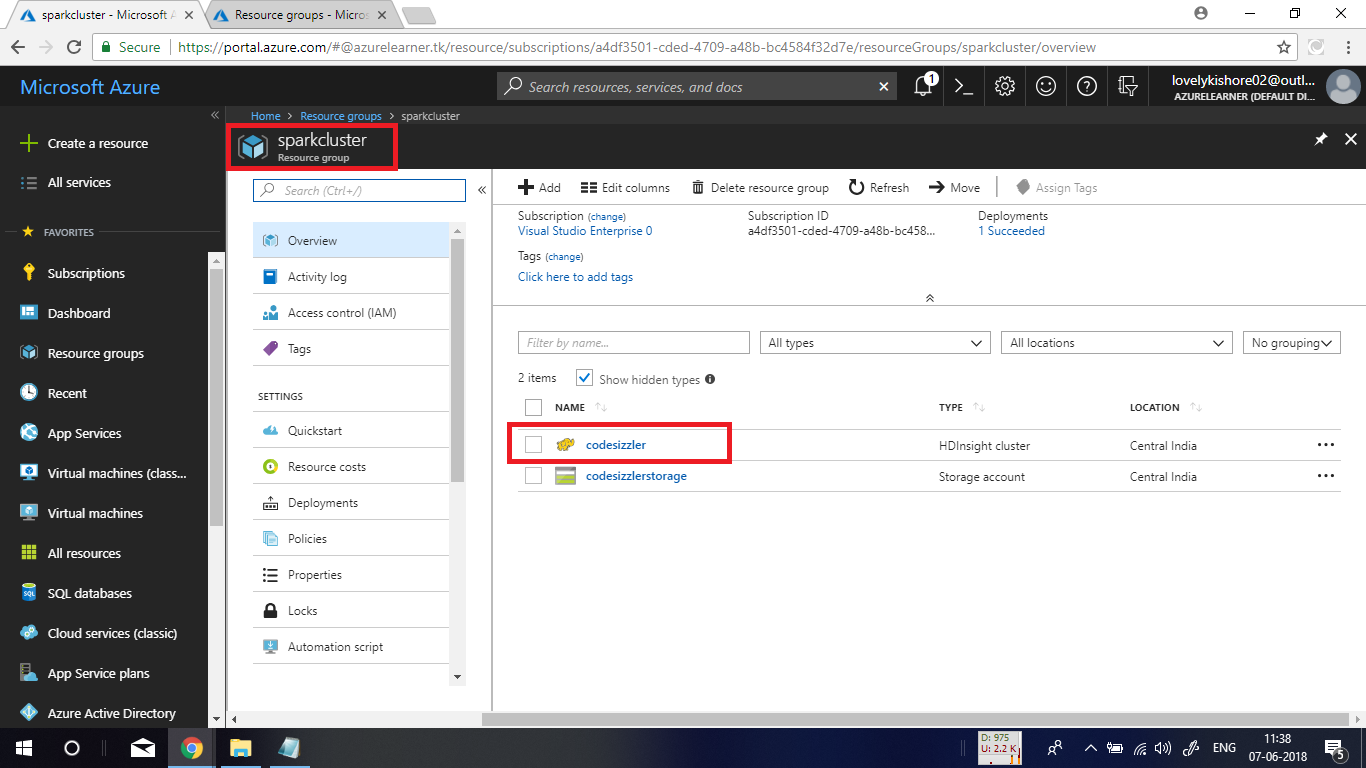
In the storage settings blade, choose **Primary Storage Type** as **Azure Storage** and **Selection Method** as **My Subscriptions.** Choose an existing storage account or create new by giving an unique name for it. Let the container name be default in there. Leave all other settings as default and click on **Next** button.



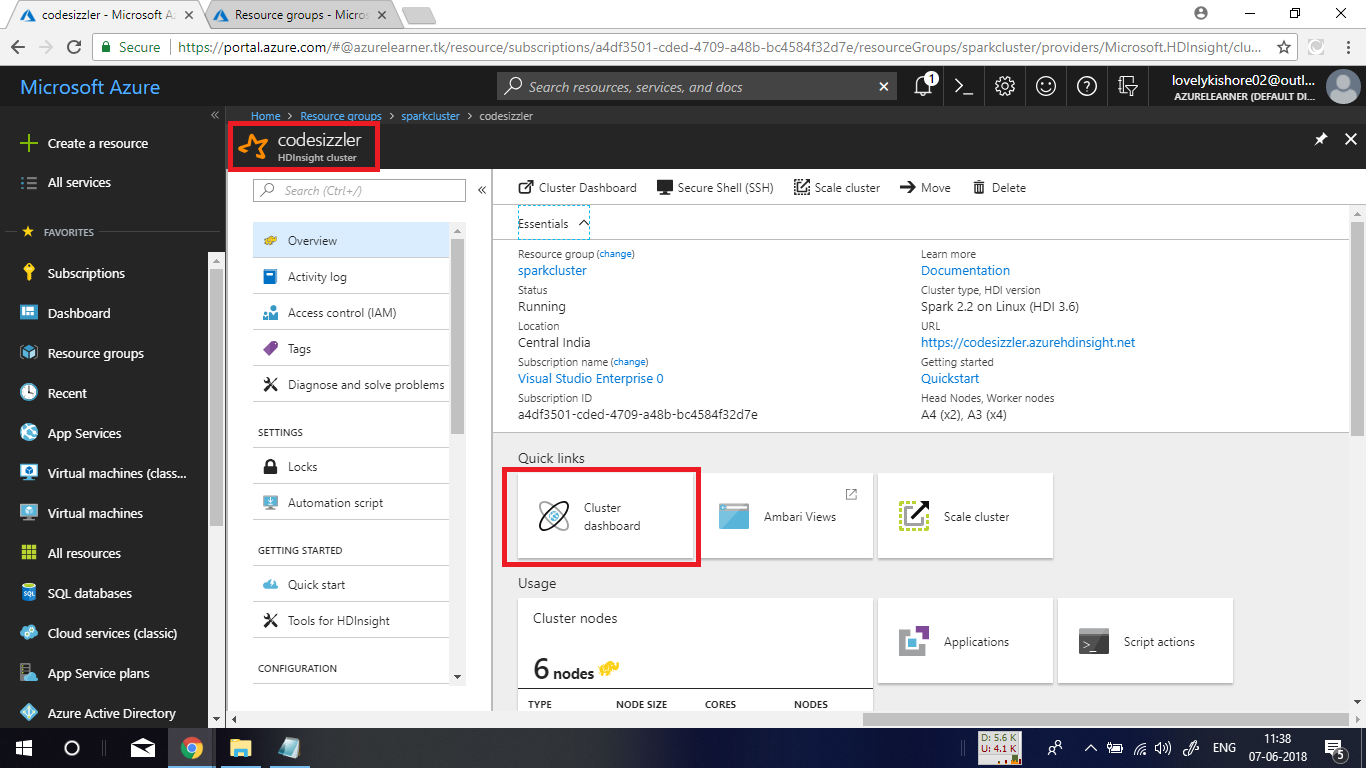
At last in the summary blade, click on **Create** button to deploy the cluster. The deployment will take few minutes. So, you need to wait for it to deploy completely.

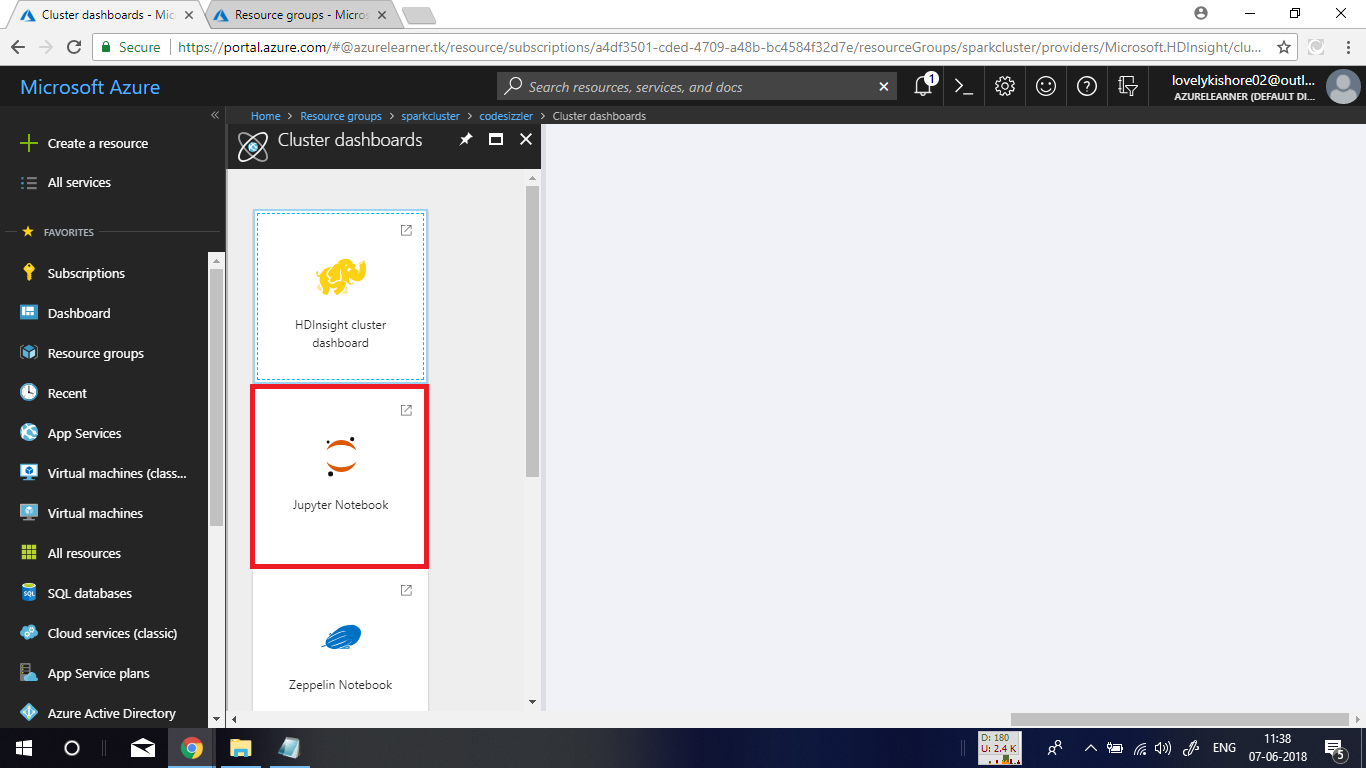


Go to the resource that you have created now. You can find it in the resource group.



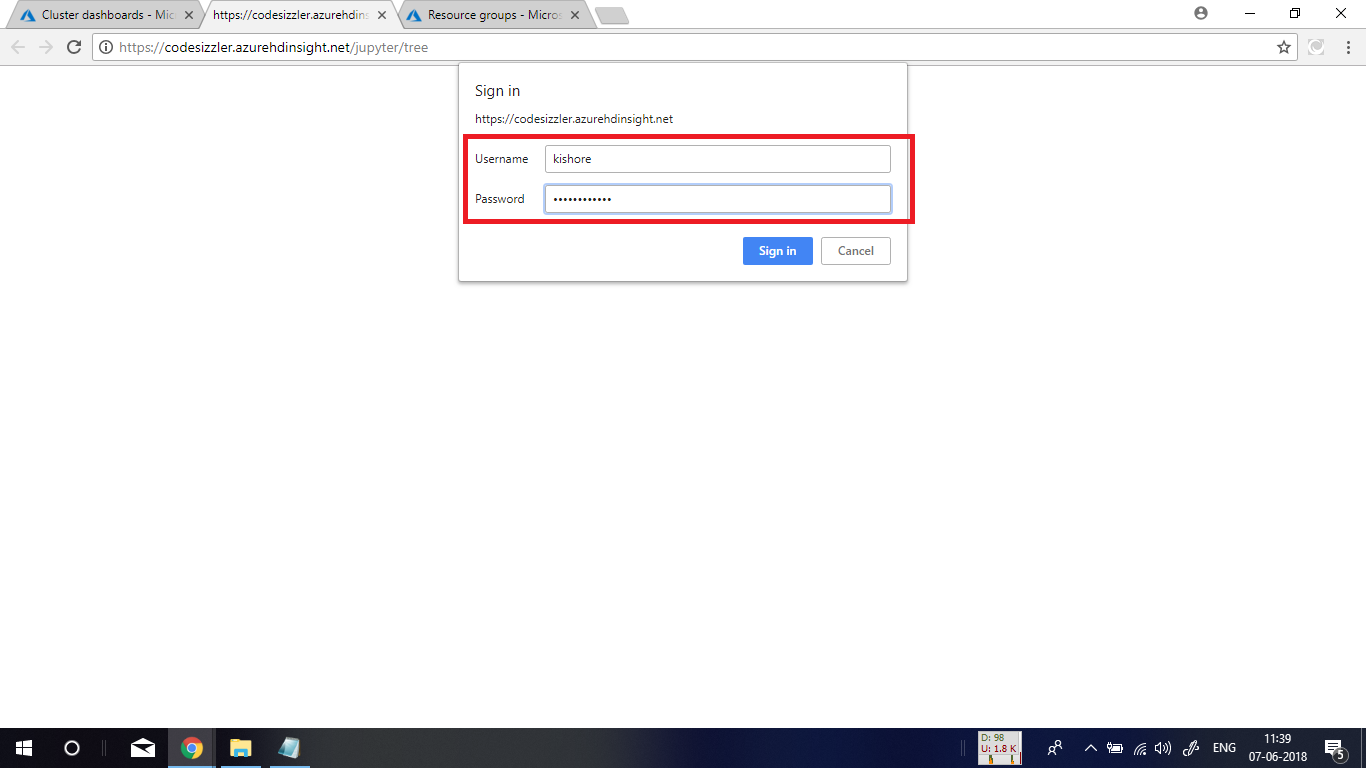
In the overview page of your cluster, click on **Cluster Dashboard.** This will show you another blade. In there, click on **Jupyter Notebook.** This will take you to new window.



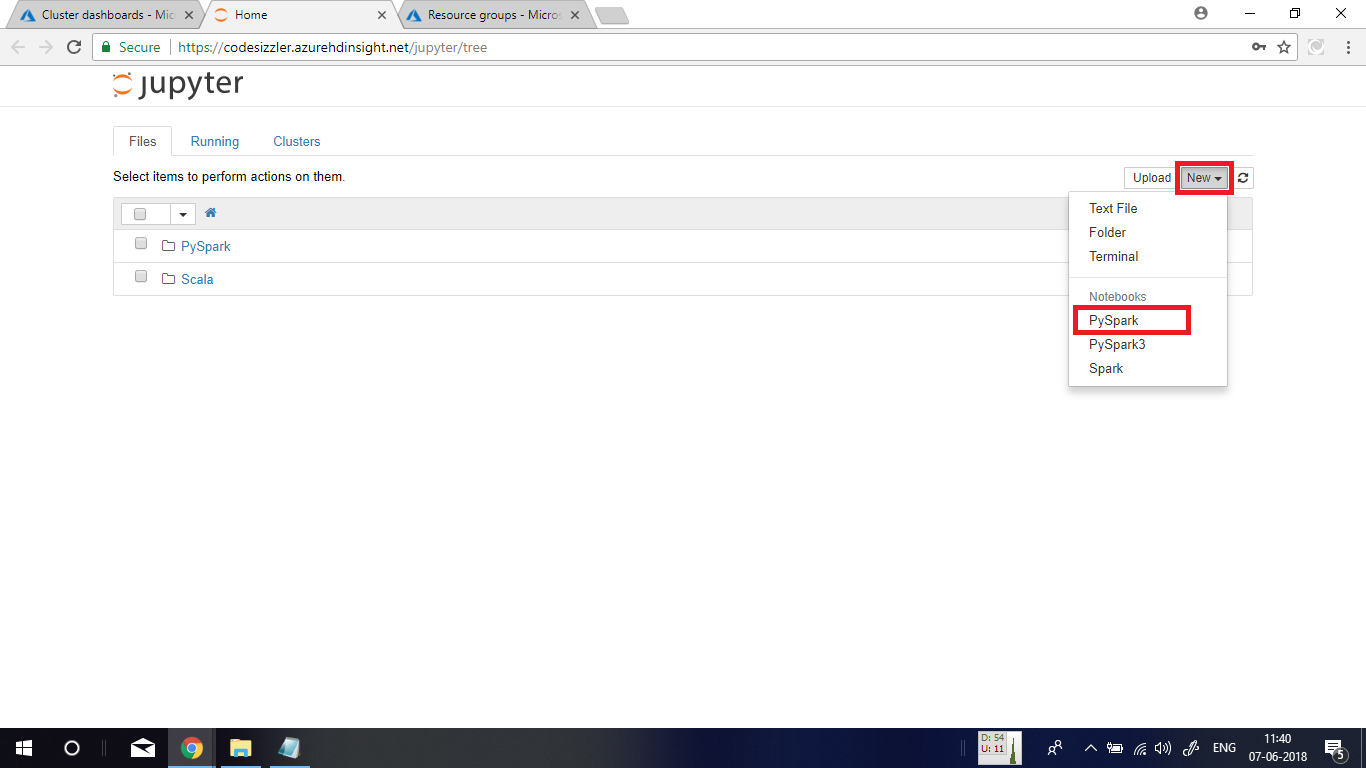


**Signing into Jupyter Notebook:**

In the new blade enter **Username** and **Password** the you gave while creating clusterand hit on **Sign In.**



After successful sign in, in the homepage of **Jupyter Notebook** that you get, click on **New** option in the top right corner and choose **PySpark** option to create new **Jupyter Notebook**.

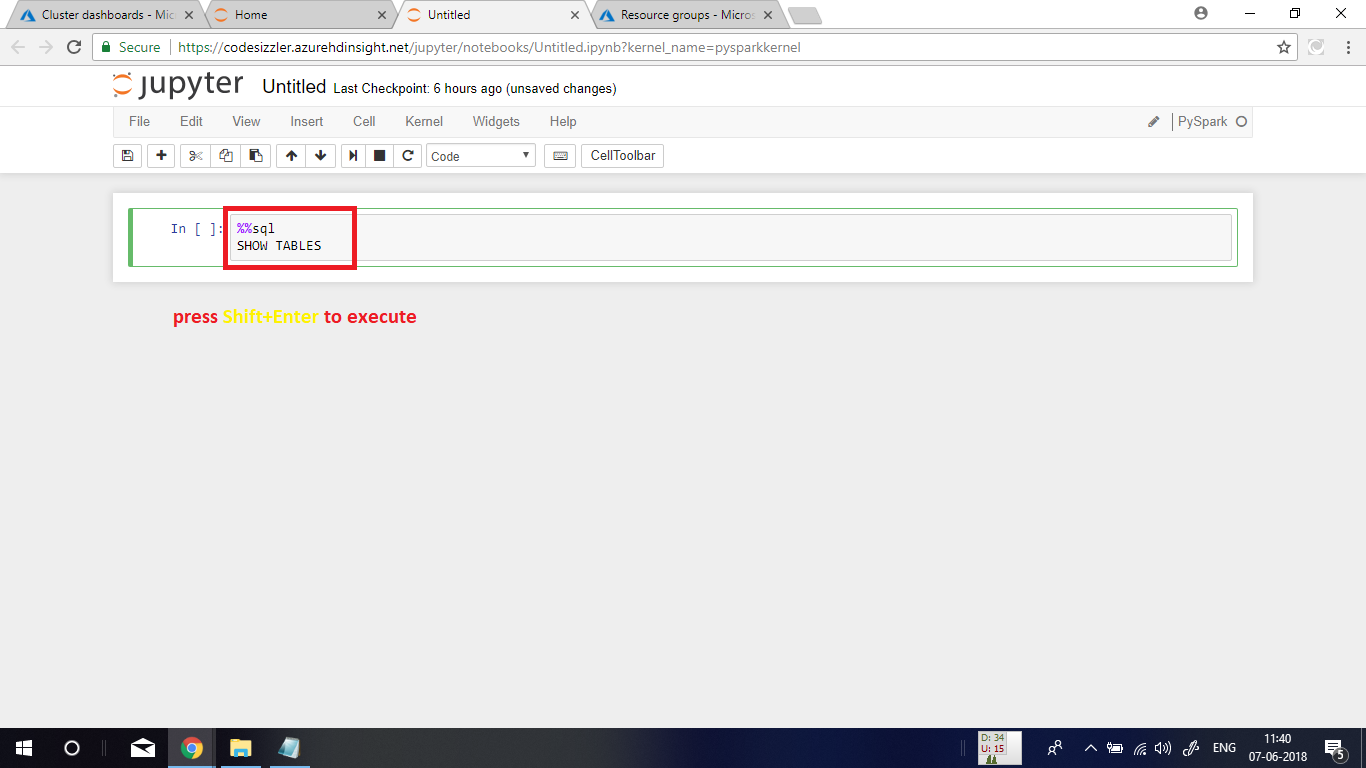


**Executing Spark SQL Statements:**

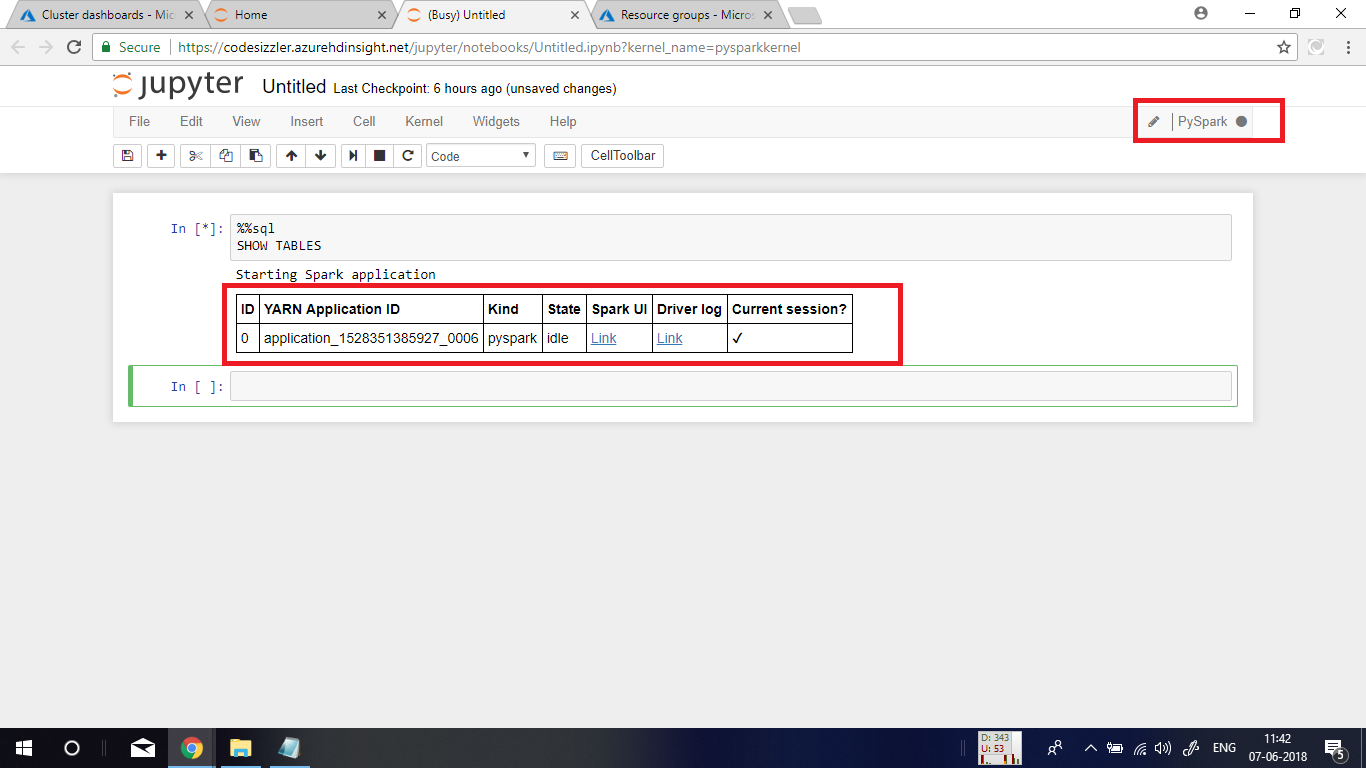
To get some data from Hive tables on the cluster, paste the below given command in the scripting pane and press **Shift+Enter** to execute it. The command is,

**%%sql**

**SHOW TABLES**



This will display you an output with a row of data. Note that if the circle in top right corner of your Jupyter Notebook window is dark, the kernel is busy compiling something. So wait for it to turn into a thin circle.



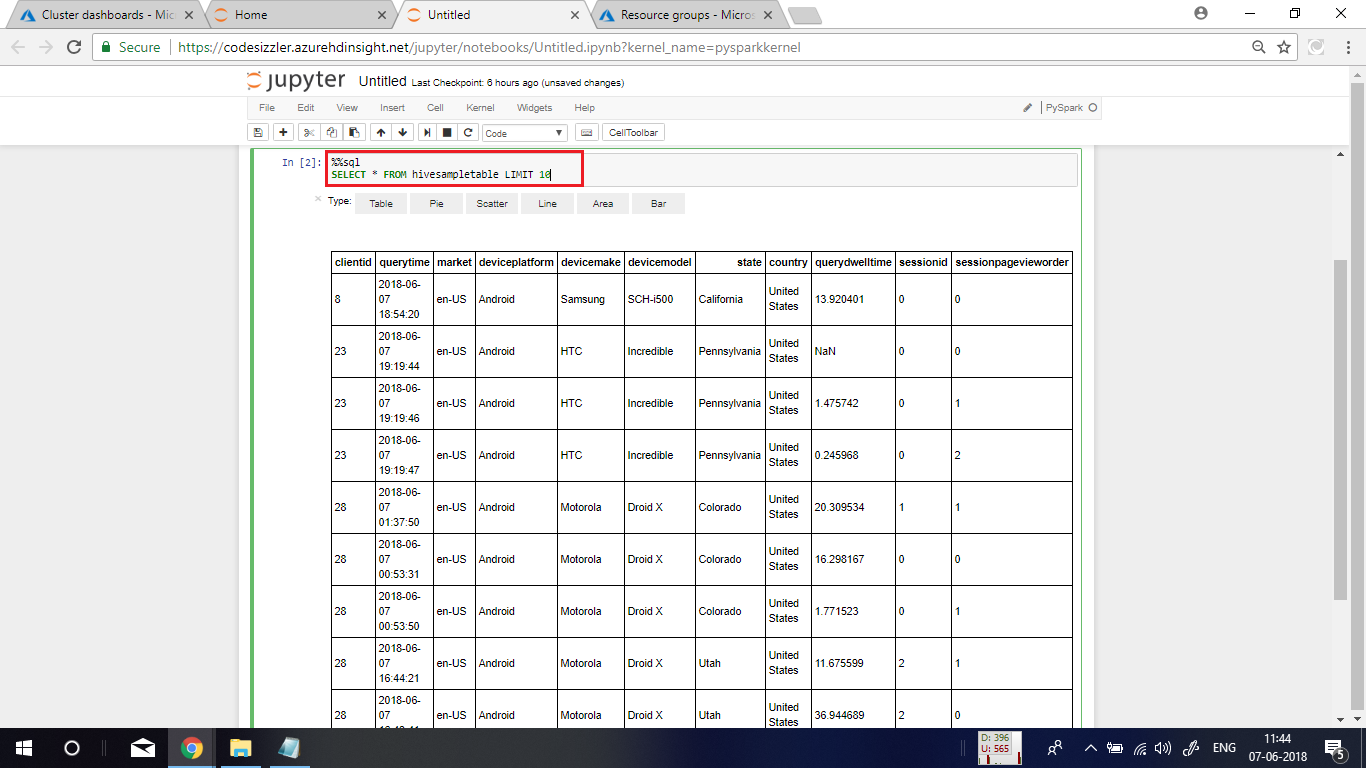
This is the output of the script that you run.



Again, run the below given query to get 10 rows from the Hive table.

**%%sql**

**SELECT \* FROM hivesampletable LIMIT 10**



**Deleting the Cluster:**

Go to the overview page of your cluster and click on delete button to delete the cluster. Usually, the HDInsight service pricing will be large. Hence make sure to delete your cluster once after you use it. The extracted data will be available in your storage account.

