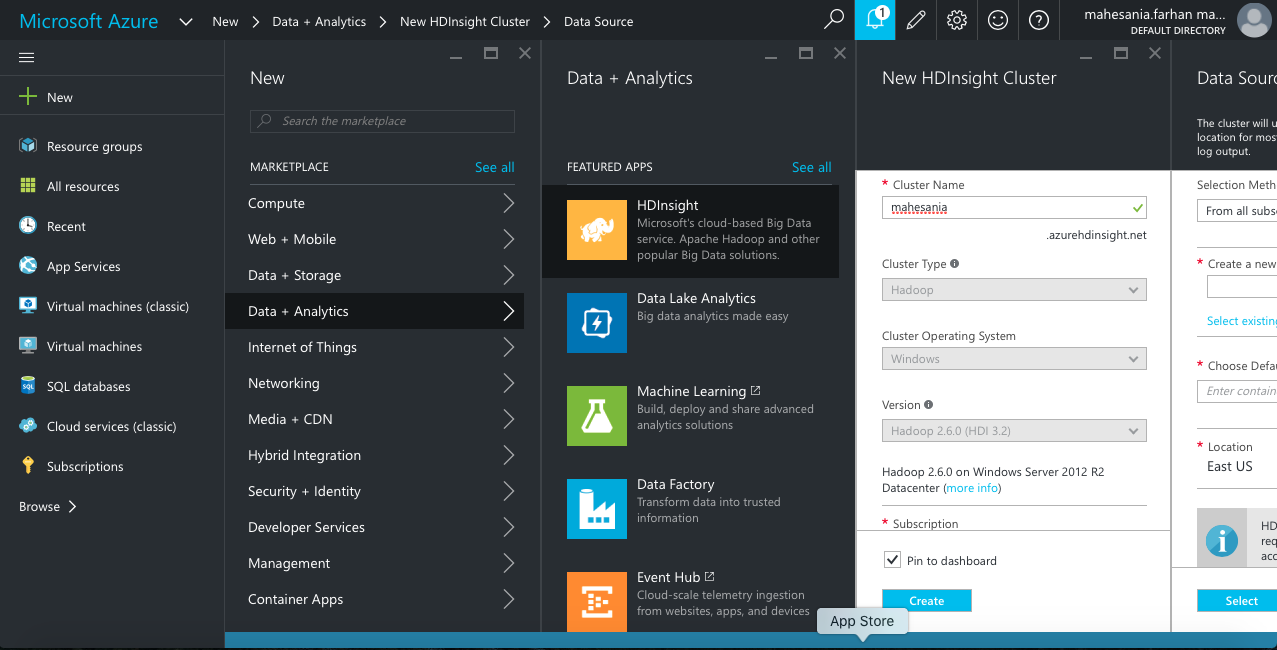
# **FOOD INSPECTION ANALYSIS PROJECT TUTORIAL**

**Prerequisites:**

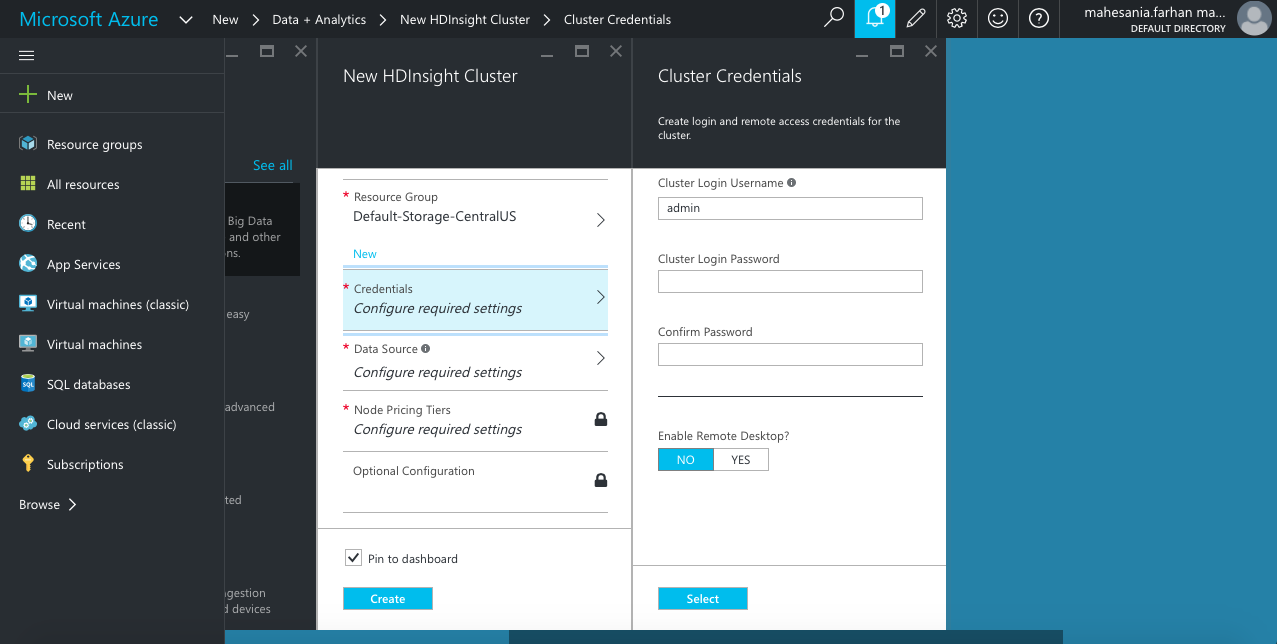
* Download CloudBerry for Azure Blob [here](http://www.cloudberrylab.com/download-thanks.aspx?prod=cbazure) .
* Download ODBC driver.
* Make a free trial account on Azure [Portal](https://portal.azure.com/).

**Steps to create HDInsight Cluster at Windows Azure (Azure id is required):**

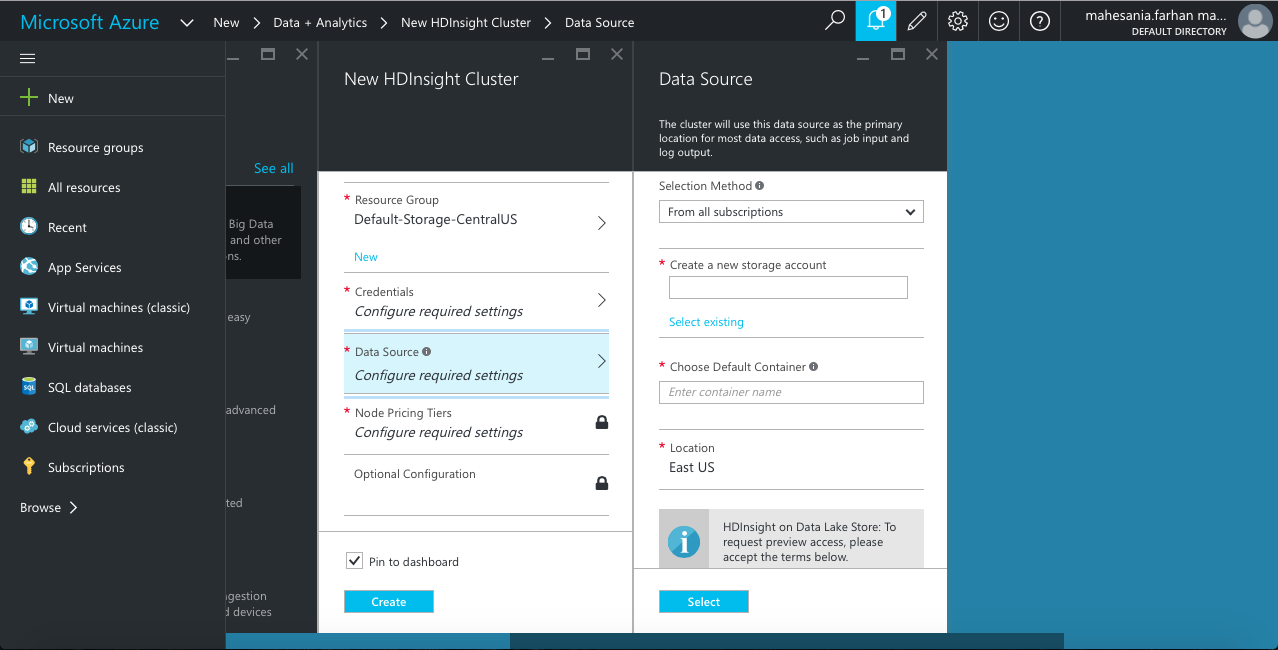
1. Click on + NEW tab on top left corner of dashboard → Data + Analytics → HDInsight.
2. Enter Cluster name, select cluster type as “Hadoop” from the dropdown box.
3. Select operating system as “Windows” from the dropdown box.



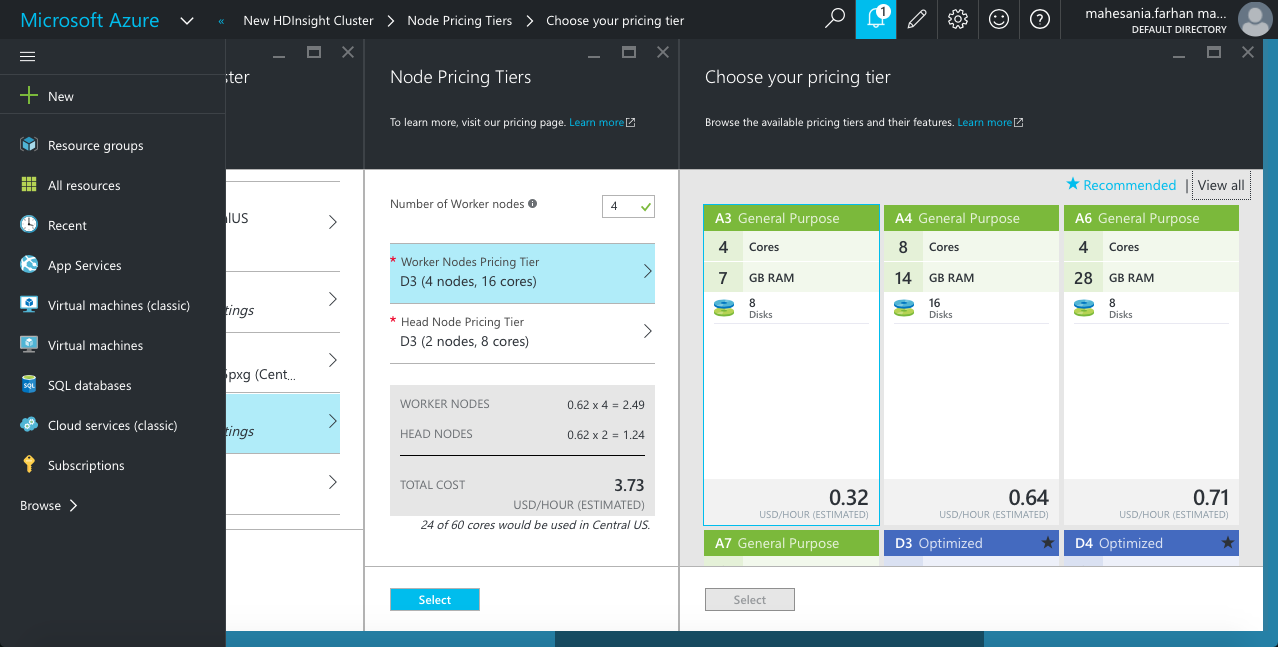
1. Click on “Credentials” tab and enter details(cluster login username, cluster login password, confirm password)



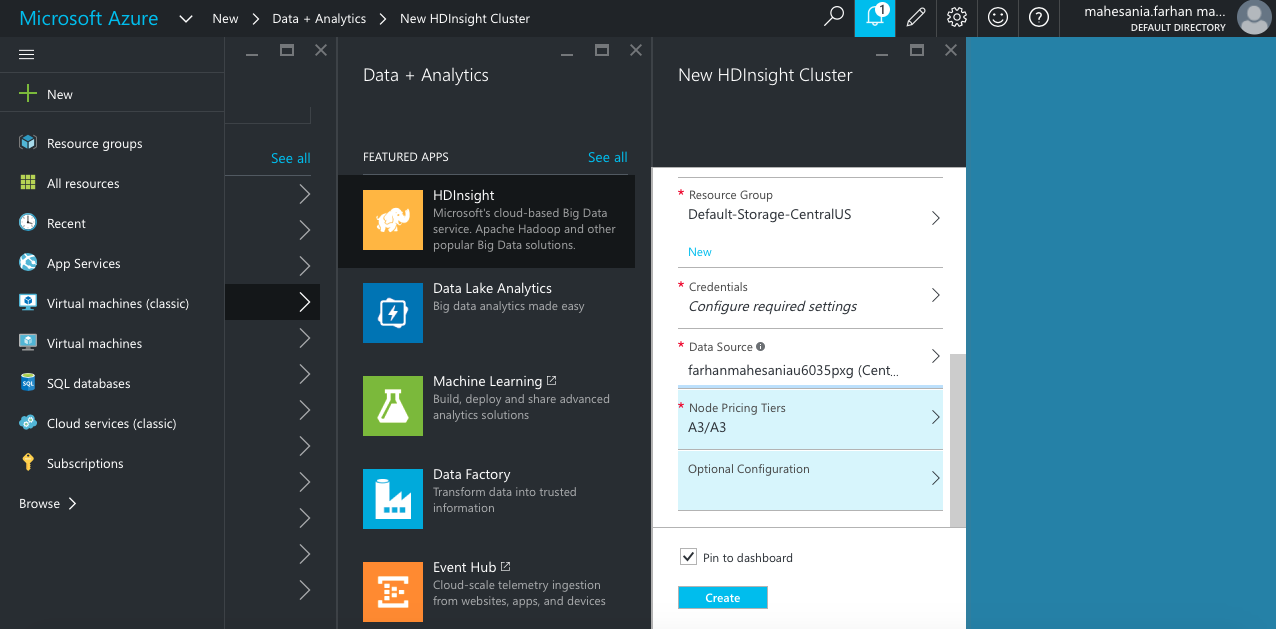
1. Click on Data Source tab to create a new data storage account or to use existing account click on “Select Existing” and choose the existing account.



1. Click on “Node Pricing Tiers Tab” click on View all and choose A3 as *Worker Node Pricing Tier* and *Head Node Pricing Tier*.



1. After entering the details and choosing the pricing tiers click on *Create* button to create a cluster. It takes a while for a cluster to get created.

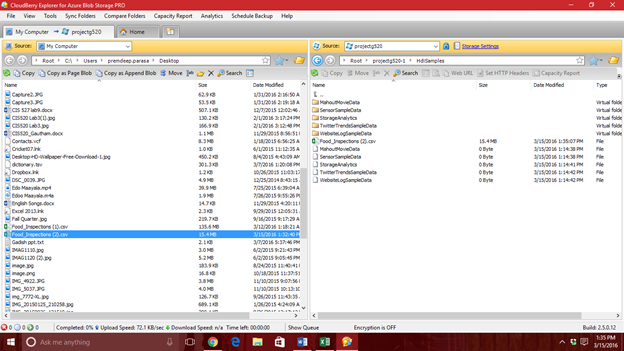


**Cloudberry for Azure Blob:**

Cloudberry is the simplest tool to upload data onto Azure. It works similarly as Windows File Explorer.

The following screenshot displays the interface of CloudBerry for Azure Blob.

1. After installing Window’s compatible Cloudberry, Open it for Azure Blob.
2. Connect the blob with your Cluster, named “projectg520”.
3. Once you connect, the root directory will open. Select “projectg520” and then, “HdiSamples”.
4. Drag and Drop your dataset file from your computer to “HdiSamples” folder.



**Running Queries in Hive:**

**Create Table:**

1. Use the following query to Create the Table in Hive:

CREATE TABLE FOOD\_INSPECTIONS

(

Inspection\_ID string,

Name string,

Licence\_No string,

Facility\_Type string,

Risk string,

Address string,

City string,

State string,

ZIP string,

Inspection\_Date string,

Inspection\_Type string,

Results string,

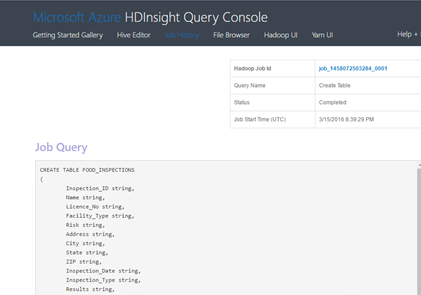
Latitude string,

Longitude string

)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

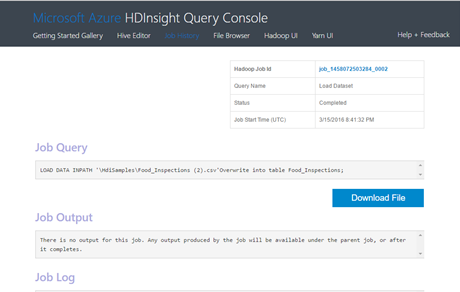
TBLPROPERTIES("skip.header.line.count"="1");



**Load Table:**

1. Once the table has been created, we need to Load the Table in Hive.
2. Use the following query to achieve it:

LOAD DATA INPATH '\HdiSamples\Food\_Inspections (2).csv'Overwrite into table Food\_Inspections;



**Select:**

Run the following queries in the Hive Editor to get results and generate the graphs in Excel.

1. Types of Risk by Facility Type :

select facility\_type, count(Risk) from food\_inspections where risk='High' group by facility\_type, risk;

1. Count of Individual risks :

select risk, count(Risk) as Risk\_count from food\_inspections group by risk;

1. Results by Inspection Types :

select results as status, count(Results) as Results\_count from food\_inspections group by results;

1. Locations covered :

select city, zip from food\_inspections group by city, zip;

1. Results of Facility Type ‘Restaurant’ :

select facility\_type, results, count(results) as results\_count from food\_inspections where facility\_type='RESTAURANT' group by facility\_type, results;

**Visualization of data:**

1. Provide credentials of Azure Cluster to ODBC.
2. Select Data tab on top.From Other Sources.
3. From Microsoft Query.
4. Hive
5. Add the columns
6. Next, and finish.
7. In Insert tab, Select power query.
8. Project the data.

**References:**

1. Dataset available here:<https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5/alt>
2. GIThub Help: <https://help.github.com>
3. Use ODBC: <https://azure.microsoft.com/en-us/documentation/articles/hdinsight-connect-excel-hive-odbc-driver/>
4. <http://doh.dc.gov/service/understanding-food-establishment-inspections>
5. <http://www.cityofchicago.org/city/en/depts/cdph/provdrs/environ_health/svcs/restaurant_food_inspection.html>