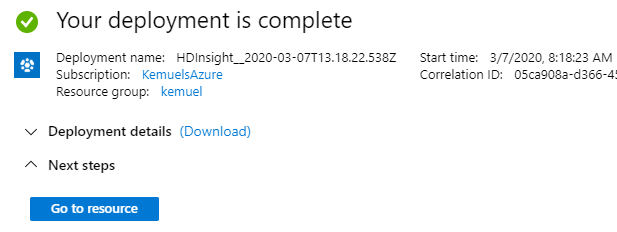
### DATA AGGREGATION DOCUMENTATION

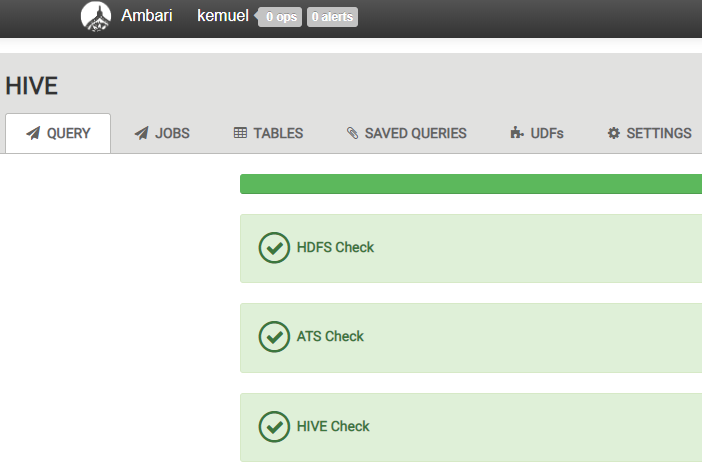
1. A Hadoop 2.7 HDInsight Cluster was re-deployed in Microsoft Azure using Virtual Machine configurations as follows
   1. Head (node size D12V2, 4 cores) with 2 nodes
   2. Work (node size D12V2 4 cores) with 1 node

* Same resource group and container was chosen as before



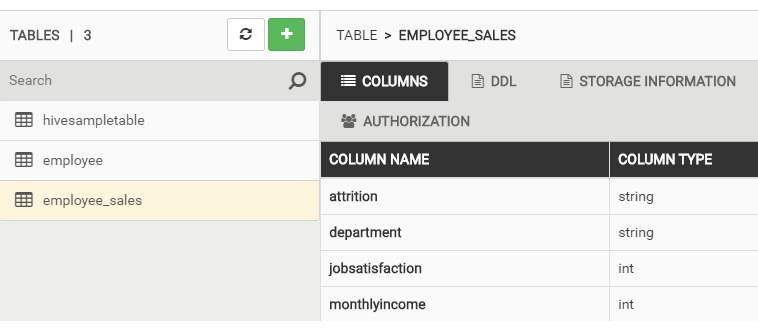
*Fig. 1 – Hadoop Cluster Deployed in HD Insight*

1. Hive (Ambari View) was used to load data from Azure storage to the HDInsight cluster, verify table exists from previous procedure (data transformation), and perform the data aggregation activities as required.



*Fig. 2- Successful System Checks in Hive on Ambari Views*

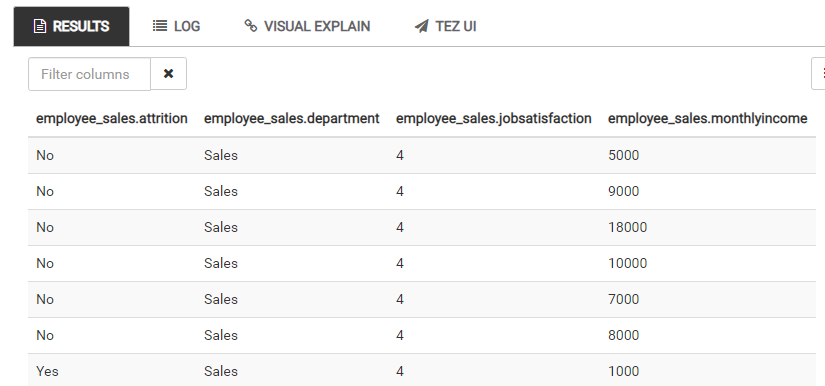
1. Table View in Ambari was used to verify the structure of the existing table and schema as shown in Fig. 3.



*Fig.3 - Schema for Existing Table in Table View*

1. SQL queries were used to verify that the data was loaded properly into the cluster and to perform various aggregations of the data. The following query was used to confirm the table exists in storage from the previous assignment. Fig. 4 shows a screenshot of the results from Hive. Note that the data was previously transformed to show only records for employees in the sales department and the income was already rounded to the nearest $1000.

SELECT \* FROM employee\_sales LIMIT 10;



*Fig. 4 – Existing table showing the first few rows of data.*

**Count – Number of Sales People with and without Attrition**

Two tables were created to show the counts of employees from department *Sales* with attrition versus those with no attrition. The SQL code and result from Hive are shown below. This demonstrates of the COUNT aggregate function in SQL.

SELECT COUNT(Attrition) AS attrition

FROM employee\_sales

WHERE Attrition LIKE "%Yes%";

SELECT COUNT(Attrition) AS Non\_Attrition

FROM employee\_sales

WHERE Attrition LIKE "%No%";

*Fig. 5 – Counts of Salespeople with Attrition (left) and No Attrition (right)*

Notice that there are more people with no attrition from the dataset (almost 4 times the number with attrition).

**Average, Min and Max – Aggregate Statistics for Each Class**

Tables were created for (i) all sales people, (ii) those with attrition, and (iii) those with no attrition, to show basic aggregate statistics in SQL (average, min and max). The SQL queries and results are shown below. The ROUND transformation function in SQL was used to show the average monthly income to the nearest dollar.

SELECT ROUND(AVG(MonthlyIncome), 0) AS average\_monthly\_income, MIN(MonthlyIncome) as min\_monthly\_income, MAX(MonthlyIncome) as max\_monthly\_income

FROM employee\_sales;

SELECT ROUND(AVG(MonthlyIncome), 0) AS average\_monthly\_income, MIN(MonthlyIncome) as min\_monthly\_income, MAX(MonthlyIncome) as max\_monthly\_income

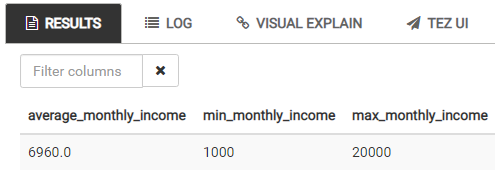
FROM employee\_sales

WHERE Attrition LIKE "%Yes%";

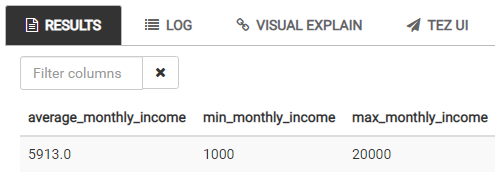
SELECT ROUND(AVG(MonthlyIncome), 0) AS average\_monthly\_income, MIN(MonthlyIncome) as min\_monthly\_income, MAX(MonthlyIncome) as max\_monthly\_income

FROM employee\_sales

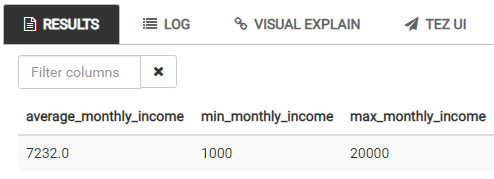
WHERE Attrition LIKE "%No%";



*Fig. 6 – Statistics of All Sales People*



*Fig. 7 – Statistics of Sales People with Attrition*



*Fig. 8 – Statistics of Sales People with no Attrition*

Notice that that average income is higher for sales people with no attrition. However, the range is the same for attrition and non-attrition categories (min and max monthly income is the same). The average statistic result could mean that monthly salary is an influential factor of employee attrition.

**Group By – Number of Sales People with and without Attrition Grouped by Monthly Income**

Two tables were created to demonstrate the use of grouping in SQL by making use of the GROUP BY query function. The objective was to count the number of sales people with attrition versus those with no attrition, by monthly income. The SQL queries to achieve this and the results are shown below.

SELECT MonthlyIncome as monthlyincome, COUNT(MonthlyIncome) AS count

FROM employee\_sales

WHERE Attrition LIKE "%Yes%"

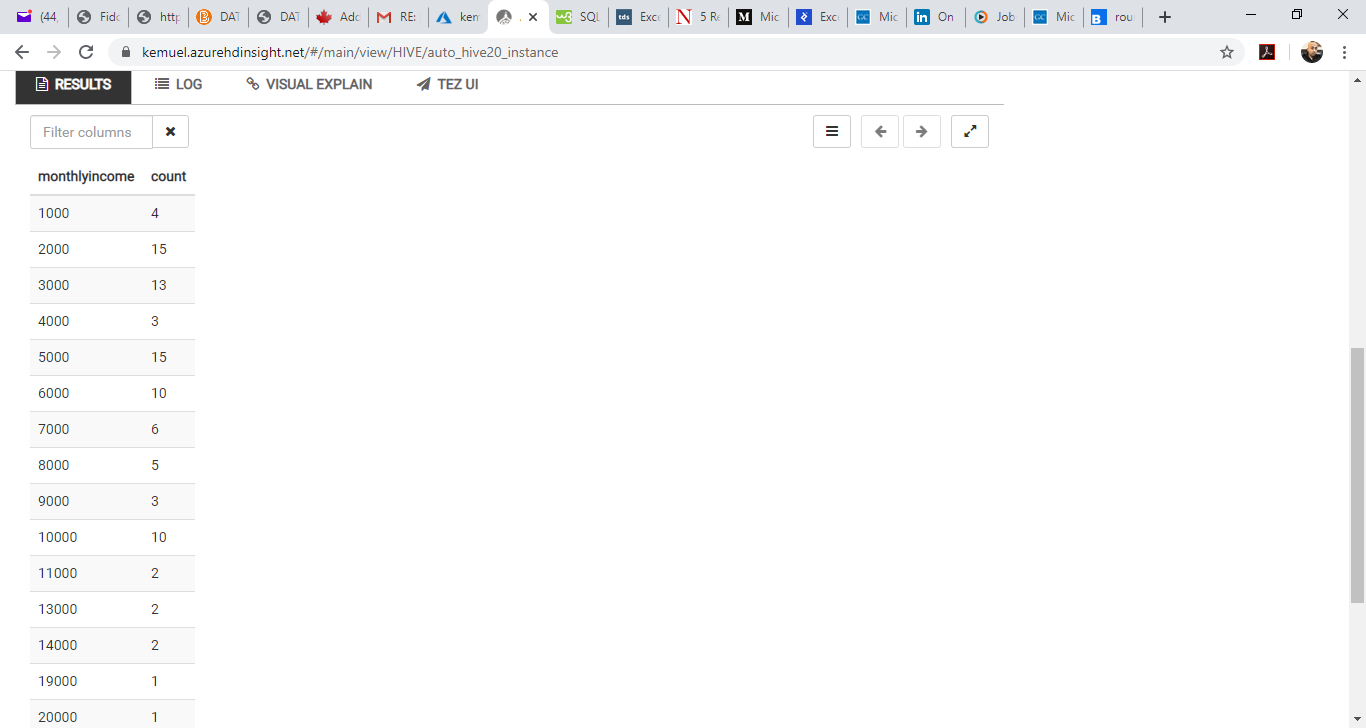
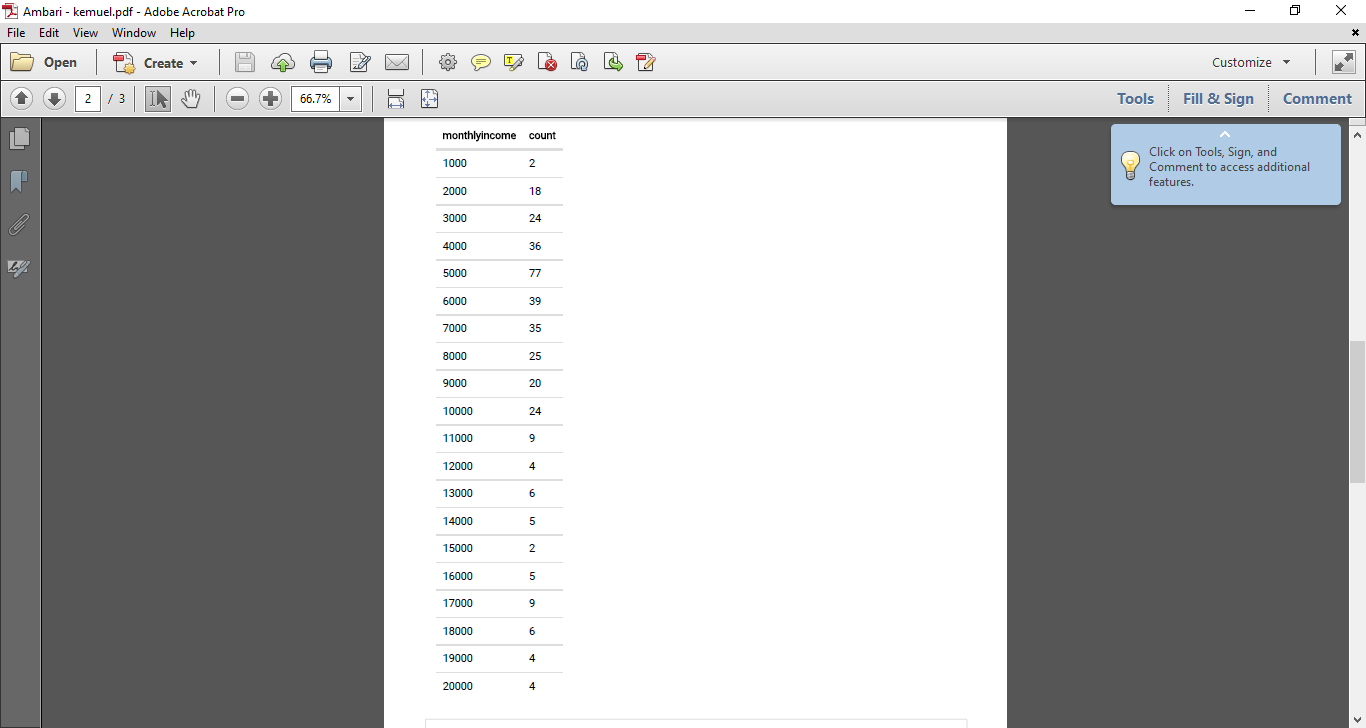
GROUP BY MonthlyIncome;

SELECT MonthlyIncome as monthlyincome, COUNT(MonthlyIncome) AS count

FROM employee\_sales

WHERE Attrition LIKE "%No%"

GROUP BY MonthlyIncome;

*Fig. 9 – Number of Sales People grouped by Monthly Income for those with Attrition (left) and no Attrition (right)*

The results help to confirm the hypothesis stated earlier that sales people with higher monthly incomes are less likely to leave their jobs, compared to those with lower incomes. However, the tables in Fig. 9 do not provide an intuitive way reveal this trend. There is need to present these results on a graph for better visualization.

**Data Source:**

IBM HR Analytics Employee Attrition & Performance (2017). Retrieved Feb 13, 2020 from <https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset/data>