**SAN FRANCISCO SALARIES ANALYSIS**

Dharmesh Soni-305093443

Ishan Fafadia-305058603

Saket Wadhera-305086930

**Table of Contents**

**Introduction**

**Table of Contents**

**Introduction**

In this Project we have chosen Dataset of San Francisco to analyze the different attributes of compensation and to find which department is best in that area and which has shown the most growth in 4 year and to best suited for analyzing in hive we have converted our data to Avro format.

And we will analyze the data in form of various charts and graph with respect to

1.Yearly Average Pay

2.Average Pay per Job Type

3 Different Compensation Regarding Job Type

**Prerequisites**

* Microsoft Azure
* Microsoft Excel
* Hive editor
* Odbc Driver

**1.Create table and load data**

**1.1** After creation of cluster, click on the cluster at the dashboard and and create the table for loading the data in the cluster

CREATE TABLE SALARY\_DATA

(

Employee\_Id INT,

EmployeeName STRING,

JobTitle STRING,

BasePay INT,

OvertimePay INT,

OtherPay INT,

Benefits INT,

TotalPay INT,

TotalPayBenefits INT,

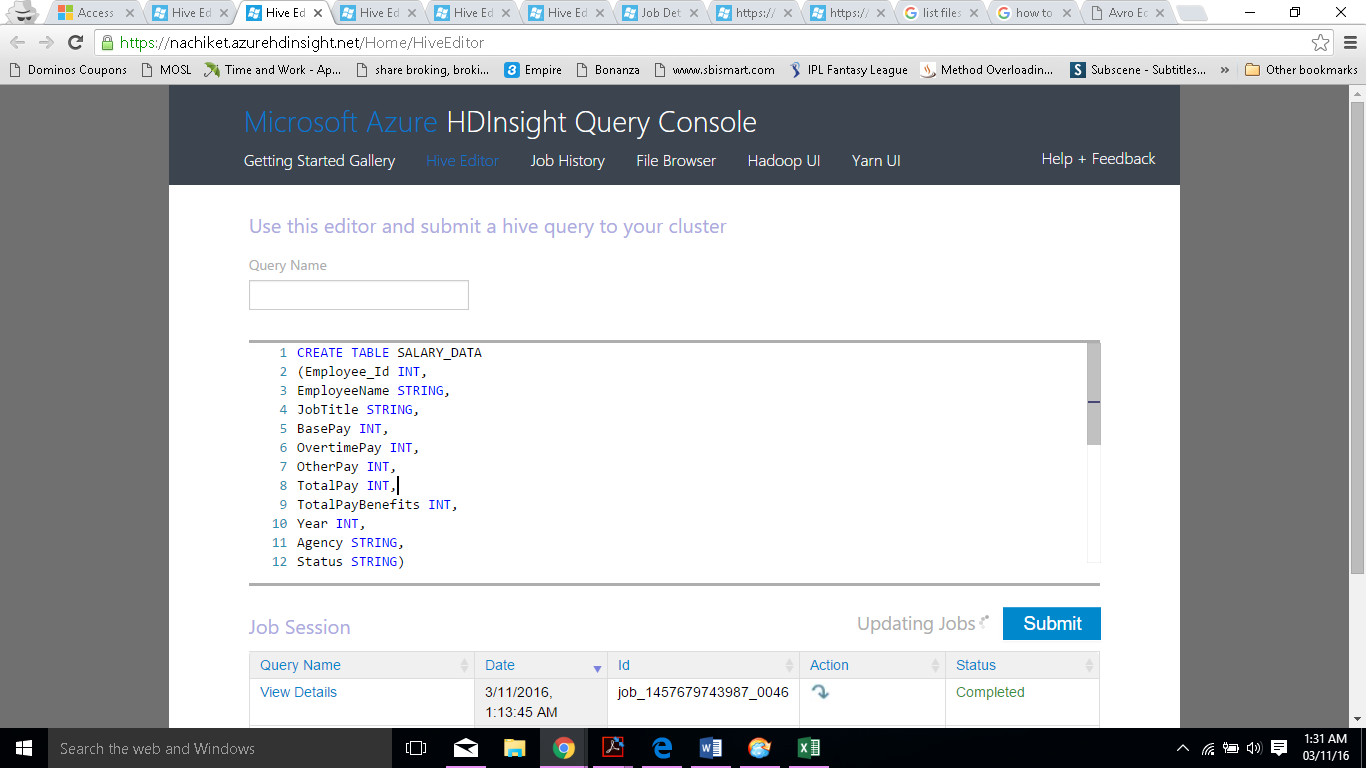
Year INT,

Agency STRING,

Status STRING)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

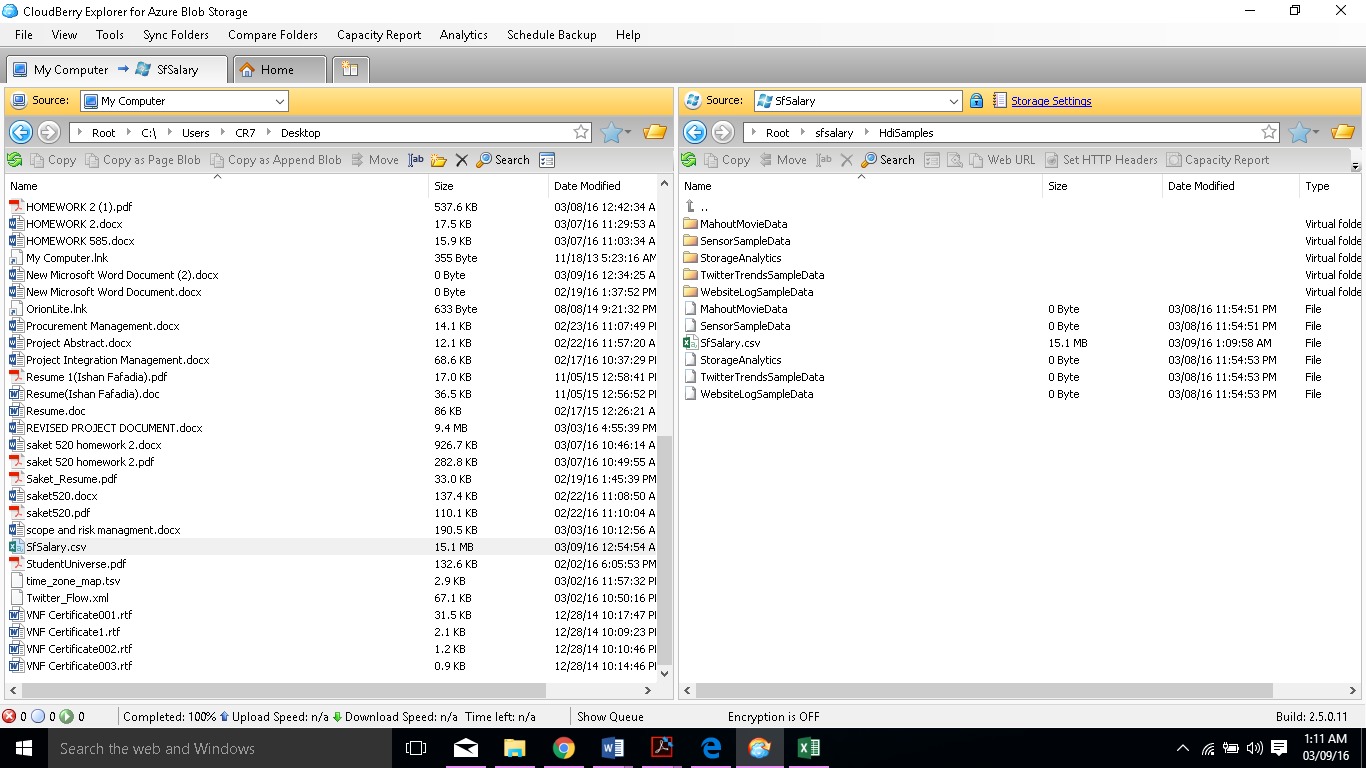
TBLPROPERTIES ("skip.header.line.count"="1");



**2.Loading Data In Cluster Using Clouberry**

To upload data in the cluster we will use Cloudberry for Azure.

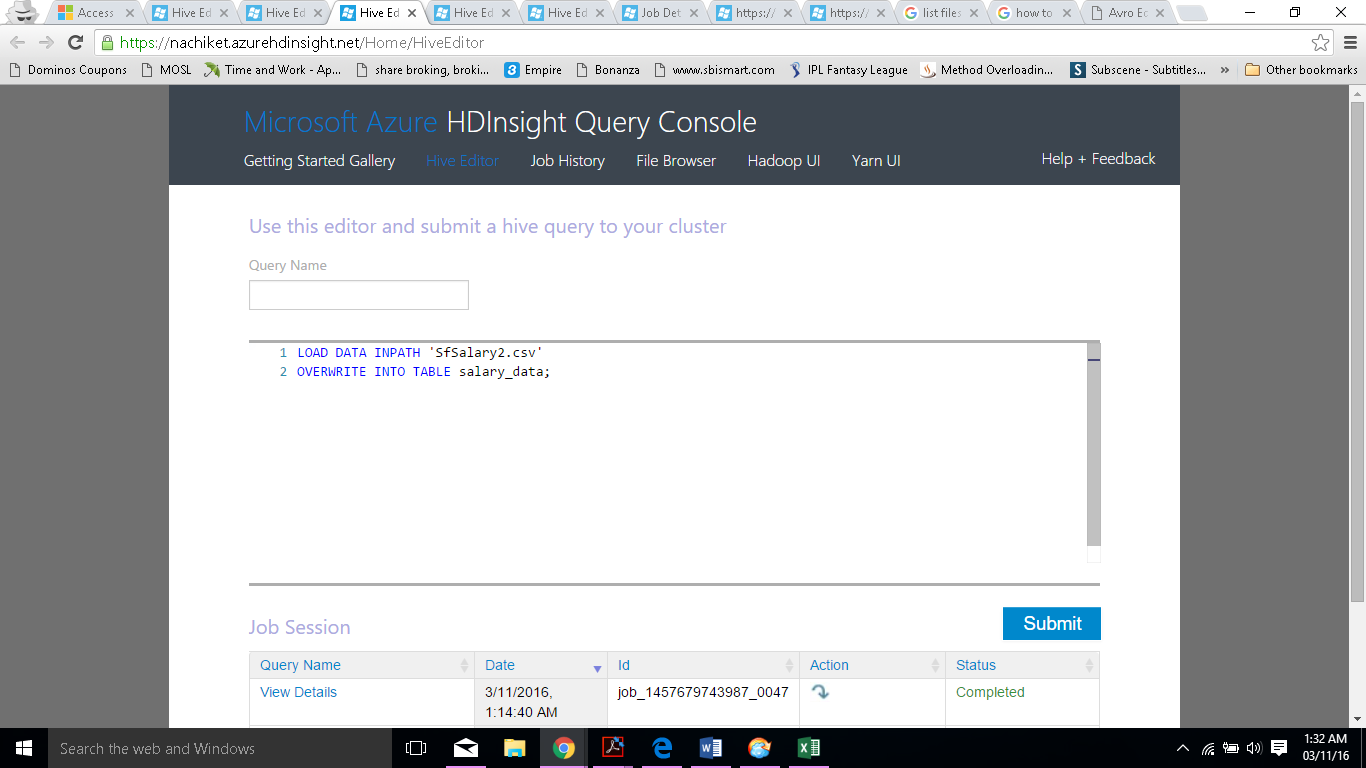
* Open the Cloudberry, once it is installed on local machine.
* Select Azure Blob In the dialog box, add the details as mentioned while creating cluster and upload file to the cluster.



**3.Load Data In Table**

LOAD DATA INPATH '\HdiSamples\SfSalary2.csv'

OVERWRITE INTO TABLE SALARY\_DATA;



**4. CONVERSION TO AVRO FORMAT**

4.1 Creating Schema For Avro Table

CREATE TABLE avro1\_table

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.avro.AvroSerDe'

STORED AS INPUTFORMAT 'org.apache.hadoop.hive.ql.io.avro.AvroContainerInputFormat'

OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io.avro.AvroContainerOutputFormat'

TBLPROPERTIES (

'avro.schema.literal'='{

"namespace": "com.saket.avro",

"name": "Sfsalary\_data",

"type": "record",

"fields":

[ { "name":"Employee\_Id","type":"int"},

{ "name":"EmployeeName","type":"string"},

{ "name":"JobTitle","type":"string"},

{ "name":"BasePay","type":"int"},

{ "name":"OvertimePay","type":"int"},

{ "name":"OtherPay","type":"int"},

{ "name":"Benefits","type":"int"},

{ "name":"TotalPay","type":"int"},

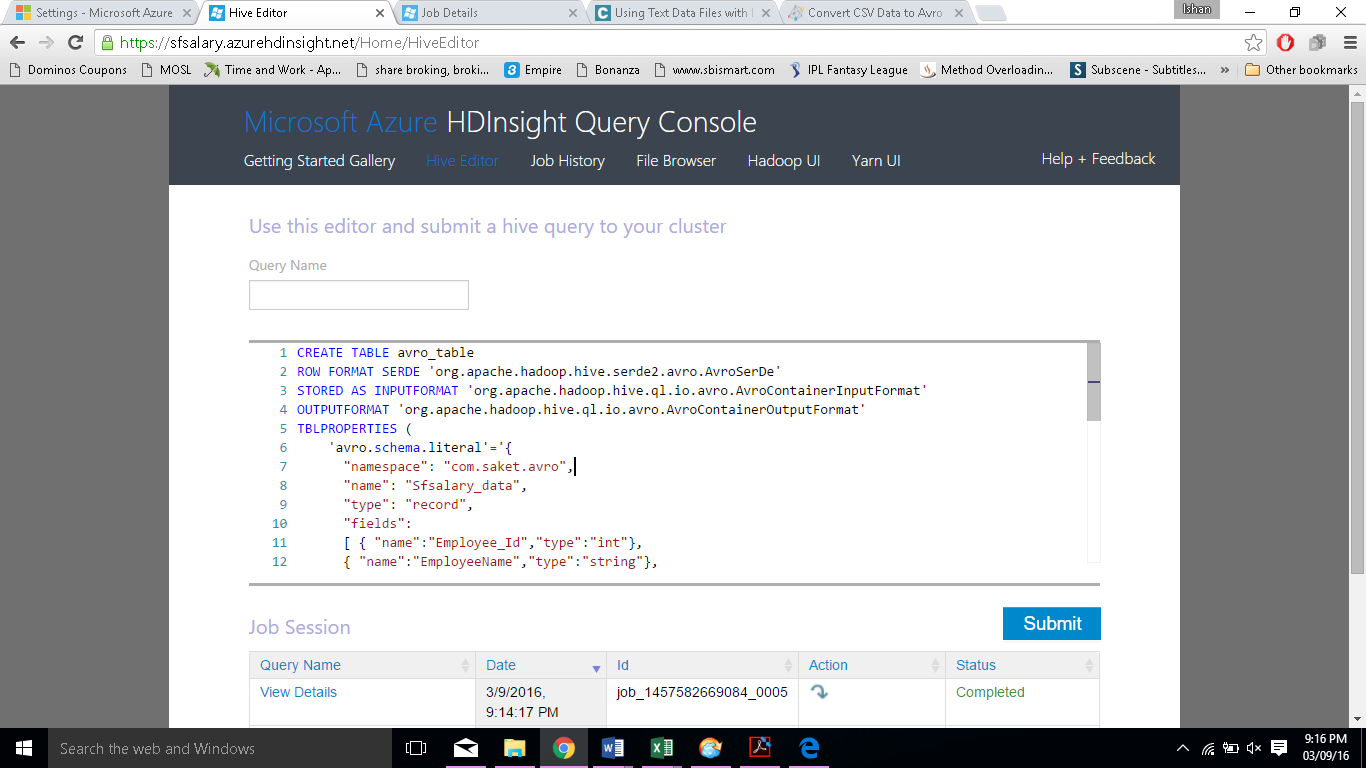
{ "name":"TotalPayBenefits","type":"int"},

{ "name":"Year","type":"int"},

{ "name":"Agency","type":"string"},

{ "name":"Status","type":"string"}]

}');

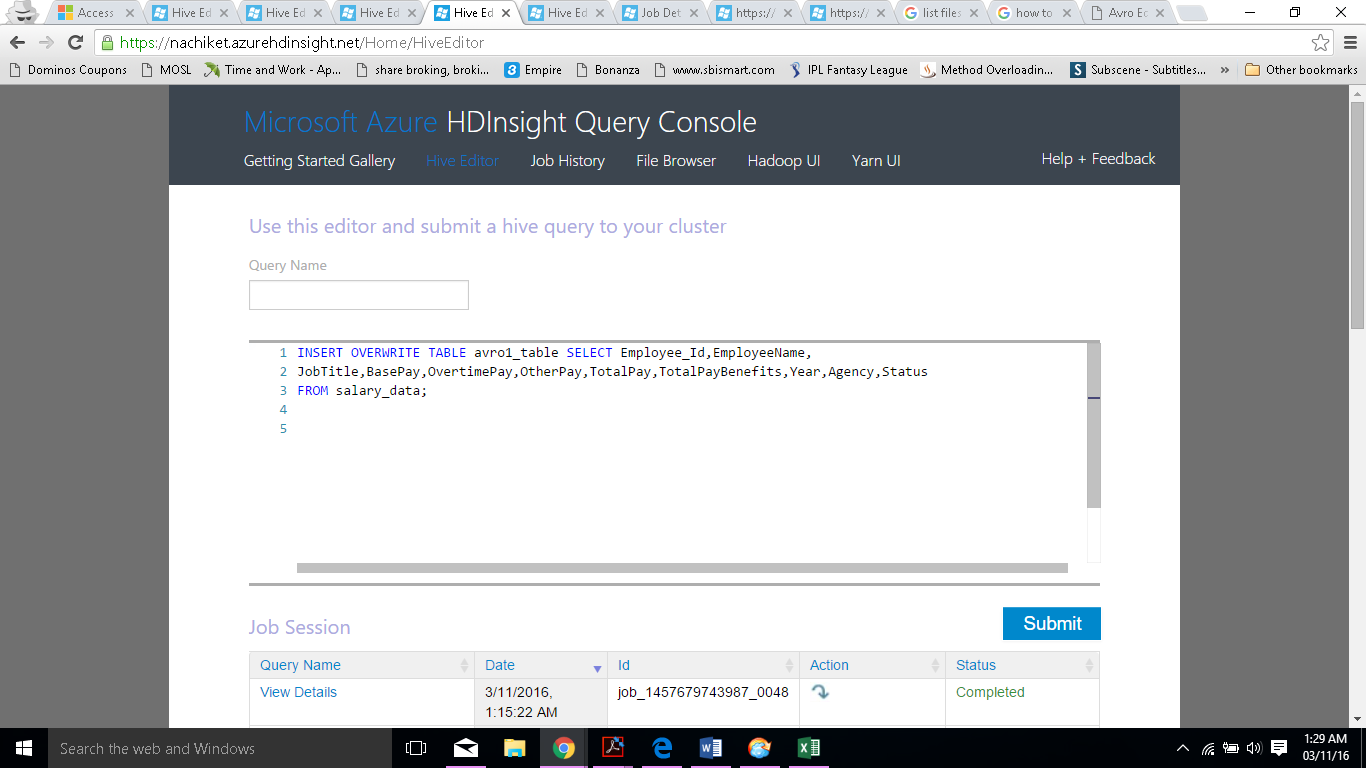


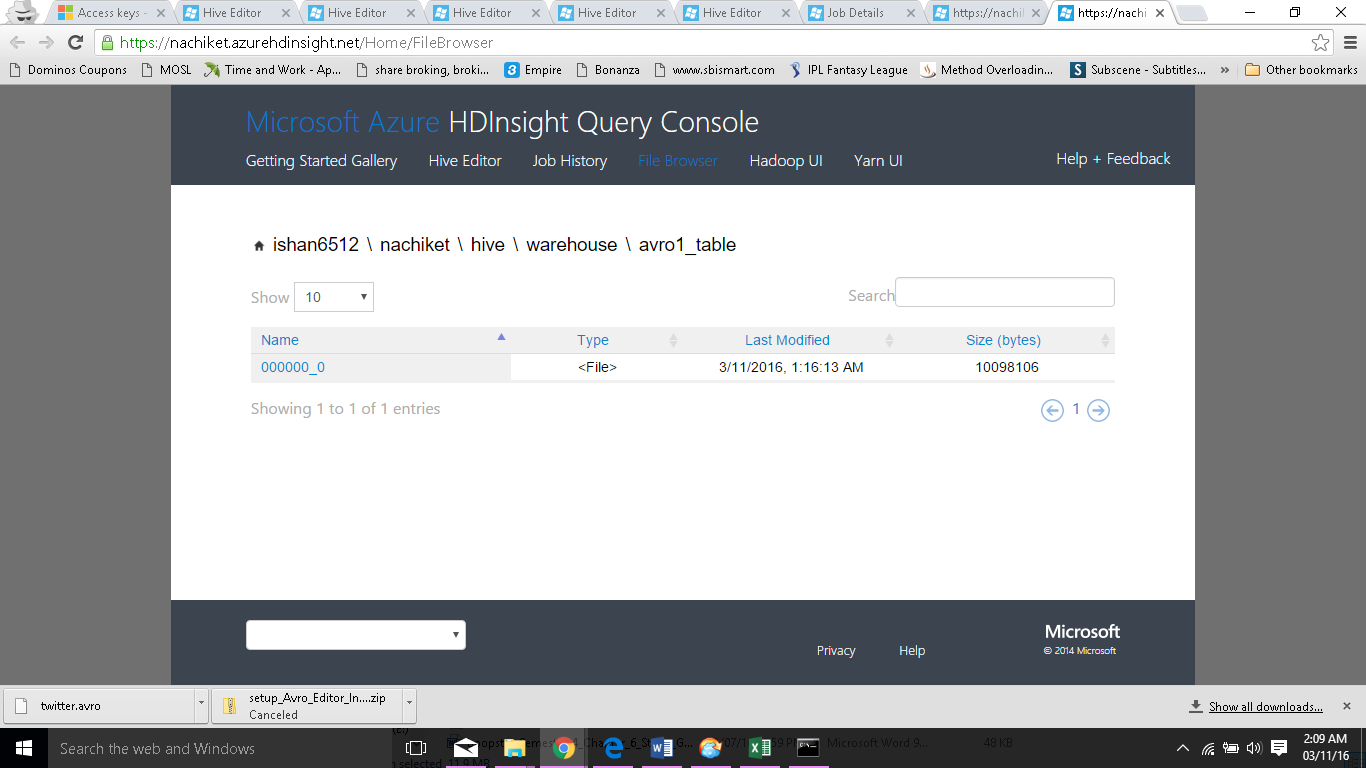
4.2 Loading Data In Avro table

INSERT OVERWRITE TABLE avro1\_table SELECT Employee\_Id,EmployeeName,

JobTitle,BasePay,OvertimePay,Benefits,OtherPay,TotalPay,TotalPayBenefits,Year,Agency,Status

FROM salary\_data;





**5.Analyzing the data**

**5.1 Maximum TotalPay By Year**

Following Query Is Run to Find That:

Create Table totalpay\_data

(

year int,

totalpay int

);

INSERT OVERWRITE TABLE totalpay\_data

Select Year,max(TotalPay) from avro\_table

group by Year order by 1;

**5.2 Benefits By Year**

Following Query Is Run to Find That:

Create Table benefits\_data

(

year int,

Benefits int

);

INSERT OVERWRITE TABLE benefits\_data

Select Year,Sum(Benefits) from avro\_table

group by year order by 1;

**5.3 Mean Salary By Year**

Following Query Is Run to Find That:

CREATE TABLE Record1\_DATA

(Year int,

Meanpay int,

Records int

);

INSERT OVERWRITE TABLE Record1\_DATA

SELECT year, percentile(cast(Totalpay as bigint), 0.5) , count(\*) Records from avro\_table

group by year order by 1;

**5.4 Mean Salary By Department**

INSERT OVERWRITE TABLE AVERAGE1\_DATA

select JobType,percentile(cast(Totalpay as bigint), 0.5),count (\*) as Records from

(

select case when upper(jobtitle) like '%FIRE%' then 'Fire'

when upper(jobtitle) like '%POLICE%' then 'Police'

when upper(jobtitle) like '%SHERIFF%' then 'Police'

when upper(jobtitle) like '%PROBATION%' then 'Police'

when upper(jobtitle) like '%SERGEANT%' then 'Police'

when upper(jobtitle) like '%MTA%' then 'Transit'

when upper(jobtitle) like '%TRANSIT%' then 'Transit'

when upper(jobtitle) like '%ANESTH%' then 'Medical'

when upper(jobtitle) like '%MEDICAL%' then 'Medical'

when upper(jobtitle) like '%NURSE%' then 'Medical'

when upper(jobtitle) like '%HEALTH%' then 'Medical'

when upper(jobtitle) like '%PYSICIAN%' then 'Medical'

when upper(jobtitle) like '%ORTHOPEDIC%' then 'Medical'

when upper(jobtitle) like '%PHARM%' then 'Medical'

when upper(jobtitle) like '%health%' then 'Medical'

when upper(jobtitle) like '%AIRPORT%' then 'Airport'

when upper(jobtitle) like '%ANIMAL%' then 'Animal'

when upper(jobtitle) like '%ARCHITECT%' then 'Architectural'

when upper(jobtitle) like '%COURT%' then 'Court'

when upper(jobtitle) like '%LEGAL%' then 'Court'

when upper(jobtitle) like '%MAYOR%' then 'Mayor'

when upper(jobtitle) like '%LIBRARY%' then 'Library'

when upper(jobtitle) like '%PARKING%' then 'Parking'

when upper(jobtitle) like '%Public Works%' then 'Public Works'

when upper(jobtitle) like '%ATTORNEY%' then 'Attorney'

when upper(jobtitle) like '%MECHANIC%' then 'Automotive'

when upper(jobtitle) like '%AUTOMOTIVE%' then 'Automotive'

when upper(jobtitle) like '%CUSTODIAN%' then 'Custodian'

when upper(jobtitle) like '%ENGINEER%' then 'Engineering'

when upper(jobtitle) like '%ENGR%' then 'Engineering'

when upper(jobtitle) like '%ACCOUNT%' then 'Accounting'

when upper(jobtitle) like '%GARDENER%' then 'Gardening'

when upper(jobtitle) like '%GENERAL LABORER%' then 'General Laborer'

when upper(jobtitle) like '%FOOD SERV%' then 'Food Service'

when upper(jobtitle) like '%CLERK%' then 'Clerk'

when upper(jobtitle) like '%PORTER%' then 'Porter'

else 'Other' end JobType,

TotalPay

from avro\_table )

JobType group by JobType;

**5.5 Salary Structure Of Different Depatments**

CREATE TABLE pay1\_DATA

(Jobtitle string,

Basepay bigint,

Overtimepay bigint,

Otherpay bigint,

Benefits bigint

)

INSERT OVERWRITE TABLE pay1\_DATA

select JobType, cast(avg(Basepay) as bigint), cast(avg(Overtimepay) as bigint),cast(avg(Otherpay) as bigint), cast(avg(Benefits) as bigint)

from

(

select case when upper(jobtitle) like '%FIRE%' then 'Fire'

when upper(jobtitle) like '%POLICE%' then 'Police'

when upper(jobtitle) like '%SHERIFF%' then 'Police'

when upper(jobtitle) like '%PROBATION%' then 'Police'

when upper(jobtitle) like '%SERGEANT%' then 'Police'

when upper(jobtitle) like '%MTA%' then 'Transit'

when upper(jobtitle) like '%TRANSIT%' then 'Transit'

when upper(jobtitle) like '%ANESTH%' then 'Medical'

when upper(jobtitle) like '%MEDICAL%' then 'Medical'

when upper(jobtitle) like '%NURSE%' then 'Medical'

when upper(jobtitle) like '%HEALTH%' then 'Medical'

when upper(jobtitle) like '%PYSICIAN%' then 'Medical'

when upper(jobtitle) like '%ORTHOPEDIC%' then 'Medical'

when upper(jobtitle) like '%PHARM%' then 'Medical'

when upper(jobtitle) like '%health%' then 'Medical'

when upper(jobtitle) like '%AIRPORT%' then 'Airport'

when upper(jobtitle) like '%ANIMAL%' then 'Animal'

when upper(jobtitle) like '%ARCHITECT%' then 'Architectural'

when upper(jobtitle) like '%COURT%' then 'Court'

when upper(jobtitle) like '%LEGAL%' then 'Court'

when upper(jobtitle) like '%MAYOR%' then 'Mayor'

when upper(jobtitle) like '%LIBRARY%' then 'Library'

when upper(jobtitle) like '%PARKING%' then 'Parking'

when upper(jobtitle) like '%Public Works%' then 'Public Works'

when upper(jobtitle) like '%ATTORNEY%' then 'Attorney'

when upper(jobtitle) like '%MECHANIC%' then 'Automotive'

when upper(jobtitle) like '%AUTOMOTIVE%' then 'Automotive'

when upper(jobtitle) like '%CUSTODIAN%' then 'Custodian'

when upper(jobtitle) like '%ENGINEER%' then 'Engineering'

when upper(jobtitle) like '%ENGR%' then 'Engineering'

when upper(jobtitle) like '%ACCOUNT%' then 'Accounting'

when upper(jobtitle) like '%GARDENER%' then 'Gardening'

when upper(jobtitle) like '%GENERAL LABORER%' then 'General Laborer'

when upper(jobtitle) like '%FOOD SERV%' then 'Food Service'

when upper(jobtitle) like '%CLERK%' then 'Clerk'

when upper(jobtitle) like '%PORTER%' then 'Porter'

else 'Other' end JobType,

Basepay,

Overtimepay,

Otherpay,

Benefits

from avro\_table )

JobType group by JobType;