Step-by-step commands:

Q1. Using Hadoop command move all those employees data into HDFS directory "/user/your_user_name/employees_data" directory.

Created a new folder “employees_data” in FTP

Uploaded “Consultantdata.txt” under “employees_data” folder in FTP

```
hdfs dfs -put employees_data
```

```
ls
```

```
cd employees_data
```

```
ls
```

Now the “Consultantdata.txt” with 943 records is present under the folder “employees_data” in HDFS

To view full file in HDFS:

```
cat Consultantdata.txt
```

To view first 10 lines of the file in HDFS:

```
cat Consultantdata.txt | head -10
```

To view last 10 lines of the file in HDFS:

```
cat Consultantdata.txt | tail -10
```

Come back to home directory

```
cd ..
```

Entering HIVE

```
hive
```

```
use roopa;
```

Q2. Create an external Hive table "employees_Table" representing this "employees_data". This table will have 5 fields id,age,gender,role and salary.

```
CREATE EXTERNAL TABLE employees_Table (id int, age int, gender char (1), role string, salary int)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘|’ STORED AS textfile;
```

```
show tables;
```

```
describe employees_Table;
```

```
load data inpath ‘/user/roopa.sondur_outlook/employees_data/’ into table roopa.employees_Table;
```

Q3. create a new bucketed table "Consultant_Table_Bucket" having 4 buckets on the field salary. This table should store the data into columnar format ORC.

```
CREATE TABLE Consultant_Table_Bucket (id int, age int, gender char (1), role string) PARTITIONED BY
(salary int) CLUSTERED BY (role) into 4 buckets ROW FORMAT DELIMITED FIELDS TERMINATED BY  ‘|’
STORED AS orcfile;
```

Warning message

```
set hive.enforce.bucketing=true;
```

Q4. Insert all those employees whose salary is greater than 5000 into bucketed table "Consultant_Table_Bucket". While inserting into "Consultant_Table_Bucket" table you need to convert "consultant" role into "Big Data Consultant" role.

```
FROM employees_Table INSERT INTO Consultant_Table_Bucket PARTITION(salary) SELECT id, age,
gender, CASE WHEN role = ‘consultant’ THEN ‘Big Data Consultant’ ELSE role END AS role, salary
WHERE salary>5000;
```

Warning message

Did some configuration settings

```
set hive.exec.dynamic.partition.mode=nonstrict;

set hive.exec.max.dynamic.partitions=10000;

set hive.exec.max.dynamic.partitions.pernode=500;

FROM employees_Table INSERT INTO Consultant_Table_Bucket PARTITION(salary) SELECT id, age,
gender, CASE WHEN role = ‘consultant’ THEN ‘Big Data Consultant’ ELSE role END AS role, salary
WHERE salary>5000;
```

Q5. Write a Hive query to find out Max, min salary of "Big Data Consultant" from the "Consultant_Table_Bucket" table.

```
SELECT MAX(salary), MIN(salary) FROM Consultant_Table_Bucket WHERE role=‘Big Data Consultant’;
```

Output: MAX(salary) = 95403, MIN(salary) = 8052

```
Total MapReduce CPU Time Spent: 22 seconds 410 msec

OK

95403 8052

Time taken: 30.222 seconds, Fetched: 1 row(s)
```

The output which I got can be seen in Hue in the below path:

/user/hive/warehouse/roopa.db/consultant_table_bucket