**Importing XML Files into Hive**

**Hive:**

Apache Hive is an open-source data warehouse system for querying and analysing large datasets stored in Hadoop files. Hadoop is a framework for handling large datasets in a distributed computing environment.

**XML File Format:**

**XML** is a **file** extension for an Extensible Markup Language (**XML**) **file format** used to create common information **formats** and share both the **format** and the data on the World Wide Web, intranets, and elsewhere using standard ASCII text. **XML** is similar to HTML.

Let us consider the following data set stored as Data1.xml

<employee>

<id>1</id>

<name>Satish Kumar</name>

<designation>Technical Lead</designation> vb

</employee>

<employee>

<id>2</id>

<name>Ramya</name>

<designation>Testing</designation>

</employee>

**Procedure**:

**Step 1:**

Convert the File into a row formatted file where each record contains a record.

**Command**:

cat Desktop/Data.xml | tr -d '&' | tr '\n' ' ' | tr '\r' ' ' | sed 's| </employee>|</employee>\n|g' | grep -v '^\s\*$' > Data.xml

**Result**:

<employee><id>1</id><name>Satish Kumar</name><designation>Technical Lead</designation></employee>

<employee><id>2</id><name>Ramya</name><designation>Testing</designation>

</employee>

**Step 2:**

Move the data into HDFS using the following command.

**Command**:

Hadoop fs –put Data.xml /user/XML

**Step 3:**

Open Hive.

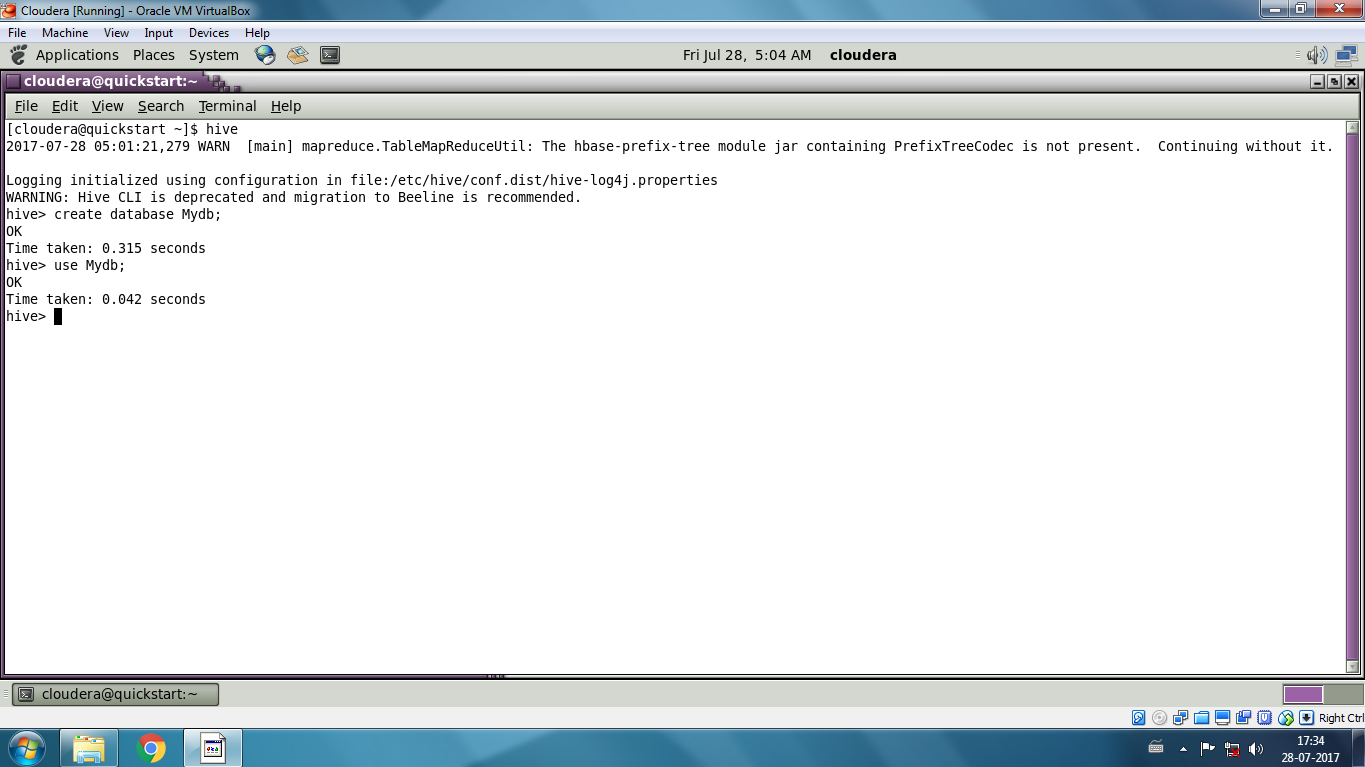
**Command:**

Hive

**Step 4:**

Create or use a database using the following command.

**Command**:

Use Mydb;

**Step 4:**

Create a duplicate table using the following command.

**Command:**

Create table duplicate (data String);

**Step 5:**

Move Data into the table using the following command.

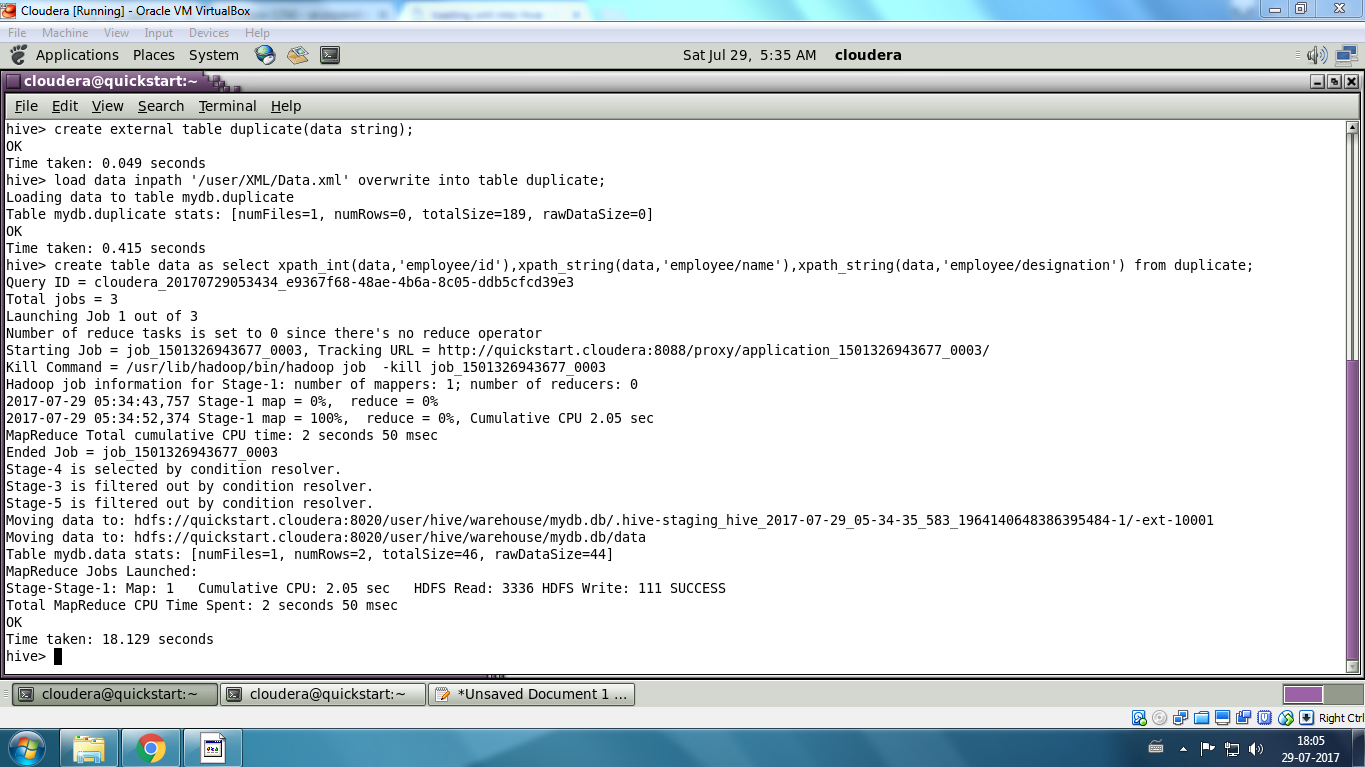
**Command:**

Load data inpath ‘/user/XML/Data.xml’ overwrite into table duplicate;

**Step 6:**

Create the actual table from the data in the duplicate table using the following command.

**Command**:

Create table Data as select xpath\_int (data,’employee/id’), xpath\_string (data,’employee/name’), xpath\_string (data,’employee/name’) from duplicate;

**Step 6:**

View the data in the table using the following commands.

Select \* from Data;