Sqoop Import Module-4

- **Importing subset of data from RDBMS**
- **Changing the delimiter during import**
- **Encoding Null values**
- **Importing entire schema or all tables**

1. Connecting to existing MySQL Database

```
mysql --user=retail_dba --password=cloudera retail_db
```

2. Show all the available tables

```
show tables;
```

3. View/Count data from a table in MySQL

```
select * from categories;
```

4. Check the currently available data in HDFS directory

```
hdfs dfs -ls
```

5. Import Single table (Subset data)

```
sqoop import \
  --connect jdbc:mysql://quickstart:3306/retail_db \
  --username=retail_dba \
  --password=cloudera \
  --table=categories \
 --warehouse-dir=categories_subset \
  --where \`category_id\`=22
```

Note: Here the ` is the same you find on ~ key

6. Check the output partition

```
hdfs dfs -cat categories_subset/categories/part-m-00000
```

7. Change the selection criteria (Subset data)

```
sqoop import    \
--connect jdbc:mysql://quickstart:3306/retail_db    \
--username=retail_dba   \
--password=cloudera  \
--table=categories \
--warehouse-dir=categories_subset_2  \
--where \`category_id\`\>22
```

8. Check the output partition

```
hdfs dfs -cat categories_subset_2/categories/part-m-00000
```

9. Use between clause (Subset data)

```
sqoop import    \
--connect jdbc:mysql://quickstart:3306/retail_db  \
--username=retail_dba  \
--password=cloudera \
--table=categories \
--warehouse-dir=categories_subset_3   \
--where "\`category_id\` between 1 and 22"
```

10. Check the output partition

```
hdfs dfs -cat categories_subset_3/categories/part-m-00000
```

11. Changing the delimiter during import.

```
sqoop import    \
 --connect jdbc:mysql://quickstart:3306/retail_db \
 --username=retail_dba  \
--password=cloudera \
--table=categories \
--warehouse-dir=categories_subset_6  \
--where "\`category_id\` between 1 and 22" \
--fields-terminated-by='|'
```

12. Check the output partition

```
hdfs dfs -cat categories_subset_6/categories/part-m-00000
```

13. Selecting subset columns

```
sqoop import    \
 --connect jdbc:mysql://quickstart:3306/retail_db \
 --username=retail_dba  \
--password=cloudera \
--table=categories \
--warehouse-dir=categories_subset_col  \
--where "\`category_id\` between 1 and 22" \
--fields-terminated-by='|' \
--columns=category_name,category_id
```

14. Check the output partition

```
hdfs dfs -cat categories_subset_col/categories/part-m-00000
```

Encoding null values

15. Inserting record with null values

```
ALTER TABLE categories modify category_name varchar(45);
```

```
INSERT INTO categories values (59,8,NULL);
```

```
select * from categories;
```

16. Now encode this null value (Replacing null string with \N value

```
sqoop import    \
 --connect jdbc:mysql://quickstart:3306/retail_db \
 --username=retail_dba  \
--password=cloudera \
--table=categories \
--warehouse-dir=categories_subset_15  \
--where "\`category_id\` between 1 and 60" \
--fields-terminated-by='|' \
--null-string='N'
```

17. Inserting record with null values

```
ALTER TABLE categories modify category_department_id int(11);
```

```
INSERT INTO categories values (60,NULL,'TESTING');
```

```
select * from categories;
```

18. Encode non string null column

```
sqoop import    \
 --connect jdbc:mysql://quickstart:3306/retail_db \
 --username=retail_dba  \
--password=cloudera \
--table=categories \
--warehouse-dir=categories_subset_17  \
--where "\`category_id\` between 1 and 61" \
--fields-terminated-by='|' \
--null-string='N' \
--null-non-string='N'
```

19. View the content

```
hdfs dfs -cat categories_subset_17/categories/part-m-00003
```

20. Import all the tables from a schema

```
sqoop import-all-tables    \
```

```
--connect jdbc:mysql://quickstart:3306/retail_db \
 --username=retail_dba  \
--password=cloudera \
--warehouse-dir=categories_subset_all_tables
```

21. View the contents

```
hdfs dfs -ls categories_subset_all_tables
```

22. Cleanup

```
delete from categories where category_id in (59,60);
```

```
ALTER TABLE categories modify category_department_id int(11) NOT NULL;
ALTER TABLE categories modify category_name varchar(45) NOT NULL;
desc categories;
```

```
drop table categories;
```

Data Science certification really needs a good and in depth knowledge of statistics cum BigData Hadoop knowledge. It also require you to have good knowledge in like the main phases of the Data Analytics Lifecycle, analyzing and exploring data with R, statistics for model building and evaluation, the theory and methods of advanced analytics and statistical modeling, the technology and tools that can be used for advanced analytics, operationalizing an analytics project, and data visualization techniques. Successful candidates will achieve the EMC Proven Professional – Data Science Associate credential. Hence to clear the real exam it realy needs very well preparation. So HadoopExam Learning Resources brings Data Science Certification Simulator with 234 Practice Questions, which can help you to prepare for this exam in lesser time. **Practice - practice - practice**! The EMC:DS E20-007 Exam Simulator offers you the opportunity to take 4 sample Exams before heading out for the real thing. Be ready to succeed on exam day!

Upcoming Releases

1. Apache Spark Training
2. Apache Spark Certification material
3. MongoDB Certification Material
4. Android Certification
5. Java Certification
6. AWS Trainings
7. Data Science Training