EXPLORING THE POWER OF SINGLE NODE CLUSTER: SPARK COMMANDS

UNLOCKING THE POWER OF SPARK DDL COMMANDS:

SPARK DATA TYPES

SYNTACTICALLY HIVE AND SPARK SQL ARE ALMOST SAME.

SPARK METASTORE SUPPORTS ALL STANDARD DATA TYPES.

NUMERIC - INT, BIGINT, FLOAT ETC
ALPHA NUMERIC OR STRING - CHAR, VARCHAR, STRING
DATE AND TIMESTAMP - DATE, TIMESTAMP
SPECIAL DATA TYPES - ARRAY, STRUCT ETC
BOOLEAN - BOOLEAN

IF THE FILE FORMAT IS TEXT FILE WITH SPECIAL TYPES, THEN WE NEED TO CONSIDER OTHER CLAUSES UNDER DELIMITED ROW FORMAT (IF WE DON'T WANT TO USE DEFAULT DELIMITERS).

SPARK STARTER

```
SPARK = SPARKSESSION.
    BUILDER.
    CONFIG("SPARK.UI.PORT", "0").
    CONFIG("SPARK.SQL.WAREHOUSE.DIR", S"/USER/WAREHOUSE").
    ENABLEHIVESUPPORT.
    MASTER("YARN").
    APPNAME("SPARK SQL - GETTING STARTED").
    GETORCREATE

SPARK2-SQL \
    --MASTER YARN \
    --CONF SPARK.UI.PORT=0 \
    --CONF SPARK.SQL.WAREHOUSE.DIR=/USER/WAREHOUSE \
    --DATABASE DBNAME
```

EXTERNAL VS MANAGED

CREATE DATABASE DATA_DEMO LOCATION '/USER/UBUNTU/DATA_DEMO.DB'

SELECT CURRENT_DATABASE()

SPARK-SQL --DATABASE DATA_DEMO TO CONNECT TO SPECIFIC DATABASE

INSIDE DATABASE WE HAVE TABLES. BELOW ARE WAYS TO GET TABLE INFO

DESCRIBE TABLE_NAME;
DESCRIBE EXTENDED TABLE_NAME
DESCRIBE FORMATTED TABLE_NAME

SPARK TABLE CREATE AND MORE : DDL

DETERMINE TABLE TYPE BASED UP ON THE FILES THAT WILL BE COPIED TO THE TABLE.

FILE FORMAT IS DELIMITED TEXT FILE: UNDERSTAND FIELD DELIMITER AS WELL.

```
STORED AS 'FILE TYPE'
ROW FORMAT
   DELIMITED FIELDS TERMINATED BY 'DELIMITER'
   COLLECTION ITEMS TERMINATED BY 'TERMINATOR'
   MAP KEYS TERMINATED BY 'TERMINATOR'
```

```
LOAD DATA INPATH '/HDFS/FILE/PATH'
  INTO TABLE TRIAL_TBL

LOAD DATA INPATH '/HDFS/FILE/PATH'
  OVERWRITE INTO TABLE TRIAL_TBL
```

```
STORED AS
FILE FORMAT FOR TABLE STORAGE, COULD BE TEXTFILE, ORC, PARQUET, ETC.
```

EXTERNAL VS MANAGED

```
CREATE EXTERNAL TABLE TRIAL_TBL (
    COL_1 INT COMMENT 'UNIQUE ID',
    COL_2 STRING COMMENT 'DATE ',
) COMMENT 'TABLE TO SAVE DATA'
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
 LOCATION '/USER/UBUNTU/EXTERNAL/DATA.CSV'
```

```
CREATE EXTERNAL TABLE TRIAL_TBL (
    COL_1 INT COMMENT 'UNIQUE ID',
    COL_2 STRING COMMENT 'DATE ',
) COMMENT 'TABLE TO SAVE DATA'
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
```

MANAGED / EXTERNAL TABLES

EXTERNAL: EXTERNAL AND SPECIFY LOCATION AS PART OF CREATE TABLE

USE EXTERNAL TABLE WHEN SAME DATASET IS PROCESSED BY MULTIPLE FRAMEWORKS SUCH AS HIVE, PIG, SPARK ETC.

DROP MANAGED TABLE, IT WILL DELETE METADATA FROM METASTORE AS WELL AS DATA FROM HDFS.

DROP EXTERNAL TABLE, ONLY METADATA WILL BE DROPPED, NOT THE DATA.

DATASOURCES & STORAGE

USING DATA_SOURCE

DATA SOURCE IS THE INPUT FORMAT USED TO CREATE THE TABLE. DATA SOURCE CAN BE CSV, TXT, ORC, JDBC, PARQUET, ETC.

ROW FORMAT CAN BE CUSTOM SERIALIZER AND DESERIALIZER. THIS WAY OF CREATING TABLE IS CALLED HIVE FORMAT TABLE CREATION. THE NECESSARY JAR FILES HAVE TO KEPT IN THE CLUSTER USING THE ADD 'LOCATION' COMMAND

BUILT-IN SERDES

AVRO (HIVE 0.9.1 AND LATER)
ORC (HIVE 0.11 AND LATER)
REGEX
THRIFT
PARQUET (HIVE 0.13 AND LATER)
CSV (HIVE 0.14 AND LATER)
JSONSERDE (HIVE 0.12 AND LATER IN HCATALOG-CORE)

DDL MANUAL FOR HIVE

HTTPS://CWIKI.APACHE.ORG/CONFLUENCE/DISPLAY/HIVE/LANGUAGEMANUAL+DDL#LANGUAGEMANUALDDL-ROWFORMATS&SERDE

THANKS FOR WATCHING

PRACTICE

PRACTICE

LIKE

SHARE

SUBSCRIBE

PRACTICE

PRACTICE

HTTPS://GITHUB.COM/INSIGHTBUILDER