**SNOWFLAKE INTEGRATION WITH PYSPARK**

Snowflake’s Data Cloud is powered by an advanced data platform provided as Software-as-a-Service (SaaS). It enables data storage, processing, and analytic solutions that are faster, easier to use, and far more flexible than traditional offerings. The Snowflake data platform is not built on any existing database technology or “big data” software platforms such as Hadoop. Instead, Snowflake combines a completely new SQL query engine with an innovative architecture natively designed for the cloud. To the user, Snowflake provides all of the functionality of an enterprise analytic database, along with many additional special features and unique capabilities.

The following steps show integration of snowflake with pyspark.

Step 1: Download the necessary jar files which are spark-snowflake\_2.11-2.8.4-spark\_2.4.jar, snowflake-jdbc-3.12.17.jar

Step 2: Creating the table in a snowflake.

--creating a table in snowflake worksheet

create table regionforsf( region\_key number(38,0),region\_name varchar(35),r\_comment varchar(152) );

select count(\*) from regionforsf;

Row count

1 0

Step 3: Wrting a pyspark dataframe to the snowflake table

In pyspark

regiondf=spark.read.option("inferSchema","true").option("header","true" ).option("delimiter",'\t').csv("file:///home/ak/Downloads/retailprj/reg ion.csv")

>>> regiondf.count()

5

regiondf.write.format("net.snowflake.spark.snowflake").options(sfURL='h ttps://fua21222.us-east-1.snowflakecomputing.com/',sfUser = 'geoinsys123',sfPassword='geoinsys@123',sfDatabase='geo\_db',sfSchema='geo\_sch ema',sfRole='SYSADMIN').option("dbtable",

"regionforsf").mode("overwrite").save()

regiondf.count()

5

34

Step 4: Reading from a snowflake table into a pyspark dataframe

spark.read.format("net.snowflake.spark.snowflake").options(sfURL='https ://fua21222.us-east-1.snowflakecomputing.com/',sfUser='geoinsys123', sfPassword='geoinsys@123',sfDatabase='geo\_db',sfSchema='geo\_schema',sfRole ='SYSADMIN').option("dbtable", "regionforsf").load()

**. Hive migration to snowflake**

Step 1: Creating storage integration inside the snowflake worksheet

--creating storage integration

create storage integration snowflake\_hiveint\_bys3

type = external\_stage

storage\_provider = s3

enabled = true

storage\_aws\_role\_arn = 'arn:aws:iam::574997813424:role/geoinsysrole' storage\_allowed\_locations = ('s3://geoinsys/anand/snowflakein'); --for viewing the details

desc integration s3\_hiveint

Step 2: Download the required connectors, jars

-we need to download a connector and have to place in hive/lib

https://search.maven.org/classic/remotecontent?filepath=net/snowflake/snowflake-hive-metastore-connector/0.5.1/snowflake-hive-metastore connector-0.5.1.jar

35

Step 3: create a file named snowflake-config.xml under hive/conf folder and inlcude the following properties.

--snowflake-config.xml

<configuration>

<<property>

<name>snowflake.jdbc.username</name>

<value>geoinsys123</value>

</property>

<property>

<name>snowflake.jdbc.password</name>

<value>geoinsys@123</value>

</property>

<property>

<name>snowflake.jdbc.role</name>

<value>ACCOUNTADMIN</value>

</property>

<property>

<name>snowflake.jdbc.db</name>

<value>RETAIL\_DBASE</value>

</property>

<property>

<name>snowflake.jdbc.schema</name>

<value>RETAIL\_SCHEMA</value>

</property>

<property>

<name>snowflake.jdbc.connection</name>

<value>jdbc:snowflake://fua21222.us-east-1.snowflakecomputing.com </value>

</property>

<property>

<name>snowflake.hive-metastore-listener.integration</name> <value>snowflake\_hiveint\_bys3</value>

</property>

</configuration>

--mapred-site.xml

<property>

<name>mapred.job.tracker</name>

<value>localhost:9010</value>

</property>

<property>

<name>fs.s3a.impl</name>

<value>org.apache.hadoop.fs.s3a.S3AFileSystem</value>

</property>

Step 4: Create external table in hive on the specified S3 location

hive> set AKIAYLYEKLCYBE7LSJSO

hive> set WQEFMcm1yYfZBOmTHb0pZLUtz8JMMMwe9ulCzxUg

hive>create external table nation\_ext\_s3(n\_nationkey int, n\_name string, n\_regionkey int, n\_comment string) row format delimited fields terminated by '\t' stored as textfile location 's3a://geoinsys/anand/project'

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

hive>set hive.resultset.use.unique.column.names=false;

select \* from nation\_ext\_s3 limit 5;

The external table created in hive is reflected in snowflake.

--in snowflake worksheet

show external tables in geo\_db.geo\_schema;

--in snowflake worksheet

Select \* from geo\_db.geo\_schema.region\_ext;