**Scala Assignment**

**How to read a text file and write it into a parquet file using scala script and spark-submit**

**Step:-1 Write the code in Scala Script in IntelliJ IDE as follows**

package org.sageit.bigdata  
  
import org.apache.spark.sql.SparkSession  
import org.apache.spark.sql.types  
import org.apache.spark.sql.{SparkSession, SaveMode}  
object TextToParquet {  
 def main(args: Array[String]): Unit = {  
 *println*("hello world")  
  
 val spark = SparkSession.*builder*().appName("TextToParquet").master("local").getOrCreate()  
 val df = spark.read.text("hdfs://cxln1.c.thelab-240901.internal:8020/user/bigdatacloudxlab14968/ssv20230314/Technology.txt")  
 df.write.mode(SaveMode.*Overwrite*).parquet("hdfs://cxln1.c.thelab-240901.internal:8020/user/bigdatacloudxlab14968/ssv20230314/output.parquet")  
  
 spark.stop()  
  
 }  
}

**Text

Description automatically generated**

**Step: 2 Create a jar file in intelliJ**

* **Click on Project name where you have written your scala code in IntelliJ on rightside tab under Maven**
* **Click on Lifecycle**
* **Click on Package and select run and execute after build**
* **Click on install and select run and execute after build**
* **The jar file will be created and stored inside the target folder in your project folder located in your PC**

**Graphical user interface, text, application

Description automatically generated**

**Step:3 Export jar file to your hdfs folder in cloudxlab ambari**

**Graphical user interface, text, email, website

Description automatically generated**

**Step:4 run the script using Spark-submit**

spark-submit \

--master yarn \

--deploy-mode client \

--class org.sageit.bigdata.TextToParquet \

--num-executors 4 \

--executor-memory 4g \

--executor-cores 2 \

--jars /hdp/current/spark2-client/jars/spark2-hdp-yarn-archive.jar,/home/bigdatacloudxlab14968/saisri\_spark/SSV20230306-1.0-SNAPSHOT.jar \

--files hdfs://cxln1.c.thelab-240901.internal:8020/user/bigdatacloudxlab14968/ssv20230314/TextToParquet.scala \

hdfs://cxln1.c.thelab-240901.internal:8020/user/bigdatacloudxlab14968/ssv20230314/Technology.txt \ hdfs://cxln1.c.thelab-240901.internal:8020/user/bigdatacloudxlab14968/ssv20230314/output.parquet

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Graphical user interface

Description automatically generated

Graphical user interface, text, email, website

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

**Step-5: To read the data present in parquet file:**

* **Go to pyspark**
* **Load the data from parquet file to a dataframe using the following command:**

df = spark.read.parquet("/user/bigdatacloudxlab14968/ssv20230314/output.parquet/part-00000-5ad4c4ec-135e-424f-af54-3a877cff7a66.snappy.parquet")

* **Display the contents using df.show()**

Text

Description automatically generated