**Question 1 – it**

**Instructions/Data Description/Output Requirements**

* Data is available in local file system under - /home/cloudera/data\_prep/nyse2010-17 – comma delimited text files
* Fields (stockticker:string, transactiondate:string, openprice:float, highprice:float, lowprice:float, closeprice:float, volume:bigint)
* Convert file format to parquet
* Save it -/user/cloudera/TSS\_SPARK/question1/output

**Question 2 – nd**

**Instructions/Data Description/Output Requirements**

* Convert data-files stored at hdfs location -$HDFS\_DP/retail\_db/orders\_avro
* into parquet file using snappy compression and save in HDFS.
* Output Requirement:
* Result should be saved in -/user/cloudera/TSS\_SPARK/question2/output
* Output file should be saved as Parquet file in Snappy Compression.

**Question 3 – nd**

**Instructions/Data Description/Output Requirements**

* Convert data-files stored at hdfs location -$HDFS\_DP/retail\_db/customer\_part\_avro –
* into tab delimited file using gzip compression and save in HDFS.
* Output Requirement:
* Result should be saved in -/user/cloudera/TSS\_SPARK/question3/output
* Output file should be saved as tab delimited file in gzip Compression.
* Sample Output:
* 21 Andrew Smith
* 111 Mary Jons

**Question 4 – nd**

**Instructions/Data Description/Output Requirements**

* Fetch all pending orders from data-files stored at hdfs location
* $HDFS\_DP/retail\_db/orders\_parquet
* and save it into json file in HDFS
* Output Requirement:
* Result should be saved in -/user/cloudera/TSS\_SPARK/question4/output
* Output file should be saved as json file.
* Output file should Gzip compressed.

**Question 5 – nd**

**Instructions/Data Description/Output Requirements**

* Instructions:
* Create a metastore table from avro files provided at below location.
* Input folder is $HDFS\_DP/retail\_db/customer\_part2\_avro
* Output Requirement:
* Table name should be customer\_parquet\_avro
* output location for hive data -/user/cloudera/TSS\_SPARK/question5/output
* Output file should be saved in parquet format using GZIP compression.