Data Ingestion to EMR from RDS and S3

(Data Ingestion)

App events

sqoop import --connect jdbc:mysql://mlc-testcapstone.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/mlctest --table app_events --target-dir /user/hadoop/mlctest/app_events/ --username student -P -m1

Brand device

sqoop import --connect jdbc:mysql://mlc-testcapstone.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/mlctest --table brand_device --target-dir /user/hadoop/mlctest/brand_device/ --username student -P -m1

Events

sqoop import --connect jdbc:mysql://mlc-testcapstone.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/mlctest --table events --target-dir /user/hadoop/mlctest/events/ --username student -P -m1

Train

sqoop import --connect jdbc:mysql://mlc-testcapstone.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/mlctest --table train --target-dir /user/hadoop/mlctest/train/ --username student -P -m1

App Labels

hadoop distcp s3a://capstone-project-mlc-metadata/app_labels_new.txt hdfs://ec2-3-87-222-170.compute-1.amazonaws.com:8020/home/hadoop/

```
[hadoop@ip-172-31-86-108 ~]$ hadoop distcp s3a://capstone-project-mlc-metadata/app_labels_new.txt hdfs://ec2-3-87-222-170.compute-1.amazonaws.com:8020/home/h
adoop/
21/01/11 17:37:15 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=
false, skipCRC=false, blocking=true, numListstatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveS
tatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3a://capstone-project-mlc-metadata/app_labels_new
.txt], targetPath=hdfs://ec2-3-87-222-170.compute-1.amazonaws.com:8020/home/hadoop, targetPathExists=true, filtersFile='null'}
21/01/11 17:37:15 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-86-108.ec2.internal/172.31.86.108:8032
21/01/11 17:37:17 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
21/01/11 17:37:17 INFO tools.SimpleCopyListing: Build file listing completed.
21/01/11 17:37:17 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
21/01/11 17:37:17 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
21/01/11 17:37:17 INFO tools.DistCp: Number of paths in the copy list: 1
21/01/11 17:37:17 INFO tools.DistCp: Number of paths in the copy list: 1
21/01/11 17:37:17 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-86-108.ec2.internal/172.31.86.108:8032
21/01/11 17:37:18 INFO mapreduce.JobSubmitter: number of splits:1
21/01/11 17:37:18 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1610340234319_0030
21/01/11 17:37:18 INFO impl.YarnClientImpl: Submitted application application_1610340234319_0030
21/01/11 17:37:19 INFO mapreduce.Job: The url to track the job: http://ip-172-31-86-108.ec2.internal:20888/proxy/application_1610340234319_0030/
21/01/11 17:37:19 INFO tools.DistCp: DistCp job-id: job_1610340234319_0030
21/01/11 17:37:19 INFO mapreduce.Job: Running job: job_1610340234319_0030
```

Label Categories

hadoop distcp s3a://capstone-project-mlc-metadata/label_categories.csv hdfs://ec2-3-87-222-170.compute-1.amazonaws.com:8020/home/hadoop/

```
[hadoop@ip-172-31-86-108 ~]$ hadoop distcp s3a://capstone-project-mlc-metadata/label_categories.csv hdfs://ec2-3-87-222-170.compute-1.amazonaws.com:8020/home
/hadoop/
21/01/11 17:40:40 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=
false, skipCRC=false, blocking=true, numListstatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveS
tatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3a://capstone-project-mlc-metadata/label_categori
es.csv], targetPath=hdfs://ec2-3-87-222-170.compute-1.amazonaws.com:8020/home/hadoop, targetPathExists=true, filtersFile='null'}
21/01/11 17:40:41 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-86-108.ec2.internal/172.31.86.108:8032
21/01/11 17:40:43 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
21/01/11 17:40:43 INFO tools.SimpleCopyListing: Build file listing completed.
21/01/11 17:40:43 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
21/01/11 17:40:43 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
21/01/11 17:40:43 INFO tools.DistCp: Number of paths in the copy list: 1
21/01/11 17:40:43 INFO tools.DistCp: Number of paths in the copy list: 1
21/01/11 17:40:43 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-86-108.ec2.internal/172.31.86.108:8032
21/01/11 17:40:43 INFO mapreduce.JobSubmitter: number of splits:1
```

Listing files which loaded from s3

```
[hadoop@ip-172-31-86-108 ~]$ hadoop fs -ls /home/hadoop/
Found 3 items
-rw-r--r--   1 hadoop hadoop   11190003 2021-01-11 17:37 /home/hadoop/app_labels_new.txt
-rw-r--r--   1 hadoop hadoop      16450 2021-01-11 17:40 /home/hadoop/label_categories.csv
drwxr-xr-x   - hadoop hadoop          0 2021-01-11 15:41 /home/hadoop/mlctest
[hadoop@ip-172-31-86-108 ~]$
```