Resume from where it failed last time

Job resumes from where it left last time from the entries put in audit table.

// originally designed to get data from one redshift table, now designed to get a file from https and load it.

Process- Main class

Audit table

Snapshot entire data for today.

(HISTORY SNAPSHOT table)

Load it for today

(CURRENT table)

Save it in hive tmp table (RAW table)

Source [HiveInserthttps or redshift]

RAW temp table

Mail

PIPELINE 1

PIPELINE 2

"wgetRead" -> reads data from external source

"temp\_archival" -> archives the read data it in temp table (over written in the next run)

"HiveInsert"->inserts the archived data in pipeline 1

"HiveSnapshotArchival"-> entire pipeline 1 table is taken snapshot and written in pipeline2

“mailStatus” -> sends status till where the job has ran and what is that steps status

Flow:

Main class is the one which needs to be triggered,

Checks audit table for run date. If no entry is present it starts from the beginning. If entry is present then checks if the last entered entry is a success , then it runs from beginning again.

If previous job had failed at **wgetRead**, then it runs from beginning again. **with new job run id**

If previous job had failed at **archival**, then it runs from beginning again. **with new job run id**

If previous job had failed at **HiveInsert**, Reads data from hive archived table and resumes from there. **with new job run id**

If previous job had failed at **HiveSnapshotArchival**, Reads data from pipeline1 table and resumes from there. **with new job run id**

If previous job had failed at mailStatus, Reads data from audit table and resend’s the data, if **HiveSnapshotArchival**  is finished for last job, then sends mail **with SAME job run id** and puts success entry for mail. Else sends fishy scenario mail and puts success entry for mail.

**// mails the status afterwards**

**// runs fresh without entry**

spark-submit --packages javax.mail:mail:1.4.7 --class org.controller.reRunnableJob.reRunnableMain --driver-memory 512m --driver-cores 2 --executor-memory 512m --executor-cores 2 --num-executors 2 file:///home/raptor/IdeaProjects/SparkLearning/build/libs/SparkLearning-1.0-SNAPSHOT.jar runDate="2020-08-02" pwd=""

// pwd for mail not given here

**//Runs from start if the previous ran job is success, even subjob is success, with new job run id**

spark-submit --packages javax.mail:mail:1.4.7 --class org.controller.reRunnableJob.reRunnableMain --driver-memory 512m --driver-cores 2 --executor-memory 512m --executor-cores 2 --num-executors 2 file:///home/raptor/IdeaProjects/SparkLearning/build/libs/SparkLearning-1.0-SNAPSHOT.jar runDate="2020-08-02" pwd=""

// pwd for mail not given here

---------

spark.sql("insert into tempDb.temp\_audit\_table values('temp\_load','HiveInsert',current\_timestamp(),current\_timestamp(),'2020-08-02','FAILED',200000908909832)")

**// Runs from hive insert, with new job run id**

spark-submit --packages javax.mail:mail:1.4.7 --class org.controller.reRunnableJob.reRunnableMain --driver-memory 512m --driver-cores 2 --executor-memory 512m --executor-cores 2 --num-executors 2 file:///home/raptor/IdeaProjects/SparkLearning/build/libs/SparkLearning-1.0-SNAPSHOT.jar runDate="2020-08-02" pwd=""

// pwd for mail not given here

-------

spark.sql("insert into tempDb.temp\_audit\_table values('temp\_load','temp\_archival',current\_timestamp(),current\_timestamp(),'2020-08-02','FAILED',200000908909832)")

**// Runs from wget, if temp archival failed new data will be feed, with new job run id**

spark-submit --packages javax.mail:mail:1.4.7 --class org.controller.reRunnableJob.reRunnableMain --driver-memory 512m --driver-cores 2 --executor-memory 512m --executor-cores 2 --num-executors 2 file:///home/raptor/IdeaProjects/SparkLearning/build/libs/SparkLearning-1.0-SNAPSHOT.jar runDate="2020-08-02"pwd=""

// pwd for mail not given here

spark.sql("insert into tempDb.temp\_audit\_table values('temp\_load','mailStatus',current\_timestamp(),current\_timestamp(),'2020-08-02','FAILED',200000908909832)")

**// Sends success mail with SAME job run id as previous job**

spark-submit --packages javax.mail:mail:1.4.7 --class org.controller.reRunnableJob.reRunnableMain --driver-memory 512m --driver-cores 2 --executor-memory 512m --executor-cores 2 --num-executors 2 file:///home/raptor/IdeaProjects/SparkLearning/build/libs/SparkLearning-1.0-SNAPSHOT.jar runDate="2020-08-02"pwd=""

// pwd for mail not given here