Data Collection Ingestion Pipeline

//creating kafka topic, pushing data

244

sh /home/raptor/IdeaProjects/KafkaGradleTest/kafkaPipeLineScripts/createTopic.sh IOTdata 2 2

sh /home/raptor/IdeaProjects/KafkaGradleTest/kafkaPipeLineScripts/randomMsgProducer.sh fol 4 50 IOTdata 2020 mugello fp1~2020-08-01

sh /home/raptor/IdeaProjects/KafkaGradleTest/kafkaPipeLineScripts/randomMsgProducer.sh col 4 50 IOTdata 2020 mugello fp1~2020-08-02

sh /home/raptor/IdeaProjects/KafkaGradleTest/kafkaPipeLineScripts/randomMsgProducer.sh sol 4 50 IOTdata 2020 mugello fp1~2020-08-03

sh /home/raptor/IdeaProjects/KafkaGradleTest/kafkaPipeLineScripts/randomMsgProducer.sh fol 4 50 IOTdata 2020 mugello fp2~2020-08-04

sh /home/raptor/IdeaProjects/KafkaGradleTest/kafkaPipeLineScripts/randomMsgProducer.sh col 4 50 IOTdata 2020 mugello fp2~2020-08-05

sh /home/raptor/IdeaProjects/KafkaGradleTest/kafkaPipeLineScripts/randomMsgProducer.sh sol 4 50 IOTdata 2020 mugello fp2~2020-08-06

sh /home/raptor/IdeaProjects/KafkaGradleTest/kafkaPipeLineScripts/randomMsgProducer.sh fol 4 50 IOTdata 2020 mugello fp3~2020-08-07

// deleting and checking topic

spark-submit --class org.util.kafka.deleteTopic --packages org.apache.kafka:kafka-clients:2.3.1,org.apache.kafka:kafka\_2.12:2.3.1,org.apache.spark:spark-sql-kafka-0-10\_2.12:2.4.4 /home/raptor/IdeaProjects/SparkLearning/build/libs/SparkLearning-1.0-SNAPSHOT.jar topicName=IOTdata zookeeperIp=localhost:3039 isSecure=false sessionTimeOutMilliSecs=200000 connectionTimeOutMilliSecs=15000 maxInFlightRequests=20 metricGroup=myGroup metricType=myType sleepMsForKafkaToDeleteTopic=60000

spark-submit --class org.util.kafka.checkIfTopicExists --packages org.apache.kafka:kafka-clients:2.3.1,org.apache.kafka:kafka\_2.12:2.3.1,org.apache.spark:spark-sql-kafka-0-10\_2.12:2.4.4 /home/raptor/IdeaProjects/SparkLearning/build/libs/SparkLearning-1.0-SNAPSHOT.jar topicName=IOTdata zookeeperIp=localhost:3039 isSecure=false sessionTimeOutMilliSecs=200000 connectionTimeOutMilliSecs=15000 maxInFlightRequests=20 metricGroup=myGroup metricType=myType

Fix Job

Batch

Streaming

Arch :

Bronze

Silver(bronze phase 2)

Kafka Topic

Audit table

Mail

Final Mail

Stats check

Batch

Gold (Ranked and Aggregation)

// reading from kafka topic

300 spark and hive 230

**// this reads data from queue and constantly pushes it into bronze (streaming)**

spark-submit --packages io.delta:delta-core\_2.12:0.7.0,org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0,javax.mail:mail:1.4.7 --class org.controller.ingestionAndDQC.readFromKafkaAndSaveToBronze --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 topic=IOTdata checkpointFolderNameRead=fromKafkaToReadStream checkpointFolderNameWrite=fromReadStreamToBatch

**// this reads data from bronze to silver (this is a batch job), this is just like bronze phase 2 only compaction happens here.**

**// previous day records are compacted and pushed into silver (bronze phase 2) in here, (processingDate -1) day's records will be processed here**

spark-submit --packages io.delta:delta-core\_2.12:0.7.0,org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0,javax.mail:mail:1.4.7 --class org.controller.ingestionAndDQC.bronzeToSilverWithCompaction --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 processingDate="2020-08-05"

**// stats for previous day data and current day data from source (bronze) vs silver**

**// stats computed for all days from processingDate-1 to history date's. If history dates recieved new data it'll be noted and mail module will send mail as failed for which dates extra data has been recieved**

**// daily and fix runtypes are available .If fix (ran after fixing extra records to silver by fix job), checks if current day bronze and silver matches, then assigns 0 to prev bronze vs curr bronze column, if daily computes stats as it is**

spark-submit --packages io.delta:delta-core\_2.12:0.7.0,org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0,javax.mail:mail:1.4.7 --class org.controller.ingestionAndDQC.bronzeVsSilverStats --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 processingDate="2020-08-05" runType=daily

**// mails the output of stats , if there is extra records in bronze for a particular date it mails the entire date's stats as output.If no isses success mail is sent**

spark-submit --packages io.delta:delta-core\_2.12:0.7.0,org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0,javax.mail:mail:1.4.7 --class org.controller.ingestionAndDQC.mailBronzeVsSilverStats --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 processingDate="2020-08-05" pwd="pwd"

**//fixing extra records**

**//If history dates recieved new data it'll be noted and mail module will send mail as failed for which dates extra data has been recieved**

**// we need to give the date in the stats table in which the failure stats is present, it'll read the latest job id and take the "extra data to be copied date" and does incremental data load with compaction(not full load, incremental)**

spark-submit --packages io.delta:delta-core\_2.12:0.7.0,org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0,javax.mail:mail:1.4.7 --class org.controller.ingestionAndDQC.fixExtraDataFromBronze --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 inputDateFromStatsTable="2020-08-05"

**// gold, does some aggregation and ranking and saves it to two different base tables**

**// overwrites if its already present for the processing date**

spark-submit --packages io.delta:delta-core\_2.12:0.7.0,org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0,javax.mail:mail:1.4.7 --class org.controller.ingestionAndDQC.silverToGold --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 processingDate="2020-08-02"

**Sample Pipeline flow tested:**

Data comes in for 1st,2nd,3rd,4th. Every day batch get’s processed. For fifth extra data comes for third. Mail is sent and then fix is done.

//244

sh /home/raptor/IdeaProjects/KafkaGradleTest/kafkaPipeLineScripts/createTopic.sh IOTdata 2 2

//300

spark-submit --packages io.delta:delta-core\_2.12:0.7.0,org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0,javax.mail:mail:1.4.7 --class org.controller.ingestionAndDQC.readFromKafkaAndSaveToBronze --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 topic=IOTdata checkpointFolderNameRead=fromKafkaToReadStream checkpointFolderNameWrite=fromReadStreamToBatch

//244

sh /home/raptor/IdeaProjects/KafkaGradleTest/kafkaPipeLineScripts/randomMsgProducer.sh fol 4 50 IOTdata 2020 mugello fp1~**2020-08-01**

**// batch loads(1st data into silver)**

spark-submit --packages io.delta:delta-core\_2.12:0.7.0,org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0,javax.mail:mail:1.4.7 --class org.controller.ingestionAndDQC.bronzeToSilverWithCompaction --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 processingDate="2020-08-02"

**// stats for previous day data and current day data from source (bronze)**

spark-submit --packages io.delta:delta-core\_2.12:0.7.0,org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0,javax.mail:mail:1.4.7 --class org.controller.ingestionAndDQC.bronzeVsSilverStats --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 processingDate="2020-08-02" runType=daily

**// mails the output of stats , if there is extra records in bronze for a particular date it mails the entire date stats as output.// need to do an incremental update on those date records in future**

spark-submit --packages io.delta:delta-core\_2.12:0.7.0,org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0,javax.mail:mail:1.4.7 --class org.controller.ingestionAndDQC.mailBronzeVsSilverStats --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 processingDate="2020-08-02" pwd="Thulz@1996"

//244

sh /home/raptor/IdeaProjects/KafkaGradleTest/kafkaPipeLineScripts/randomMsgProducer.sh fol 4 50 IOTdata 2020 mugello fp1~**2020-08-02**

**// batch loads(2st data into silver)**

spark-submit --packages io.delta:delta-core\_2.12:0.7.0,org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0,javax.mail:mail:1.4.7 --class org.controller.ingestionAndDQC.bronzeToSilverWithCompaction --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 processingDate="2020-08-03"

**// stats for previous day data and current day data from source (bronze)**

spark-submit --packages io.delta:delta-core\_2.12:0.7.0,org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0,javax.mail:mail:1.4.7 --class org.controller.ingestionAndDQC.bronzeVsSilverStats --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 processingDate="2020-08-03" runType=daily

**// mails the output of stats , if there is extra records in bronze for a particular date it mails the entire date stats as output.// need to do an incremental update on those date records in future**

spark-submit --packages io.delta:delta-core\_2.12:0.7.0,org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0,javax.mail:mail:1.4.7 --class org.controller.ingestionAndDQC.mailBronzeVsSilverStats --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 processingDate="2020-08-03" pwd="Thulz@1996"

//244

sh /home/raptor/IdeaProjects/KafkaGradleTest/kafkaPipeLineScripts/randomMsgProducer.sh fol 4 50 IOTdata 2020 mugello fp1~**2020-08-03**

**// batch loads(3st data into silver)**

spark-submit --packages io.delta:delta-core\_2.12:0.7.0,org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0,javax.mail:mail:1.4.7 --class org.controller.ingestionAndDQC.bronzeToSilverWithCompaction --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 processingDate="2020-08-04"

**// stats for previous day data and current day data from source (bronze)**

spark-submit --packages io.delta:delta-core\_2.12:0.7.0,org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0,javax.mail:mail:1.4.7 --class org.controller.ingestionAndDQC.bronzeVsSilverStats --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 processingDate="2020-08-04" runType=daily

**// mails the output of stats , if there is extra records in bronze for a particular date it mails the entire date stats as output.// need to do an incremental update on those date records in future**

spark-submit --packages io.delta:delta-core\_2.12:0.7.0,org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0,javax.mail:mail:1.4.7 --class org.controller.ingestionAndDQC.mailBronzeVsSilverStats --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 processingDate="2020-08-04" pwd="Thulz@1996"

//244

sh /home/raptor/IdeaProjects/KafkaGradleTest/kafkaPipeLineScripts/randomMsgProducer.sh fol 4 50 IOTdata 2020 mugello fp1~**2020-08-04**

sh /home/raptor/IdeaProjects/KafkaGradleTest/kafkaPipeLineScripts/randomMsgProducer.sh fol 4 50 IOTdata 2020 mugello fp1~**2020-08-03 // extra data for yesterday (Job must fail and fix must happen)**

**// batch loads(1st data into silver)**

spark-submit --packages io.delta:delta-core\_2.12:0.7.0,org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0,javax.mail:mail:1.4.7 --class org.controller.ingestionAndDQC.bronzeToSilverWithCompaction --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 processingDate="2020-08-05"

**// stats for previous day data and current day data from source (bronze)**

spark-submit --packages io.delta:delta-core\_2.12:0.7.0,org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0,javax.mail:mail:1.4.7 --class org.controller.ingestionAndDQC.bronzeVsSilverStats --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 processingDate="2020-08-05" runType=daily

**// mails the output of stats , if there is extra records in bronze for a particular date it mails the entire date stats as output.// need to do an incremental update on those date records in future**

spark-submit --packages io.delta:delta-core\_2.12:0.7.0,org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0,javax.mail:mail:1.4.7 --class org.controller.ingestionAndDQC.mailBronzeVsSilverStats --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 processingDate="2020-08-05" pwd="Thulz@1996"

**// failure attachment**

**//fixing extra records**

**//If history dates recieved new data it'll be noted and mail module will send mail as failed for which dates extra data has been recieved**

**// we need to give the date in the stats table in which the failure stats is present, it'll read the latest job id and take the "extra data to be copied date" and does incremental data load with compaction(not full load, incremental)**

spark-submit --packages io.delta:delta-core\_2.12:0.7.0,org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0,javax.mail:mail:1.4.7 --class org.controller.ingestionAndDQC.fixExtraDataFromBronze --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 inputDateFromStatsTable="2020-08-05"

**// stat’s table saved input of failure in 5th .Takes failure dates from the stats table automatically**

**// stats after the fix has happened**

spark-submit --packages io.delta:delta-core\_2.12:0.7.0,org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0,javax.mail:mail:1.4.7 --class org.controller.ingestionAndDQC.bronzeVsSilverStats --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 processingDate="2020-08-05" runType=**fix**

//////////////////////////////////////////////////////////

Silver to gold and final status mail.

**// daily job, reads data from silver, does aggregation and ranking and stores in 2 diff location**

**Loads day -1 data**

spark-submit --packages io.delta:delta-core\_2.12:0.7.0,org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0,javax.mail:mail:1.4.7 --class org.controller.ingestionAndDQC.silverToGold --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 processingDate="2020-08-02"

**//checks for previously ran job for silver to gold in current day and sends the status as mail. If failed sends mail with attachment**

spark-submit --packages io.delta:delta-core\_2.12:0.7.0,org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0,javax.mail:mail:1.4.7 --class org.controller.ingestionAndDQC.mailFinalFinishStatus --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 processingDate="2020-08-03" pwd="Thulz@1996"