Scala Programming Language Project

Data management track iti intake 41

Yusuf M. khalifa

2021

# Business Use Case Definition:

Telecom Company X is aiming to increase the data usage of its customer base as a key objective for 2021.

So X decided to target their data customer base with specialized campaigns to upsell their data packages.

X also partnered up with various famous applications to give the customers extra promotions to increase the traffic on the applications.

X is looking for aspiring data engineers to transform such strategic decision into a reality. Aspiring data engineers will work with streaming data capturing the data usage traffic for X customers.

Since the customer base generates ridiculous amount of traffic, X is only interested in a subset of the base “Segment” to try the new initiative on.

To insure a good customer experience, X wants to target the segment with **only one campaign a day per application**.

# Data Definition:

1. DATA\_yyyyMMddHHmm.csv
   1. Contains a data record per line indicating one session. Field are ordered as the following:
      1. Msisdn: customer dial.
      2. Service Identifier: Id for the service/application
      3. Traffic Volume: amount of download traffic used in the session
      4. Timestamp : timestamp of the transaction
2. SEGMENT.csv
   1. Contains list of dials X Company interested to track their usage
3. RULES.csv
   1. Contains the rules which defines the rules for sending a notification per service identifier/application. Fields are ordered as the following:
      1. Service Identifier: Id for the service/application
      2. Service Description : Service/application name
      3. Start time : the lower bound of usage tracking window
      4. End time : the upper bound of usage tracking window
      5. Total volume : the volume amount that needs to be exceeded to mark a valid rule to send the notification

Data Generation

(Create directory paths before running the commands):

1. Generate the rules   
   
2. Generate the segment



1. Generate the streaming data



# Deliverables:

One [fat jar](https://stackoverflow.com/questions/19150811/what-is-a-fat-jar) which takes multiple parameters in this following order

1. Data directory path
2. Segment data directory path
3. Rules data directory path
4. Output directory path – One directory per application ( Contains the records which passed rules criteria successfully )

The program needs to have the following characteristics:

1. Fault tolerant (input errors shouldn’t break down your software)
2. Smooth Recovery (start processing again from the last checkpoint)

You are eligible to use your favorite framework/s to achieve the task.

# Sample input & Output:

1. Data

0124637855,5,367,20200417193743

0127368267,7,421,20200417193743

**0126545098**,1,315,20200417193743

0123979727,5,16,20200417193743

0127074387,6,111,20200417193743

**0126545098**,1,372,20200417193744

**0126545098**,2,222,20200417193744

1. Segment

0123917827

**0126545098**

0128419091

1. Rules

1,Twitter,8,20,147

2,WhatsApp,9,18,68

3,Youtube,10,20,99

1. Output

0126545098,1,315,20200417193743

0126545098,2,222,20200417193744