**Name:** cust\_data

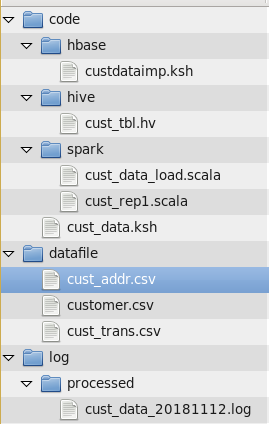
This process will load .csv file from hdfs location and combine data source after filtered null values and save it hdfs location and load data in hbase. By using hbase table load data in spark and processing report

e.g. Extract customers within an income range (eg. 100k-150k) and spent at least $1000 in a month (eg. 201809)

To run above process execute shell script cust\_data.ksh.

**Deployment Instruction**

1. Create following folders in local directory under customer and copy source code in respective directory



1. Create following folder in HDFS

* customer
* inside customer create following folder
  + output
  + processed

1. create table in HBase

Perform following steps

1. $hbase shell
2. hbase(main):008:0> create ‘customer’, ‘cust\_info’
3. hbase(main):008:0> create ‘customer\_trans’,’cust\_trans’
4. create external tables in hive reference by hbase

$hive –f /customer/code/hive/cust\_tbl.sql

1. copy/move data file to hdfs directory

$ cd /customer/datafile

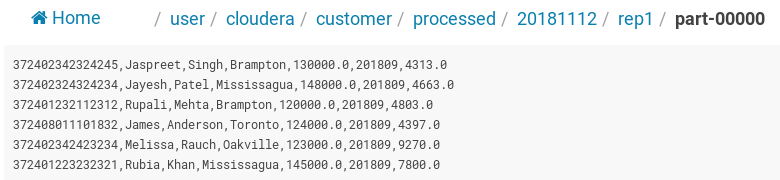
$hdfs dfs -copyFromLocal \*.csv customer/

1. run shell script

$ cd /customer/code

$./cust\_data.ksh –s 1,4

**Note: I am facing issue to setup scala on ecplise in my linux VM so here I have used spark-script instead of API. In production environment we need to use spark-api instead of spark script and for deployment need to use spark-submit command**

1. after successful execution of all 4 steps following file generated
   1. Report output generated in customer/processed/20181112/rep/part-00000 

Note: initially data is generated in customer/output/rep1/part-0000 once stage 4 runs in shell script data moved to customer/processed/20181112 (today’s date) folder

* 1. Log is generate in local folder /customer/log/processed/cust\_data\_20181112.log file

