DS503 Big Data Management

Project 1

Tianyu Li(tli) & Xiaosong Wen (xwen2)

There are three folders in our zip: Java, MapReduce and Pig.

The Java folder contains java programs for customer and transaction, and the main class to generate 50,000 customers and 5M transactions with random value. The result will be saved in Customers.csv and Transactions.csv file. These codes are using Java 1.6.

Use the command:

hdfs dfs -put ~/project1/\*.csv /user/hadoop/input/

To uploading the two csv files into Hadoop file system

The MR folder contains java code for the MapReduce Queries. The version of Hadoop is 3.1.2

The first query need input/Customers.csv as input argument;

The second and third queries need input/Customers.csv and input/Transactions.csv as input argument;

The last two quries need input/Transactions.csv as input argument;

The Pig folder contains four pig query scripts. They all take the Customers.csv and Transactions.csv generated from the previous Java folder as Input. The version of PIG is 0.13.