Assignment 4

Reduce side Join Analytics:

A common situation in many companies is that transaction records are kept separate from the customer data. There is, of course, a relationship between the two; usually a transaction record contains the unique ID of the customer through which the sale was performed.

In the Hadoop world, these would be represented by two types of data files: one containing records of the customer IDs and information for transactions (ERP data file provided with this problem), and the other would contain the full data for each customer (CRM data file provided with this problem).

Frequent tasks require reporting that uses data from both these sources; say, for example, we wanted to see the total number of transactions and total value for customer but do not want to associate it with an anonymous ID number, but rather with a name. This may be valuable when customer service representatives wish to call the most frequent customers—data from the sales records—but want to be able to refer to the person by name and not just a number.

We can perform the report explained in the previous section using a reduce-side join.

Expected steps:

1. Copy CRM file into HDFS directory say “/CRM” and copy ERP file into HDFS directory say “/ERP” and run below command
2. To run this program you need to pass 3 parameters CRM file directory location, ERP directory location and output file directory location

Example:

To run this program

```
hadoop jar Zigsaw_ReduceJoin.jar edu.zigsaw.ReduceJoin  /CRM  /ERP
/output_reduce
```