**Activity 3**

Let us look at another example, but this time, the data will be read from a CSV file instead of a text file.

The CSV file that we will be using contains the following data:

|  |  |
| --- | --- |
|  | Product,Price,Payment\_Type,Name,City,State,Country |
|  | Product1,1200,Mastercard,carolina,Basildon,England,United Kingdom |
|  | Product1,1200,Visa,Betina,Parkville,MO,United States |
|  | Product1,1200,Mastercard,Federica e Andrea,Astoria,OR,United States |
|  | Product1,1200,Visa,Gouya,Echuca,Victoria,Australia |
|  | Product2,3600,Visa,Gerd W ,Cahaba Heights,AL,United States |
|  | Product1,1200,Visa,LAURENCE,Mickleton,NJ,United States |
|  | Product1,1200,Mastercard,Fleur,Peoria ,IL,United States |
|  | Product1,1200,Mastercard,adam,Martin ,TN,United States |
|  | Product1,1200,Mastercard,Renee Elisabeth,Tel Aviv,Tel Aviv,Israel |
|  | Product1,1200,Visa,Aidan,Chatou,Ile-de-France,France |
|  | Product1,1200,Diners,Stacy,New York ,NY,United States |
|  | Product1,1200,Amex,Heidi,Eindhoven,Noord-Brabant,Netherlands |
|  | Product1,1200,Mastercard,Sean ,Shavano Park,TX,United States |
|  | Product1,1200,Visa,Georgia,Eagle,ID,United States |

To create the file, use the following command:

# Copy the contents of the file shown in the previous slide into this

$ vim sales.csv

# Once the file is ready, put it in the HDFS

$ hdfs dfs -put ./sales.csv /user/root/

First, to load the input file in the script, we use the LOAD command:

-- Load the CSV file

salesTable = LOAD 'hdfs:///user/root/sales.csv' USING PigStorage(',') AS (Product:chararray,Price:chararray,Payment\_Type:chararray,Name:chararray,City:chararray,State:chararray,Country:chararray);

In this line, we are loading a CSV file in from the HDFS, therefore, we are also setting the delimiter used.

We are also, assigning the datatypes for each of the columns that are present in the CSV file. **chararray in Pig is the same as String in Java**.

In the next line, we will group the data using the column that we want the data of. In this example, we are using the country column:

-- Group data using the country column

GroupByCountry = GROUP salesTable BY Country;

In the next line, for each tuple in 'GroupByCountry', we are generating the resulting string of the form of: Name of Country: No. of products sold

-- Generate result format

CountByCountry = FOREACH GroupByCountry GENERATE CONCAT((chararray)$0, CONCAT(':', (chararray)COUNT($1)));

And finally, we are storing the result in the HDFS with the STORE command:

-- Save result in HDFS folder

STORE CountByCountry INTO 'salesOutput' USING PigStorage('\t');

salesCSV.pig (cntd)

This line creates a directory called **"salesOutput"** in your HDFS user's home directory.

i.e., if the pig script is executed as the **root** user, then it creates **/user/root/salesOutput/** on the HDFS.

-- Load the CSV file

salesTable = LOAD 'hdfs:///user/root/sales.csv' USING PigStorage(',') AS (Product:chararray,Price:chararray,Payment\_Type:chararray,Name:chararray,City:chararray,State:chararray,Country:chararray);

-- Group data using the country column

GroupByCountry = GROUP salesTable BY Country;

-- Generate result format

CountByCountry = FOREACH GroupByCountry GENERATE CONCAT((chararray)$0, CONCAT(':', (chararray)COUNT($1)));

-- Save result in HDFS folder

STORE CountByCountry INTO 'salesOutput' USING PigStorage('\t');

**Pig Operators**

There are several Apache Pig operators to perform several types of operations.

The types of operators are:

* Loading and Sorting
* Filtering
* Grouping and Joining
* Sorting
* Combining and Splitting
* Diagnostic Operators

Solution :

=========

|  |
| --- |
| -- Load the CSV file |
|  |

|  |
| --- |
| salesTable = LOAD 'hdfs:///user/root/deepa/sales.csv' USING PigStorage(',') AS (Product:chararray,Price:chararray,Payment\_Type:chararray,Name:chararray,City:chararray,State:chararray,Country:chararray); |
|  |

|  |
| --- |
| -- Group data using the country column |
|  |

|  |
| --- |
| GroupByCountry = GROUP salesTable BY Country; |
|  |

|  |
| --- |
| -- Generate result format |
|  |

|  |
| --- |
| CountByCountry = FOREACH GroupByCountry GENERATE CONCAT((chararray)$0, CONCAT(':', (chararray)COUNT($1))); |
|  |

|  |
| --- |
| -- Save result in HDFS folder |
|  |

STORE CountByCountry INTO 'salesOutput' USING PigStorage('\t');