

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | |
|  | | | Parallel And Distributive Computing | | | | |  | | |
|  | |  | | | | | | |  | |
|  | | | | |  |  | | | | |
|  | | | | | Project Report |  | | | | |
|  | | | | | 05 December 2022—PDC—Dr. Nausheen Shoaib |  | | | | |
|  | | | |  | | |  | | | |
|  |  | | | | | | | | |  |

**PROJECT REPORT**

**Group Members:**

1. Abdul Ahad Shaikh (20K-0319)
2. Ali Jodat (20K-0155)
3. Basil Ali Khan (20K-0477)

**Problem Statement:**

Calculating the Number of Products and Sales by Country using Java Program that utilizes Map-Reduce Framework

**Task and Data:**

The task is specifically focused on finding number of products and sum of sales per country given the input file data set SalesJan2009.csv.



**Data Set Description:**

Column 1: Transaction on date

Column 2: Product

Column 3: Price

Column 4: Payment Type

Column 5: Name

Column 6: City

Column 7: State

Column 8: Country

Column 9: Account Created

Column 10: Last Login

Column 11: Latitude

Column 12: Longitude

**Platform, Framework and Tools Description:**

Hadoop

Hadoop is a distributed computing Framework developed and maintained by The Apache Software Foundation written in Java. Hadoop consists of HDFS and Map-Reduce and is generally deployed in a group of machines called cluster. Initially, GFS and Map-Reduce were built to empower Google Search. HDFS stands for Hadoop Distributed File System and is used to store data across multiple disks. Map-Reduce is a way to parallelize Data processing tasks.

Map-Reduce Algorithm

Map-Reduce Algorithm consists of Map() procedure that performs filtering and sorting of input data and Reduce() performs summary\aggregate function per (key, value) pair.

Java

Java is a popular programming language, created in 1995. It is owned by Oracle. It is a simple programming language. Java makes writing, compiling, and debugging programming easy. It helps to create reusable code and modular programs. Java is a class-based, object-oriented programming language and is designed to have as few implementation dependencies as possible. A general-purpose programming language made for developers to write once run anywhere that is compiled Java code can run on all platforms that support Java. Java applications are compiled to byte code that can run on any Java Virtual Machine. The syntax of Java is similar to C/C++.

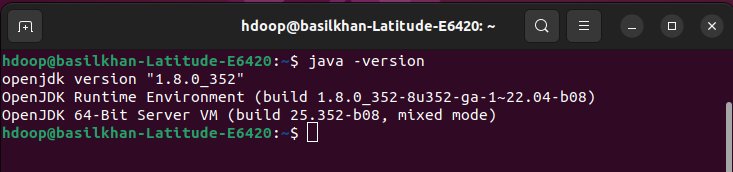
**Methodology:**

The main idea of this problem's solution is to use the same Key for every row with the same country name. In addition, the value used at each mapper will be the price (sales) of that row, which corresponds to the Key country.

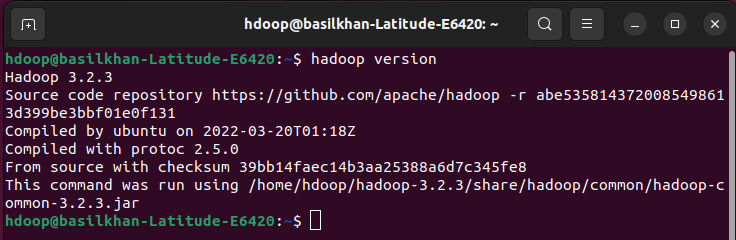
In addition, since for this task we would like to output multiple values for each key, the code utilizes a custom made class that implements the Writeable Interface. A custom Hadoop writable data type which needs to be used as value field in Map-Reduce programs must implement Writable interface.

**Setup and Implementation:**

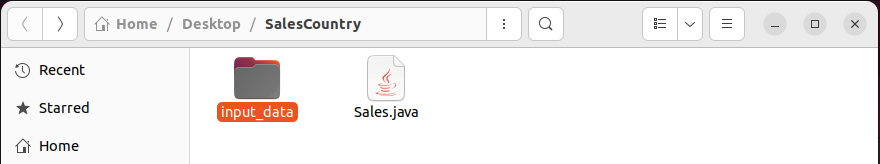
* *Java Version*



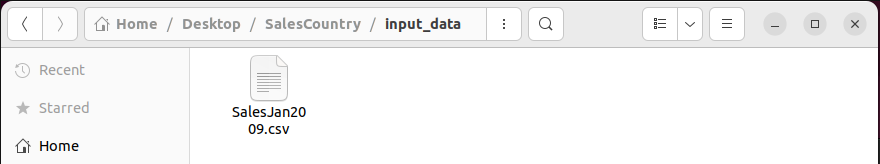
* *Hadoop Version*



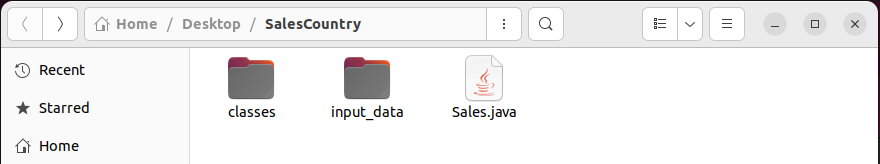
* *Creating new folder for input data*



* *Copying data set .csv file in pervious created folder*



* *Creating folder to hold Java class files*



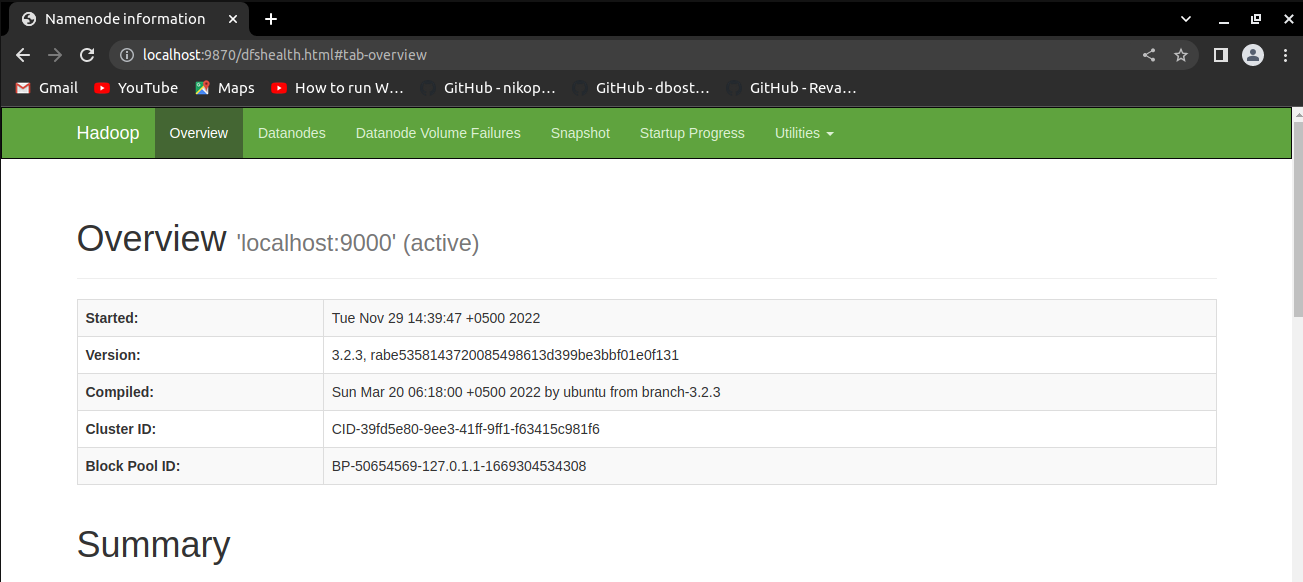
* *Starting start-dfs.sh and start-yarn.sh*

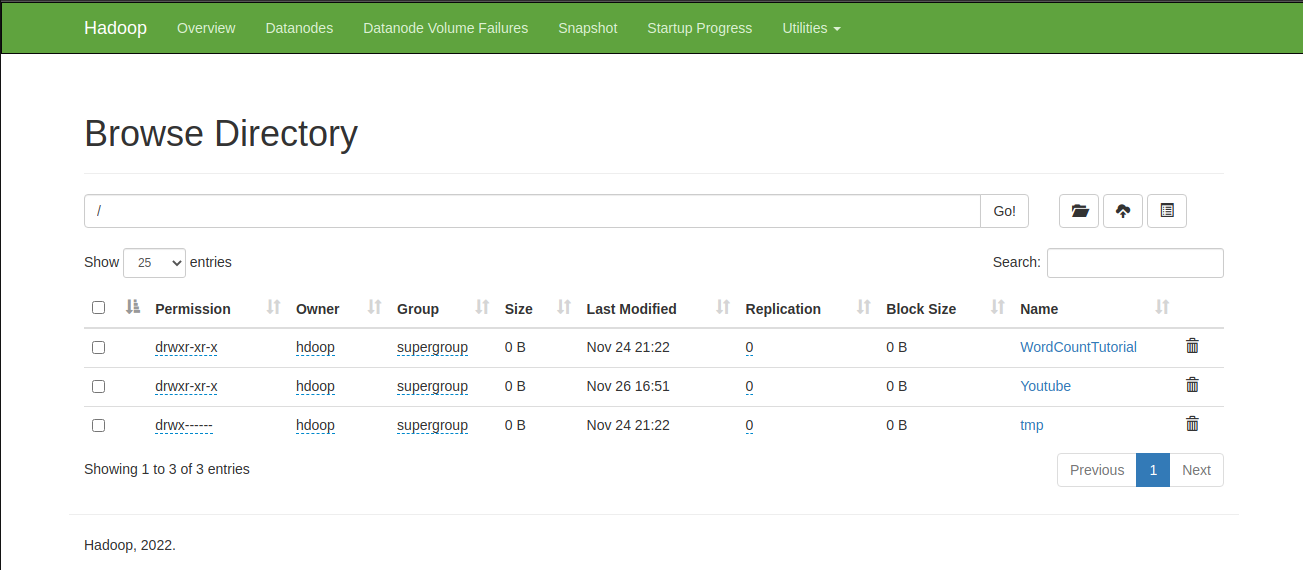
*start-dfs.sh - Starts the Hadoop DFS daemons, the name node and data node.*

*Start-yarn.sh- Starts the Hadoop YARN daemons, the resource and node managers.*

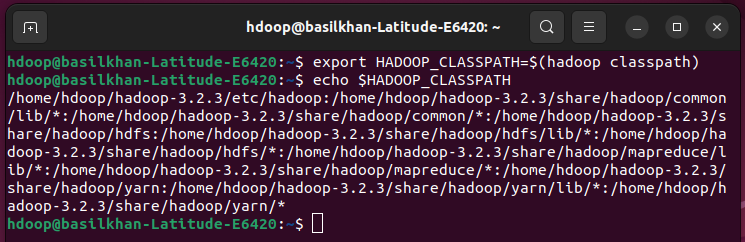


* *For checking services to start we go to: localhost:9870*

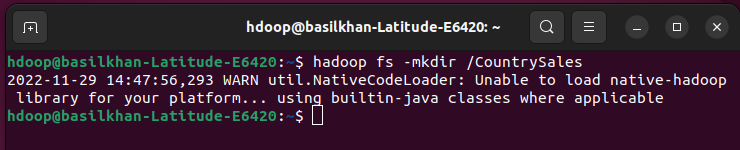


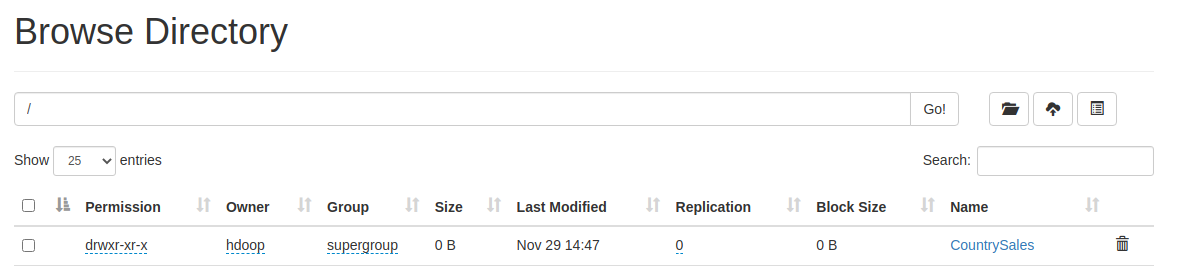


* *Setting HADOOP\_CLASSPATH environment variable and making it sure that it is set correctly*

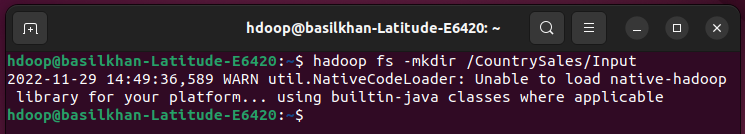


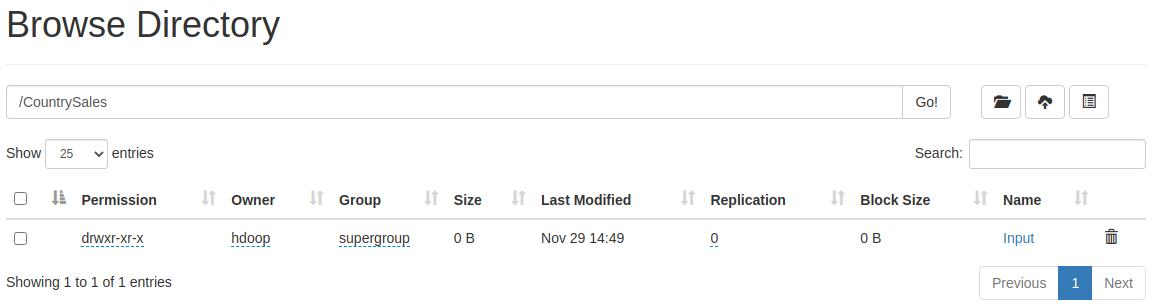
* *Creating a directory on Hadoop file system*



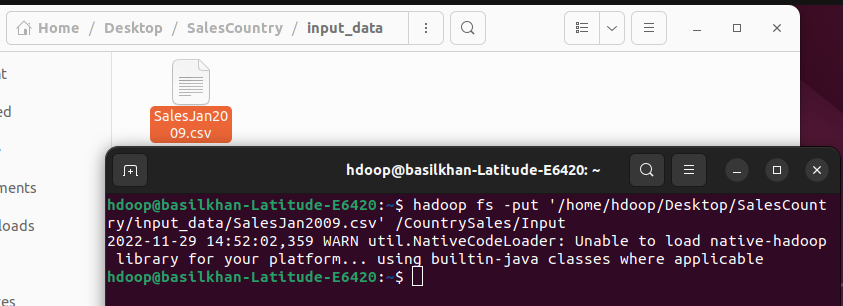


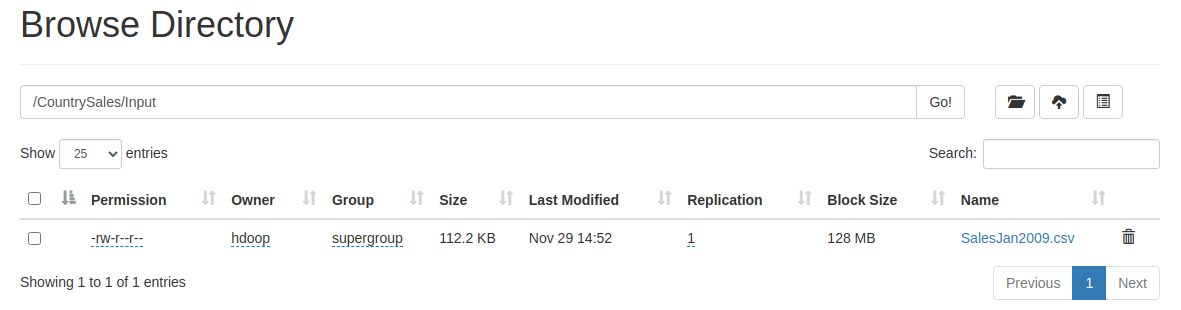
* *Creating a directory inside it for input data set*



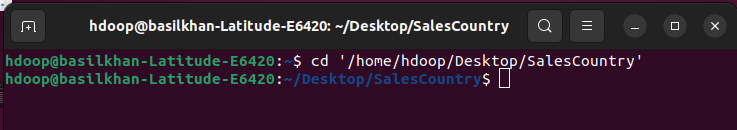


* *Uploading the input file to that HDFS directory*

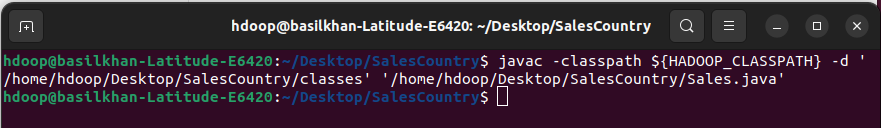


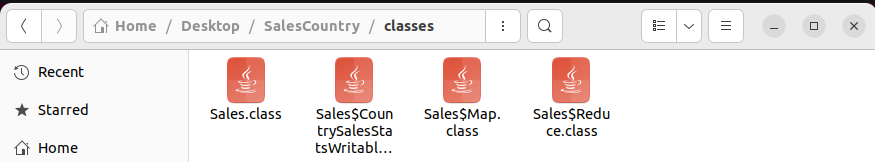


* *Changing current directory to local machine directory*

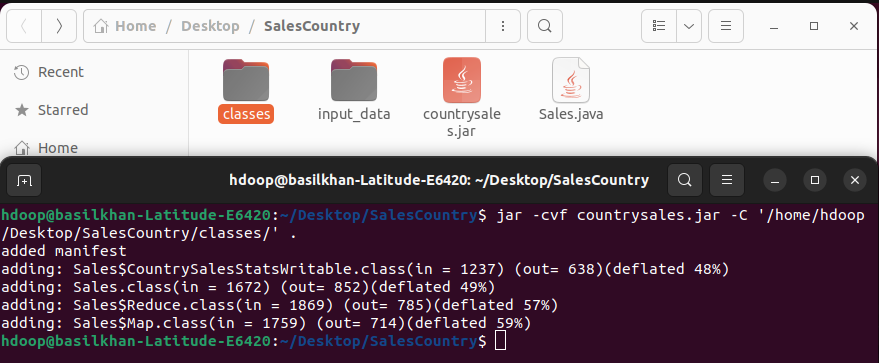


* *Compiling the Java code and checking files*

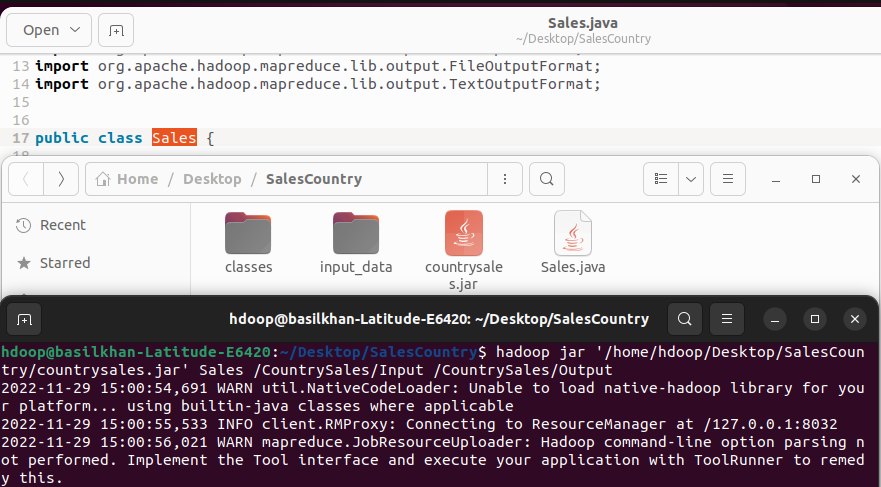


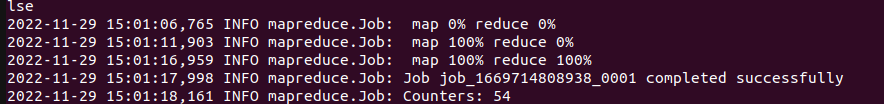


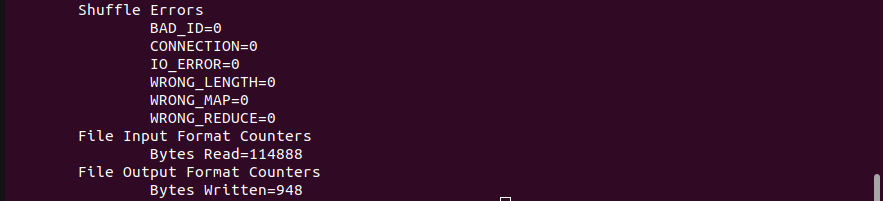
* *Putting output files in one JAR file*

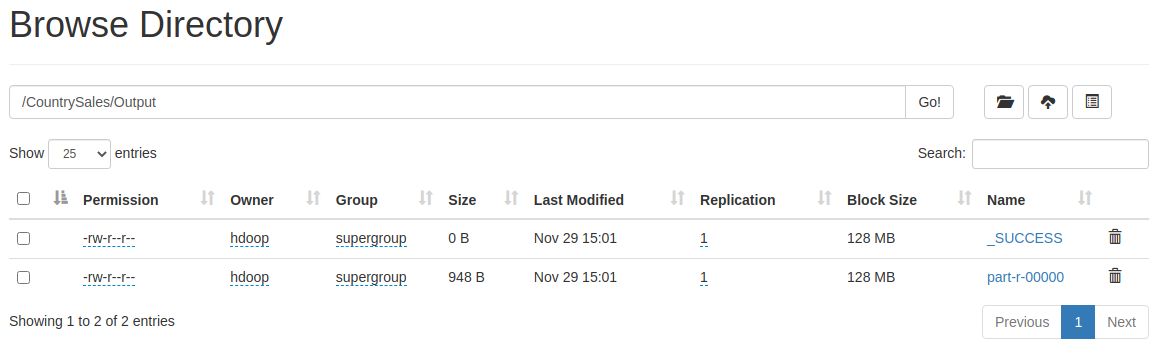


* *Running JAR file on Hadoop*

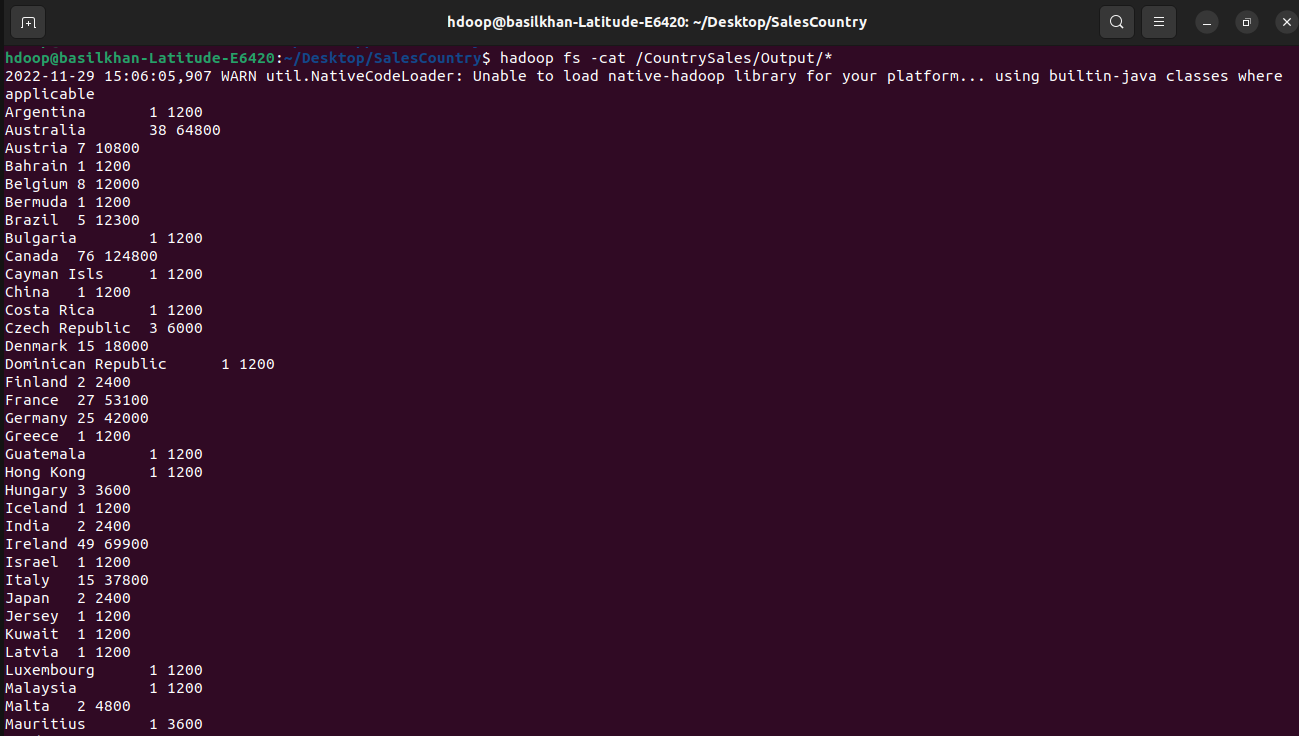




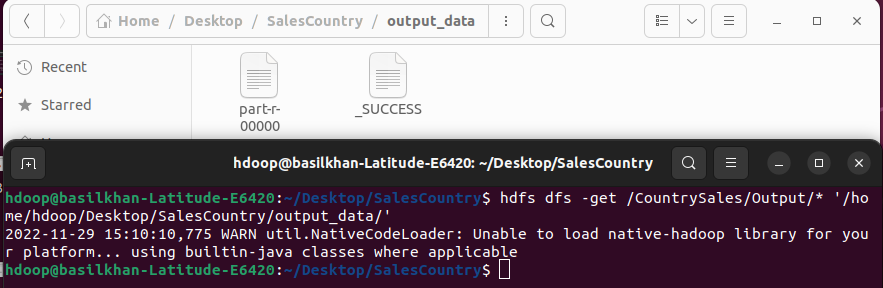




* *Output*

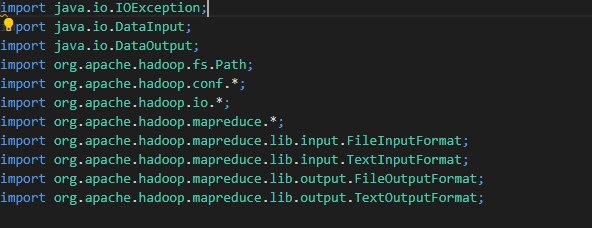


* *Copying output files from HDFS environment to local machine directory*



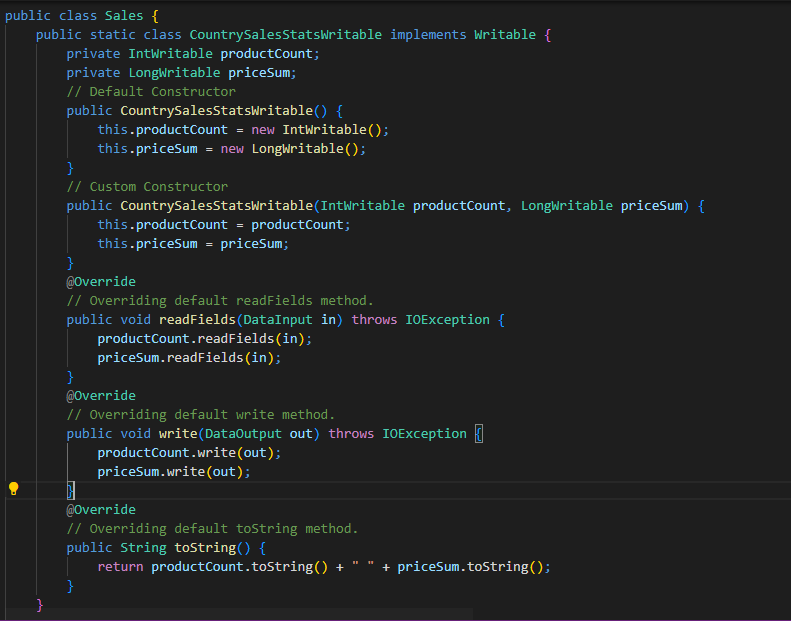


**Libraries and Packages:**



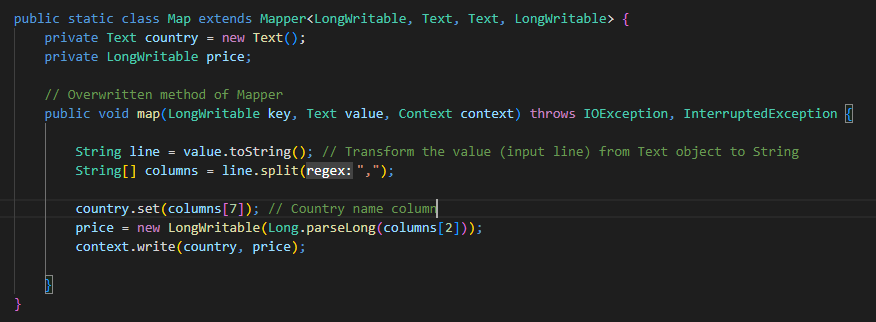
**Code Snippets and Explanation:**

* *The objects of this class will be utilized, since for this task we would like to output multiple values for each key. A custom Hadoop writable data type which needs to be used as value field in Map-Reduce programs must implement Writable interface*



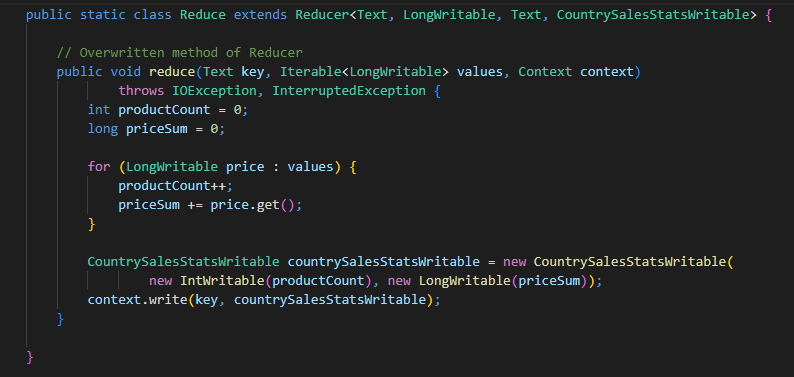
* *Input types to Mapper are: <LongWritable, V:Text>*

*Output types from Mapper are: <Text, V:LongWritable>*

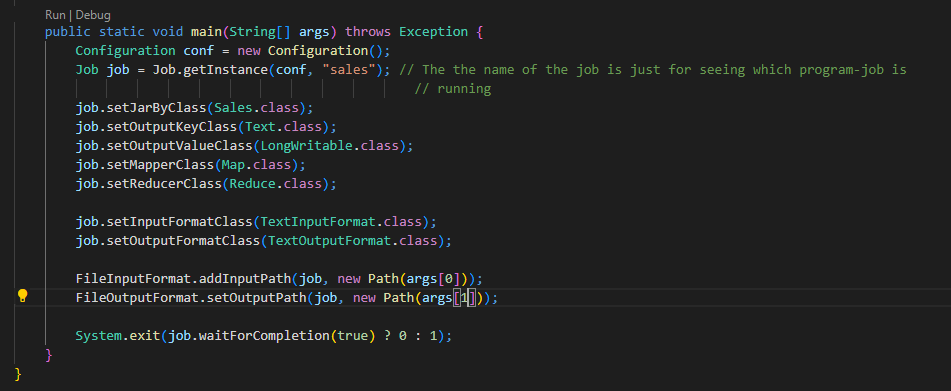


* *Input types to Reducer are: <Text, V:list(LongWritable)>*

*Output types from Reducer are: <Text, V:CountrySalesStatsWritable>*



* *Main function*



**Conclusion:**

This project implements Hadoop Map-Reduce algorithm on the Country Sales data and display the results in output file in Hadoop file system.

**Data Set and Reference Links:**

* <https://www.cs.ucy.ac.cy/courses/DSC511/data/SalesJan2009.csv>
* <https://medium.com/edureka/mapreduce-tutorial-3d9535ddbe7c>
* <http://hadooptutorial.info/creating-custom-hadoop-writable-data-type/>