**Birla Institute of Technology & Science, Pilani**

**Work Integrated Learning Programmes Division**

**Third Semester 2020-2021**

**Assignment 1**

Course No. : S2-20\_DSECLZG522

Course Title : Big Data Systems

Using Hadoop for Historical Sales Data Analysis

# Contribution & Team

**Group 079**

|  |  |  |
| --- | --- | --- |
| S.No | Name | BITSIDNO |
| 1 | Yashodeep Vaidya | 2019HC04042 |
| 2 | Sharad K Laad | 2019HC04034 |
| 3 | Vijay Jaiswal | 2019HC04049 |

# Table of Contents

[Contribution & Team 0](#_Toc80990018)

[Table of Contents 1](#_Toc80990019)

[Purpose of study – Assignment Details 2](#_Toc80990020)

[Solution: 3](#_Toc80990021)

[Question 1 & 3 3](#_Toc80990022)

[Java Class Structure 3](#_Toc80990023)

[Implementation Code 3](#_Toc80990024)

[Command to Run the Driver 3](#_Toc80990025)

[Taking User Input: Item Type 4](#_Toc80990026)

[Running Logs: 4](#_Toc80990027)

[Summary of MapRaduce Job: 4](#_Toc80990028)

[Question 2 6](#_Toc80990029)

[Java Class Structure 6](#_Toc80990030)

[Implementation Code 6](#_Toc80990031)

[Command to Run the Driver 6](#_Toc80990032)

[Taking User Input: Item Type 7](#_Toc80990033)

[Running Logs: 7](#_Toc80990034)

[Summary of MapRaduce Job: 7](#_Toc80990035)

[Question 4 9](#_Toc80990036)

[Java Class Structure 9](#_Toc80990037)

[Implementation Code 9](#_Toc80990038)

[Command to Run the Driver 9](#_Toc80990039)

[Taking User Input: Year 10](#_Toc80990040)

[Running Logs: 10](#_Toc80990041)

[Summary of MapRaduce Job: 10](#_Toc80990042)

# Purpose of study – Assignment Details

Purpose of this report and study is to answer following questions

**Business context:**

A large multi-national retail chain has sales orders data across regions and different sales channels for a large variety of item types. The business team wants to use this data to analyze various aspects of sales - e.g. top selling items in a region, regions with maximum profit in a certain item type, if there is a significant difference in revenue in two item types across regions etc.

**Problem statement:**

As the data analytics team, use the sales transaction data set with about 100K records to answer these questions below —

1. Average unit\_price by country for a given item type in a certain year
2. Total units\_sold by year for a given country and a given item type
3. Find the max and min units\_sold in any order for each year by country for a given item type. Use a custom partitioner class instead of default hash based.
4. What are the top 10 order id for a given year by the total\_profit

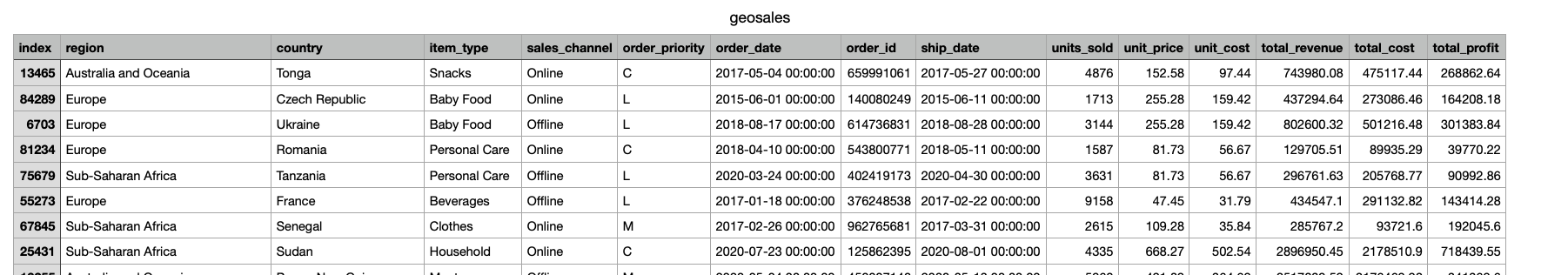
You have to show the above analysis working on a Hadoop system using map reduce code, preferably in Java or Python. You can do data preparation steps as required before running a MapReduce job to answer these questions above.

**Submission materials:**

* Working documented code

Data set:

[https://drive.google.com/drive/folders/1fJyk6GE13xw0P-OKWwW8fxa-ny1G0TPp?usp=sharing](https://drive.google.com/drive/folders/1fjyk6ge13xw0p-okwww8fxa-ny1g0tpp?usp=sharing)



# Solution:

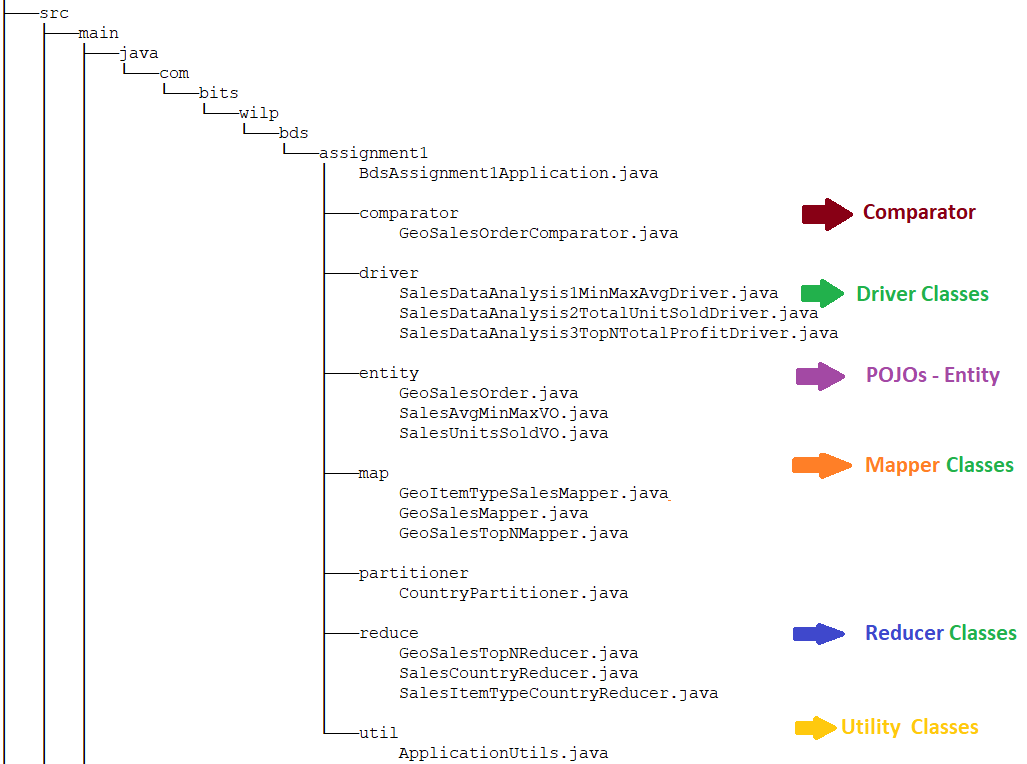
## Question 1 & 3

1. Average unit\_price by country for a given item type in a certain year
2. Find the max and min units\_sold in any order for each year by country for a given item type. Use a custom partitioner class instead of default hash based.

**Answer 1&3:**

Since the input and parameter were same for Q1 and Q3 we have used same implementation for both.

### Java Class Structure



### Implementation Code

Attached with this document.

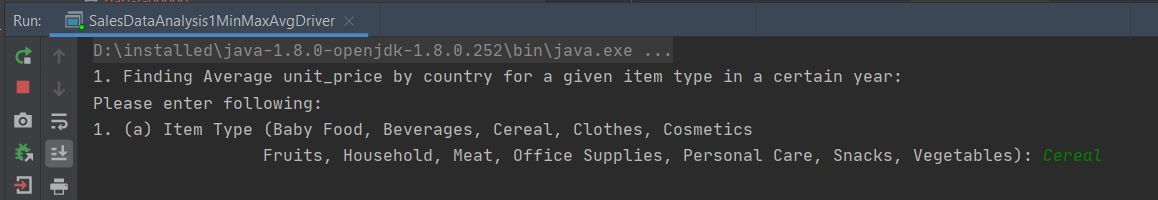
### Command to Run the Driver

***java com.bits.wilp.bds.assignment1.driver.SalesDataAnalysis1MinMaxAvgDriver geosales.csv geosales\_output\_1***

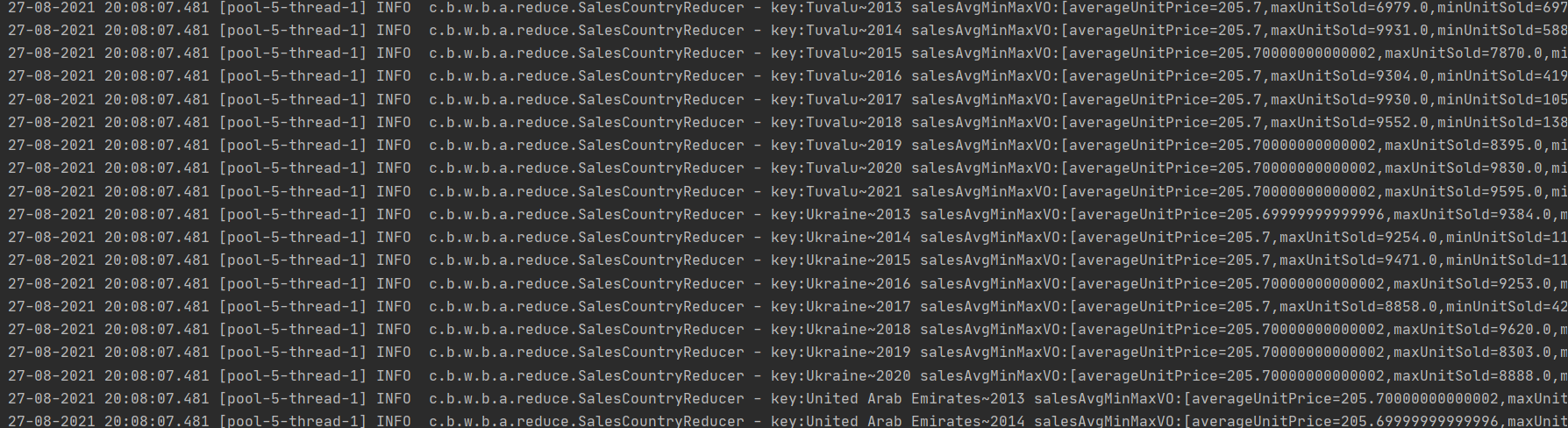
Here :

1. **geosales.csv:** input file having historical sales data
2. **geosales\_output\_1:** Output folder where the analysis results will be saved

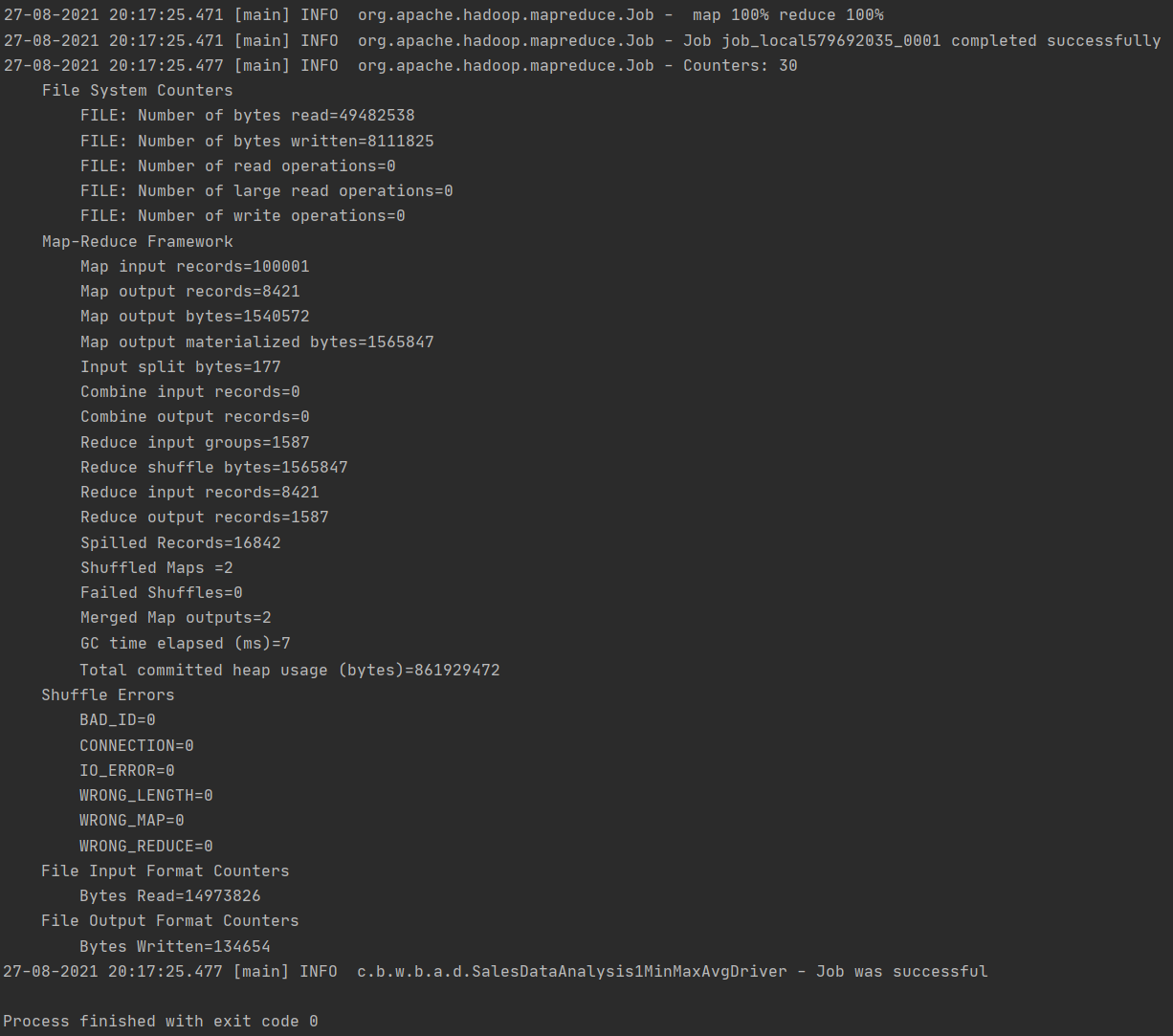
### Taking User Input: Item Type

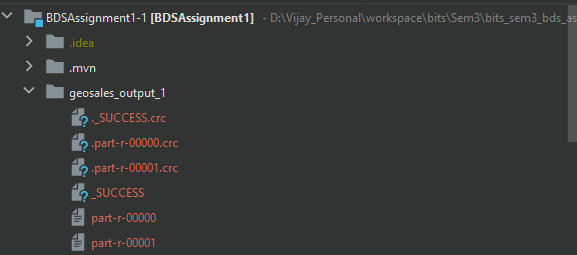


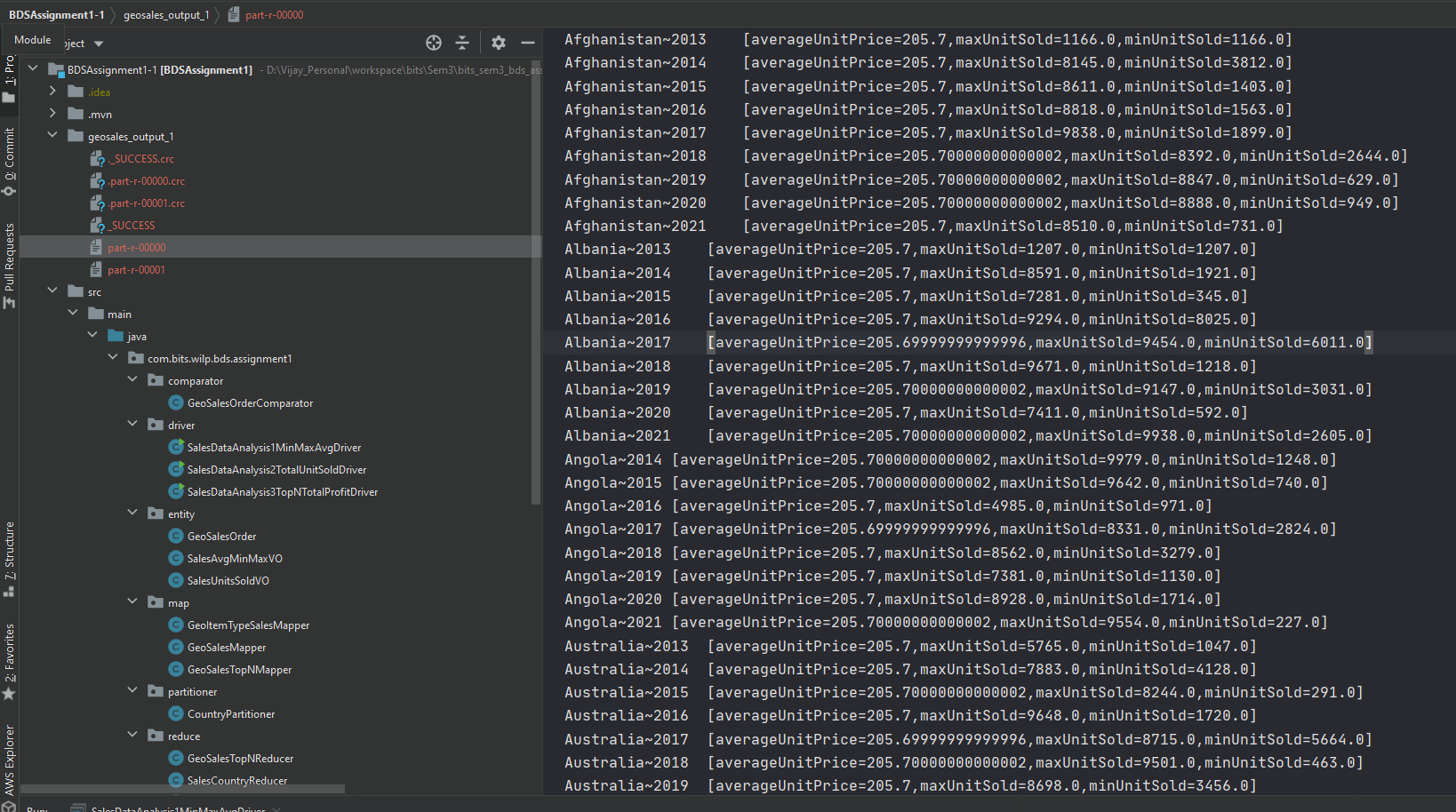
### Running Logs:



### Summary of MapRaduce Job:



Output Folder Created with Analysis: 

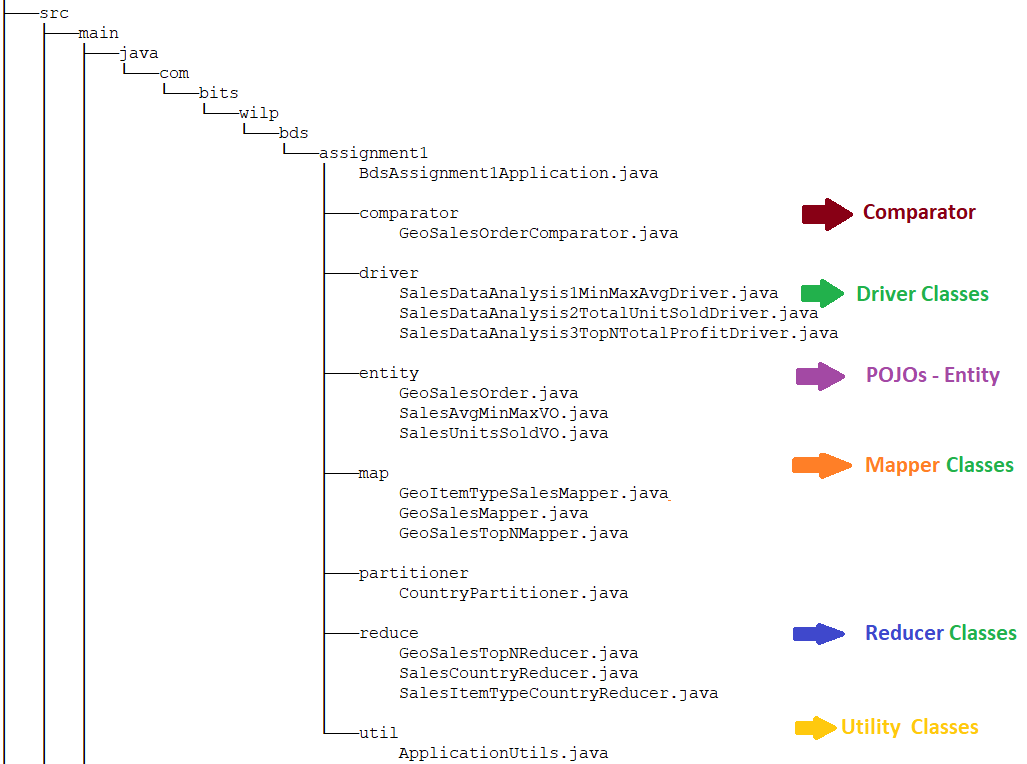


## Question 2

1. Total units\_sold by year for a given country and a given item type

**Answer 2:**

### Java Class Structure



### Implementation Code

Attached with this document.

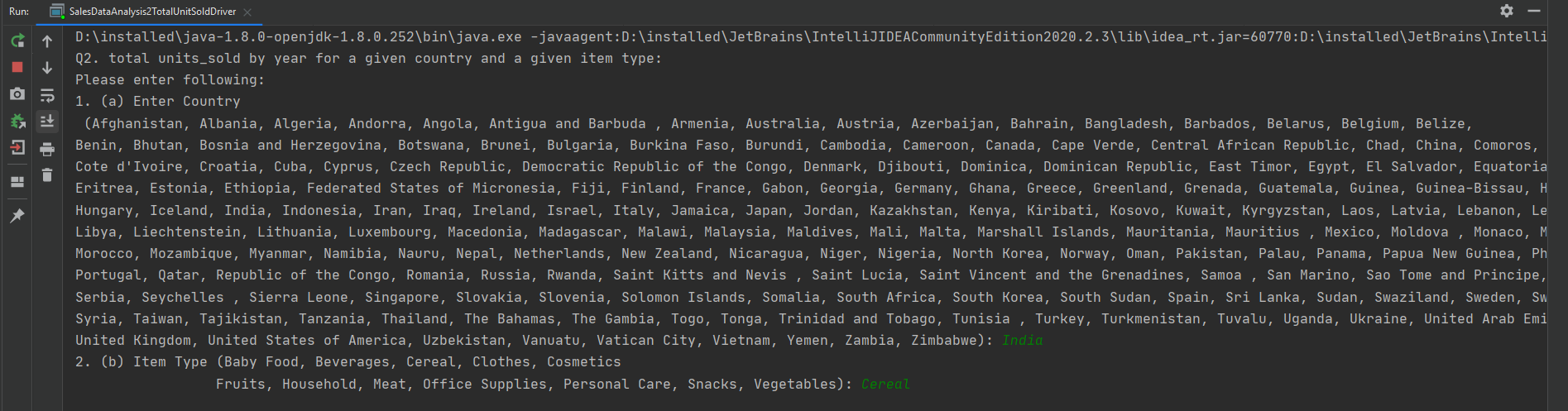
### Command to Run the Driver

***java com.bits.wilp.bds.assignment1.driver.SalesDataAnalysis2TotalUnitSoldDriver geosales.csv geosales\_output\_2***

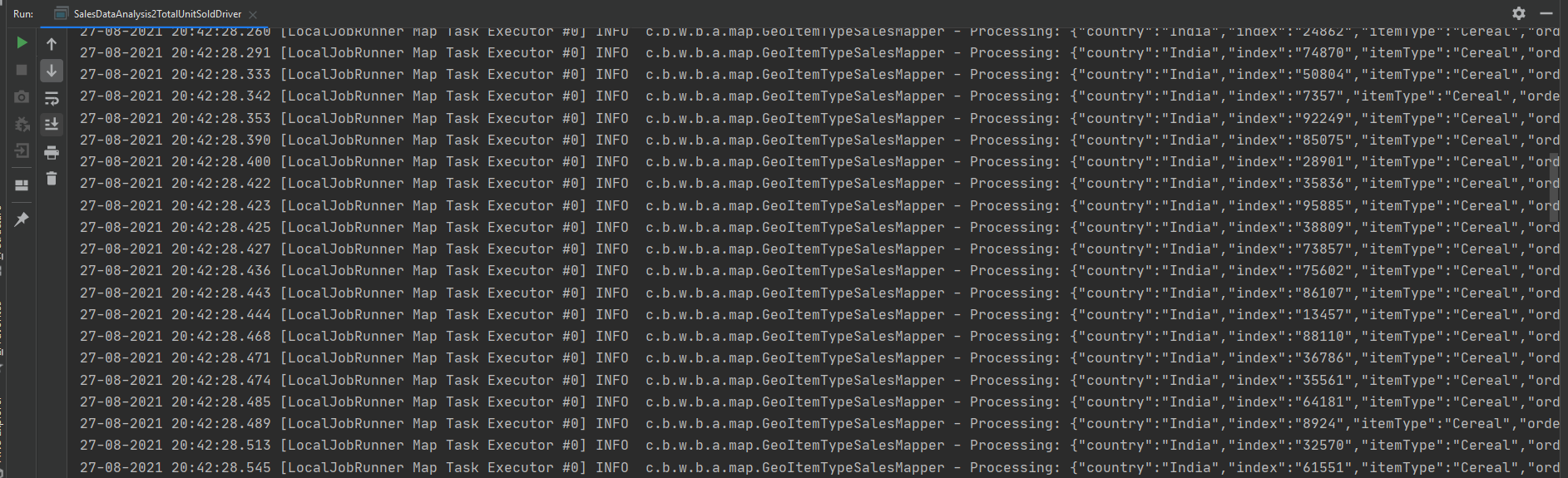
Here :

1. **geosales.csv:** input file having historical sales data
2. **geosales\_output\_2:** Output folder where the analysis results will be saved

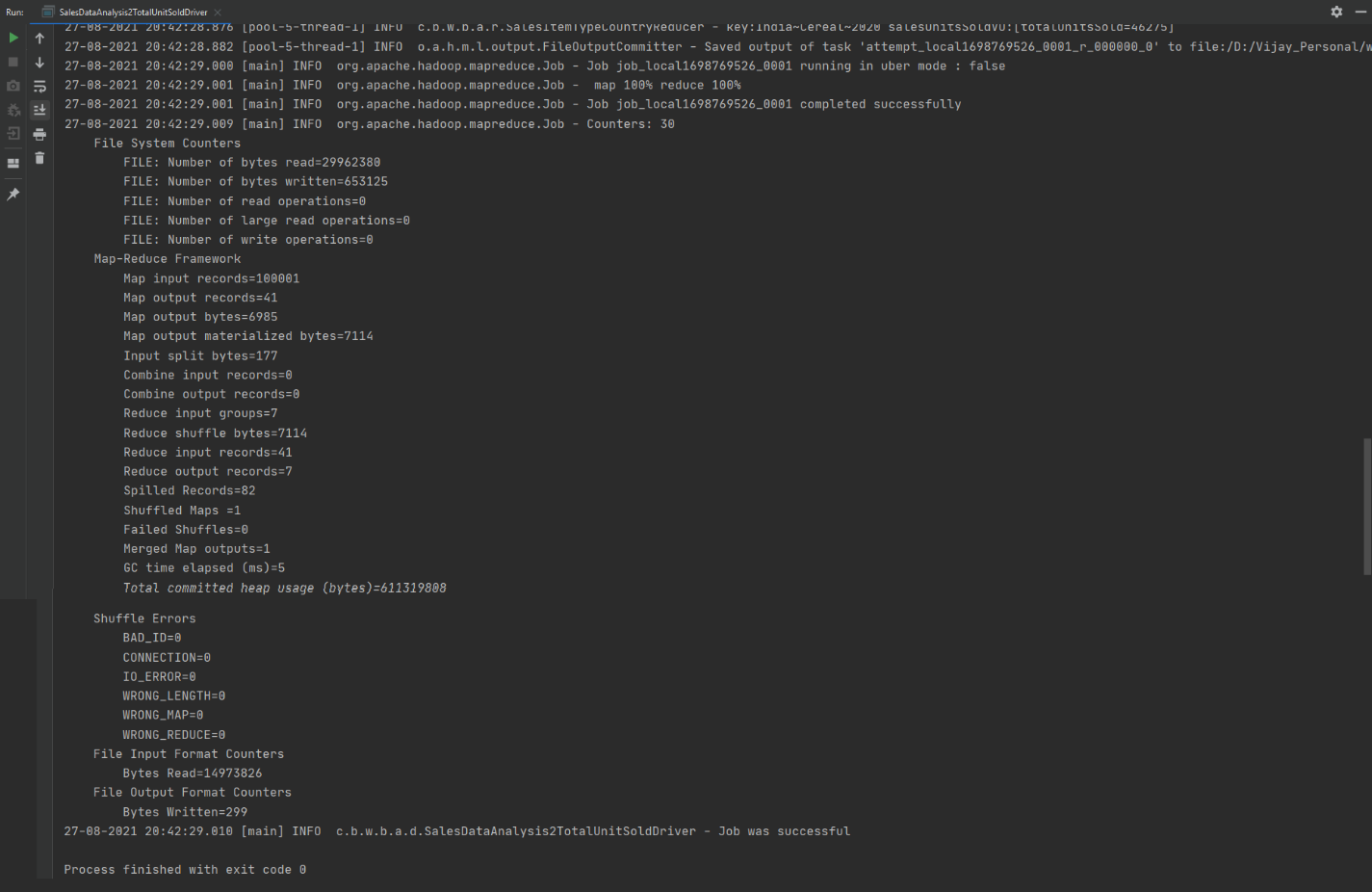
### Taking User Input: Item Type

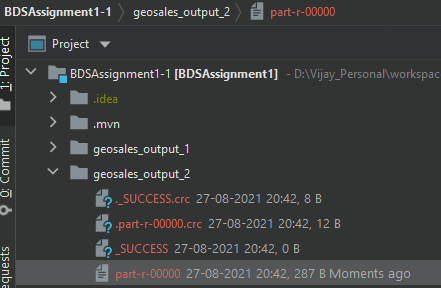


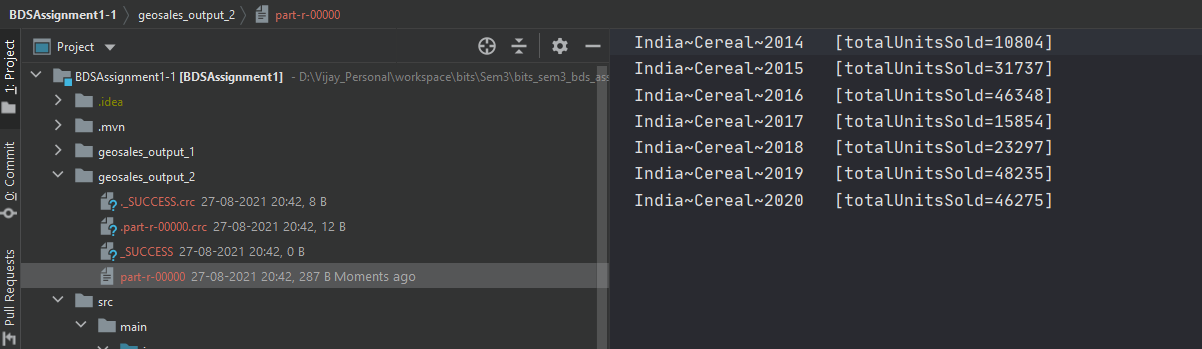
### Running Logs:



### Summary of MapRaduce Job:



Output Folder Created with Analysis: 

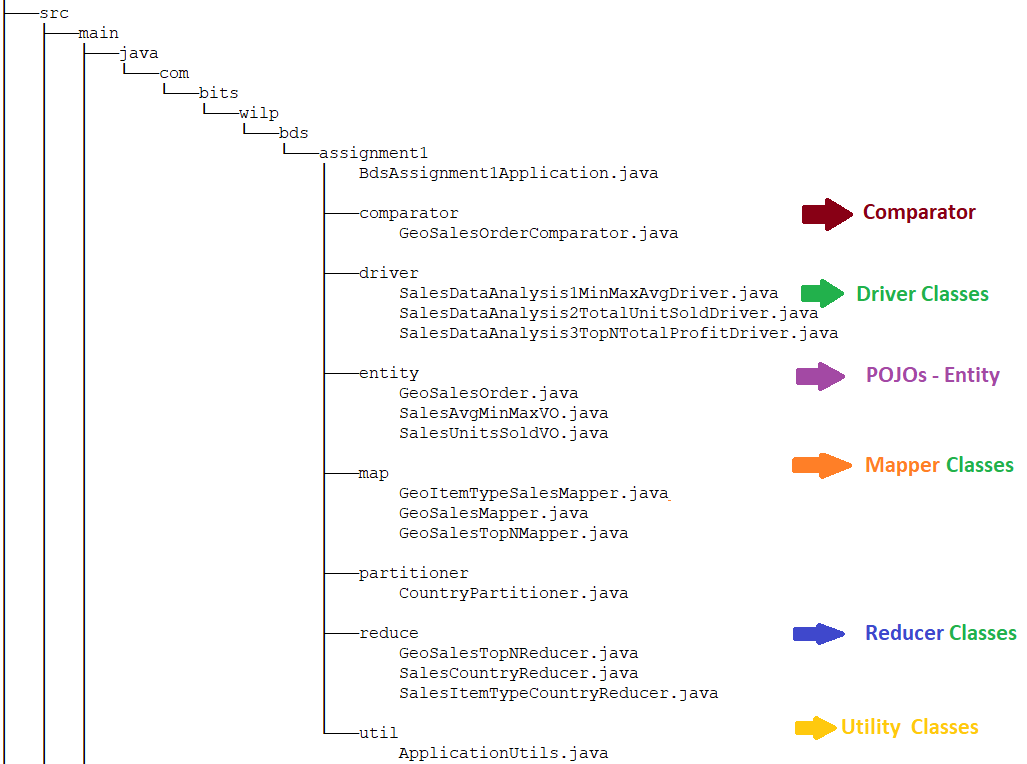


## Question 4

1. What are the top 10 order id for a given year by the total\_profit

**Answer 4:**

### Java Class Structure



### Implementation Code

Attached with this document.

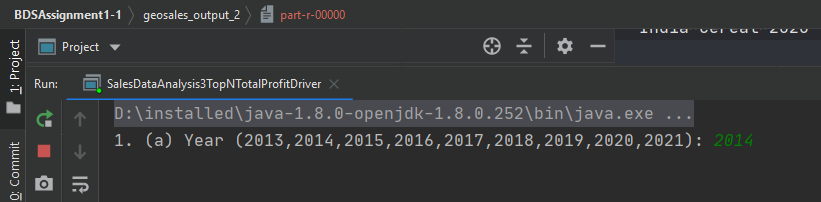
### Command to Run the Driver

***java com.bits.wilp.bds.assignment1.driver.SalesDataAnalysis3TopNTotalProfitDriver geosales.csv geosales\_output\_3***

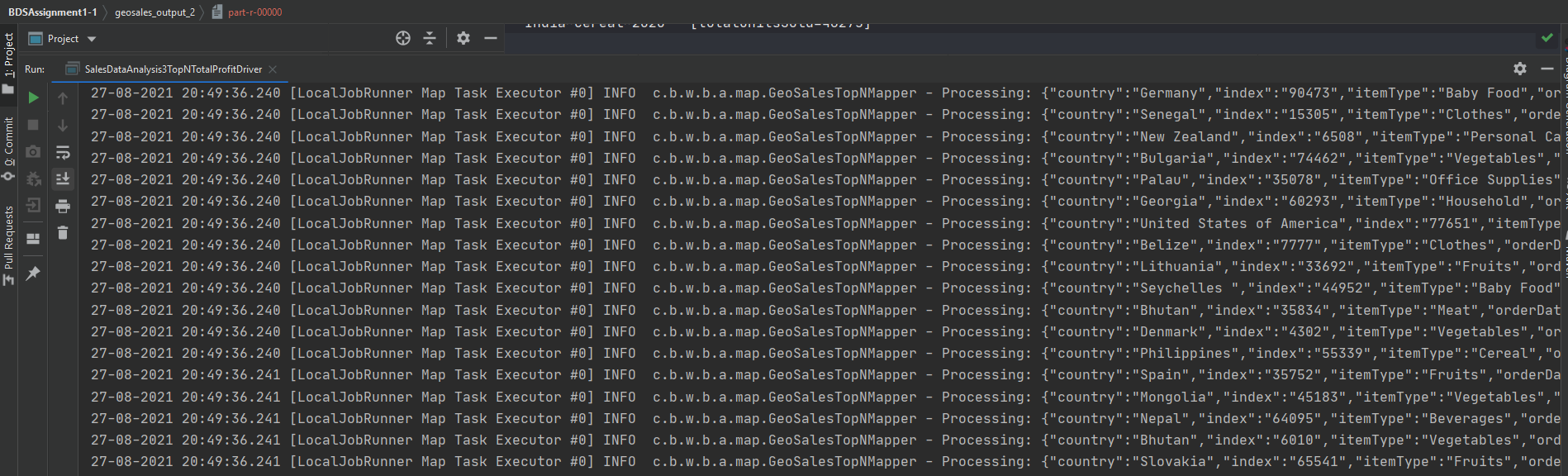
Here :

1. **geosales.csv:** input file having historical sales data
2. **geosales\_output\_3:** Output folder where the analysis results will be saved

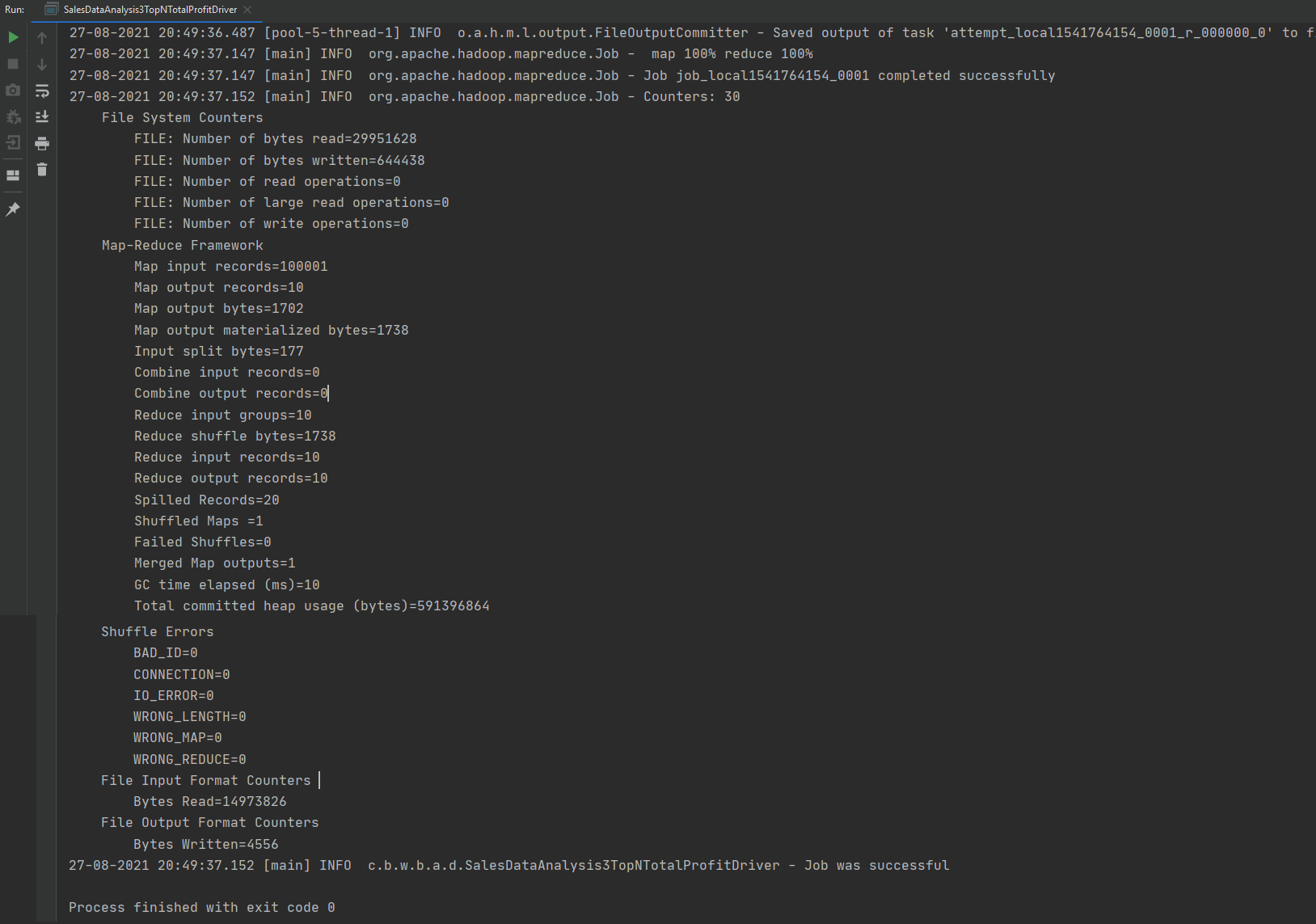
### Taking User Input: Year



### Running Logs:



### Summary of MapRaduce Job:



Output Folder Created with Analysis: 