Name : Deore Ketan Balkrushna

Roll No. : BE-A-31

Subject : Data Mining and Warehousing

# Experiment No. : 1

## Title :

For an organization of your choice, choose a set of business processes. Design star / snow flake schemas for analysing these processes. Create a fact constellation schema by combining them. Extract data from different data sources, apply suitable transformations and load into destination tables using an ETL tool. For Example: Business Origination: Sales, Order, Marketing process.

## Objectives :

Understands the basis of Star/Snowflake/fact constellation schema and learn the Rapid Miner tool for performing various operation on built-in or external datasets.

## Hardware Requirement :

Pentium or higher processor, 2GB RAM and 500 GB HDD.

## Software Requirement :

Rapid Miner

## Theory :

### What does ETL mean?

ETL stands for Extract, Transform and Load. An ETL tool extracts the data from different RDBMS source systems, transforms the data like applying calculations, concatenate, etc. and then load the data to Data Warehouse system. The data is loaded in the DW system in the form of dimension and fact tables.

## Extraction

- A staging area is required during ETL load. There are various reasons why staging area is required.
- The source systems are only available for specific period of time to extract data. This period of time is less than the total data-load time. Therefore, staging area allows you to extract the data from the source system and keeps it in the staging area before the time slot ends.
- Staging area is required when you want to get the data from multiple data sources together or if you want to join two or more systems together. For example, you will not be able to perform a SQL query joining two tables from two physically different databases.
- Data extractions' time slot for different systems vary as per the time zone and operational hours.
- Data extracted from source systems can be used in multiple data warehouse system, Operation Data stores, etc.
- ETL allows you to perform complex transformations and requires extra area to store the data.
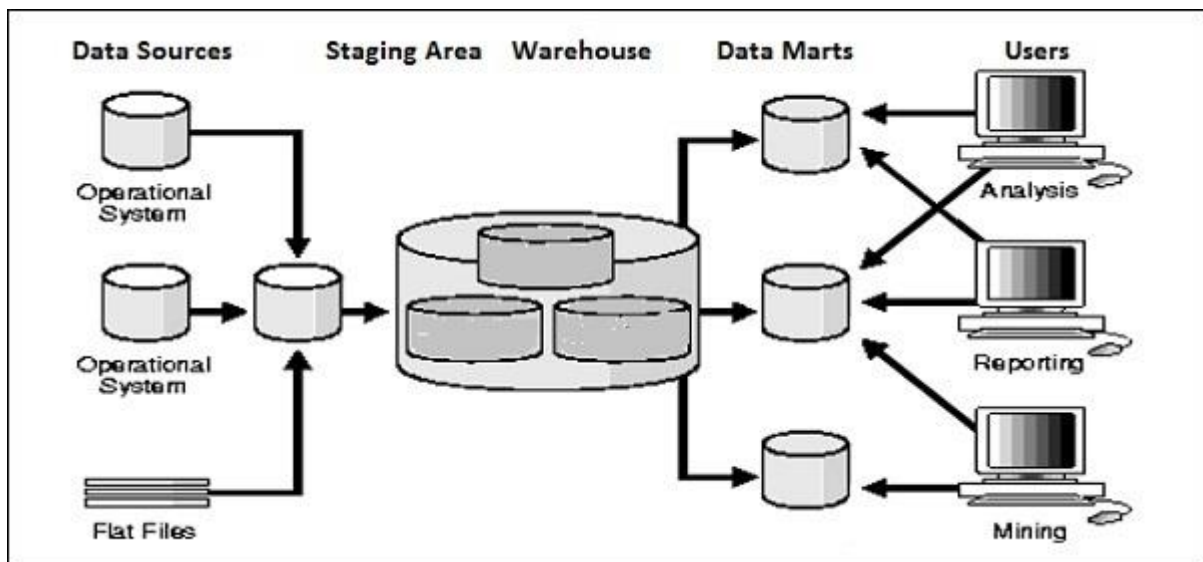


## Transform

In data transformation, you apply a set of functions on extracted data to load it into the target system. Data, which does not require any transformation is known as direct move or pass through data.

You can apply different transformations on extracted data from the source system. For example, you can perform customized calculations. If you want sum-of-sales revenue and this is not in database, you can apply the SUM formula during transformation and load the data.

For example, if you have the first name and the last name in a table in different columns, you can use concatenate before loading.

**Load**

    During Load phase, data is loaded into the end-target system and it can be a flat file or a Data Warehouse system.
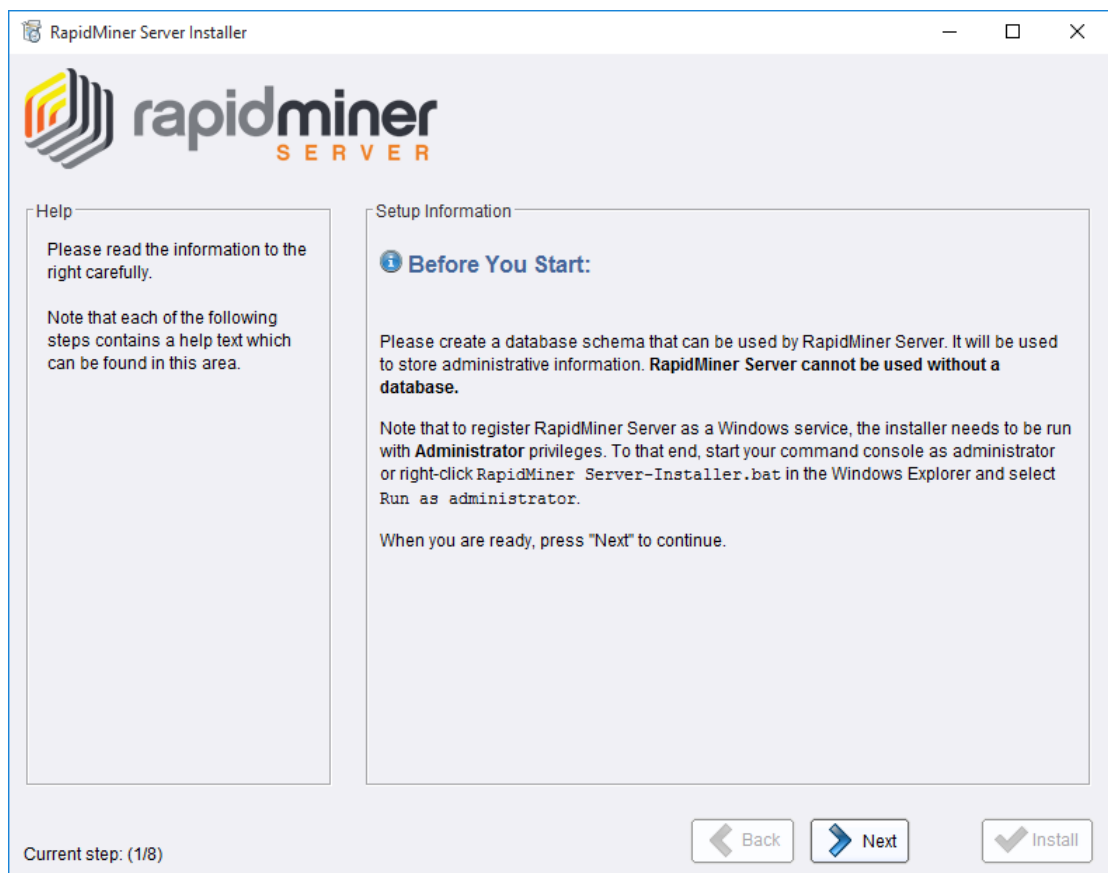
## Rapid Miner :

    Rapid Miner is a world-leading open-source system for data mining. It is available as a stand-alone application for data analysis and as a data mining engine for the integration into own products. Rapid Miner is now Rapid Miner Studio and Rapid Analytics is now called Rapid Miner Server.
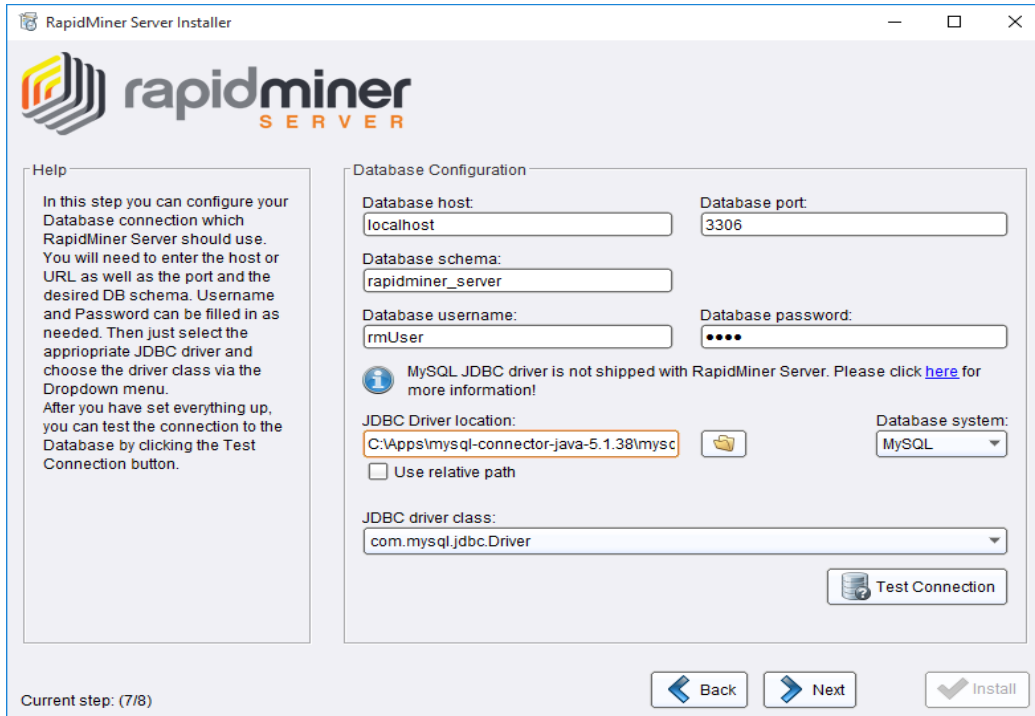
## Steps for Installation :

1. Download Rapid Miner Server
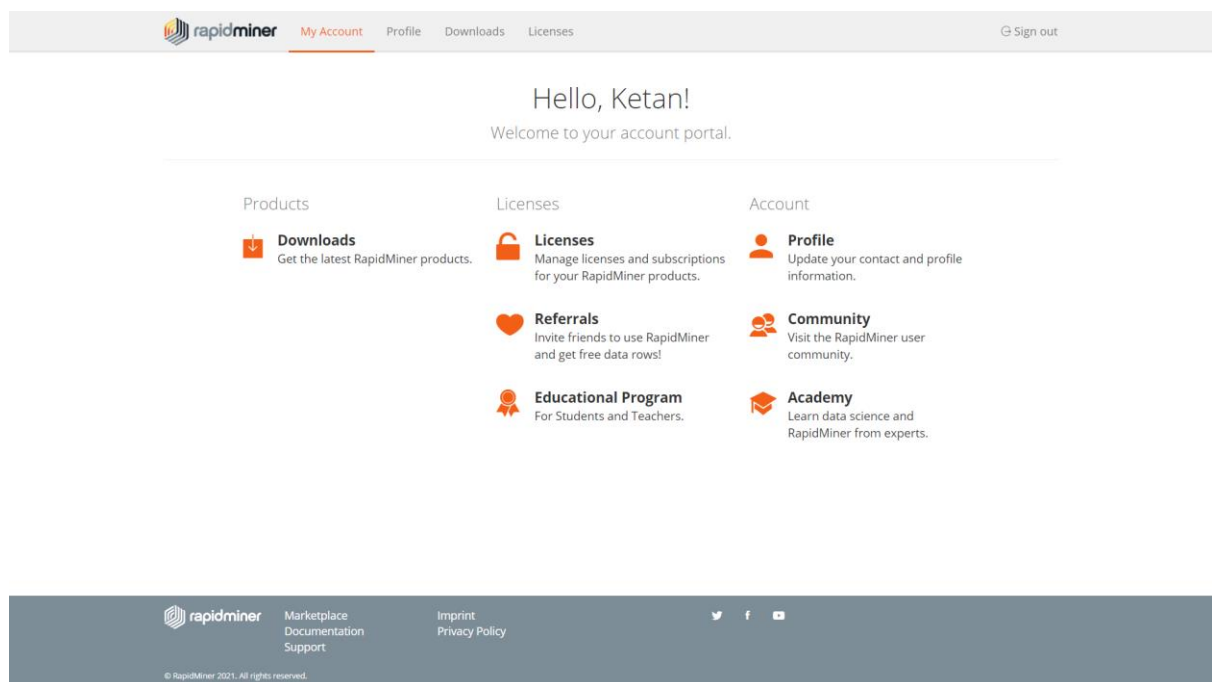
2. Installing Rapid Miner Server



3. Configure Rapid Miner Server Settings

4. Configuring Rapid Miner server's database connection



5. Installing Radoop Proxy



6. Completing the Installation of Rapid Miner Server

7. Installation of Rapid Miner Studio
And choose Installation location
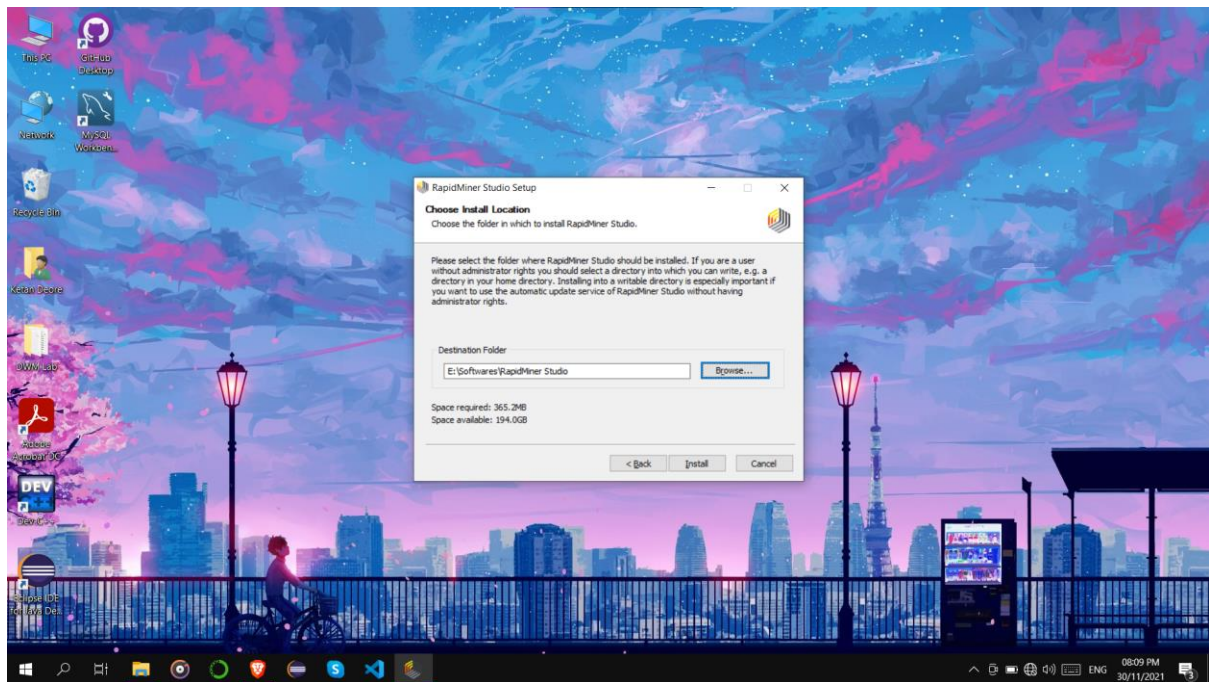
8. Complete Installation and Launch the Studio
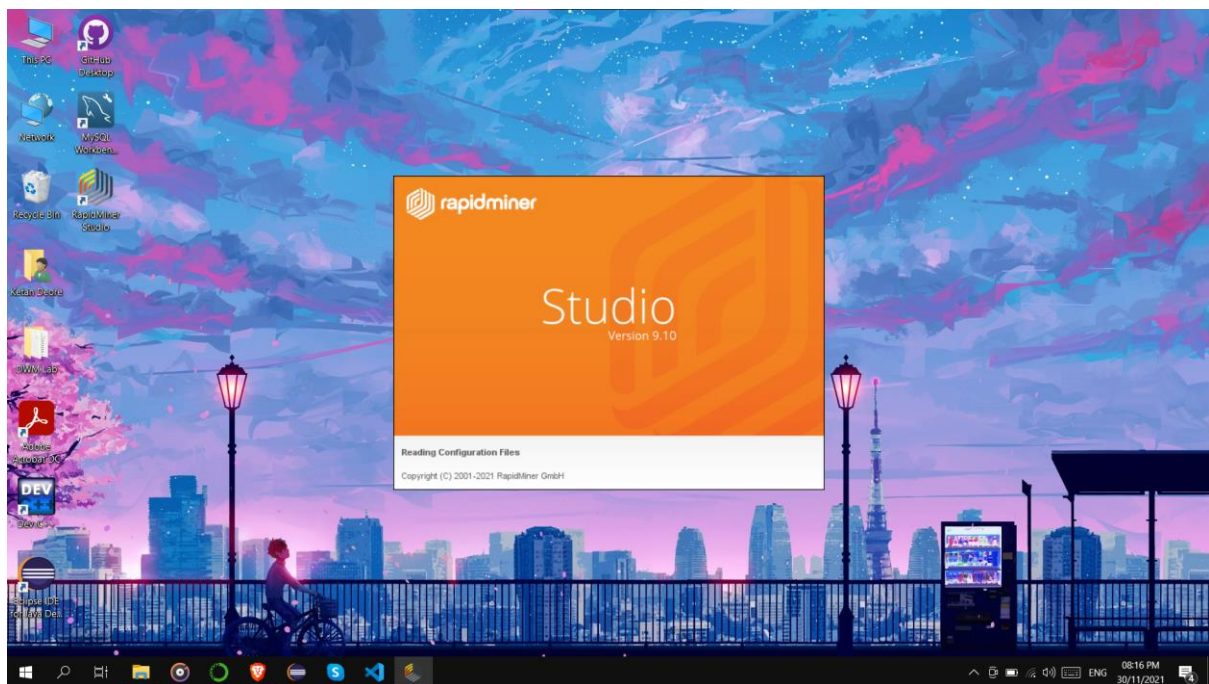


## Data Warehousing Schemas :

1. Star Schema
2. Snowflake Schema
3. Fact Constellation

**Star Schema :**

For example, as you can see in the above-given image that fact table is at the center which contains keys to every dimension table like Deal_ID, Model ID, Date_ID, Product_ID, Branch_ID & other attributes like Units sold and revenue.

**Snowflake Schema :**

A Snowflake Schema is an extension of a Star Schema, and it adds additional dimensions. It is called snowflake because its diagram resembles a Snowflake.

The dimension tables are normalized which splits data into additional tables. In the following example, Country is further normalized into an individual table.

| Star Schema | Snow Flake Schema |
|---|---|
| Hierarchies for the dimensions are stored in the dimensional table. | Hierarchies are divided into separate tables. |
| It contains a fact table surrounded by dimension tables. | One fact table surrounded by dimension table which are in turn surrounded by dimension table |
| In a star schema, only single join creates the relationship between the fact table and any dimension tables. | A snowflake schema requires many joins to fetch the data. |
| Simple DB Design. | Very Complex DB Design. |
| De-normalized Data structure and query also run faster. | Normalized Data Structure. |
| High level of Data redundancy | Very low-level data redundancy |
| Single Dimension table contains aggregated data. | Data Split into different Dimension Tables. |
| Cube processing is faster. | Cube processing might be slow because of the complex join |
| Offers higher performing queries using Star Join Query Optimization. Tables may be connected with multiple dimensions. | The Snow Flake Schema is represented by centralized fact table which unlikely connected with multiple dimensions. |

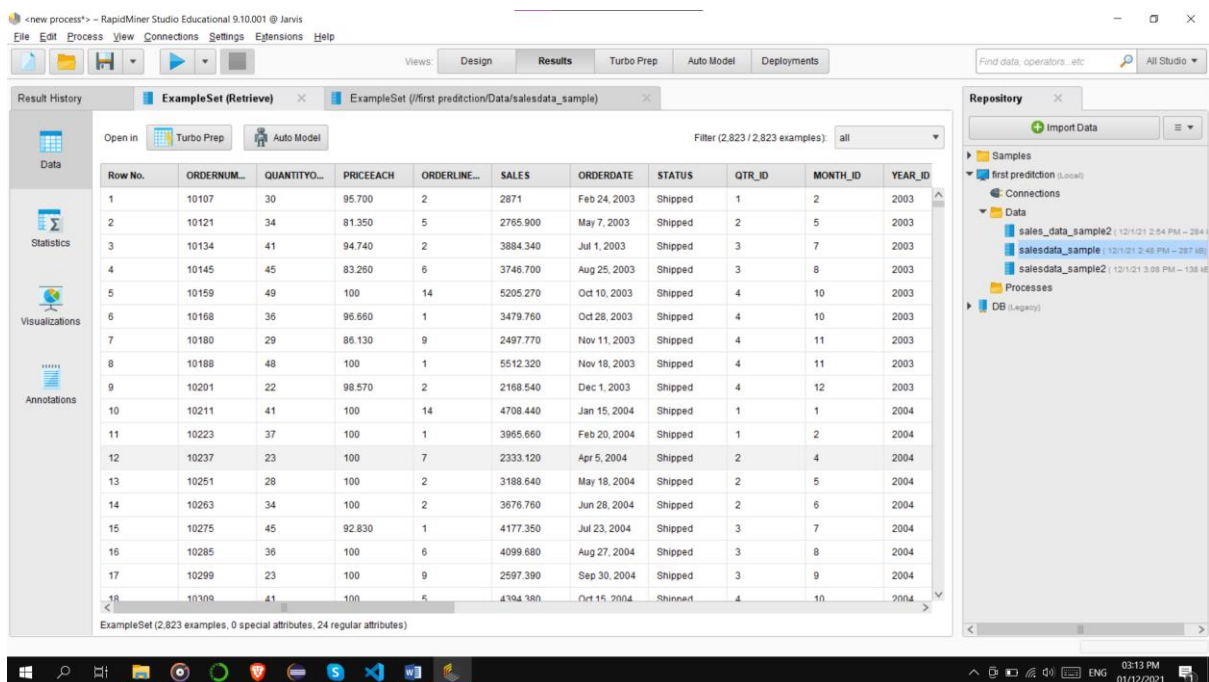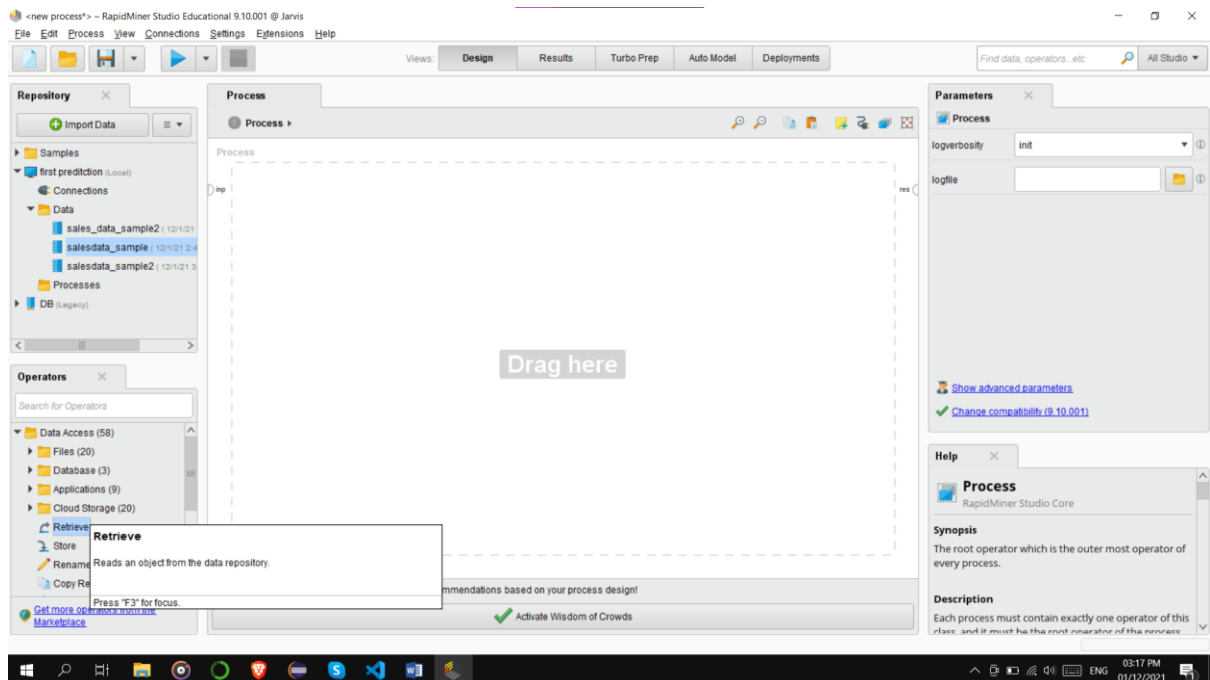## 1. Design Model
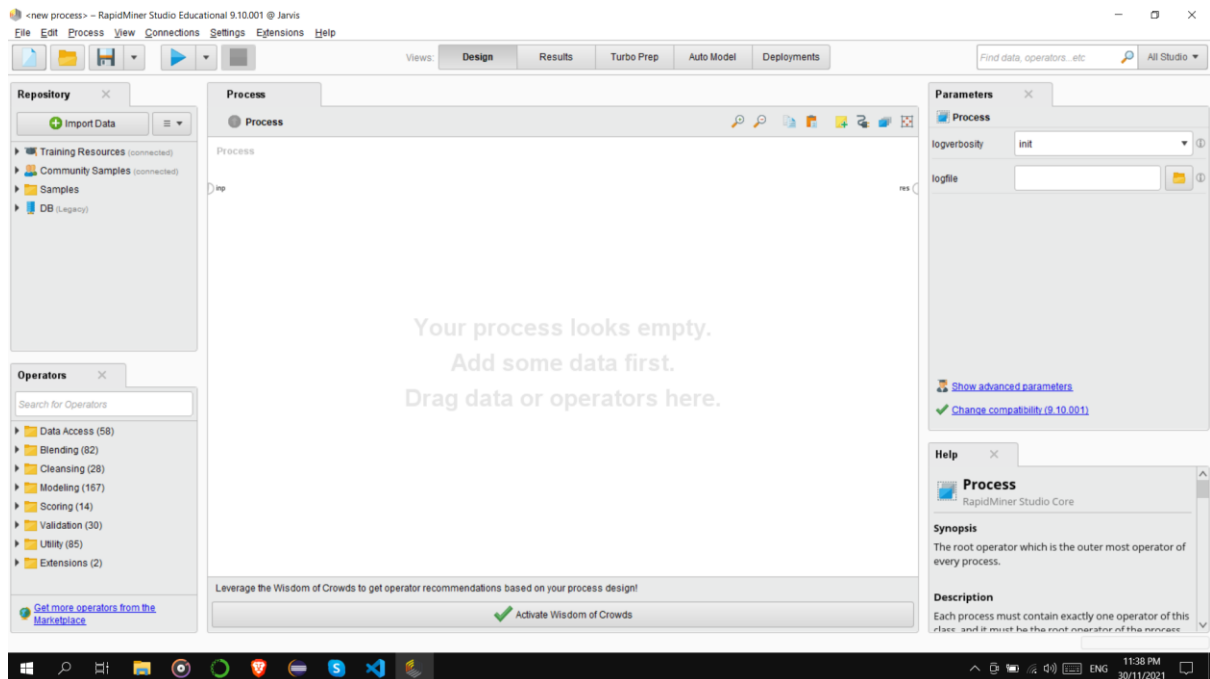


## Step 1 - Import Data from Source
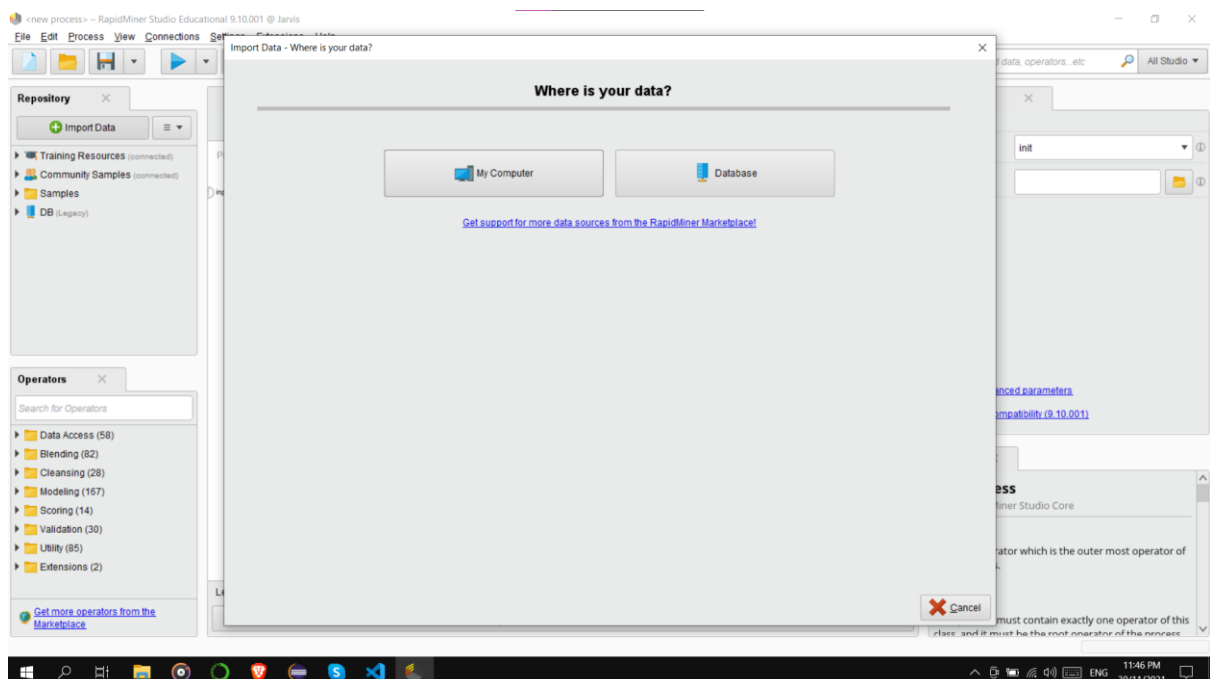
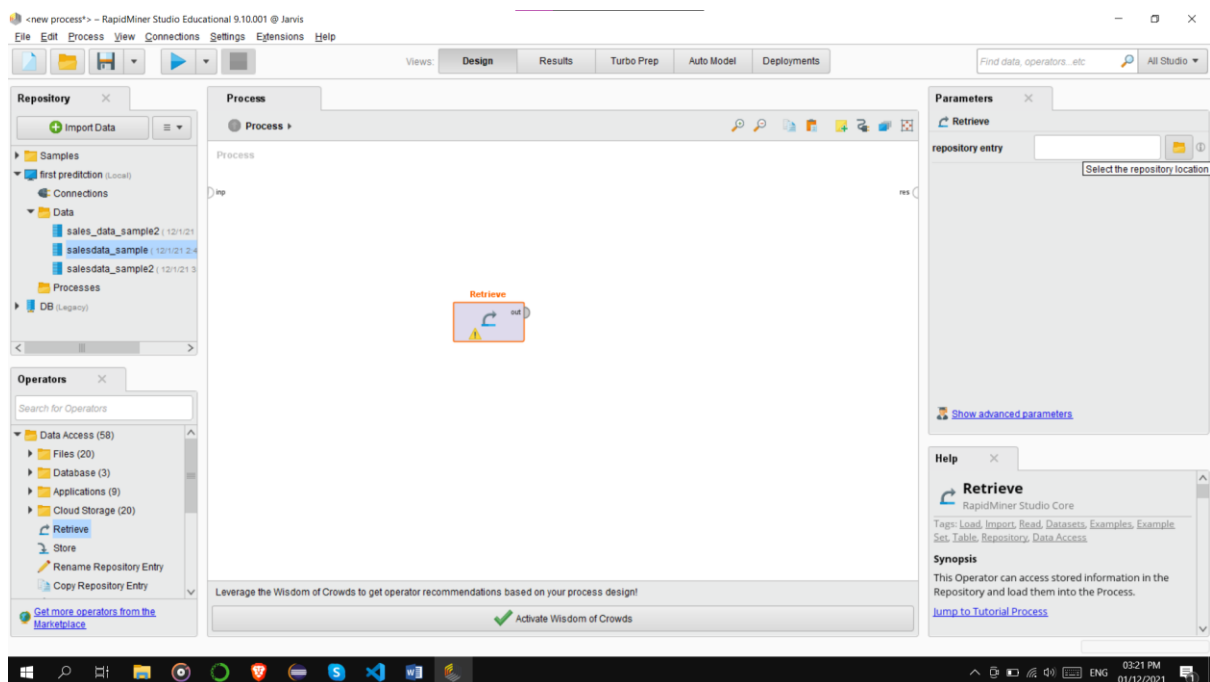## Step 2 - Select data Location
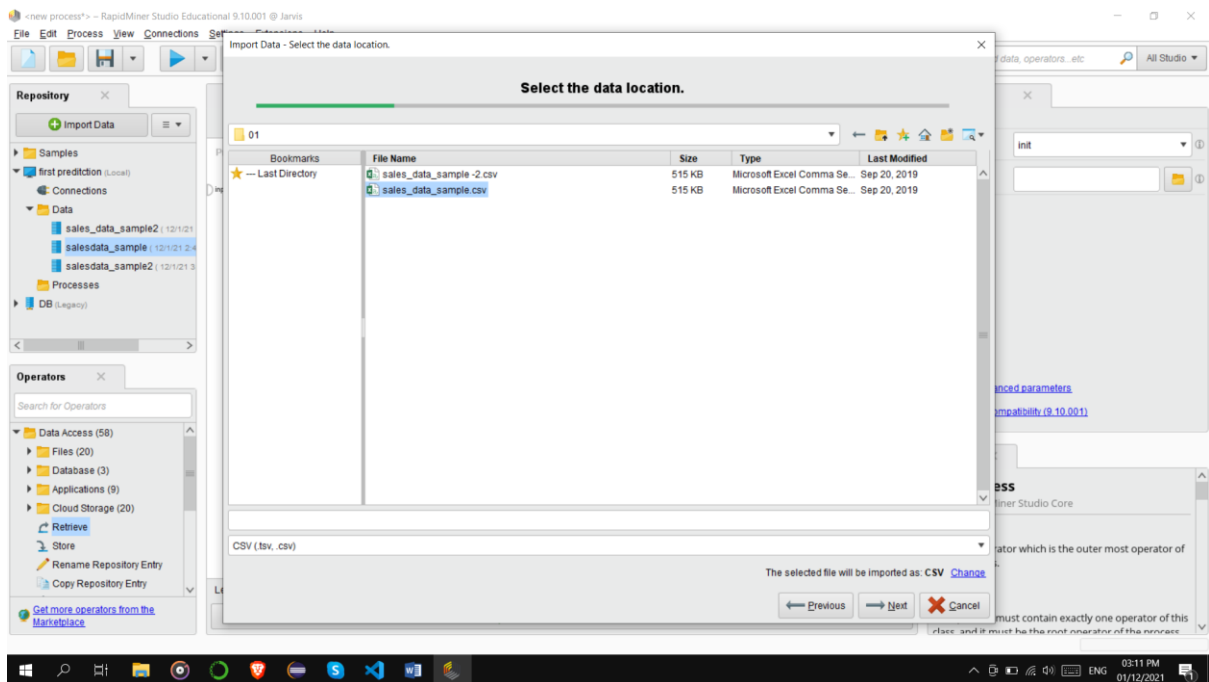


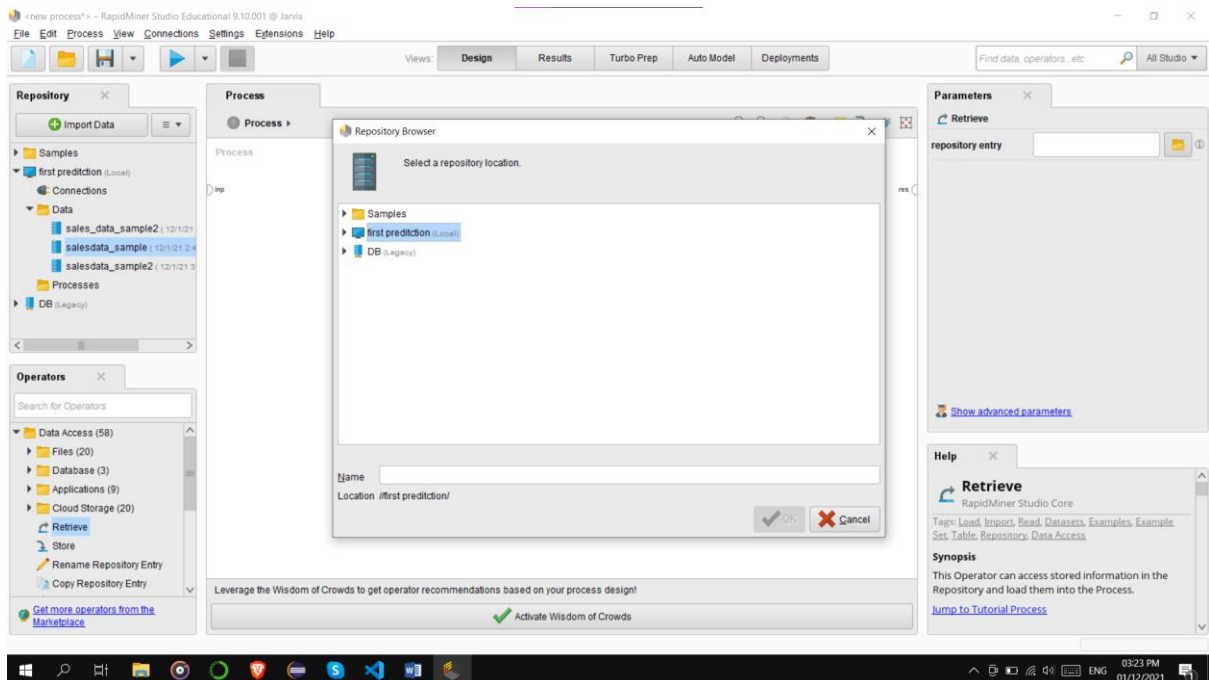## Step 3 - Open Dataset regarding business

Step 4 - Click on retrieve operator drag in process view,
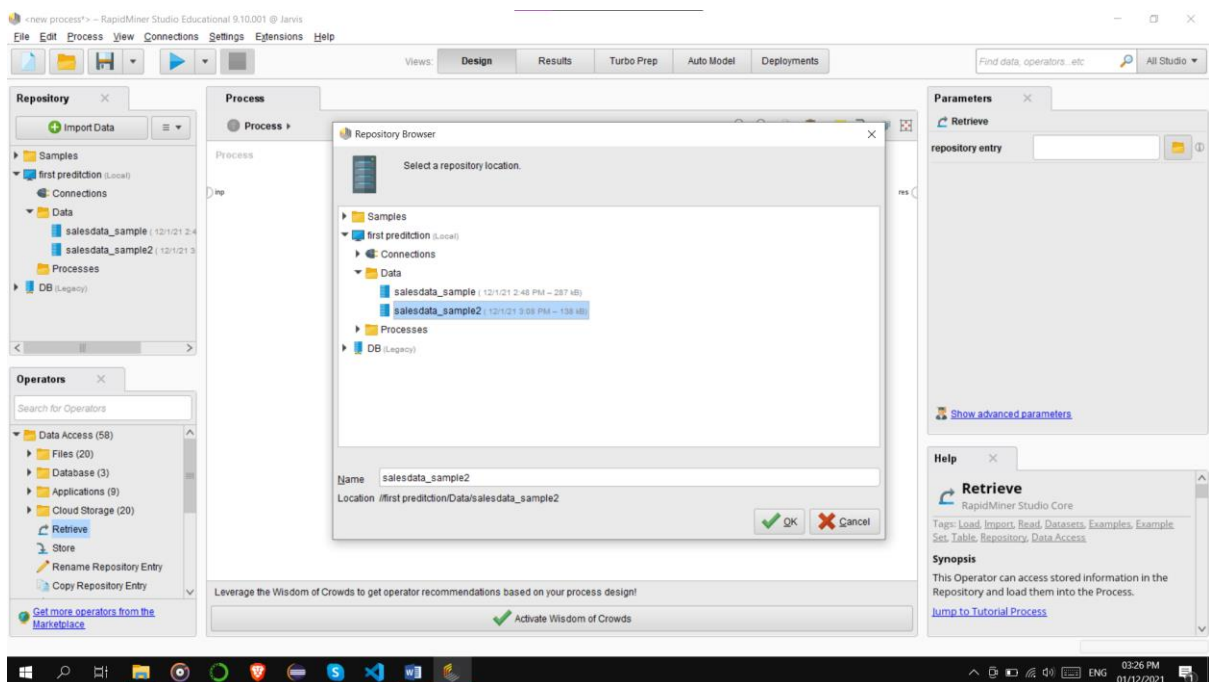
It has input and out Operator.
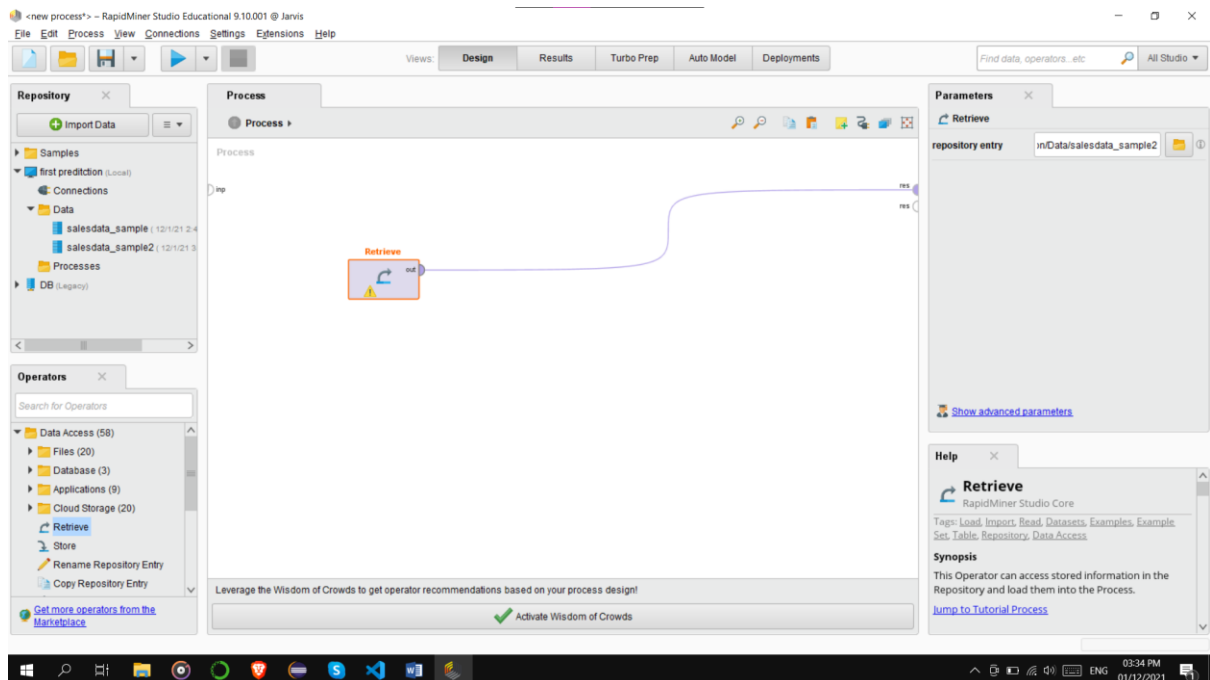


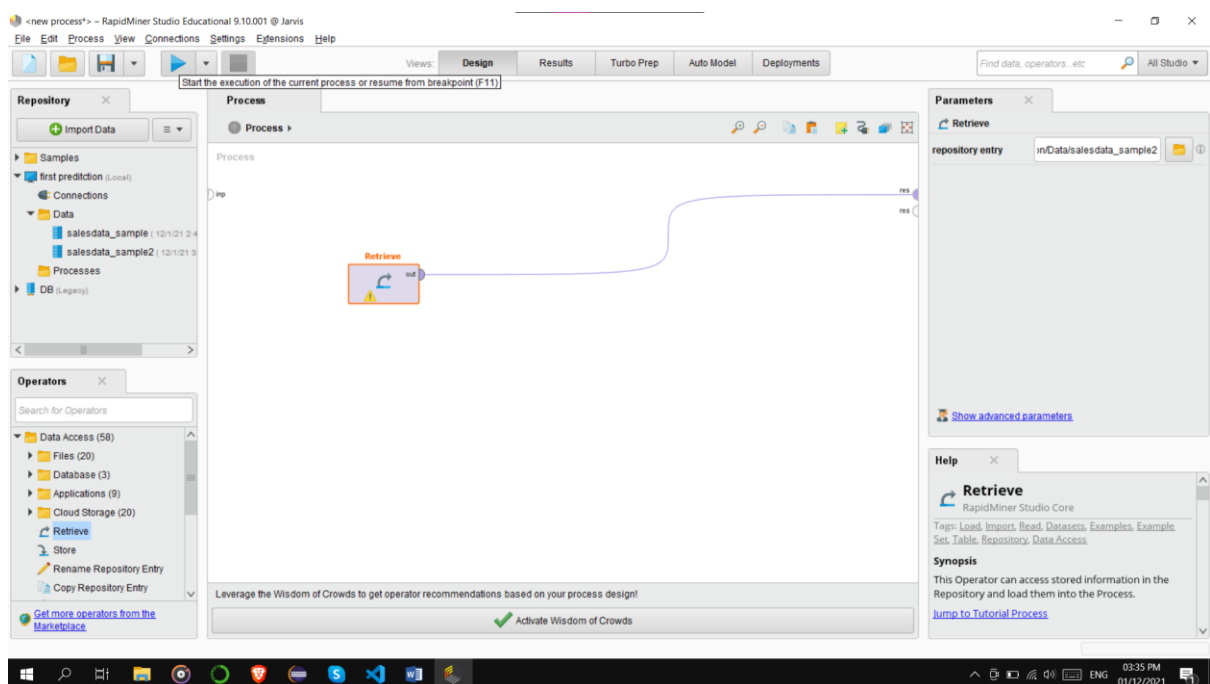Step 5 - Click on Repository Entry

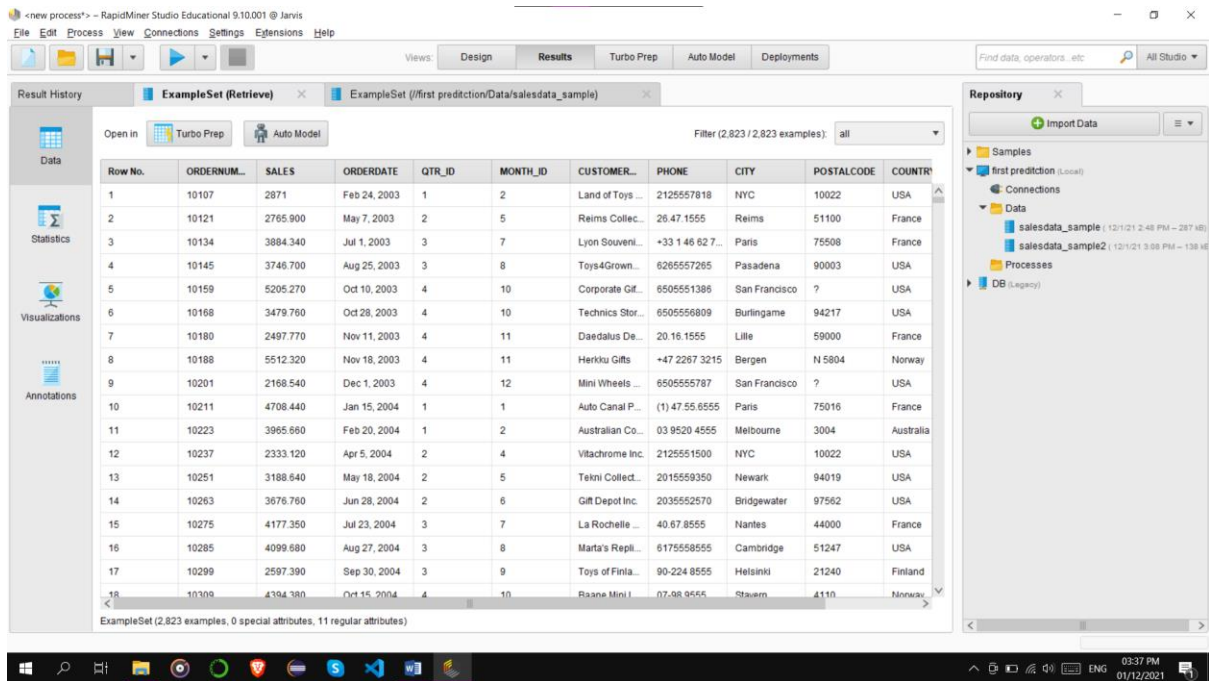## Step 6 - Select Local Repository



## Step 7 - Select updated dataset

# Step 8 - Join out operator to result operator
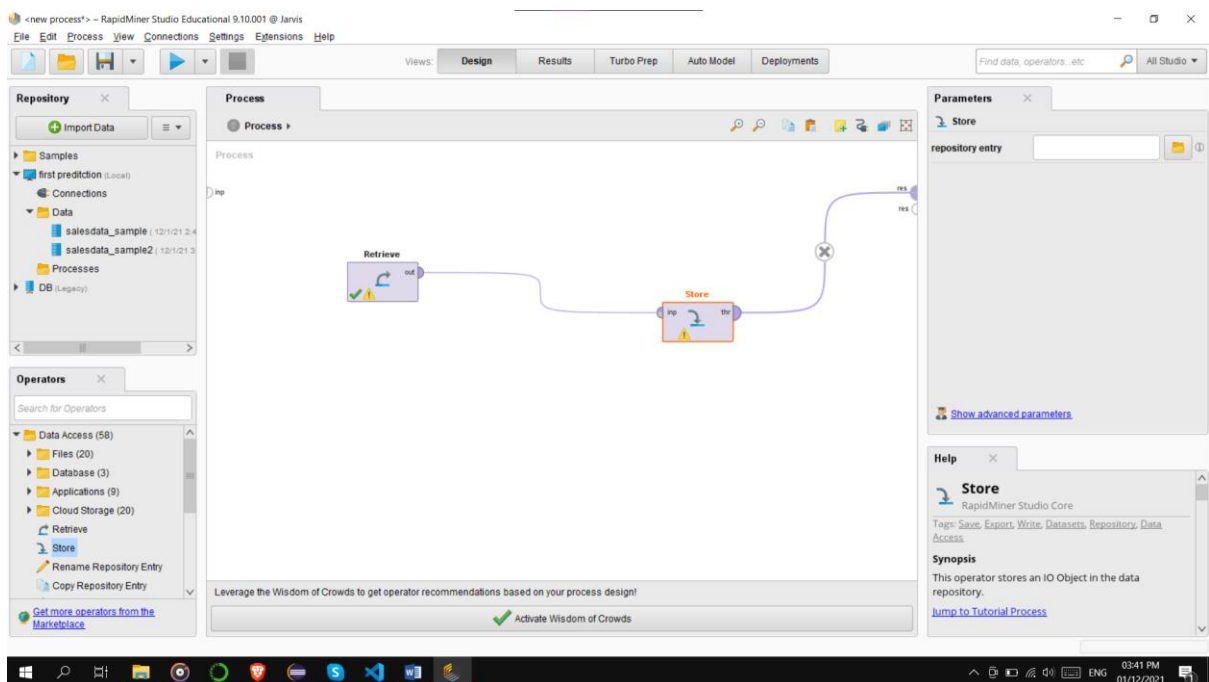


# Step 9 - Start Execution of Current Process
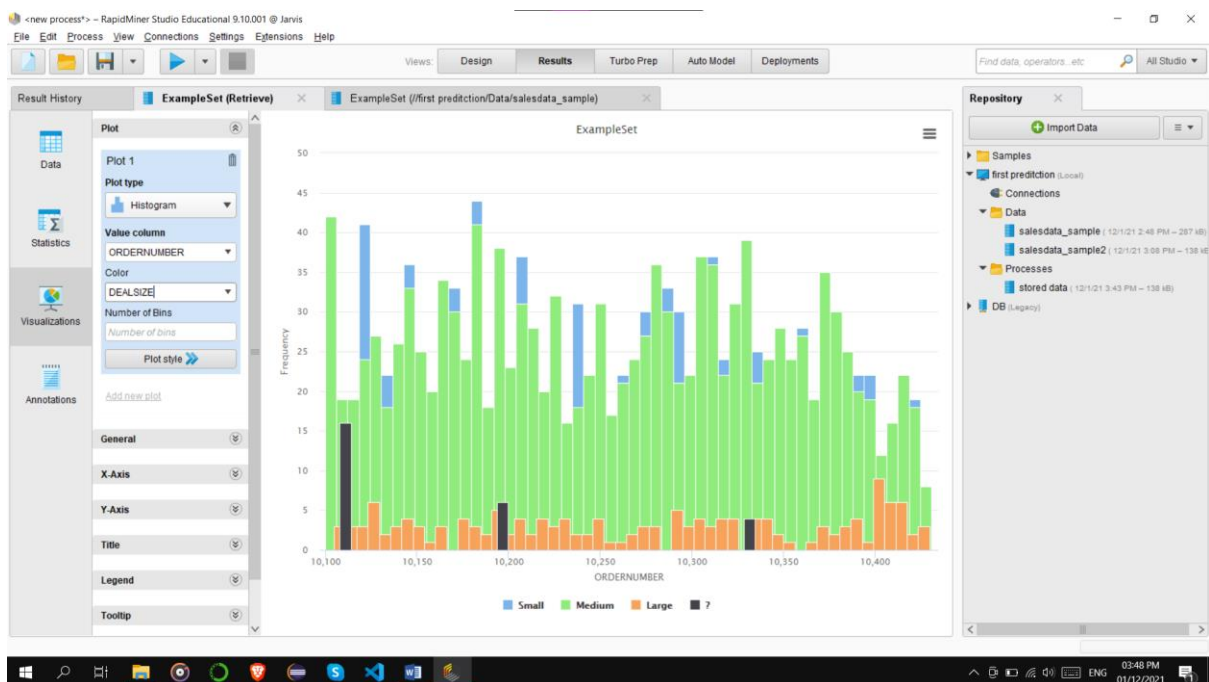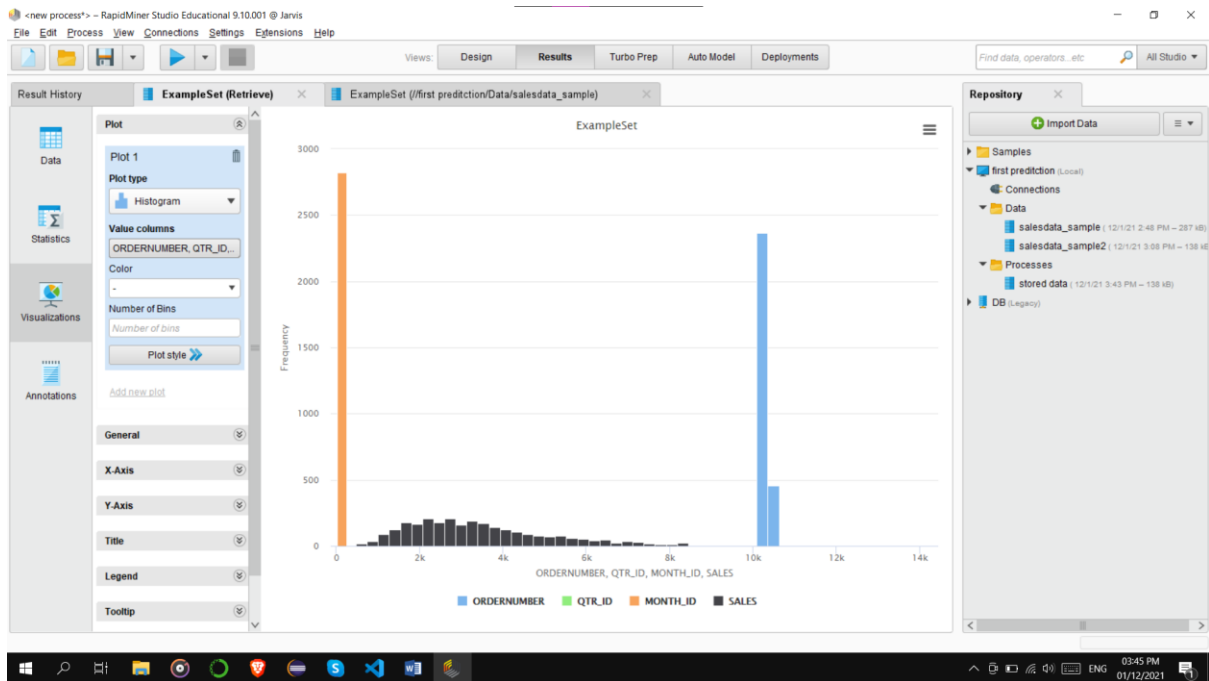
## Step 10 - Output Result Generated after Execution of Current Process



## Step 11 - Now add Store operator and connect it to result operator

Step 12 - You can also plot histograms or other charts of dataset
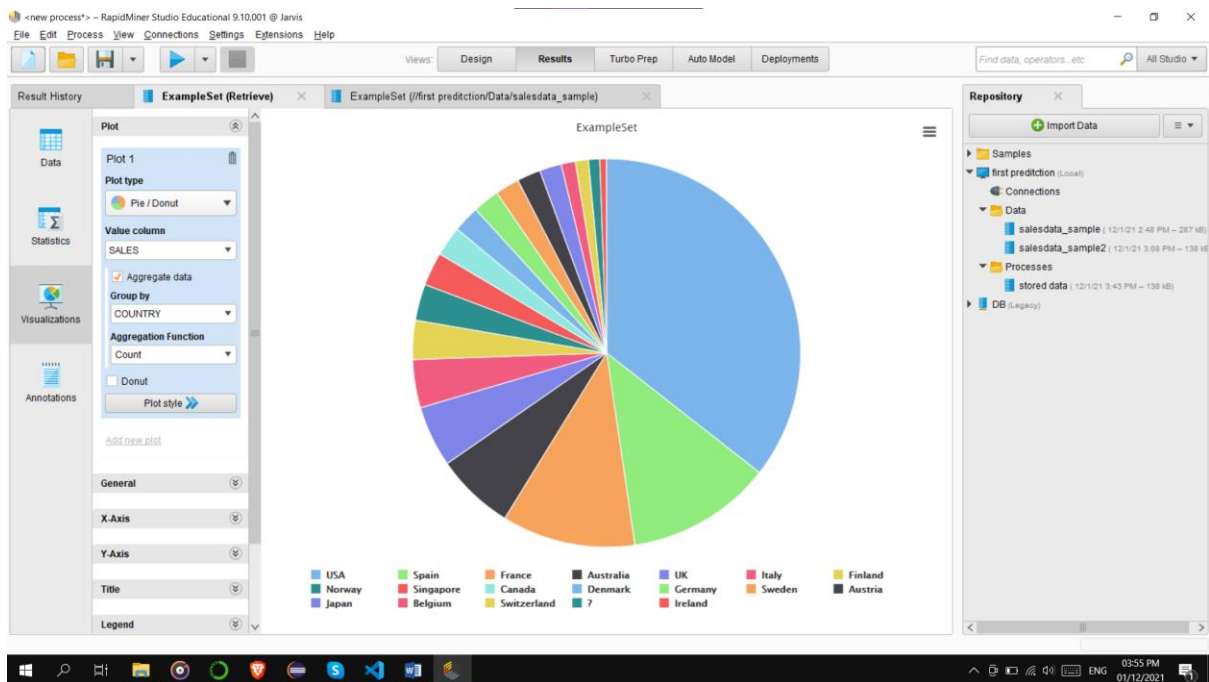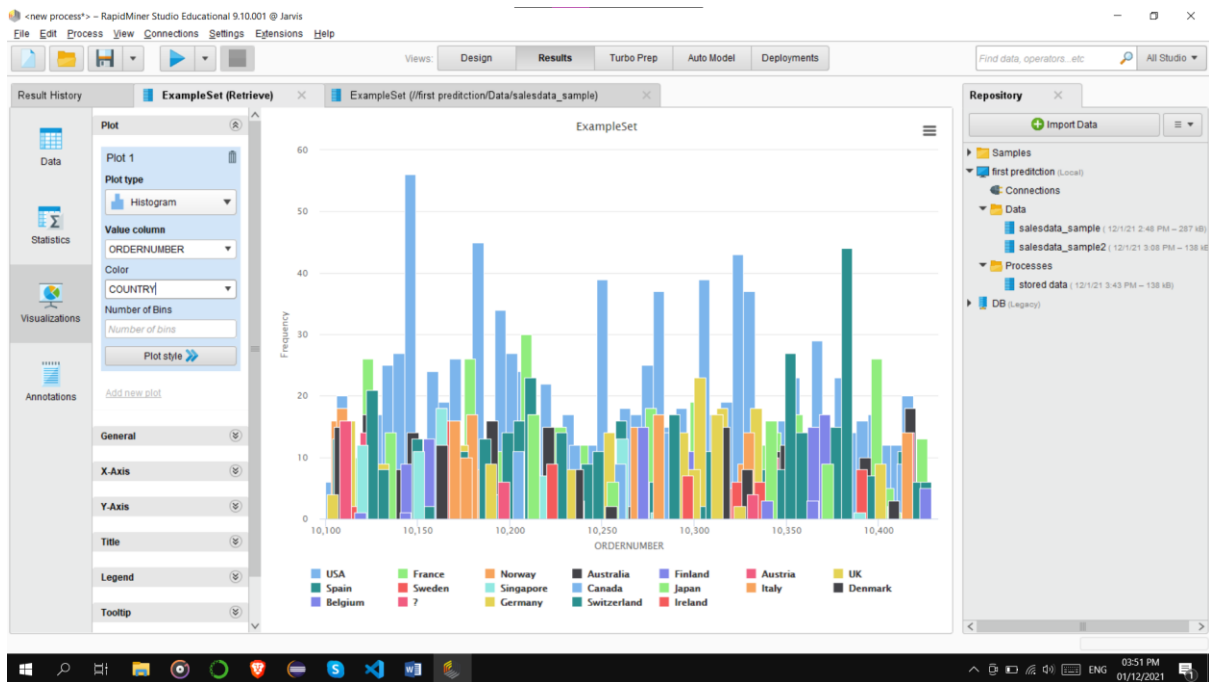
Conclusion :

Hence, we are able to study the Rapid Miner Tool, from which we can perform the ETL operations on the datasets and can perform analysis on those datasets.