

Name : Harshvardhan M Deshmukh

Roll No. : BE-A-35

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

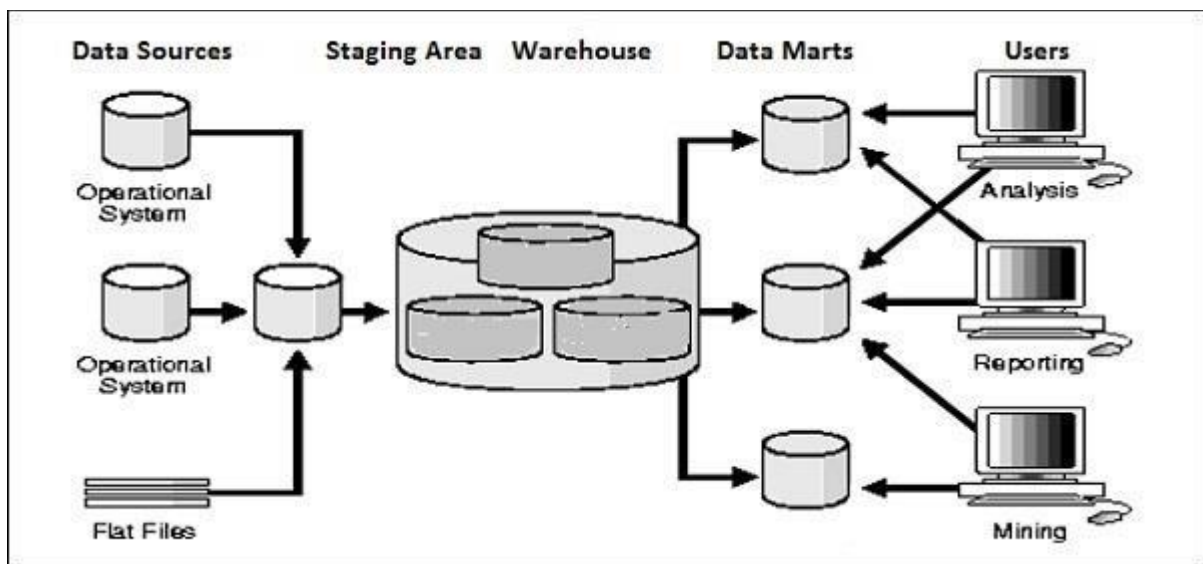
Theory :

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

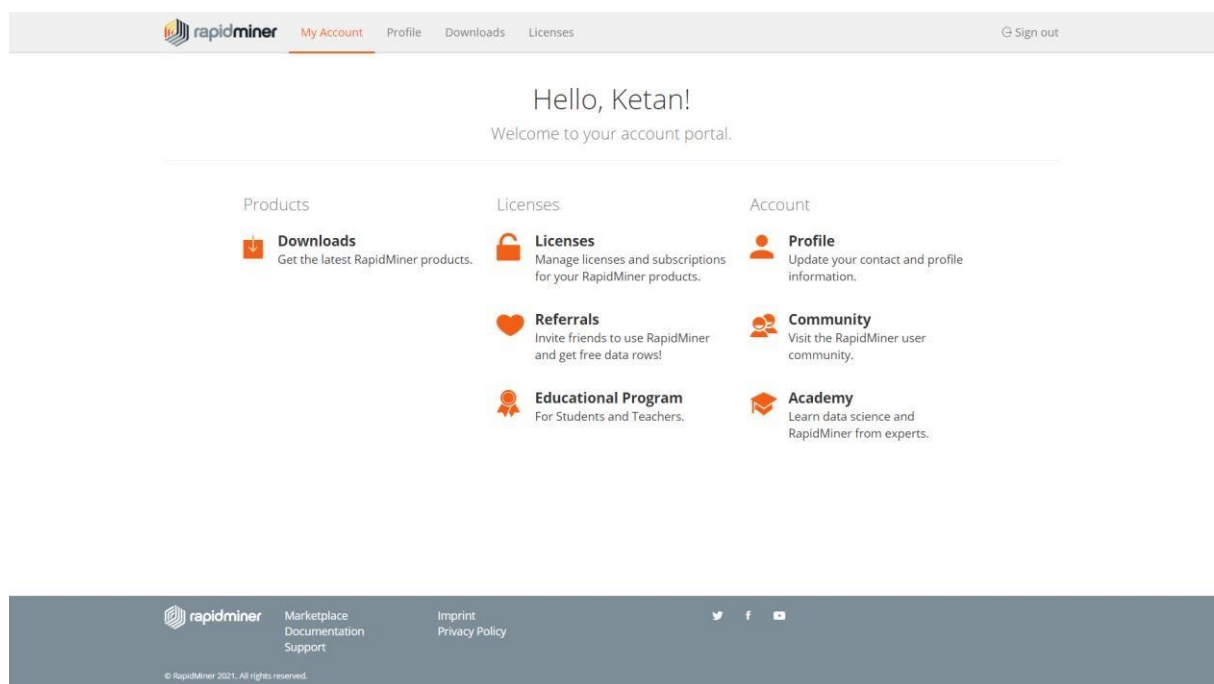
During Load phase, data is loaded into the end-target system and it can be a flat file or a Data Warehouse system.

Rapid Miner :

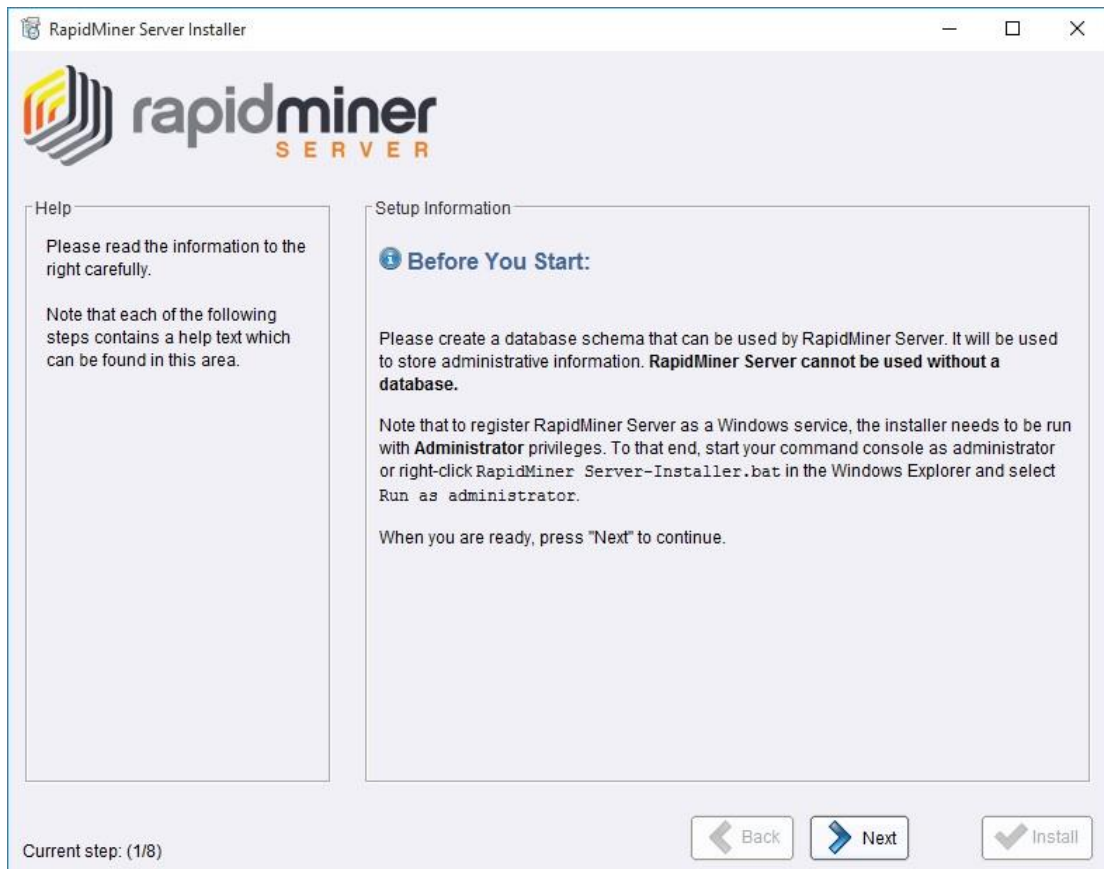
Rapid Miner is a world-leading open-source system for data mining. It is available as a stand-alone application for data analysis and as a data mining engine for the integration into own products. Rapid Miner is now Rapid Miner Studio and Rapid Analytics is now called Rapid Miner Server.

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



Help

In this step you can configure your Database connection which RapidMiner Server should use. You will need to enter the host or URL as well as the port and the desired DB schema. Username and Password can be filled in as needed. Then just select the appropriate JDBC driver and choose the driver class via the Dropdown menu. After you have set everything up, you can test the connection to the Database by clicking the Test Connection button.

Database Configuration

Database host: Database port:

Database schema:

Database username: Database password:

MySQL JDBC driver is not shipped with RapidMiner Server. Please click [here](#) for more information!

JDBC Driver location:  Database system:

☐ Use relative path

JDBC driver class:



Current step: (7/8) Back Next Install

5. Installing Radoop Proxy



Help

In this step you can enable the Radoop Proxy component that provides a proxy server for RapidMiner Studio users with Radoop extension installed. If you want to know more about Radoop Proxy and its use cases, please visit the <http://docs.rapidminer.com> website.

Radoop Proxy Installation

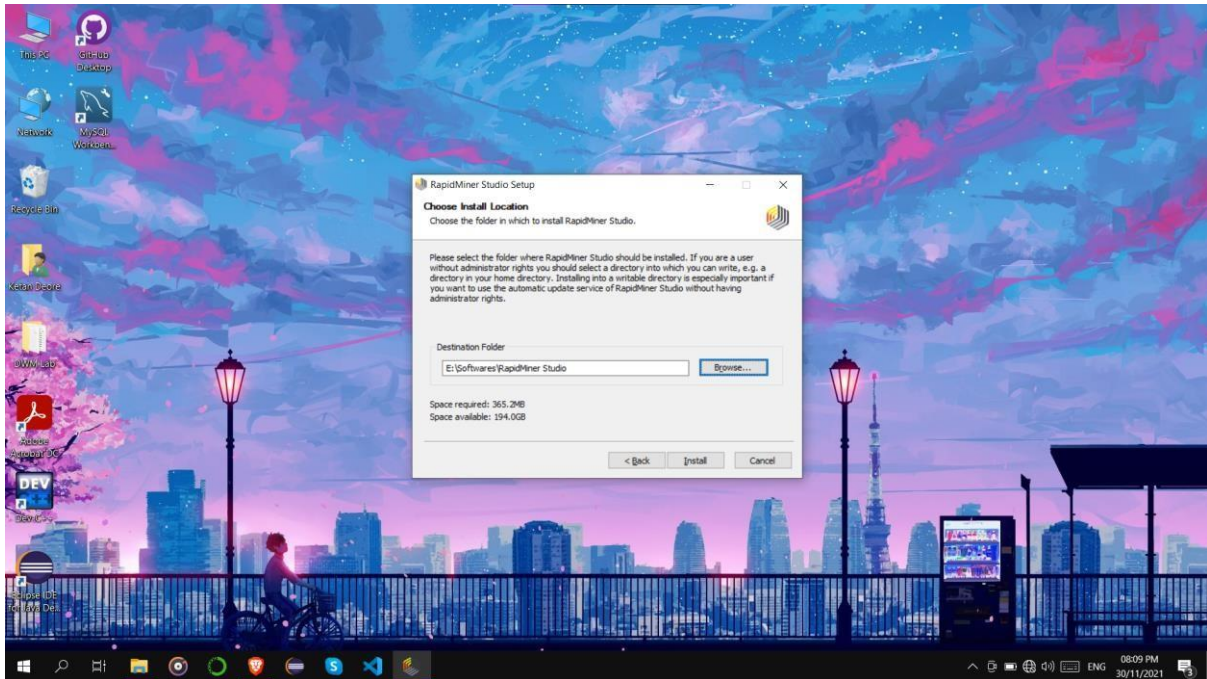
☐ Enable Radoop Proxy

Port:

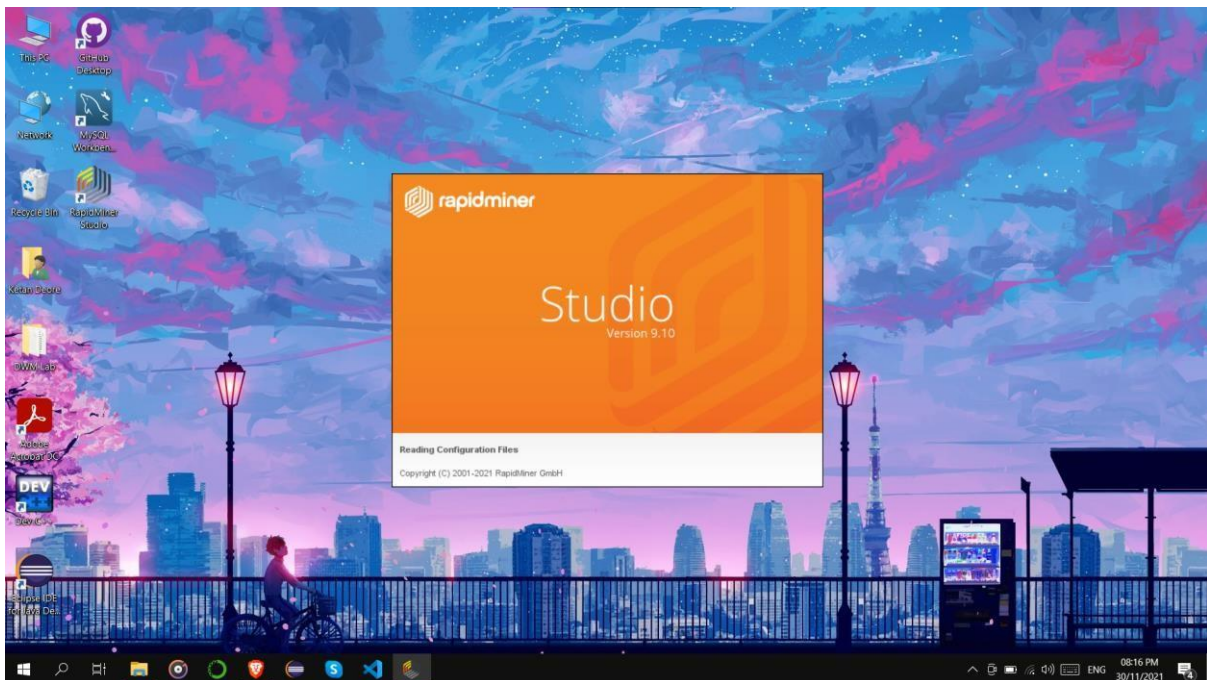
Current step: (9/9) Back Next Install

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.

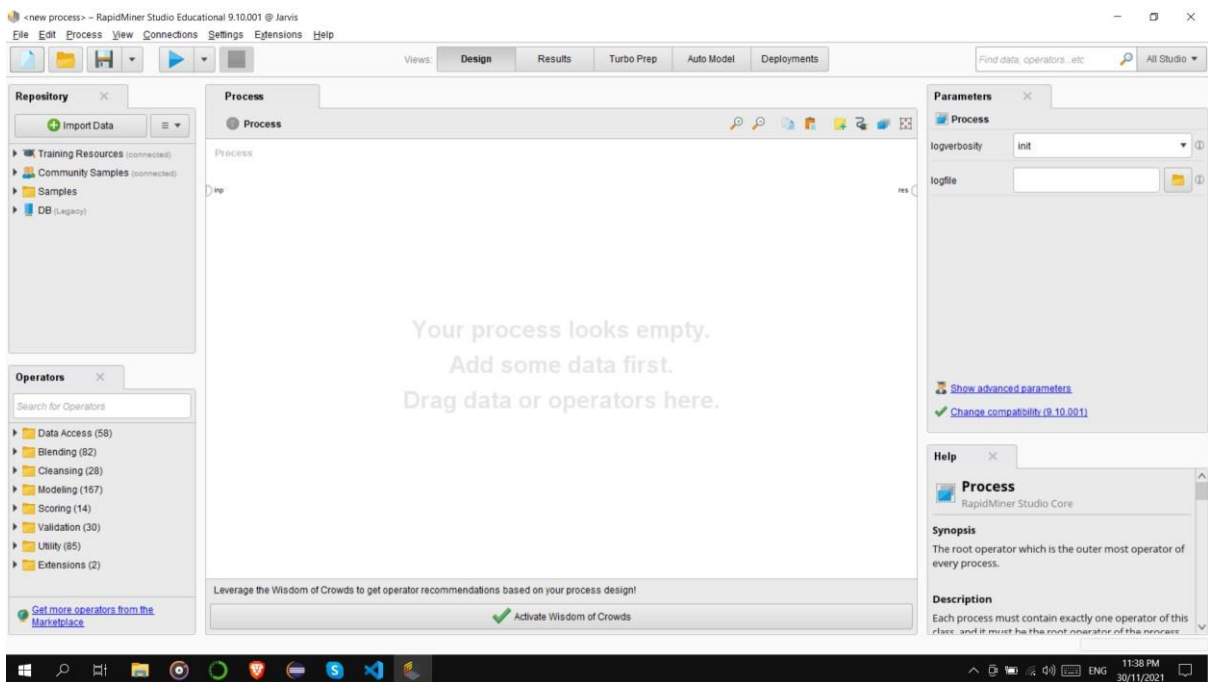
Snowflake Schema :

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.

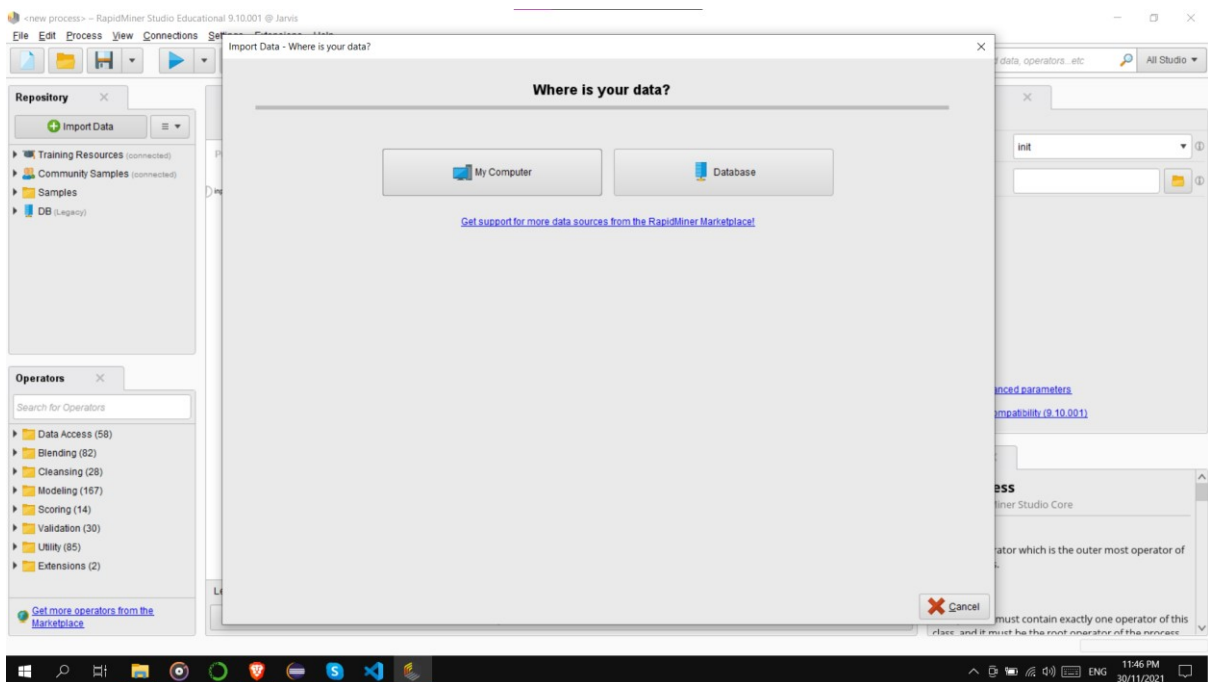
The dimension tables are normalized which splits data into additional tables. In the following example, Country is further normalized into an individual table.

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.

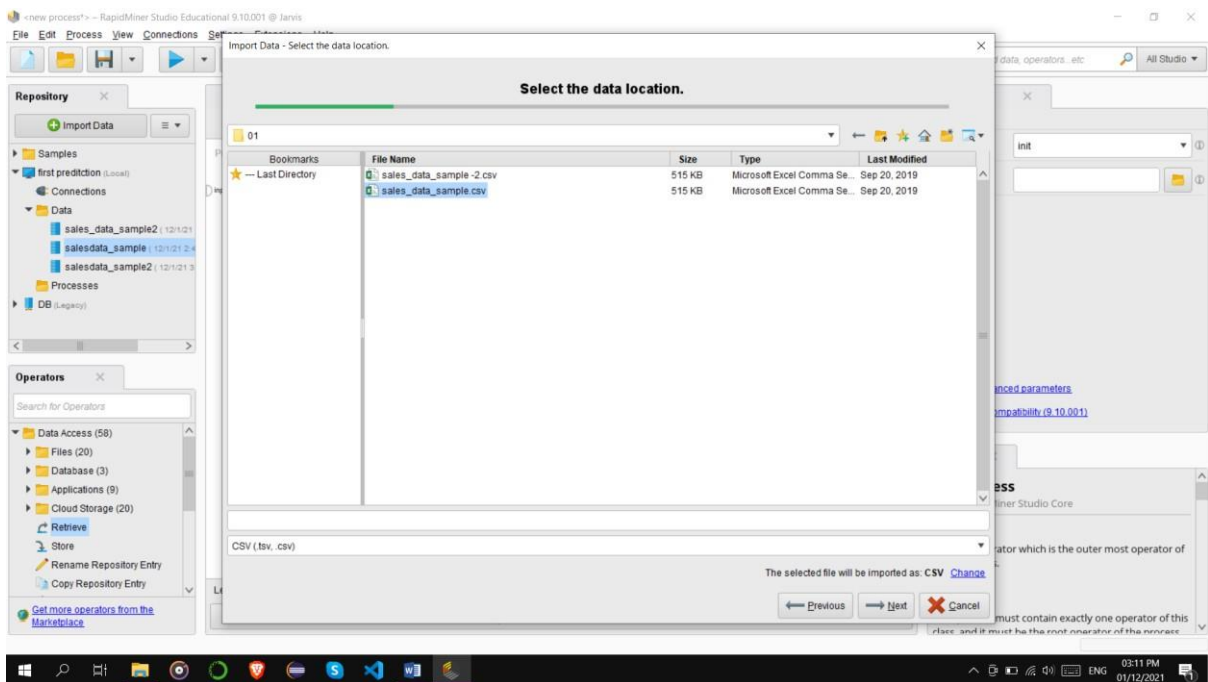
1. Design Model



Step 1 - Import Data from Source



Step 2 Select data Location



Step 3 - Open Dataset regarding business

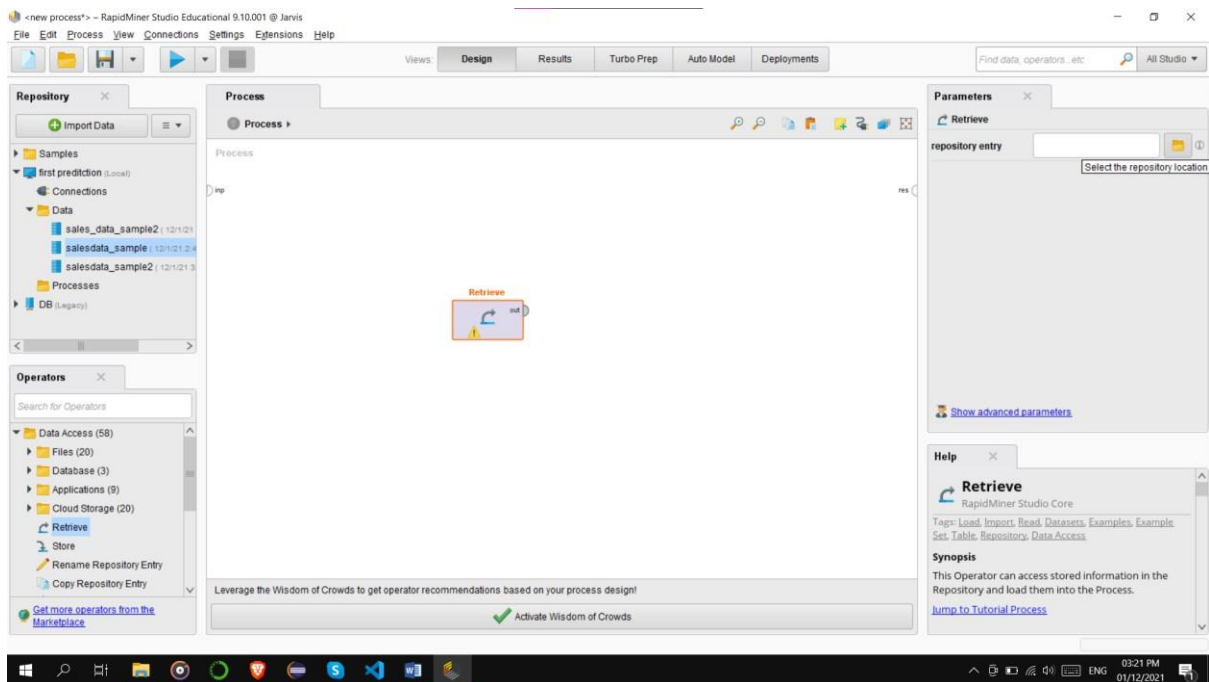
The screenshot shows the 'Results' view in RapidMiner Studio. The table displays 18 rows of sales data with columns: Row No., ORDERNUM..., QUANTITY..., PRICEEACH, ORDERLINE..., SALES, ORDERDATE, STATUS, QTR_ID, MONTH_ID, and YEAR_ID. The data includes various order numbers, quantities, prices, and dates from 2003 to 2004.

Row No.	ORDERNUM...	QUANTITY...	PRICEEACH	ORDERLINE...	SALES	ORDERDATE	STATUS	QTR_ID	MONTH_ID	YEAR_ID
1	10107	30	95.700	2	2871	Feb 24, 2003	Shipped	1	2	2003
2	10121	34	81.350	5	2765.900	May 7, 2003	Shipped	2	5	2003
3	10134	41	94.740	2	3884.340	Jul 1, 2003	Shipped	3	7	2003
4	10145	45	83.260	6	3746.700	Aug 25, 2003	Shipped	3	8	2003
5	10159	49	100	14	5205.270	Oct 10, 2003	Shipped	4	10	2003
6	10168	36	96.660	1	3479.760	Oct 28, 2003	Shipped	4	10	2003
7	10180	29	86.130	9	2497.770	Nov 11, 2003	Shipped	4	11	2003
8	10188	48	100	1	5512.320	Nov 18, 2003	Shipped	4	11	2003
9	10201	22	98.570	2	2168.540	Dec 1, 2003	Shipped	4	12	2003
10	10211	41	100	14	4708.440	Jan 15, 2004	Shipped	1	1	2004
11	10223	37	100	1	3965.660	Feb 20, 2004	Shipped	1	2	2004
12	10237	23	100	7	2333.120	Apr 5, 2004	Shipped	2	4	2004
13	10251	28	100	2	3188.640	May 18, 2004	Shipped	2	5	2004
14	10263	34	100	2	3676.760	Jun 28, 2004	Shipped	2	6	2004
15	10275	45	92.830	1	4177.350	Jul 23, 2004	Shipped	3	7	2004
16	10285	36	100	6	4099.680	Aug 27, 2004	Shipped	3	8	2004
17	10299	23	100	9	2597.390	Sep 30, 2004	Shipped	3	9	2004
18	10309	41	100	5	4394.380	Oct 15, 2004	Shipped	4	10	2004

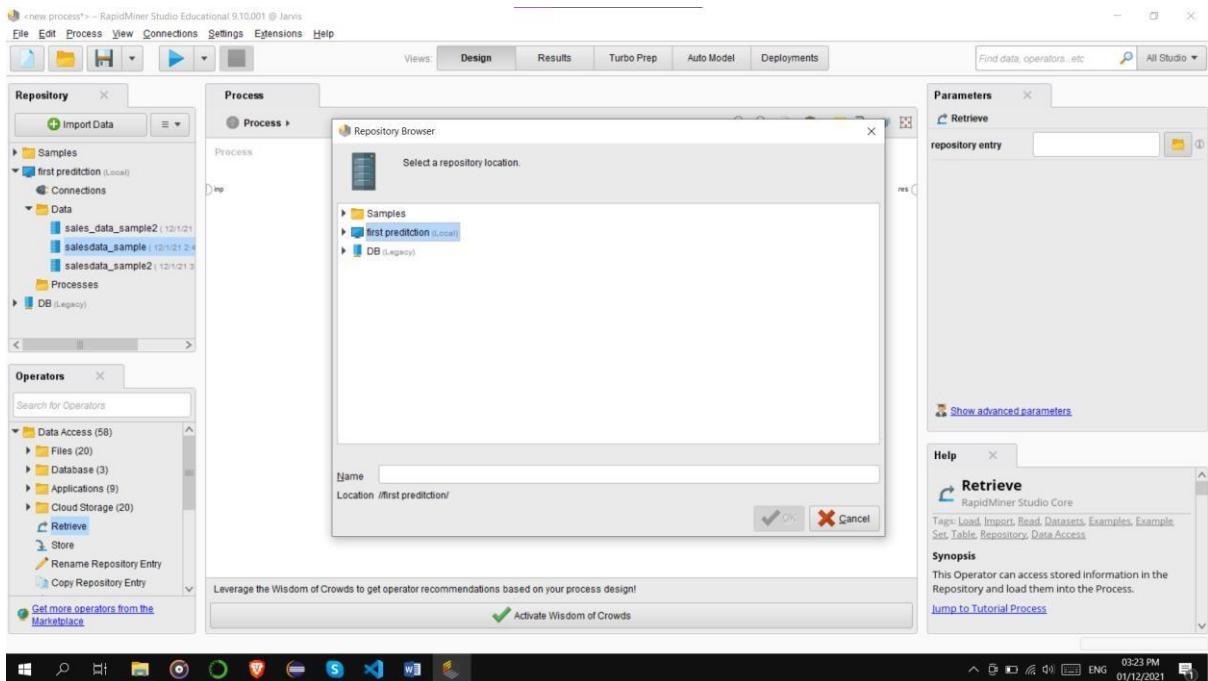
Step 4 Click on retrieve operator drag in process view,
It has input and out Operator.

The screenshot shows the 'Design' view in RapidMiner Studio. The 'Operators' panel on the left lists various operators, and the 'Retrieve' operator is highlighted. The 'Process' canvas in the center is empty, with a 'Drag here' prompt. The 'Parameters' panel on the right shows settings for the 'Process' operator, including 'logverbosity' and 'logfile'. The 'Help' panel on the bottom right provides information about the 'Process' operator.

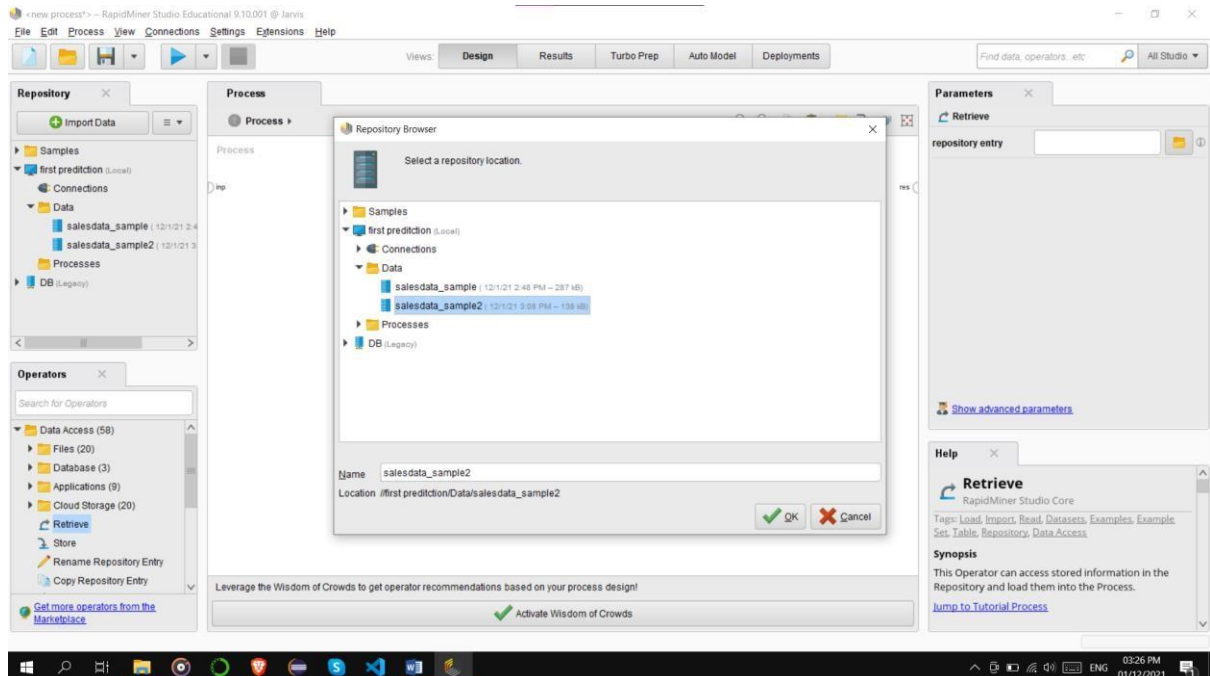
Step 5 - Click on Repository Entry



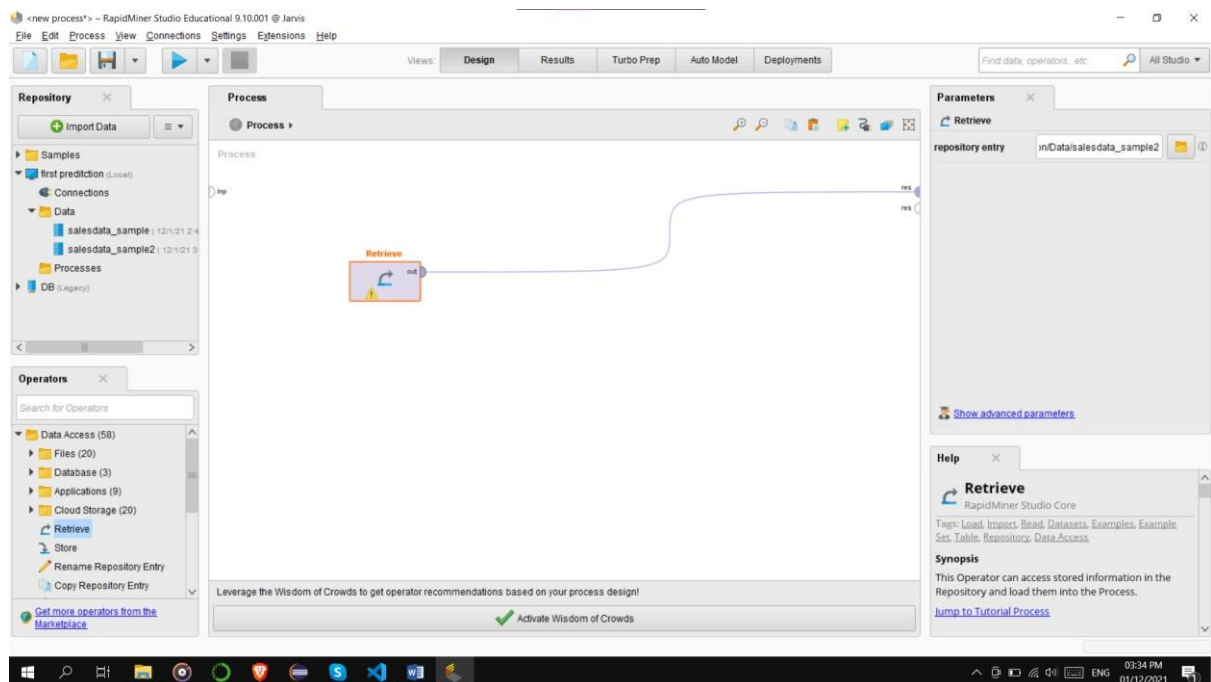
Step 6 Select Local Repository



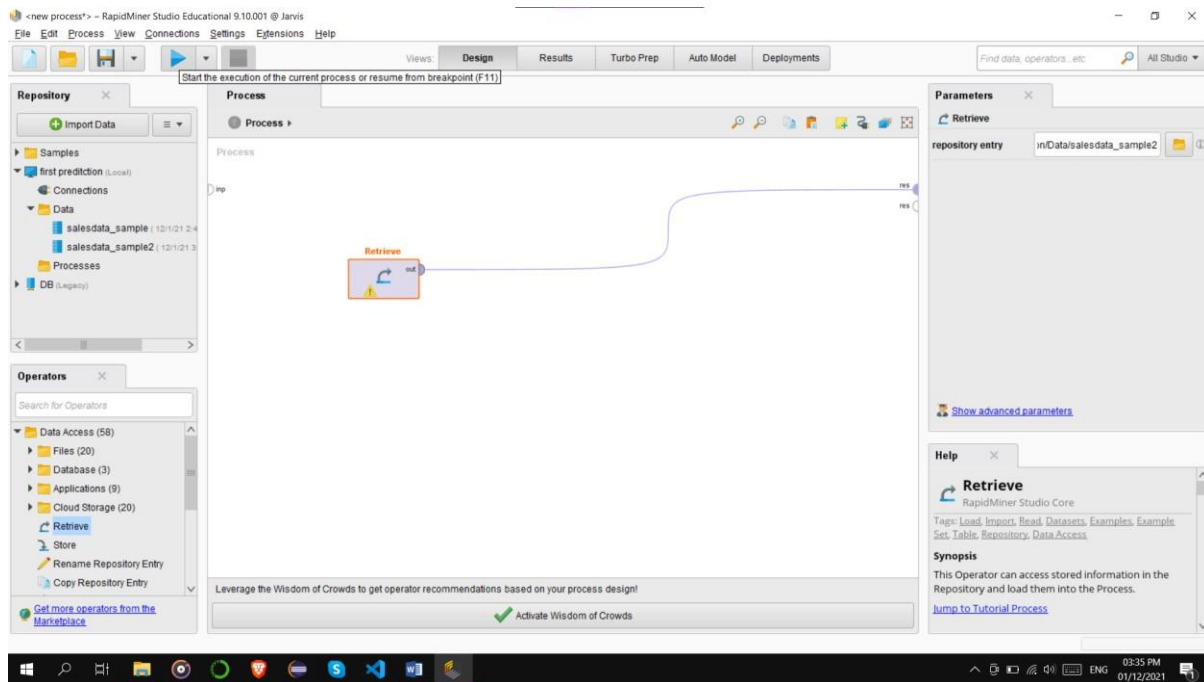
Step 7 - Select updated dataset



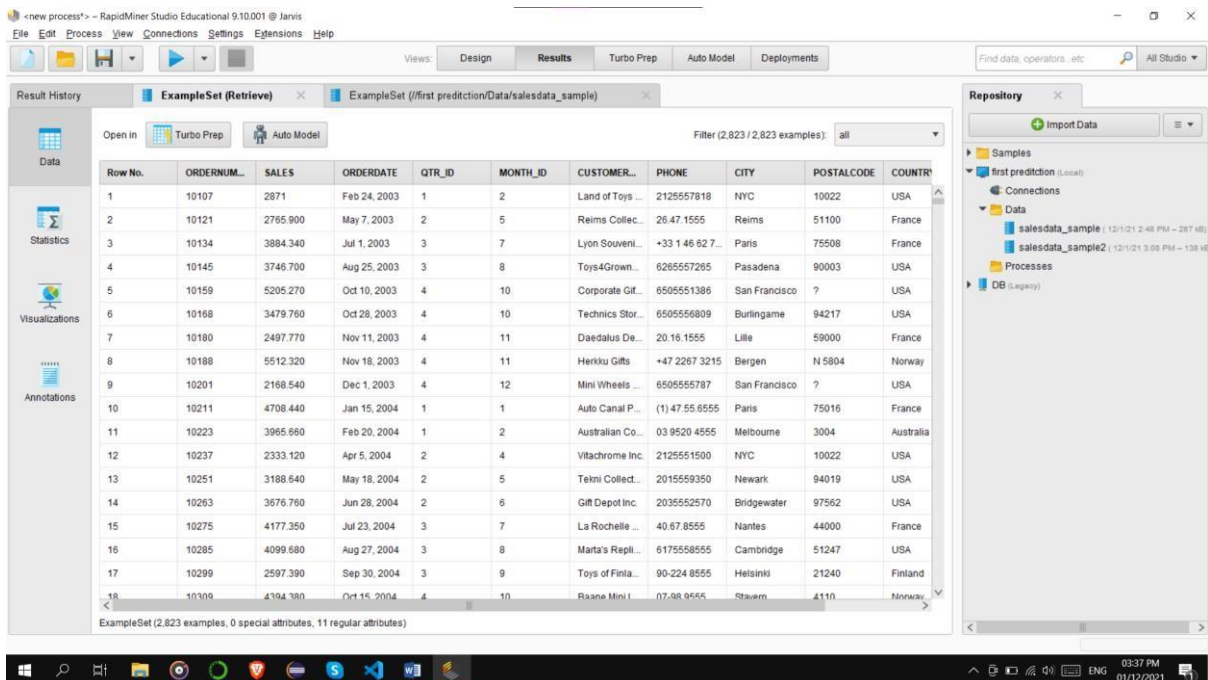
Step 8 Join out operator to result operator



Step 9 - Start Execution of Current Process



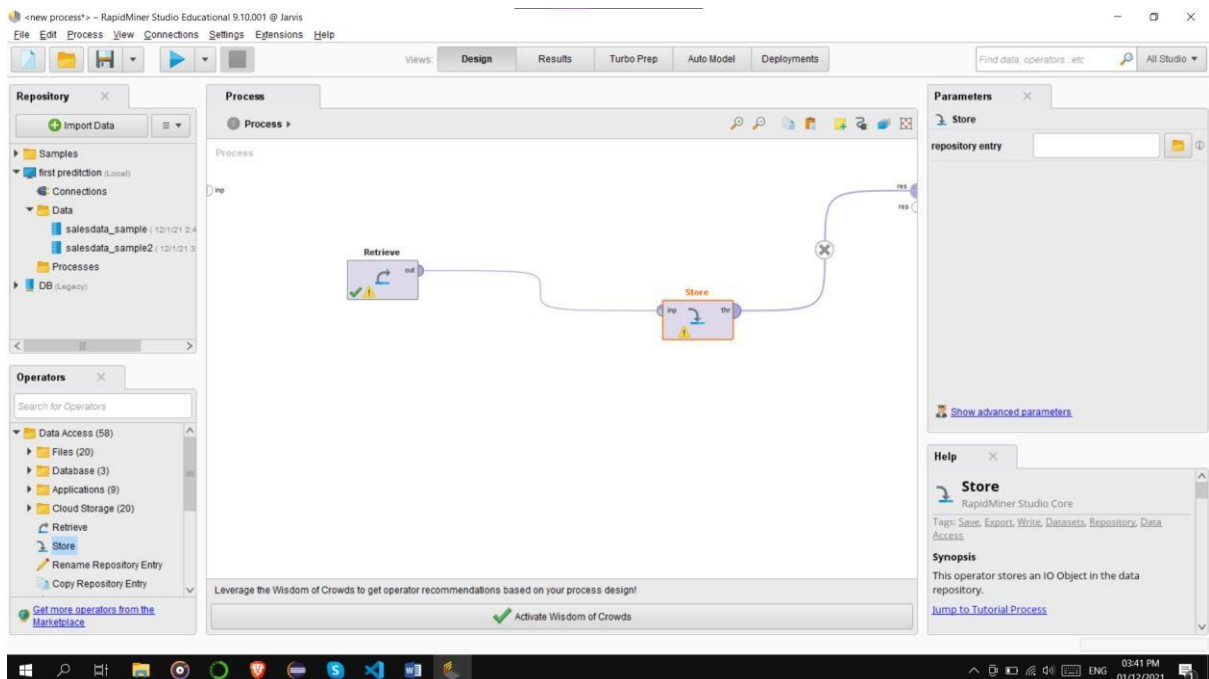
Step 10 - Output Result Generated after Execution of Current Process



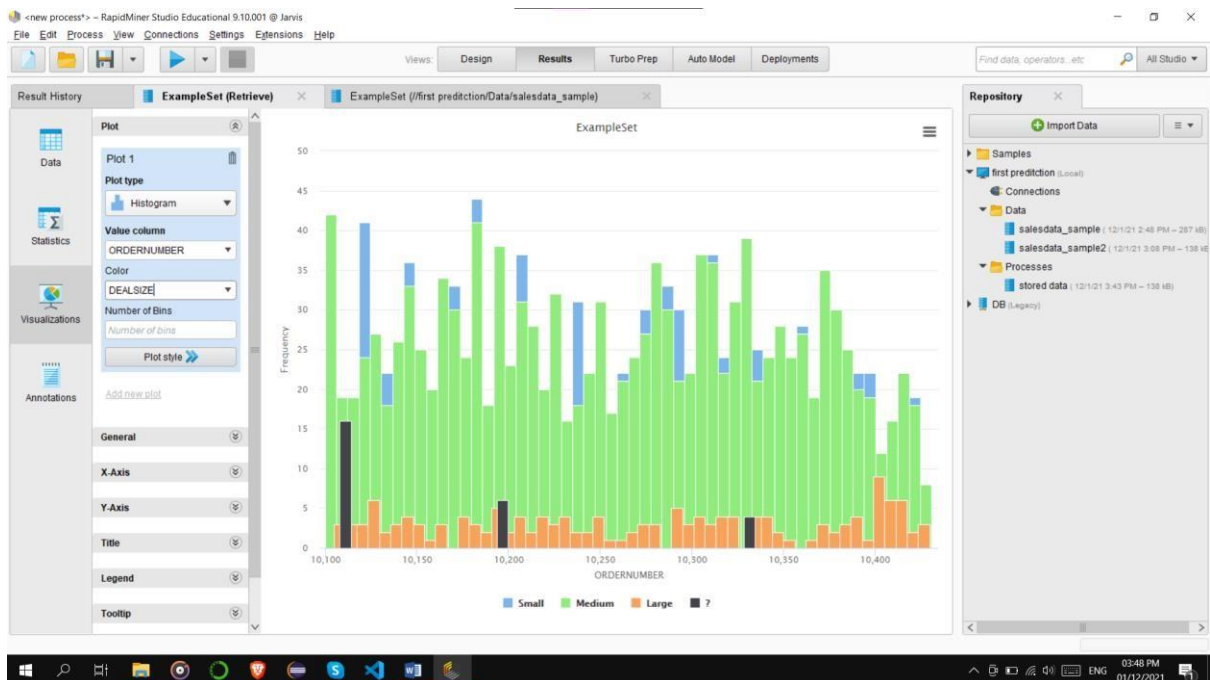
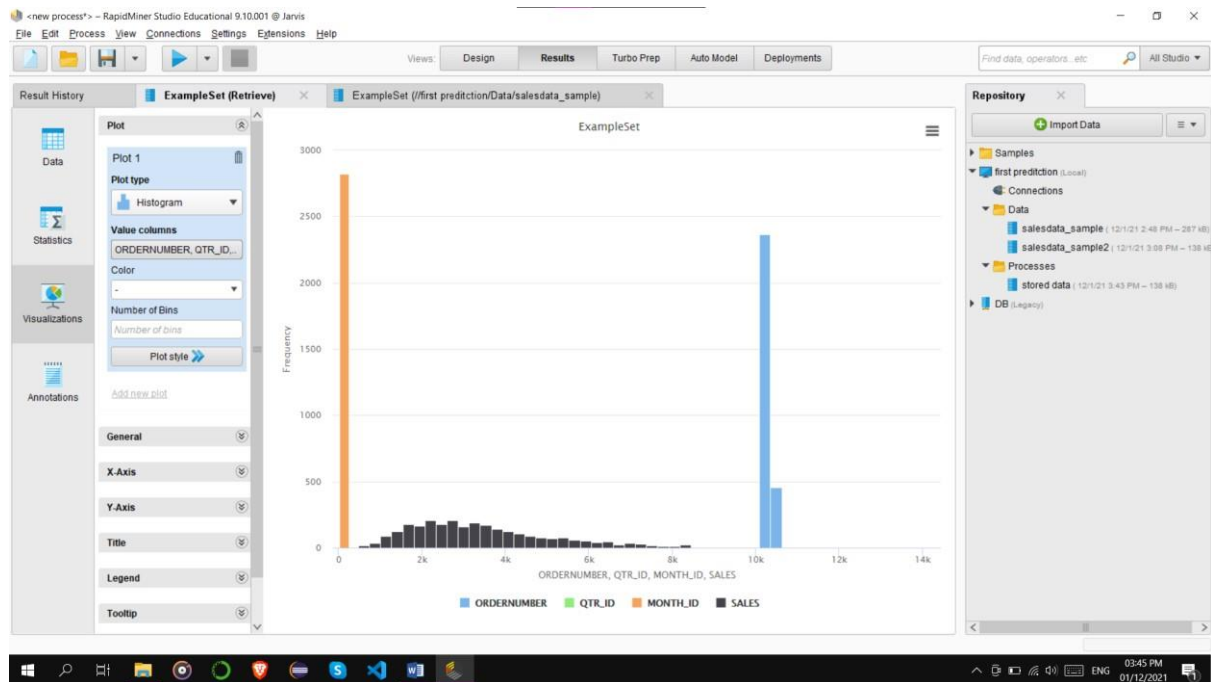
The screenshot shows the RapidMiner Studio interface with the 'Results' tab selected. The main window displays a table of sales data with 18 rows and 11 columns. The table is titled 'ExampleSet (Retrieve)' and 'ExampleSet (/first prediction/Data/salesdata_sample)'. The table contains the following data:

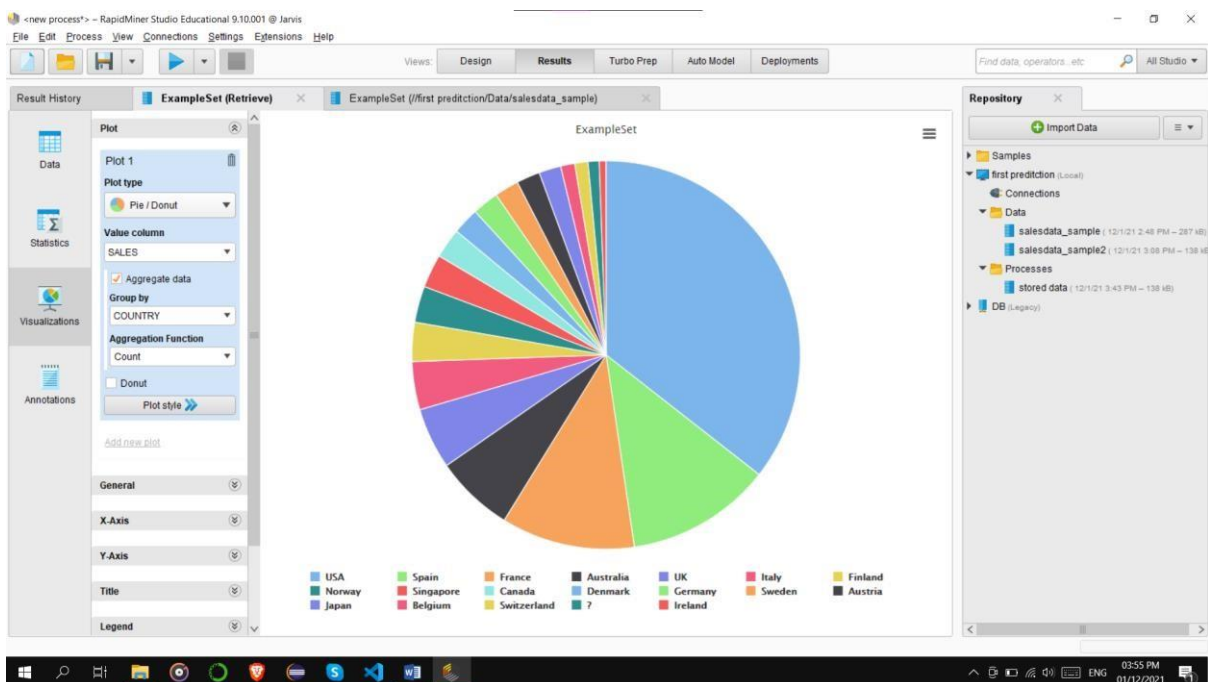
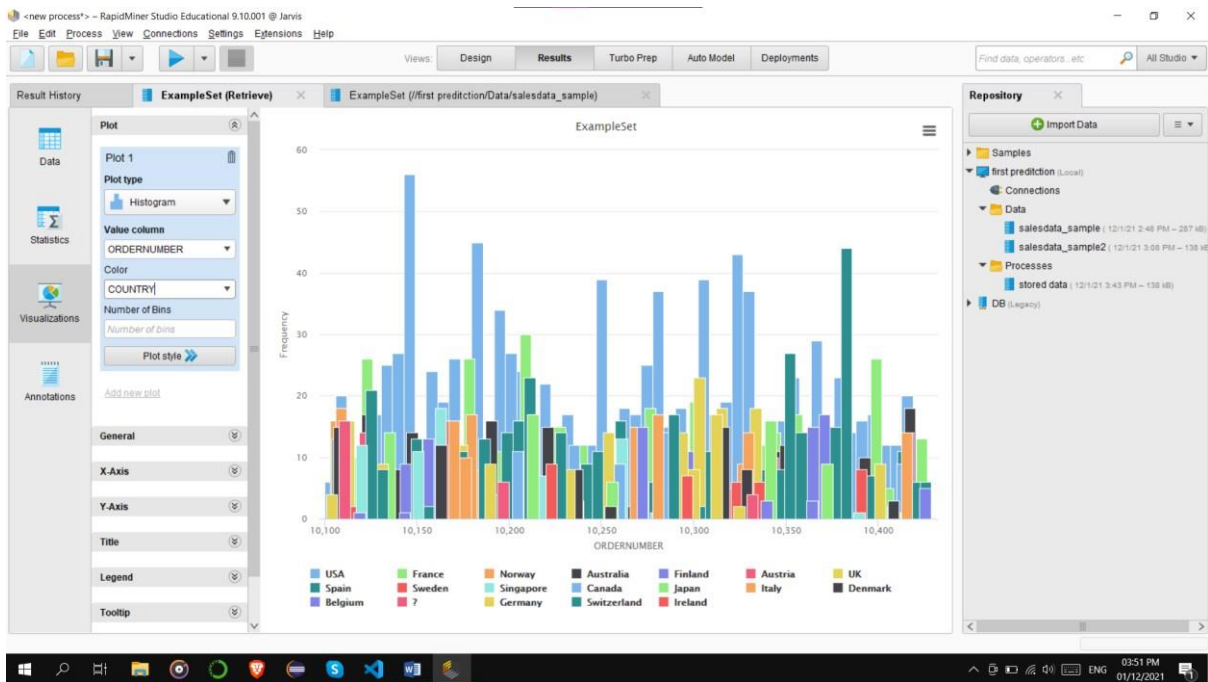
Row No.	ORDERNUM...	SALES	ORDERDATE	QTR_ID	MONTH_ID	CUSTOMER...	PHONE	CITY	POSTALCODE	COUNTRY
1	10107	2871	Feb 24, 2003	1	2	Land of Toys ...	2125557818	NYC	10022	USA
2	10121	2765.900	May 7, 2003	2	5	Reims Collec...	26.47.1555	Reims	51100	France
3	10134	3884.340	Jul 1, 2003	3	7	Lyon Souveni...	+33 1 46 62 7...	Paris	75508	France
4	10145	3746.700	Aug 25, 2003	3	8	Toys4Grown...	6265557265	Pasadena	90003	USA
5	10159	5205.270	Oct 10, 2003	4	10	Corporate Gif...	6505551385	San Francisco	?	USA
6	10168	3479.760	Oct 28, 2003	4	10	Technics Stor...	6505556809	Burlingame	94217	USA
7	10180	2497.770	Nov 11, 2003	4	11	Daedalus De...	20.16.1555	Lille	59000	France
8	10188	5512.320	Nov 18, 2003	4	11	Herku Gifts	+47 2267 3215	Bergen	N 5804	Norway
9	10201	2168.540	Dec 1, 2003	4	12	Mini Wheels ...	6505555787	San Francisco	?	USA
10	10211	4708.440	Jan 15, 2004	1	1	Auto Canal P...	(1) 47.55.6555	Paris	75016	France
11	10223	3965.660	Feb 20, 2004	1	2	Australian Co...	03 9520 4555	Melbourne	3004	Australia
12	10237	2333.120	Apr 5, 2004	2	4	Vitachrome Inc.	2125551500	NYC	10022	USA
13	10251	3188.640	May 18, 2004	2	5	Tekni Collect...	2015559350	Newark	94019	USA
14	10263	3676.760	Jun 28, 2004	2	6	Gift Depot Inc.	2035552570	Bridgewater	97562	USA
15	10275	4177.350	Jul 23, 2004	3	7	La Rochelle ...	40.67.8555	Nantes	44000	France
16	10285	4099.680	Aug 27, 2004	3	8	Marta's Repli...	6175558555	Cambridge	51247	USA
17	10299	2597.390	Sep 30, 2004	3	9	Toys of Finla...	90-224 8555	Helsinki	21240	Finland
18	10308	4194.180	Oct 15, 2004	4	10	Baana Mini I...	07-98 8444	Stavem	4110	Norway

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.