

Name: Kunal Kharalkar

Roll No.: A-52

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

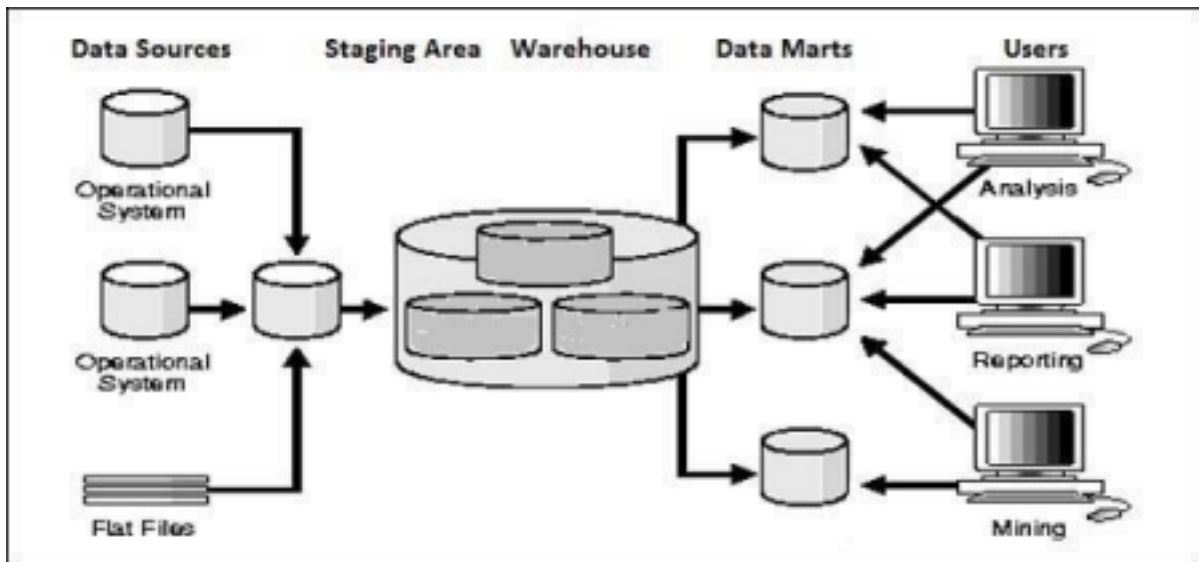
Theory :

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

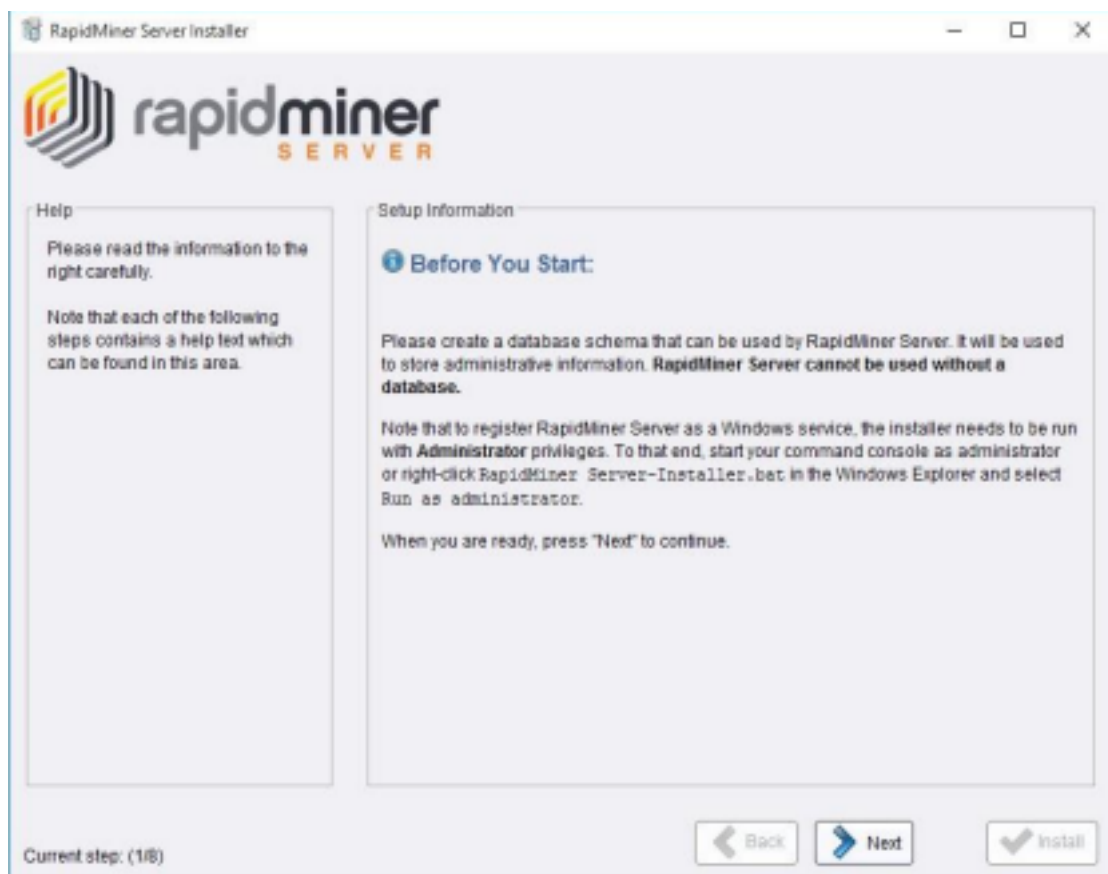
During Load phase, data is loaded into the end-target system and it can be a flat file or a Data Warehouse system.

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

RapidMiner Server Installer

rapidminer SERVER

Help

In this step, you can specify a host name and port under which clients, foremost RapidMiner Studio, will connect to RapidMiner Server. Therefore, you must choose a valid hostname. If you check "Bind to this hostname only", RapidMiner Server will listen only on the respective network interface.

Furthermore, you can assign the amount of memory utilized by RapidMiner Server (in MB) and optionally register it as a Windows service.

If you do not have the JAVA_HOME Environment variable set, you need to specify your Java directory.

Server Settings

Hostname: ☐ Bind to this hostname only

Port for web interface: Internal Port:

Server web interface will be available at <http://rapidminer.example.com:8080>

Server Memory (in MB):

Number of bundled Job Containers: Memory per Job Container (in MB):

RapidMiner Server will allocate memory up to **4,096 MB** (System: **20,354 MB**)

☒ Register as Windows service (needs administrator privileges)

Service ID: Service Name:

JAVA_HOME folder: 

Current step: (6/6) Version: 8.0.0

[< Back](#) [Next >](#) [Install](#)

4. Configuring Rapid Miner server's database connection

RapidMiner Server Installer

rapidminer SERVER

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:

[Test Connection](#)

Current step: (7/8)

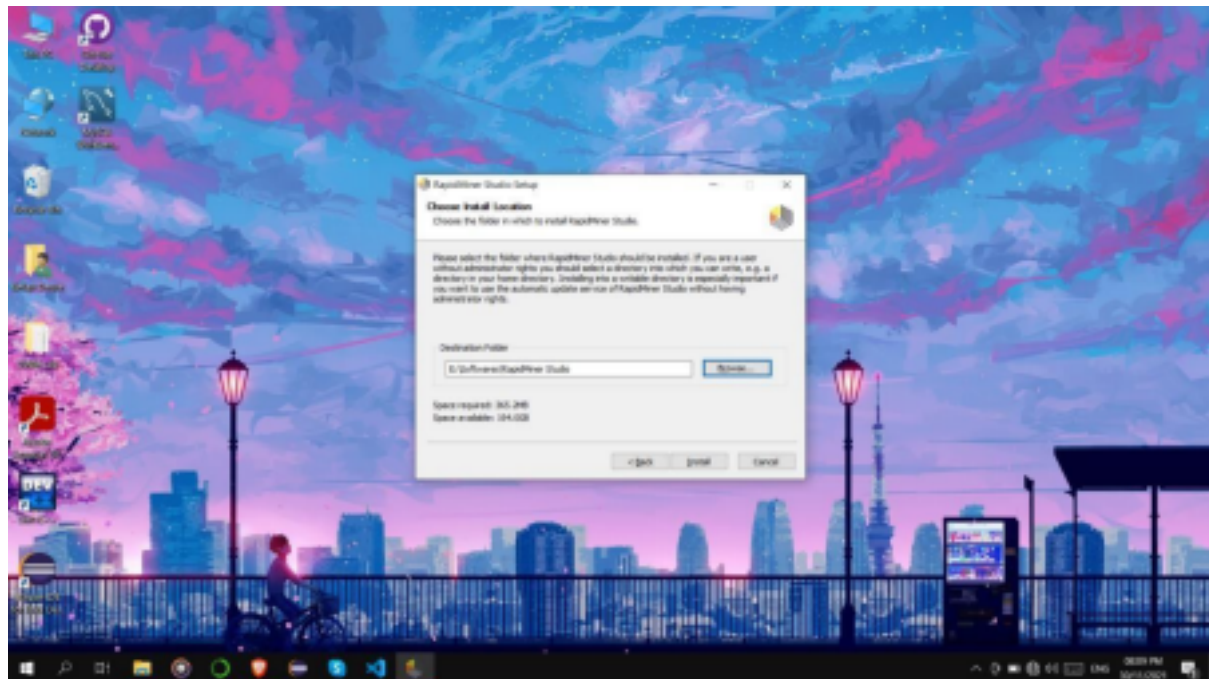
[< Back](#) [Next >](#) [Install](#)

5. Installing Radoop Proxy

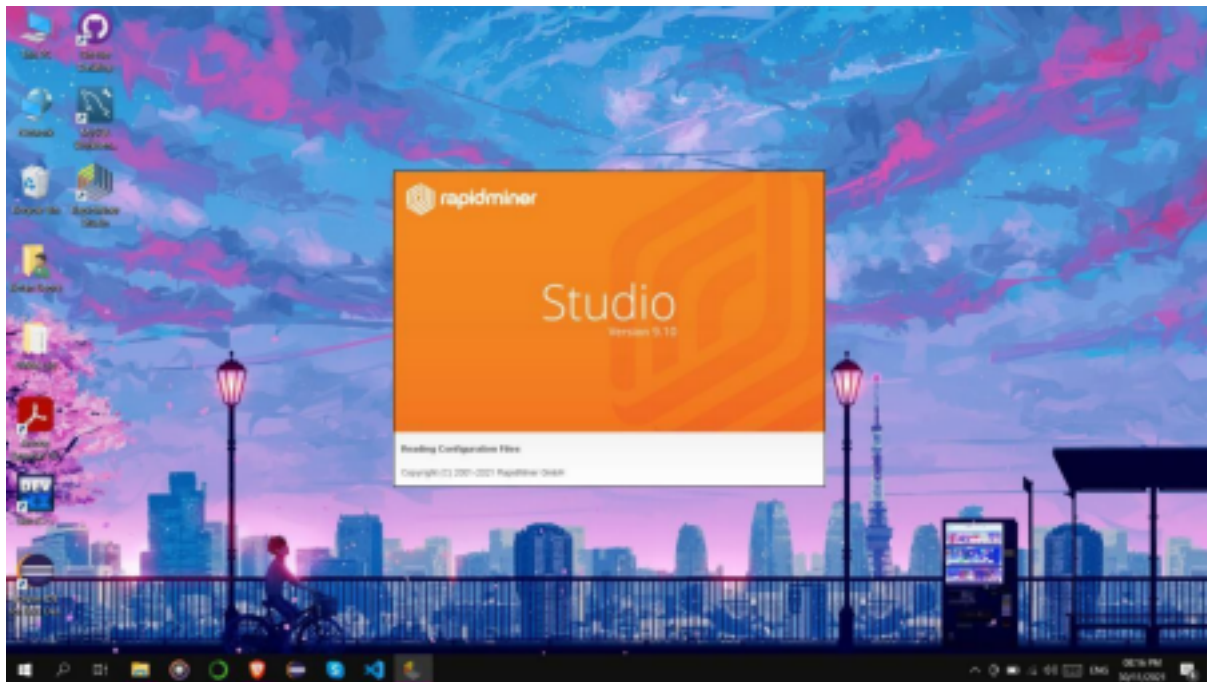


6. Completing the Installation of Rapid Miner Server

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

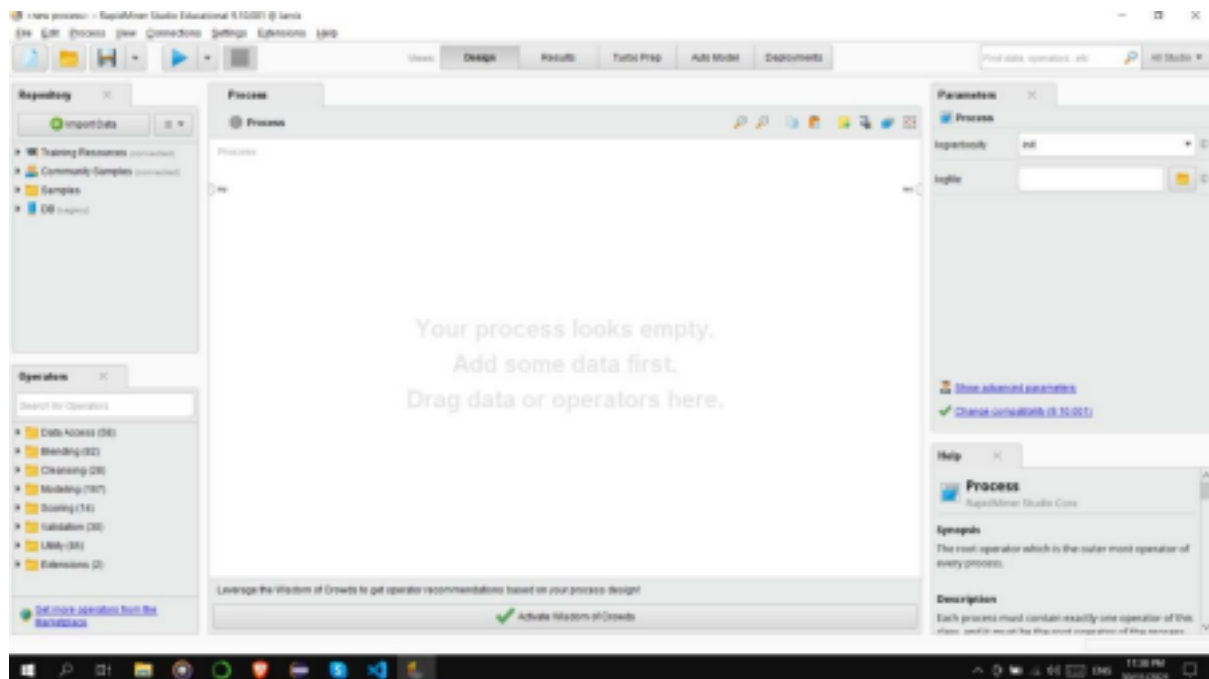
Hierarchies for the dimensions are stored in

the dimensional table. Hierarchies are divided into separate tables.

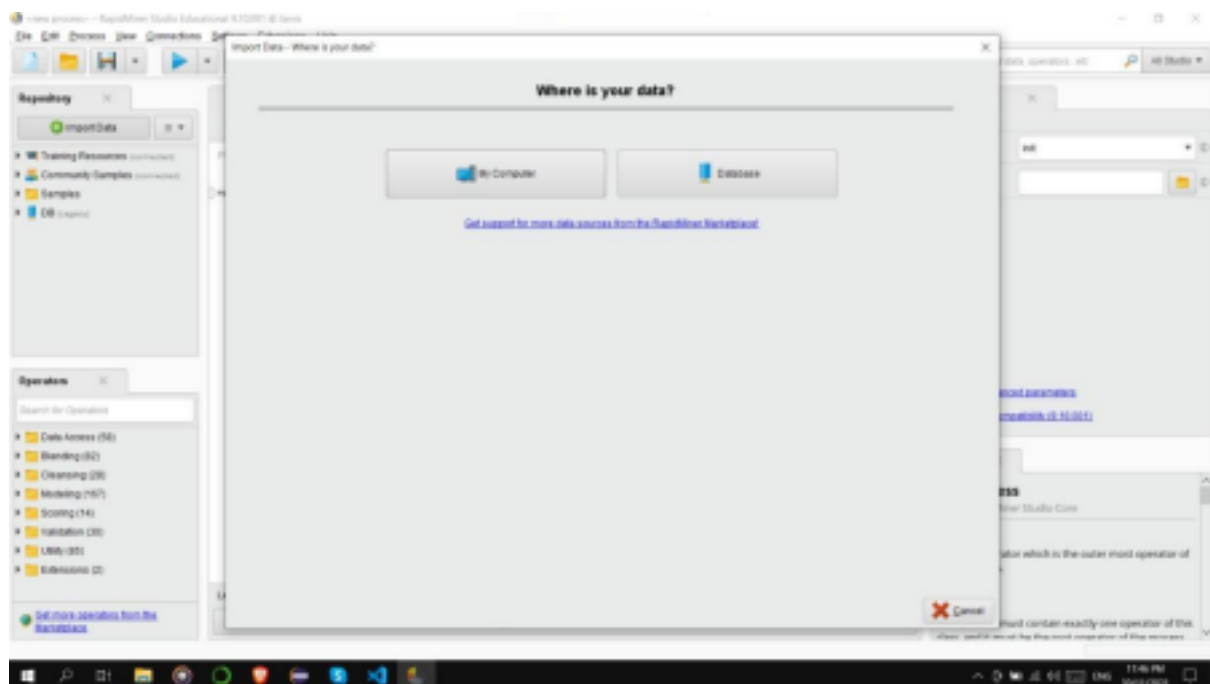
It contains a fact table surrounded by dimension tables.	One fact table surrounded by dimension table which are in turn surrounded by dimension table
--	--

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.

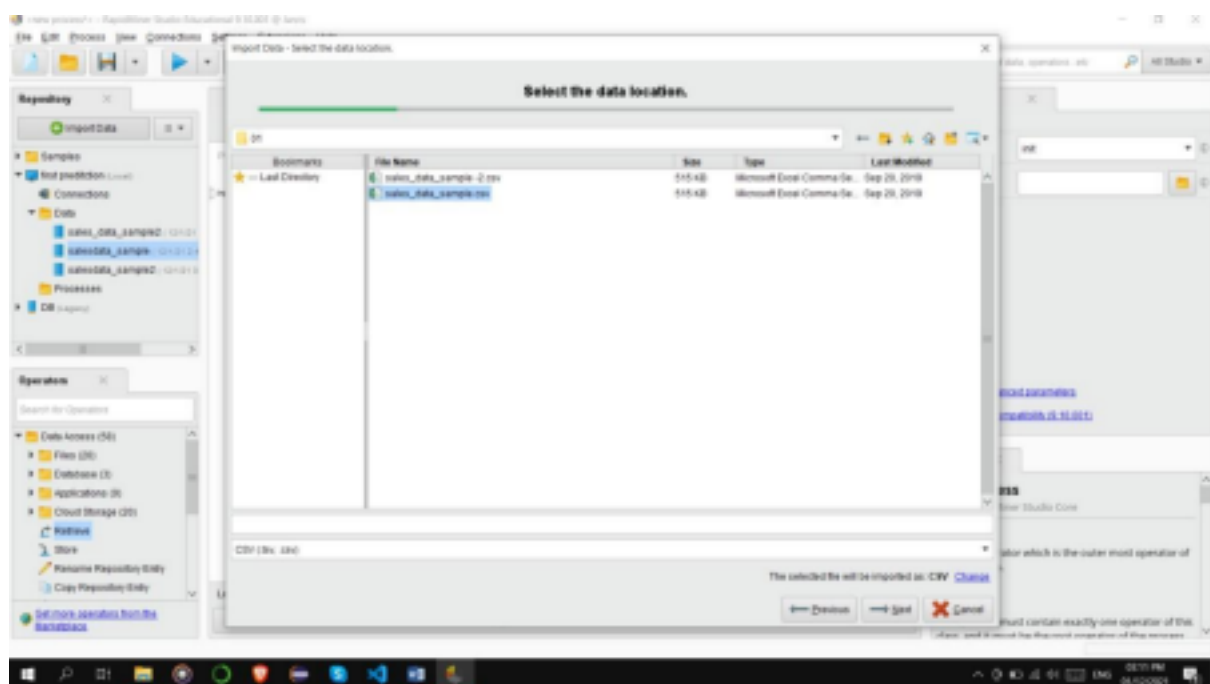
1. Design Model



Step 1 - Import Data from Source



Step 2 - Select data Location



Step 3 - Open Dataset regarding business

Row No.	ORDERNO.	QUANTITY	PRICEACH	ORDERNO.	SALES	ORDERDATE	STATUS	GER_ID	MONTH_ID	YEAR_ID
1	10107	30	81.700	2	2074	Feb 24, 2003	Shipped	1	2	2000
2	10121	36	81.200	6	2769.800	May 7, 2003	Shipped	2	5	2000
3	10138	41	84.740	2	3884.240	Jul 1, 2000	Shipped	3	7	2000
4	10145	45	83.200	6	3746.700	Aug 25, 2003	Shipped	3	8	2000
5	10159	49	100	14	5295.270	Oct 10, 2003	Shipped	4	10	2000
6	10168	36	86.600	1	3479.760	Oct 28, 2003	Shipped	4	10	2000
7	10180	29	86.130	9	2487.770	Nov 11, 2003	Shipped	4	11	2000
8	10188	48	100	1	5012.320	Nov 18, 2003	Shipped	4	11	2000
9	10201	22	86.570	2	2788.540	Dec 1, 2003	Shipped	4	12	2000
10	10211	41	100	14	4798.440	Jan 15, 2004	Shipped	1	1	2004
11	10225	37	100	1	3685.800	Feb 26, 2004	Shipped	1	2	2004
12	10237	23	100	7	2303.120	Apr 9, 2004	Shipped	2	4	2004
13	10251	28	100	2	3788.640	May 18, 2004	Shipped	2	5	2004
14	10263	34	100	2	3876.760	Jun 28, 2004	Shipped	2	6	2004
15	10275	45	82.830	1	4777.350	Jul 23, 2004	Shipped	3	7	2004
16	10285	36	100	6	4089.960	Aug 27, 2004	Shipped	3	8	2004
17	10296	23	100	9	2587.380	Sep 30, 2004	Shipped	3	9	2004
18	10308	43	100	6	4794.780	Oct 15, 2004	Shipped	4	10	2004

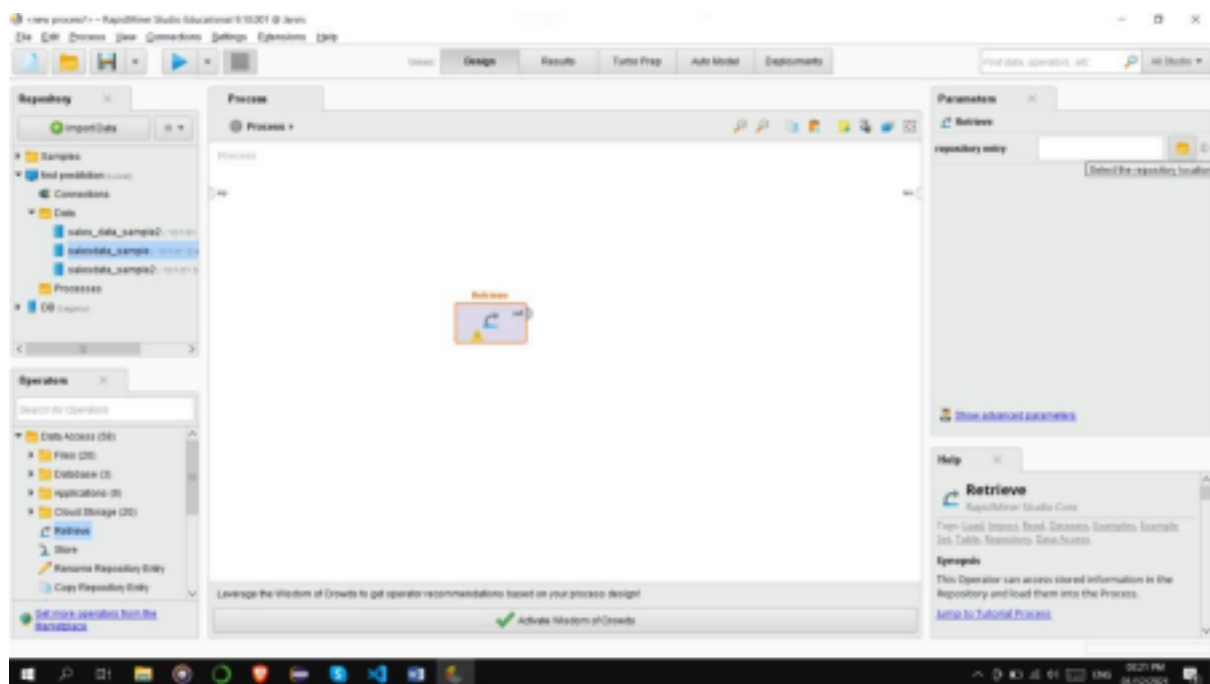
Step 4 - Click on retrieve operator drag in process view,
It has input and out Operator.

Retrieve
Reads an object from the data repository.

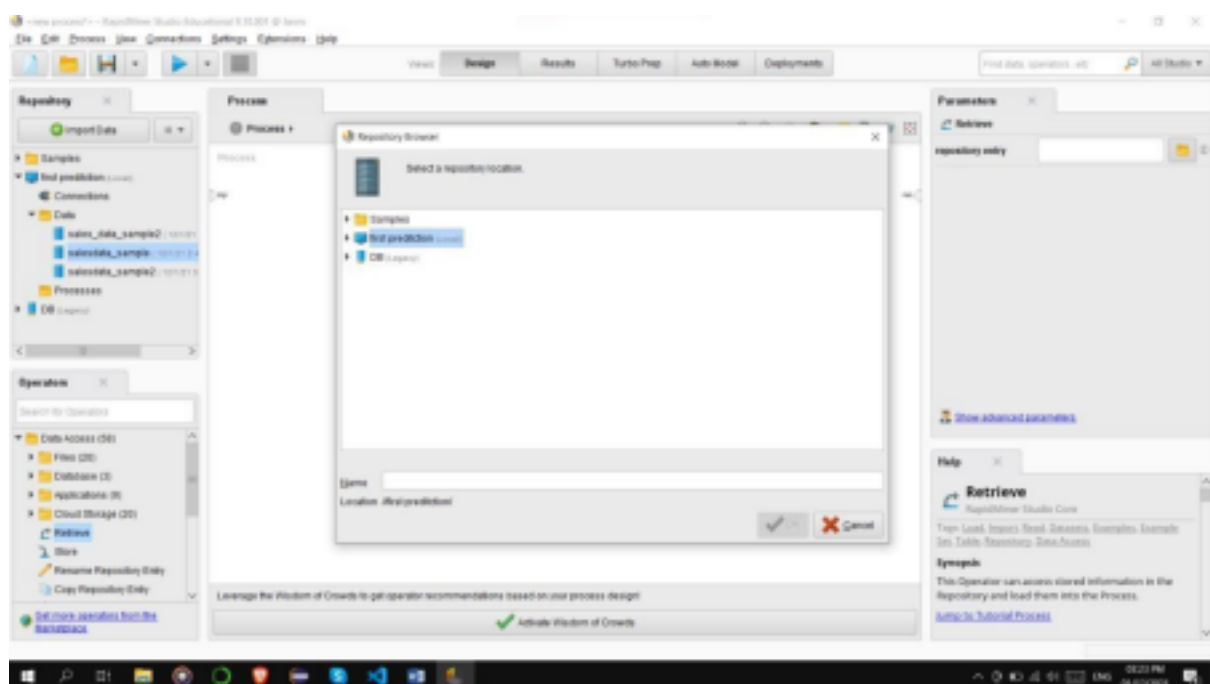
Parameters
Superiority: all
logfile: [empty field]

Help
Process
Rapidminer Studio Core
Synopsis
The most operator which is the outer most operator of every process.
Description
Each process must contain exactly one operator of this class, and it must be the most operator of the process.

Step 5 - Click on Repository Entry



Step 6 - Select Local Repository



Step 7 - Select updated dataset





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.