ISBN: 978-93-88901-23-9

Data Warehousing

ER. KUMARI DEEPIKA
DR. UJWALA SURYAWANSHI
DR. MAHENDRA KONDEKAR
DR. SANTOSH CHOWHAN
DR. SHRUTI BHARADWAJ
ER. JYOTI BHOSALE

First Edition: 2022

Data Warehousing

(ISBN: 978-93-88901-23-9)

Authors

Er. Kumari Deepika
Assistant Professor,
Symbiosis Institute of Computer Studies and Research, Symbiosis International Deemed University, Pune, India

Dr. Santosh Chowhan
Associate Professor,
Department of Data Science and Analytics, School of Sciences, JAIN (Deemed-to-be University), JC Road, Bangalore

Dr. Ujwala Suryawanshi
Assistant Professor,
Rajarshi Shahu Mahavidyalaya,
Latur, Maharashtra, India

Dr. Shruti Bharadwaj
Assistant Professor,
United college of Engineering and Research,
Prayagraj. U. P.

Dr. Mahendra Kondekar
Incharge Principal,
Marathwada Institute of Technology, Cidco,
Aurangabad, Maharashtra, India

Er. Jyoti Bhosale
Assistant Professor,
Vilasrao Deshmukh Foundations Group of Institution, Latur

Bhumi Publishing

2022

First Edition: December, 2022

ISBN: 978-93-88901-23-9

Published by:

Bhumi Publishing,

Nigave Khalasa, Kolhapur 416207, Maharashtra, India

Website: www.bhumipublishing.com

E-mail: bhumipublishing@gmail.com

Book Available online at:

https://www.bhumipublishing.com/book/

PREFACE

This technical book is intended not only for the students but IT professionals also. In this book, the core concepts of data warehousing and its practical approach for implementation have been expalined. In the former sections, it is discussing the need and way of evolution of the Data warehouse system. In later sections, what, why, the components of data warehouse, and the implementation approach of building a data warehouse have been discussed. Authors has also focused on explaining the main objective (to generate pre-defined and ad-hoc reports i.e presentation of data to users) of a data warehouse once it is built. As we worked on performance optimization of ETL Process while pursuing my M.Tech. degree, so through this book we have tried best to provide the insights of the data warehousing. After reading this book, the readers are able to understand theoretical and practical components to design and build a data warehouse.

- Authors

ACKNOWLEDGEMENT

Er. Kumari Deepika is the first author of this book. Deepika did Bachelor of Engineering in Information Technology from Nagpur University. Thereafter she qualified GATE and completed Master of Technology in Computer Science and Technology from Central University of Punjab. During her MTech she worked on Performance Optimization on ETL process as dissertation research project. She has qualified UGC NET twice. Currently she is working as an Assistant Professor in Symbiosis Institute of Computer Studies and Research, Symbiosis International Deemed University, Pune, India. Along with her job, she is also pursuing part time PhD from J C Bose University of Science and Technology.

First of all, I would like to say my very special thanks to my mother and Father who kept supporting me and encouraging me to study and be an achiever in life. A big credit goes to my husband who has strongly supported me to continue my study after marriage and even at the time of pregnancy. He always appreciated my efforts highly, and inspired me to go beyond crossing all the obstacles in the way.

A big thanks to my lifetime three best friends Shailja Pant, Dr. Sarika Sharma and Shruti Bhardwaj who spent their valuable time whenever I needed a friend to talk and discuss the problems. A good friend is someone who truly understands your feeling, and realizes yourself when you lose your momentum. Dr. Sarika Sharma is someone who has played a role of my mentor and guided me at every step of my job.

My beloved friends are also contributed their work in this book,

Dr. Santosh Chowhan

Dr. Ujwala Suryawanshi

Dr. Mahendra Kondekar

Dr. Shruti Bharadwaj

Er. Jyoti Bhosale

Without their efforts and work this book was not completed, special thanks to all of them.

This work dedicated to My Papa, Mummy, Husband, Dr. Sarika Sharma, Shailja Pant and Dr. Shruti Bharadwaj.

- ***Authors***

Time Frame for Curriculum Delivery

Sr. No.	Topic	Hours
1.	Introduction, Data Warehousing, Comparative Study with ER Modeling	3
2.	Data Warehousing Project Management and Requirements Gathering	3
3.	Technical Architecture and Design of Datawarehouse	4
4.	Dimension Modelling and Implementation	5
5.	Extraction, Transformation, Load Understanding, Different Source, Transformations and Targets	6
6.	Understanding Techniques to Improve Query Performance	5
7.	Materialized Views and Partitioning	4
8.	OLAP and Business Intelligence	4
9.	Deployment, Support and Expansion of the Data Warehouse System	5
10.	Project	6

UNIT I: BASICS OF DATA WAREHOUSING

Learning Objectives

After going through this chapter, you will be able to:

- Differentiate between Strategic information and tactical decisions
- Understand the reasons for IT professionals fail to provide useful insight
- Define Data Warehouse
- Understand the need of DWH

1.1 Introduction

Data is the key to standing competitive in the business. It works like fuel that helps a vehicle(business) run. Every organization that runs a business has a huge volume of data which is scattered throughout the different storage systems it has. It means data is available in the system but the challenge is how to use that data to stay competitive along with other competitors. If they use their data wisely to extract information that helps them make strategies and plans to run their business efficiently, they will certainly have the edge over their competitors. Data warehousing is the solution to gather all their scattered data in one centralized system and perform a different kind of analysis to extract useful information.

1.2 How data is different from Information

Data Warehousing (DWH) is the combination of two words "Data" and "Warehouse". So, first of all, we have to understand what is Data. Nowadays dealing with data is a human being's need as it plays a vital role in day-to-day life. Raw and isolated facts about an entity (real-world object) i.e recorded are termed as 'Data' or we can say that unprocessed data as unprocessed means not meaningful (no direct outcome) and Information is the processed data, meaningful outcome extracted from the data. In analogy to this, raw foods or ingredients are considered as data and cooked food is termed information. For example, suppose a teacher delivers his/her classes, from the teacher's perspective it is the information that is gathered by his/her experience of understanding the different resources, and at the same time from the student's perspective it is the data that they are collecting, their brain process that data to make it information. The collection of similar/related data is known as Data Base.

1.3 Relate Data with Database

Table 1: Data and Database

DATA	DATABASE
Audio	Songs.pk
Video	Youtube
Image	Instagram
Map	Google Map

1.4 What is a Database Management System?

Database Management System is the software used to manage the database as it acts as the interface between user and Data Base files. It provides users a platform that is both convenient from their view and efficient from the perspective of a machine, to use for storing data and retrieving data from it. The management of data involves both the structure it provides to the users for the storage of their data and the mechanisms it provides for the manipulation of information.

1.5 Theme for DWH

With the passage of time, the volume of data grows. Since data is scattered throughout the different storage systems within the organizations and proven useless because of not stored in a centralized manner in proper schema and format. Whenever we talk about huge volumes of data (in Terabytes or more) and its variety i.e. data from different sources (either homogeneous or heterogeneous systems) in different formats. To store and process such a high volume of data with varieties, organizations need to shift their focus toward Data Warehousing.

1.6 How Strategic Information is different from tactical Decision

In order to understand the term Data Warehousing, we need to understand the concept of Strategic information. Strategic information is something that helps management executives take decisions to make the organization's functions operate in an efficient way.

Now we will take the example of a hospital management system. In the Hospital Management system, there are different kinds of users. There is a database system used to store data for booking an appointment with a doctor. In it, some patients book a new appointment that is entered into the database or some modify the booked appointment that is updated into the system. Thus, the overall consistency of the database is maintained with respect to time. The above activities are some of the operational aspects i.e. day-to-day operations or transactions of the hospital. The users who are using this database are involved in the operational aspects of the system. These users deal with actions like a request for an appointment, or cancellation of the appointment e.t.c. Such operations come under tactical decisions. The tactical decision is something that is planned to achieve a particular task. On the other side, there are some users who are dealing with strategic decisions. Strategic decisions are different from tactical decisions qualitatively with respect to nature. Strategic decisions are related to the identification of long-term or overall aims and interests and the means of achieving them. Let us move to other aspects of the database system other than tactical decisions, which are taken for addressing operational issues of any system. We can understand this from an example of a Hospital Management System.

Suppose an area is cancer prone so the hospital located in that area would require a full-fledged oncology department that can cater to the patient's needs efficiently. To take such decisions come under the strategic decision.

1.7 Need for DWH

Organizations have an ample amount of data but their Information systems are not able to draw useful insights i.e. useful strategic information from this. Since Data is scattered throughout the different storage systems in different formats so it is not suitable for users to perform a different kind of analysis on the whole data set to extract useful insights from them. All these scattered data need to be brought into one storage system with a suitable structure and in a common format so that analysis and calculations can be performed conveniently on them. Data Warehousing technology provides the solution to integrate data in a centralized storage system so that data can be used intelligently to assist the Decision support system in taking strategic decisions. The goal of DWH is to provide the right information to the right people at the right time.

1.8 Why Information Crisis Persist

Information crises persist because of two reasons

- The data in an organization resides in various disparate systems (multiple platforms) in different formats with diverse structures. But for proper decision-making on overall corporate strategies and objectives, we need integrated data from all systems.
- Data needed for making strategic decisions must be available in a common format that enables executives and managers to analyze trends in order to lead their companies in the right direction.
    - For this, they need to review the data from different business viewpoints.
    - The humongous amount of operational systems' data cannot be readily used to spot trends.
    - Operational data is event-driven i.e. record the details of every transaction that happens.
    - This data cannot be used to state the prevailing trend in the market.
    - Operational data cannot be directly used for reviewing data from different angles.

1.9 Reasons for IT Professionals Fail to Provide Useful Insights

However, when the scenario of the usage of operational data changes to that data used in making strategic decisions then the aspects of storage, design, and modeling get changed. As for usage scenario changes, a separate data storage system is required to handle all this. The implementation of another data storage system would require a considerable amount of cost and effort in terms of time, money, and manpower. So once again the question comes if the data usage scenario changes, can't we change the design to handle it instead of going for DWH? Can IT professionals provide solutions by changing the design to extract required data from existing sources to help us get strategic information?

The answer to the above questions is no because it does not look feasible to develop a design and solution to perform that many data integrations from heterogeneous data sources.

IT professionals have put their best efforts to overcome the information crisis but the factors that were responsible for the inability to provide strategic information prior to data warehousing.

- IT received too many ad-hoc requests for a variety of reports but they were not able to generate all such reports to fulfill requests within the assigned period with the limited resources.
- Requests were not only numerous but also kept changing over time.
- Users want more reports subsequently to expand and understand earlier reports
- The users indulged themselves in the spiral of asking for more and more supplementary reports thereby increasing the IT load.
- The users are completely dependent on IT to provide the information, as they could not access the information directly in an interactive manner.

Case Scenario: Renal Transplant Success rate.

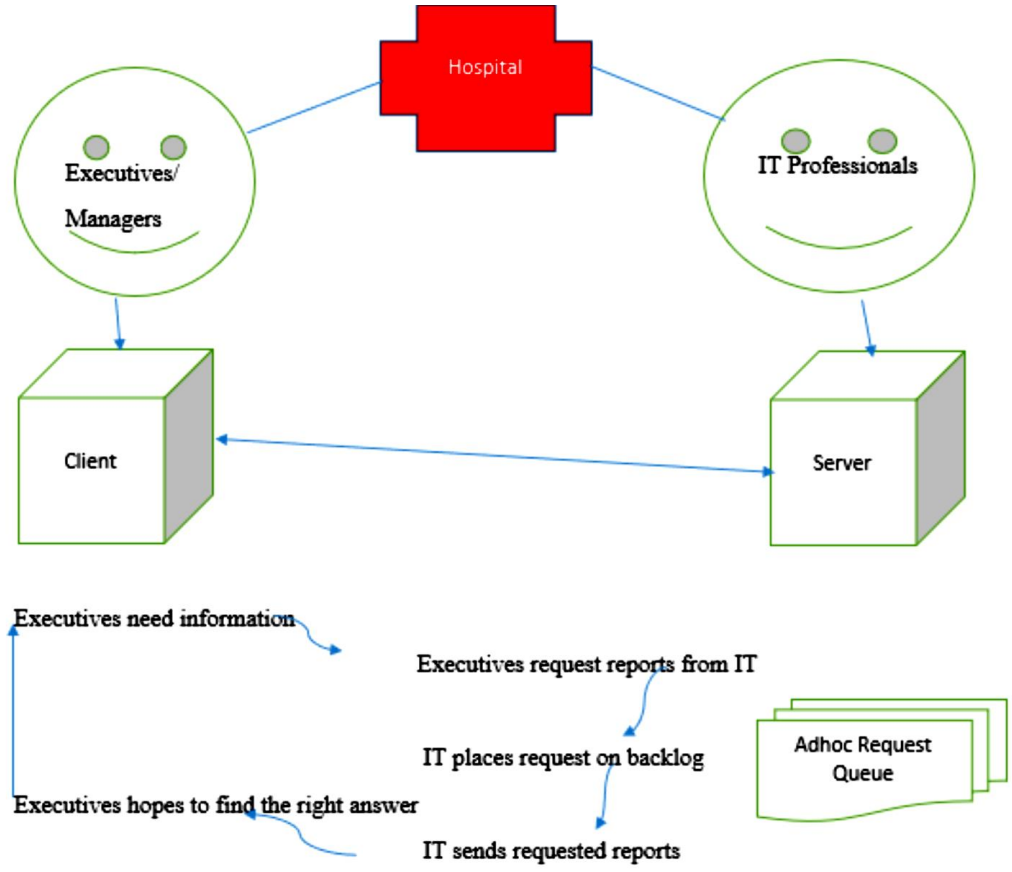


Fig 1.1: The spiral chain of asking reports by executives & managers and providing of information by the IT

In this figure, let's take a scenario from the hospital management system. In the hospital management system suppose we will try to find the strategic information regarding renal transplant success rate.

For this one end, the user is business executives/ managers who act as clients, and on the other side users are IT professionals who act as a server. Executives need strategic information for the analysis that's why they request reports from IT. For processing one request IT professionals apply some aggregation operation in this scenario, IT professionals look up and aggregate donor medical history, patient hereditary case, and so on. For one information executives request subsequent requests. These requests are ad-hoc. Due to an overload of ad-hoc requests, and a lack of resources in terms of limited timeframe and human resources. It places requests on backlog and maintains an ad-hoc request queue. IT processed from the queue and send requested reports to the executives. The spiral of asking for reports by executives and managers and providing of information by IT.

1.10 Separate DWH Environment is Business Need

As we saw in the last section that IT was unable to cater to the reporting requirements of the business with the current data setup. Organizations need to set up another separate environment for flexible and conclusive analysis of data so that managers and executives

Would be able to make strategic decisions efficiently. For the above-stated reasons, a new type of system environment comes into the picture i.e. Data Warehousing, for providing data for analysis and helping in strategic decisions, discerning (discrete) trends, and monitoring performance.

As we earlier discussed Data Warehousing is the combination of two words 'Data' and 'Warehousing'. The operational data needs to be stored in a separate database system in a common format so that useful insights i.e. strategic information can be drawn from data to help decision-makers take strategic decisions. Data Warehouse would be a separate Database that will support the strategic aspect of the organization. Let's discuss the word 'Warehouse' in the next section.

1.11 What is Warehouse?

By Warehouse, we just mean the storage where we keep our inventory stocks. We elaborate on the term Warehouse to a little more extent than we can say that warehouse is where stocks from several different sources are going to be stored. A large number of stocks are maintained in the warehouse. A warehouse is needed for strategic importance. A warehouse is just an alternate term for inventory management. Inventory management is part of supply chain management. Inventory management deals with both the raw products and the desired or final product. It's not like that warehouse is just dumping or storing areas for goods or products. Suppose there is a pharmaceutical company named Moon Pharma that are having clients all over India. The location of the warehouse is of prime importance depending upon several factors like, what is the cost of transportation, logistic coordination, and so on. This cost counts as the overhead. An analogy, suppose there is a hospital named Mercury Health having branches or centers across the country. Different branches maintain their records in different formats. If

records are not integrated and maintained in a common format then it is hard to retrieve information. If it is not maintained properly then data may be ambiguous and inconsistent. So retrieving the information from them may be overhead on the cost in terms of time and space.

1.12 Define Data Warehouse

Data Warehouse is a kind of warehouse of data elements that are captured from different operational data sources. The whole set of data elements is of strategic importance.

Strategic decisions are based on the data elements that have been gathered from distinct sources.

Let us Sum Up

In this chapter, we have learned about the basics of Data Warehousing. Data is the oil that businesses need to run their vehicle smoothly. To be standing competitive in the market, the available data needs to be analyzed properly to extract information that can help the organizations make strategic decisions. Scattered data across multiple data storage systems in different formats doesn't help so the need for a Data warehouse becomes indispensable for the business. We also discussed the concept of warehouse and inventory management. The goal of DWH is to provide the right information to the right people at the right time so that the right decisions can be taken in order to lead the business in the right direction.

Unit End Questions

Q1. What do you understand by Strategic Information? Give suitable examples. Also, write down some of the characteristics of Strategic Information. For a hospital, name five types of strategic objectives.

Q2. Explain the term Information Crisis.

Q3. As you have seen, a pharmacy collects huge amounts of data through its operational systems. Name any four types of transaction data that are likely to be collected by the pharmacy through its daily operations.

Q4. Differentiate between operational Systems and informational systems

Q5. Give reasons why operational systems are not useful for making strategic decisions.

Q6. Explain the factors which lead to the growth and usage of data warehouses

Q7. Data Warehousing is the only viable means to resolve the information crisis and to provide strategic information. Justify the statement.

UNIT II: OLTP AND OLAP

Learning Objectives

After going through this chapter, you will be able to:

- Differentiate between OLTP and OLAP
- Define the term Granularity
- Define Data Warehouse
- Understand the features of DWH
- Understand the different approaches for the development of DWH

2.1 Introduction

OLTP and OLAP are the two data storage systems that cater to the data needs of the various kinds of users dealing with that organization. Every organization has to deal with the two aspects of business i.e. tactical and strategic aspects. OLTP systems help organizations to deal with tactical aspects whereas OLAP systems help them to deal with strategic aspects. In this chapter, you will study the difference between these two storage systems OLTP and OLAP, about the data granularity, and the definition and features of the Data Warehouse. You will also get to know about the different approaches used for the development of the Data Warehouse.

2.2 Compare and Contrast Between OLTP and OLAP

In the last section, we have studied what is DWH, and why we need DWH. In this section, we will discuss Online Analytical Processing (OLAP) and how it is different from Online Transactional Processing (OLTP) with respect to different parameters. OLAP is the alternate term for DWH. The two terms OLAP and DWH are interchangeable throughout this book.

Let's move on to compare and contrast OLTP and OLAP through scenarios of the hospital management system. Operational Data that is discussed in the last section is the data that is involved in the operation of a particular system.

For example, Patient 'X' (having patient id 'P101') admitted to Ward 'Z' (having ward id 'W11') pays the bill for test 'Y' (having test id 'T0876') prescribed by the doctor 'W' (having doctor id 'D483').

We see in this case data about patient id, patient name, doctor id, doctor name, ward id, ward name, test id, test name, the charge of the test so on. All these data elements are all operational data elements for performing the operation of paying the bill for the 'X' test by the patient named 'Y'. Operational data are subjected to frequent updates and queries. Since frequent insert and update operations are performed on data of OLTP, it needs to be normalized. Normalized data means the removal of data redundancy and prevention of different anomalies like Insert, update, and delete anomalies. In order to normalize the data, a big table is decomposed into many smaller tables. The tables in the OLTP system need to be normalized because the insert/update

operations performed on them are very frequent in nature such as 50 k operations a day. All these operations should be optimized so that it produces a satisfactory response time. For example, suppose a patient pays his bill at the bill desk, the data entry operations should be completed within fractions of seconds. Otherwise, the user wouldn't be happy using this system.

The queries used to manipulate, store, and fetch data from operational systems are point queries.

A point query is termed a query executed on an individual tuple.

For Example, what is the consultation charge for Doctor 'W'?

What is the charge of the test 'X"?

There is a particular point or a specific tuple that is targeted using the record identifier as the key which is used to access all relevant information required for these operations.

On the other side, OLAP concentrated on a group of data for the analytical queries as data that are archived for a long period of time over different operational data sets that tell something about the overall trends of operation that have been happening. We can understand this more clearly with this example, suppose data were collected from all branches of the hospital for the past ten years to find out the trend for emergency time, and emergency cases. Which location is suitable for the new branch of the hospital etc.

The result set of the above-mentioned queries helps in taking strategic decisions. Such queries are termed analytical queries as they are having analytical nature rather than operational or transactional nature.

The data used for analytical queries are subjected to infrequent modifications. In order to perform analysis on OLAP data requires a huge amount of data and aggregation of those data.

Computation of the ideal age of donor for successful renal transplant case. For this analysis, a huge amount of aggregation over a large data set based on all cases of renal transplant needs to be Integrated.

Hence the performance issues come in the OLAP system mainly when the complex query is executed.

On the other hand, performance issues arise in OLTP systems when data manipulation operations are performed. The performance metric for the OLTP system is the transaction throughput (number of transactions per unit time) whereas that in the OLAP system is the query throughput (number of queries executed per unit time)

DWH provides an interactive response system to its users that's why it is also called an Online Analytical Processing System.

A DWH is a platform that provides a reliable and flexible environment to manage such kinds of historical data.

Table 2.1: summarize the difference between OLTP and OLAP

Parameters	OLTP	OLAP
Process	It is an Online Transactional System. It manages database modification.	OLAP is an Online Analysis and data retrieving process.
What the Data	Reveals a snapshot of ongoing business processes.	Multi-dimensional views of various kinds of business activities.
Purpose of Data	Used to run business	Used to analyze business
Source of Data	Operational Data; OLTPs are the original source of the data.	Consolidation data; OLAP data comes from the various OLTP databases.
Query	Insert, Update and Delete information from the database.	Based on Select commands to aggregate data for reporting.
Processing Speed	Typically, very fast	Depending on the amount of data involved batch data refreshes and complex queries may take many hours.
Space Requirements	Can be relatively small if historical data is archived	Larger due to the existence of aggregation structures and history data; requires more indexes than OLTP.
Response	Its response time is in milliseconds.	Response time in seconds to minutes.
Performance Metric	Transaction throughput is the performance metric.	Query throughput is the performance metric.
Database Design	ER modeling	Schema-Star Schema and Snowflake Schema
Orientation	Application-oriented	Subject Oriented
Data Integrity	OLTP database must maintain data integrity constraints.	OLAP database does not get frequently modified. Hence, data integrity is not an issue.
Target Audience/Spectators	It is a market-oriented process.	It is a customer-oriented process.
Productivity	It helps to increase users' self-service and productivity.	Help to increase the productivity of the business analysts.

2.3 Define DWH

A well-known American scientist named William. H.(Bill) Inmon is recognized as the father of Data Warehousing. He defined the term Data Warehousing as

"A DWH is a subject-oriented, integrated, non-volatile and time-variant collection of data in support of management's decisions."

- Subject-Oriented: A data warehouse is used to store and analyze data of a particular subject area(line of business).

  e.g., "Claim" can be a subject area in the health care domain. It will handle the operation of claim processing of the patient.

  It can be understood by switching to the scenario of health care. Let's assume That patient "X" has medical insurance from "Y" for the service "Z" from '1st Jan 2019-31st Dec 2019'.the above said is the overall claim processing operation of the patient. "claim" can be the business subject area to handle the above said claim processing operation. From the above scenario, we can conclude that OLTP is application-oriented or application-centric as it is used to run business, on the other part OLAP is subject-oriented as it is used to analyze the business.
- Integrated data: A data warehouse integrates data from multiple data sources.

  e.g. Two sources A and B may have different ways of identifying a patient, but in a data warehouse, patient data will be stored in one standard format.

  There are two sources, in source A, the format for naming the patient is the first name, and the last name while in source B, the format for naming the patient is the full name. Both sources identify the same patient in different ways However in DWH, patient data will be stored in one standard format.
- Non-volatile: Once data is in the data warehouse, it will not change. DWH analyzes the business through historical trends. So, historical data in a data warehouse should never be altered. Only Read (Batch Update) or fetch operation applied to DWH, update access is restricted in it. OLAP is used as a decision support system. But update access in OLTP as it is used for operational System Applications.
- Time variant: Historical data is kept in a data warehouse. For example, one can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse. On the contrary side, only current values in OLTP. In short "OLAP is historical snapshots of operational data in DWH. Thus DWH provides the Right information to the Right people at Right Time.

2.4 What is Granularity?

Granularity is defined in terms of the level of detail of data.

We clearly understand the term through some examples. In healthcare, Electronic Health records is a patient-centric digital records. If the record per transaction, then it is in the lowest granularity as it is the finest data. However, the records are aggregated and then moved towards the higher order granularity from the lower granularity. This can be explained through the scenario. Let's assume the charge of a consultant for a patient in a ward in one visit is "X" and suppose a consultant visits twice for that patient in a day so if the charge of that consultant for that patient in that ward is aggregated day-wise. We can take another example if the charge of that

consultant for all the patients in that ward in one day. It is again aggregated day-wise and also accumulated charge of that ward as a consultation charge. It flows from lower order to higher order of granularity.

2.5 Different Approaches for the development of DWH

Bill Inmon practiced the top-down approach from centralized data warehousing to data marts. Another practioner named Ralph Kimball practiced the bottom-up approach from a number of data marts to data warehouses. To understand these approaches and how is it different we have to understand data mart first.

Now switch to the term Data Mart, Data Mart is the subset of Data Warehouse or we can say DWH is the superset of Data Marts. Data Mart is focused on a specific functional unit of a system or we can say it is a business process-centric storage system. In healthcare, finance, human resource, administration, etc are data marts. In a top-down approach i.e from DWH to Data Mart, centralized rules and control are inherently architected as it's not a union of disparate data marts for the single. It takes longer to build DWH with high exposure to the risk of failure as it is the central storage of data about all the contents.

On the other side, the bottom-up approach practiced by Ralph Kimball i.e from data mart to DWH is an inherently incremental approach with a lesser risk of failure as compared to a top-down approach. In this approach, a data mart is built separately one after another towards the DWH but it may lead to the redundancies in data of data marts. Due to redundant data, inconsistencies and incompatibility can also occur.

Oncology, urology, nephrology and so on are the various departments. So there would be separate data mart for each department, and clubbing them to form DWH.

Practically we have to combine both approaches top down as well as bottom up approaches.

Let us Sum Up

In this chapter, we have discussed about the two different data storage systems i.e. OLTP (Online Transactional Processing System) and OLAP (Online Analytical Processing System) that organizations maintain to handle their different kind of data needs. We have also understood the difference between these type of data stores based on different parameters. We then defined the data warehouse and granularity of data. Granularity of data means level of detail of data to be stored in the data warehouse. The Lower level of granularity leads to the higher level of detail and vice versa. We have also the different approaches for data warehouse development i.e. Top-down and Bottom-up approach.

Unit End Questions

Q1. Define Data Warehouse According to Bill Inmon

Q2. List out the differences between OLTP and OLAP?

Q3. What are the advantages and disadvantages of Top-down Approach and Bottom-up Approach?

UNIT III: DATA WAREHOUSING SYSTEM AND COMPONENTS

Learning Objectives

After going through this chapter, you will be able to:

- Identify the components of Data Warehousing
- Understand the properties of DWH architecture
- Identify the Distinct layers of dataflow in DWH
- Understand the DWH Architecture and its types

3.1 Introduction

In this section, we shall discuss the Data Warehousing system, its structure and architecture. Structure is the organization and arrangement of interrelated elements. The first part of this section deals with various components of the DWH. Architecture is the end-to-end process from fetching the data from various source system to the end users for analytics, reporting and querying through DWH. The later part of this section deals with architecture.

3.2 Components of Data Warehousing

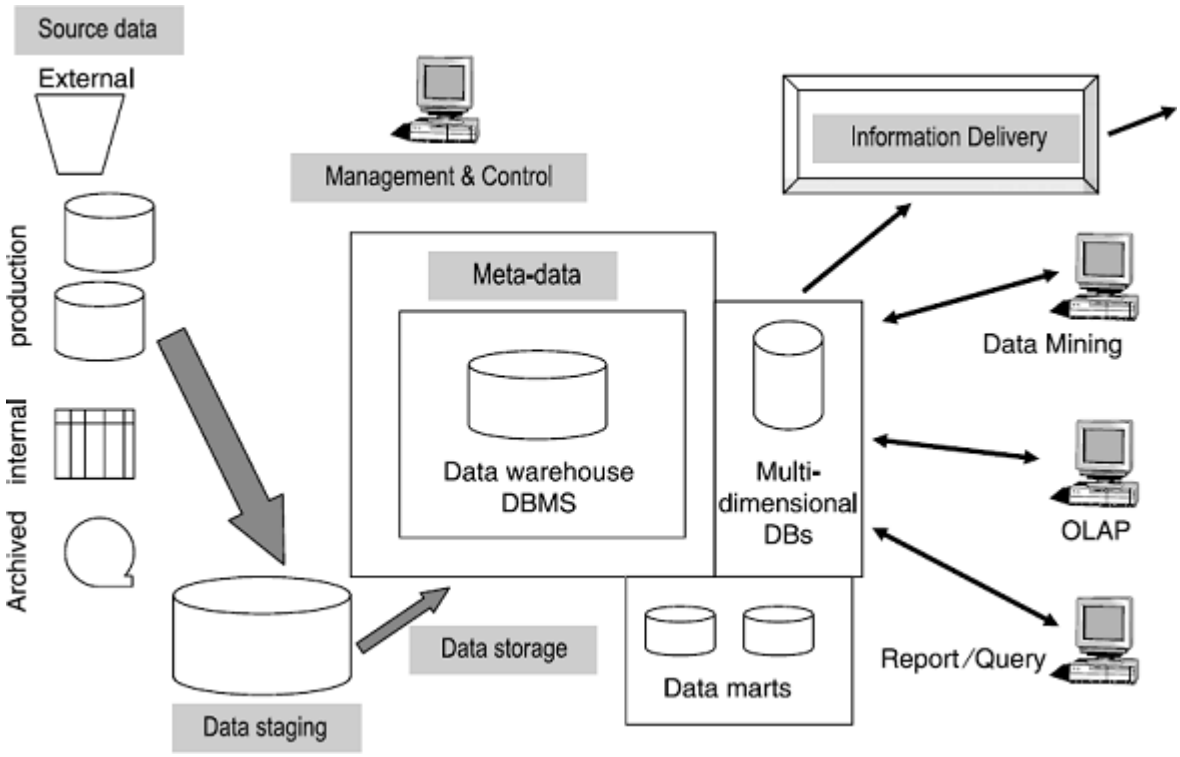


Fig 3.1: Components of Data Warehouse

3.3.1 Source Data Component

There are four broad categories of source data feeding the data warehouse

(a) Production data (b) Internal data

(c) Archived data (d) External data

3.3.1.1 Production Data

- There are various transactional (Operational) systems in place within an enterprise. Production data is generated when various kind of transactions performed within those systems.
- Data is disparate in nature
- It is a challenging task to standardize and transform the data from the various production systems because of disparity in data. Conversion and integration of data to make it to useful data so that it can be stored into Data Warehouse.
- To integrate the data from various sources so that it becomes valuable and meaningful once loaded in DWH.

  Ex- Appointment schedule, registration charge etc.

3.3.1.2 Internal Data

- Some data stored by users in documents, spreadsheets is considered as internal data. Patient profiles and data in departmental databases also fall into this category.
- Such data is not well structured, so it adds additional complexity to the processing of data. It is challenging task to transform and integrate those data so that it can be stored to DWH.
- Scheduling the acquisition of internal data poses some challenges.

  Ex- staff Profiles, Patient profiles or stakeholders' profiles so on

3.3.1.3 Archived Data

- Operational systems must have good response time. They are used to capture the transactions and run the current business.
- It can't hold the data for long period (Usually for more than 2-3 months) so old data is transferred to the backup storage systems.
- It depends on the circumstances in organization that dictate the how often and which portions of the operational databases to be archived for storage.
- Archived data storage is used to get the historical data.
- DWH storage system is used to keep the historical data for a long period of time. Historical data is used in performing analysis over time.
- This type of data is useful for discerning patterns and analyzing trends. Ex-pharmaceutical company-expired medicine –archived

3.3.1.4 External Data

- Not only internal data but external data is also needed to understand the different market trends.

- External agencies as well as national statistical offices working on collection of data provides specific data related to the industry and customers that is used by executives for statistical computations related to their industry.
- Data related to market shares of other competitors is also useful
- To compare the financial indicators of business with the standard values to check the performance status.
- External agencies for example - Indian Health Service (IHS), Centers for Medicare and Medicaid Services (CMMS), Agency for Healthcare Research and Quality (AHRQ)

3.3.2 Staging Area:

Staging Area is a temporary storage area that lies in between the Source systems and Data warehouse. Data is stored temporarily here for further processing. Now the question arises here, why is this temporary storage area is needed? Since business is operated from the various geographic locations so every operational system can't allow to fetch data at the same time. Variation in business cycles and data processing cycles, network resource limitations, and other location or geographical factors, it is not possible to extract data from all the systems exactly at the same time, so a separate temporary storage area is required. That gives the data processing team a flexibility to schedule the fetch of data from different operational systems at different time and store it in the staging area for further processing of data. Data is overwritten every time in that area. A pharmaceutical company named "Sun" having centers throughout the world. They are having different time zone. Suppose a branch in Pittsburgh, PA, USA, comes under Eastern Daylight Time another branch in Pune, MH, India following GMT. Both are having different time zone, so both are having different production times or down time as the difference in their time zone is 10HRS. Pune is 10 HRS ahead of Pittsburg. W/O effecting the production, we can fetch the data in different timings and store in staging area. And also integrated in one standard format [Modelling and Optimization of ETL processes in Data Warehouse by Nitin Anand and Manoj Kumar].

3.3.3 ETL Process:

ETL stands for Extraction, Transformation, and Load process. In order to make this process smooth, an ETL tool is used to perform the processing. ETL tool is usedextract the data from different source systems, transform the data by applying all the necessarylogical operations like arithmetic calculations, Sorting, filtering concatenations, etc. and then load the data into the Data Warehouse system.

ETL Tools: Talend, Clover DX, Informatica Power Center, Microsoft SSIS, Oracle Data Warehouse builder etc.

Transformations: Aggregator, Expression, sorter, filter etc

3.3.4 Separate Storage Component:

Operational systems must have good response time on order to deal with business operations. Data repositories are used to store data of operational systems are normalized (Usually 3 rd normal form), it can contain data for a short span of time i.e. current data for fast and efficient processing. On the other hand, the data repository for DWH contains huge volume of historical data for long period of time for analysis. Since both the systems contradict each other in terms of volume and time period so it becomes necessary to have a separate storage component to store the historical data in different structures suitable for quick analysis.

The data in the operational databases is changeable in nature as it changes from moment to moment. Data that is used for analysis should not be volatile in nature and it should represent snapshots for specified periods.

3.3.5 Information Delivery Components:

Set of tools and technologies that are used to fetch data from data warehouse and provide the report to the users in their desired format. DWH Architecture is the diagrammatic representation of end-to-end flow of data i.e. from source system to the end user for analysis through Data Warehouse.

3.4 Properties of DWH architecture

There are five properties of DWH architecture. They are as follows:

3.4.1 Separation- There should be separation between processing of analytical and that of transactional. If there is no separation between them then there is effect of analytical queries on transactional workloads.

3.4.2 Scalability- Hardware and software Architecture should be scalable for the accommodation/adjustment of the upgradation of the volumes of data which has to be managed and requirements of the number of users.

3.4.3 Extensibility- Without redesigning the whole system there should be extensible to perform new operations by slight modifications.

3.4.4 Security- strategic data is stored in DWH, so data should be secure as these data used for analyzing the business.

3.4.5 Administration- there should not be complex that much to administer is complicated.

3.5 Distinct layers of dataflow in DWH:

Data is flown from various different layers till it reaches to the end user from its origin i.e source. The different layers are mentioned as follows:

3.5.1 Data Source Layer- Transactional data are captured in many different source systems and this layer represents the distinct sources of data

3.5.2 Data Extraction Layer- In this layer data is extracted from data source and that are loaded into the DWH. In this layer, required cleansing operations on the data are done but not major transformations on the data are done here.

3.5.3 Staging Area- It is already discussed prior. It is the area where raw data is stored prior to the data warehouse such that subsequent processing and integration on the data.

3.5.4 Extract Transform Layer (ETL)- In this layer, data is fetched from source systems, required logic is applied to transform the data according to business rule and load the processed data in DWH. This layer is responsible for movement of data from Source to DWH.

3.5.5 Data Storage Layer- In this layer transformed and cleansed data are to be stored.

There are three types of storage acts as entities based on its scope. They are as follows:

(a) Data Warehouse (DWH)- already discussed.

(b) Data Mart (DM)- already discussed.

(c) Operational Data Store (ODS) – it is the storage of data that is used mainly for operational or transactional kind of reporting and it also acts as source for Enterprise Data Warehouse. It generally contains data for relatively smaller time period than DWH. It takes transactional data from various production systems and integrates it loosely. It is subject oriented, integrated and time variant but not non-volatile.

3.5.6 Data Logic Layer- In this layer business rules applied to the data that affects the presentation of the reports but not affected data transformation rules.

3.5.7 Data Presentation Layer- In this layer, information is presented graphically, tabular form to the users through OLAP tools and reporting tools. If jobs are created and scheduled to execute periodically then jobs will run and generate reports that will be sent to the users' mailbox.

3.5.8 Meta Data Layer- Meta data is the information about the data objects that are stored in DWH system such as what are the data objects DWH contains what are the transformation applied to the data objects, number of times ETL jobs executed, sources, access procedures, data staging, users, data mart schema and so on. It is the logical/Conceptual Data model.

3.5.9 System Operations Layer- In this layer information regarding the status of ETL job, performance of the system, history of accessing the system and how DWH operates etc.

3.6 DWH Architecture

There are various architectures of Data Warehouse based on the components and their arrangements within the system.

3.6.1 DWH: Basic

In this architecture, data from various source system is fetched and loaded directly to the Data Warehouse. Distinct information management systems of the organization will take data from the Data Warehouse and use it. So, this kind of architecture is suitable for the business that is not spanned across multiple locations, and allow ETL processes to fetch data at the same time.

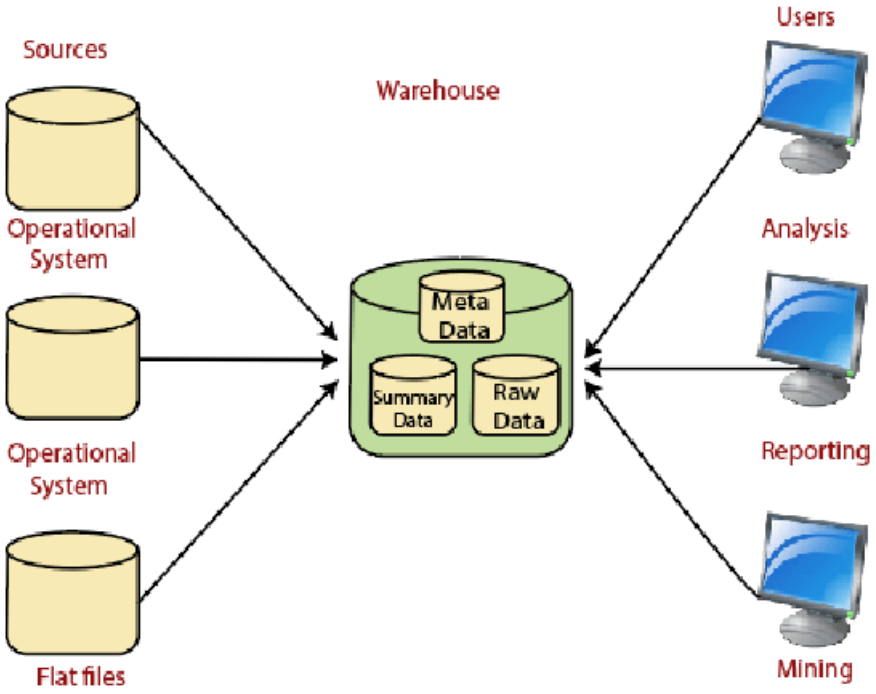


Fig 3.2: Architecture of a Data Warehouse

3.6.2 DWH: with staging area

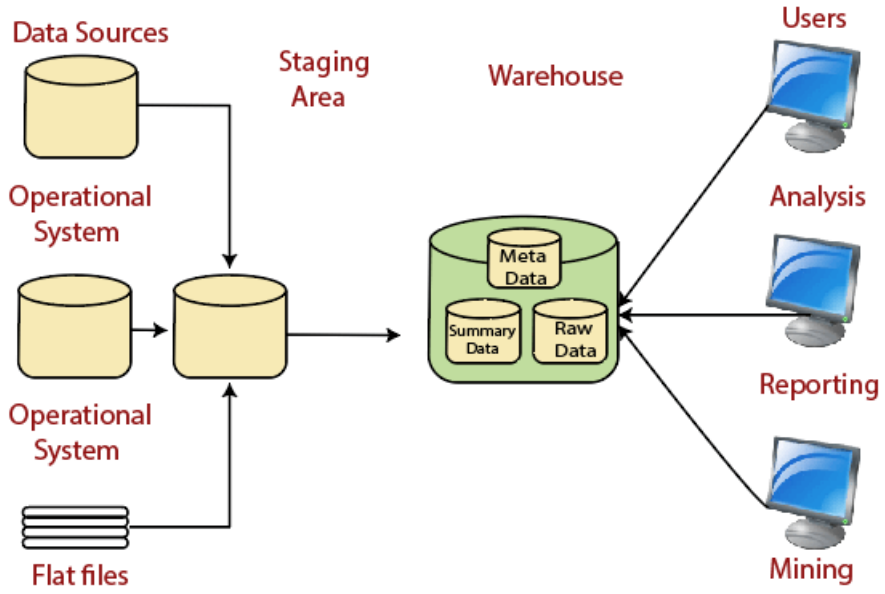


Fig 3.3: Architecture of a Data Warehouse with a Staging Area

In this architecture data is fetched from the different source systems at the time when it allows ETL Processes to fetch data from their storage area, and then all the logical data operations and integrations of Data take place and load the processed data to the Data Warehouse. This type of

architecture is suitable for the organization that have the business spanned across the multiple different geographical locations, and they have the different downtime to allow ETL processes to fetch data from their data storage area.

3.6.3 DWH: with staging area and data marts

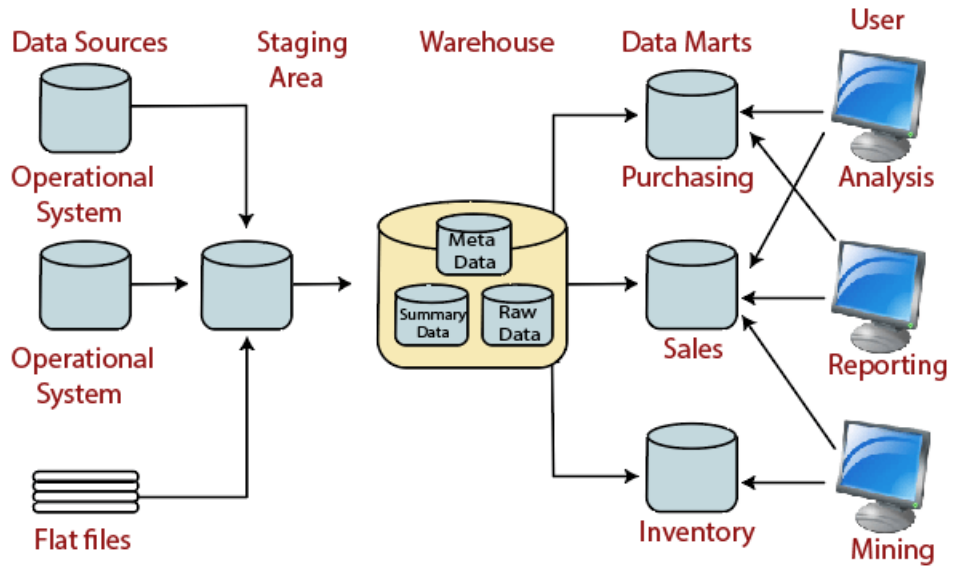


Fig 3.4: Architecture of a Data Warehouse with a Staging Area and Data Marts

In this architecture data is fetched from the different source systems at the time when it allows ETL Processes to fetch data from their storage area, and then all the logical data operations and integrations of Data take place and load the processed data to the Data Warehouse. The different data marts are used to store data for separate business units, and provide data to the corresponding reporting systems to generate the reports.

3.7 Type of DWH architecture

There are three types of DWH architecture and they are as follows:

3.7.1 1-tier architecture-

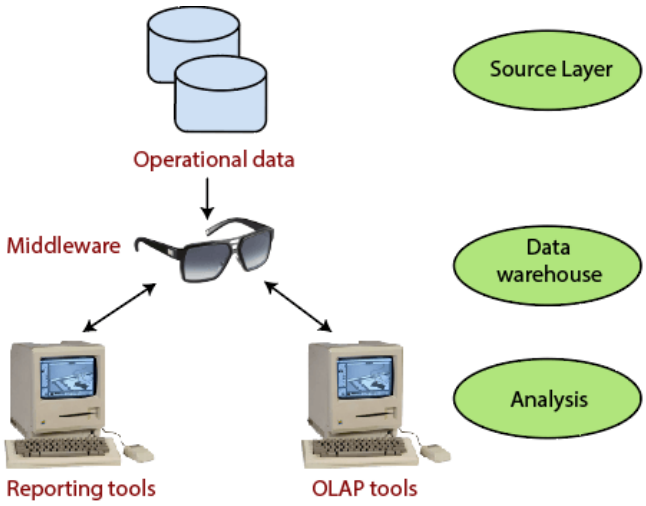


Fig 3.5: Single-Tier Data Warehouse Architecture

In this architecture, only a single layer is present physically i.e source layer so it is also called as single tier architecture. Here, DWH worked as virtually by implementing as the multidimensional view of the transactional data. This architecture lacks in maintaining the separation between transactional queries and analytical queries, and analytical queries increase the loads of transactions. That's why it is not widely used. So, it is effective for organization having need for small volumes of data.

3.7.2 2-tier architecture- In this architecture, separation is maintained between source layer and DWH by four subsequent Data Flow stages. They are as follows:

(a) Source layer (b) Staging layer

(c) DWH layer (d) Analysis

The limitation of this architecture in terms of numbers of users as it lacks in scalability.

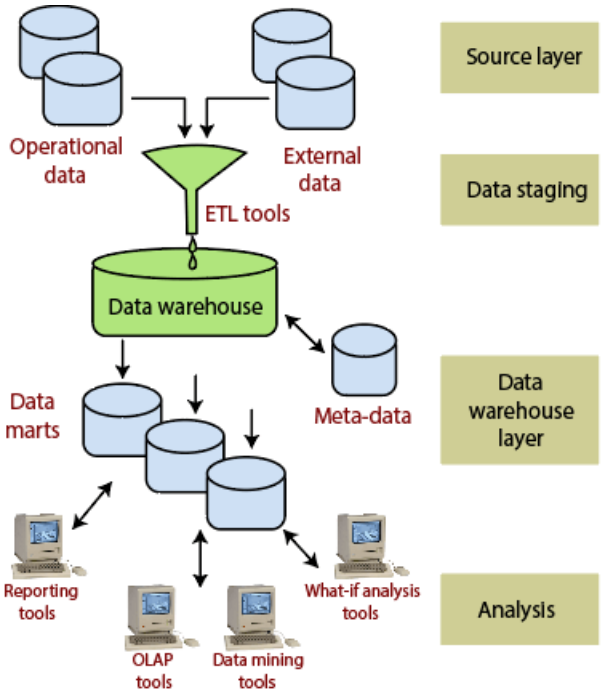


Fig 3.6: Two-Tier Data Warehouse Architecture

3.7.3 3-tier architecture- In this architecture, there is structured flow for data from raw data to the processed data.

There are three tiers in this architecture. They are as follows:

(a) Bottom tier (Database server) - in this tier, data is extracted from transactional database used for end user applications like analysis, reporting, and so on.

(b) Middle tier (OLAP Server) - in this tier, data is transformed for analysis, and querying. It can work on either Relational Online Analytical Processing (ROLAP) or Multidimensional Online Analytical Processing (MOLAP). In ROLAP, there is mapping of operations on multidimensional data

to relational operations. In MOLAP, it directly implements of multidimensional data and operations. (ROLAP & MOLAP will be explained in detail in later section of the book).

(c) Top tier (Client Layer)- In this tier, Client is to be provided access tools for data analysis, querying, reporting, and data mining.

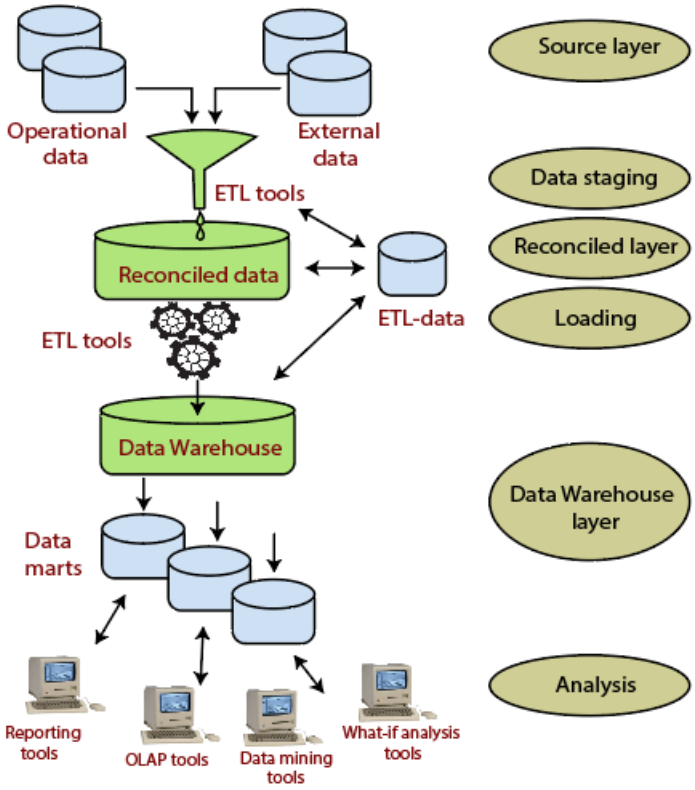


Fig 3.7: Three-Tier Architecture for a Data Warehouse System

Let Us Sum Up

In this chapter, we have studied about end-to-end architecture of the data warehouse. It has various components such as Source systems, staging area, ETL layer, Data warehouse storage area, and Information delivery components. We have also understood the different kinds of source data along with their characteristics, and other components which form the different layers of the data warehouse. We have also discussed the varying architecture i.e. 1-Tier,2-Tier, and 3-Tier Architecture of Data warehouse based on depending upon elements of an organization.

Unit End Questions

1. What are the components of Data Warehousing? Draw a neat diagram and explain them briefly?
2. What are the different types of Source System for designing a Data Warehouse for an organization? Explain them briefly?

UNIT IV: MODELING

Learning Objectives

After going through this chapter, you will be able to:

- Define Data Modeling
- List the roles and responsibilities of Data Modeler in the project
- Define Dimensional Modeling
- Understand different elements of Dimensional Modeling
- Understand the different schemas for designing DWH

4.1 Introduction:

Data model is the abstract representation of design aspect of Data Warehouse. It lays the foundation for the build of structural components of storage of data in Data Warehouse. It does not only specify the storage structures but also specifies the relation among those components. After having good understanding of requirements of an organization's data need, business analysts and data modelers work on the design of model of DWH and a data model is the outcome of that phase. In this chapter, you will study what a data model is, the roles and responsibilities of a data modeler in a project, and the concept of data modeling. You will also understand the different components of dimensional modeling as well as different schemas used for designing a Data Warehouse.

4.2 What is Data Model?

Data Models are the abstraction of the usage and representation of the data through data objects, operations that applied on data objects and integrity rules applied on objects. Data model converts the requirements into visual representation for the flow of business process.

Data Modeling involves

(a) Problem-First and foremost we identify a Problem

(b) Data Requirements-whenever a problem area is identified, the most important step is Requirement Gathering

(c) Conceptual Data model- It includes entities and relationships and does not contain detailed level of information about attributes. At this level, the data modeler attempts to identify the highest-level relationships among the different entities. It is independent of the platform.

(d) Logical Data model- It is based on conceptual data model, includes entities, their attributes and relationships among them. At this level, the Data modeler describes the data in detail without regard to how they will be physically implemented in the database. It is independent of the technology.

In data warehousing, it is common practice for the conceptual data model and the logical data model to be combined into a single step.

(e) Physical Data model- It includes tables, columns, keys, data types, validation rules, database triggers, stored procedures, domains, access constraints and indices for fast data access. At this level, the data modeler will specify how the logical data model will be realized in the database schema. It is platform/technology dependent.

The physical data model should be de-normalized to meet performance requirements based on the nature of the database. If the nature of the database is transactional then it is not de-normalized. De-normalization is common in Data Warehouses.

4.3 Roles and Responsibilities of Data Modeler in the project

Role of data modeler in the project right from the beginning

Data modeler is required in a project right from the beginning. For small projects, a Data modeler plays a dual role of a Data Modeler as well as a Business Analyst. Whereas, in medium to large size projects, the separate business analyst as well as the data modeler should be present. They should both work in closely with each other to avoid any mismatch in expectations.

Presence of Data Modeler for the entire duration

A project needs a modeler throughout the duration of the project. Data Modeler needs to understand the timelines of the entire project. Data Modeling is not a kind of one-time activity for few days but it is an iterative process and goes hand- In-hand with the project lifecycle. Though a major chunk of the Data model is completed in the design phase, it is bound to be changes due to customization needs of the Extract, Transform, and Load (ETL) and reporting layers. The data modeler needs to understand requirements completely from a project end-to-end perspective.

Customers' Involvement in Project

Customers should be involved in project right from beginning. Their involvement from early in the design phase make sure that the data and system requirements are documented and well understood.

Managing the expectation mismatch

Few important guidelines must be followed to avoid expectation mismatch while gathering requirements from the customers. They are as follows.

- Understand the whole system, and then simultaneously concentrate on the portions that require changes or modifications or a new development.
- Ask specific questions during the interviews to help define the requirements.
- Use prescribed six sigma methodology questions like who, what, why, how, where, when, etc.
- Explain the timelines and effort estimations very clearly
- Document the requirements and have them regularly reviewed by the customer.

Consider a requirement from a client organization stating that they need customer data information for business purpose.

By using the six-sigma methodology, we can ask questions like:

(a) Who would be using this customer data?
(b) What kind of customer data will be stored in the warehouse?
(c) Why do they need the customer data?
(d) How will the customer data be located into the data warehouse?
(e) From where does the customer data comes?
(f) When will the customer access the data?

Few more possible questions

(a) What is the purpose of the customer data being added into the warehouse?
(b) Who are all the project stakeholders involved?
(c) What are the Extract, Transform, and Load (ETL) and Reporting strategies?
(d) What are the other system requirements like hardware, software, specific technology etc?

4.4 Define Dimensional Modeling

In the previous section, we already have discussed both OLTP, OLAP system are different, and they are having different approaches for database design. They are different with respect to accessibility, which in terms of visibility of the structure of database from the end-user perspective and in terms of its performance.

In OLTP, database structure is hidden and accessible to the end users in form of data entry and query. However, in OLAP users need to understand how data is structured for effective analysis for them. In addition to the access of the system by whom, also which type of access to the system affects the performance of system is kept in mind while designing the system.

In OLTP, users update data to the database in response to the transaction events of the business. However, in OLAP, it is regarded as read-only database as it is batch updated through ETL process. Therefore, DWH is de-normalized. When database with normalized architecture is used, there is a need to join number of smaller tables when data is retrieved as every join is costly which means analysis (OLAP) query would run for a long time and if every analyst is running such queries, as a data warehouse is an enterprise database, execution of query will suffer with performance bottlenecks.

Limitations in ER Modeling

Entity- Relationship (ER) Modeling is the design technique for OLTP system. There are some limitations in ER Modeling as follows:

- Complex in nature
- Not user friendly by business people
- Does not contain complete history i.e difficult to maintain all the history of the transactions
- Slow query performance
- Difficult to prepare reports and analysis as not optimized for reporting.
- It models business processes and not the information

- It contains the current state of the database and not the history data.

Objectives of Dimensional Modeling

Dimensional Modeling is the technique for designing data warehouse. The objectives of dimensional modeling are

- To produce database structures that are easy for end users to understand and write queries against.
- Optimize query performance (as opposed to update performance)

4.5 Elements of Dimensional Modeling

In OLAP System, Data is stored in two types of tables i.e. Dimension and Fact tables. Dimensions are the business objects, while facts are the business measurements.

4.5.1 Dimension Table: It contains the detailed and descriptive information of the business objects that constitute the business. It is mostly static in nature, and gets changed slowly. E.g. Customer, Product, Location, Doctor, Patient, Ward, Test etc.

4.5.2 Fact Table: It contains the measurements data. Once the business functions, transactions are generated and data related to those transactions are stored to the fact table. It generally contains the Keys of associated Dimensions and mostly the numerical data generated out of transactions. E.g. Suppose a patient registers himself in the Hospital XYZ and Pays RS.1000 as registration fees, then a fact record is generated for that transaction with PatientId, HospitalId, Department_id, and Transaction id as dimension keys and RegistrationFees as measurements.

Types of Dimensions:

4.5.1.1 Conformed Dimension: Conformed dimensions are the dimensions which once built in the model can be reused with multiple times with different fact tables. Dimensions are confirmed when they are the same dimension or when one dimension is strict rollup of another. Same dimensions mean it should have exactly same set of primary keys and have the same number of records.

Strict Rollup: when one dimension is strict roll up of another which means to confirm dimension can be combined into a single logical dimension by creating a union of the attributes. Eg. Patient is the conformed dimension that is shared with multiple facts such as diagnosis fact, billing fact so on.

Advantages of Conformed Dimension:

- It supports incremental development approach.
- Easy and cheap maintenance
- It significantly reduces the complexity of extraction and loading.
- Answers business questions that is related to cross data mart.
- Cross data mart is the data marts that share the conformed dimension.

4.5.1.2 Junk Dimension/Dirty Dimension: junk dimension is simply a dimension that stores the junk attributes. It is a collection of flags, transactional ports and text attributes that are not related to any particular dimension. Eg. Transaction info dimension for pharmacy.

Table 4.1: Transaction info dimension for pharmacy

Transaction info dimension		
Transaction Info Key	**Transaction Type**	**Payment Type**
1	Regular Sale	Cash
2	Regular Sale	Check
3	Regular Sale	Credit
4	Regular Sale	Debit
5	Refund	Cash
6	Refund	Check
7	Refund	Credit
8	Refund	Debit
9	No Sale	Cash
10	No Sale	Check
11	No Sale	Credit
12	No Sale	Debit

It provides structure for related code, indicators and descriptors. It provides simplified design that already has multiple dimensions. It captures the context of specific transaction.

4.5.1.3 Degenerate Dimension: Degenerate Dimension is a generated dimension key in the fact table which doesn't have its own dimension table. It is directly related to an event stored in the fact table but it is not eligible to be stored in the separate dimension table. Eg. Invoice number of the bills generated for the patient. Suppose a patient pays the amount of bills then for each transactions the invoice number is generated to keep track of the transaction.

4.5.1.4 Static Dimension: These are the dimensions whose data are not extracted from the data source but are created in the context of DWH. Stored procedure is used to generate data for these dimensions. Eg-the time dimension-it contains day, week, month, quarter, year, decade etc.

4.5.1.5 Slowly Changing Dimension(SCD)- Some attributes of dimensions are subjected to change over time so the concept of SCD is to deal with the changes in attributes so depending on the business requirements on whether history of changes for a particular attribute have to be preserved or not in DWH. This is called a SCD.

Types of SCD

SCD Type 1- It doesn't maintain any history of the changes. If the changes are made for any attribute then existing records will be updated with modifications.

Record Time: 1st October 2018

OLTP

D_ID	D_Name	Designation	Salary
100	ABC	Apprentice Trainee	45,000

OLAP

D_Key	D_ID	D_Name	Designation	Salary
1	100	ABC	Apprentice Trainee	45,000

Record Time: 1st October 2019

OLTP

D_ID	D_Name	Designation	Salary
100	ABC	Surgeon	75,000

OLAP

D_Key	D_ID	D_Name	Designation	Salary
1	100	ABC	Surgeon	75,000

SCD Type 2- it maintains the complete history of the changes. Changes can be tracked in SCD Type 2 Dimensions because a new record is created for each change.

Record Time: 1st October 2018

OLTP

D_ID	D_Name	Designation	Salary
100	ABC	Apprentice Trainee	45,000

OLAP (based on Status Flag)

D_Key	D_ID	D_Name	Designation	Salary	Status
1	100	ABC	Apprentice Trainee	45,000	1

OLAP (based on Start Time and End Time)

D_Key	D_ID	D_Name	Designation	Salary	Start_Time	End_Time
1	100	ABC	Apprentice Trainee	45,000	1st October 2018	NULL

Record Time: 1st October 2019

OLTP

D_ID	D_Name	Designation	Salary
100	ABC	Surgeon	75,000

OLAP (based on Status Flag)

D_Key	D_ID	D_Name	Designation	Salary	Status
1	100	ABC	Apprentice Trainee	45,000	0
2	100	ABC	Surgeon	75,000	1

OLAP (based on Start Time and End Time)

D_Key	D_ID	D_Name	Designation	Salary	Start_Time	End_Time
1	100	ABC	Apprentice Trainee	45,000	1st October 2018	30th September 2019
2	100	ABC	Surgeon	75,000	1st October 2019	NULL

SCD Type 3- Patial history is maintained in SCD Type 3 Dimension. All the changeable attributes are stored into two columns, one is for previous value and another is for the current value.

Record Time: 1st October 2018

OLTP

D_ID	D_Name	Designation	Salary
100	ABC	Apprentice Trainee	45,000

OLAP

D_Key	D_ID	D_Name	Previous Designation	Current Designation	Previous Salary	Current Salary
1	100	ABC	NULL	Apprentice Trainee	NULL	45,000

Record Time: 1st October 2019

OLTP

D_ID	D_Name	Designation	Salary
100	ABC	Surgeon	75,000

LAP

D_Key	D_ID	D_Name	Previous Designation	Current Designation	Previous Salary	Current Salary
1	100	ABC	Apprentice Trainee	Surgeon	45,000	75,000

Synthetic Keys/Surrogate Keys:

It is a key that is used as a primary key in the dimension table in data warehouse. It doesn't come from source but is generated using a sequence generator either in Extract Transform Load (ETL) process or the script used to load data into dimension table. such keys are needed because the organization don't want to expose the natural key outside and for SCD type 2 dimension, the record for one particular dimension entity is repeated for any change, and this case the natural key doesn't solve the purpose of uniqueness. So a new key is required for uniqueness and surrogate key is generated. Eg.

D_Surrogate_Key	D_ID	D_Name	Designation	Salary	Start_Time	End_Time
1	100	ABC	Apprentice Trainee	45,000	1st October 2018	30th September 2019
2	100	ABC	Surgeon	75,000	1st October 2019	NULL

Types of Facts

4.5.1.6 Fully Additive Facts- A fact is said to be fully additive fact if it could be summed up across all the dimensions.

Eg. In pharmaceutical company Sales amount of a medicine is an additive fact, because it could be summed up along all the dimensions present in the fact table; time, store, product. Sales amount for all 7 days in a week represent the total sales amount for that week.

4.5.1.7 Semi Additive Facts- A fact is said to be semi additive fact if it could be summed up most of the dimension but not all.

Eg suppose a patient gets admitted in a hospital named Chandryan for two months and at the end of the day the total amount due on him for the day is recorded in patient bill table. Here total amount due is semi additive fact measure.

4.5.1.8 Non-Additive Facts-A fact is said to be non-additive if it couldn't be summed up any of the dimension. All ratios are non-additive.

Eg suppose a pharmaceutical company 'SUN" manufactures ten types of drugs and the profit margin for each of the drugs is a non-additive fact.

4.5.1.9 Factless Fact Table-fact less fact table is the table that contains only dimensional keys, not contains facts as it captures events for the information rather than computation. Fact table captures the transactions happened but there are certain entities for which there is no event occurs i.e no transactions so these dimension entities are not captured in fact table. So negative reporting is not possible from the fact table. To perform the negative reporting the fact less fact table is required.

Eg Suppose a medicine is not sold throughout the year in a location, only looking at fact table we can't find which medicine is not sold in which location because fact table has not had this record as this med is not sold in that location that's why we needed a factless fact table that identifies these scenarios and capturing such scenarios would help the concerning team focus on sell of the particular product by making some business strategy such as giving some discount on the product etc.

4.5.1.10 Multiple Facts for Single Process

It is worth to note that there is more than one fact table exists for one process. Number of fact tables depend on different lines of operations being performed under the process.

Eg Let's take a scenario of online pharmacy name "JIO". A customer named "POPE" placed an order (Order id 100) for some medicines for hypertension (Amlopres AT, Losar H, Losartan 50 mg). For this process, two facts are generated. One for sold and another for shipment as both activities having different timestamp. Hence measures for both can't be stored in the same fact table. When medicine is sold at the same time they were not shipped, so for the shipping timestamp we can't enter value NULL in the fact table.

Thus multiple fact tables can generate for the single process.

4.6 Different schemas for designing DWH

4.6.1 Star Schema: Star Schema is the schema that consists of fact, and dimension tables and relationships among them. It looks like a star in which fact is placed at the center, and surrounded by the dimensions. The primary-foreign key relationship from dimension to fact is shown by a link. Referential integrity constraint is maintained between fact and dimensions in which the primary key belongs to dimension table and foreign key from a fact table. Combination of all these foreign keys become the primary key of the fact. Dimension table provide the basis for analyzing data in the fact table. Dimension answer "who", "what", "when", "where", "how" and "why" questions about the business events stored in the fact table.

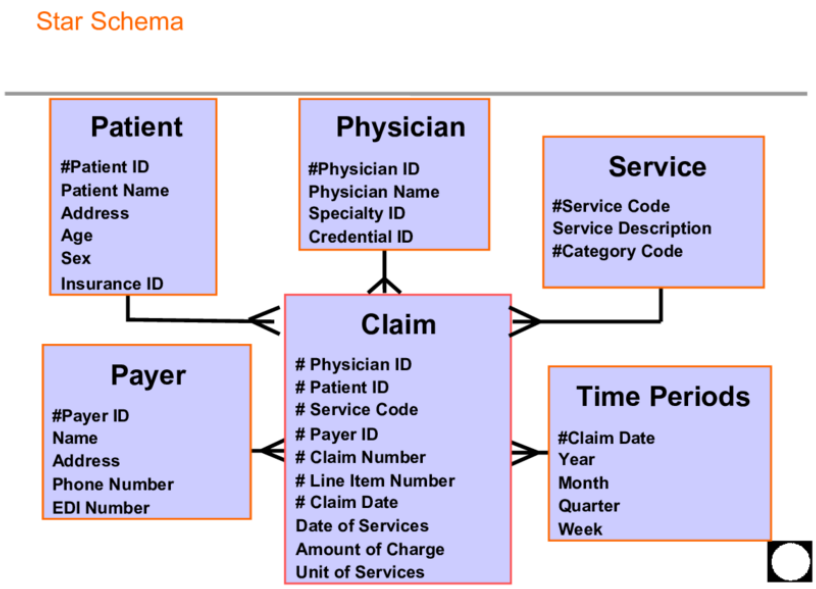


Fig 4.1: Star Schema

Patient Dimension (Who) Physician Dimension (Where)

Time Periods Dimension (When) Service Dimension (What)

Claim Fact

4.6.2 Snowflake Schema: A snowflake schema has the same kind of relationship between fact and dimensions as star schema. The only difference here is that the dimension tables are fully normalized. It is named snowflake because it looks like a snowflake because of decomposition of one de-normalized dimension into many normalized dimensions

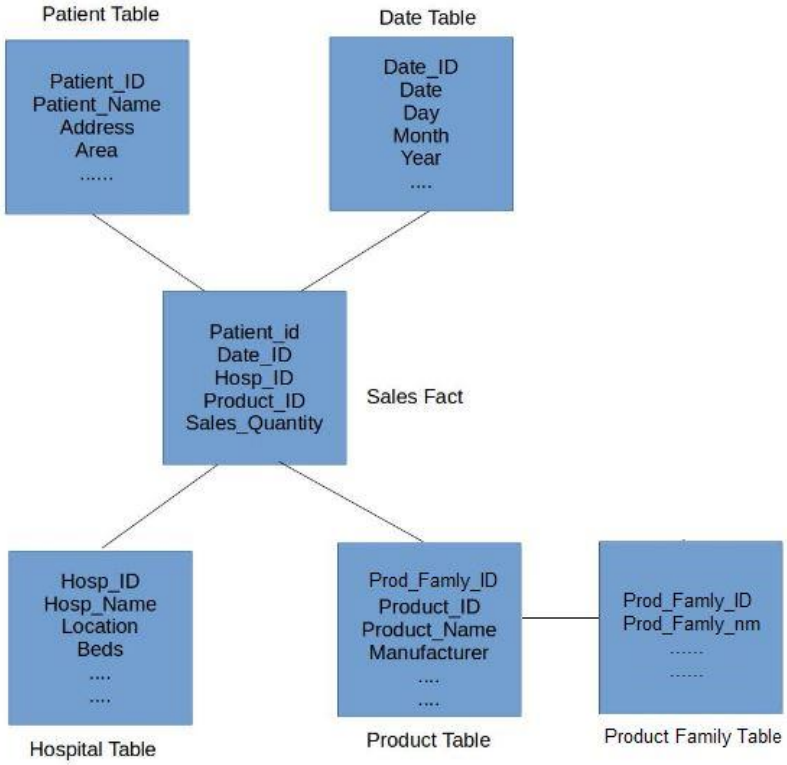


Fig 4.2: Snowflake Schema

4.6.3 Galaxy Schema/Fact Constellation Schema: A galaxy schema is the collection of multiple star schemas in which multiple facts are connected to their respective dimensions. Since it is a collection of star schemas so looks like a galaxy that's why it is called galaxy schema.

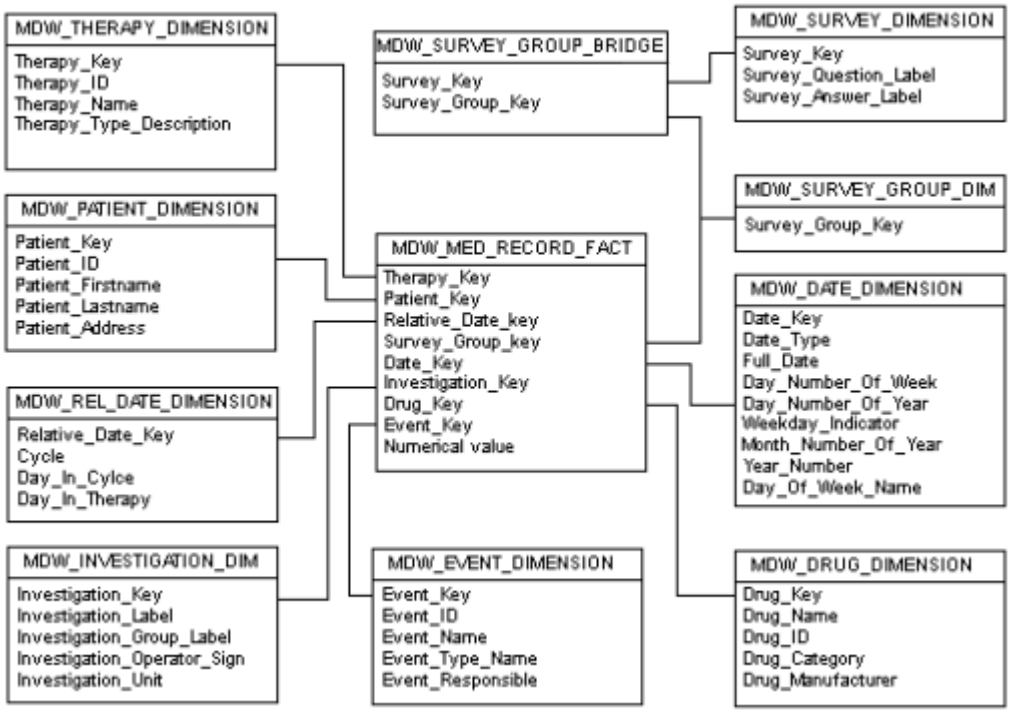


Fig 4.3: Galaxy Schema/Fact constellation

4.7 Bus Matrix and Information Package Diagram

Bus Matrix is the representation of facts and dimension in matrix form i.e 2-Dimensional Array.

	Dimension 1	Dimension 2	Dimension …	Dimension N
Fact 1	X			X
Fact 2	X	X		
Fact….				
Fact N		X		X

Fig 4.4 Bus Matrix

In this fig, Dimensions are the columns and Facts are the rows. Dimension i is the part of Fact i then it is marked in the intersection cell of Dimension i and Fact i. Dimension 1 is the part of Fact 1 and Fact 2. Dimension 2 is the part of Fact 2 and Fact N.

	Date	Patient	Physician	Employee	Facility	Diagnosis	Procedure	Payer
Clinical Events								
Patient Encounter Workflow	X	X	X	X	X	X		
Procedures	X	X	X	X	X	X	X	
Physician Orders	X	X	X		X	X		
Medications	X	X	X			X		
Lab Test Results	X	X	X	X	X	X	X	
Disease/Case Management Participation	X	X	X	X	X	X		
Patient Reported Outcomes	X	X	X		X	X	X	
Patient Satisfaction Surveys	X	X	X		X	X	X	
Billing/Revenue Events								
Inpatient Facility Charges	X	X	X		X	X	X	
Outpatient Professional Charges	X	X	X		X	X	X	
Claims Billing	X	X	X		X	X	X	X
Claims Payments	X	X	X		X	X	X	X
Collections and Write-Offs	X	X	X	X	X	X	X	X

Fig 4.5: Healthcare organization's bus matrix as provided by Kimball
(Ralph Kimball & Margy Ross; The Data Warehouse Toolkit;
Third Edition Wiley Publishers 2013)

4.8 Aggregate table

Aggregate table is used to store the aggregate value of fact measures calculated on the basis of one or more dimensions. It is also called summary tables. It helps in improving the performance of aggregate queries. So instead of calculating aggregated fact measures at run time, it can be fetched directly from the aggregated fact table, store once and used whenever required. It is restricted to additive facts only.

Eg. Summary of sales of drugs by region, by product, by category.

We can understand it more clearly through a scenario. Let's suppose in healthcare data warehouse has the following characteristics-

Billing Fact Table-Patient_ID, Day_Id, Ward_Id, Treatment_id, Transaction_Id, Dialysis_Id, Consultation_Id, Bill_amount

A nephrology ward having kidney dialysis patients were admitted. Each patient goes for dialysis every day. Finance department wants to know the bill amount received from the patients in that ward in a month.

Number of days: 30

Number of consultation visit per day: 2

Number of patients in a nephrology ward: 30

Dialysis per patient per day

Dialysis charge: Rs 3,000

Consultation charge: Rs 1,000

Approximate number of records=

30 patients * 30 days * 1 dialysis per day per patient * 2 consultation visit

=1800 records

Bill Amount= 1800 records * (1000 (consultation charge) + 3000 (dialysis charge))

=1800 records * 4000 Charge

==72,00,000

Now, we can see how a query is generated in normal circumstance and using an aggregate. Aggregation of data is done for a month.

Normal

Here, the financiers will issue a query similar to:

SELECT SUM(BILLING_AMOUNT)

FROM BILLING_FACT

WHERE WARD_ID=10;

[WARD_ID=10: NEPHROLOGY WARD]

Assuming that the DW scans 10 records per second, the approximate time to complete the query will be

Query time = number of records/scan rate

= 1800 records /10 records per second

=180 seconds

Aggregate

An aggregate table will be built, which summarizes the billing_fact by patient.

The aggregate may be defined as follows :

Create table billing_fact_aggregate as

Select Patient_Id, sum(billing_amount) as billing_amount

From billing_fact

Group by patient_id;

The size of the new table will be only 30 records (Number of patients=30)

The SQL needed to get the answer is

Select sum(billing_amount)

From billing_fact_aggregate

Where ward_id=10;

Query time= number of records / scan rate

=30 records / 10 records per second

=3 seconds

Types of Aggregate table

4.7.1 One way aggregate table- In one way aggregate table, the aggregation of fact values is done on the basis of one dimension. Eg the month dimension displays the time dimension rolled up to hold time information at month level.

4.7.2 Two way aggregate table- In two way aggregate table, the aggregation of fact values is done on the basis of two dimensions.

4.7.3 N way aggregate table- In n way aggregate table, the aggregation of fact values is done on the basis of n dimensions.

4.8 Operations on OLAP

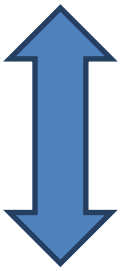
Drilling- drilling means to move from one level to another level in order to explore different aspect of the business outcome. It can be done in two different ways:

4.8.1 Rolling up/Drilling up- it means rollup a measure in some way. In drilling up dimensional details are removed to show the reports on aggregated level.

4.8.2 Drilling Down- it means the further break out a measure in some way. In drilling down dimensional details are added to explore the reports on more detailed level.

Quaterly Medicine Sales summary		
Region	Units_Sold	Revenue
Northeast		
Southeast		
Central		
Northwest		
Southwest		

Drilling Up



Drilling Down

Quaterly Medicine Sales summary			
Region	State	Units_Sold	Revenue
Northeast	Assam		
	Sikkim		
	Arunachal Pradesh		
	Meghalaya		
	Tripura		
	Mizoram		
	Nagaland		
Southeast			
Central			
Northwest			
Southwest			

Fig 4.6: Operations on Dimensions and Facts: Drilling up & Drilling Down

Let us Sum up

In this chapter, we have studied about the concept of data model, its importance in a DWH project, and the roles and responsibilities of data modeler in project. Data model is the abstract representation of structural components of a Data warehouse, it lays foundation for the build of DWH. After having good understanding of data needs of an organization, a data modeler works with business analysts to create a data model. We then defined the dimensional modeling and elaborated the elements of dimensional modeling i.e. Fact and Dimension. We have also learnt the concepts of different schemas used for designing a DWH, Bus matrix and Information package diagram, and Aggregate table and OLAP operations.

Unit End Question

Q1. Explain Star Schema and snowflake schema with example?

Q2. What are the shortcomings of OLTP design?

Q3. What is aggregate fact table, its types and its advantages? Explain with example.

Q4. Fact table and its different types, explain them briefly?

Q5. Define Fact less fact table with one example, why is it needed?

Q6. Define Dirty dimension with an example.

Q7. Define Conformed dimension with an example.

Q8. Define Degenerate dimension with an example.

Q9. Solve

Suppose there are 10 products 4 stores and 1000 customer. Each customer purchases 1 piece of each product every month from all the 4 stores. How many records will be populated in fact table in 1 year? How many facts and dimensions tables are required in this scenario? Write the name of fact and dimension?

Q10. **Answer the following**

a) Define fact and aggregate fact table. What are the types of fact table and aggregate fact table? Explain them briefly with example.

b) Suppose two records for an employee are coming from source on 1st Jan 2015 and 2nd Jan 2018.

1st Jan 2015

Stud_id	Sname	Program_id	Program_name	Address
1000	Queen	10	MCA	A/12, street # 3

2nd Jan 2018

Stud_id	Sname	Program_id	Program_name	Address
1000	Queen	20	MBA	B/238, street # 5

Explain SCD and its types and draw SCD 1, SCD 2(on date) and SCD 3 after processing of data from 3rd Jan 2018.

2. Answer the following questions

a. Below is the attendance fact table structure

STUDENT_ID	Time_ID	Course_ID
100	01-JAN-19	C10
101	01-JAN-19	C10
-	-	-
199	01-JAN-19	C10

Suppose there are 100 students in a class and 1 course Attendance is taken every day except holidays. There are 100 holidays in a year. How many records will be populated in attendance table in the year 2019? Which type of fact table is this?

b. What is the role of Staging Area in Data Warehousing, explain in brief?

c. Why is the separate storage component needed to keep data in Data Warehouse?

d. Explain Star Schema and snowflake schema with example?

e. What are the shortcomings of OLTP design?

3. Solve (1 * 6 = 6 Marks)

Suppose two records for an employee are coming from source on 1st Jan 2018 and 2nd Jan 2019.

1st Jan 2018

Emp_id	Ename	Designation	Salary	Dept_number	Address
100	King	Junior Engineer	25000	10	A/12, street # 3

2nd Jan 2019

Emp_id	Ename	Designation	Salary	Dept_number	Address
100	King	Engineer	35000	10	B/238, street # 5

Draw SCD 1, SCD 2(on date) and SCD 3 table after processing of data from 2nd Jan 2019.

Suggested Reading

1. Paulraj Ponniah; Data Warehousing Fundamentals for IT Professionals; Second Edition Wiley Publishers 2015.
2. Ralph Kimball & Margy Ross; The Data Warehouse Toolkit; Third Edition Wiley Publishers 2013.

UNIT V: DATA TRANSFORMATION PROCESS FUNCTIONS

Learning Objectives

After completing this chapter, you will be able to

- Understand the logical transformation of data
- Understand the formatting of data
- Understand the data quality

In the last section, we have focused on the different aspects of data modeling. This activity is performed in the design phase. After design, we will switch into implementation phase. This section deals with implementation part of the DWH life cycle. In implementation, we have to focus on the Extract Transform Load (ETL) process. Extraction-Transformation-Loading (ETL) is the process of fetching relevant data from many different source systems, applying the required business and functional logic and loading it into Data Warehouse.

ETL perform data extraction from various resources, transform the data into a format suitable for Data warehouse, and load it into DW storage area. ETL process in data warehousing technologies is responsible for performing the following tasks from pulling data out of the distinct source systems to placing it into a data warehouse:

- Data is extracted from the distinct source systems; the extracted data is then converted into a unified format that is suitable for data warehouse format data and ready for further transformation processing.

Transformation of the data may involve the below tasks

- Business requirements given by organization are applied
- Columns are cleansed
- Rows are filtered out
- Combining multiple columns into one column
- Splitting operation of one column into multiple columns
- Data from heterogeneous sources are joined together
- Performing transpose operation
- Performing simple and complex data validations
- Transformed data is loaded into data warehouse or data marts

The important functionalities of an ETL tool may be summarized in the following tasks, which include majorly as:

- To identify the relevant information at the source side;
- To extract the relevant information;

- To customize, and integrate the information coming from multiple distinct sources into a common format; To cleanse the resulting data set, on the basis of data and business rules given, and
- To propagate the transformed data to the Data Warehouse and/or data marts

An ETL process has to finish its execution in a limited time window so that the most current information would be consumed by decision-making persons to make the useful business strategic decisions; otherwise, the DW would remain unavailable to decision makers and its users as well due to huge volumes of data manipulated by typical DWs. It becomes very crucial to optimize the whole ETL execution so that the huge volumes could be processed in a definite and specific time frame given by clients.

ETL Difficulties:

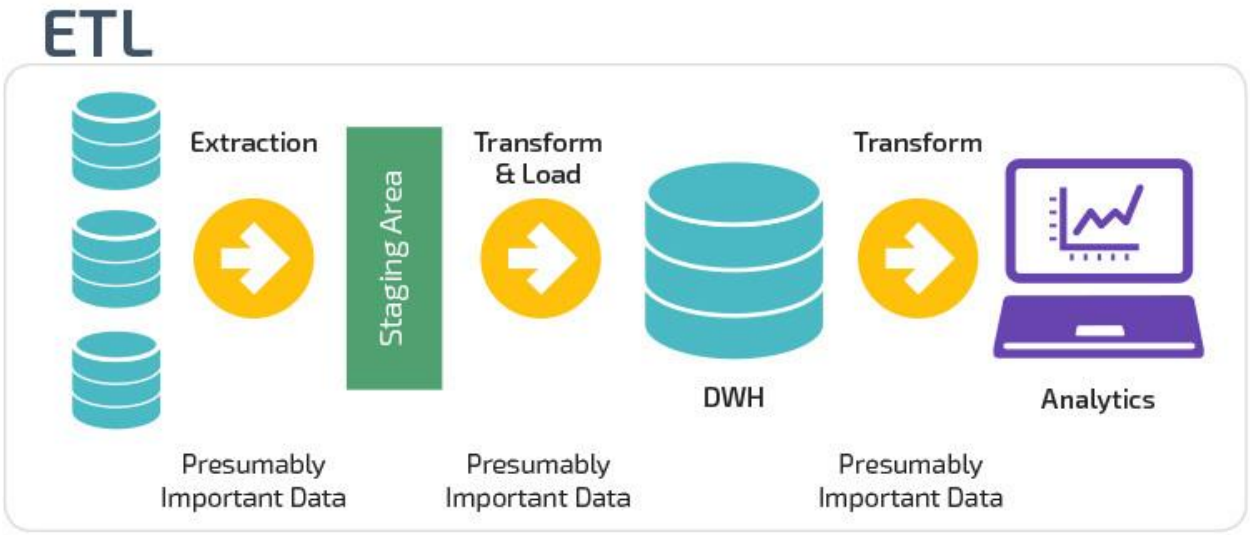
Source System are heterogeneous in nature as they run on multiple platforms and different operating systems. Some source systems are legacy applications running on obsolete database technologies. In DWH, historical information is critical and difficult to preserve the changes in the value of data on operational system. In some old legacy system that are evolved over time, in that case quality of the data is doubtful. As the conditions of the business changes with respect to time, the structure of the source system also changes for this ETL process design should be flexible enough to accommodate those changes. Some source system stored data in the format that is not interpretable conveniently for the user. Data format is inconsistencies. It is highly likely that representation of same data is different in different system. So, data format consistencies is also one of the major problems that needs to be addressed in ETL processing.

To handle all these problems makes the ETL process a time-consuming task.

ETL Steps:

DIAGRAM

ETL Architecture



Data Extraction:

Data Extraction is the process in which data is pulled from distinct source systems. Data is extracted in two different ways.

1. Initial Extraction for one time load- For the first-time load, the full extraction has to be done because all data available in the source system needs to be loaded into the target system.
2. On Going Extractions for incremental Load-For each subsequent load, the changes in the data happened after the last load is to be fetched and loaded into the target system. This helps in reducing the number of records to be processed. It also enhances the performance of data loading.

Issues faced in the process of Data Extraction

1. Identification of source – to identify source application and source structure

Source identification process

(a) For the analysis, each data item of metrics or facts is listed in the fact table.
(b) Attributes of each dimension from all dimension are listed.
(c) The source system and source data item is to be found for each data item of the target.
(d) The preferred source is to be chosen among the multiple source for one data element.
(e) For a single target field, multiple source field is is to be identifies and consolidation rules is to be formed.
(f) For multiple target fields single source field is to be identified and splitting rules is to be established.
(g) Default values is to be ascertained.
(h) For incomplete data, source data is to be inspected.

2. Extraction methods- it can be either manual or tool based. This depends on the source system as whether the connector for that system is available in the ETL tool or not. Connector for some legacy system is not compatible with ETL tools so manual extraction is required.
3. Frequency of Extraction- how frequently data is to be extracted. It depends on the requirements, when data should be available in the target system that is whether it is near real time or deferred, so data extraction can be performed in two ways, they are
    a) Immediate Data Extraction- In immediate data extraction data is captured through transaction logs, database triggers, and using application programs to capture changes in source application.
    b) Deferred Data Extraction- In deferred data extraction data is captured based on date and time stamp of the last extraction and by comparing files with the files extracted last time.
4. Time window-For each source of the data, the specific time is scheduled for the extraction.

5. Execution order of job-determine the dependencies of jobs on each other. If the job is dependent then they must be executed in sequential manner, in case of no dependency (i.e independent) these jobs can be run in parallel mode i.e execution at the same time.
6. Handling the exception- it is the strategy in which the bad records of the source is handled. In exceptional handling, the logic is developed to correct the error records or to notify the source team about bad records in their system. Source team is responsible to make the necessary corrections in those records to make them suitable for processing in future. Business requirements to make it suitable for the target system.

Validations during Extractions as follows:

1. The source data records are to be reconciled
2. No spam/unwanted data is loaded
3. Data type is to be checked
4. All types of duplicate/fragmented data are to be removed
5. all the keys are in place or not are to be checked.

Transformation is the process which is responsible for the most complex part of processing in ETL as it deals with:

- Business and functional specifications.
- Specifying the data types for each and every field that has been selected
- Joining tables
- Create derivative fields using expression
- Create fact and dimension tables
- Creating a sequence for surrogate key

Validations during Transformation

1. Filtering – specific columns and rows are to be filtered for loading
2. Data is to be standardized as per business rules and lookup tables.
3. Different character sets (UTF-8, ASCII, UNICODE so on) may be used across different source system so proper handling of character set conversion is needed.
4. In different source systems, metrics are measured in different units. So, there is a need to convert it into a common unit for the purpose of analysis.
5. Threshold value of the data is to be checked for example; age cannot be more than two digits.
6. Flow of data from the staging area to the intermediate tables is to be validated
7. Fields should not be NULL.
8. Cleaning-as per cleansing rules, data is to be scrubbed and enriched. For example, mapping Gender Female to 0 or Gender Male to 1 and Transgender to 2
9. Split a column into multiples and merging multiple columns into a single column.
10. rows and columns are to be transposed

11. data is to be merged by using lookups
12. Complex data validations-these operations are processor intensive operations so if first some columns are failed then the remaining validations must not be performed for that record.

Data Loading- Loading is the process that performs loading data into the DW. Data loading can be done in different ways-

Initial Load- For the first time loading the data into the DWH is termed as initial load. When DWH systems goes live so all the data available in the source system has to be transferred to the DWH storage system. At this time initial load is performed to do this.

Incremental Load- In incremental load, loading the data into DWH in periodic manner. Business proceeds with its functions and generates data continually so in order to move the generated data on a periodic basis, incremental load is required to do this.

Full Refresh- if some hidden bugs is unearthed after go live, or there is some logic change comes as per business requirements then in order to handle such scenarios the loaded data in DWH storage system has to be erased completely and full load to be done. Why not updating of records instead of reloading completely. Huge volumes of data are stored in DWH tables and updating so many records may be very costly operations in terms of resources (memory, CPU utilization etc.) and time consumption. So generally, reload operation is preferred. But suppose if only a chunk of records of small volumes needs to be changed for example less than 50K records so in that case updating of data is preferred.

Verification during loading

1. The key field data is to be ensured neither missing nor null.
2. That combined values and calculated measures are to be checked.
3. Data is to be checked in dimension table as well as history table.
4. BI reports on the loaded fact and dimension table is to be checked

The important functionalities of an ETL tool may be summarized in the following tasks, which include majorly as:

- To identify the relevant information at the source side;
- To extract the relevant information;
- To customize, and integrate the information coming from multiple distinct sources into a common format;
- To cleanse the resulting data set, on the basis of data and business rules given, and
- To propagate the transformed data to the Data Warehouse and/or data marts.

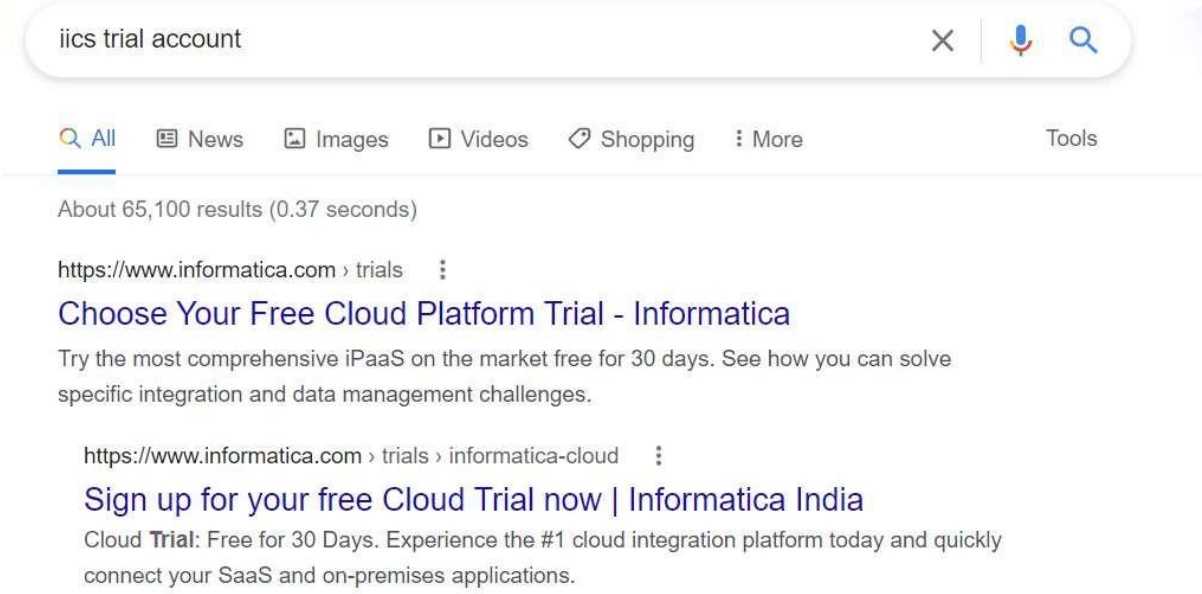
ETL Tools available in the market

- Informatica
- Talend
- Oracle Warehouse Builder

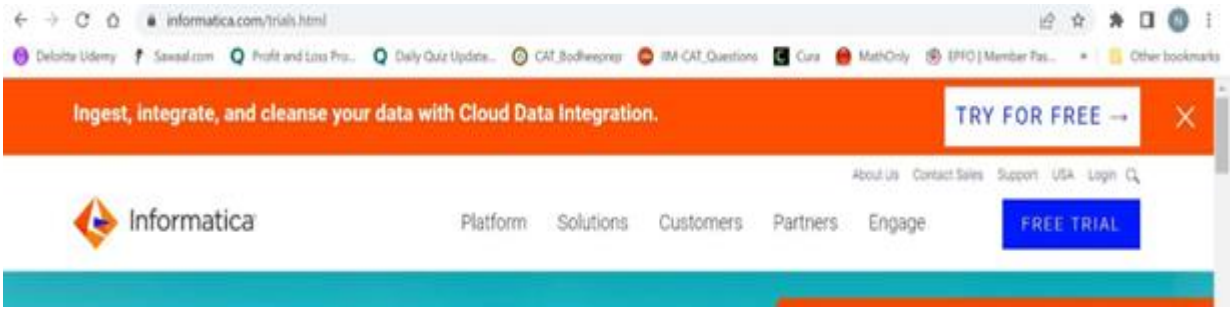
- Oracle Data Integrator
- Ab- Initio
- SSIS
- Data Stage
- SAP-Business Object Data Integrator
- Sybase ETL
- Pentaho

IICS Setup steps:

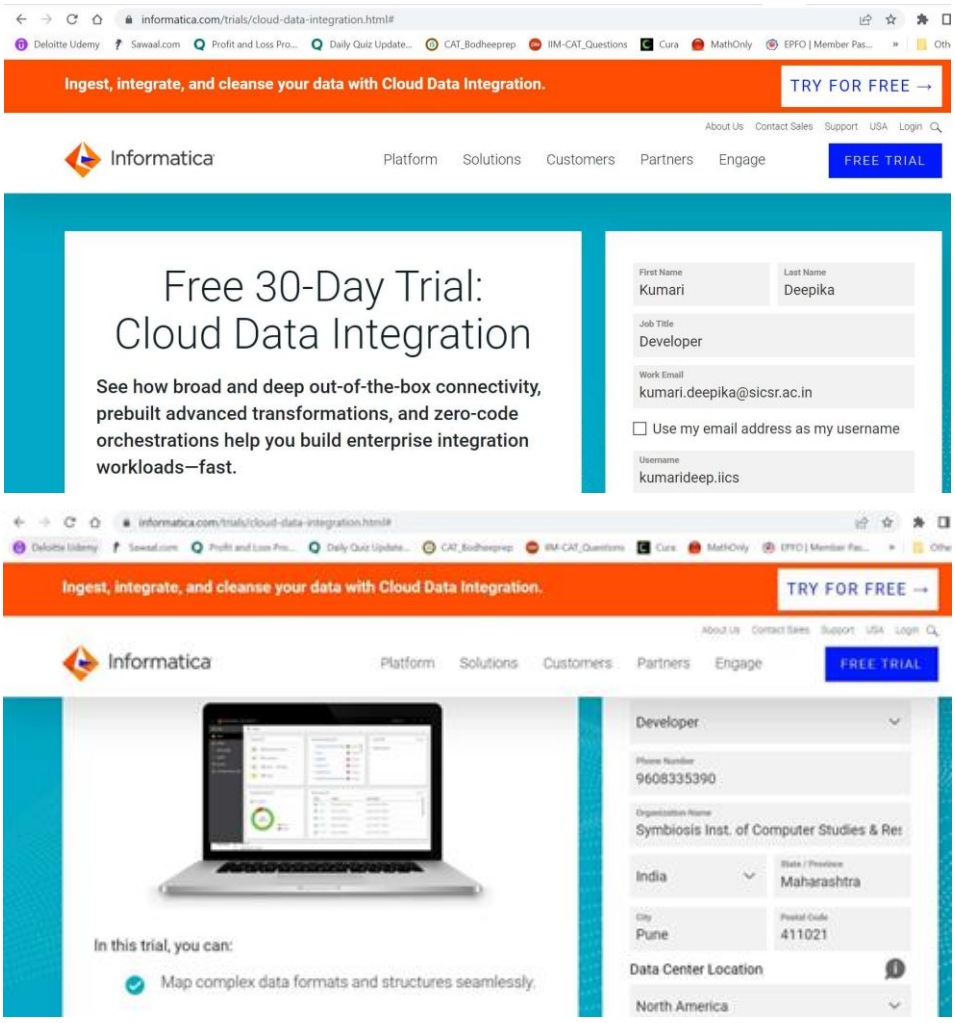
1. Open www.google.com
2. Search IICS trial Account.



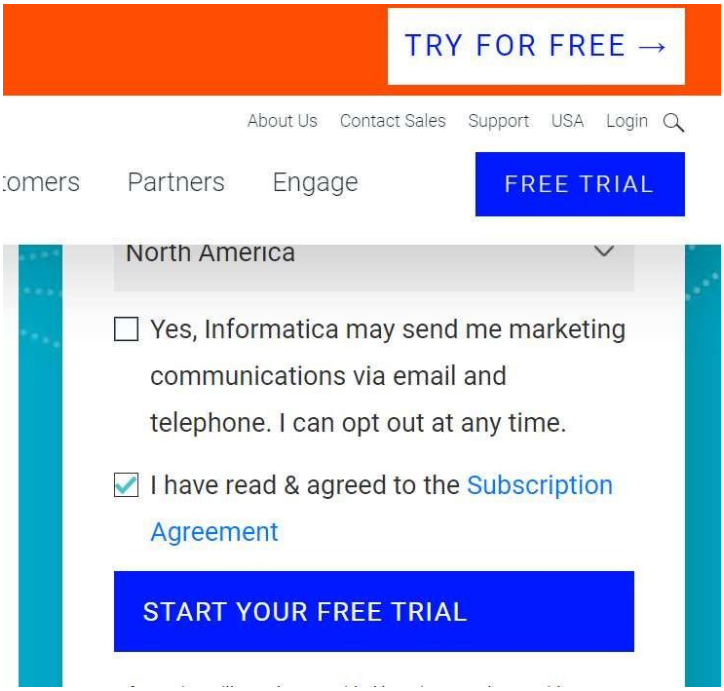
3. Click the first link, and then click on **TRY FOR FREE** option.



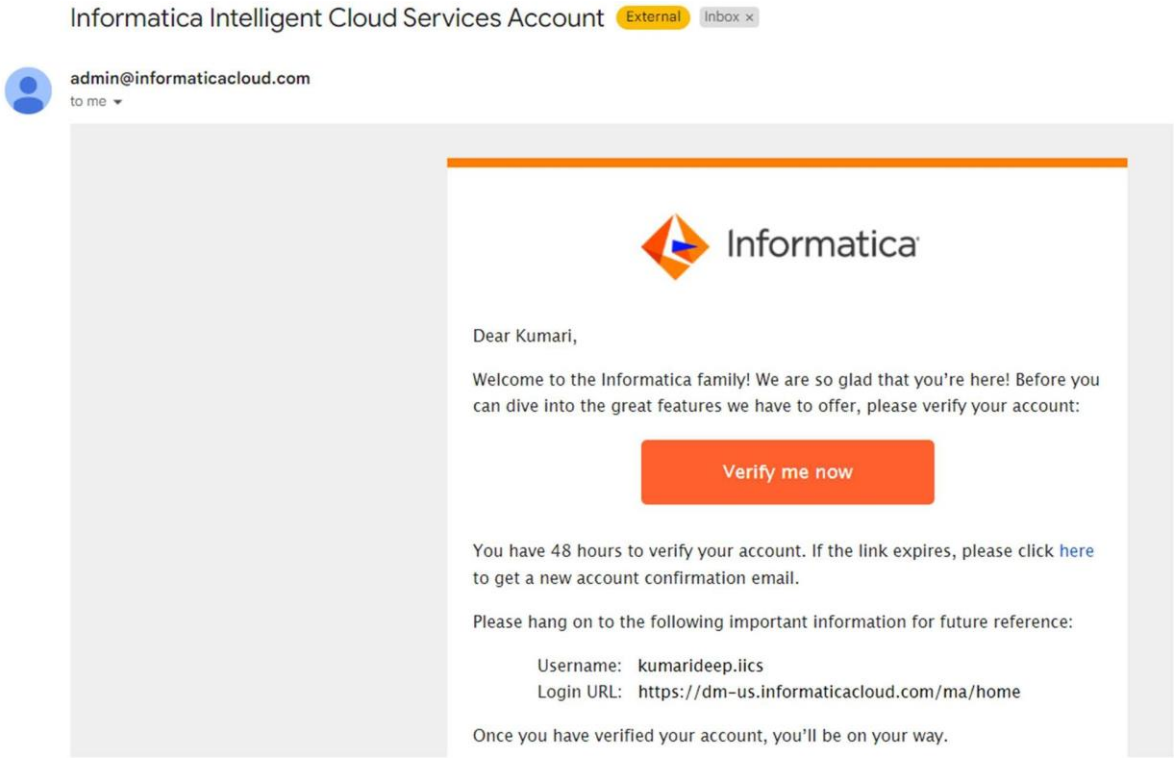
4. Put your First Name, Last Name, Job Title, Work Email etc. and uncheck box **Use my Email Address as my Username.** Put your Username of your choice.



5. Click on **Start your Free Trial.**

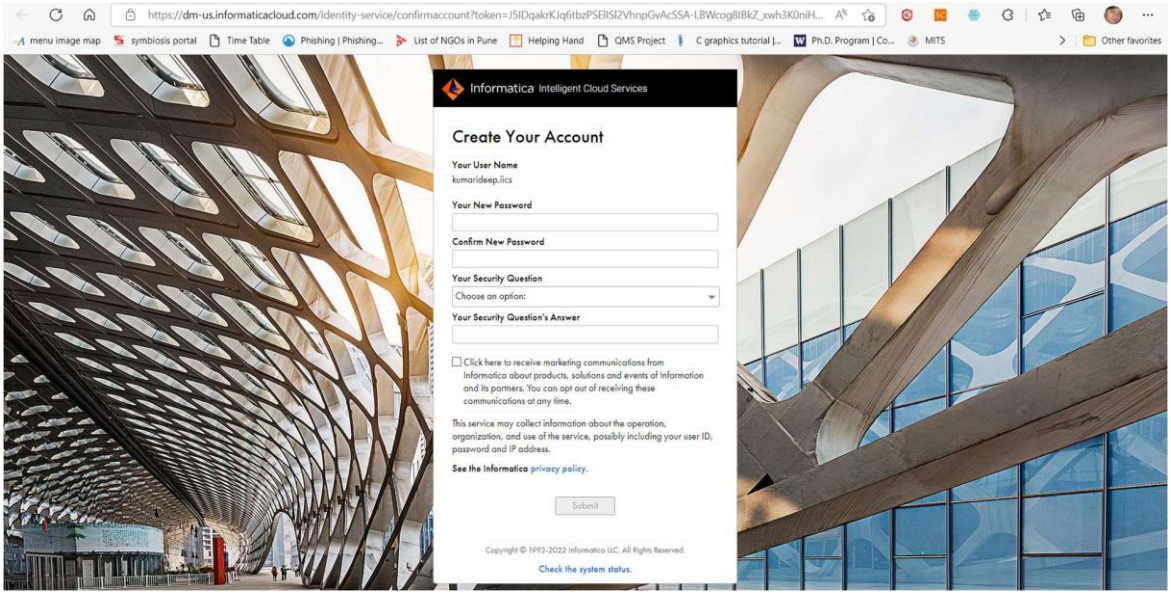


6. You will get an email from Informatica to set up your password in order to activate your account.

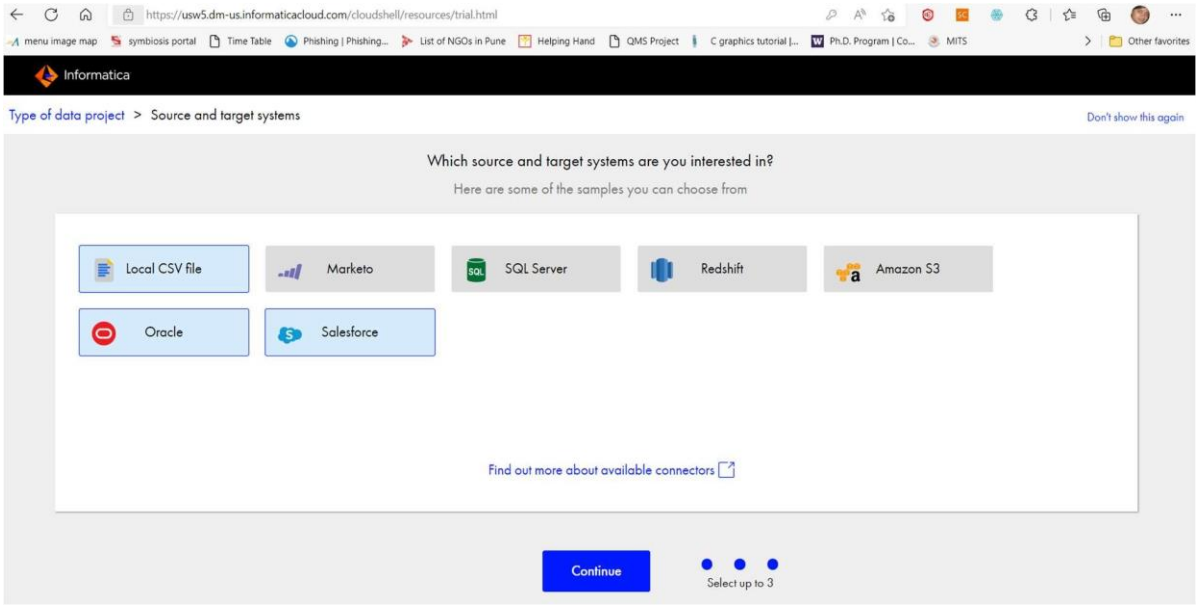


7. Click on verify me now.

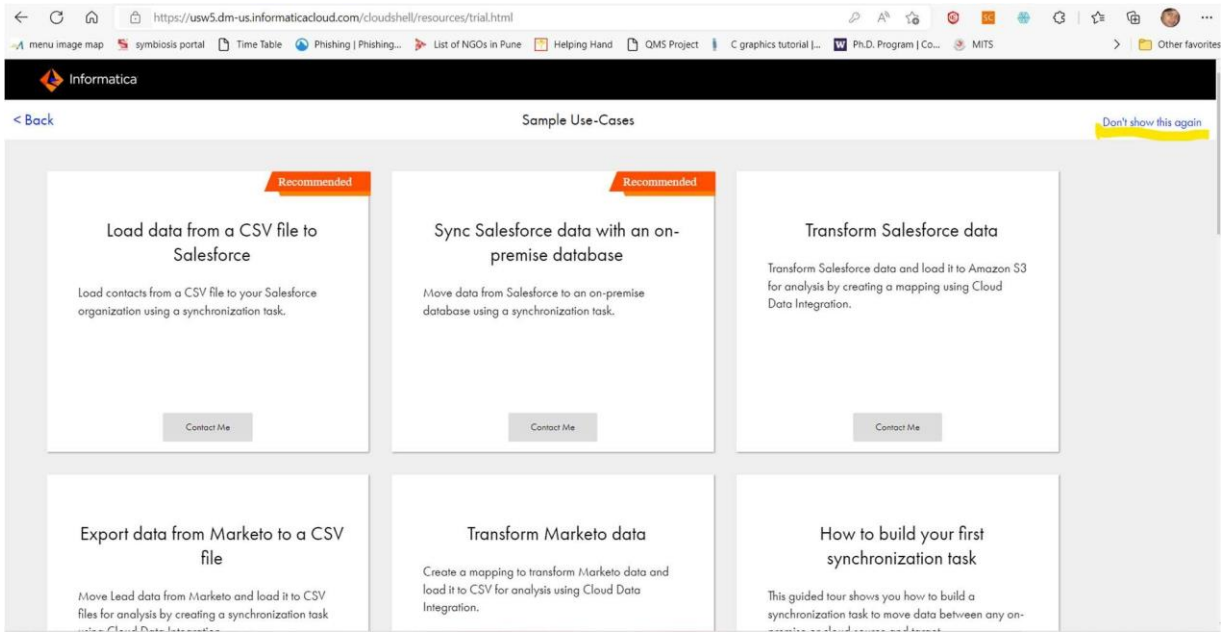
8. Set up your password by filling in the below details



9. Once you set up your password, you have to select three types of data sources for integration.

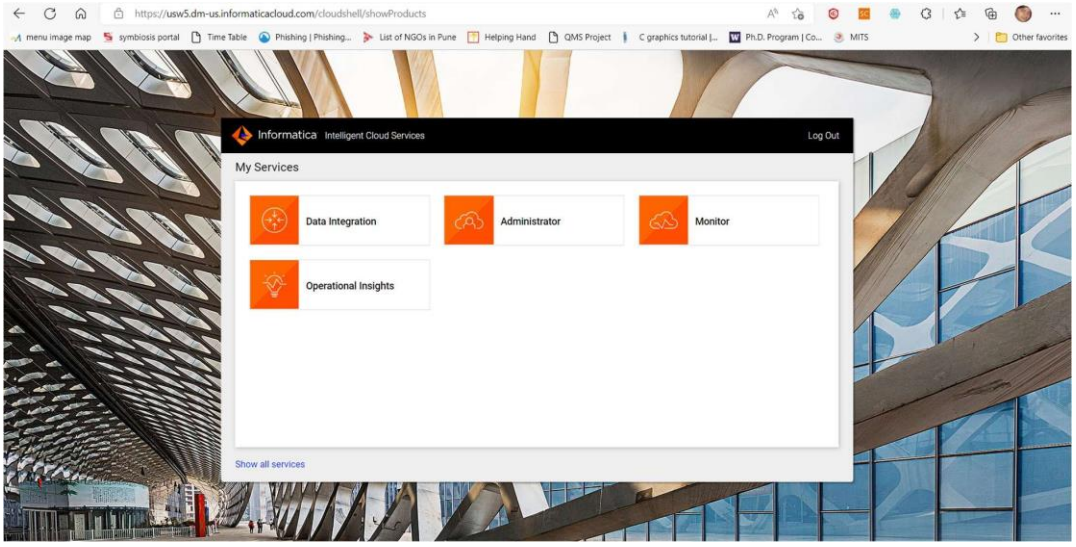


10. Click on continue.

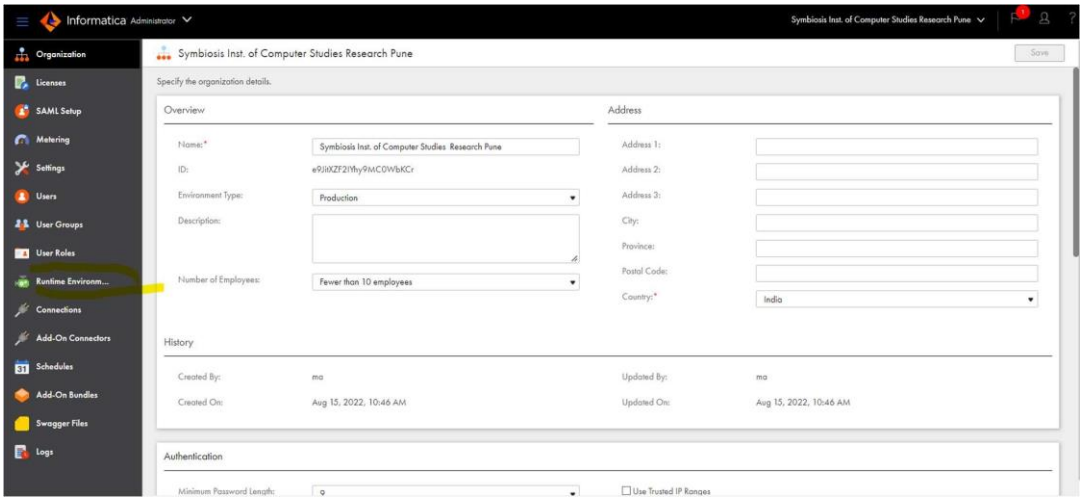


11. Don't show this again.

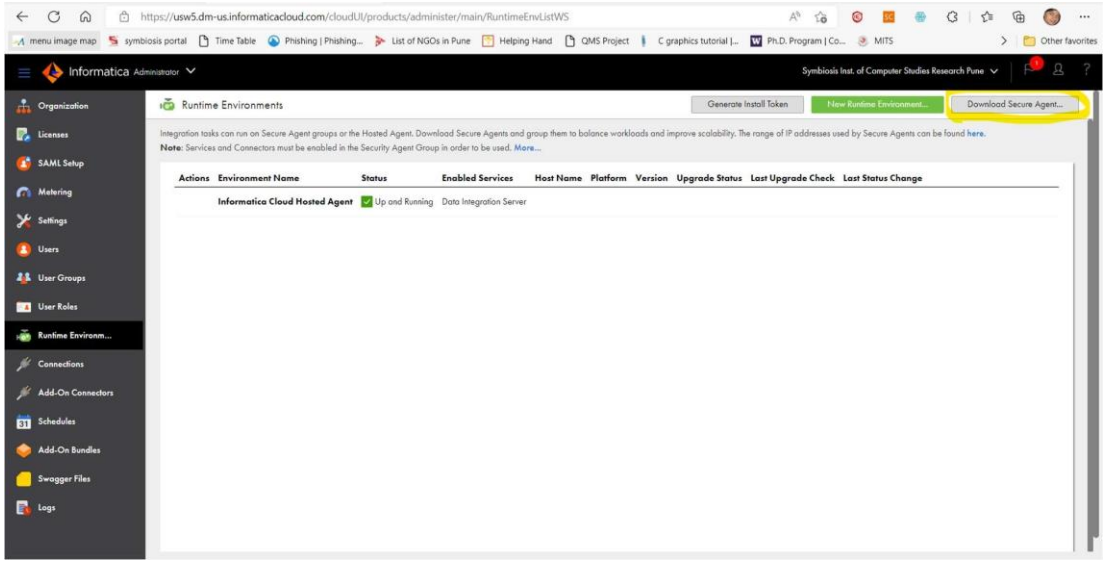
12. Go to the Administrator



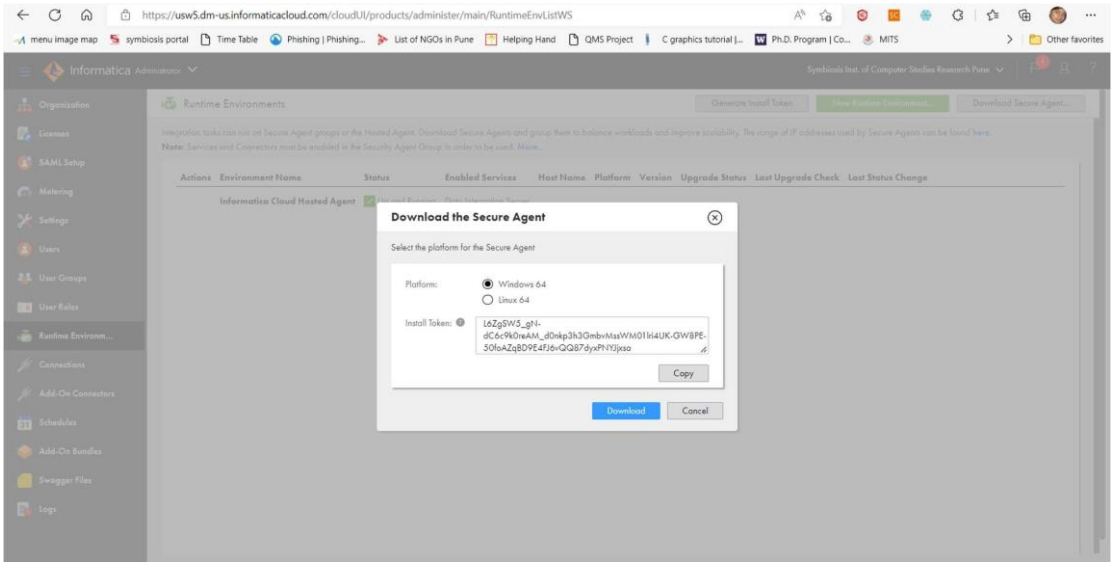
13. Click on runtime environment



14. Download and install secure agent.



15. Copy the install token and keep it in a notepad.



L6ZgSW5_gN-dC6c9k0reAM_d0nkp3h3GmbvMssWM01lri4UK-GW8PE-50foAZqBD9E4FJ6vQQ87dyxPNYJjxsa

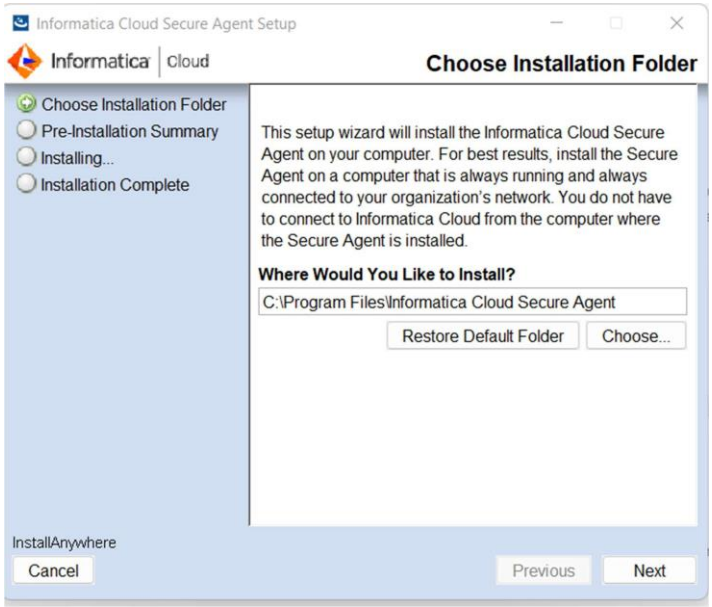
16. Click on download

Q. Why do we need a secure agent?

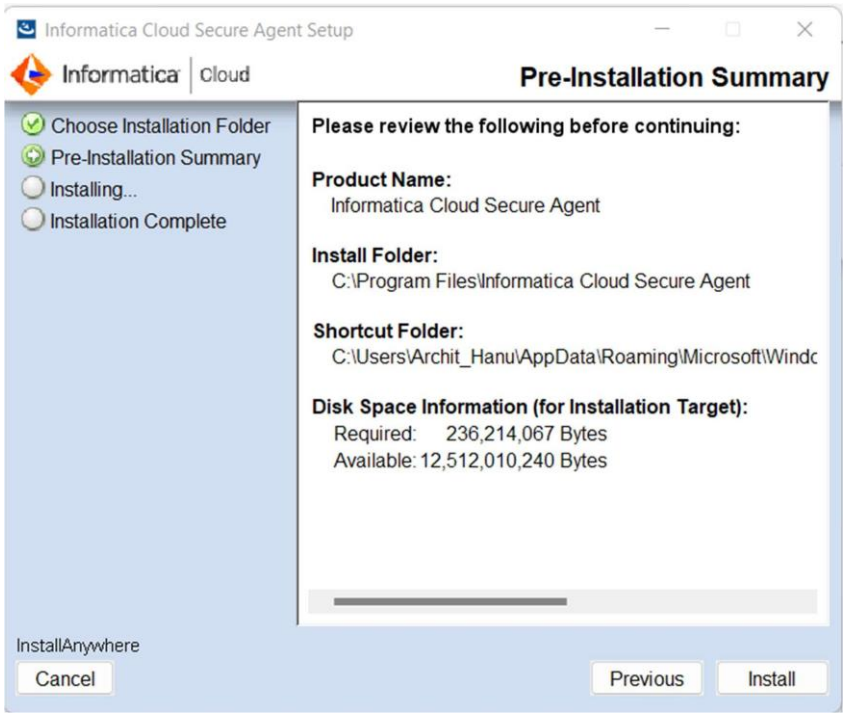
Ans: secure agent is a local run time environment. When we create iics account, it provides us one shared cloud run time environment but that is not able to read and write files or database tables available in our local machine. If we need to do read or write from our local then secure agent is needed.

17. Go to download folder , where secure agent is downloaded and double click on the executable file(.exe)

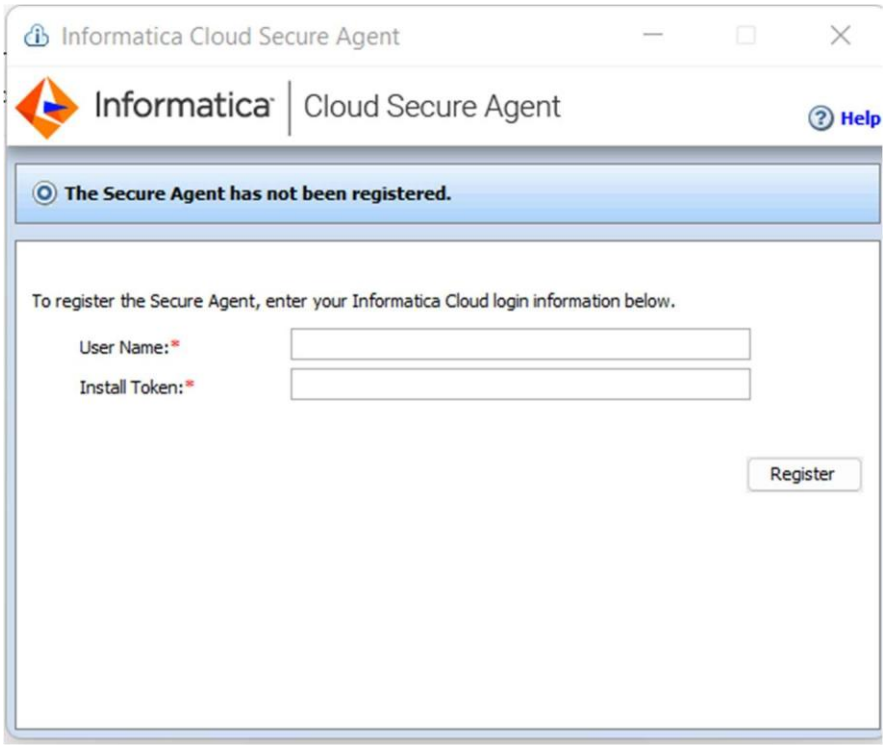
18. Your installation will start, click on next



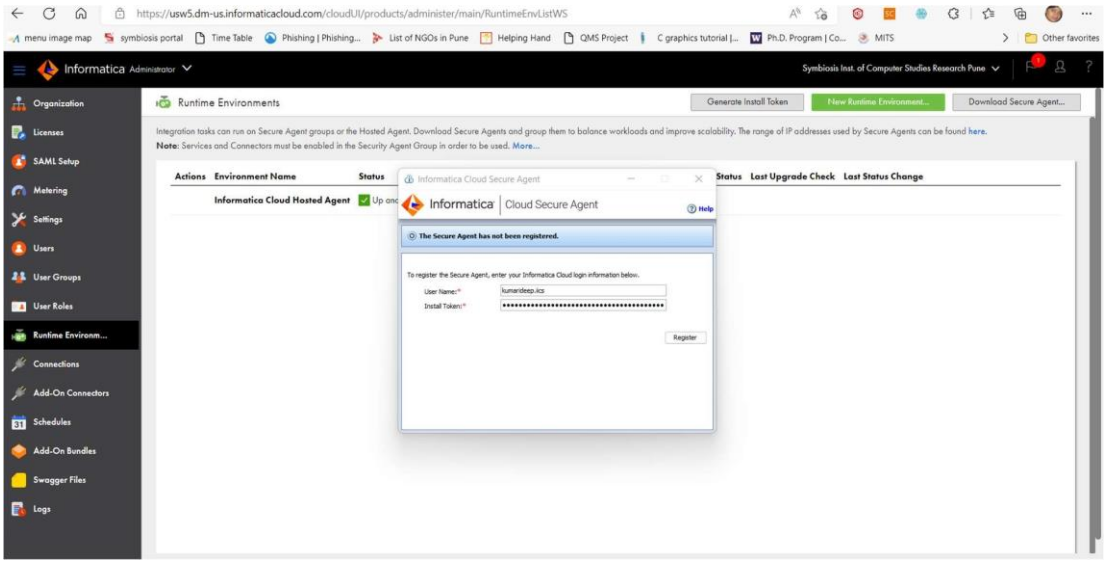
Click on install



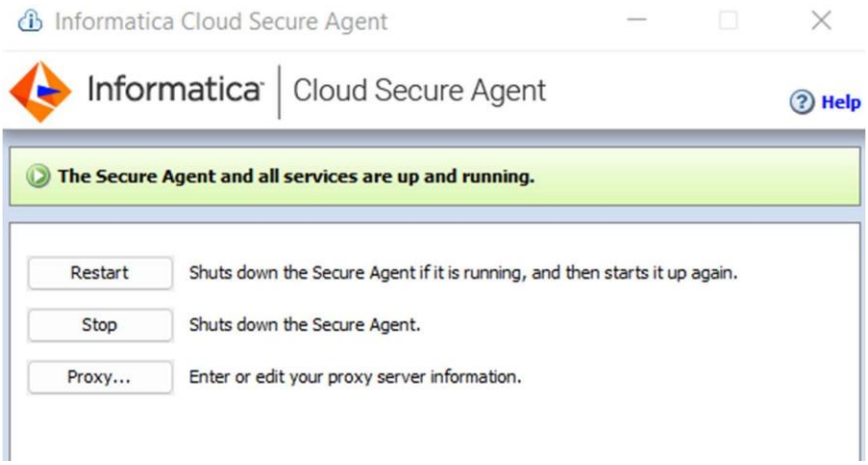
Once the installation of the secure agent is done, thereafter a popup window will come. Wehave to put our user's name and install token there.



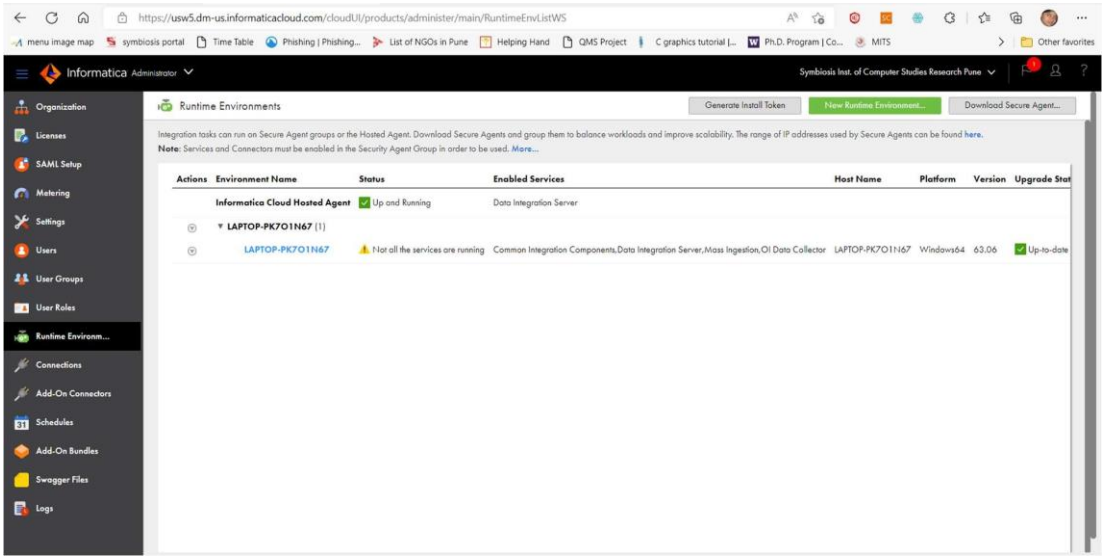
19. Click on register



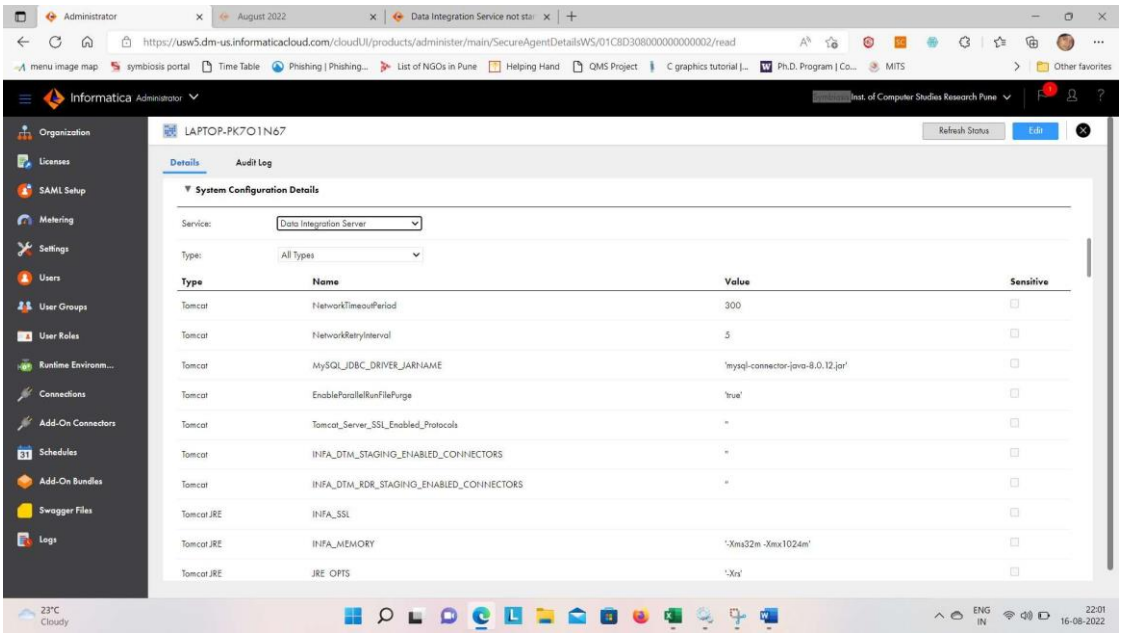
Click on restart



20. Go and refresh iics run time environment page. It will show the new secure agent installed there.

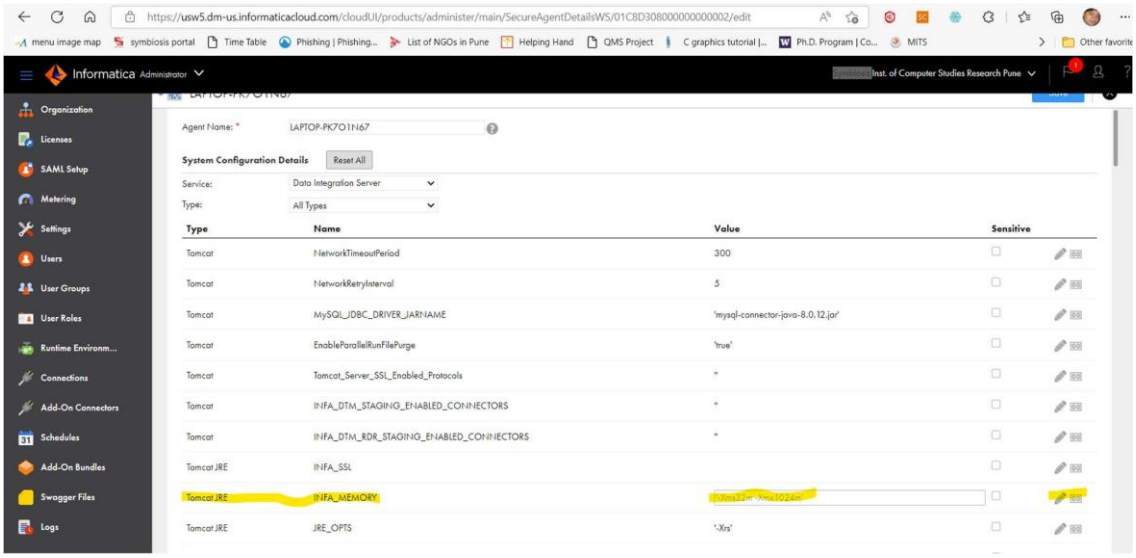


Issue: Data integration server service is not starting upResolution:

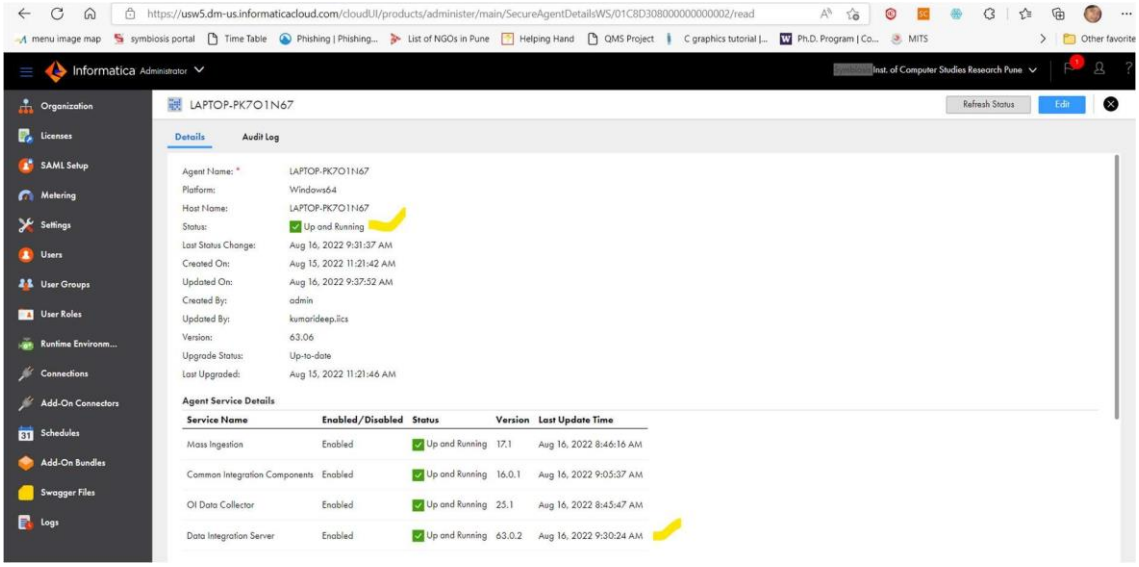


Go to system configuration Details and select data integration server from the drop-down menu ofservices

Click on edit and change the value of the INFA_Memory component from 512 to 1024.



After editing click on save button.



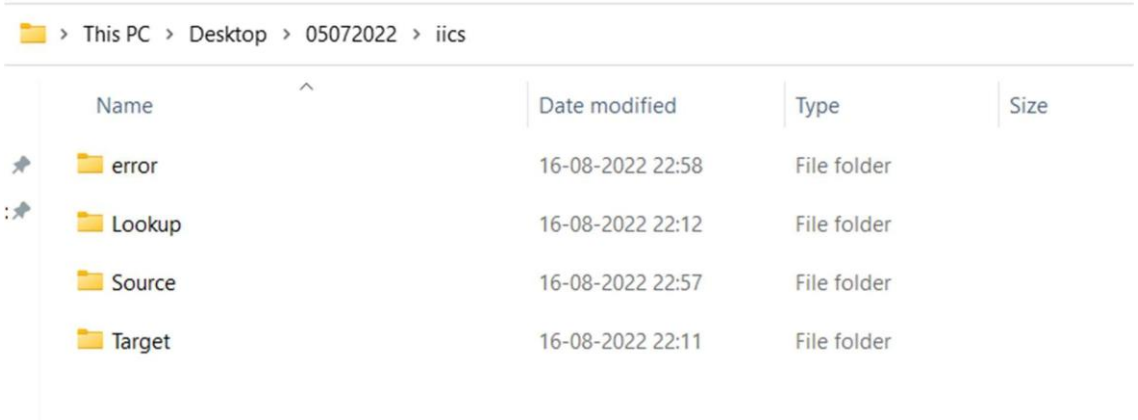
ETL Components

ETL tool is a must for organizations which are of enormous size and having many applications serving their different IT needs. Data transformation Tools (ETL) tools are generally needed for enterprise's "data consolidation" and "data integration" needs. It is very true that tools are not the only way to achieve needed objective. But they provide rapid project development and easier code maintenance.

ETL has three major components.

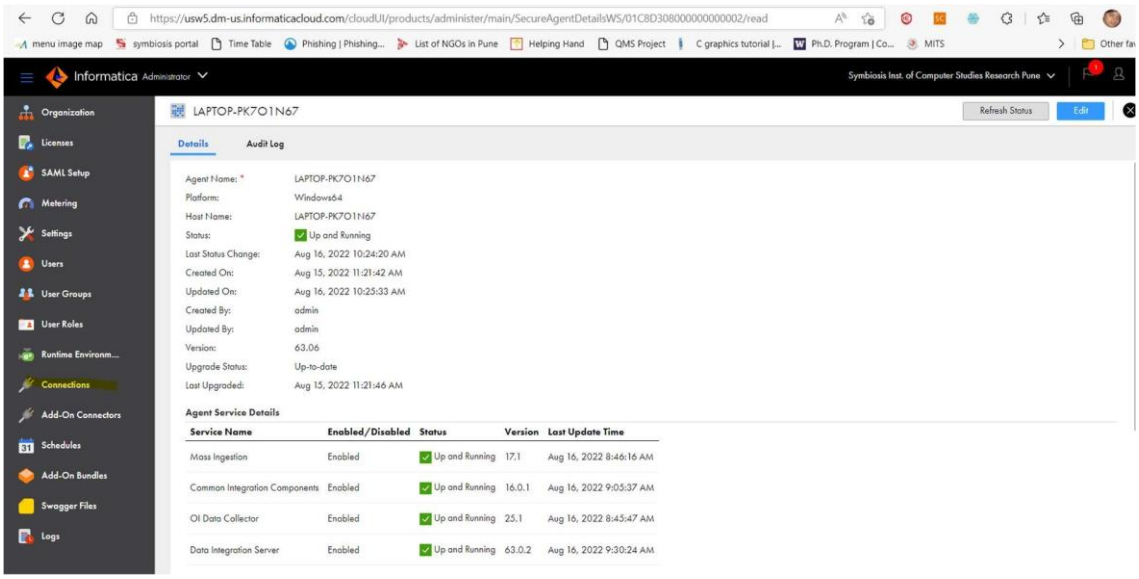
1. **Source:** Source system contains the transaction data generated from the business execution. Source structure is imported into the ETL tool and used in the ETL mapping development.
2. **Transformation:** Transformations are the objects which receive, modify, and pass data to the downstream flow for further processing. The purpose of the transformation is to modify the source data as per the requirement of the target system.
3. **Target:** Target system contains the historical data for years. It caters to the reporting needs of the business. Target structure is imported into the ETL tool and used in theETL mapping development.

Code development Create directory structure

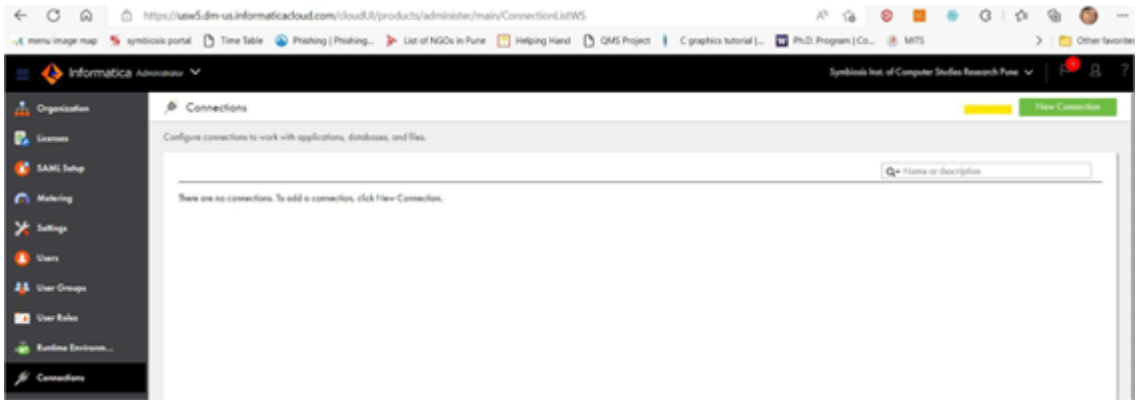


Create iics connections for all the folders created for iics development work.

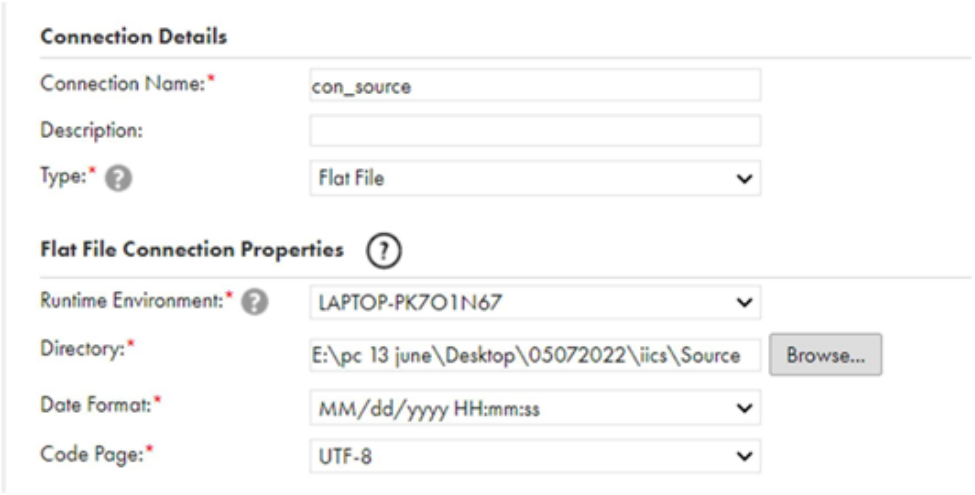
a) Go to administrator services, and click on connections



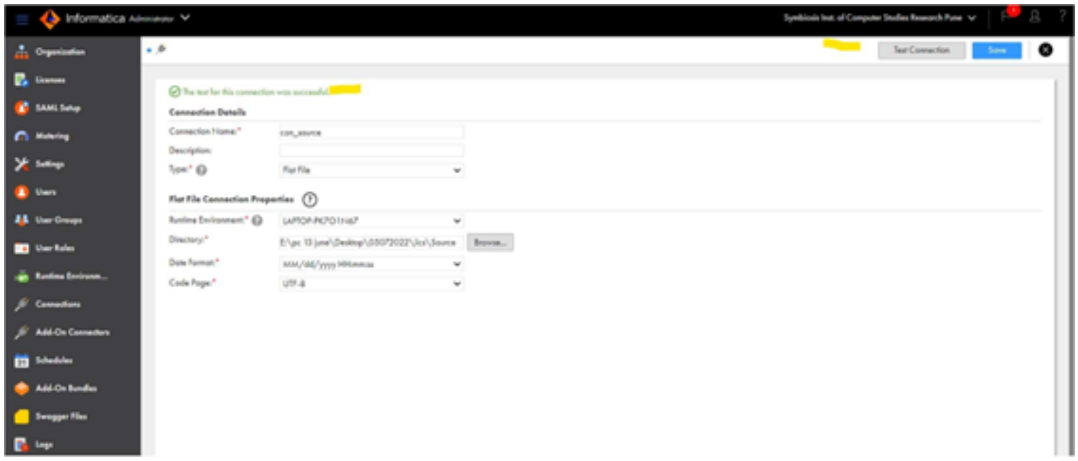
Click on New connection



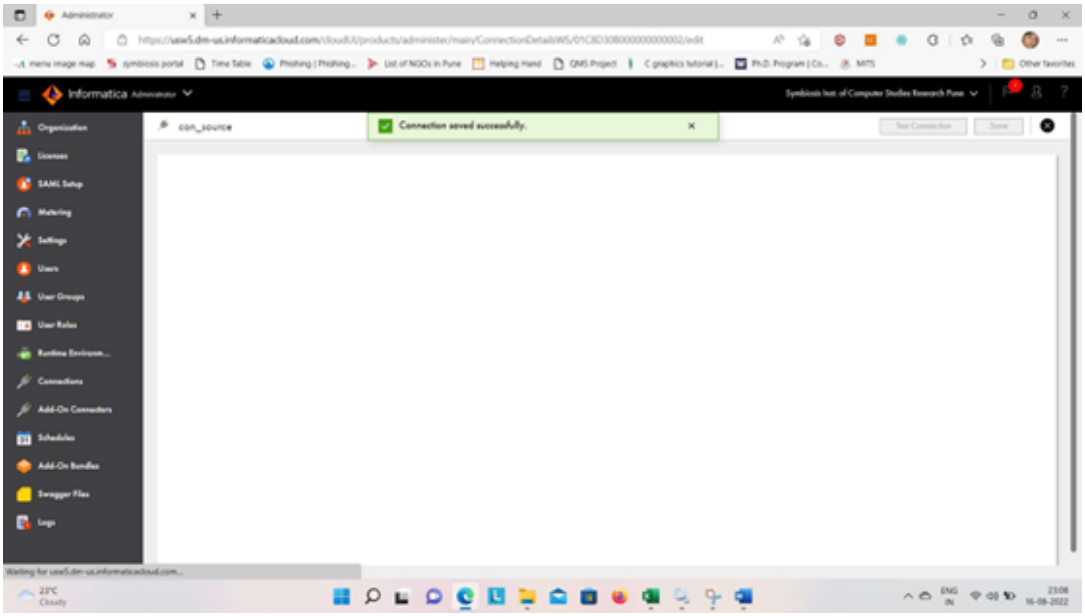
Set connection name, connection type connection properties etc. to create the connection



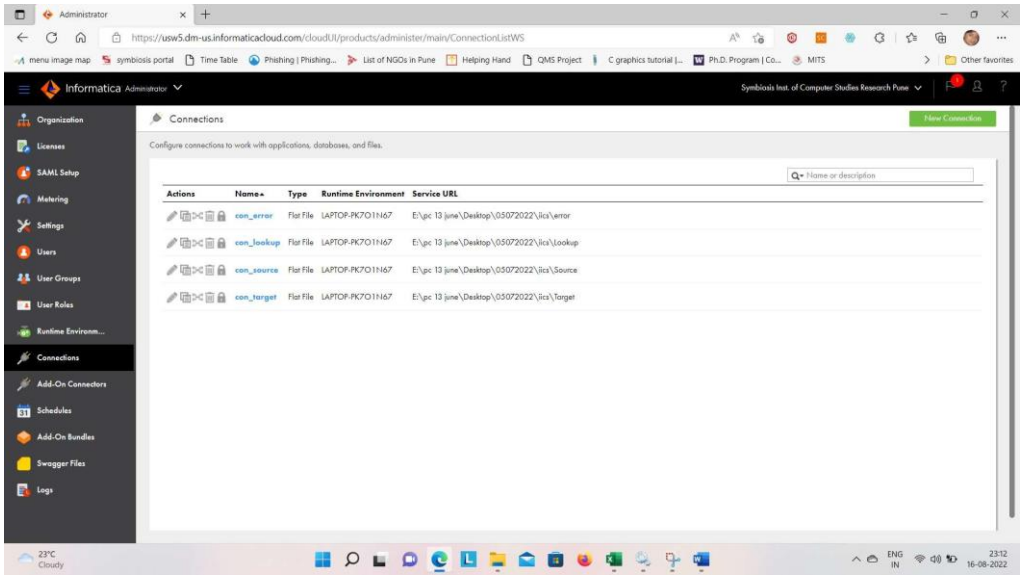
Click on test connection. If the result is successful then save otherwise check the values entered forconnection attributes and provide correct values for all attributes.



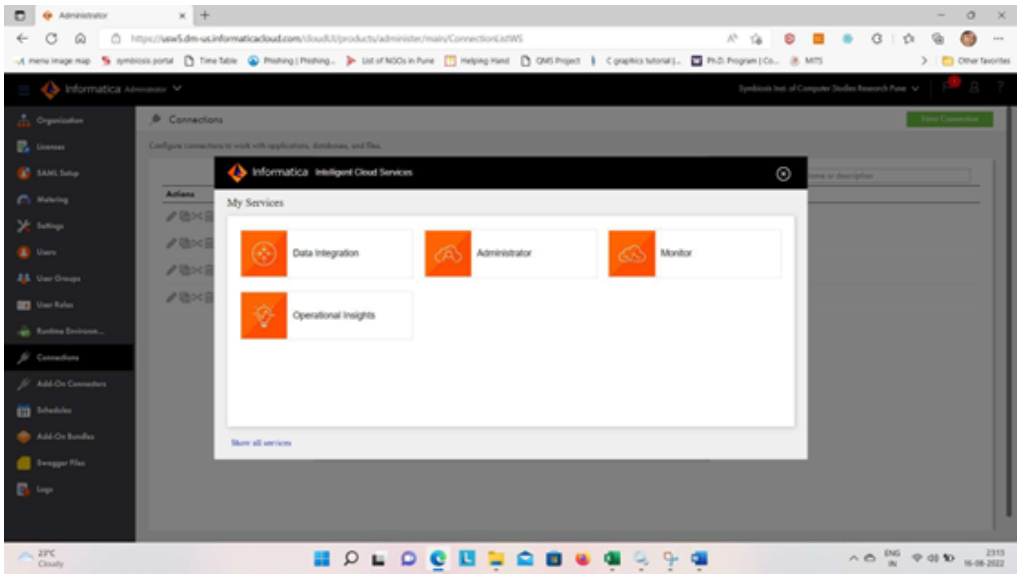
Save the connection



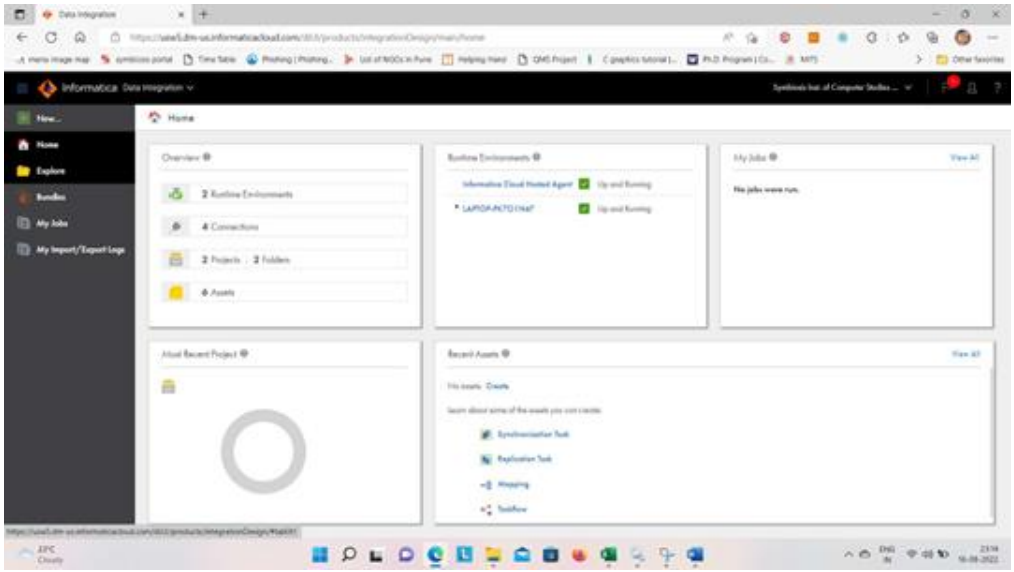
Note: do the same for the target



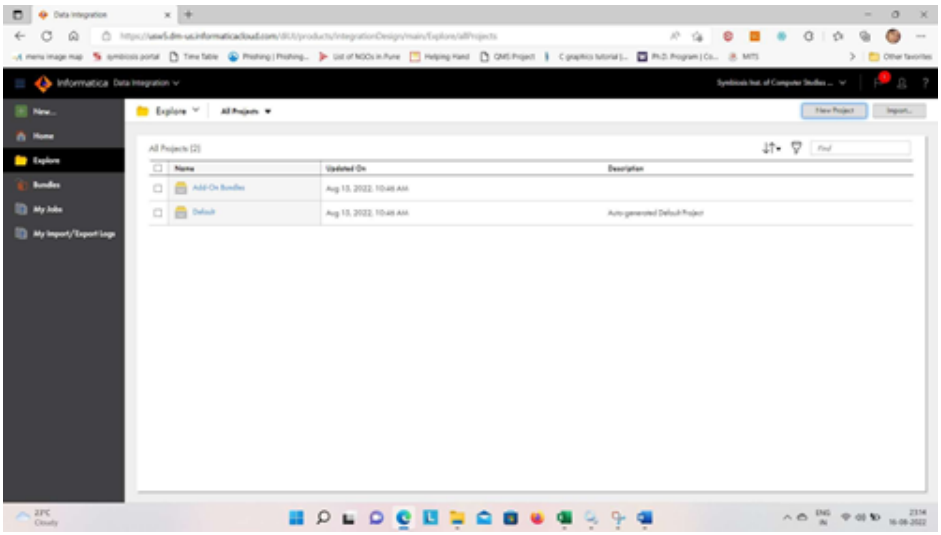
Switch onto Data integration



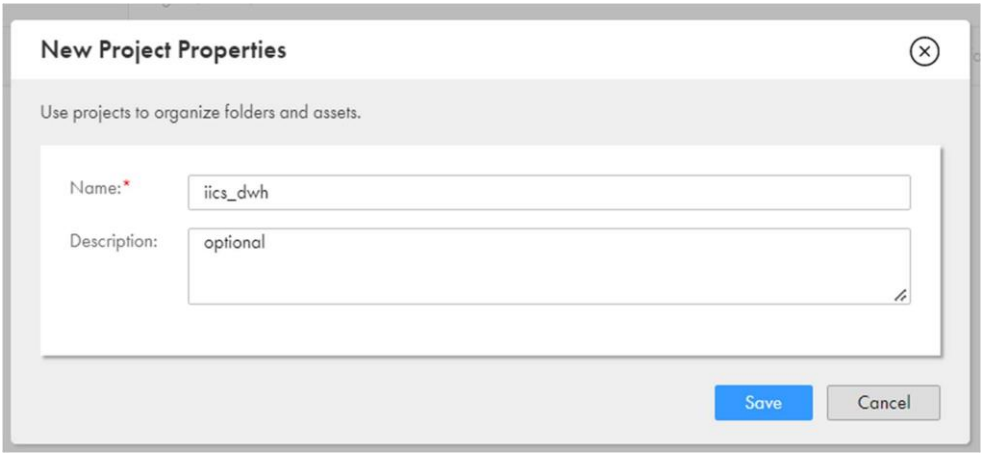
Click on explore



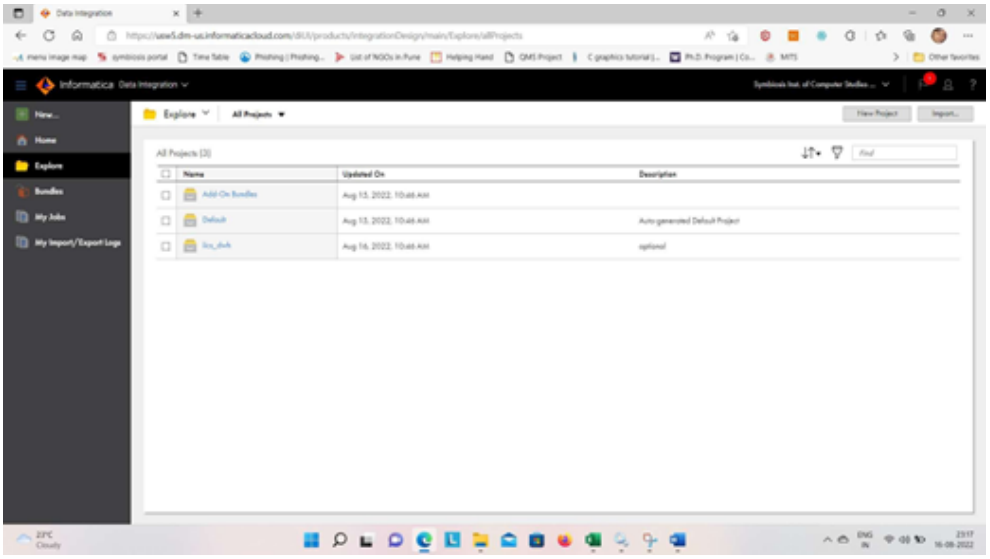
Click on a new project



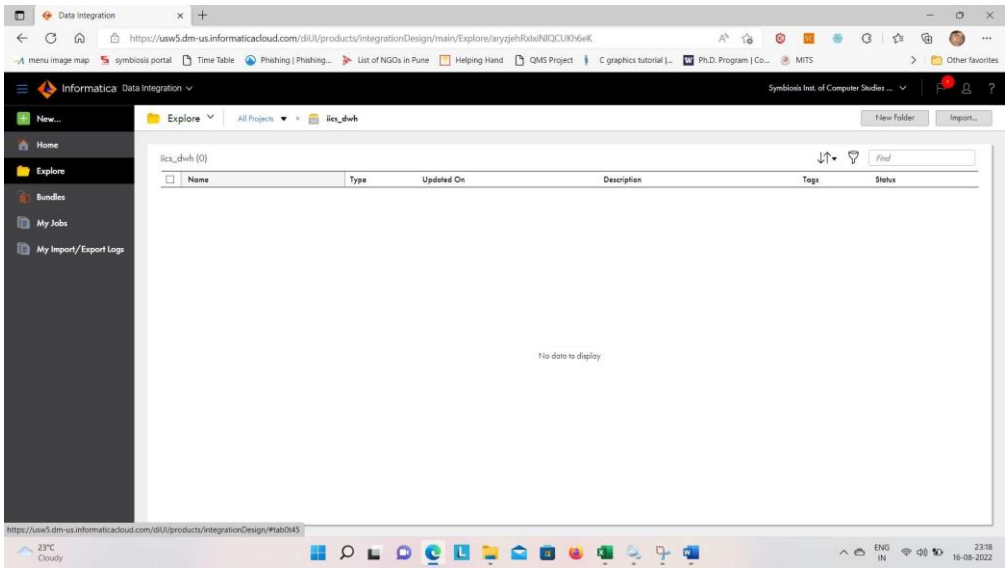
Provide a name and description for the project, and save it.



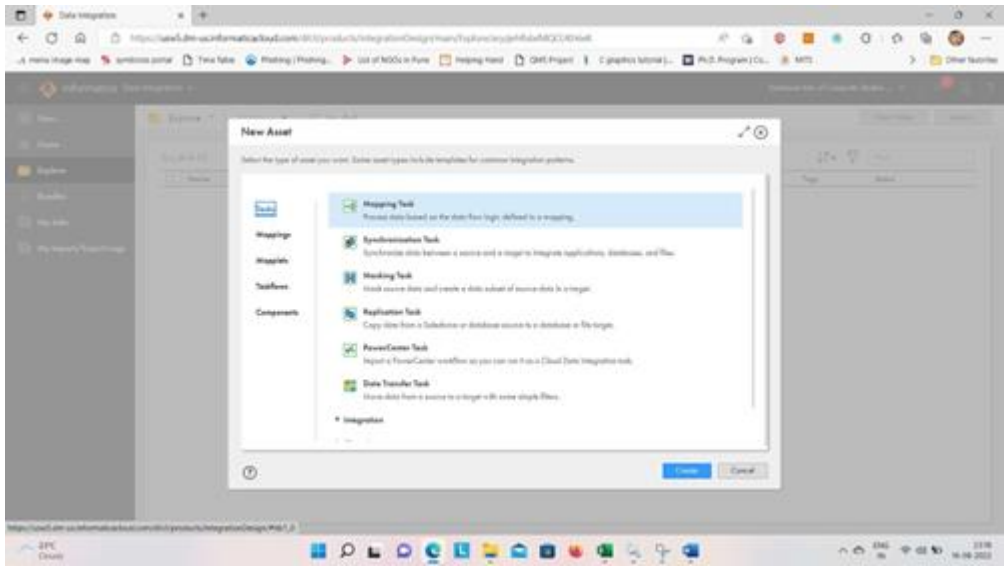
Open the folder

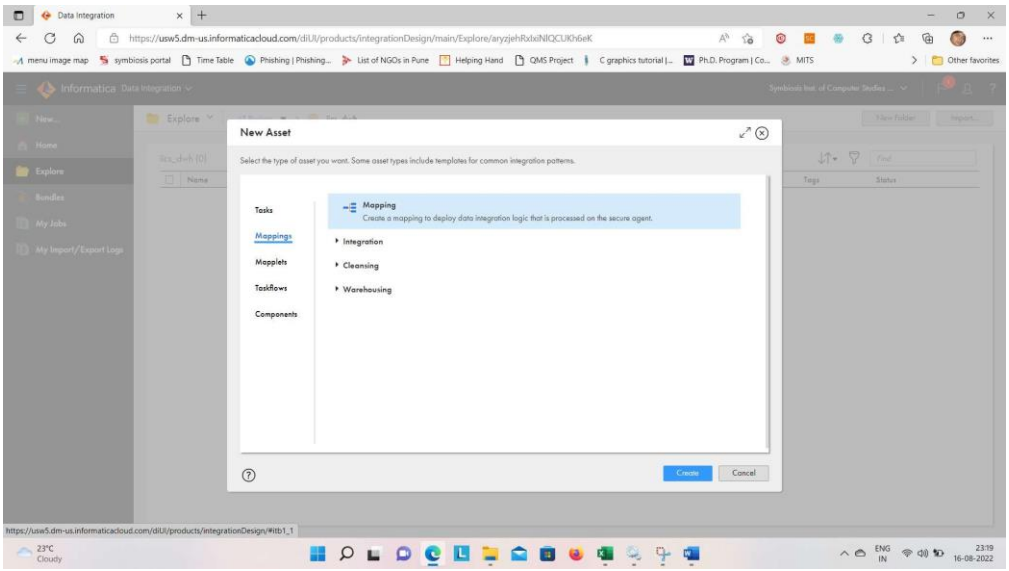


Click on new

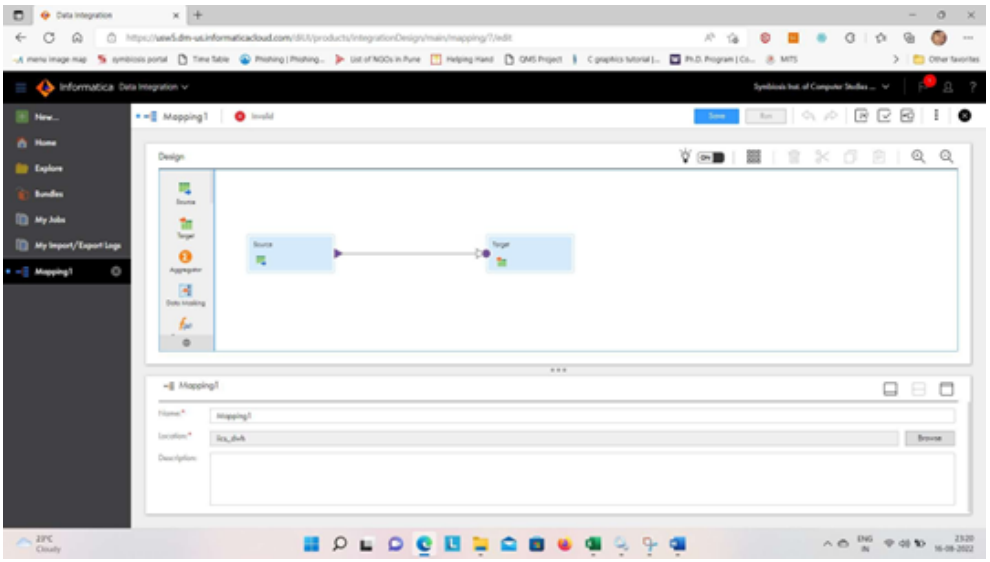


Select mapping and then click on create.

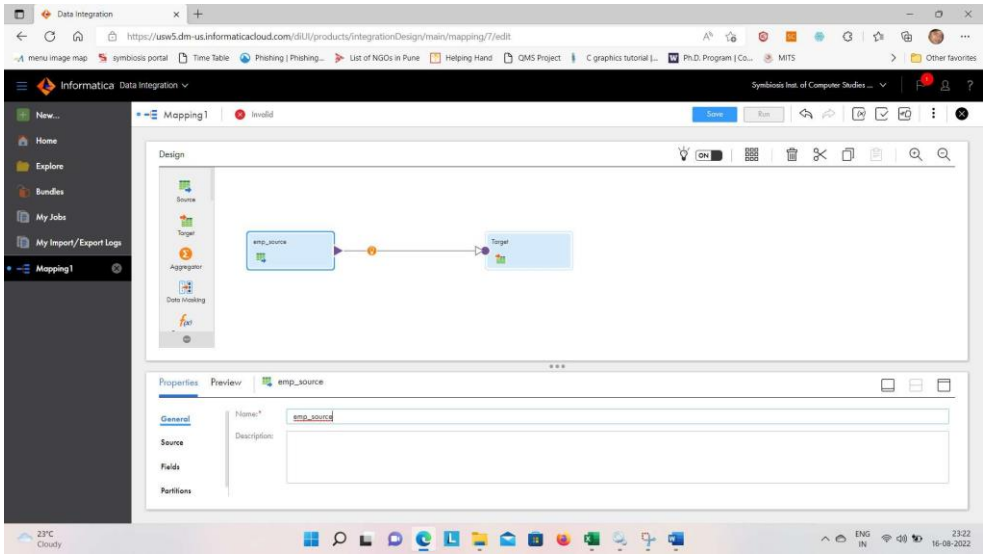




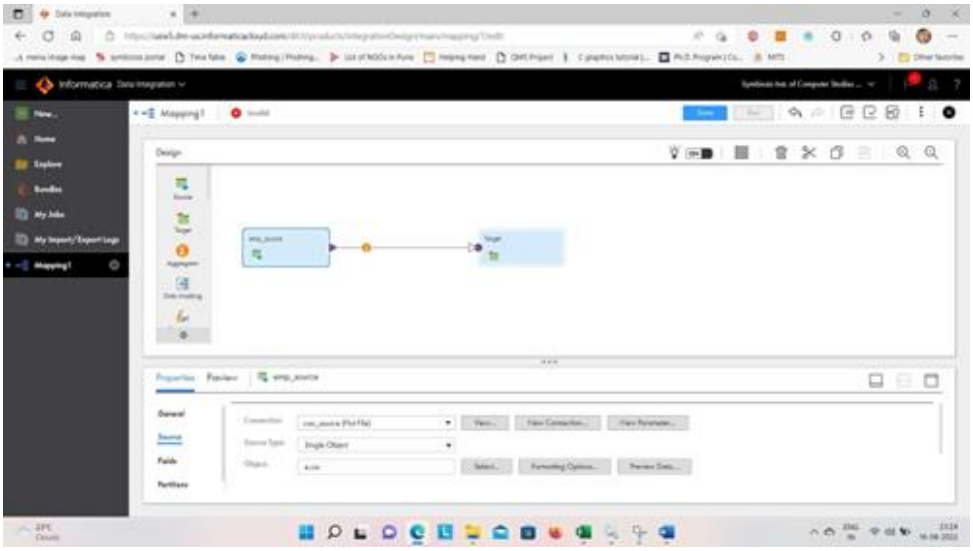
Provide mapping name



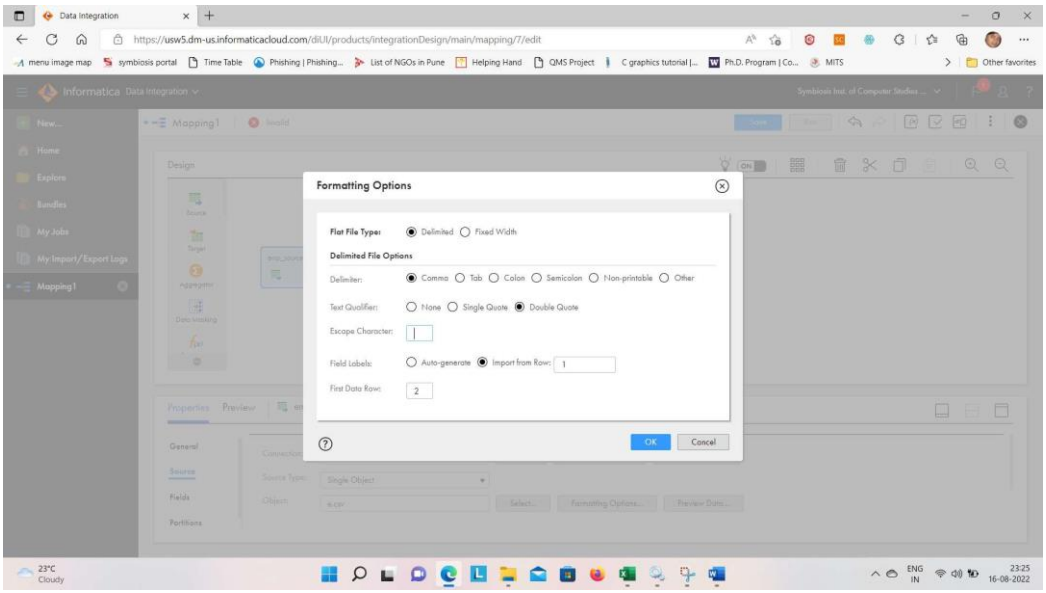
Click on the source box. In general, provide the name of the source



Click on the source and provide the connection name, connection type, and source file.

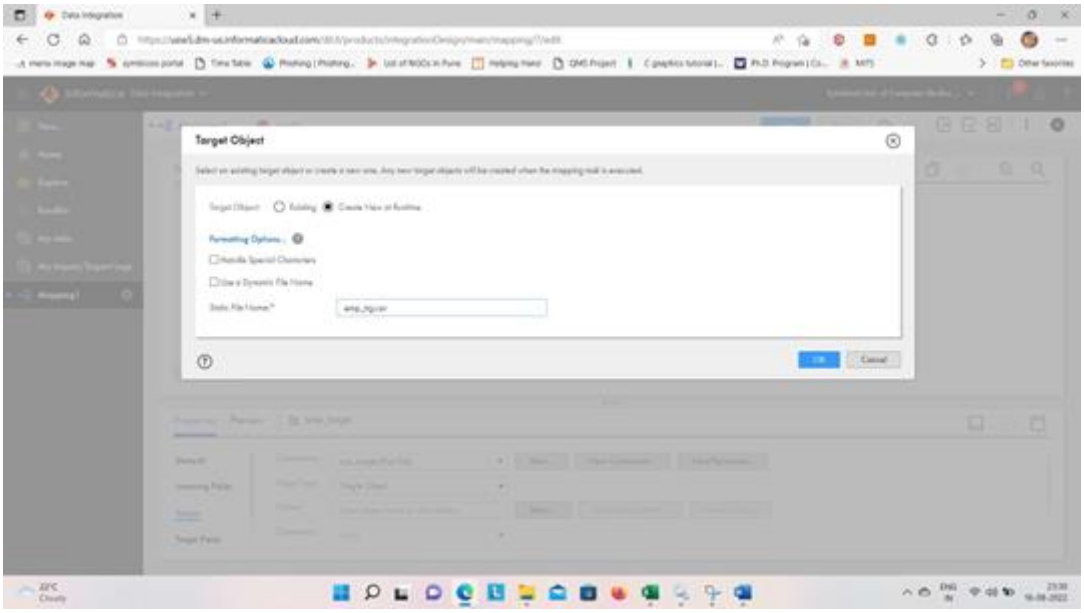


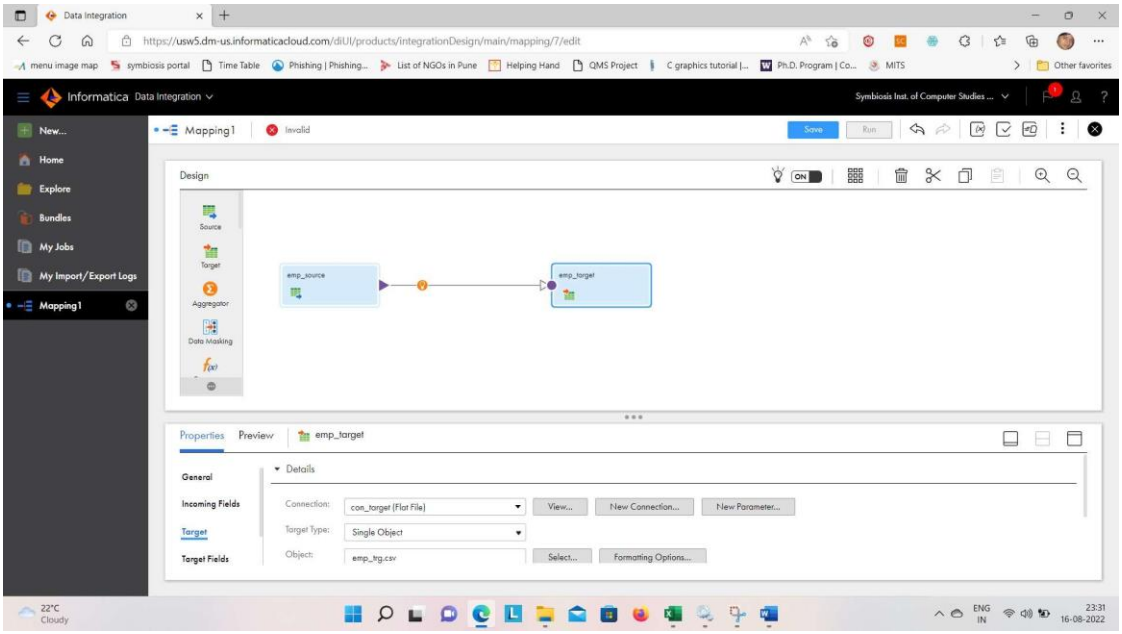
Go to the formatting option and select the properties of the file as per your file schema.



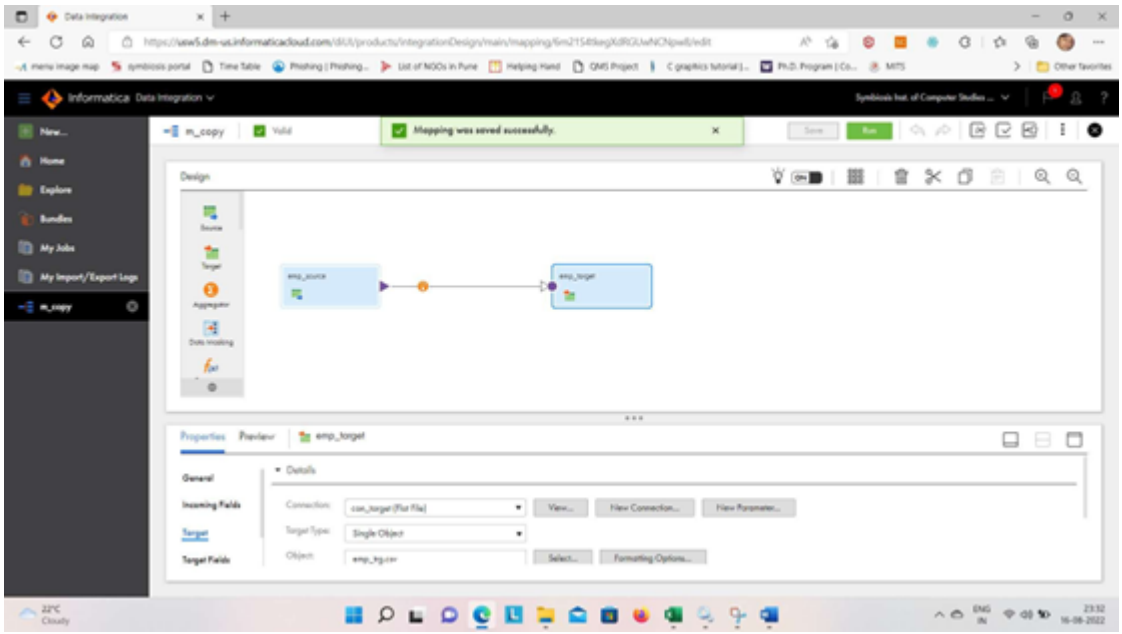
Perform the same steps for the target.

Note: If the target structure is as same as the incoming fields that get passed to the target, then create a run time option is used to create the target but for a different structure a target schema needs to beput in the target folder, and it has to be imported as the target.





For create run-time target option, no field mapping is required.



Save your mapping.

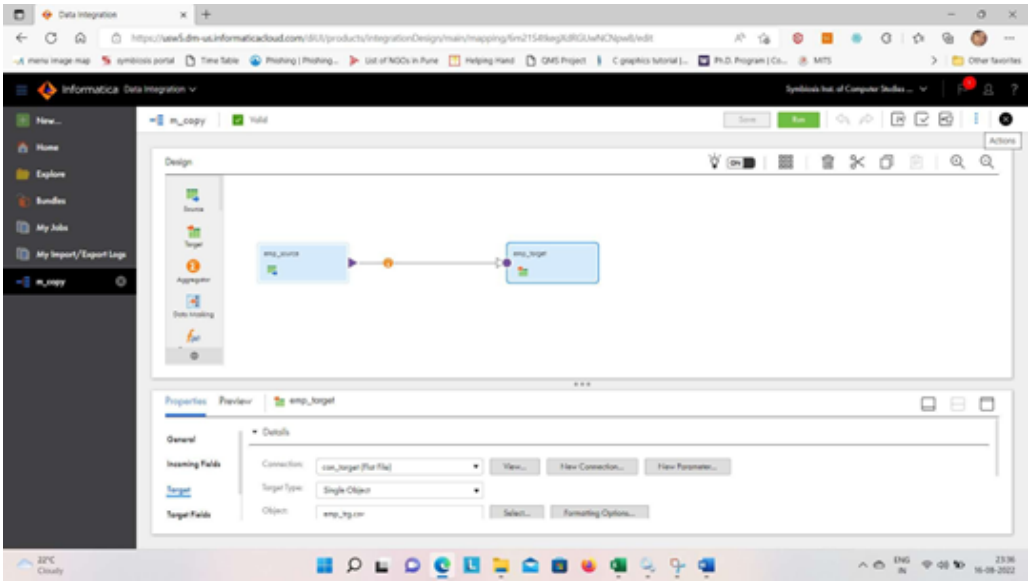
Steps to create mapping task

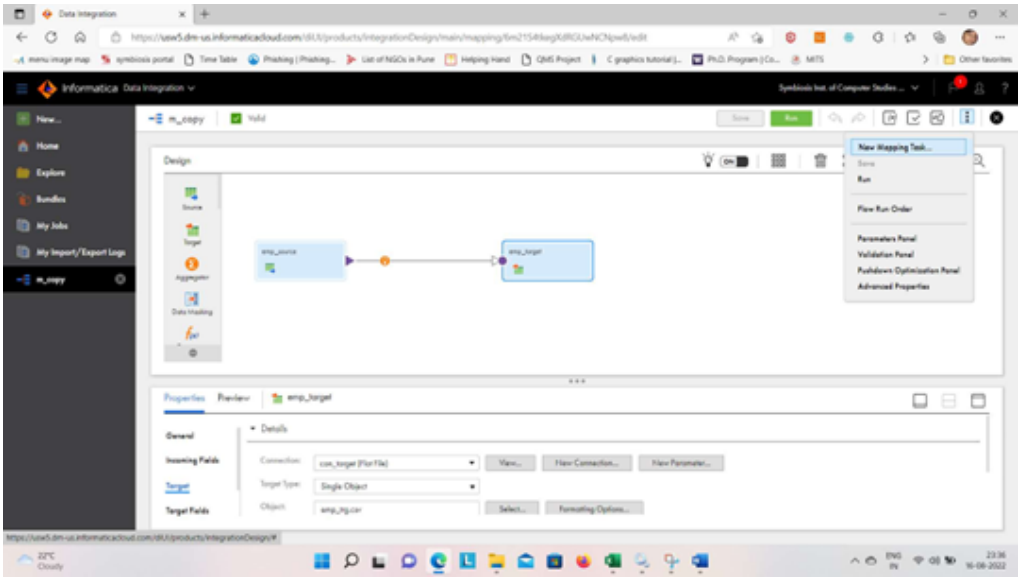
A mapping task is created for each mapping in iics. It is required to set additional properties that can't be set in mapping.

For example- send success and failure email in case of success or fail of job.Execute some pre and post commands.

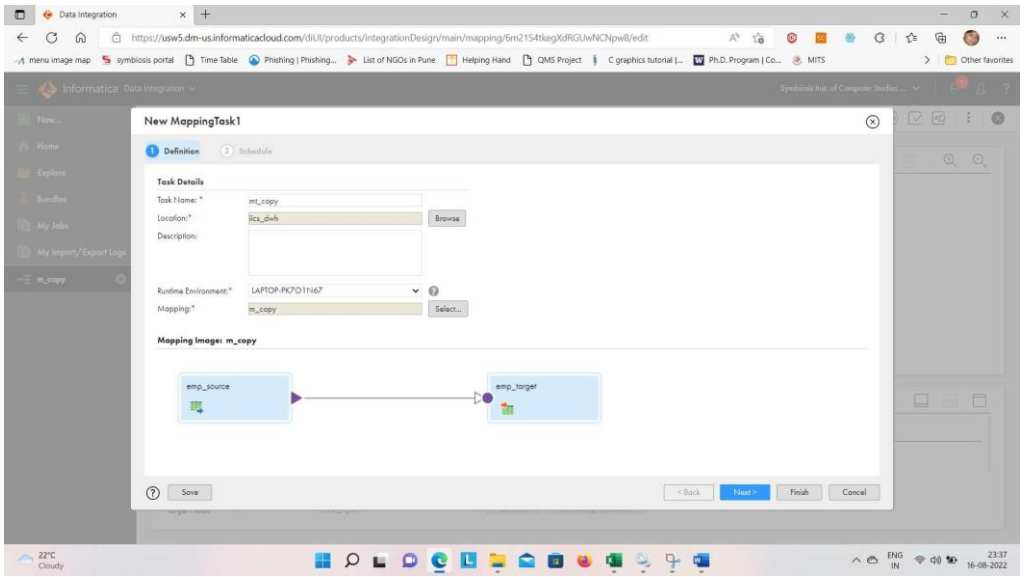
To set memory related properties, error handling properties etc.

Click on three dots available in right hand upper corner.

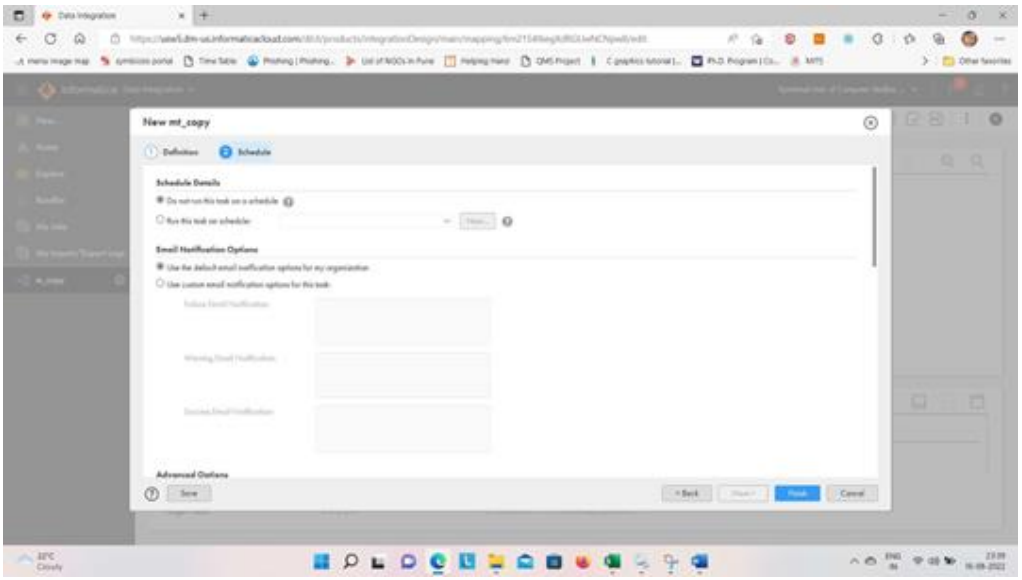


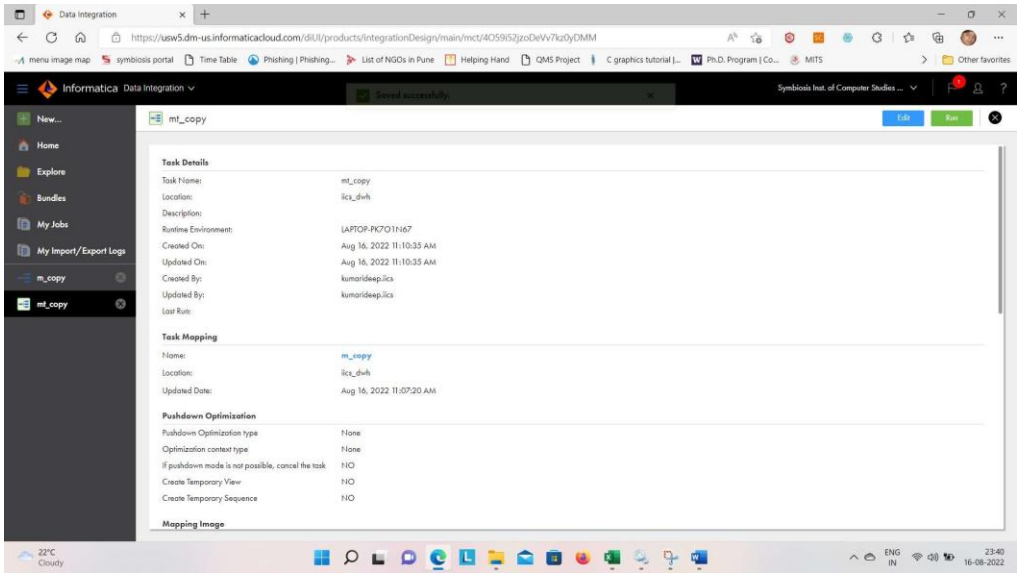


Set the values for attributes and click on next

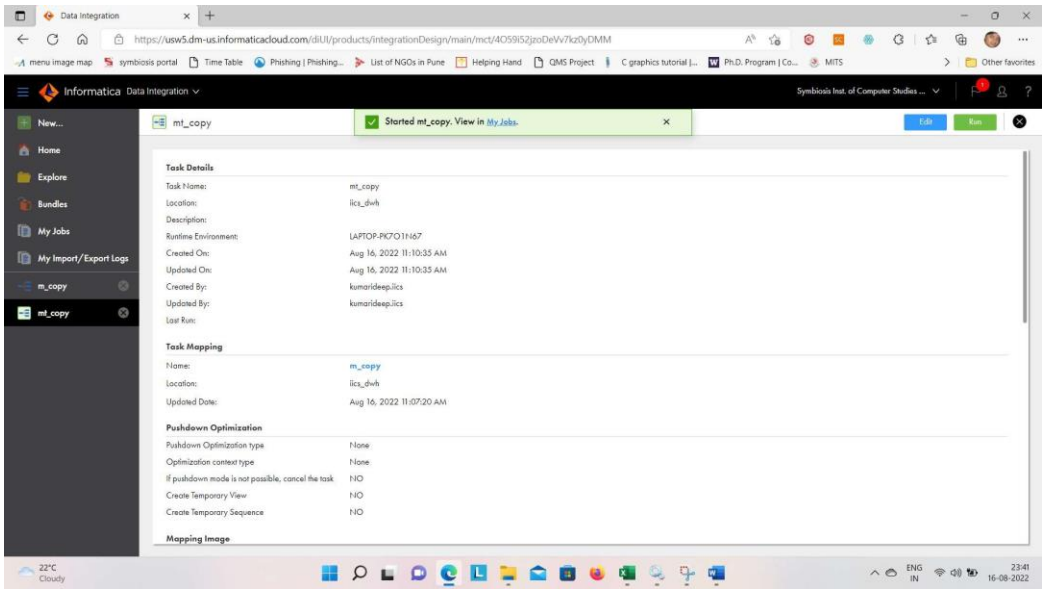


Set your desired properties and click finish.

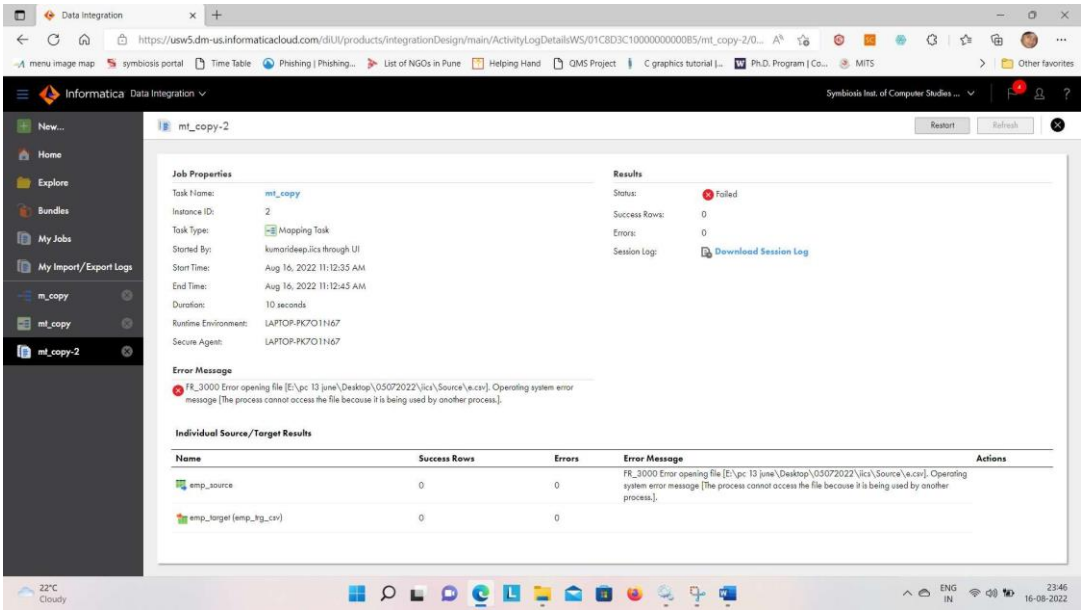




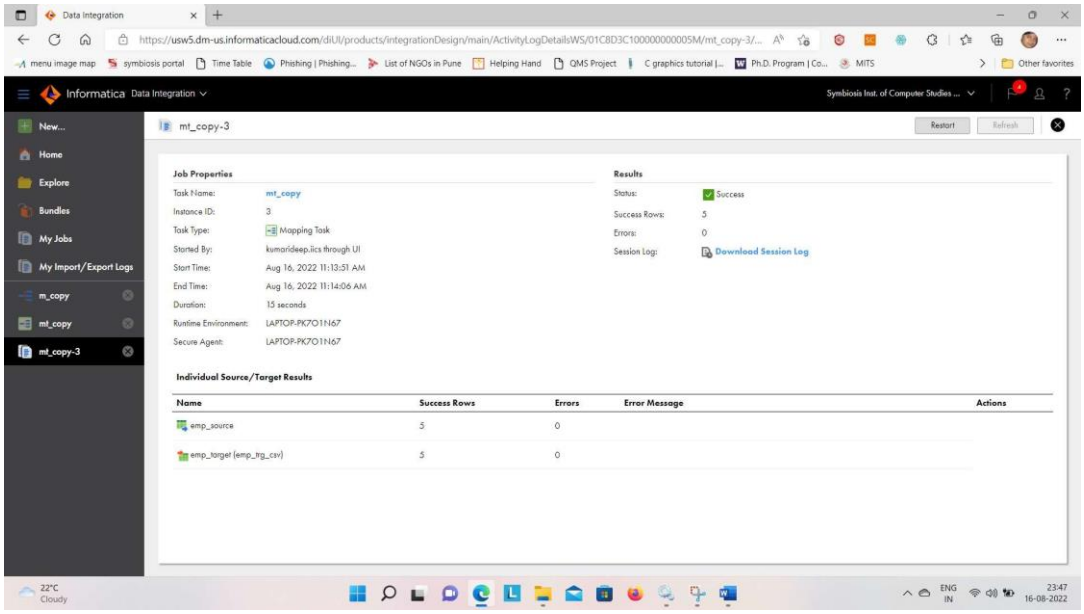
Run the mapping task and switch on to my jobs to monitor the execution



Mapping execution may result in either success or failure. In case of failure, the error is checked andto be resolved by the developer by making required changes.



Success scenario



Validate data in target folder

e_id	e_first_na	e_last_nar	e_sal	d_no
100	Shailja	Pant	5000	10
101	Shruti	Baradwaj	10000	20
102	Jyoti	Pant	7000	10
103	Shilpi	Chitragupt	9000	20
104	Mansi	Mathur	15000	30

source

e_id	e_first_na	e_last_nar	e_sal	d_no
100	Shailja	Pant	5000	10
101	Shruti	Baradwaj	10000	20
102	Jyoti	Pant	7000	10
103	Shilpi	Chitragupt	9000	20
104	Mansi	Mathur	15000	30

Target

Source and target data is being matched then data validation for copy operation is passed.

Transformations in Informatica

Transformations are the objects which are used in ETL mappings to implement the required transformations of data as per business and functional requirements. On the basis of changing the number and type of records, Transformations are classified into two categories.

1. **Active Transformation**: An active transformation changes the number of rows that pass through a transformation. Or, it changes the row type.For example, the Filter transformation is active, because it removes rows that do not meet the filter condition. The Update Strategy transformation is active, because it flags rows for insert, delete, update, or reject.
2. **Passive Transformation:** A passive transformation does not change the number of rowsthat pass through the transformation, maintains the transaction boundary, and maintains the row type.

On the basis of connectivity, Transformations are classified into two categories.

1. **Connected Transformation:** A connected transformation is connected to either the other transformation or the target.
2. **Unconnected Transformation:** An unconnected transformation is not connected to other transformations in the mapping. An unconnected transformation is called within another transformation, and returns a value to that transformation.

Different Transformations

1. Expression Transformation:

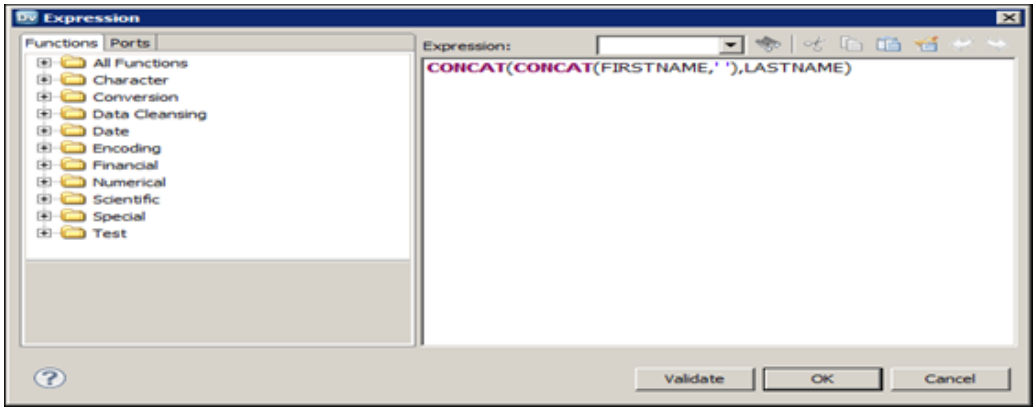


Figure of Expression editor

The Expression Transformation is a passive and connected transformation. It is used to perform non aggregate calculations. Row level functions and operators such as concatenation, multiplication e.tc. are used within Expression to derive new columns with modified data (Transformation Guide, 2011).

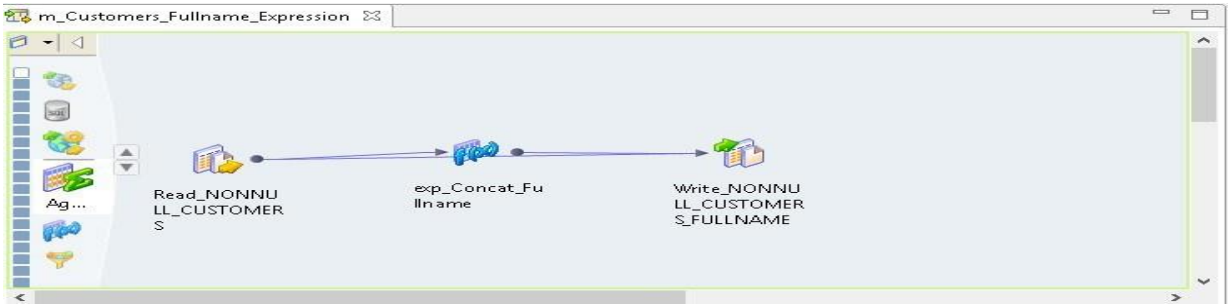


Figure of Expression Transformation in Mapping

Aggregator Transformation:

The Aggregator Transformation is an active and connected transformation that is used to perform aggregate calculations such as SUM, MAX, MIN, AVG, FIRST, and LAST e.t.c. on a group data. An aggregate expression can include conditional clauses and non-aggregate functions. It can also include one aggregate function nested within another aggregate function.

The Aggregator transformation is used to perform calculations on groups. The Expression transformation is used to perform calculations on a row-by-row basis only (Transformation Guide, 2011).

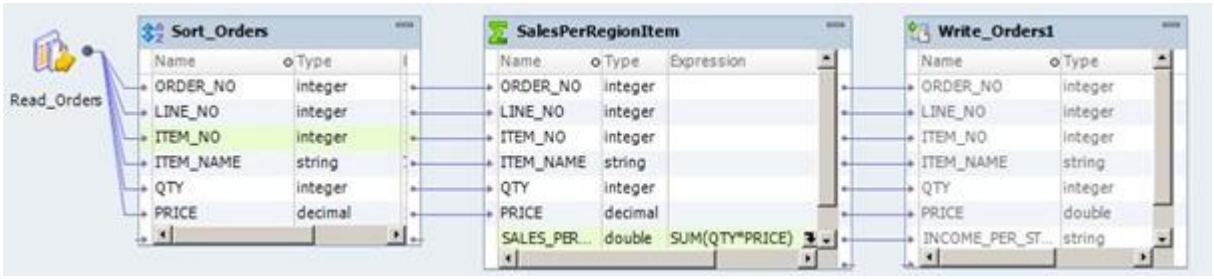


Figure of Aggregator Transformation in Mapping

Filter Transformation:

The Filter transformation is an active and connected transformation. The Filter transformation is used to filter out rows in a mapping. As an active transformation, the Filter transformation may change the number of rows passed through it.

The Filter transformation allows rows that meet the specified filter condition to pass through. Itdrops rows that do not meet the condition (Transformation Guide, 2011).

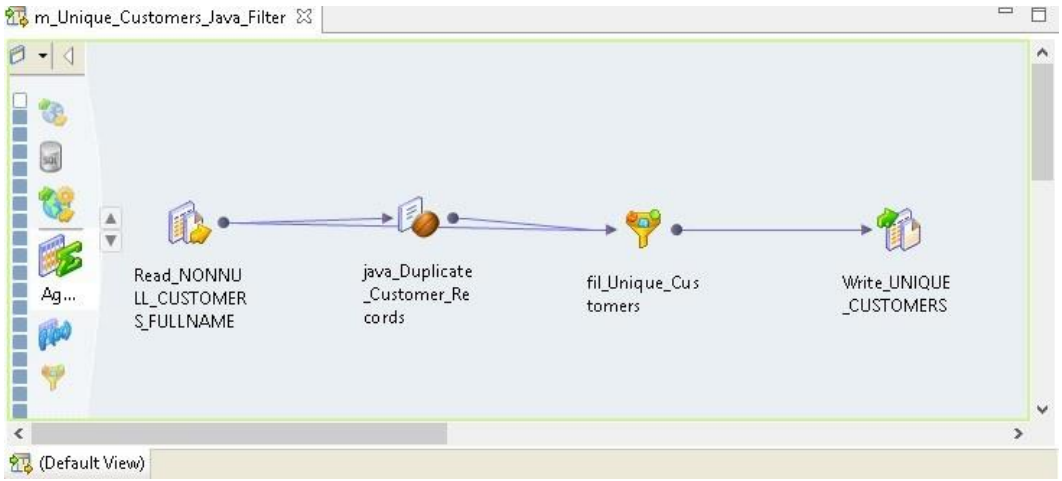


Figure of Filter Transformation in Mapping

Router Transformation:

The Router Transformation is an active and connected transformation. The router is equivalent to multiple filter transformations. It routes data into multiple output groups based on one or more conditions. Route the output groups to different transformations or to different targets inthe mapping.

A Router transformation is similar to a Filter transformation because both transformations use acondition to test data. A Filter transformation tests data for one condition and drops the rows of data that do not meet the condition. A Router transformation tests data for one or more conditions and can route rows of data that do not meet any of the conditions to a default output group (Transformation Guide, 2011).

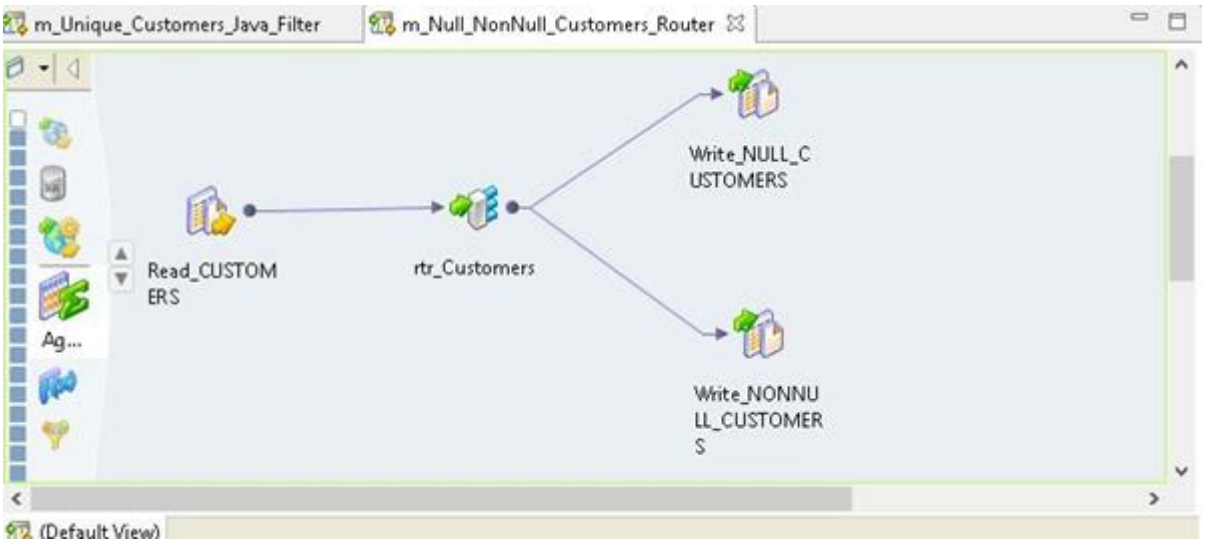


Figure of Router Transformation in Mapping

Lookup Transformation:

Lookup transformation is a passive transformation. It can be connected as well as unconnected. The PowerCenter Server queries the lookup source based on the lookup ports in the transformation. It compares Lookup transformation port values to lookup source column values based on the lookup condition. Pass the result of the lookup to other transformations and a target. The Unconnected Lookup transformation is not connected to the mapping flow, it is called from the other transformation through an expression port just like a function in programming languages. Unconnected lookup transformations return a value to the calling expression. If the same lookup is to be used for multiple times, unconnected lookup transformation is preferred (Transformation Guide, 2011).

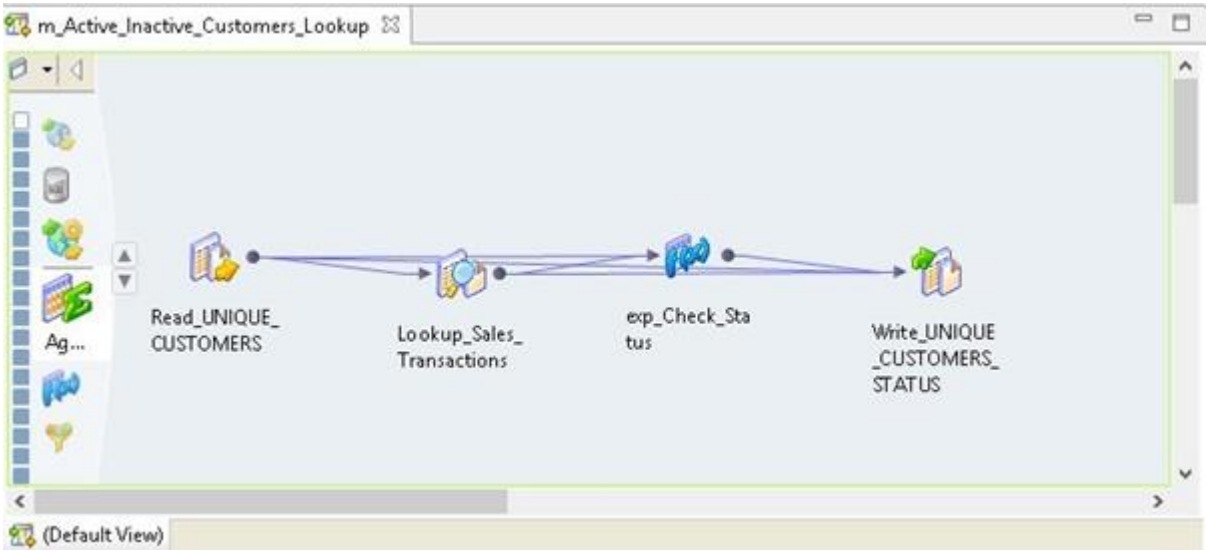


Figure of Lookup Transformation in Mapping

Joiner Transformation:

The Joiner Transformation is an active and connected transformation. The joiner transformation provides us the options to perform join operations in Informatica. The joins are created using Joiner transformation is similar to joins performed in the database. There are two different sources in Joiner transformation, one is called master and the other is called detail.

There are four types of joins can be performed using joiner transformation.

- Normal Join
- Master outer Join
- Detail outer Join
- Full outer Join

At least one join condition needs to be mentioned. On the basis of join conditions, join operation is performed and transformed data are produced (Transformation Guide, 2011).

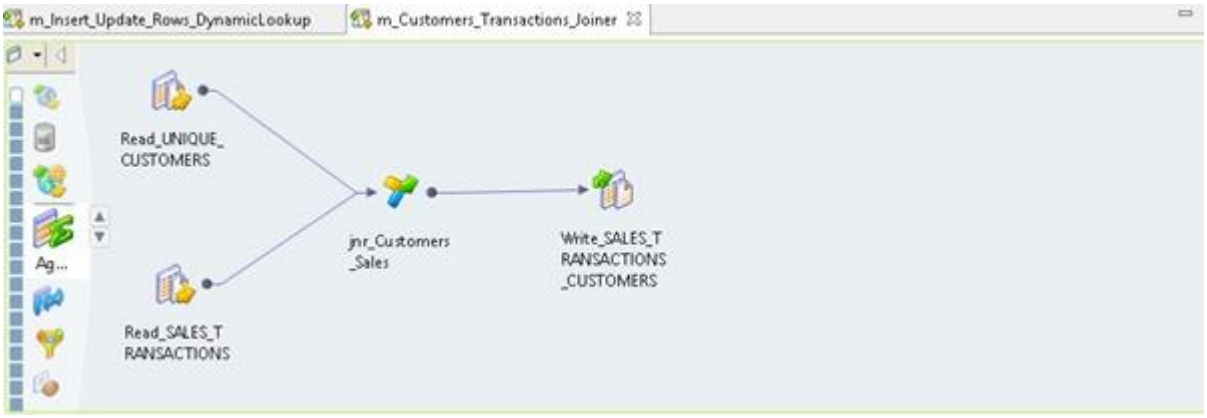


Figure of Joiner Transformation in Mapping

Rank Transformation:

The Rank Transformation is an active and connected transformation. It is used to find top or bottom n ranked values. Rank transformation is performed on a group, if no group option is provided, it considers the entire dataset as one group and returns rows having a top or bottom n-ranked values (Transformation Guide, 2011).

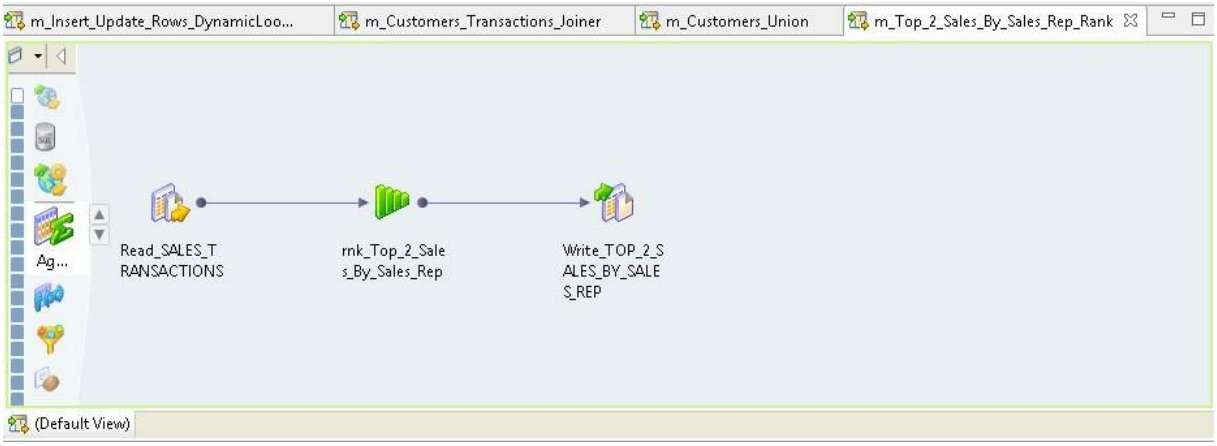


Figure of Rank Transformation in mapping

Sequence Geneartor Transformation:

The Sequence Geneartor Transformation is a passive and connected transformation. It is used to generate numbers in a sequence like 1,2,3, and so on. Whenever we need to generate numbers in a sequence for each and every row, we use Sequence Generator transformation in Informatica (Transformation Guide, 2011).

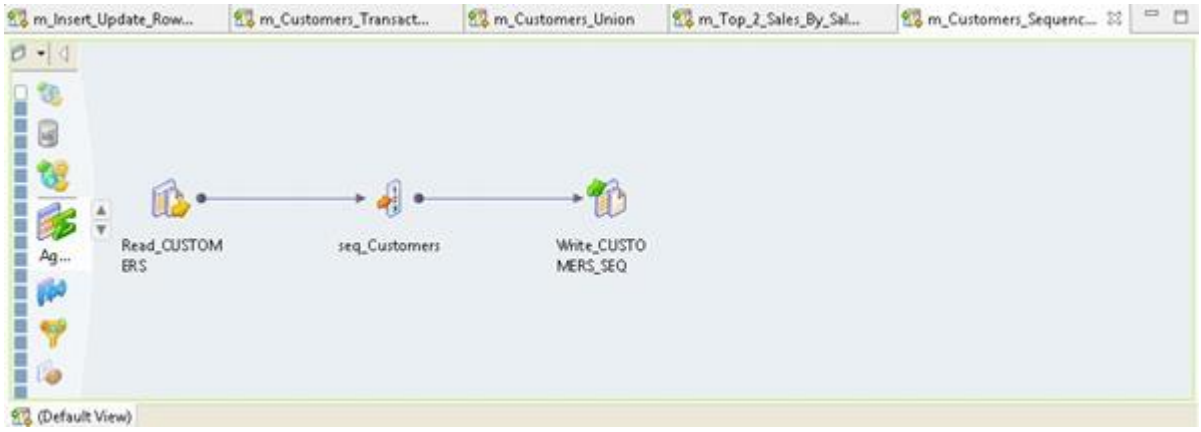


Figure of Sequence Generator Transformation in mapping

Sorter Transformation:

Sorter transformation is an active and connected transformation used to sort the data. It sorts data in ascending or descending order based on specifying the sort key. We can specify one more sort key and configure each sort key port to sort either in ascending or descending order. We can also configure the order of ports in which the sort operation is to be performed.

The sorter transformation is used to sort numbers and case sensitive data as well, it can also remove duplicate data by checking the distinct option (Transformation Guide, 2011).

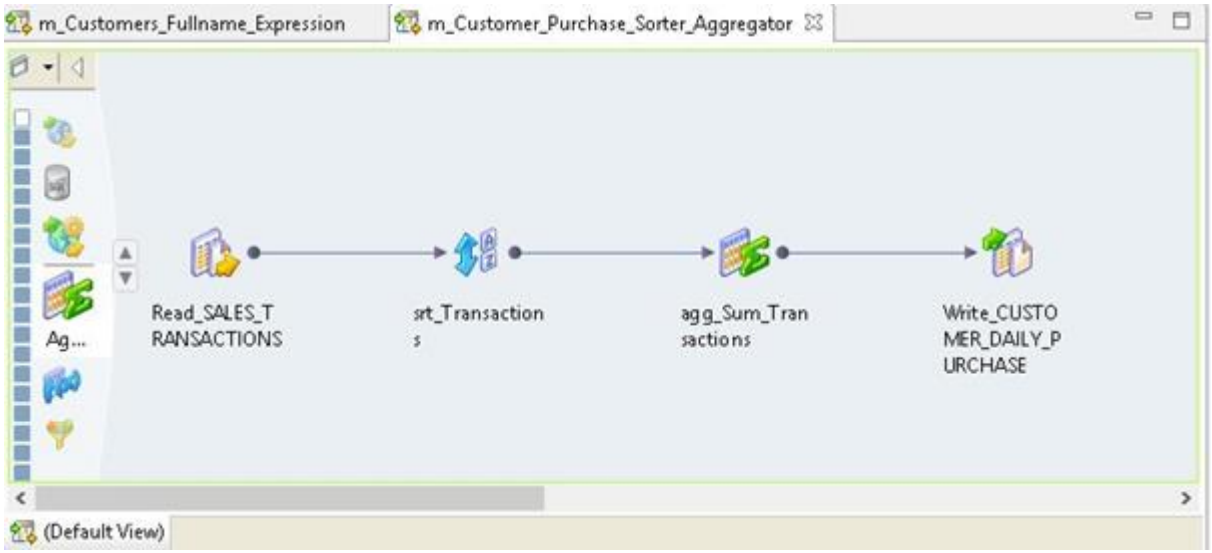


Figure of Sorter Transformation in Mapping

Union Transformation:

The Union Transformation is an active and connected transformation. It is a multi group transformation. It acts like Union All operations performed at the database level. There are multiple input groups created during Union transformation that can receive data from multiple sources and make a unified output data set. To be processed in Union transformation, every source has the

compatible data structure, i.e. number of ports/columns and their data types must be in the same format (Transformation Guide, 2011).

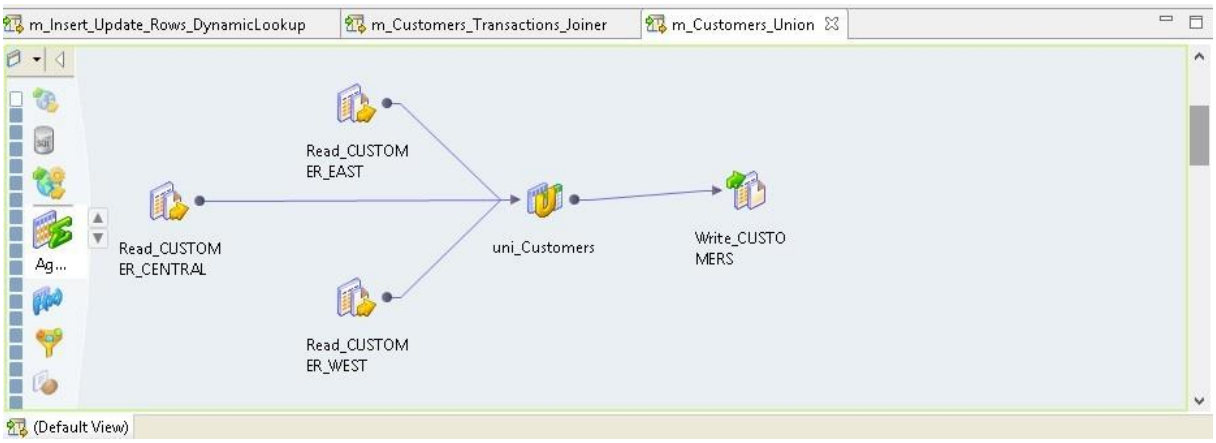


Figure of Union Transformation in Mapping

Update Strategy Transformation:

Update Strategy Transformation is an active and connected transformation. Operations other than insert like update and delete can be performed using update strategy transformations. If there is a requirement in which new records have to be inserted and existing records have to update then Update strategy transformation with lookup transformation is used to flag records for insert and update operations. Delete and reject operations can also be performed using update strategy transformation (Transformation Guide, 2011).

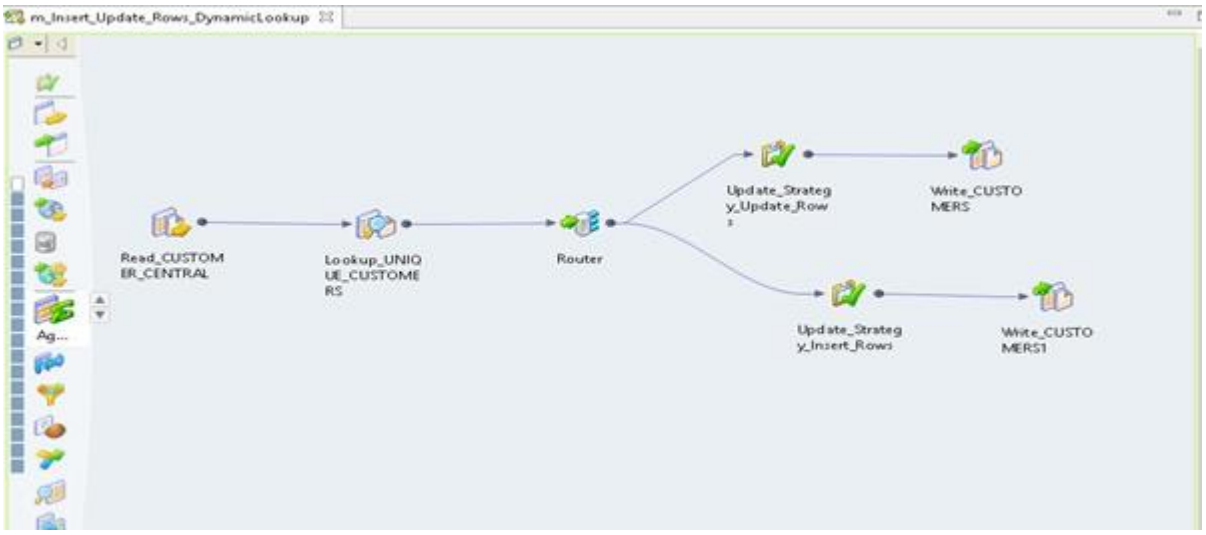


Figure of Update Strategy Transformation in Mapping

Mapping Creation in Informatica:

Mapping: A mapping is a set of source and target definitions linked to transformation objects that defines the rules for data transformations. Mappings represent the data flow between sources and

targets. When the server runs a session, it uses the instructions configured in the mapping to read, transform, and write data.

Mapping is created in **a Mapping Developer tool** under Informatica PowerCenter Designer. Every mapping must contain the following components.

- **Source definition:** Describe the characteristics of source table or file.
- **Transformation:** Modifies data before writing it to target. Use different transformation objects to perform different functions.
- **Target definition:** Defines the target table or file.
- **Links:** Connects sources, targets, and transformations so the server can move the dataas it transforms it

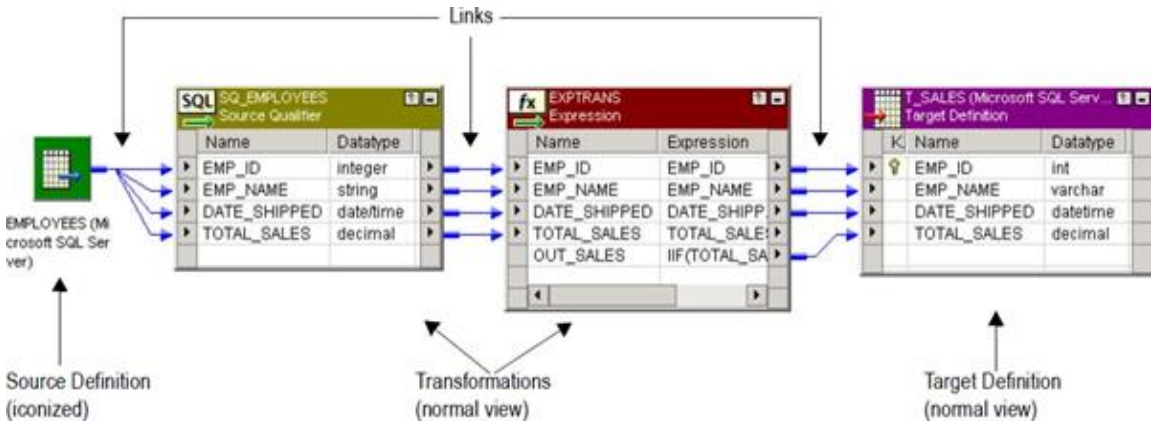


Figure of Sample Mapping Snapshot

Reporting in DWH:

In the previous sections we have studied on DWH concepts and different aspects of ETL processing. So once our DWH storage system is built and the ETL Jobs are ready to populate data in DWH tables then the next step is to develop a reporting solution for the DWH to deliver the overall DWH applications to the client so the need of reporting is to provide information to the business users in some graphical representation format for their strategic decision making. As a matter of fact, it is the reporting layer that gives the actual Return of Investment (ROI) to the user because users see data in the form of reports and dashboards. The more efficient reporting solutions we delivered to the user, the more satisfied users we have. A perfectly built DWH with maintaining high quality of data, proper ETL solution in place will be just worthless for the users or the whole organizations. If a good reporting solution is not provided with them.

Reports and dashboards are developed by using front end tools in this step. User needs different types of reports to support their different decision activities. So various reporting tools are available in the market to cater the reporting needs of the end user.

Reporting tool is used to produce reports based on data, as well as applying various operations such as filter, aggregations, sorting, rank etc. and the output format of the result. OLAP

tools provides users the intuitive way to view organizations data. Users can navigate through hierarchies and dimensions with click of mouse buttons. It also facilitates users to drill down, drill up levels in each dimension to change their view of data. Pre aggregate data is stored in multidimensional databases.

Data mining tools are hot commodities nowadays. It uses a variety of statistical and AI algorithms to analyse data in a various way. Reporting tools are easy to use as they provide point and click tools that generate SQL queries in background to query relational data stored in the DWH. It also makes easy for users to read reports of the retrieved data.

There are several types of reports created as below:

Standard and static reports-these are the reports that report designer defines with fixed layout when it is created. They are generated on request made by the end user, scheduled to refresh periodically, and made available to the end user on the web server

1. **Ad-hoc reports-** Ad-hoc reports are created on demand by the users. They are designed and developed either from scratch or using the template of other standard reports.
2. **OLAP reports-** These are the fixed design reports designed and developed by the report designer. Various operations like slicing, dicing, pivot or rotation, drill-up/roll-up, drill down can be performed to extract more general information. It can be generated either on periodical refresh or a request made by an end user.
3. **Dashboards-** Dashboard is the graphical representation of the charts, maps and different indicators. It can be either static or interactive reports. It contains a number of graphs and charts, high level aggregated data, comparisons of the business strategic data and the performance indicators. The components of dashboard is refreshed periodically on specified time intervals.

Reporting tools:

There are several BI reporting, dashboard and data visualization tools available in the market. They are as below:

1. Oracle BI OBIEE
2. Tableau Business Intelligence
3. IBM Cognos
4. SAS Business Intelligence
5. Pentaho
6. SAP Business Objects Web Intelligence (Webl)
7. SAP BW
8. Microsoft BI Platform
9. SAP Lumira
10. Jasper Soft
11. Dashboard Designer
12. Crystal Reports
13. Tibco Spotfire
14. Qlick View
15. Necto

REFERENCES:

Agarwal, S., D, T. M., & Tiwari, I. (2013a). Foundation of data, data warehousing, data mining and e governance (pp. 11–30). River Publishers.

Agarwal, S., D, T. M., & Tiwari, I. (2013b). Roadmap to e governance data management, data center, data warehousing and data mining (pp. 1–10). River Publishers.

Amanzougarene, F., Chachoua, M., & Zeitouni, K. (2016). A new approach of imprecision management in qualitative data warehouse (W. Shi, B. Wu, & A. Stein, Eds.; pp. 101– 118). CRC Press-Taylor & Francis Group.

Amin, W., Kang, H. P., & Becich, M. J. (2010). Data management, databases, and warehousing (M. Ochs, J. Casagrande, & R. Davuluri, Eds.; pp. 39–71). SPRINGER.

https://doi.org/10.1007/978-1-4419-5714-6_3

Anonymous. (1998). Data warehousing and data mining for telecommunications. Database, 21, 94.

Banerjee, S., & Davis, K. C. (2009). Modeling data warehouse schema evolution over extended hierarchy semantics (S. Spaccapietra, E. Zimanyi, & I. Song, Eds.; Vol. 5530, pp. 72–96). Springer-Verlag Berlin.

Bentayeb, F., Loudcher, S., Maiz, N., Harbi, N., Mahboubi, H., Boussaid, O., Favre, C., & Darmont, J. (2012). Innovative approaches for efficiently warehousing complex data from the web (pp. 26–52). IGI GLOBAL. https://doi.org/10.4018/978-1-61350-038-5.ch002

Biehl, R. E. (2016). Biomedical data warehousing (pp. 1–14). CRC Press-Taylor & Francis Group.

Combi, C., Oliboni, B., & Pozzi, G. (2009). Modeling and querying temporal semistructured data warehouses (S. Kozielski & R. Wrembel, Eds.; Vol. 3, pp. 299–323). Springer.

Conventional data warehouses (pp. 75–131). (2008). Springer-Verlag Berlin.

Designing conventional data warehouses (pp. 245–306). (2008). Springer-Verlag Berlin.

F, V. R. (2012). Data virtualization for business intelligence systems: Revolutionizing data integration for data warehouses (pp. 1–275). Elsevier Science BV.

Fasel, D., & Shahzad, K. (2012). Fuzzy data warehouse for performance analysis (pp. 217–251). IGI Global. https://doi.org/10.4018/978-1-4666-0095-9.ch010

Gillespie, T. (1996). The data warehouse toolkit: Practical techniques for building dimension data warehouse - Kimball, R. Library Journal, 121, 102.

Gillespie, T. (1998). Oracle 8 data warehousing. Library Journal, 123, 126.

Golfarelli, M., & Rizzi, S. (2013). Data warehouse testing (pp. 91–108). IGI Global. https://doi.org/10.4018/978-1-4666-2148-0.ch005

Gray, P. (1997a). Developing the data warehouse. Information Systems Management, 14, 82–86.

Gray, P. (1997b). The data warehouse toolkit: Practical techniques for building dimensional data warehouses. Information Systems Management, 14, 82–86.

Gray, P. (1997c). Using the data warehouse. Information Systems Management, 14, 82–86.

Gupta, P. (2019). Data warehousing, data mining, & OLAP. JIMS8M-the Journal of Indian Management & Strategy, 24, 64.

Harrington, J. L. (2009). Data warehousing (pp. 351–361). Elsevier Science BV.

Homayouni, H., Ghosh, S., & Ray, I. (2019). Data warehouse testing (A. Memon, Ed.; Vol. 112, pp. 223–273). Elsevier Academic Press Inc. https://doi.org/10.1016/bs.adcom.2017.12.005

Kozielski, S., & Wrembel, R. (Eds.). (2009). New trends in data warehousing and data analysis (Vol. 3, pp. 1–345). SPRINGER. https://doi.org/10.1007/978-0-387-87431-9

Krishnan, K. (2013a). Data warehousing in the age of big data (pp. 1–346). Morgan Kaufmann Pub Inc.

Krishnan, K. (2013b). Data warehousing in the age of big data introduction (pp. XIX–XXIII). Morgan Kaufmann Pub Inc.

Krishnan, K. (2013c). Data warehousing revisited (pp. 127–145). Morgan Kaufmann Pub Inc.

Krishnan, K. (2013d). Integration of big data and data warehousing (pp. 199–217). Morgan Kaufmann Pub Inc.

Marotta, A., Ruggia, R., Gonzalez, L., Serra, F., Etcheverry, L., Martirena, E., & Rienzi, B. (2012). Quality management in web warehouses (pp. 1–25). IGI Global. https://doi.org/10.4018/978-1-61350-038-5.ch001

Mason, D. (2003). Data warehousing and web engineering. Electronic Library, 21, 615. https://doi.org/10.1108/02640470310509216

Morzy, T. (2007). Temporal semistructured data models and data warehouses (pp. 277–297). Idea Group Publishing.

Nebot, V., Berlanga, R., Perez, M., Aramburu, J., & Pedersen, T. B. (2009). Multidimensional integrated ontologies: A framework for designing semantic data warehouses (S. Spaccapietra, E. Zimanyi, & I. Song, Eds.; Vol. 5530, pp. 1–36). Springer-Verlag Berlin.

Pedersen, T. B. (2009). Warehousing the world: A vision for data warehouse research (S. Kozielski & R. Wrembel, Eds.; Vol. 3, pp. 1–17). Springer.

Pedersen, T. B. (2014a). Data analytics: Exploiting the data warehouse (pp. 329–383). Springer-Verlag Berlin. https://doi.org/10.1007/978-3-642-54655-6_9

Pedersen, T. B. (2014b). Logical data warehouse design (pp. 121–178). Springer-Verlag Berlin. https://doi.org/10.1007/978-3-642-54655-6_5

Pedersen, T. B. (2014c). Physical data warehouse design (pp. 233–284). Springer-Verlag Berlin. https://doi.org/10.1007/978-3-642-54655-6_7

Pedersen, T. B. (2014d). Querying data warehouses (pp. 179–230). Springer-Verlag Berlin. https://doi.org/10.1007/978-3-642-54655-6_6

Pedersen, T. B. (2014e). Spatial data warehouses (pp. 427–473). Springer-Verlag Berlin. https://doi.org/10.1007/978-3-642-54655-6_11

Pedersen, T. B. (2014f). Trajectory data warehouses (pp. 475–506). Springer-Verlag Berlin. https://doi.org/10.1007/978-3-642-54655-6_12

Raffaeta, A., Leonardi, L., Marketos, G., Andrienko, G., Andrienko, N., Frentzos, E., Giatrakos, N., Orlando, S., Pelekis, N., Roncato, A., & Silvestri, C. (2013). Visual mobility analysis using t-warehouse (pp. 1–22). IGI GLOBAL. https://doi.org/10.4018/978-1-4666-2148-0.ch001

Rahman, N. (2012). Saving DBMS resources while running batch cycles in data warehouses (pp. 118–132). IGI GLOBAL. https://doi.org/10.4018/978-1-4666-1752-0.ch009

Silvers, F. (2012). Data warehouse ROI (pp. 1–11). CRC Press-Taylor & Francis Group.

Spatial data warehouses (pp. 133–179). (2008). Springer-Verlag Berlin.

Sumathi, S., & Esakkirajan, S. (2007). Data mining and data warehousing (Vol. 47, pp. 415– 475). Springer-Verlag Berlin.

Tabb, L. (2005). Data Mining and Data Warehouses (pp. 165–175). Blackwell Science Publ. Temporal data warehouses (pp. 181–243). (2008). Springer-Verlag Berlin.

ten Hompel, Michael, & Schmidt, T. (2008). Data model of a WMS the example of my WMS. (pp. 255–294). Springer-Verlag Berlin.

Tupper, C. D. (2011). Data Warehouses II (pp. 321–336). Morgan Kaufmann Pub Inc.

Vaisman, A. A., & Zimanyi, E. (2013). Trajectory data warehouses (C. Renso, S. Spaccapietra, & E. Zimanyi, Eds.; pp. 62–82). Cambridge Univ Press.

Zhao, Z., Li, J., Ye, Y., Liu, Y., & Liu, Y. (2016). The research and application of data warehouse's model design the data warehouse's model design for the decision support system of hospital drugs (J. Hung, N. Yen, & K. Li, Eds.; Vol. 375, pp. 337–351). Springer. https://doi.org/10.1007/978-981-10-0539-8_34

Zygmunt, A., Valenta, M. A., & Kmiecik, T. (2005). The specifics of dedicated data warehouse solutions (K. Zielinski & T. Szmuc, Eds.; Vol. 130, pp. 283–293). IOS Press.

Data Warehousing (ISBN: 978-93-88901-23-9)

About Authors

Er. Kumari Deepika is the first author of this book. Deepika did Bachelor of Engineering in Information Technology from Nagpur University. Thereafter she qualified GATE and completed Master of Technology in Computer Science and Technology from Central University of Punjab. During her MTech she worked on Performance Optimization on ETL process as dissertation research project. She has qualified UGC NET twice. Currently she is working as an Assistant Professor in Symbiosis Institute of Computer Studies and Research, Symbiosis International Deemed University, Pune, India. Along with her job, she is also pursuing part time PhD from J C Bose University of Science and Technology.

Dr. Santosh Chowhan is working as an Associate Professor at Department of Data Science and Analytics, School of Sciences, JAIN (Deemed-to-be University), JC Road, Bangalore-560027. He has a vast experience in the subject of Data Science. He has attended and presented his work in many conferences, seminars, workshops of National and International level.

Dr. Ujwala S Suryawanshi completed M.Sc. (CA) degree from Swami Ramanand Teerth University Campus, Nanded and pursued Ph.D. in 2020 in Computer Science from SRTM University, Nanded. The main topic of PhD is Brain Tissue Segmentation Using 3D Magnetic Imaging Resonance. She is presently working as Assistant Professor in Rajarshi Shahu Mahavidyalaya, Latur, and Maharashtra, India.

Dr. Shruti Bharadwaj is working as an Assistant Professor at United college of Engineering and Research, Prayagraj. U. P. She has decade of teaching experience. She has published several research papers in various reputed National & International journals. She attended and presented her research work in many conferences, seminars, workshops of National and International level.

Dr. Mahendra H Kondekar completed MCA degree in 1999 from Government Engineering College, Aurangabad and pursued Ph.D. in 2013 in computer science SRTM University, Nanded. The main topics of research are Biometrics Technology, Fuzzy & Neural network, and data analysis. He is presently working as Incharge Principal in Marathwada Institute of Technology, Cidco, Aurangabad, Maharashtra, India.

Er. Jyoti D Bhosale completed Diploma in IT, BE in IT, ME in Computer Networking, PGD in Disaster Management, Pursuing PhD. She is presently working as an Assistant Professor at Computer Engineering Department, VDF Group of Institution Latur (Maharashtra). She has 12 years of teaching experience. She has published several research papers in the field of Computer Engineering, Disaster Management field in various reputed National & International journals.

Bhumi Publishing

ISBN 978-93-88901-23-9
9 789388 901239