## DATA WAREHOUSING WITH IBM CLOUD DB2 WAREHOUSE

### Data warehouse

The definition of the data ware house focuses on the data storage .it means to retrieve and analyse the data to extract, transform and load data and to manage the dictionary are also considered essential components of a data warehousing system.

### lets understand data warehouse architecture

1.operational database layer

An organisation’s enterprise resource planning system fall into this layer

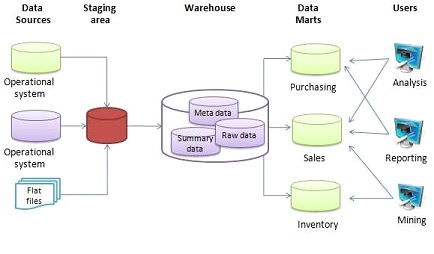
2.Data access layer

The interface between the operational and informational access layer tools to extract, transform, load data into the warehouse fall into this layer

3.Metadata layer

The data dictionary this is more detailed data about data stored

4.The data access for reporting and analysing and the tools for reporting and analysing data business intelligence tools fall into this category



### Lets see the objectives of data warehouse

1. The primary objectives of data warehouse system is to provide consolidated, flexible, meaningful data storage to the end user for reporting and analysing
2. Achieving a constant and efficient connection to the data source
3. Harnessing efficient data cleaning and loading technologies to the warehousing system this process is known as Data Transformation Services
4. Data storage is cleaned and stored data will have to be portioned summarized and stored for efficient query and analysis
5. Finally tools necessary for query, analysis and reporting on data

### Data Warehouse Design

A data warehouse is a single data repository where a record from multiple data sources is integrated for online business analytical processing (OLAP). This implies a data warehouse needs to meet the requirements from all the business stages within the entire organization.

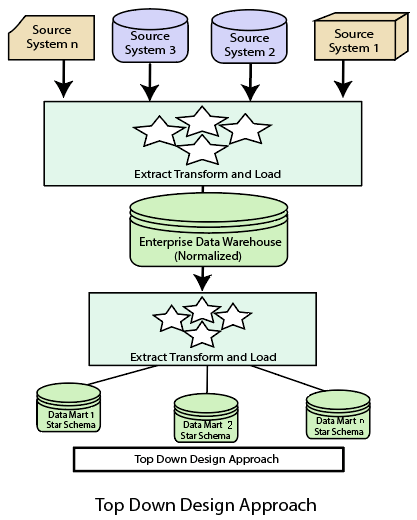
Data warehouse design takes a method different from view materialization in the industries. It sees data warehouses as database systems with particular needs such as answering management related queries.

There are two approaches

* top-down approach
* bottom-up approach

### Top-down Design Approach

In the "Top-Down" design approach, a data warehouse is described as a subject-oriented, time-variant, non-volatile and integrated data repository for the entire enterprise data from different sources are validated, reformatted and saved in a normalized (up to 3NF) database as the data warehouse.

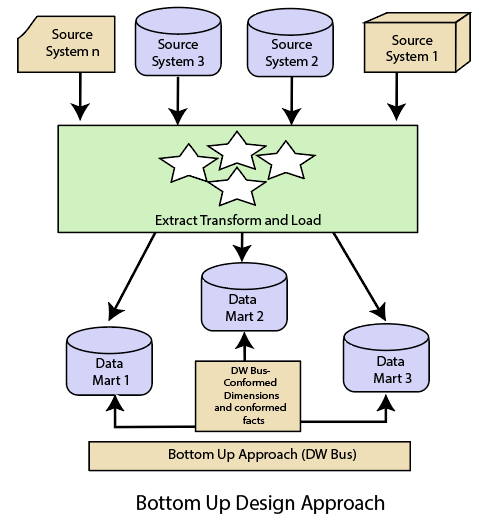


### ****Advantages of top-down design****

* Data Marts are loaded from the data warehouses.
* Developing new data mart from the data warehouse is very easy.

### Bottom-Up Design Approach

In the "Bottom-Up" approach, a data warehouse is described as "a copy of transaction data specifical architecture for query and analysis," term the star schema. In this approach, a data mart is created first to necessary reporting and analytical capabilities for particular business processes (or subjects).



### ****Advantages of bottom-up design****

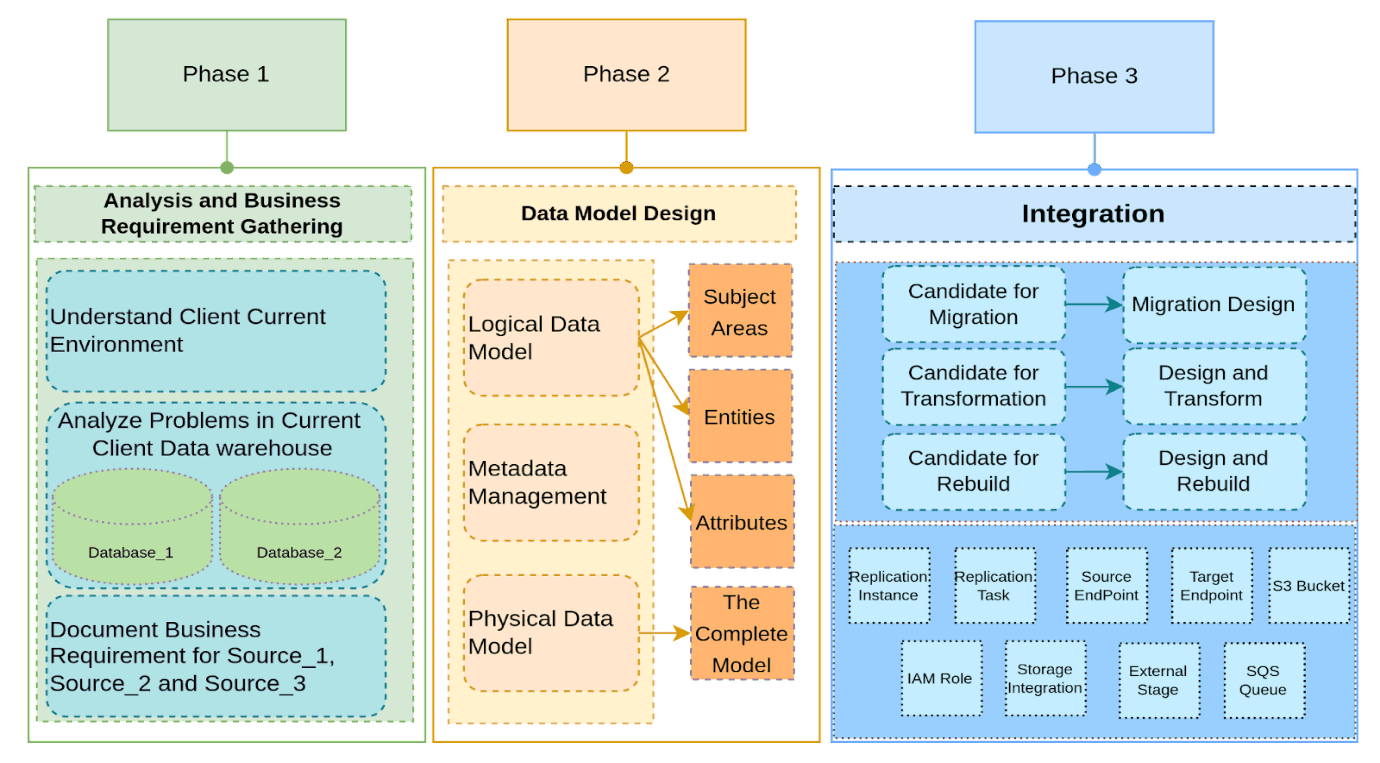
* Documents can be generated quickly.
* The data warehouse can be extended to accommodate new business units.

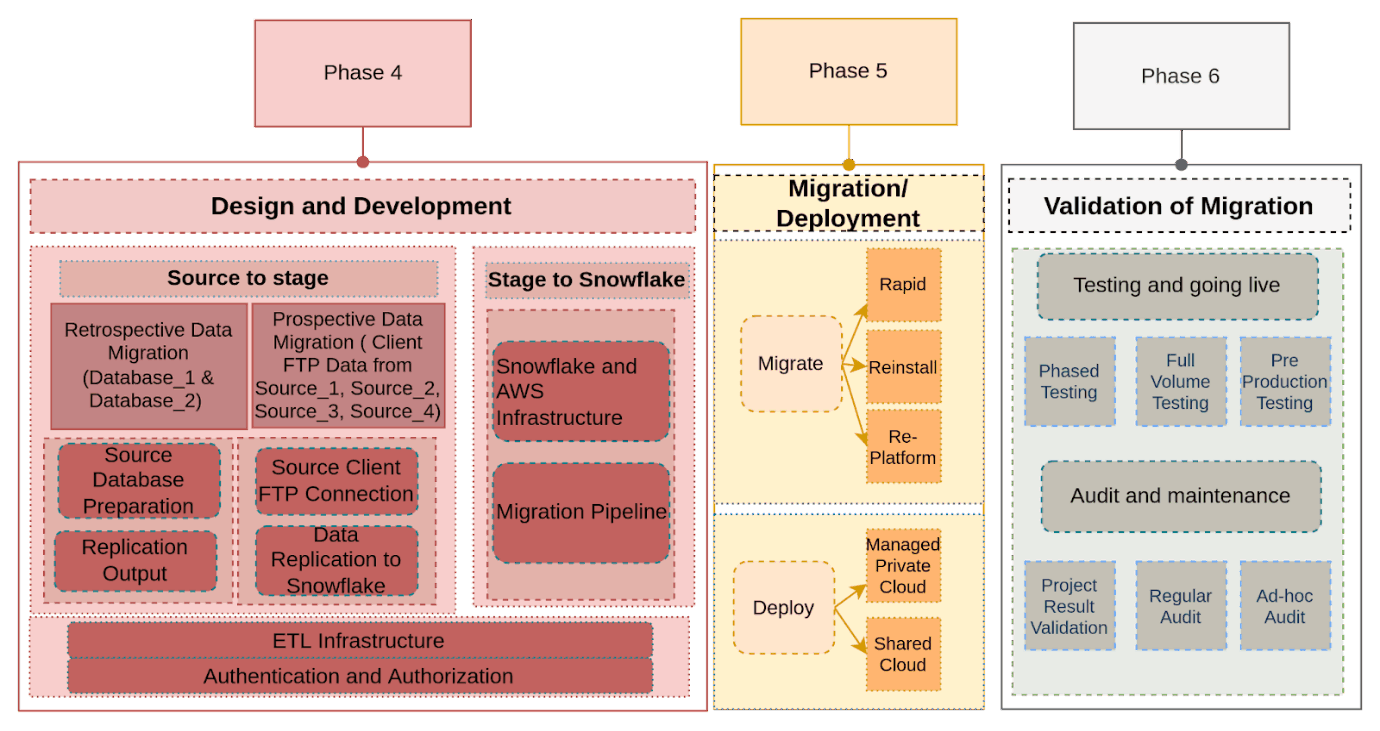
### Data Warehouse Design Summarized

The key ideas about data warehouse design that we hope you take away from this post are:

* 1. Before designing, identify business needs, align those needs to functional requirements, and secure buy-in from key business and technical leaders.
  2. Partitioning your warehouse into development, testing, and production is key for enabling efficient data warehousing management.
  3. Designing the technical components of your data warehouse requires:
     + Modelling your extraction, transformation, and loading process.
     + Designing the front-end visualization component (BI tools)
     + Optimization at the data and the query level
  4. Data warehousing is a continually evolving process requiring additional optimization and use case development.

### Phases Involved in Data Warehousing





### Phase 1: Analysis and Business Requirement Gathering

This phase involves understanding the client’s current environment and analysing the problems that are occurring in the current data warehouse if any. It focuses on analyzing the business requirements that the clients have and preparing relevant documents for the same.

### Business Requirements Session

Understanding the client’s current environment, database, and infrastructure

* + - Determining the scope of the project in relation to business objectives
    - Discovering future needs and current needs by diving deep into the data sources and optimizing them
    - Creating a disaster recovery plan in the case of system failure
    - Thinking about each layer of security
    - Anticipating compliance needs and mitigating regulatory risks
    - Analysing problems in the client’s existing data warehouse staging and production databases
    - Document details of the current environment and problems, along with data sources and dataflow in the data warehouse

### Migration of Data

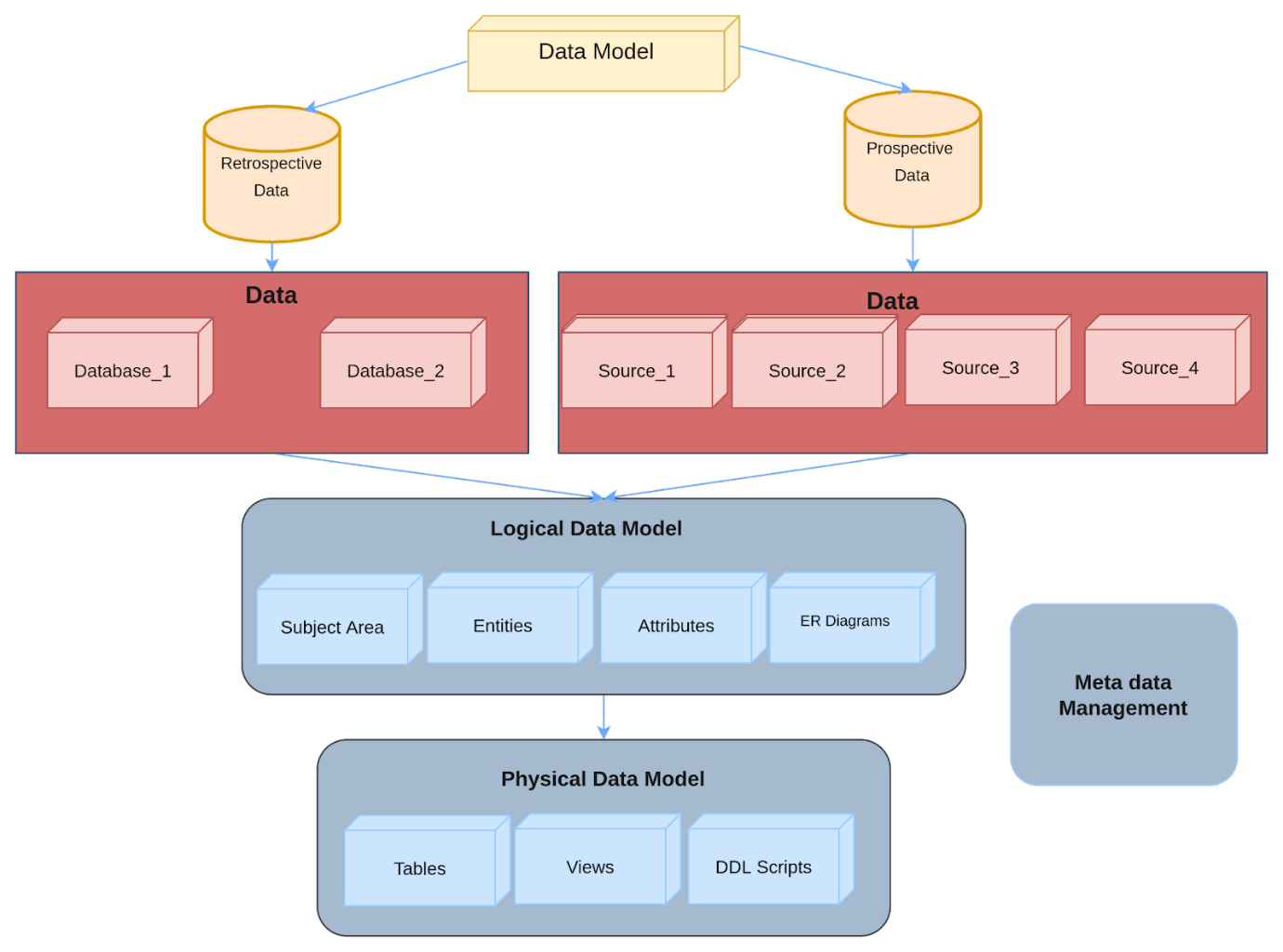
Data migration is one important activity that aims toward cost reduction and lessening the operational burden of executing licensed OLAP database systems on-premises.

### Phase 2: Data Model Design

Designing an apt data model is crucial to a successful data warehouse. Data modelling is a process of visualizing data distribution in a warehouse.

The steps involved in creating a data model design are:

* Analyse new data warehouse services (Snowflake) and create data models according to the new data warehouse
* Create a Logical Data Model
  + Define Subject Areas, Entities, and Attributes in Logical Data Model.
  + Understand Tables and relations between tables and create ER Diagrams
* Create Physical Data Model
  + Add storage tables and views as needed.
  + Generate a DDL script for the physical data model
* Metadata Management
  + Metadata contains all the data about the data warehouse
  + Metadata will also contain all the information about the ETL processing which will help us to track the ETL easily and monitor them without any trouble. There will be monitoring tables that keep track of each file move.



### Phase 3: Integration

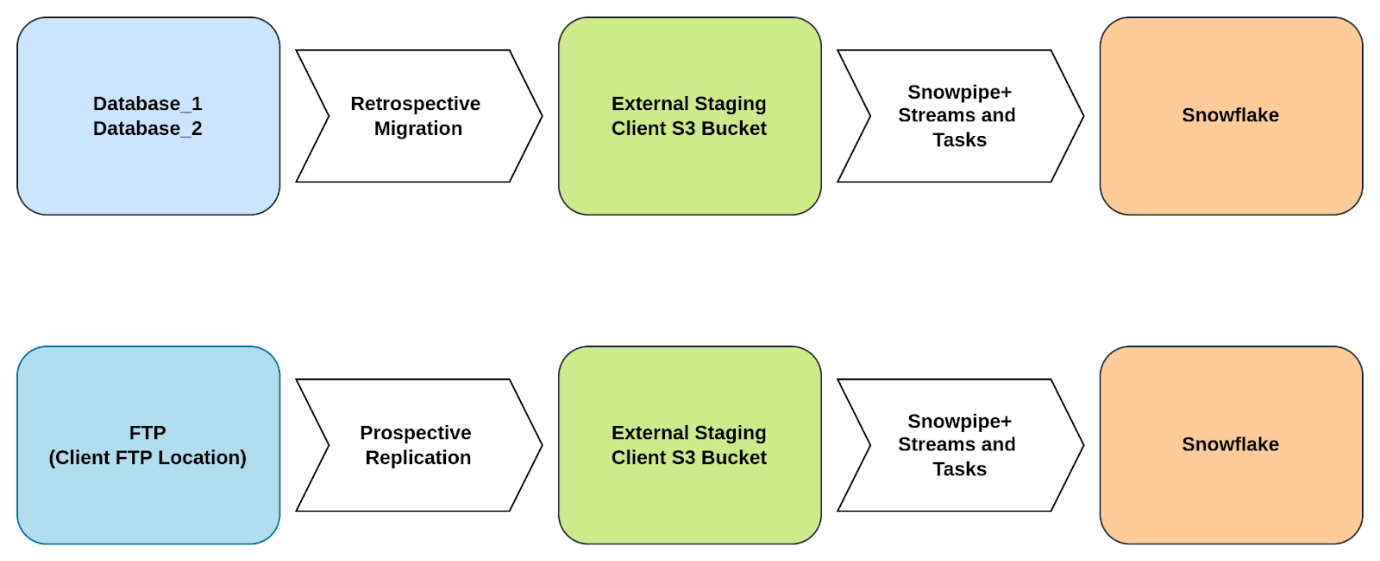
Integration of all the key elements is important for a successful migration, this will decide the path for successful data movement.

* + Identify the migration candidates from current sources
  + Transformation design mapping for all relevant candidates
  + Identify candidates for a rebuild and redevelopment

### Phase 4: Design and Development

Once the client data warehouse design is ready, we will work towards the mapping of the source and data warehouse and document them to develop the pipeline for migrating the retrospective and prospective data.

All the data sources, data from the existing data warehouse and staging and production databases, and other sources will be moved to the new data warehouse staging layer (S3 Bucket) and then to the Snowflake data warehouse.



### Phase 5: Migration and Deployment

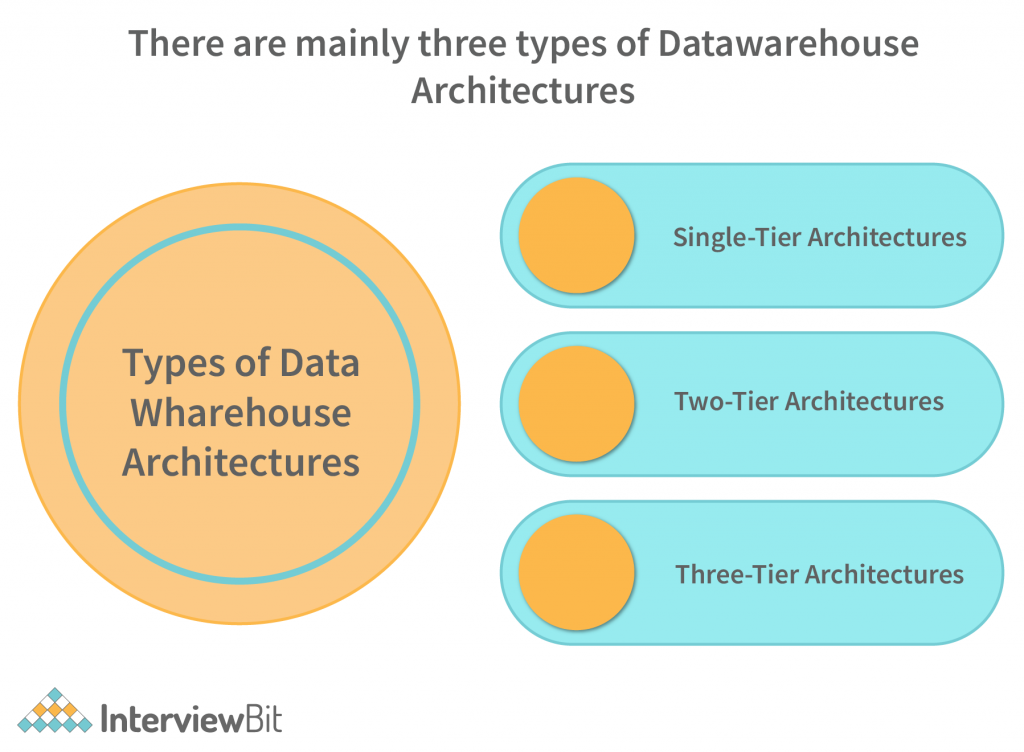
The migration process will take care of the smooth transition of all components from the development environment to UAT and production. It could be classified into rapid migration, reinstalling, and re-platform.

### Phase 6: Validate Migration

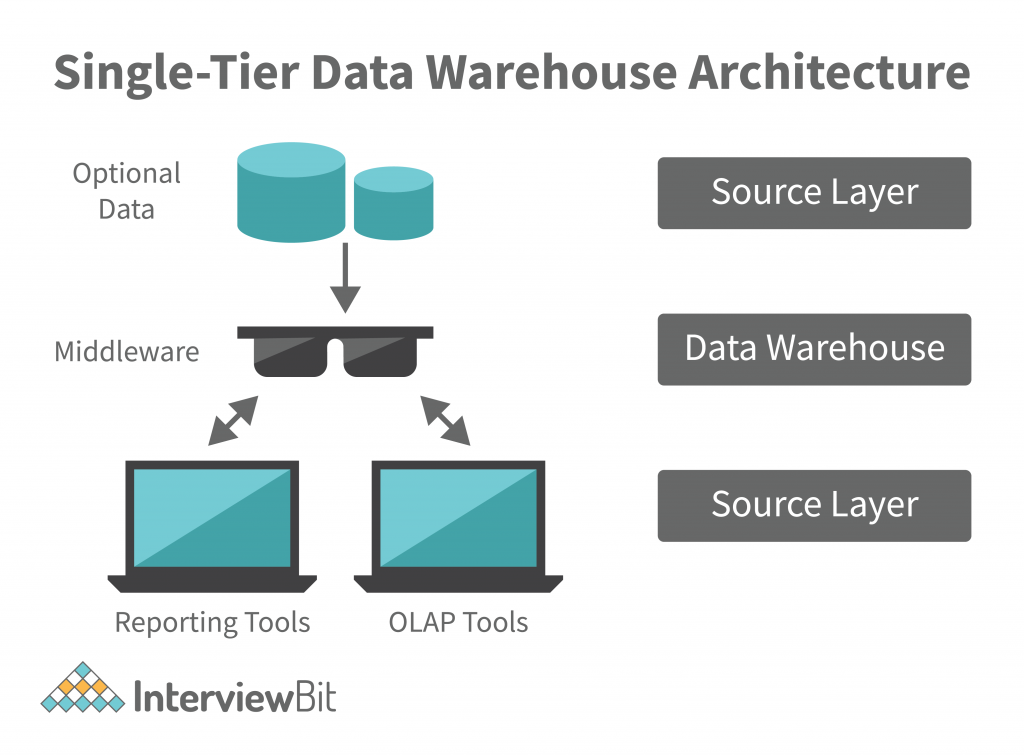
Testing should start from the moment we begin manipulating data and continue throughout all further stages. When the new system starts working, it is important to validate the project results and monitor the system’s performance in the long run.

### Data Warehouse Architecture

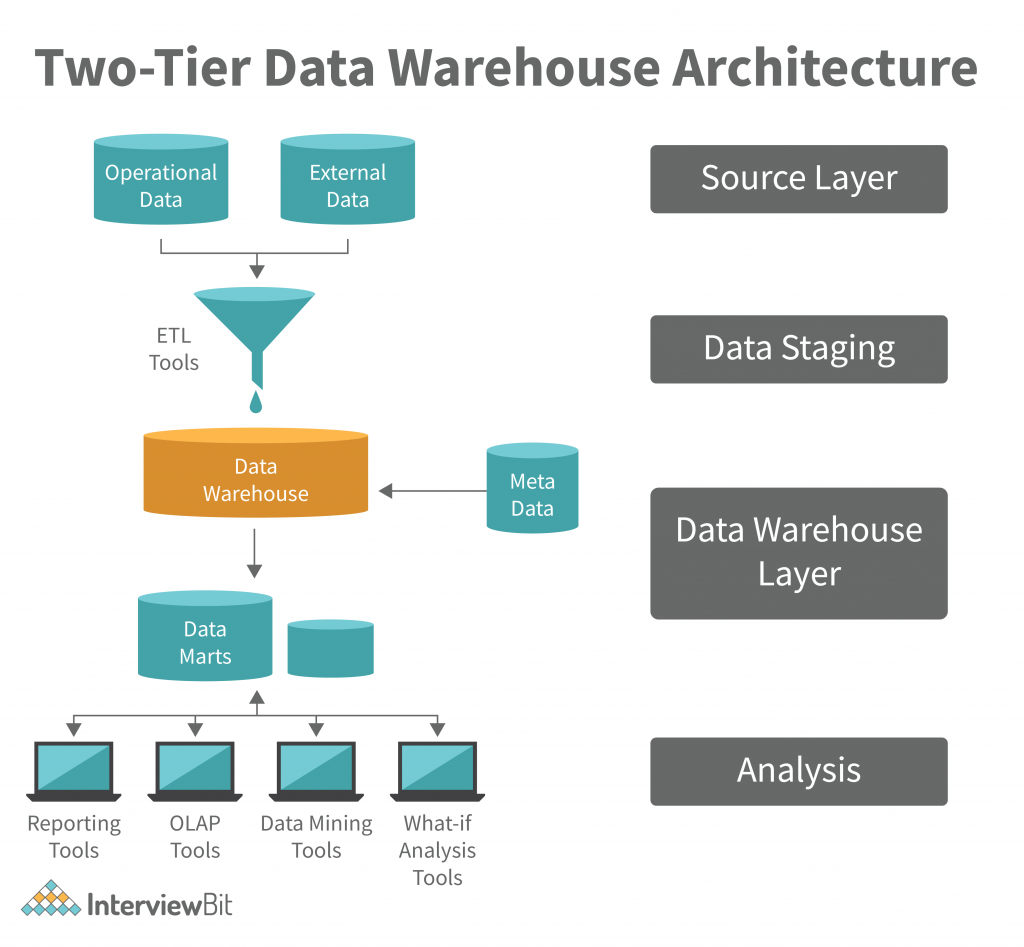
There are three approaches to creating a data warehouse layer: Single tier, two-tier, and three-tier.



* **Single-tier architecture:** A single-layer structure aimed at keeping data space minimal. This structure is rarely used in real life.



* **Two-tier architecture:** Data warehouse is the aggregation of data in a format that is easy to transform and load into a database.



* **Three-Tier Data Warehouse Architecture:** The Top, Middle, and Bottom Tiers of this Architecture of Data Warehouse are collectively referred to as the Top Tier.

1. The bottom tier of the Datawarehouse is a relational database system. This database system typically contains a relational database system. Back-end tools clean, transform, and load data into this layer.
2. A middle tier OLAP server is either ROLAP or MOLAP-based. It abstracts OLAP from the end user by serving as a middle tier OLAP server
3. The front-end client layer of the top-tier is important because it is the first point of interaction with the data. It is where data is presented to the end user, and decisions are made with the data.

### Data Integration

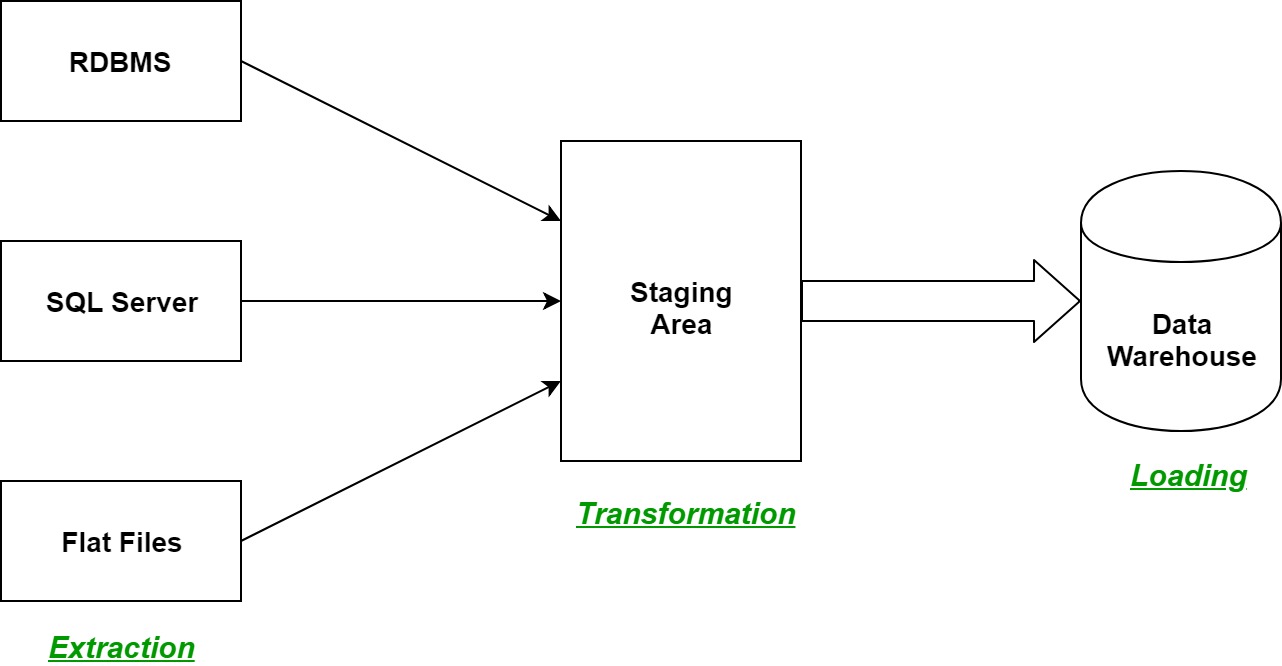
Data integration can be considered as one of the main components in the [Data management](https://www.simplilearn.com/what-is-data-management-article) process. It is the process of collecting and consolidating data from all sources into one single dataset or [data warehouse](https://www.simplilearn.com/data-warehouse-article)

### Data Integration Methods and Strategies

1. **Manual data integration**: Data managers must manually conduct all phases of the integration, from retrieval to presentation.
2. **Middleware data integration**: Middleware, a type of software, facilitates communication between legacy systems and updated ones to expedite integration.
3. **Application-based integration**: Software applications locate, retrieve, and integrate data by making data from different sources and systems compatible with one another.
4. **Uniform access integration**: A technique that retrieves and uniformly displays data, but leaves it in its original source.
5. **Common storage integration**: An approach that retrieves and uniformly displays the data, but also makes a copy of the data and stores

### ETL Process in Data Warehouse

1. ETL stands for Extract, Transform, Load and it is a process used in data warehousing to extract data from various sources, transform it into a format suitable for loading into a data warehouse, and then load it into the warehouse. The process of ETL can be broken down into the following three stages:
2. **Extract**: The first stage in the ETL process is to extract data from various sources such as transactional systems, spreadsheets, and flat files. This step involves reading data from the source systems and storing it in a staging area.
3. **Transform**: In this stage, the extracted data is transformed into a format that is suitable for loading into the data warehouse. This may involve cleaning and validating the data, converting data types, combining data from multiple sources, and creating new data fields.
4. **Load**: After the data is transformed, it is loaded into the data warehouse. This step involves creating the physical data structures and loading the data into the warehouse.
5. The ETL process is an iterative process that is repeated as new data is added to the warehouse. The process is important because it ensures that the data in the data warehouse is accurate, complete, and up-to-date. It also helps to ensure that the data is in the format required for data mining and reporting.



### Data exploration

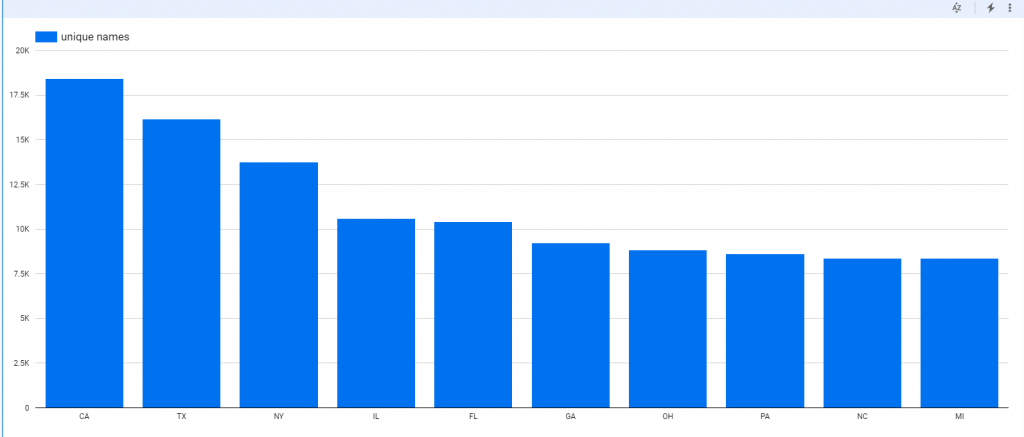
**Data exploration** is the process of analysing datasets to find patterns and relationships, and is sometimes more formally referred to as **exploratory data analysis (EDA)**.

patterns and relationships, and is sometimes more formally referred to as **exploratory data analysis (EDA)**.

### Data exploration techniques

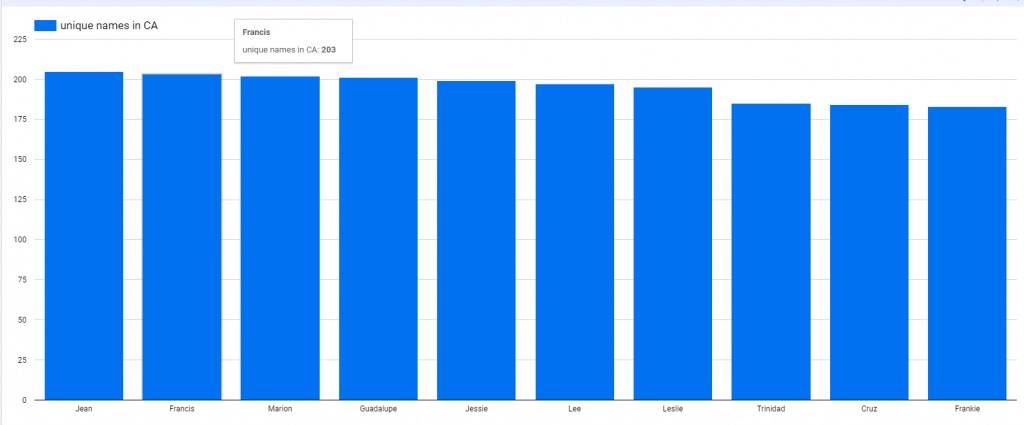
* Unique value count

It’s a first thing which can be useful during exploration, showing how many unique values are included per categorical column. This will give us a general idea of what the data is about.

**

* Frequent value count

Detecting how frequently individual values occur in a specific column. This will give an insight into the content of each categorical variable.

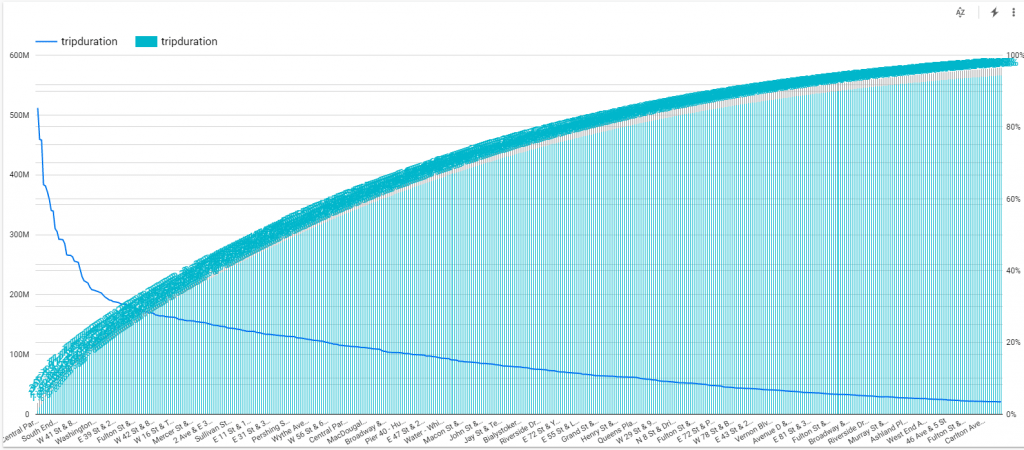
**

* Variance

For numeric values we have plenty of easy ways to get some basic information like minimum, maximum, or variance are very useful. Variance gives a good indication about spread of the values across specific variable.

* Pareto analysis

This data exploration technique allows you to focus on what is really  important. 80-20 pareto rule can effectively show the level where we are observing significant meaning of specific value.

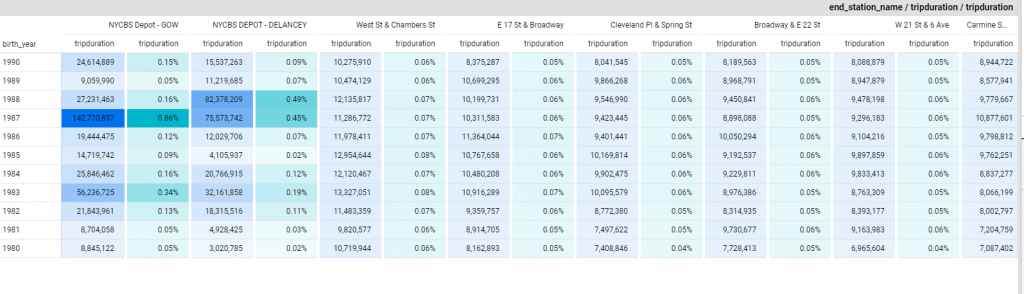
**

* Histogram

Gives information about the range of values falling in the majority sector. It indicates any skew data as well as minimum and maximum ones.

* Correlation heat-map

Correlation basically means association between two things. It is useful to express relationships between different columns in data. One of the best ways to see correlation between numeric columns is heat-map. Correlation, however, might be useful with various types of data.

**

### Data warehouse enables data architect to deliver actionable insights

A data warehouse is a centralized repository for storing and managing large volumes of data from various sources, making it a crucial tool for data architects to deliver actionable insights. Here's how a data warehouse enables data architects to accomplish this:

1. Data Integration: Data architects can use data warehouses to integrate data from diverse sources such as transactional databases, external APIs, spreadsheets, and more. This consolidation of data simplifies the process of data analysis by providing a unified and consistent view of the data.
2. Data Cleaning and Transformation: Data in its raw form may contain errors, inconsistencies, or missing values. Data architects can design data transformation processes within the data warehouse to clean and preprocess the data, ensuring its quality and reliability for analytical purposes.
3. Historical Data Storage: Data warehouses store historical data, allowing data architects to analyse data trends and patterns over time. This historical context is crucial for making informed decisions and identifying long-term insights.
4. Structured Data: Data in a data warehouse is typically organized into structured tables, making it easier for data architects to query and analyse the data using SQL or other querying languages. This structured format simplifies the extraction of relevant insights.
5. Performance Optimization: Data architects can design data warehouses with performance in mind. They can create indexes, partitions, and aggregates to optimize query performance, ensuring that data analysts can access and analyse data quickly.
6. Data Security and Governance: Data warehouses offer robust security and governance features. Data architects can implement access controls, encryption, and auditing to protect sensitive data and ensure compliance with data privacy regulations.
7. Scalability: Data warehouses can scale both vertically and horizontally to accommodate growing data volumes and analytical workloads. This scalability ensures that the data architecture can support the evolving needs of the organization.
8. Business Intelligence Tools Integration: Data architects often integrate data warehouses with business intelligence (BI) tools, reporting tools, and data visualization platforms. This integration enables data analysts and business users to create reports, dashboards, and visualizations to derive actionable insights from the data.
9. Data Modelling: Data architects design data models within the data warehouse to represent the relationships between different data entities. These models help ensure that the data is organized logically and is easy to work with.
10. Ad Hoc Querying: Data architects can create data structures that facilitate ad hoc querying, allowing data analysts and business users to explore the data in real-time and discover insights on the fly.
11. Data Aggregation: Data warehouses support the pre-aggregation of data, which is especially useful for summarizing large datasets and accelerating query performance. This enables users to quickly obtain aggregated insights.
12. Metadata Management: Data architects establish metadata repositories within data warehouses to document data definitions, lineage, and relationships. This metadata aids in understanding the data and its context, which is critical for deriving meaningful insights.
13. Monitoring and Maintenance: Data architects monitor the performance and health of the data warehouse, ensuring that it continues to deliver reliable and actionable insights. They can implement automated processes to manage backups, updates, and data consistency.