CHENNAI INSTITUTE OF TECHNOLOGY
(Autonomous)
(Affiliated to Anna University, Approved by AICTE, Accredited by NAAC & NBA)
Sarathy Nagar, Kundrathur, Chennai – 600069, India.

LECTURE NOTES

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CCS341 DATA WAREHOUSING

III CSE/VI SEMESTER

CCS341 -Data warehousing

UNIT 1 – Introduction to Data warehouse

Data warehouse Introduction:

A **data warehouse** is designed to support the management decision-making process by providing a platform for data cleaning, data integration, and data consolidation. A data warehouse contains subject-oriented, integrated, time-variant, and non-volatile data.

Characteristics of Data warehouse:

1. **Subject Oriented:** A data warehouse is often subject-oriented because it delivers may be achieved on a particular theme which means the data warehousing process is proposed to handle a particular theme that is more defined. These themes are often sales, distribution, selling. etc.
2. **Time-Variant**: When the data is maintained via totally different intervals of time like weekly, monthly, or annually, etc. It founds numerous time limits that are unit structured between the big datasets and are command within the online transaction method (OLTP). The time limits for the data warehouse are extended than that of operational systems. The data resided within the data warehouse is predetermined with a particular interval of time and delivers information from the historical perspective. It contains parts of time directly or indirectly.
3. **Non-volatile:** The data residing in the data warehouse is permanent and defined by its names. It additionally means that the data in the data warehouse is cannot be erased or deleted or also when new data is inserted into it. In the data warehouse, data is read-only and can only be refreshed at a particular interval of time. Operations such as delete, update and insert that is done in a software application over data is lost in the data warehouse environment. There are only two types of data operations that can be done in the data warehouse:
    - Data Loading
    - Data Access
4. **Integrated:** A data warehouse is created by integrating data from numerous different sources such that from mainframe computers and a relational database. Additionally, it should also have reliable naming conventions, formats, and codes. Integration of data warehouse benefits in the

successful analysis of data. Dependability in naming conventions, column scaling, encoding structure, etc. needs to be confirmed. Integration of data warehouse handles numerous subject-oriented warehouses.

Need for Data warehouse:

1. **Business User:** Business users require a data warehouse to view summarized data from the past. Since these people are non-technical, the data may be presented to them in an elementary form.
2. **Store historical data:** Data Warehouse is required to store the time variable data from the past. This input is made to be used for various purposes.
3. **Make strategic decisions:** Some strategies may be depending upon the data in the data warehouse. So, data warehouse contributes to making strategic decisions.
4. **For data consistency and quality:** Bringing the data from different sources at a commonplace, the user can effectively undertake to bring the uniformity and consistency in data.
5. **High response time:** Data warehouse has to be ready for somewhat unexpected loads and types of queries, which demands a significant degree of flexibility and quick response time.

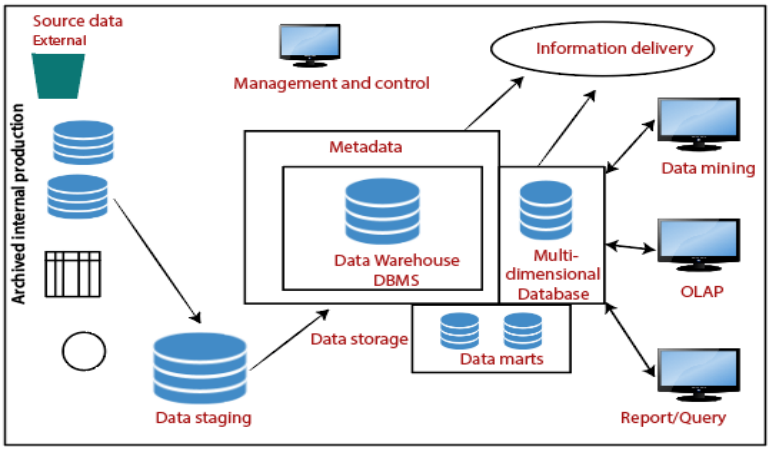
Benefits of Data warehouse:

1. **Decision Support**: Enables informed decision-making by providing a comprehensive view of organizational data.
2. **Business Intelligence:** Supports business intelligence and analytics, allowing organizations to gain insights into their operations, customer behavior, and market trends.
3. **Data Quality:** Improves data quality through integration, cleansing, and validation processes.
4. **Historical Analysis:** Facilitates historical analysis, helping organizations understand trends, patterns, and performance over time.

Data warehouse Components:

There are five main components,

1. Databases
2. ETL
3. Metadata
4. Query tool
5. Data mart



Components or Building Blocks of Data Warehouse

Databases:

The central database is the foundation of data warehousing environment. This database is implemented on the RDBMS technology.

It is one of the first component of data warehouse, some of the data warehouse databases are

- Relational database
- Analytical database
- Central database
- Cloud based database

1. Relational database:

This database consists of data in the form of rows and columns which collectively form table.

2. Analytical database:

This database helps sustain and manage storages in analytics part.

3. Central database:

It keeps all the databases related to business organization and makes it easier for analysts to build the report.

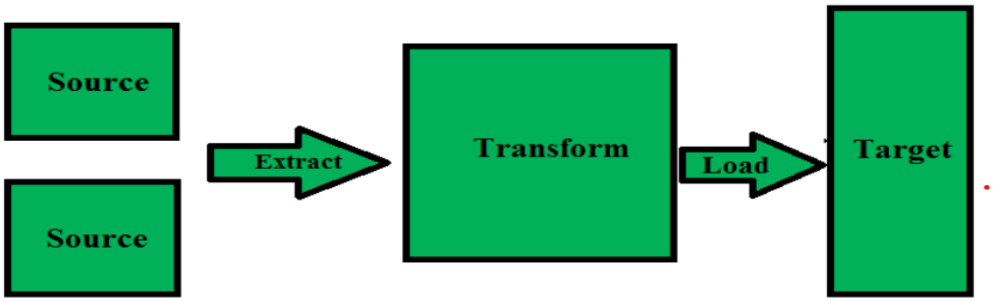
4. Cloud based database:

This database consists of data hosted on cloud. So that it does not acquire any hardware system for establishing data warehouse.

Extraction, Transform and Load (ETL):

It is a data integration process in which data is extracted from various resources and it is transformed into a suitable format and then loaded into a data warehouse.

1. **Extract:**
   It is the process of extracting raw data from all available data sources such as databases, files, ERP, CRM or any other.
2. **Transform:**
   The extracted data is immediately transformed as required by the user.
3. **Load:**
   The transformed data is then loaded into the data warehouse from where the users can access it.



Metadata:

Metadata is data about your data. It specifies the source, usage, values, and other features of the data sets in the data warehouse.

Business metadata: The data that shows a readily comprehensible standpoint of the information in the warehouse.

Technical metadata: The data utilized by programmers and managers when performing administration tasks and warehouse development. It describes how to access data – including where it resides and how it is structured.

Query Tools:

These are the components that allow users to interact with the data warehouse to get relevant data from it.

Some of the tools are,

1. Query and reporting tool
2. Application development tool
3. Data mining tool
4. Online analytics tool

1. Query and reporting tool are classified into 2 types:

   Managed query tool

   This is used to protect the end user from SQL query related complexities by adding a security layer between the database and end user.

   Reporting query tool

   This tool is used for developing business report that the end user can use it at affordable cost.

2. Application development tool: This is a graphical data access environment which integrates OLAP tools with data warehouse and can be used to access all database systems.

3. Data mining tool: They are used to discover knowledge from data warehouse. The data can also be used for data visualization and data correction purpose
4. OLAP Tools: Allow users to interactively analyze multidimensional data cubes.A multidimensional database structure that allows for efficient analysis of data. OLAP cubes facilitate complex queries and provide a multidimensional view of data.

Data mart:

- It is an alternative to data warehousing or acts as a mini data warehouse. It uses less time and money to create. It might be a collection of summarized, denormalized or aggregated data.
- It is concerned with specific business unit, by limiting the data to a particular business unit. (ex: sales dept)
- It has high security by limiting the visibility of non-essential data to the department eliminates the chance of that data being used irresponsibly.
- It has high speed as there will be less data in the data mart, the processing overhead is decreased. This means that queries will run faster.

Operational database Vs data warehouse:

Operational Database	Data Warehouse
Operational database are designed to support high volume transactional processing.	Data warehousing are designed to support high volume analytical processsing.
Operational database are process oriented or application oriented.	Data warehouse are subject oriented.
Operational database are concerned with current data.	Data warehousing are concerned with historical data.
It contains detailed data only.	It contains summarized and detailed data.
It is used for decision making and operational reporting.	It is used for long term decision making and management reporting.
It is used at the operational level.	It is used at the managerial level.
Data within operational systems are updated regularly according to the need.	New data may be added regularly and once added it is rarely changed.
Less no of data are accessed	Large no of data are accessed.

It is created for Online Transaction Processing(OLTP.)	It is created for Online Analytical Processing(OLAP).
It supports thousands of concurrent clients.	It supports a few concurrent clients relative to OLTP.

Data Warehouse Architectures:

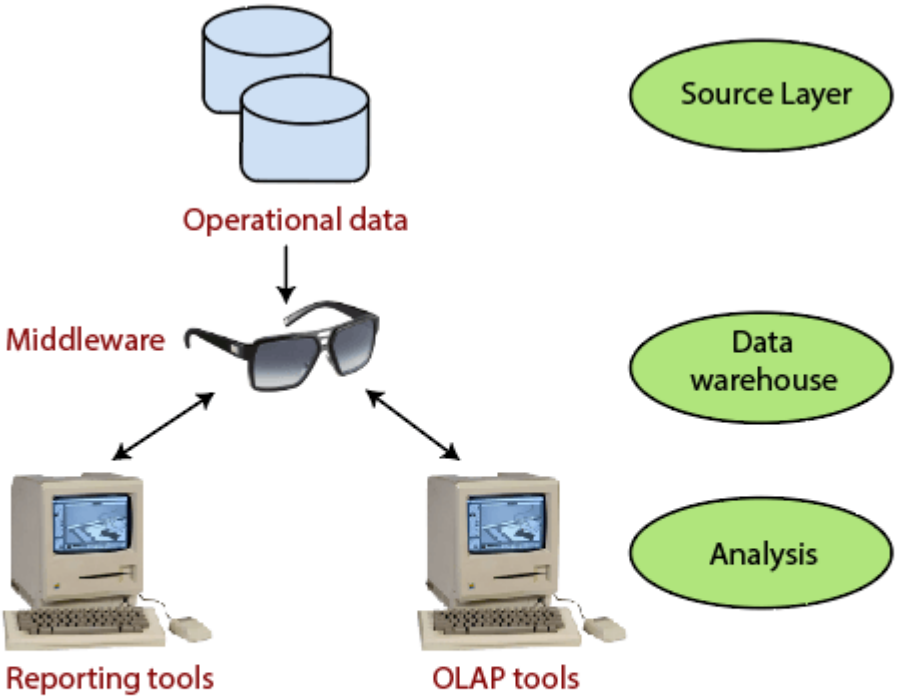
There are three types of data warehouse architecture, they are

1. Single-tier architecture
2. Two-tier architecture
3. Three-tier architecture

Single-tier architecture:

Single-Tier architecture is not periodically used in practice. Its purpose is to minimize the amount of data stored to reach this goal. It removes data redundancies.

In this method, data warehouses are virtual. This means that the data warehouse is implemented as a multidimensional view of operational data created by specific middleware, or an intermediate processing layer.

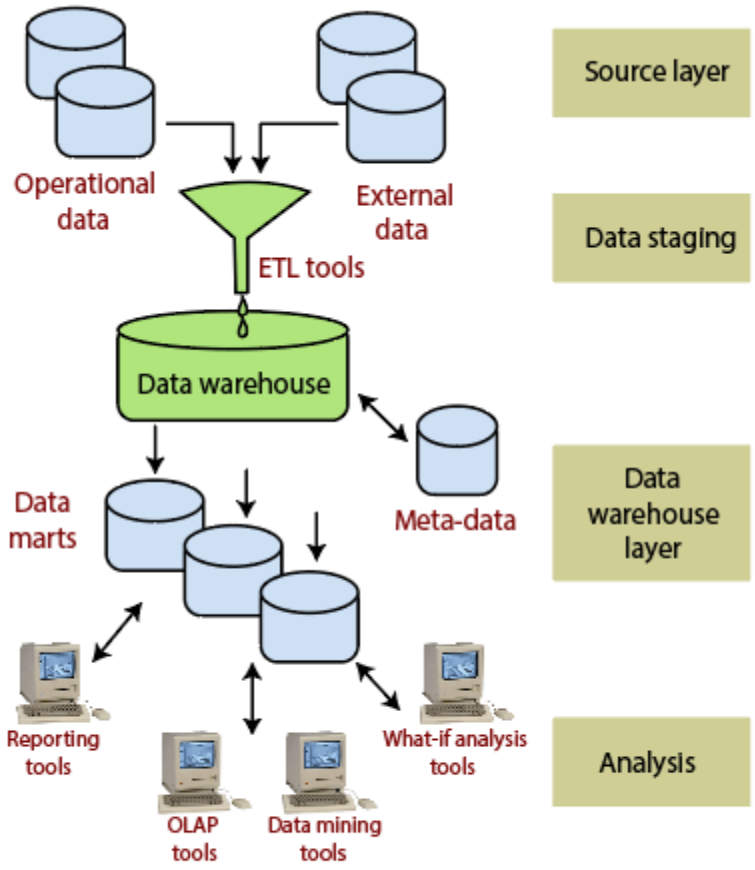


Single-Tier Data Warehouse Architecture

The vulnerability of this architecture lies in its failure to meet the requirement for separation between analytical and transactional processing. Analysis queries are agreed to operational data after the middleware interprets them. In this way, queries affect transactional workloads.

Two-tier architecture:

The requirement for separation plays an essential role in defining the two-tier architecture for a data warehouse system.



Two-Tier Data Warehouse Architecture

Although it is typically called two-layer architecture to highlight a separation between physically available sources and data warehouses, in fact, consists of four subsequent data flow stages:

1. **Source layer:** A data warehouse system uses a heterogeneous source of data. That data is stored initially to corporate relational databases or legacy databases, or it may come from an information system outside the corporate walls.
2. **Data Staging:** The data stored to the source should be extracted, cleansed to remove inconsistencies and fill gaps, and integrated to merge heterogeneous sources into one standard schema. The so-named **Extraction**, **Transformation**, and **Loading Tools (ETL)** can combine heterogeneous schemata, extract, transform, cleanse, validate, filter, and load source data into a data warehouse.

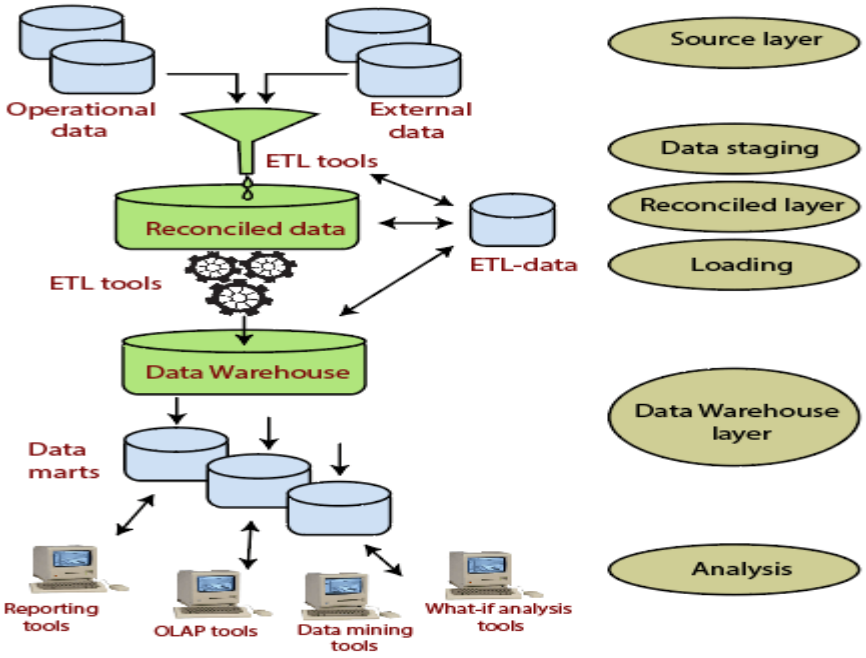
3. **Data Warehouse layer:** Information is saved to one logically centralized individual repository: a data warehouse. The data warehouses can be directly accessed, but it can also be used as a source for creating data marts, which partially replicate data warehouse contents and are designed for specific enterprise departments. Meta-data repositories store information on sources, access procedures, data staging, users, data mart schema, and so on.
4. **Analysis:** In this layer, integrated data is efficiently, and flexible accessed to issue reports, dynamically analyze information, and simulate hypothetical business scenarios. It should feature aggregate information navigators, complex query optimizers, and customer-friendly GUIs.

Three-Tier Data warehouse Architecture

The three-tier architecture consists of the source layer (containing multiple source system), the reconciled layer and the data warehouse layer (containing both data warehouses and data marts). The reconciled layer sits between the source data and data warehouse.

The main advantage of the **reconciled layer** is that it creates a standard reference data model for a whole enterprise. At the same time, it separates the problems of source data extraction and integration from those of data warehouse population. In some cases, the **reconciled layer** is also directly used to accomplish better some operational tasks, such as producing daily reports that cannot be satisfactorily prepared using the corporate applications or generating data flows to feed external processes periodically to benefit from cleaning and integration.

This architecture is especially useful for the extensive, enterprise-wide systems. A disadvantage of this structure is the extra file storage space used through the extra redundant reconciled layer. It also makes the analytical tools a little further away from being real-time.



Three-Tier Architecture for a data warehouse system

Three-Tier Data Warehouse Architecture

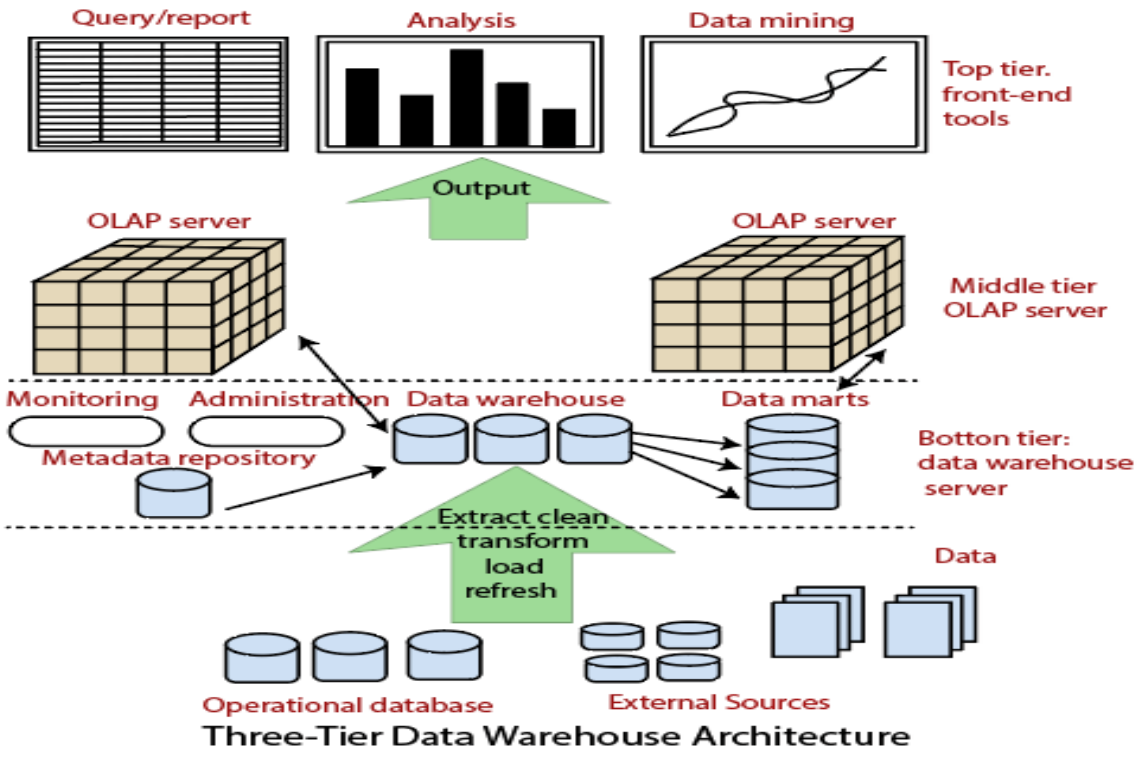
Data Warehouses usually have a three-level (tier) architecture that includes:

1. Bottom Tier (Data Warehouse Server)
2. Middle Tier (OLAP Server)
3. Top Tier (Front end Tools).

A **bottom-tier** that consists of the **Data Warehouse server**, which is almost always an RDBMS. It may include several specialized data marts and a metadata repository.
Data from operational databases and external sources (such as user profile data provided by external consultants) are extracted using application program interfaces called a gateway. A gateway is provided by the underlying DBMS and allows customer programs to generate SQL code to be executed at a server. **Examples** of gateways contain **ODBC** (Open Database Connection) and **OLE-DB** (Open-Linking and Embedding for Databases), by **Microsoft**, and **JDBC** (Java Database Connection).

A **middle-tier** which consists of an **OLAP server** for fast querying of the data warehouse.
The OLAP server is implemented using either
(1) A **Relational OLAP (ROLAP) model**, i.e., an extended relational DBMS that maps functions on multidimensional data to standard relational operations.
(2) A **Multidimensional OLAP (MOLAP) model**, i.e., a particular purpose server that directly implements multidimensional information and operations.

A **top-tier** that contains **front-end tools** for displaying results provided by OLAP, as well as additional tools for data mining of the OLAP-generated data. The overall Data Warehouse Architecture is as follows,



Three-Tier Data Warehouse Architecture

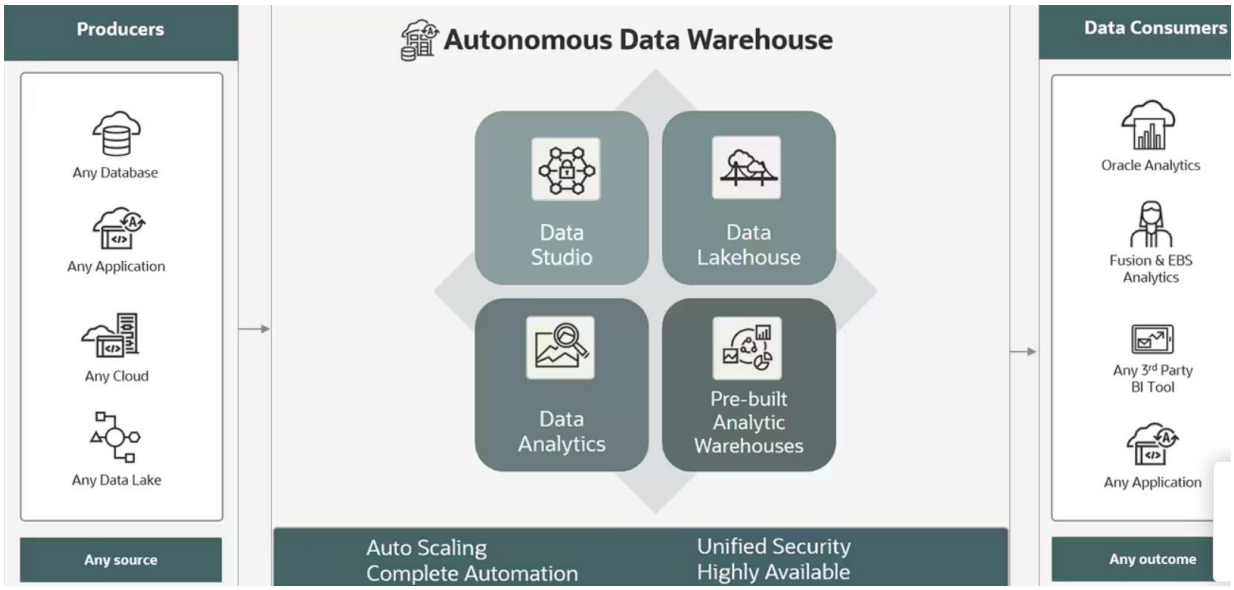
The **metadata repository** stores information that defines DW objects. It includes the following parameters and information for the middle and the top-tier applications:

1. A description of the DW structure, including the warehouse schema, dimension, hierarchies, data mart locations, and contents, etc.
2. Operational metadata, which usually describes the currency level of the stored data, i.e., active, archived or purged, and warehouse monitoring information, i.e., usage statistics, error reports, audit, etc.
3. System performance data, which includes indices, used to improve data access and retrieval performance.
4. Information about the mapping from operational databases, which provides source RDBMSs and their contents, cleaning and transformation rules, etc.
5. Summarization algorithms, predefined queries, and reports business data, which include business terms and definitions, ownership information, etc.

Autonomous Data warehouse:

ADW are robust data warehousing solutions that feature enhanced data quality, data security, fully autonomous, and business intelligence analytics to streamline a company's decision-making process.

Architecture of Autonomous Data warehouse:



The data warehouse on a cloud, the concept has recently emerged as an alternative to conventional or traditional, on-premises data warehousing and Cloud-based data warehouse.
The market is inundated with Data Warehouse solutions, be it on-premises or cloud. The most common use cases for using a data warehouse are

- The need to centrally store all business-critical data
- Single source for analyzing web, mobile, CRM, and other applications

- Tools that can dive deeper than traditional analytics tools by querying raw data with SQL
- Providing multiple people access to the same data set simultaneously

Need for Autonomous Data warehouse:

- ADW is highly automated.
- ADW is hosted on Oracle Cloud it has easy use of installation, bringing up the environment, patching the environment, and networking side of it, protecting and encrypting the data.
- There is no involvement of network or database admin at any level.
- ADW can handle complex SQL queries.
- The user of ADW, focus on data sets, business use cases that the data sets can be applied to, and extracting the useful information and using it for future use. Hence it saves a lot of time by using it for strategizing and delivering solutions rather than planning and maintaining the foundation.
- Automatic scaling, auto-tuning, auto-patching are some other features in ADW.
- Autonomous Data Warehouse is auto secure. It autonomously encrypts data at rest and in motion (including backups and network connections), protects regulated data, applies all security patches, enables auditing, and performs threat detection.

Benefits of autonomous data warehouse:

Automated Management

ADW offers automated provisioning, configuration, back-up as well as zero downtime patching and recovery from failure without human intervention

Elastic Scaling

ADW is truly elastic. ADW auto-scales without any human action i.e. it can scale compute independently from storage and get exactly the number of CPUs needed without being constrained by fixed models.

High Performance

ADW is not powered by generic hardware but by Exadata, Oracle's engineered system that is the best platform to run Oracle databases.

Comprehensive Data Protection

Data is automatically protected and the built-in Oracle Data Safe makes it easy to discover sensitive data, mask it, evaluate security risks, and implement security controls. And privileged users cannot access other users' data with Oracle's Database Vault technology.

Converged Database

ADW is a converged database allowing you to store data from all formats in one single database and one single source of truth.

Flexible Deployment

ADW is available both in Oracle Cloud Infrastructure and in the customer's data center.

Snowflake:

Snowflake is a cloud-based advanced data platform system, provided as Software-as-a-Service (SaaS). Snowflake provides features of data storage from AWS S3, Azure, Google Cloud, processing complex queries and different analytic solutions. The analytic solutions provided by Snowflake are faster, easy to use and more flexible than traditional databases and their analytics features.

Need for Snowflake:

- There is no hardware neither virtual nor physical to select, install, configure or manage from client side.
- There is no software to install, configure or manage to access it.
- All ongoing maintenance, management, upgrades and patching are owned by Snowflake itself.

Features of Snowflake:

- Simple, reliable data pipelines in multi languages like Java, Python, PHP, Spark, Ruby etc.
- Secured access, very good performance and security of data lake.
- Zero administration for tool, data storage and data size.
- Simple data preparation for modeling with any framework.
- No operation burden to build data intensive applications.
- Share and collaborate live data across company's ecosystem.

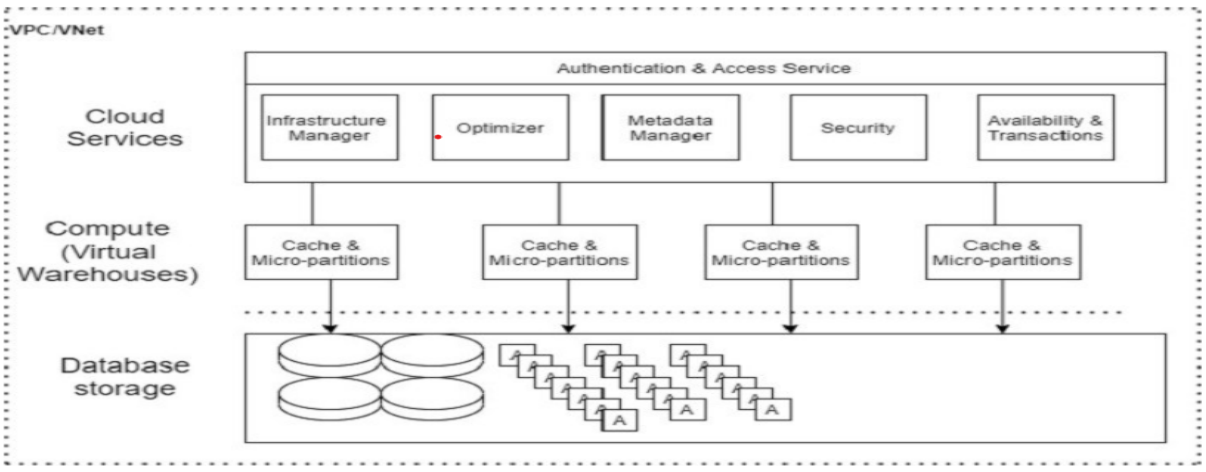
Data Architecture of snowflake:

Snowflake data architecture re-invents a new SQL query engine. It is designed for the cloud only. Snowflake has central data repository for storage of structured and semi-structured data. These data can be accessed from all available compute nodes in the Snowflake platform. It uses virtual warehouse as compute environment for processing the queries. While processing queries, it utilizes multi-cluster, micro-partitioning and advanced cache concepts. Snowflake's cloud services are responsible to provide end to end solution to the user like logging validation of user to result of select queries.

Snowflake's data architecture **has three main layers** −

- Database Storage
- Query Processing
- Cloud Services

Following is the **data architecture** diagram of Snowflake −



Database Storage

Snowflake supports Amazon S3, Azure and Google Cloud to load data into Snowflake using file system. User should upload a file (.csv, .txt, .xlsx etc.) into the cloud and after they create a connection in Snowflake to bring the data. Data size is unlimited, but file size is up to 5GB as per cloud services. Once data is loaded into Snowflake, it utilizes its internal optimization and compression techniques to store the data into central repository as columnar format. The central repository is based on cloud where data stores.

Snowflake owns responsibilities to all aspects of data management like how data is stored using automatic clustering of data, organization and structure of data, compression technique by keeping data into many micro-partitions, metadata, statistics and many more. Snowflake stores data as data objects and users can't see or access them directly. Users can access these data through SQL queries either in Snowflake's UI or using programming language like Java, Python, PHP, Ruby etc.

Query Processing

Query execution is a part of processing layer or compute layer. To process a query, Snowflake requires compute environment, known as "Virtual Warehouse" in Snowflake's world. Virtual warehouse is a compute cluster. A virtual warehouse consists of CPU, Memory and temporary storage system so that it could perform SQL execution and DML (Data Manipulation Language) operations.

- SQL SELECT executions
- Updating of data using Update, Insert, Update
- Loading data into tables using COPY INTO <tables>
- Unloading data from tables using COPY INTO <locations>

However, the number of servers depends on size of virtual warehouses. For example, XSmall warehouse has 1 Server per cluster, while a Small Warehouse has 2 Servers per cluster and it gets double on increasing the size such as Large, XLarge, etc.

While executing a query, Snowflake analyzes the requested query and uses the latest micro-partitions and evaluates caching at different stages to increase performance and decrease the time for bringing the data. Decrease the time means less credit is used of a user.

Cloud Services

Cloud Service is the 'Brain' of the Snowflake. It coordinates and manages activities across Snowflake. It brings all components of Snowflake together to process user requests from logging validation to deliver query's response.

The following services are managed at this layer −

- It is the centralized management for all storage.
- It manages the compute environments to work with storage.
- It is responsible for upgrades, updates, patching and configuration of Snowflake at cloud.
- It performs cost-based optimizers on SQL queries.
- It gathers statistics automatically like credit used, storage capacity utilization
- Security like Authentication, Access controls based on roles and users
- It performs encryption as well as key management services.
- It stores metadata as data is loaded into the system.

Autonomous Data warehouse verses Snowflake:

S.No	Autonomous Data warehouse	Snowflake
1.	No DBA expertise required to provision, backup, secure, or patch	Requires manual tuning and specialized DBA expertise for cluster set and storage options
2.	Runs on Exadata backbone with enterprise class capabilities	Runs on generic AWS or Azure infrastructure.
3.	High availability with fully automated backups and RAC for compute node failures.	Availability upto nine hours per year and requires premier support for HA at additional cost.
4.	Auto applies security patches and upgrades with complete tenant and data isolation	Shared tenancy and no isolation of operational users from customer application data.
5.	Single data warehouse – scaling compute or storage is not constrained by fixed blocks.	DBA must setup single or multi clusters resulting in over provisioning and higher costs.
6.	Advanced enterprise grade analytics tools built in with complete oracle cloud integration.	Must rely on third parties for BI, Data integration and advanced analytics.

Modern Data Warehousing:

A Modern Data Warehouse is a cloud-based solution that gathers and stores the information. Organizations can process this data to make intelligent decisions. Various organizations use a Modern Data Warehouse to improve their finances, human resources, and operations business processes.

Components of Modern Data Warehouse (Modern Data Warehouse Pyramid):

There are five different components of a Modern Data Warehouse.
Level 1: Data Acquisition
Data acquisition can come from a variety of sources such as:

- IoT devices
- Social media posts
- YouTube videos
- Website content
- Customer data
- Enterprise Resource Planning
- Legacy data stores

Level 2: Data Engineering
Next the data need to be uploaded into the data warehouse. Data engineering uses pipelines and ETL (extract, transform, load) tools.

Level 3: Data Management Governance
After that need to evaluate the quality of the data then need to steward that data because security and privacy must be considered.
Data governance helps ensure the quality of the info by stewarding, prepping, and cleaning the data to ensure it is ready for analysis.

Level 4: Reporting and Business Intelligence
Once the data is clean, start using factory analysis to take that raw material(data) and turn it into a finished good (business intelligence).

Level 5: Data Science
Modern Data Warehouse is about more than seeing the information; it's about using the data to make smarter decisions. There are several different programs to help you leverage the data, including:

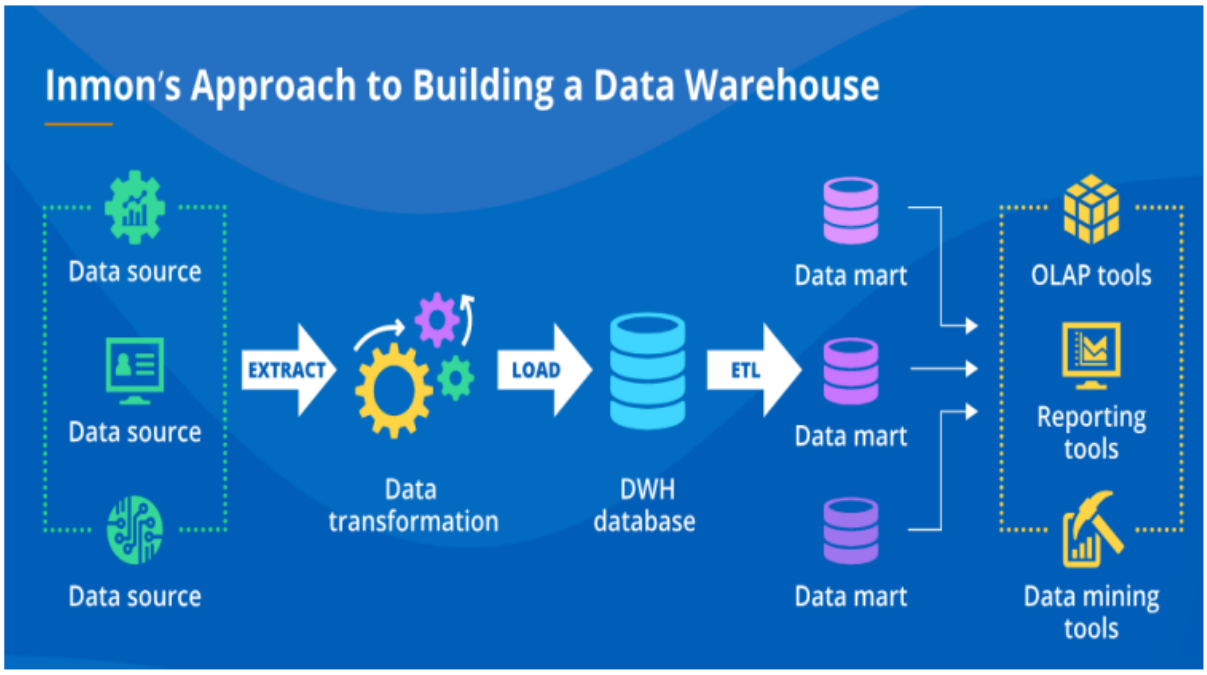
- AI
- Deep learning
- Machine learning
- Statistical modeling
- Natural language processing (NLP)

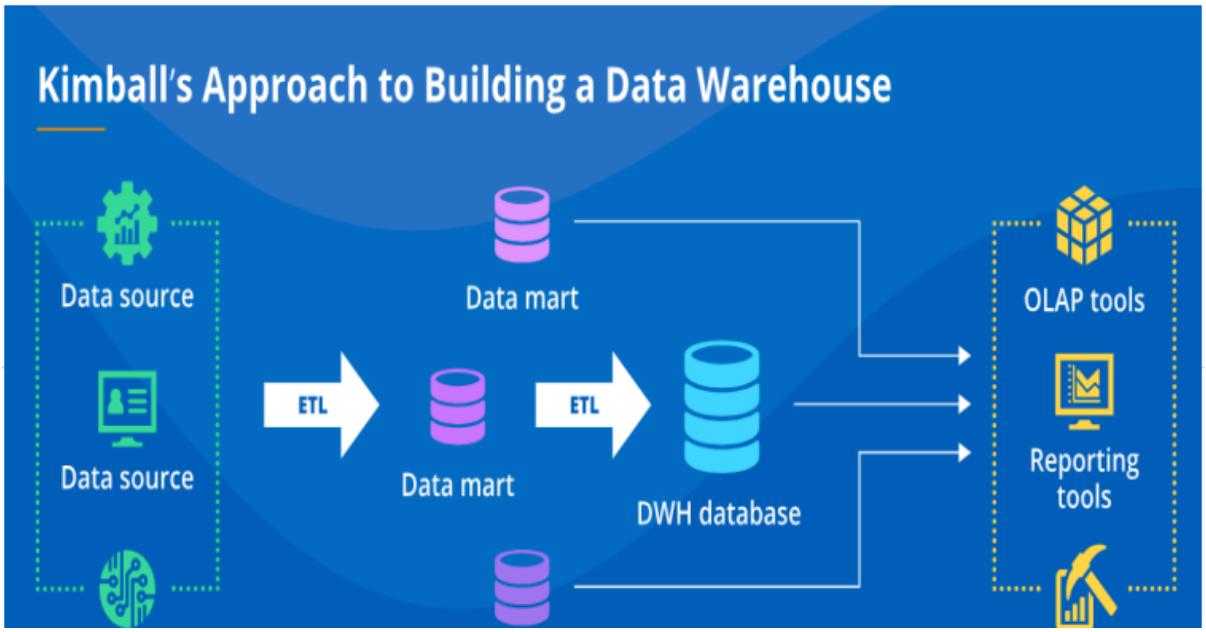
Benefits of modern data warehouses:

- Lower upfront costs
- Less maintenance
- Faster speeds
- More flexibility
- Easier to scale

STEPS IN BUILDING A DATA WAREHOUSE:

1. Identify Business Requirements: Understand the goals and objectives of the organization, and determine the data requirements for decision-making.
2. Data Source Identification: Identify and gather data from various internal and external sources, including operational databases, legacy systems, spreadsheets, and external data feeds.
3. Data Cleaning and Transformation: Cleanse and transform the raw data to ensure consistency, accuracy, and compatibility with the data warehouse schema. This may involve tasks like removing duplicates, correcting errors, and standardizing formats
4. Data Modeling: Design the data warehouse schema, including dimensions, facts, and relationships between them. Common modeling techniques include star schema and snowflake schema.
5. ETL Process: Extract, transform, and load the data into the data warehouse. This involves extracting data from source systems, transforming it according to the data warehouse schema, and loading it into the appropriate tables.
6. Metadata Management: Implement metadata to document and manage the data assets in the data warehouse. Metadata provides information about the source, structure, and meaning of the data, facilitating data governance and usage.
7. OLAP Cube Design: Design OLAP cubes for multidimensional analysis and reporting. OLAP cubes provide pre-aggregated data views for faster querying and analysis.
8. Implementation: Construct the data warehouse using the chosen technologies, such as relational databases, cloud-based platforms, or hybrid solutions. Test the data warehouse to ensure accuracy, performance, and usability





Information Needed to Support DBMS Schemas for Decision Support:

- Business Requirements: Understanding the specific analytical needs of the organization, including key performance indicators (KPIs), metrics, and reporting requirements.
- Data Sources: Identifying and accessing relevant data sources, including operational databases, external data feeds, and historical data archives.
- Data Quality: Assessing the quality of the data, including completeness, accuracy, consistency, and timeliness. Data quality issues must be addressed through cleansing and transformation processes.
- Data Model: Designing an appropriate data model for the decision support system, including dimensional modeling techniques such as star schema or snowflake schema.
- Metadata: Documenting metadata to provide context and understanding of the data, including data lineage, definitions, and relationships.
- ETL Processes: Implementing ETL processes to extract, transform, and load data into the decision support system, ensuring data consistency and integrity.
- Security and Privacy: Implementing security measures to protect sensitive data and ensure compliance with regulatory requirements.
- Scalability and Performance: Designing the DBMS schemas to support scalable and efficient querying and analysis, including indexing strategies, partitioning, and optimization techniques.