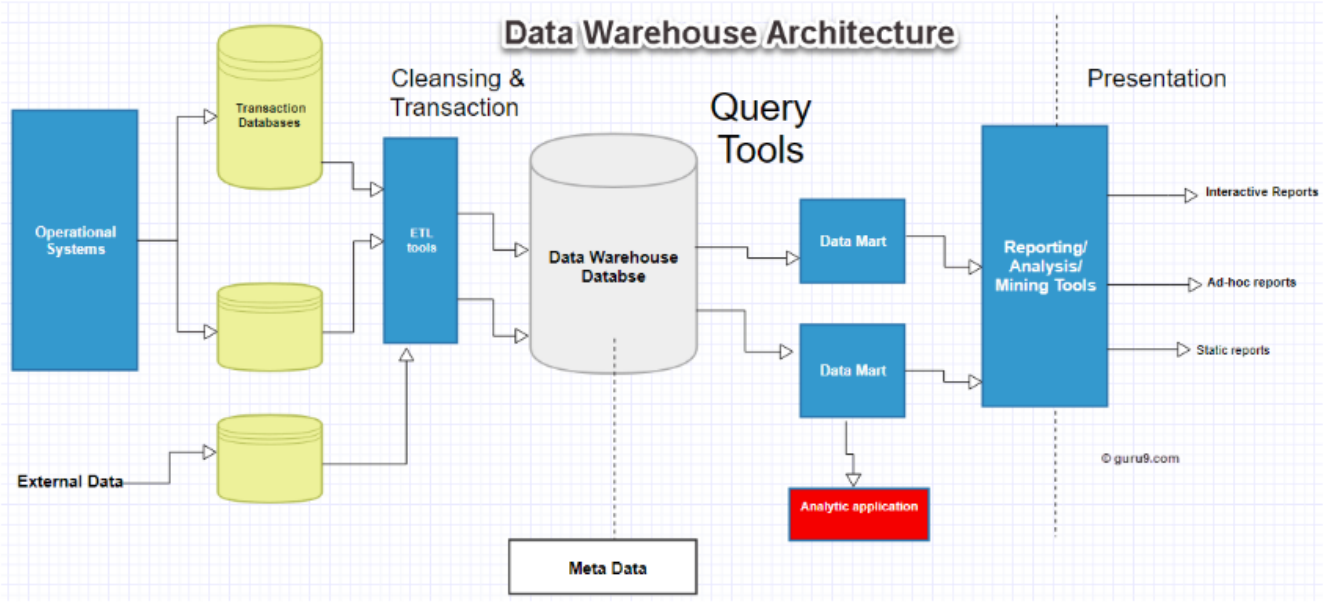
1st Internals Paper

1. Explain the concept of Data Warehouse

A data warehousing is a technique for collecting and managing data from varied sources to provide meaningful business insights.

Datawarehouse Components

We will learn about the Datawarehouse Components and Architecture of Data Warehouse with Diagram as shown below:



Data Warehouse Architecture

2. What is Extract, Transform and Load? Explain

ETL stands for Extract Transform and Load. ETL is a data integration process that combines data from multiple data sources into a single, consistent data store that is loaded in a DW. It is often used to build a data warehouse.

Extract: Extract is the process of fetching (reading) the information from the database. At this stage, data is collected from multiple or different types of sources which can be structured or unstructured.

Transform: Transform is the process of converting the extracted data from its previous form into the required form. Data can be placed into another database.

This can involve Filtering, Cleaning, De-duplication, Validation, Authenticating the data.

Load: Load is the process of writing the data into the target database.

- ETL takes place during off-hours when traffic on the source systems and the DW is at its lowest.

Need of ETL

There are many reasons the need for ETL is arising:

- ETL helps the companies to analyze their business data for making critical business decisions.
- Data warehouse provides a shared data repository.
- ETL provides a method of moving data from various sources into a data warehouse.
- As the data sources change, the data warehouse will automatically update.
- ETL helps to migrate the data into a data warehouse.

3. Explain various components of business intelligence.

The primary components of Business Intelligence:

- **OLAP (Online Analytical Processing)**

This is the component of BI that executives use to sort and select clusters of data for monitoring purposes.

- **Advanced Analytics:**

This component generates useful stats related to specific products and services.

- **Real Time BI**:

This has become a very important component of BI which helps business owners to respond to real time trends. It can help marketing professionals create better deals and limited time offers.

- **Data Warehousing:**

This component allows business managers and owners to analyze different sets of data and analyze patterns that can help at making changes to drive performance.

- **Data Sources:**

This is the component related to the storage of several forms of data. It contains the raw data that can be processed using software and then actionable information can be generated. It also helps at making decisions based on facts at all the organizational levels. The information generated from the data sources can be used throughout the organization in various divisions.

4. Define Metadata and explain metadata types.

A data of a data / Data that provides information about other data is known as Metadata.

Types of Metadata

- Operational Metadata:

Operational Metadata provides information about how the data is being used and the data lifecycle. It describes who has access to use it, who is using it, when it was created, and when it will be due to retire

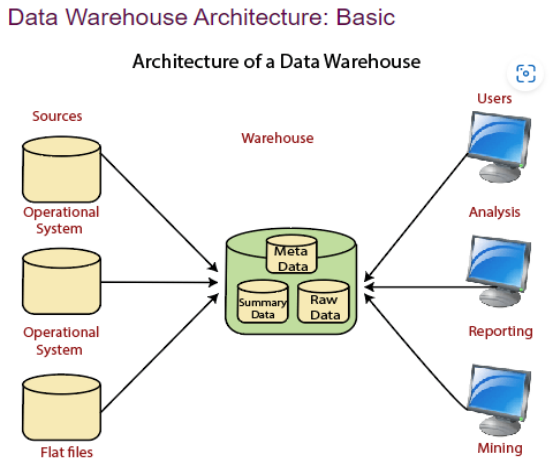
- Extraction and Transformation Metadata

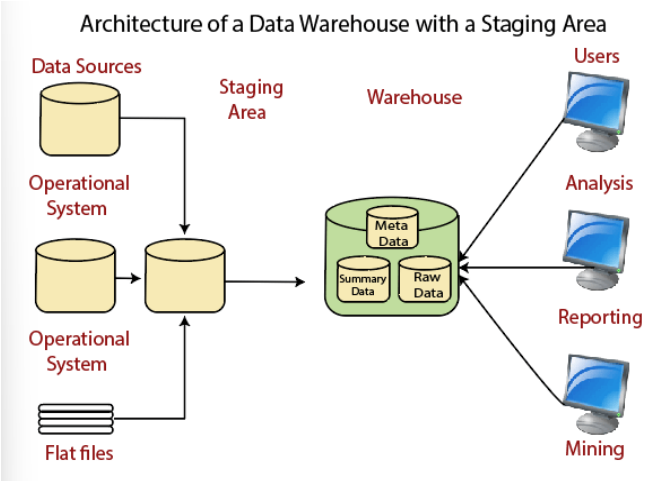
Data on data removal from the source systems are included in extraction and transformation metadata, including frequencies extraction, extraction methods and data extraction regulations for business purposes. This category also provides details on all transformations in the data staging area.

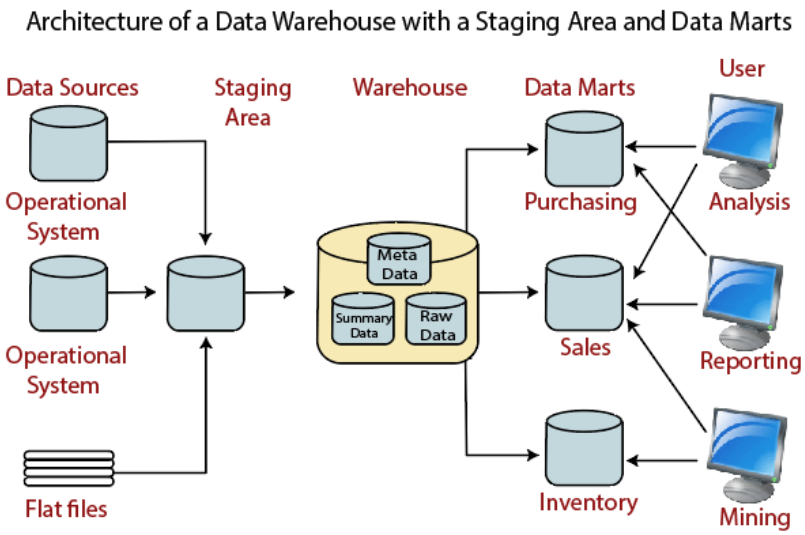
- End-User Metadata

The end-user metadata is the navigational map of the data warehouses. It enables the end-users to find data from the data warehouses. The end-user metadata allows the end-users to use their business terminology and look for the information in those ways in which they usually think of the business.

5. Data warehouse architecture.







6. Explain multidimensional data model with a neat diagram.

Multidimensional data model in data warehouse is a model which represents data in the form of data cubes. It allows to model and view the data in multiple dimensions and it is defined by dimensions and facts. Multidimensional data model is generally categorized around a central theme and represented by a fact table.

A multidimensional database allows to rapidly and reliably providing data-related responses to complicated market questions. The Multidimensional Data Model can be defined as a way to arrange the data in the database, to help structure and organize the contents of the database. The Multidimensional Data Model can include two or three dimensions of objects from the database structure, versus a system of one dimension, such as a list.

(Diagram Important (Error 404: Diagram not found))

7. Differentiate fact tables and dimension table.

S.NO	Fact Table	Dimension Table
1.	Fact table contains the measuring of the attributes of a dimension table.	Dimension table contains the attributes on that truth table calculates the metric.
2.	In fact table, There is less attributes than dimension table.	While in dimension table, there is more attributes than fact table.
3.	In fact table, There is more records than dimension table.	While in dimension table, there is less records than fact table.

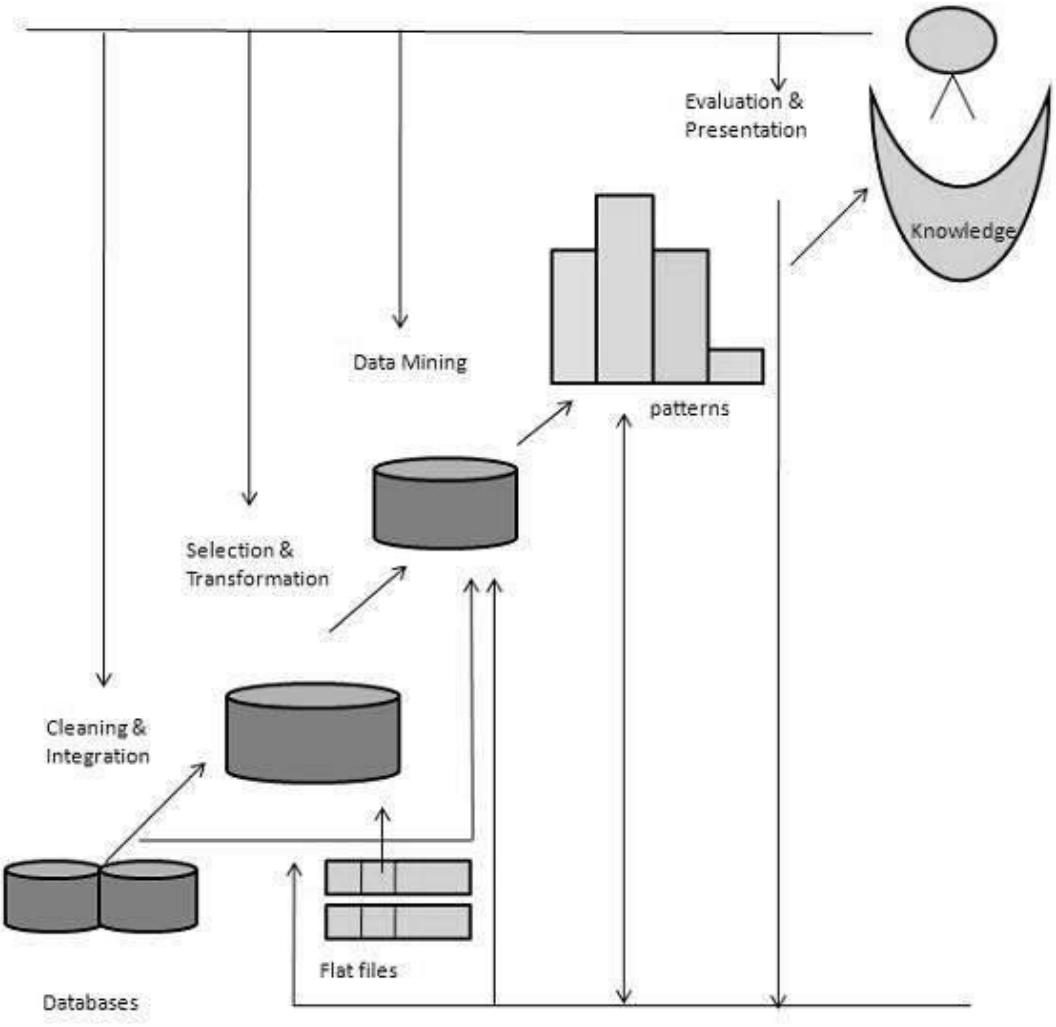
S.NO	Fact Table	Dimension Table
4.	Fact table forms a vertical table.	While dimension table forms a horizontal table.
5.	The attribute format of fact table is in numerical format and text format.	While the attribute format of dimension table is in text format.
6.	It comes after dimension table.	While it comes before fact table.
7.	The number of fact table is less than dimension table in a schema.	While the number of dimensions is more than fact table in a schema.
8.	It is used for analysis purpose and decision making.	While the main task of dimension table is to store the information about a business and its process.

2nd Internals Paper

1. Explain with a neat diagram the steps involved in the knowledge discovery in database process.

Here is the list of steps involved in the knowledge discovery process −

- **Data Cleaning** − In this step, the noise and inconsistent data is removed.
- **Data Integration** − In this step, multiple data sources are combined.
- **Data Selection** − In this step, data relevant to the analysis task are retrieved from the database.
- **Data Transformation** − In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- **Data Mining** − In this step, intelligent methods are applied in order to extract data patterns.
- **Pattern Evaluation** − In this step, data patterns are evaluated.
- **Knowledge Presentation** − In this step, knowledge is represented.



2. (Study the problem)

3. What is meant by Outlier? How these outliers are detected using data mining.

In data mining, outliers are data points that deviate significantly, or in simpler terms are "far away", from the rest of the data point.

Outliers can be in both the univariate and multivariate forms.

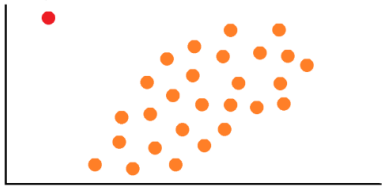
- Univariate outliers are observations that significantly deviated values from the distribution of one variable.
- Multivariate outliers are extreme values issued from multiple variables.

Types:

a) Global outliers

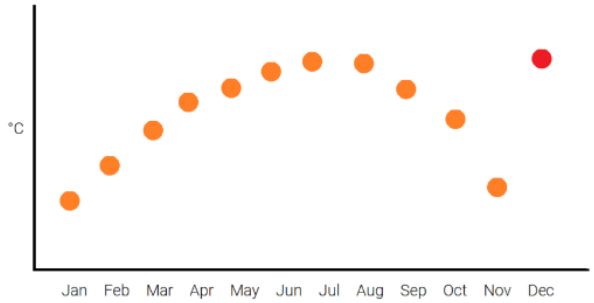
Global outliers or points anomalies are data points that deviate strongly from the rest of the points. If you plot them it would come to you as quite "obvious". Most, if not all, outliers' detection techniques attempt to identify global outliers.

An example of global outliers could be a very large order received in a day or a spike in a time series.



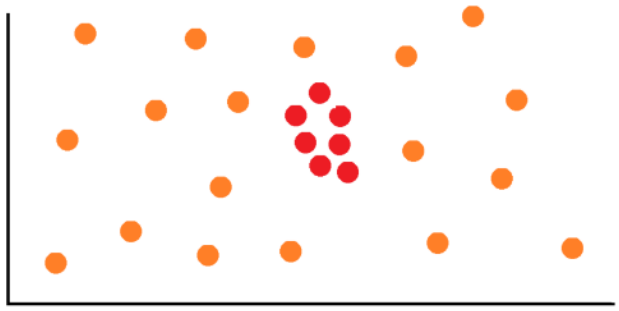
b) Contextual outliers

Contextual outliers, also called conditional outliers, are extreme observations that deviate from the rest of the observations based on a specific condition



c) Collective outliers

Collective outliers are a subset of observations whose values as a group deviate significantly from the rest of the observations.



4. Explain the data pre-processing techniques in detail.

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

Data Preprocessing Techniques:

- **Data cleaning** can be applied to remove noise and correct inconsistencies in the data.
- **Data integration** merges data from multiple sources into coherent data store, such as a data warehouse.
- **Data reduction** can reduce the data size by aggregating, eliminating redundant features, or clustering, for instance. These techniques are not mutually exclusive; they may work together.
- **Data transformations**, such as normalization, may be applied.

Need for preprocessing:

Incomplete, noisy and inconsistent data are common place properties of large real world databases and data warehouses.

Incomplete data can occur for a number of reasons:

 Attributes of interest may not always be available

 Relevant data may not be recorded due to misunderstanding, or because of equipment malfunctions.

 Data that were inconsistent with other recorded data may have been deleted.

 Missing data, particularly for tuples with missing values for some attributes, may need to be inferred.
 The data collection instruments used may be faulty.
 There may have been human or computer errors occurring at data entry.
 Errors in data transmission can also occur.
 There may be technology limitations, such as limited buffer size for coordinating synchronized data transfer and consumption.
 Data cleaning routines work to ―clean‖ the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.
 Data integration is the process of integrating multiple databases cubes or files. Yet some attributes representing a given may have different names in different databases, causing inconsistencies and redundancies.
 Data transformation is a kind of operations, such as normalization and aggregation, are additional data preprocessing procedures that would contribute toward the success of the mining process.
 Data reduction obtains a reduced representation of data set that is much smaller in volume, yet produces the same (or almost the same) analytical results.

5. Explain Data Objects and Attribute types with example.
(Couldn't find a proper answer for this ☹)

6. Explain the four types of Machine Learning.

Machine learning is a subset of AI, which enables the machine to automatically learn from data, improve performance from past experiences, and make predictions. Machine learning contains a set of algorithms that work on a huge amount of data. Data is fed to these algorithms to train them, and on the basis of training, they build the model & perform a specific task.

Types:

1) **Supervised Machine Learning** is based on supervision. It means in the supervised learning technique, we train the machines using the "labelled" dataset, and based on the training, the machine predicts the output.

   Supervised machine learning can be classified into two types of problems, which are given below:

- Classification

  Classification algorithms are used to solve the classification problems in which the output variable is categorical, such as "Yes" or No, Male or Female, Red or Blue, etc. The classification algorithms predict the categories present in the dataset. Some real-world examples of classification algorithms are Spam Detection, Email filtering, etc.

Some popular classification algorithms are given below:

- Random Forest Algorithm
- Decision Tree Algorithm
- Logistic Regression Algorithm
- Support Vector Machine Algorithm

- Regression

  Regression algorithms are used to solve regression problems in which there is a linear relationship between input and output variables. These are used to predict continuous output variables, such as market trends, weather prediction, etc.

  Some popular Regression algorithms are given below:

  - Simple Linear Regression Algorithm
  - Multivariate Regression Algorithm
  - Decision Tree Algorithm
  - Lasso Regression

2) **Unsupervised Machine Learning** is different from the Supervised learning technique; as its name suggests, there is no need for supervision. It means, in unsupervised machine learning, the machine is trained using the unlabelled dataset, and the machine predicts the output without any supervision.

   In unsupervised learning, the models are trained with the data that is neither classified nor labelled, and the model acts on that data without any supervision.

   The main aim of the unsupervised learning algorithm is to group or categories the unsorted dataset according to the similarities, patterns, and differences.

   Categories of Unsupervised Machine Learning

   Unsupervised Learning can be further classified into two types, which are given below:

   - Clustering
   - Association

   1.. Clustering

   The clustering technique is used when we want to find the inherent groups from the data. It is a way to group the objects into a cluster such that the objects with the most similarities remain in one group and have fewer or no similarities with the objects of other groups.

Some of the popular clustering algorithms are given below:

- K-Means Clustering algorithm
- Mean-shift algorithm
- DBSCAN Algorithm
- Principal Component Analysis

2.. Association

Association rule learning is an unsupervised learning technique, which finds interesting relations among variables within a large dataset. The main aim of this learning algorithm is to find the dependency of one data item on another data item and map those variables accordingly so that it can generate maximum profit. This algorithm is mainly applied in Market Basket analysis, Web usage mining, continuous production, etc.

3) **Semi-Supervised Machine Learning** is a type of Machine Learning algorithm that lies between Supervised and Unsupervised machine learning. It represents the intermediate ground between Supervised (With Labelled training data) and Unsupervised learning (with no labelled training data) algorithms and uses the combination of labelled and unlabelled datasets during the training period.
4) **Reinforcement Learning** works on a feedback-based process, in which an AI agent (A software component) automatically explore its surrounding by hitting & trail, taking action, learning from experiences, and improving its performance. In reinforcement learning, there is no labelled data like supervised learning, and agents learn from their experiences only.

7. Explain Naïve Bayesian classification in detail with example
(Send help☹)