COMPUTER ENGINEERING DEPARTMENT

ASSIGNMENT NO-01

SUB: Data Warehousing & Mining

COURSE: T.E. **Year: 2020-2021** **Semester: VI**
DEPT: Computer Engineering
SUBJECT CODE: CSC603 **SUBMISSION DATE: 04/05/2021**

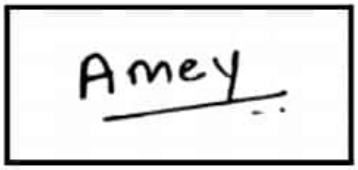
==========

Name: Amey Thakur **Roll No.: 50**

Class: TE Comps B-50 **ID: TU3F1819127**

Assignment No 1

Sr. No.	Question	CO mapping
1	Explain the architecture of the data warehouse, approaches to designing a data warehouse.	CO1
2	Write a short note on Metadata and its types.	CO1
3	Explain major steps of ETL in detail with a diagram.	CO2
4	Explain OLAP operations in detail with an example.	CO2
5	Explain the KDD process and different data mining techniques.	CO3
6	Explain different Data transformation and Data extraction techniques.	CO3



AMEY B 50 Amey :

Page No.: Date: YOUVA

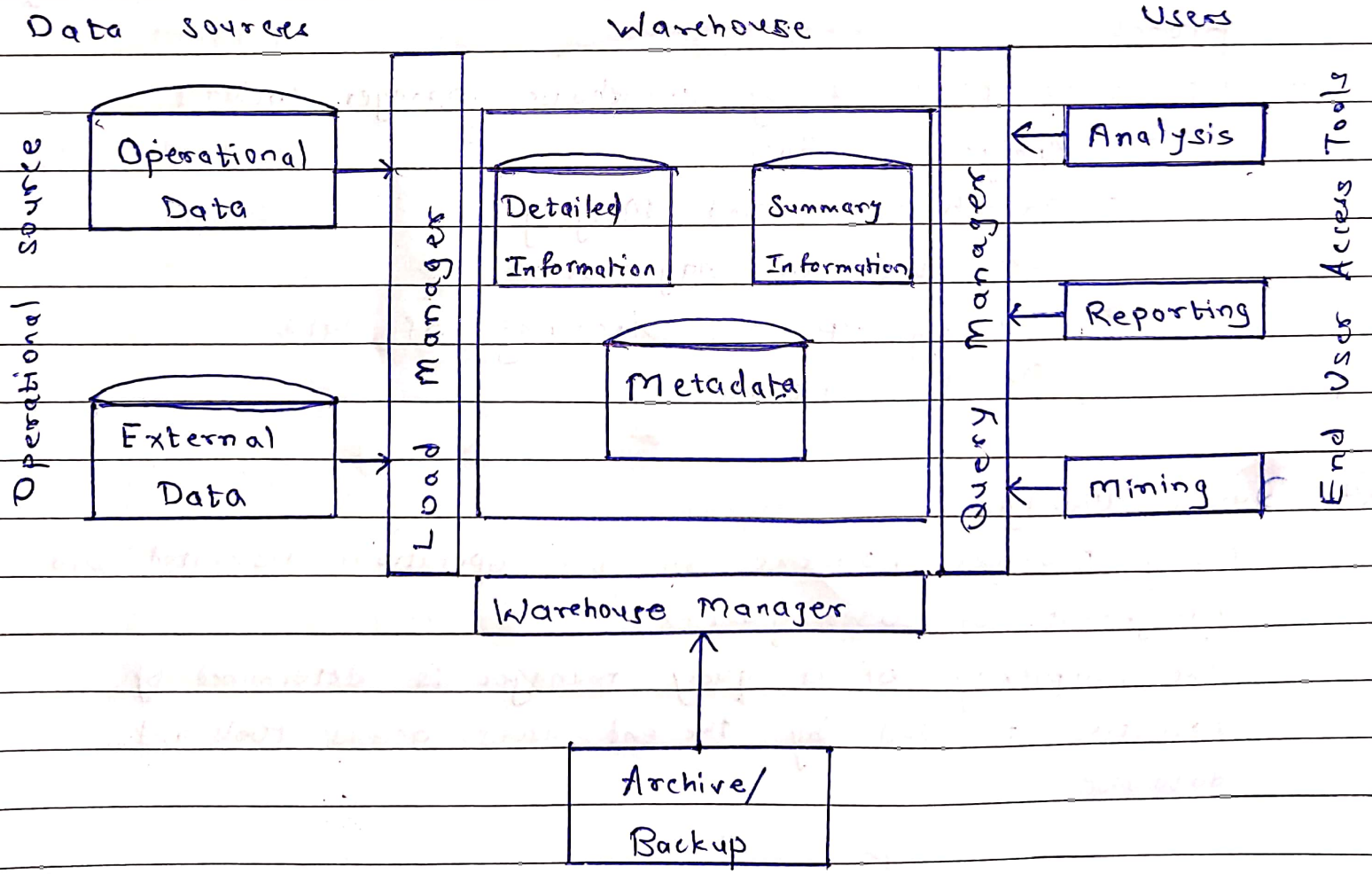
Q1 Explain the architecture of the data warehouse, approaches to design a data warehouse.

Ans:

Data Warehouse

- Data warehouse is constructed by integrating data from multiple heterogeneous sources.
- It is integrated, subject oriented, time variant and non-volatile collection of data.
- It was defined by Bill Inmon in 1990.
- Data warehouse is a system used by reporting and data analysis.
- It is constructed as a core component of business intelligence.

Architecture of Data Warehouse



AMEY B 50 Amey

Page No.:
Date:
YOUVA

Data warehouse architecture consist of following components:

① Operational source
- Operational source is a data source consists of operational data and external data.
- Data can come from Relational DBMS like Oracle, Informix.

② Load Manager
- Load manager performs all the operations required to Extract and Load data
- The size and complexity of the load manager varies between specific solutions from one data warehouse to other.

③ Warehouse Manager
- Warehouse Manager is responsible for the warehouse management process.
- Operations performed by warehouse manager includes:
  → Analysis of data
  → Transformation and Merging
  → Generation of Aggregation
  → Backing up and Archiving of Data
  → De-Normalization

④ Query Manager
- Query manager performs all the operations associated with management of user queries.
- The complexity of a query manager is determined by facilities provided by the end users access tools and database.

⑤ Detailed data

- It is used to store all the detailed data in the database schema.
- Detailed data is loaded into the data warehouse to supplement the aggregated data.

⑥ Summarized data

- Summarized data is a part of data warehouse that stores predefined aggregations.
- These aggregations are generated by the warehouse manager.

⑦ Archive and Backup data

- The detailed and summarized data are stored for the purpose of archiving and backup.
- The data is transferred to storage archives such as magnetic tapes or optical disks.

⑧ Metadata

- Metadata is basically Data stored above Data
- It is used for extraction and loading process, warehouse management process and query management process

⑨ End User Access Tools

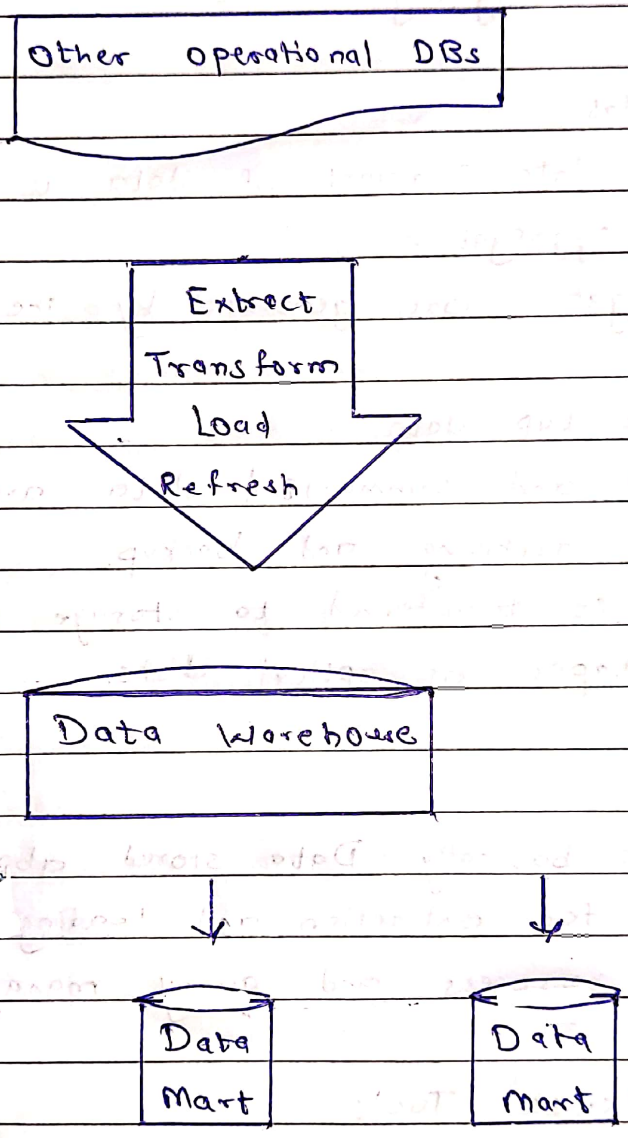
- End User Access Tools consist of Analysis, Reporting and Mining.
- The users interacts with warehouse using end user access tools.

AMEY B 50 Amey

Page No.:
Date:
YOUVA

Data Warehouse Design Approaches.

① Top Down Approach

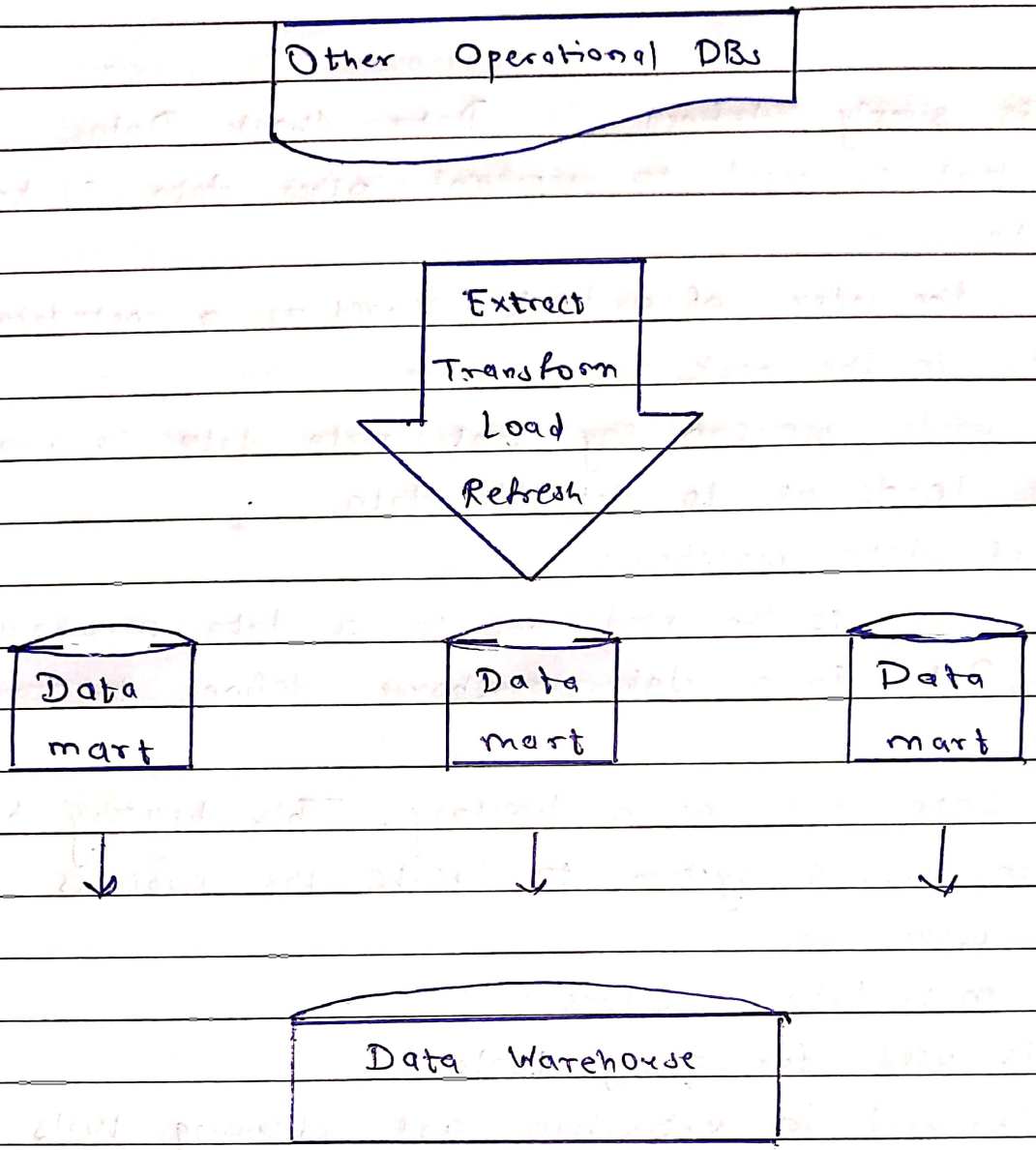


- In this approach, the data flow begins with data extraction
- This data is then loaded into staging area
- It is then transferred to operational Data store (ODS).
- Sometimes the ODS step is skipped, if it is replication of operational database.
- Data is loaded into data warehouse in a parallel process to avoid extracting it from the ODS.
- Then the data mart is loaded with data
- Finally OLAP environment is available for the users

AMEY B 50 Amey

Page No.:
Date:
YOUVA

Bottom up Approach.



- The position of data warehouse and data mart are reversed in bottom up approach
- The data flow begins with extraction of data from operational databases into the staging area
- Data is then loaded into ODS.
- The data in ODS is appended to or refreshed by the fresh data.
- It is then processed to fit into the data mart structure
- The data from the data mart is then extracted to the staging area aggregated, summarized and so on.
- It is then loaded into the data warehouse
- Finally it is made available to the end users for analysis.

AMEY B 50 Amey

Page No.:
Date:
YOUVA

Q2. Write a short note on meta data and its types.

Ans:

Metadata

- Metadata is simply defined as Data About Data
- The data that is used to represent other data is known as metadata
- For example, the index of a book serves as a metedata for the contents in the book.
- In other words, we can say that meta data is summarized data that leads us to detailed data
- In terms of data warehouse,
  1. Meta Data is the road map to a data warehouse.
  2. Meta Data in a data warehouse defines the warehouse objects.
  3. meta Data acts as a directory. This directory helps the decision support system to locate the contents of a data warehouse.
- Roles of metadata includes -
  1. It is used for query tools
  2. It is used in extraction and cleansing tools
  3. It is used in reporting tools
  4. It is used in transformation tools.
  5. It plays an important role in loading functions.

AMEY B 50 Amey

Page No.:
Date:
YOUVA

Types of meta data

① Operational metadata
- In data warehouse, data comes from several operational systems of the enterprise.
- Different source system contains different data structures.
- The data elements selected for the data warehouse have various field lengths and data types.
- So the information of operational data source is given by operational metadata.

② Extraction and Transformation Metadata.
- It contains the information about extraction of data from heterogeneous source system
- It also contains the information about data transformation in data staging area.

③ End user Metadata.
- It is the navigational map of data warehouse
- It enables the end user to find information from the data warehouse.
- The end user meta data allows the end user to use their own business terminology and look for information in those ways in which they normally think of the business.

AMEY B 50 Amey

Page No.: YOUVA
Date:

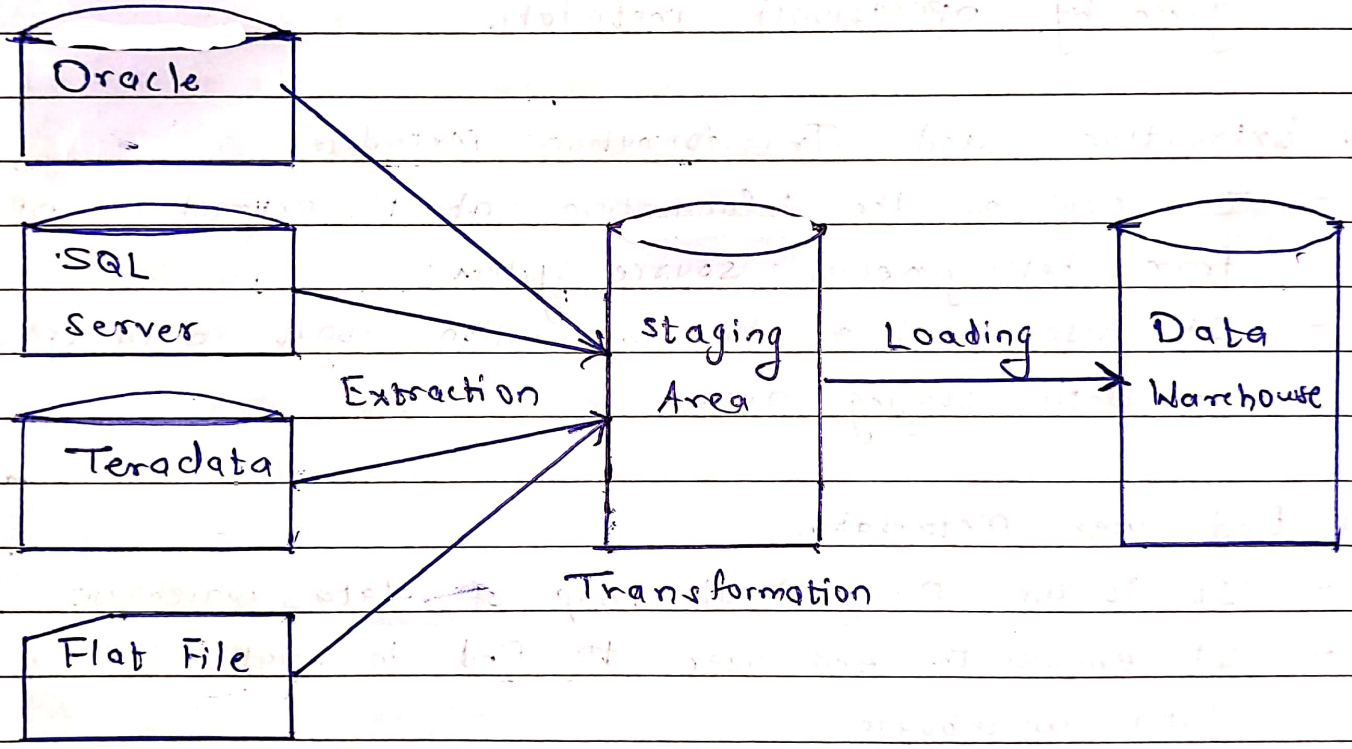
Q3. Explain major steps of ETL in detail with a diagram

Ans:

ETL

- ETL stands for Extract, Transform and Load.
- It is a process in data warehousing responsible for pulling data out of the source systems and placing it into a data warehouse.

ETL Process



- Above figure represents the ETL process.
- ETL process is the way to move and prepare data for data analysis.
- ETL process involves the following tasks:
  1. Extracting the data from different sources
  2. Transforming the data
  3. Loading

AMEY B 50 Amey

Page No.:
Date:
YOUVA

① Extracting the data from different sources
- This is the first step in ETL process.
- Different data sources can be RDBMS or files.
- In this step, Data is extracted from source system
- Data is also made accessible for further processing.
- The main objective of the extraction step is to retrieve all required data from source system.
- The extraction step should be designed in a way that it does not negatively affect the source system
- Most data projects consolidate data from different source systems.
- Each separate source uses a different format.
- Common data-source formats includes RDBMS, XML.
- Thus the extraction process must convert the data into a format suitable for further transformation

② Transforming the data
- This may involve cleaning, filtering, validating, applying business rules.
- In this step, certain rules are applied on the extracted data
- The main aim of this step is to load the data to the target database. in a cleaned and general format.
- This is because when the data is collected from different sources each source will have their own standards.
- In some cases data does not need any transformations and here the data is said to be rich data, direct move or pass through data

③ Loading

- This is the final step in the ETL process.
- In this step, the extracted data and transformed data is loaded to the target database.
- In order to make data load efficient, it is necessary to index the database and disable constraints before loading the data.
- All the three steps in ETL process can be run parallel.
- Data extraction takes time and so the second step of transformation process is executed simultaneously.
- This prepares data for the third step of loading.
- As soon as some data is ready, it is loaded without waiting for completion of the previous steps.

AMEY B 50 Amey

Page No.: Date: youva

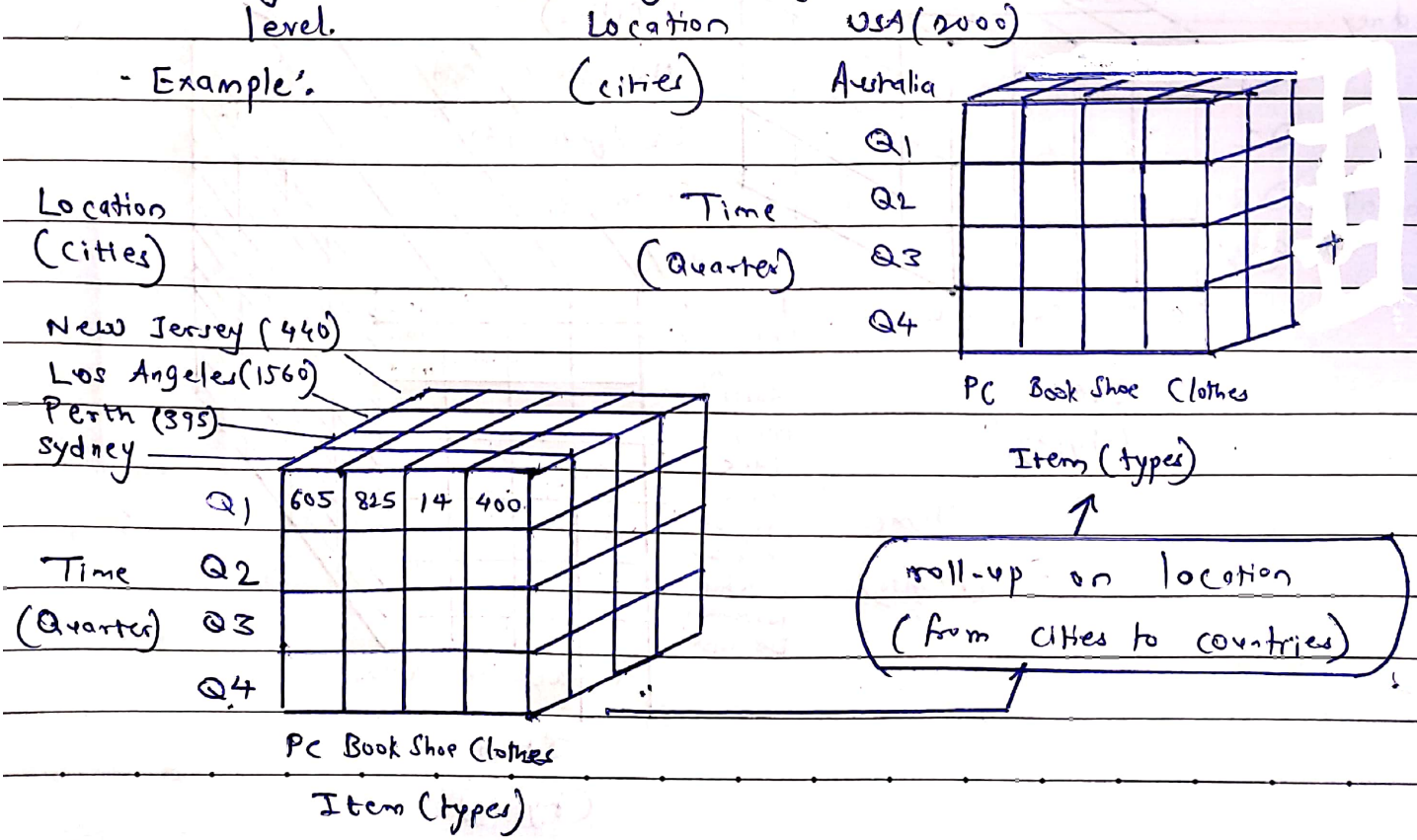
Q4. Explain OLAP operations in detail with an example.

Ans:

OLAP Operations

- OLAP operations are implemented to retrieve the information from data warehouse into OLAP multi-dimensional databases.
- Since OLAP servers are based on multidimensional view of data, so OLAP operations are performed in multidimensional data
- List of OLAP operations.
  1. Roll-up
  2. Drill-down
  3. Slice and Dice
  4. Pivot (rotate)

1. Roll-up
   - Roll-up is also known as "consolidation" or "aggregation".
   - The roll-up operation can be performed in two ways
     1. Reducing dimensions
     2. Climbing up concept hierarchy. Concept hierarchy is a system of grouping things based on their order or level.
   - Example:



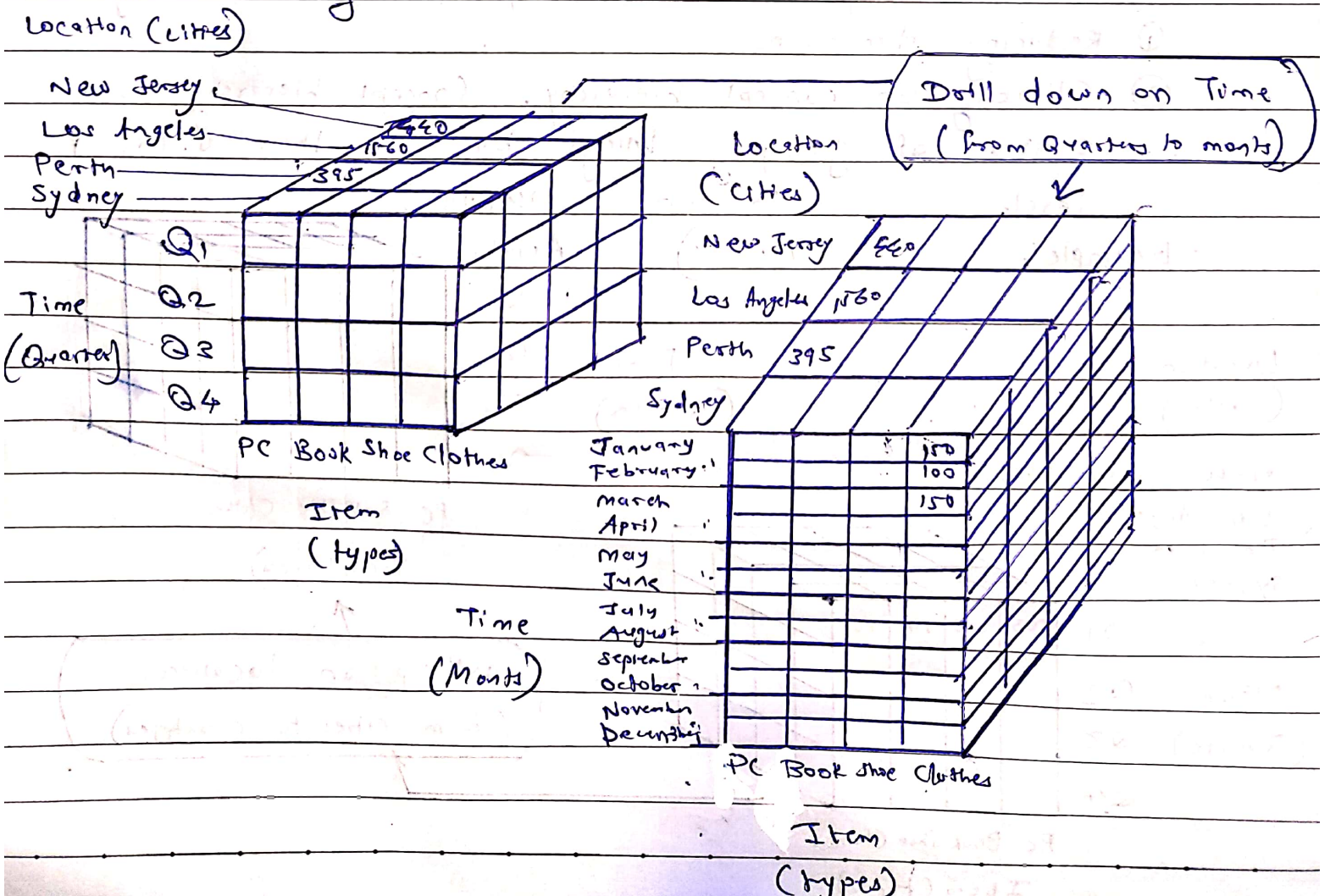
AMEY B 50 Amey:

Page No.:
Date:
YOUVA

- In this example, cities New Jersey and Los Angeles and rolled up into country USA
- The sales figure of New Jersey and Los Angeles are 440 and 1560 respectively. They become 2000 after roll up.
- In this aggregation process, data is location hierarchy moves up from city to the country
- In the roll up process at least one or more dimensions need to be removed, Quarter dimensions is removed

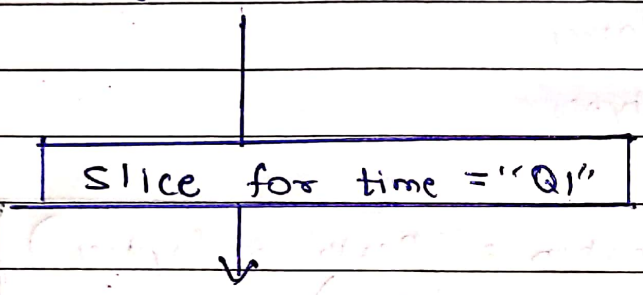
(2) Drill - down

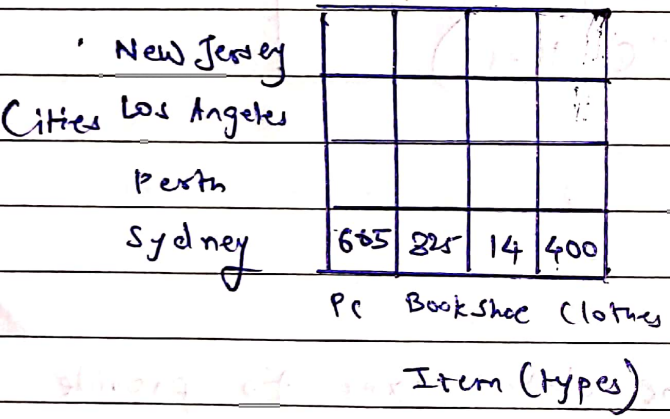
- In drill-down data is fragmented into smaller parts
- It is the opposite of the roll-up process. It can be done via
  (1) moving down the concept hierarchy
  (2) Increasing a dimension



Consider the above diagram
- Quarter Q1 is drilled down to months January, February and March.
- Corresponding sales are also registers.
- In this example, dimension months are added

③ Slice
- Here, one dimension is selected, and a new sub cube is created.





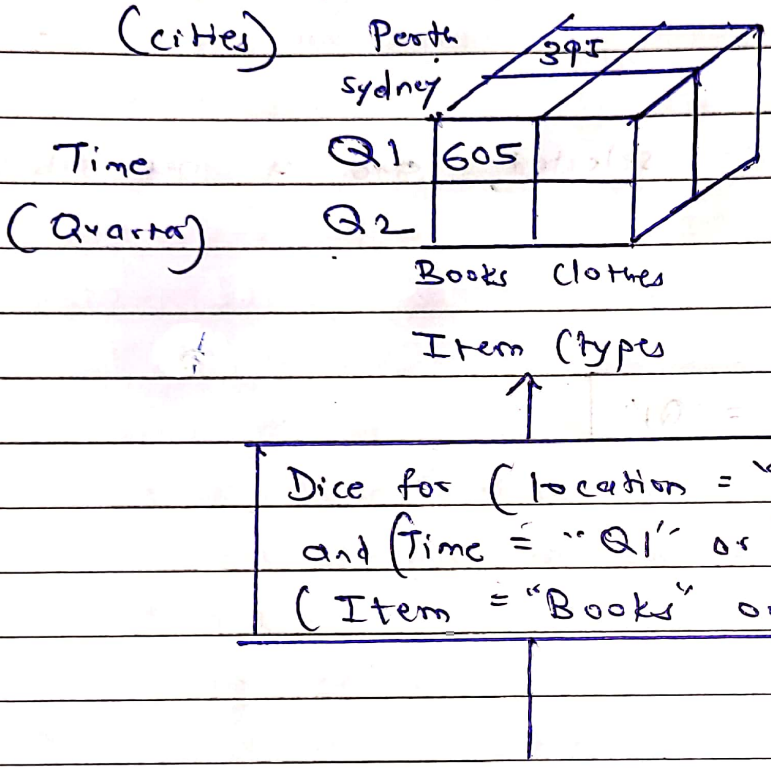
- Dimension Time is sliced with Q1 as the filter
- A new cube is created altogether.

AMEY B 50 Amey

Page No.:
Date:
YOUVA

Dice

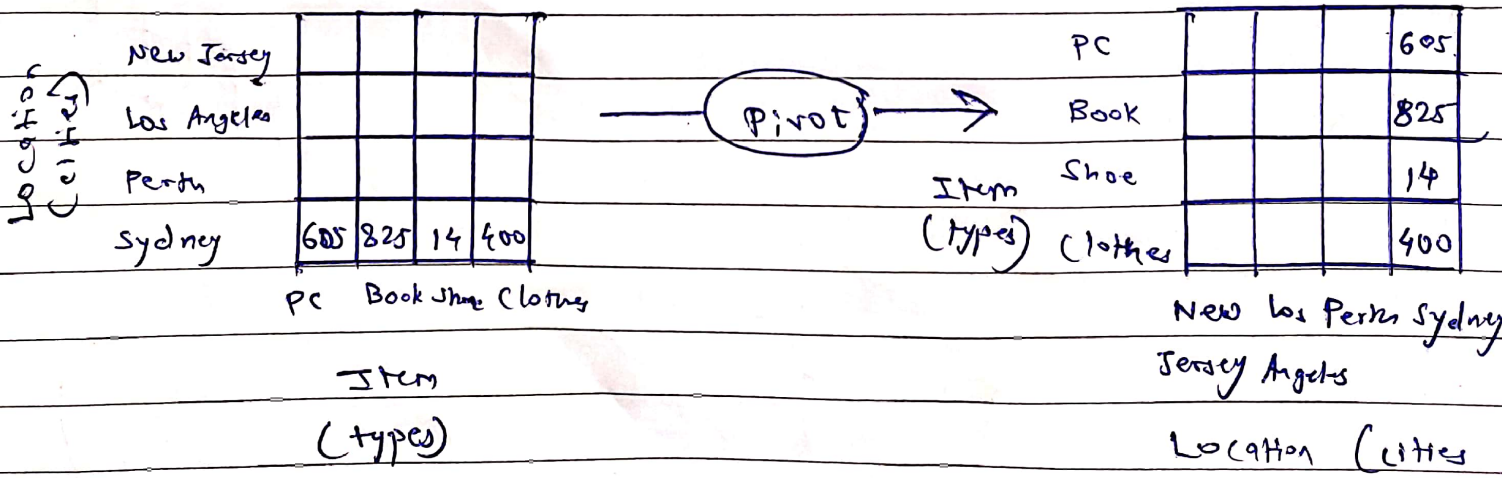
- This operation is similar to a slice
- The difference in dice is you select 2 or more dimensions that result in creation of a sub-cube



Dice for (location = "Perth" or "Sydney") and (Time = "Q1" or "Q2") and (Item = "Books" or "Clothes")

(4) Pivot

- In pivot, you rotate the data axes to provide a substitute presentation of data
- In following example, the pivot is based on item (types).



AMEY B 50 Amey

Page No.:
Date:
YOUVA

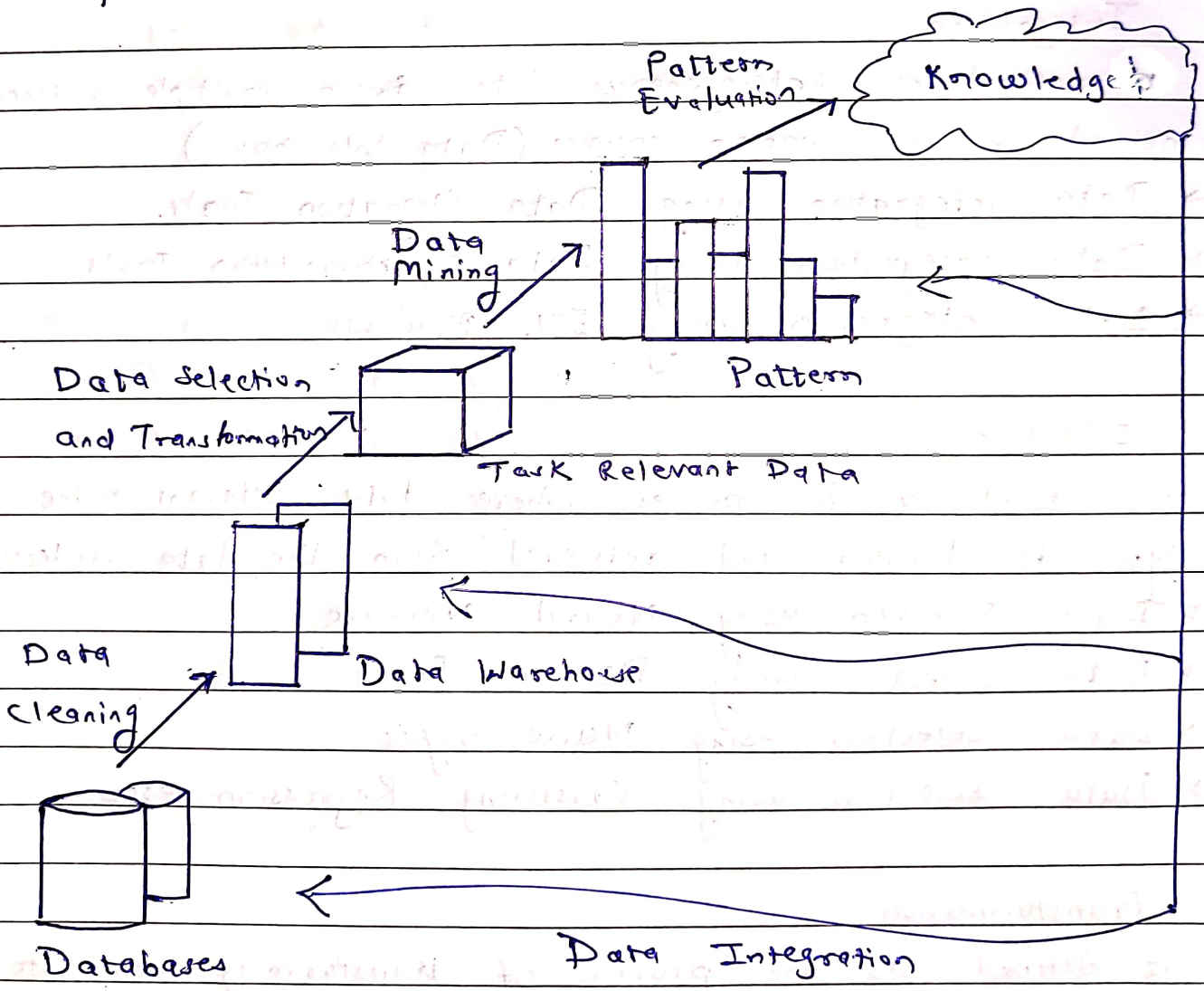
Q5. Explain the KDD process and different data mining technique

Ans:

KDD Process

- Data mining is also known as Knowledge Discovery in Databases, refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data stored in databases.

Steps involved in KDD process



KDD Process

AMEY B 50 Amey

Page No.:
Date:
YOUVA

① Data Cleaning
- It is defined as removal of noisy and irrelevant data from collection
  → Cleaning in case of Missing values
  → Cleaning noisy data, where noise is a random or variance error.
  → Cleaning with data discrepancy detection and data transformation tools.

② Data Integration
- It is defined as heterogeneous data from multiple sources combined in a common source (Data Warehouse)
  → Data integration using Data Migration Tools.
  → Data integration using Data Synchronization Tools.
  → Data integration using ETL process.

③ Data Selection
- It is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.
  → Data Selection using Neural Network
  → Data Selection using Decision Trees.
  → Data Selection using Naive bayes
  → Data Selection using Clustering, Regression, etc.

④ Data Transformation
- It is defined as the process of transforming data into appropriate form required by mining procedure
- It is a two step process
  (1) Data Mapping - Assigning elements from source base to destination to capture transformations.
  (2) Code generation - Creation of the actual transformation program.

⑤ Data Mining
- It is defined as clever techniques that are applied to extract patterns potentially useful.
  → Transform task relevant data into patterns
  → Decides purpose of model using classification or characterization

⑥ Pattern Evaluation
- It is defined as identifying strictly increasing patterns representing knowledge based on given measures.
  → Find interestingness score of each pattern
  → Uses summarization and visualization to make data understandable by user.

⑦ Knowledge Representation
- It is defined as technique which utilizes visualization tools to represent data mining results.
  → Generate reports
  → Generate Tables
  → Generate discriminant rules, etc

AMEY B 50 Amey

Page No.: YOUVA
Date:

Data Mining Techniques

(1) Classification

- This analysis is used to retriere important and relevant information about data and metadata This data mining method helps to classify data in different classes.

(2) Clustering

- Clustering analysis is a data mining technique to identify data that are like each other. This process helps to understand the differences and similarities between the data.

(3) Regression

- Regression analysis is a data mining technique to identify and analyze the relationship between variables. It is used to identify the likelihood of a specific variable, given the presence of other variables

(4) Association Rules

- This data mining technique helps to find the association between two or more items. It discovers a hidden pattern in the data set.

(5) Outer Detection

- This type of data mining technique refers to observation of data items in the data-set which do not match an expected pattern or expected behavior. This technique can be used in variety of domains such as intrusion detection, fraud detection, etc.

AMEY B 50 Amey

Page No.:
Date:
YOUVA

(6) Sequencial Patterns

- This data mining technique helps to discover or identify similar patterns or trends in transaction data for certain period.

(7) Prediction

- Prediction has used a combination of the other techniques of data mining like trends, sequential patterns, clustering, classification, etc. It analyzes past events or instances in a right sequence for predicting a future event.

AMEY B 50 Amey.

Page No.:
Date:
YOUVA

Q6. Explain different data transformation and data extraction techniques.

Ans:

Data Transformation Techniques

① Smoothing

- It is a process that is used to remove noise from the dataset using some algorithms. It allows for highlighting important features present in the dataset. It helps in predicting the patterns. When collecting data, it can be manipulated to eliminate or reduce any variance or any other noise form.
- The concept behind data smoothing is that it will be able to identify simple changes to help predict different trends and patterns. This serves as a help to analysts or traders who need to look at a lot of data which can often be difficult to digest for finding patterns that they wouldn't see otherwise.

② Aggregation

- Data collection or aggregation is the method of storing and presenting data in summary format. The data may be obtained from multiple data sources to integrate these data sources into a data analysis description. This is the crucial step since the accuracy of data analysis insights is highly dependent on the quantity and quality of the data used. Gathering accurate data of high quality and a large enough quantity is necessary to produce relevant results.
- For example, Sales data may be aggregated to compute monthly and annual total amounts.

AMEY B 150 Amey

Page No.: YOUVA
Date:

③ Discretization

- It is a process of transforming continuous data into set of small intervals. Most Data Mining activities in the real world require continuous attributes. Yet many of the existing data mining frameworks are unable to handle these attributes.
- Also, even if a data mining task can manage a continuous attribute, it can significantly improve its efficiency by replacing a constant quality attribute with its discrete values.
- For example, (1-10, 11-20) (age:- young, middle age, senior).

④ Attribute construction

- Where new attributes are created and applied to assist the mining process from the given set of attributes. This simplifies the original data and makes the mining more efficient.

⑤ Generalization

- It converts low level data attributes to high level data attributes using concept hierarchy.
- For example, Age initially in numeric form (22, 25) is converted into categorical value (young, old).

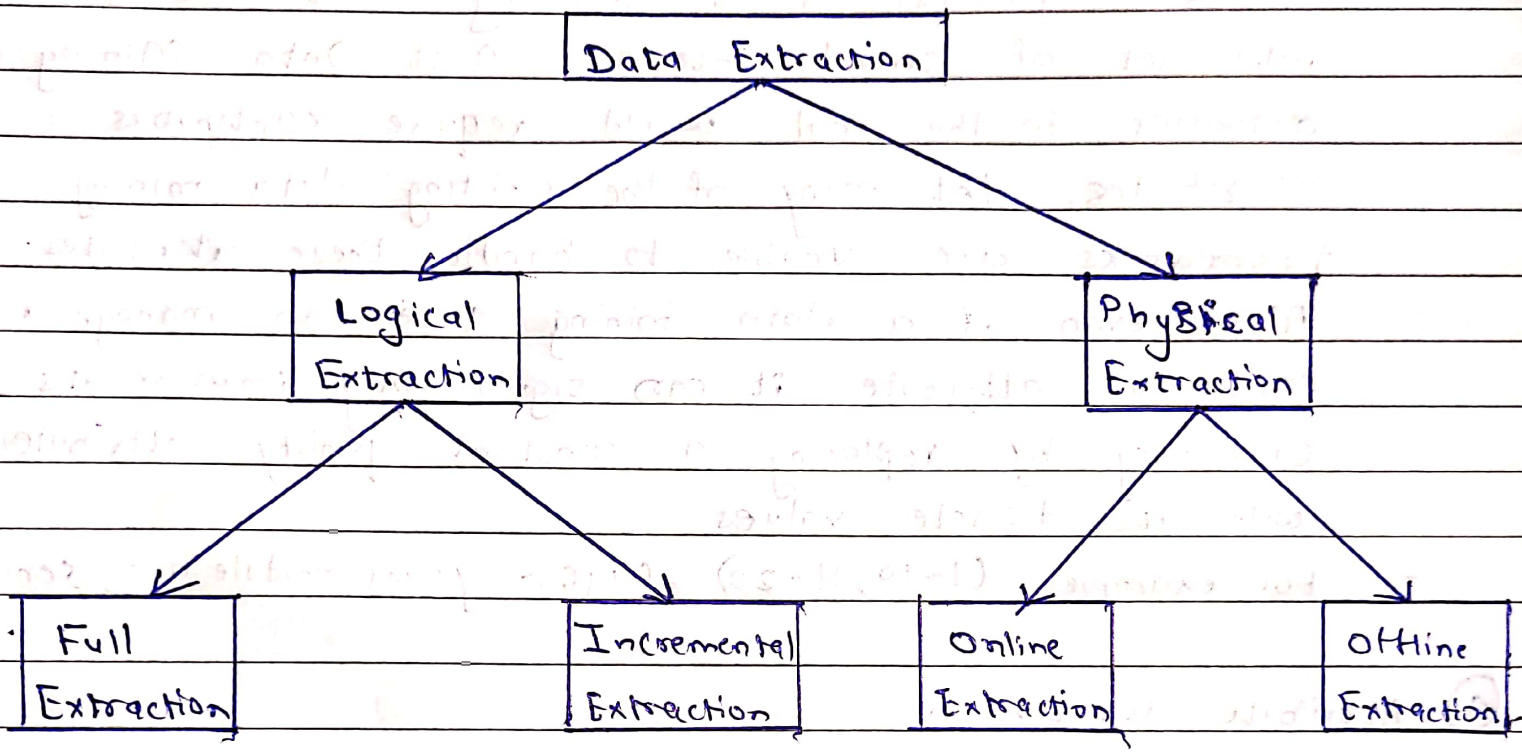
⑥ Normalization

- Data normalization involves converting all data variable into a given range.
- Techniques that are used for normalization are:
  1. Min-Max Normalization
  2. Z-Score Normalization
  3. Decimal Scaling

AMEY B 50 Amey

Page No.: YOUVA
Date:

Data Extraction Techniques



There are two types of data extraction techniques

1. Logical Extraction
2. Physical Extraction.

Logical Extraction.

1. Full Extraction

- In this method, data is completely extracted from the source system. The source data will be provided as-is and no additional logic information is necessary on the source system. Since it is complete extraction, so no need to track source system for changes.
- For example, exporting complete table in the form of flat file.

AMEY B 50 Amey

Page No.:
Date:
YOUVA

(2) Incremental extraction

- In incremental extraction, the changes in source data need to be tracked since the last unsuccessful extraction. Only these changes in data will be extracted and then loaded. Identifying the last changed data itself is the complex process and involve many logic.
- You can detect the changes in the source system from the specific column in the source system that has the last changed timestamp. You can also create a change table in the source system, which keeps track of the changes in the source data.

Physical Extraction

(1) Online Extraction

- In this process, extraction process directly connect to the source system and extract the source data.

(2) Offline Extraction

- The data is not extracted directly from the source system but is staged explicitly outside the original source system.
- You should consider the following structures
  → Flat Files: Generic Format
  → Dump Files: Database specific File