* Types of Digital Data =

(i) Structured Data.

(ii) Semi- Structured Data

(iii) Unstructured Data

(i) Structured Data = Structured data is a data where elements are addressable for effective analysis. It has been organized into a formatted repository that is typically a database. It concern all data which can be stored in database SQL in table with rows and columns. They are having relational key and can easily mapped into pre-designed fields.

Eg.- Relational data.

(ii) Semi- Structured Data = Semi- Structured data is information that does not reside in a relational database but that have some organizational properties that makes it easier to analyze. Semi- structured exist to save space

Eg.- XML data.

(iii) Unstructured data - Unstructured data is a data which is not ~~[illegible]~~ organized in a predefined manner or does not have a pre-defined data model, thus it is not a good fit for a mainstream relational database. So, for unstructured data, there are alternative platforms for storing and managing.

Eg.- Word, PDF, Text.

Scanned by CamScanner

Scanned by CamScanner

* Diff =

	Structured Data	Semi-Structured Data	Unstructured Data
(i)	Based on relational database table.	(i) It is based on XML / RDF	(i) It is based on character and binary data.
(ii)	Matured transaction and various concurrency techniques.	(ii) Transaction is adapted from DBMS not matured.	(ii) No transaction management and no concurrency.
(iii)	Versioning over tuples, row, tables.	(iii) Versioning over tuples or graph is possible	(iii) Versioned as whole.
(iv)	It is schema dependent and less flexible.	(iv) It is more flexible than structured data but less flexible than unstructured data.	(iv) It is very flexible and there is absence of schema
(v)	It is very difficult to scale DB schema	(v) It's scaling is simpler than structured	(v) It is very scalable.
(vi)	very robust	(vi) New technology, not very spread	(vi) NA.

* Data warehouse =

A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and ad / hoc queries and decision making. Data warehousing involves data cleaning, data integration and data consolidation.

Goals of Data warehouse =

(i) Improving integration
(ii) Speeding up response time
(iii) Faster and more flexible reporting
(iv) Recording changes to build history
(v) Increasing data quality
(vi) Unburdening operational systems.

Scanned by CamScanner
Scanned by CamScanner

(vii) Increasing recognizability
(viii) Increasing findability.

- Advantages =

(i) It allows business users to quickly access critical data from some sources all in one place
(ii) Data warehouse provides consistent information on various cross functional activities.
(iii) Data warehouse helps to integrate many sources of data to reduce stress on the production system
(iv) Data warehouse helps to reduce total turnaround time for analysis and reporting.
(v) It saves users time of retrieving data from multiple sources.

- Disadvantages =

(i) Not an ideal option for unstructured data
(ii) Data warehouse can be outdated relatively quickly
(iii) Creation and implementation of data warehouse is surely a time confusing affair.
(iv) Data warehouse is too complex for average users.
(v) Organizations need to spend lots of their resources for training and implementation purposes.

* Data Mining =

Data Mining refers to the extraction of useful information from a bulk of data or data warehouses. The result of data mining is the patterns and knowledge that we gain at the end of extraction process. Data Mining is also known as Knowledge discovery or knowledge extraction. Nowadays data mining is used in almost all the places where a large amount of data is stored & processed

Eg = Banks

Scanned by CamScanner

Scanned by CamScanner

* Advantages of Data Mining =

- Data Mining technique helps companies to get knowledge-based info.
- Data Mining helps organizations to make the profitable adjustments in operation and production
- Data Ming is a cost-effective and efficient solution compared to other statistical data applications.
- Data Mining helps with decision making process
- It can be implemented in new systems as well as existing platforms.

* Disadvantages of data mining =

- There are companies which may sell useful information of their customer to others for money.
- Many data mining analytic s/w is difficult to operate and requires advance training to work on.
- Selection of correct data mining tool is a difficult task
- Data mining techniques are not accurate

* Knowledge discovery process =

Below are the list of steps involved :-

(i) Data cleaning - In this step noise and inconsistent data is removed

(ii) Data integration - In this step multiple data sources are combined

(iii) Data selection - In this step data relevant to the task are retrieved from the database

(iv) Data Transformation - In this step data is transformed or consolidated into forms appropriate for mining

(v) Data Mining - In this step data patterns are evaluated

(vi) Pattern Evaluation = Pattern are extracted by different methods.

(vii) Knowledge Presentation - In this step knowledge is represented

Scanned by CamScanner

Scanned by CamScanner

* What type of data can be mined =

(i) Flat files
(ii) Relational databases
(iii) Data warehouse
(iv) Transactional Databases
(v) Multimedia Databases
(vi) Spatial databases
(vii) World Wide Web (WWW)
(viii) Time series Databases

* Data Mining Vs Query Tools =

- Query tools make it very easy to build queries without even having to learn a database specific query language, whereas Data Mining is a technique which deals with extracting useful and previously unknown information from raw data.
- The main difference is that in order to use query tools the users need to know exactly what they are looking for while data mining is used mostly when the user has a vague idea about what they are looking for.

* Diff =

	DATA WAREHOUSING	DATA MINING
(i)	It is a system which is designed for analytical analysis instead of transactional work	(i) Data mining is the process of analyzing data patterns.
(ii)	Data is stored periodically	(ii) Data is analyzed regularly.
(iii)	Data warehousing is solely carried out by engineers	(iii) Data mining is carried by business users with the help of engineers.
(iv)	Data warehousing is the process of pooling all relavant data together	(iv) Data mining is considered as a process of extracting data from large data sets

Scanned by CamScanner
Scanned by CamScanner

Q= How to make a data structured ?

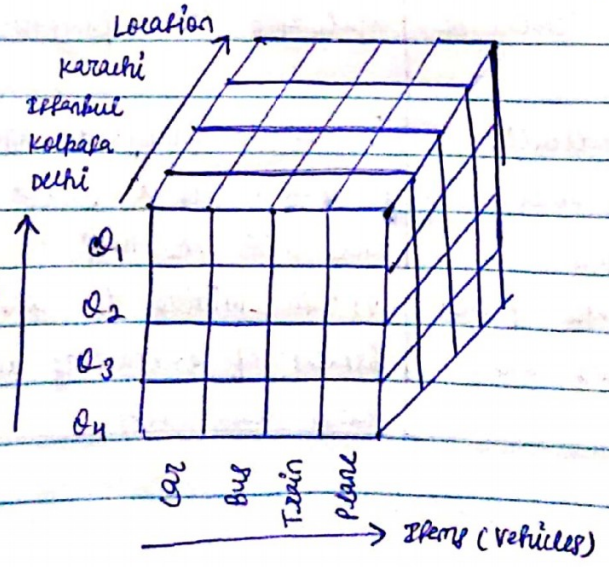
⇒ • Using metadata convert the unstructured data to structured data
• Then perform indexing
• Retrieval, updation, deletion is easy in structured data.

* Diff b/w OLAP & OLTP

	OLAP	OLTP
(i)	Online Analytical processing	(i) online Transactional Processing
(ii)	complex queries	(ii) Simple queries
(iii)	1 Lakh record size	(iii) 10 - 100 record size
(iv)	Read only	(iv) Read / Write
(v)	Long transaction	(v) Short transactions
(vi)	Tables in OLAP database are not normalized	(vi) Tables in OLTP database are normalized (3NF).
(vii)	Processing time is more in OLAP	(vii) Processing time is comparatively less in OLTP.

* OLAP Operations -

OLAP stands for online Analytical Processing. It is a software technology that allows user to analyze information from multiple db at the same time. OLAP database are divided into one or more cubes called as Hyper-cubes.



Scanned by CamScanner
Scanned by CamScanner

Operations =

There are five operations that can be performed on OLAP cube :-

(i) **Drill Down =**

In drill down operation, the less detailed data is converted into highly detailed data. It can be done by :-

- Moving down in the concept heirarchy.
- Adding a new dimension.

In the cube given in overview section, the drill down operation is performed by moving down the concept hierarchy of Time dimension (Quarter > Month)

(ii) **Roll Up =**

It is just opposite of drill down operation. It performs aggregation on the OLAP cube. It can be done by :-

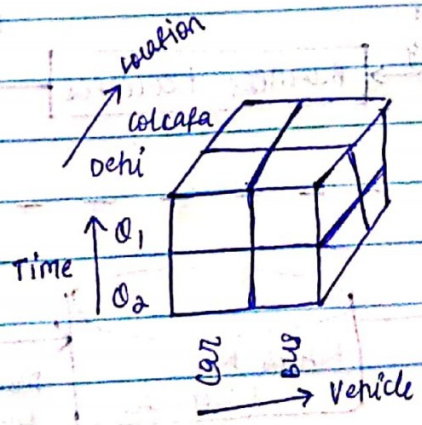
- Climbing up in the concept hierarchy.
- Reducing the dimensions.

In the cube given in the overview section, the roll up operation is performed by climbing up the concept heirarchy of Location (City > Country)

(iii) **Dice =**

It selects a sub-cube from OLAP cube by selecting two or more dimensions. In the cube given in the overview section, a sub-cube is selected by using foll dimensions

- Location = Delhi or Kolkata
- Time = Q_1 or Q_2
- Item = Car or Bus

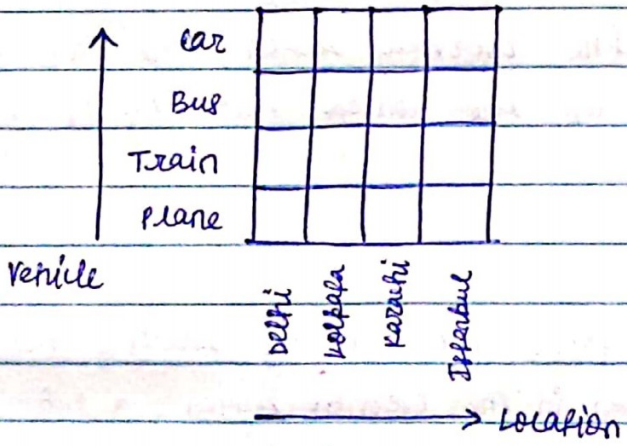


Scanned by CamScanner

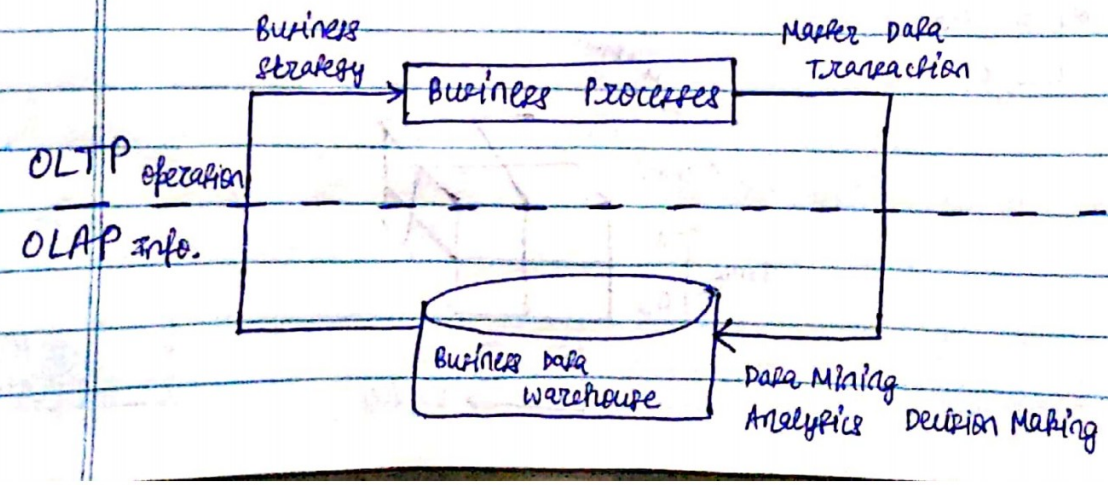
Scanned by CamScanner

(iv) **Slice** = It selects a single dimension from the OLAP cube which results in a new sub-cube creation. In the cube given in the overview section, Slice is performed on the dimension Time Q1

(v) **Pivot** = It is also known as rotation operation as it rotates the current view to get a new view of the representation in the sub cube obtained after Slice operation



* **Data Models** =

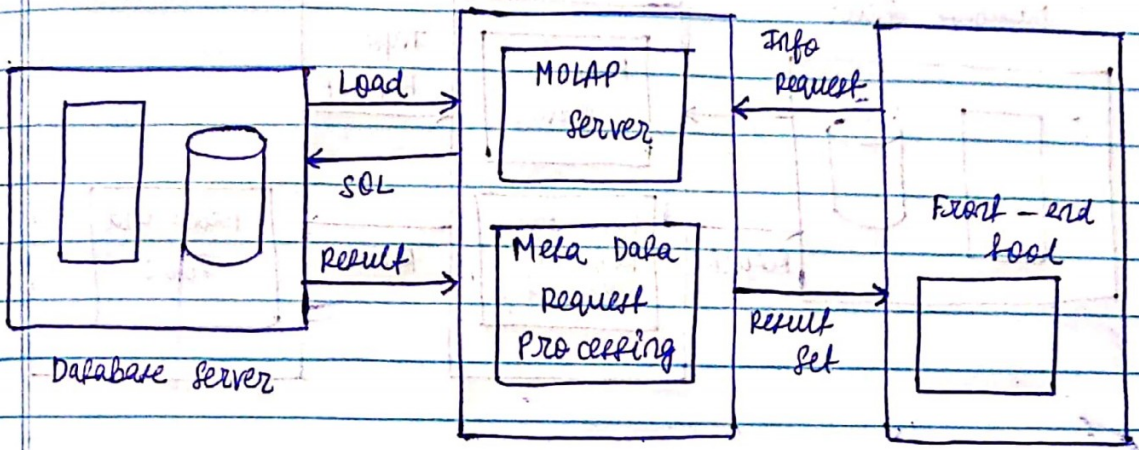


Scanned by CamScanner

Scanned by CamScanner

* MOLAP =

- Multidimension OLAP (MOLAP) is a classical OLAP that facilitates data analysis by using a multidimensional data cube.
- Data is pre-computed, pre summarized and stored in MOLAP.
- In MOLAP operations are called processing
- The storage utilization is low if the data set is sparse.
- MOLAP tools remove complexities of designing a relational database.



* Advantages =

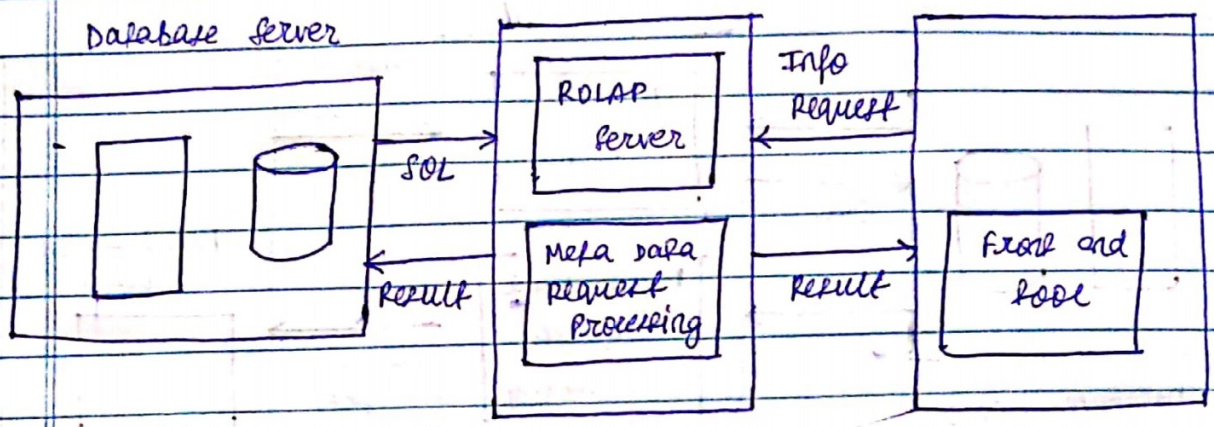
- It can store, manage and analyze considerable amount of multidimensional data
- Fast query performance
- Smaller sizes of data as compared to relational databases
- Helps user to analyze larger, less defined data.
- MOLAP cubes are built for fast data retrieval and are optimal for slicing and dicing operations

* Disadvantages -

- MOLAP is less scalable than ROLAP.
- MOLAP introduces data redundancy.
- MOLAP is not capable of containing detailed data.
- It can handle only limited amount of data, therefore it is impossible to include a large amount of data in a cube itself.

Scanned by CamScanner
Scanned by CamScanner

* **ROLAP =**
  - Relational OLAP (ROLAP) works with data that exist in a relational database.
  - It allows multidimensional analysis of data and is the fastest growing OLAP



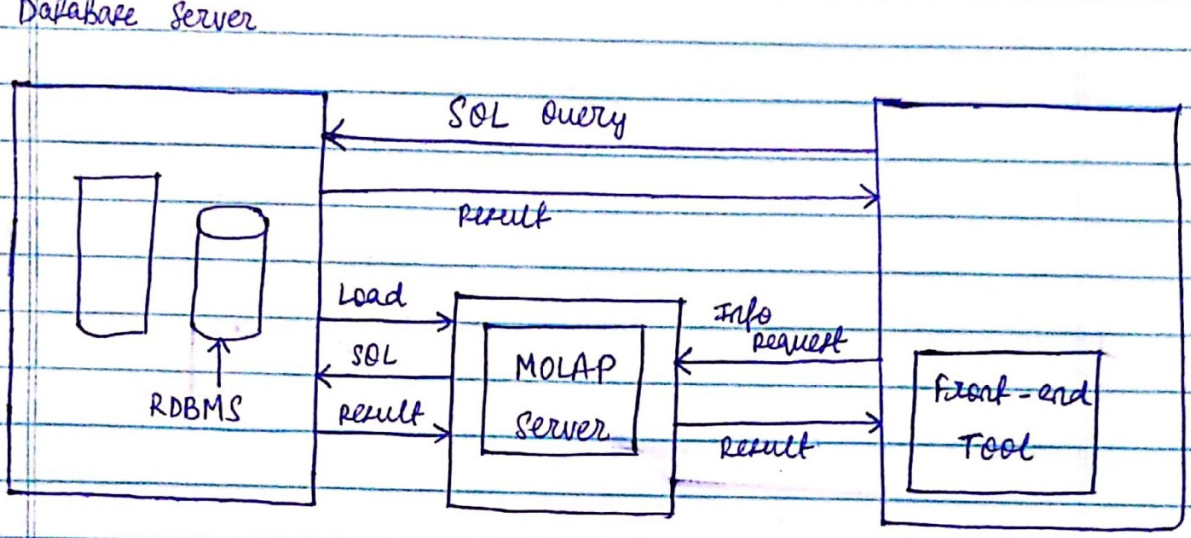
* **Advantages =**
  - ROLAP servers can be easily used with existing RDBMS.
  - Data can be stored efficiently.
  - ROLAP tools do not use pre-calculated data cubes
  - ROLAP offer scalability for manging large volume of data.

* **Disadvantages =**
  - Poor query performance
  - ROLAP needs high utilization of manpower, software and hardware
  - Some limitations of scalability.

Scanned by CamScanner
Scanned by CamScanner

* HOLAP =

- Hybrid OLAP (HOLAP) is a mixture of both ROLAP and MOLAP. It offers fast computation of MOLAP and higher scalability of ROLAP.
- HOLAP uses two databases.
- Detailed information is stored in relational database.



* Advantages =

- This kind of OLAP helps to economize the disk space.
- Hybrid OLAP uses cube technology which allows faster performance for all types of data
- ROLAP are instantly updated and HOLAP users have access to this real-time instantly updated data.

* Disadvantages =

- There are higher chances of overlapping
- HOLAP supports both MOLAP and ROLAP tools and applications thus it is very complicated.

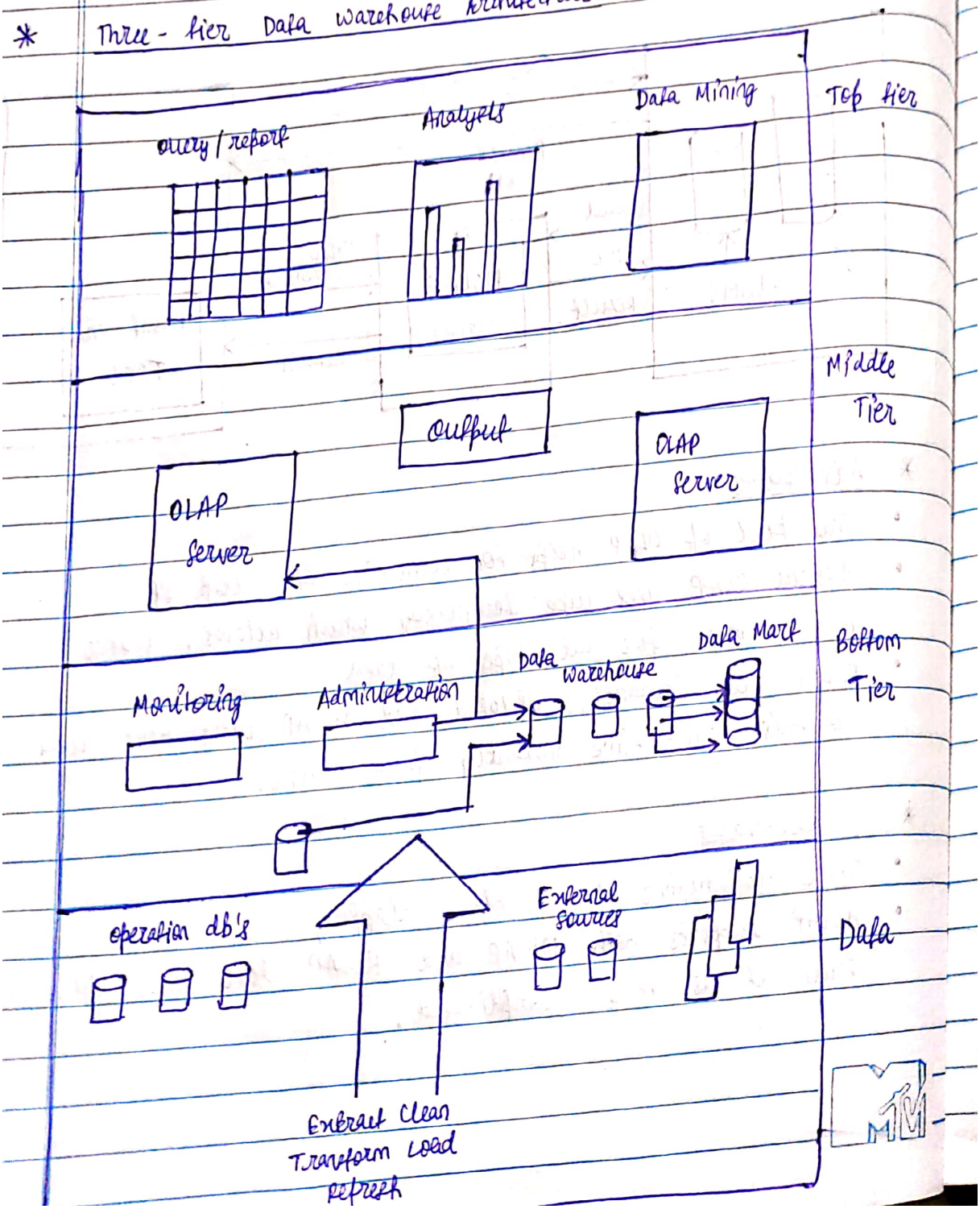
Scanned by CamScanner

Scanned by CamScanner

* Roles of OLAP tools in B.I architecture =

(i) Business focussed calculations
(ii) Business focussed Multidimensional data
(iii) Trustworthy Data and Calculation
(iv) Speed of thought Analysis
(v) Flexible, self-service reporting

* Three-tier Data warehouse Architecture =



Scanned by CamScanner

- **Bottom Tier** = The bottom tier of the architecture is the data warehouse database server. It is the relational database system. We use back-end tools and utilities to feed data into the bottom tier.

- **Middle Tier** = In Middle Tier we have OLAP server that can be implemented in the following ways :-

(i) By Relational OLAP (ROLAP)

(ii) By Multidimension OLAP (MOLAP)

- **Top Tier** = This tier is the front-end client layer. This layer holds the query tools and reporting tools, analysis tools and data mining tools.

* **Data warehouse Models** =

(i) Virtual warehouse

(ii) Data Mart

(iii) Enterprise warehouse

(i) **Virtual warehouse** =

The view over an operation data warehouse is known as virtual warehouse. It is easy to build a virtual warehouse. Building a virtual warehouse requires access capacity on operational db servers.

(ii) **Data Mart** = → Dependent data mart
→ Independent data mart
→ Hybrid data mart.

Data Mart contains a subset of organization wide data. This subset of data is valuable to specific groups of an organization.

Eg:- Marketing data marts may contain data related to items, customers and sales.

- Features =

Data marts are flexible

Scanned by CamScanner

(ii) Data Marts are small in size

(iii) Data Marts are customized by department.

(iv) The life cycle of data mart is complex

(iii) **Enterprise warehouse =**

- An enterprise warehouse collects all the information and the subjects spanning an entire organization
- It provides us enterprise - wide data integration
- This information can vary from a few gigabytes to hundreds of gigabytes or beyond.

* **Kimball vs Inmon in data warehouse =**

	Kimball	Inmon
(i)	Tactical requirements	(i) Strategic requirements
(ii)	Individual business requirements	(ii) Enterprise wide integration
(iii)	KPI, business performance measures are the structure of data	(iii) Data that meet multiple and varied information and non metric data
(iv)	Source systems are quite stable	(iv) Source systems have high rate of change
(v)	Small team	(v) Bigger team
(vi)	Low start - up cost	(vi) High start up cost.

Scanned by CamScanner