Unit III : NO SQL Databases.

3.1 Introduction.

1. History:-
- The term NOSQL was first used by Carlo Strozzi in the year 1998.
- He mentioned this name for his open source db sys. in which there was no provision of SQL query interface.
- In the early 2009, at conference held in USA, NOSQL was comes into picture and actually comes in pratice.

2. Overview.
- NOSQL is not a RDBMS.
- NOSQL is specially designed for large amount of data stored in distributed environment.
- The important feacture of NOSQL is, it is not bounded by the table schema restrictions like RDBMS.
- NOSQL generally avoids join operation.

3. Need:-
- In real time, data requirements are changed a lot. Data is easily available with facebook, Google, Twitter and others.
- The data that includes user information, social graphs, geographic location data and other user-generated content.
- To make use of such abundant resource and data, it is necessary to work with a technology which can operate such data.
- SQL databases are not ideally designed to operate such data.
- NoSQL databases specially designed for operating huge amount of data.

4. Advantages:-
1. Good resource scalability
2. Lower operational cost
3. Supports semi-structure data.
4. No static schema.
5. Supports distributed computing
6. Faster data processing.
7. No complicated relationships
8. Relatively simple data models.

5. disadvantages:-
1. Not a defined standard.
2. Limited query capabilites.

6. Companies working with NOSQL.
1. Google 2. facebook 3. LinkedIN 4. McGraw-H Educatio

3.2 Four Types of NOSQL database.

1. Key-value store databases:-
- This is very simple NOSQL db.
- It is specially designed for storing data a schema free data.
- Such data is stored in a form of data along with indexed key.

- this type is generally used when you need quick performance for basic create-Read-update-delete operations and data is not connected

Hamsterdb.
riak, redis, oracle, emcached

Example
- storing and retrieving session information for a web page.
- storing user profile and preferences
- storing shopping cart data for ecommerce

Limitations:-
- It may not work well for complex queries attempting to connect multiple relations of data.
- It data contains lot of many-to-many relationships, a key-value stor is likely to show poor performance.

Examples:-
- Cassandra, Azure Table Storage (Ars), Dynamodb.

ex of unstructure data for user records

Key 1 : ID : 123	First name : Ganesh

Key 2 :	email : abc@gmail.com	Age 37

Key 3 :	facebook ID : xxxx	Pass : xxx	Name : Max

2. Column store database:-
- Instead of storing data in relational tuples (table rows), it is stored in cells grouped in columns.
- It offers very high performance and a highly scalable architecture.

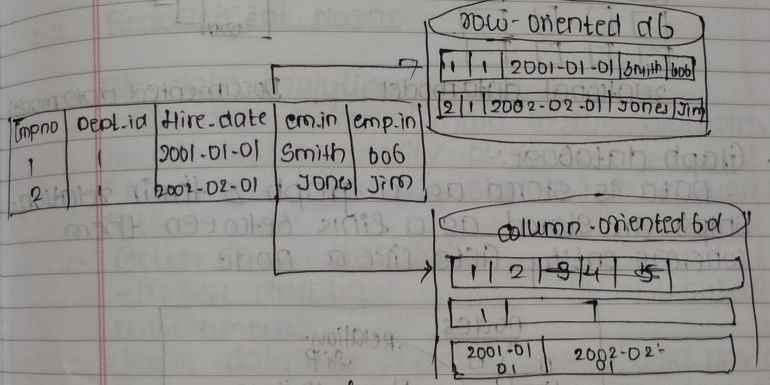


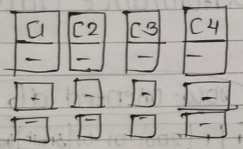
fig:-

Example.
1. HBase 2. Big table 3. Hyper table.

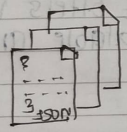
3. Document database:-
- Document database works on concept of Key value stores where "documents" contains all of complex data.
- Every documents contains a unique key, which to retrieve the documents.
- key is used for storing, retrieving and managing document-oriented information a

known as semi-structured data.

Examples:-
- Mongodb
- CouchDB

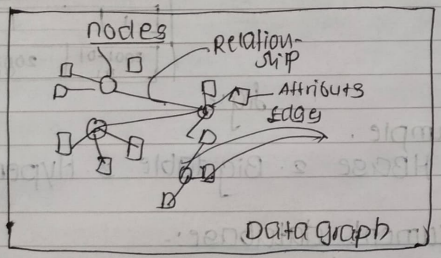


relational datamodel



Documented data model

4. Graph database:-
- Data is stored as a graph & their relationships are stored as a link between them whereas entity acts like a node



Example:-
1) Neo4j , 2) Polyglot

5. Comparison.

	Datamodel	Performance	Scalibility	Flexibility
1.	Key-value store	High	High	High
2.	column store	High	High	moderate
3.	document store	High	variable (High)	high
4.	graph database	valuable	variable	high

1.2 Benefits of NoSQL.

1. Big data Analytics:-
- Big data is one of main feature promotes growth and popularity of NOSQL
- NoSQL has good provision to handle such big data.

2. Better data Availability:-
- NoSQL database works with distributed environments.
- NoSQL database environments should provide good availability across multiple data servers
- NoSQL database supply high performance.

3. Location Independence.
- No SQL database can read and write database regardless of location of database operation.

1.3 CAP Theorem (Brewer's theorem)

CPA theorem states three basic requirements of NOSQL database to design a distributed architecture.

a) Consistency:- database remain consistent state like before, even after execution of

b) Availability:- It indicates that NOSQL system is always available without any downtime

c) partition tolerance:- this means that the system continues to function even the communication failure happens between servers ie if one serve fails, other server will take over.

there are many combinations of NOSQL rules.

1. CA
- It is a single site cluster.
- All nodes are always in contact.
- Partitioning system can block the system

2. CP:-
Some data may not be accessible always still it may be consistent or accurate.

3. AP:-
- System is available under partitioning
- some part of the data may be ~~in~~ inconsistent.

1.4 BASE Properties for NOSQL.

1.5 NOSQL Key value database: MongoDB.

1. NOSQL:-
- NOSQL means "Not only SQL", NOSQL database can use SQL-like query concept.
- It includes all databases that is not a traditional relational RDBMS.
- NOSQL databases are more specialized for various types of data types.
- It is more efficient and better performing than RDBMS servers is most instances.

2. MongoDB:-
- open source document dbms.
- one of the popular NOSQL db, written in C++.
- High performance, high availability and automatic scaling are the important features.
- It has its own ad-hoc query language with rich features set.
- large number of use cases can be easily handl
- Mobile application, Content Management Sys E-commerce, Gaming application, Analytics, Ach ing and logging are the application areas of MongoDB.
- Multi-document transactions are not possib in MongoDB.
- It provides atomic operations on a single document.
- the biggest advantage of MongoDB is that a cache memory it automatically uses a free memory available on the machine.

3. Mongo features:-
- High performance as compared to traditional SQL.
- Good support for embedded data models.
- faster query processing using index support from embedded documents and arrays.
- Higher availability.
- Provides horizonal scalability.
- provides automatic sharding techniques.

4. Collect & document:-
- Collection is equivalent to a table in RDBMS.
A collection is a group of documents which exists within a single database.
collection is schemless.
- Different fields can be created with different documents in a collection. But typically, all the documents in a single collection are of similar or related purpose.
- A document is set of key-value pairs. Documents have dynamic schema.

RDBMS terminology with mongoDB terminology.

RDBMS	MongoDB.
Database	Database
Table	Collection
Tuple/row	document
Column	field
Table join	Embedded documents
primary key	primary key (default key. id provided by mongodb itself.

5. Document database:-
In Mongodb document is used to store data. Each document consists of field and values.
- Mongodb documents and JSON objects are similar to each other.
- other documents arrays etc can be included in the field values.

Documents have following advantages:-
- Due to embedded documents and arrays, joins are avoided. Hence less expensive.
- Document uses dynamic schema.
- In popular programming language documents are native data types.

- A mongodb documents is a follows:-
```
{ movie: "Hanuman"
  Ticket: 100
  Screen: 02
}
```

6. Mongodb schema:-
- It use dynamic schemas. Documents in a collection may have different set of fields. Due to this polymorphism can be easily used. (with defining the structure.
- Adding new field and deleting existing field can be easily done.

7. Comparison MongoDB and RDBMS

MongoDB	RDBMS
1. Schema less	1. requires schema
2. Complex joins are avoided	2. complex joins are used
3. documed based db	3. table based db.
4. It has its own ad-hoc query language with rich features set	4. RDBMS has SQL
5 Easy to scale	5 Not easy to scale.

1.6. Comparative study of RDBMS and NOSQL (SQL vs NOSQL)

SQL dbses are RDBMS; whereas NOSQL database are non-relational database.

1. Data storge:-

- SQL databases stores data in a table whereas NOSQL databases are ~~non-relational database~~. stores data as document based, key-values pairs graph databases or wide-column stores.
- SQL data is stored in form of tables with some rows.
- NOSQL data is stored as collection of key-value pair or documents or graph based data with no standard schema definations.

2. Database Schema:-

- SQL databases have predefined schema which cannot be change very frequently whereas NOSQL databases have dynamic schema which can be change any time for unstructured data.

3. Complex Queries:-

- SQL databases provides standard platform for running complex query.
- NOSQL does not provide any standard environment for running complex query.
- NOSQL are not as powerful as SQL query language.

SQL	NOSQL
1. Structured Query Language	1. Not Only SQL or Non-relational database.
2. SQL is a declarative query language	2. This is Not a declarati query language.
3. SQL db works on ACID properties • Atomicity • Consistency • Isolation • Durability	3. NOSQL database follows the Brewers CAP theorem, • Consistency • Availability • Partition Tolerance
4. Structured & organized data	4. Unstructured and unre cable data.
5. Relational db is table based	5. Key-value pair storage column store, document st graph db.
6. Data and its relationship are stored in seperate tables.	6. NO-pre-defined schem
7. Tight consistency	7. Eventual consistency
8. ex. MySQL, Oracle, MS SQL, SQlite, DB2.	8. ex, Mongodb, big tabl Neo4j, couchdb, cassandra HBase.

17 Advantages of NoSQL.

1. The growth of big data.
2. Continuous Availability of data.
3. location independence
4. Modern transactional capabilities.
5. Flexible data models.
6. Better Architecture.
7. Analytics and Business Intelligence.

18 MapReduce.

- Map reduce is a data processing paradigm for condensing large volume of data into useful aggregated results.
- Mongodb uses mapReduce command for map-reduce operations.
- generally used for processing large data sets.

mapReduce command. (syntax).

```
>db.collection.mapReduce(
function() {emit(key,value);}, //map function
function(key,values) {return reduceFunction},{
    out: collection,
    query: document,
    sort: document,
    limit: number
})
```

The map-reduce function first queries the collection, then maps the result documents to emit key-value pairs, which is then reduced based on the keys that have multiple values.

In the above syntax

- map is a javascript function that reduces or groups all the documents having the same key.
- out specifies the location of the map reduce query result.
- reduce is a javascript function that reduces or groups all the documents having the same key.
- query specifies the optional selection criteria for selecting documents.
- sort specifies the optional sort criteria.
- limit specifies the optional maximum number of documents to be returned

example of mapReduce:-

```
>db.posts.mapReduce(
  function() { emit(this.user_id,1); },
  function(key,values) {return Array.sum(values)},
  {
    query:{status:"active"},
    out:"post_total"
  }
);
```

output:-

```
{
    "result" : "post_total",
    "timeMillis" : 9,
    "counts" : {
        "input" : 4,
        "emit" : 4,
        "reduce" : 2,
        "output" : 2
    },
```

1.9 HIVE

- HIVE is a datawarehouse infrastructure tool.
- It processes structured data in HDFS. Hive structures data into tables, rows, columns and partitions.
- It resides on top of Hadoop.
- It is used to summarize big data, analysis of big data.
- It is suitable for online Analytical application processing
- It supports ad hoc queries. It has its own SQL language called HiveQL or HQL.
- Primitive datatypes like integers, floats, doubles and strings are supported by HIVE.
- Associative arrays, lists, structs etc can be used
- Serialize API and Deserialized API are used to store and retrieve data
- HIVE is easy to scale and has faster processing

2.0 Architecture of HIVE.

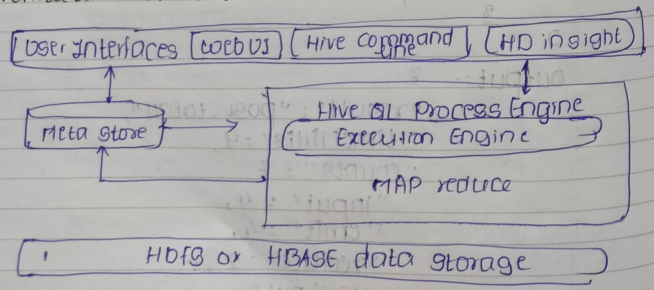


fig : Hive Architecture.

this information cannot be changed further as these are the standard components.

1. User Interface :- Hive supports Hive web UI, Hive command line, and Hive HD through which user can easily process queries.
2. Meta Store : Hive stores Meta data, schema etc, in respective database servers known as metastores.
3. HiveQL Process Engine :- HiveQL is used as querying language to get information from metastore. It is an alternative to MapReduce Java program. HiveQL query can be written for MapReduce job.
4. Execution Engine :- Query processing and result generation is the job of Execution engine. It is same as that of MapReduce results.
5. HDFS or HBASE : Hadoop distributed file system or HBASE are the data storage techiniques to store data into file system.

2.1 Working of Hive

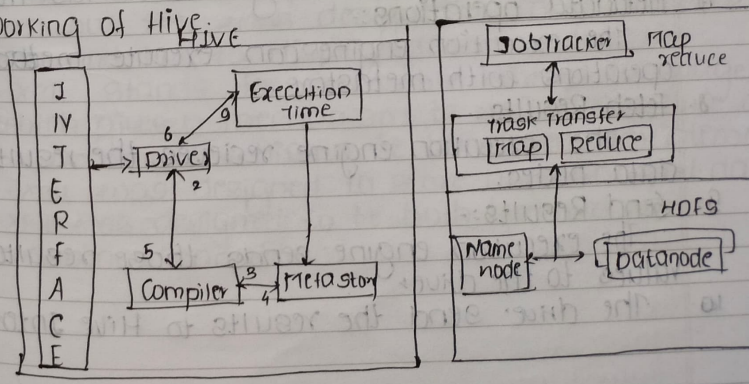


fig Hive & Hadoop communication

Command line or web UI sends query to JDBC or ODBC driver to execute.

1. Get plan:- With the help of query compiler driver checks the syntax and requirement of query.
2. Get Metadata:-
The compiler sends metadata request to Metastore for getting data.
3. Send metadata:-
Metastore sends the required metadata as a response to the compiler.
4. Send plan:-
The compiler checks the requirements and resends the plan to the driver. Thus the parsing and compiling of a query is complete.
5. Execute plan:- The driver sends the execute plan to the execution engine.
6. Execute Job:- The execution engine sends the job to JobTracker. JobTracker assigns it to TaskTracker.
7. Metadata operations:-
The execution engine can execute metadata operations with metastore.
8. Fetch Results:-
The execution engine recieves the results from data nodes.
9. Send Results:-
The execution engine sends those resultant values to the driver.
10. The driver send the results to Hive Interface.

2.2 Hive Data models.

The Hive data models contain the following components:-

- Databases
- Tables
- Partitions
- Buckets or clusters

Partitions:- The table is divided into a smaller parts based on the value of a partition column. Then on these slices of dataqueries can be made for faster processing.

Buckets:- Buckets give extra structure to the data that may be eised for efficient queries. Different data required for queries joined together. Thus queri can be evaluated quickly

2.3 Programming language - XML.

- Introduction:-

- XML is a markup language is very much li HTML but XML was designed to carry (transfer data and not to display data.
- XML stands for Extensible Markup languag which gives a mechanism to define structure a document which to be transfered over inter
- XML was designed to store and transport d
- XML was designed to be both humand and machine-readable.

XML plays an important role in many different IT systems.
- XML is often used for distributing data over the internet.
- It is important (for all types of software developers) to have a good understanding of XML

2.5 XML does not do anything.

```
<note>
<to> Tove </to>
<from> Jani </from>
<heading> Reminder </heading>
<body> don't forget me this weekend! </body>
</note>.
```

The xml above is quite self-descriptive
- It has sender and receiver information
- It has a heading
- It has a message body.

But still, the xml above does not DO anything. XML is just information wrapped in tags.

Someone must write a piece of software to send, receive, store, or display it:

Note
To: Tove
From: Jani
Reminder
Don't forget me this weekend!

2.6 • difference between XML and HTML
- XML - was designed to carry data - with focus on what data is
- HTML - was designed to display data - with focus on how data looks.
- XML tags are not predefined like HTML tags are

2.7 Valid XML documents.
XML documents which satisfy given below conditions are called as XML documents.
a) conditions:-
- ~~The doc~~ must be well formed
- element used in the start & end tag pairs must follows the specified structure
- This structure is specified in a seperate XML DTD (Document Type defination) file or XML schema file

b) DTD Syntax:-
- start with a name is given to the root tag of the document.

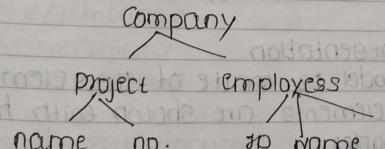


fig An XML DTD file called company.

XML document the basic object is XML and it can be represented as hierarchical data model or tree data model.

2.9 :- Structuring elem concepts used to construct an XML document

a) Element :-

- An element is a group of tags data values that can contain character data, child elements or a mixture of both
- Elements can be two type simple and complex
- The elements are constructed from other elements by nesting them is called complex element.
- The elements contain data values are called as simple elements.

b) Attributes :-

- Additional information that describes elements

c) " there are some additional concepts used in XML, such as entities, identifiers and references

- Tree representation

XML model is made of two elements

i) Complex elements are shown with help of internal nodes.

ii) Simple elements are shown by leaf nodes.

Hence, XML model is called as a tree model or hierarchical model.

example:

as tree representation :-

- Simple elements : <price>, <amount>
- Complex elements : <drink>, <snacks> etc

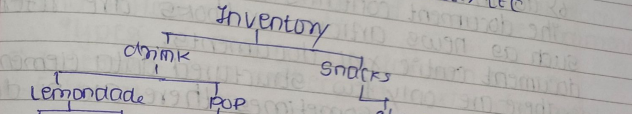


fig : tree model for above code.

b) Textual representation :-

- whenever value of the STANDALONE attribute in an XML document is set to "YES" such XML docum is known as schemaless XML documents

ex
```
<inventory>
  <drink>
    <lemonade>
      <price> $ 2.50 </price>
    </lemonade>
  </drink>
</inventory>
```

3.1. Types of XML documents :-

a) Data-centric XML documents

- The document contains many small data item that follow a specific structure and hence it may be extracted from a structured database are called as data-centric XML documents

- In order to exchange and display data over internet the document is formatted as XML documents.

b) document-centric XML documents.
- The document contains large amounts of texts, such as news articles or books are referred as document centric XML documents.
- There are only few structured data elements in these documents. Sometimes there are no data elements in such documents.

c) Hybrid XML documents:-
- If both types of data are used simultaneously in document then it is refered as hybrid XML document.
- In such documents some part that contain structured data and other parts are predominantly unstructured.

3.2 XML- structured data or semi structured data.
- Data centric XML documents sometimes considered as semi-structured data or some time considered as structured data.
- Structured data if an XML document is written as per predefined XML schema or DTD.
- semi structured if an XML allows documents that do not ~~contain~~ conform to any particular schema.

3.3 XML Document Type Definition (DTD).
- DTD describes the rules for analysing an XML documents.
- A DTD ensures the contents of an XML documents conform to expected rules the document should follow.

Types of DTD as per location of DTD.
a) Internal: DTD embedded in XML document.
b) External: DTD as a separate external file.

Types of DTD declarations:-
a) element type b) Attribute list c) Entity d) Notation

a) element type declarations:- defines one of the kin of elements you can use that is, one of the tag types.
b) Attributes list declarations. identifies which elements have attributes they may have attrib values and optional attributes
- parts, = i) name ii) type iii) An optional defau

3.4 Advantage of DTD:-
- allow an XML parser to validate an XML document against its definition. This allows th online validation of an XML document before it is processed.
- It can extend capabilities of an XML documen by allowing further processing ~~ea~~ possibilities.
- DTD can serves as an automotive documen

disadvantage of DTD.
a> It is written in a different (non xml) syntax
b> It has no supports for namespaces
c> It only offers extremly limited data typing

Programming language - JSON.
- JSON stands for JavaScript object Notation
- basically used for data transmission in the web applications.
- is the extension of JS language
- can be used for storing and exchanging data
- can be used as an alternative to XML.
- is lightweigh, text based data-interchange format. Text can be read and use as a data format by any programming language.
- is self-describing and easy to understand
- has file name extension .json
- supports name-pair values, arrays, lists, vector, sequence etc.

JSON code

```
{"movies": [
{"name": "Tom and Jerry", "year": "2000"},
]}.
```

JSON data types
i> Number:- defined as a double precision floating-point format.
- Octal formats, hexadecimal formats, NaN, Infinity are not used.

ex: var jnum = (marks 69)
- Here jnum is the json object name given by use

ii> Boolean - used to indicate either true or false value

ex: var jnum = { name: 'sachin', marks: 69, firstclass: true }

iii> null: null means empty, ex: var c = null;
iv> json value: may include string, boolean value integer, float null etc

```
ex = var x = 2;
     var c = "sachin";
     var d = null;
```

v> string: sequence of characters {

ex: var jnum = { name: 'sachin' }

vi> Array: collection of similar data type.
vii> object: uses name/value pairs.
All the keys should be different from each ot

```
ex: {
    "screen": 05"
    "movie": John carter",
    "ticket": 200
}
```

viii> whitespace: used between token pairs to ma code more readable

ex: var c = "sachin"

for collecting & storing info. of a particular organization.
- goal of using a dw is to have an efficient way of managing information and analyzing data.

Data warehouse components:-
1. Source Data components
- In DW source data is coming from various heterogeneous data sources grouped into four broad categories.
a) Production data.
- this data is coming from the various heterogeneous (different) operational system of organization
- Based on data requirements in the data warehouse,
we may choose segments of data from the different operational systems.
ex:- In case of ABC shoppers stop all sales data can be considered as production data

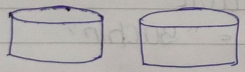


fig: representation of production data

b) Internal data:-
- In any organization will have various documents like customer, profiles, registers, employee data and sometimes even getional databases.
- data is generated by internal organizational operations.
ex - In big organizations employee's personal data may be treated as internal data, or in sales database customer profile is an internal data

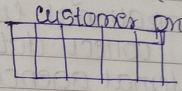


fig: representation of internal data

c) Archived data.
- the main task of operational systems is to run the current bussiness operations.
- In every operational system, we will periodically fetch the old data and store it as archived files.
- As per corporate requirements. we will decide how often and which portions of the operational databases are archived for storage.
ex: Daily transaction will be archived during night time in case of daily sales database

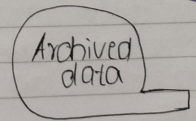


fig: Representation of archieve da

Datawarehouse Architecture.

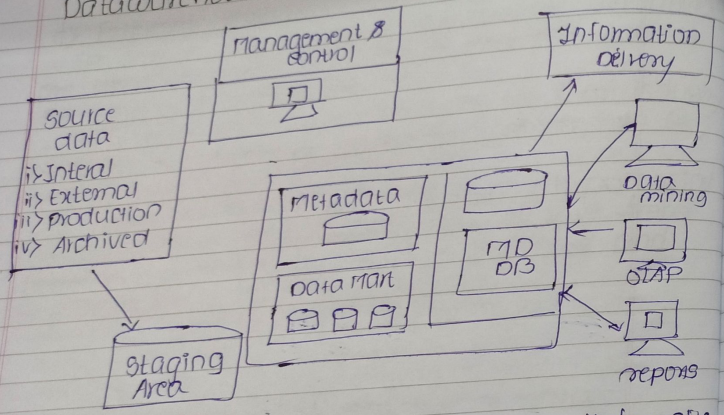


fig: Architectural components in the 3 major areas

1. Intro.
- It is primarily based on the business processes of a business enterprise taking into consideration the data consolidation across the business enterprise with adequate security, data modelling and organization, extent of query requirements.
- Meta data management and application, warehouse staging area planning for optimum bandwidth utilization and full technology implementation.

2. main areas of data warehouse:-
a) Data acquistion
b) Data storage
c) Information delivery.

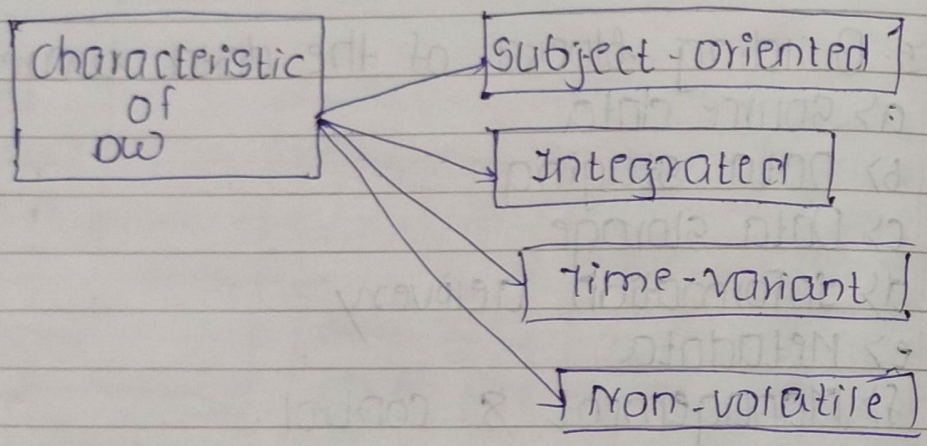
3. Building blocks of the data warehouse.
a) source data
b) Data staging
c) Data storage
d) Information delivery
e) Metadata
f) Management & control.

4. DW Architecture:-
- In order to set up this information delivery system, we need different building blocks.
- These building blocks are arranged together in the most optimal way to serve intended target
- Architecture, in the context of an organization's dw efforts, is a conceptualization of how the dw is built.
- DW relates all components (which has definit functions and provides specific services together) so as to make fully functional data warehous
- Architecture is the proper arrangment of th components.
- we can build a data warehouse with software and hardware components.
- To suit the organizational requirements, w need to arrange these building blocks in a maximum benefit.

Datawarehouse - characteristic, Advantages & Disadvantages

1. characteristic

a. subject - oriented data:-



a. Subject - Oriented

- DW is organized around subject such as sales, product, customer etc.
- It focuses on modeling & analysis of data for decision makers like managers, CEO higher hierarchy in organizations

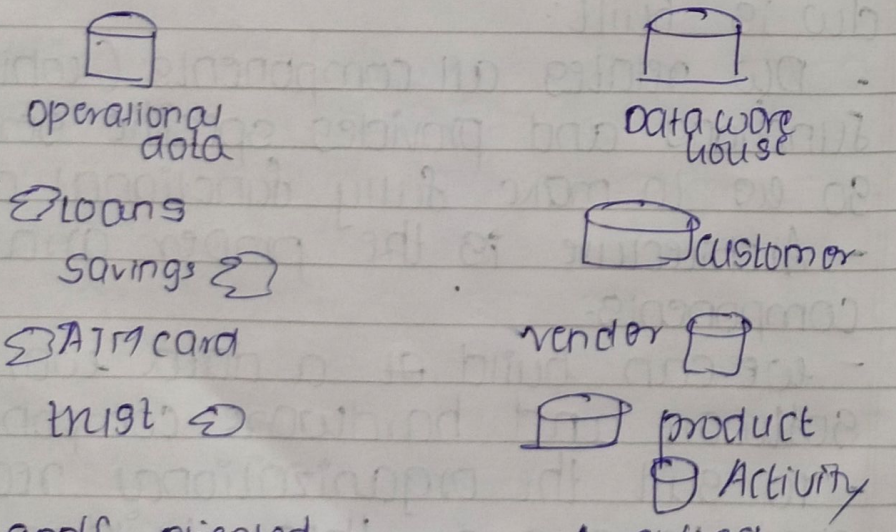


fig : Subject - Oriented data

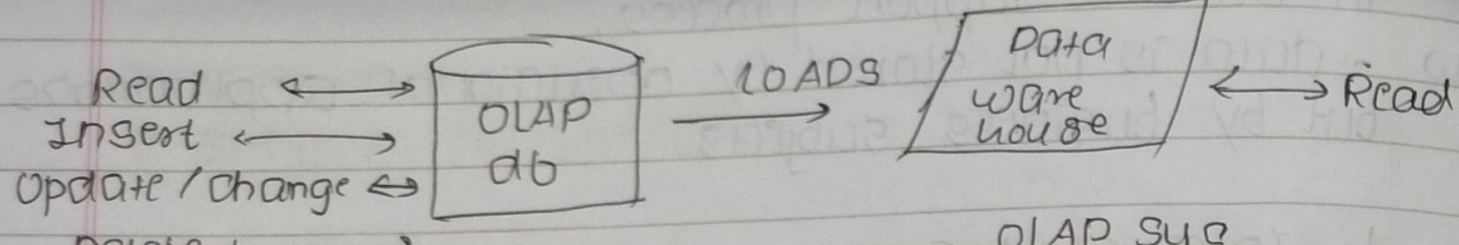


fig: Non-volatile data.

The dw pt is a physically seperate data storage, which is transformed from the source operational RDBMS.

The operational updates of data do not occur in dw. ie update, insert & delete operations are not performed.

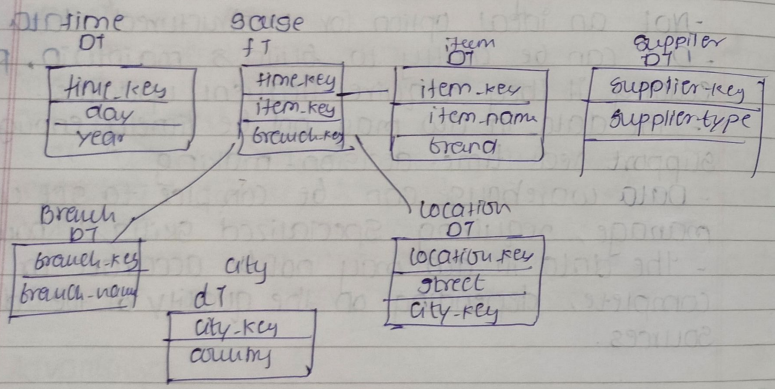
- It usually requires two procedures in data accessing: initial loading of data and access to data.
- Non volatile defines that once entered into the warehouse, and data should not change.

Advantages of DW.

- DW allows business user to quickly access critical data from some sources all in one place.
- DW provides consistent information on various cross functional activities.
- It is also supporting ad-hoc reporting & query.
- DW helps to integrate many sources of data to reduce stress on the production system.
- DW helps to integrate ~~many source of~~ help to

- dimension table contains the set of attributes
- There is a fact table in center (Big).
- It contains the key to each of four dimensions
- The fact table also contains the attributes, namely dollar sold & units sold

② Snowflake schema



- Some dimension tables in the snowflake schema are normalized
- normalization splits up the data into additional tables.
- Unlike Star schema, the dimensions tables in a snowflake schema are normalized.
- For ex, the item dimension table in star schema is normalized & split into two dimension tables, namely item & supplier table

supplier key is linked to the supplier DT. The supplier DT contains the attributes supplier key and supplier-type.

4 OLAP and types of OLAP Architecture
a) ROLAP b) MOLAP c) HOLAP.

online Analytical processing (OLAP) Server is based on the multidimensional data model.
- It allows manager, and analysts to get an insight of the information through fast, consistent, and interactive access to information.

Types of OLAP Servers.
1. Relational OLAP (ROLAP)
2. Multidimensional OLAP (MOLAP)
3. Hybrid OLAP (HOLAP)

1. ROLAP.
- ROLAP servers are placed between relational back-end server and client front-end tools.
- To store and manage warehouse data, ROLAP uses relational or extended-relational DBMS.

ROLAP includes:-
- Implementation of aggregation navigation logic
- Optimization for each DBMS back end.
- Additional tools and services.

2. MOLAP.

MOLAP uses array-based multidimensional storage engines for multidimensional view of data.

- With multidemensional data stores, the storag utilization may be low if the data set is sparse - ∴, many MOLAP server use two levels of da storage representation to handle dense and spa data sets.

3. HOLAP:-

Hybrid MOLAP is a combination of both ROLAP and MOLAP.

It offers higher scalability of ROLAP and faster communication of MOLAP.

- HOLAP servers allows to store the large data volumes of detailed information.

- The aggregations are stored seperately in MOLA store.

⑤ OLAP operations:-

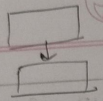
a> roll up b> Drill down c> slice d> dice.
e> pivot (rotate).

roll up

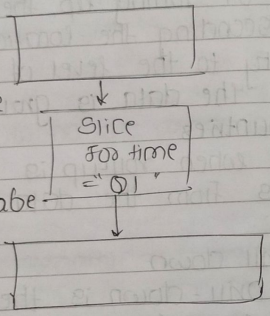
roll-up performs aggregation on a data cube in any of the following ways:-

- By climbing up a concept hierarchy for a dimension.
- By dimension reduction.

3. Slice:-
The slice operation select one particular dimension from a given cube and provides a new sub-cube.



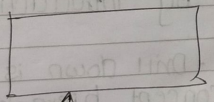
Here slice is performed for "dimension "time" using the criterion time = "Q1"

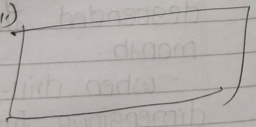


- It will from a new sub-cube by selecting one or more dimensions

4. Dice:-
Dice selects two or more dimensions from a give cube and provides a new sub-cube

We dice operation on the cube based following selection criteria involves 3 dimensions

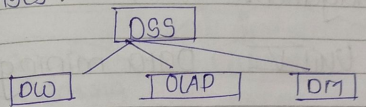




- (location = "Toronto" or "Vancouver")
- (time = "Q1" or "Q2")
- (item = "Mobile" or "Modem")

4. Pivot:- it is known as rotation.
It rotates the data axes in view in order to provide an alternative presentation of data.

6. Decision support system:-
DSS is a kind of computerized information systems that support decision making activities.
- DSS provides each kind of decision information as well as solution of commercial question for the enterprises.



- There are two-kinds of DSS system
1. thinks so long as the system has certain supports to the decisions is DSS.
2. thinks DSS is interactive computer-based system which can help the decision-maker using data and model to solve non-structure questions.

- The majority is not DSS. bcoz most of them do not help to solve non-structure problems
- Some are not interactive:- some databases become the model stores house and they are not entire

DW:- DW does not like traditional database which faces the service level. It mainly faces the high level application and carries on the decision support. Thus, in fact, it is expansion of DSS database.

interactive way and can promptly purpose information from mutative data and not too complete data the movement which reveates with enterprise management.

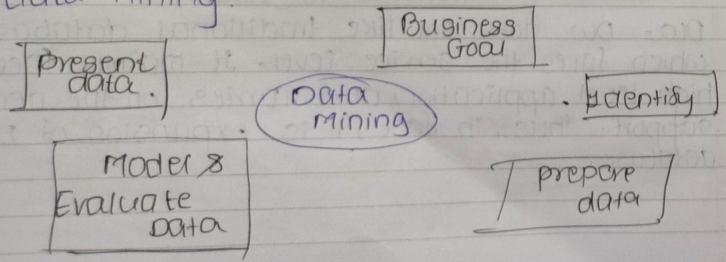
Unit V Data mining

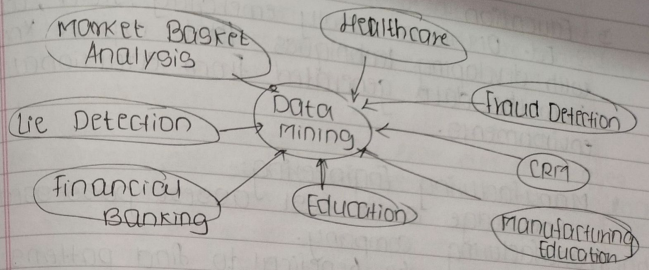
1. Data mining & it's application.

- DM is widely used in diverse areas.
- There are a number of commerical data mining system available today and yet there are many challenges in the field.

DM helps banks work with credit rating and anti-fraud systems, analyzing customer financial data, purchasing transactions, and card transactions.

[The process of extracting information to identify patterns, trends, and useful data that would allow the business to take the data driven decision from huge sets of data is called data mining.





1. Healthcare :-

DM in healthcare has excellent potential to improve the health system.

- It uses data and analytics for better insights and to identify best practices that will enhance health care services and reduce costs.
- Analysis use DM approches such as ML, multi-dimensional db, Data visualization, soft computing & statistics.
- DM also enables healthcare insures to recogn Fraud & abuse.

2. Market Basket Analysis.

- MBA is a modeling method based on hypothe
- If u buy a specific group of products, then
are more likely to buy another group produces
- Using a different analytical comparison of results between various stores, between custom in different demographic groups can be done.

3. Education:-
- Ed. DM is a newly emerging field, concerned with developing techniques that explore knowledge from the data generated from eductional environments.

4. Manufacturing Engineering:-
Knowledge is the best asset possessed by a manufacturing company.
- DM tools can be benefical to find patterns in a complex manufacturing process.
- It can also be used to ~~fore~~ forecast the product development period, cost, and expectations among the other tasks.

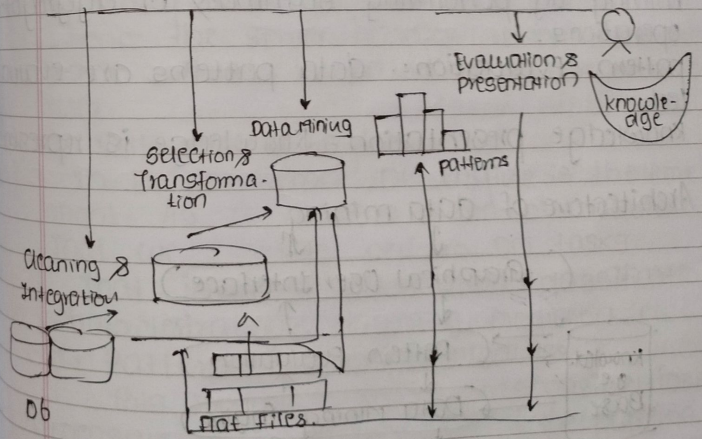
5 Customer Relationship management:-
CRM is all about obtaining & holding customers also enhancing customer loyalty & implementing customer-oriented strategies.
To get a decant relationship with the customer, a business organization needs to collect data and analyze the data.
With DM technologies, the collected data can be used for analytics.

6. Fraud detection:-
- Billions of dollars are lost to the action of frauds.
- An ideal fraud detection system should protect the data of all the users.

- Supervised methods consist of a collection of sample records, and these records are classified as fradulent or non-fradulent.

7. Lie Detection:-
Law enforcement may use dm techniques to investigate offences, monitor suspected terorist communication etc,
this technique include text mining also, and it seeks meaningful patterns in data, which is usually meaningful patter in data, which is usually unstructured text.

8. KDD process with diagram.



06

DM Architecture system contains too many components. That is a data source, data warehouse server, data mining engine, and knowledge base.

a. Data source.

There are so many documents present. That is db, data warehouse, WWW,
- That are the actual source of data.
- Sometimes, data may reside even in plain text files or spreadsheet.
- WWW or the Internet is another big source of data.

b. DB or DW (warehouse) Server.

- The db server contains the actual data that is ready to be processed.
- Hence the server handles retrieving the relevant data. That is based on DM request of the user.

c. Data Mining Engine.

In DM system DM engine is the core component. As It consists a no. of modules. That we used to perform DM tasks.
- That includes association, classification, characterization, clustering, prediction, etc.

d. Pattern Evaluation modules.

This module is mainly responsible for the measure of interestingness of the patterns. For this, we use a threshold value. Also, it interacts with the DM engine.

Their's main focus is to search towards intersting patterns.

e. Graphical user interface.
- we use this interface to communicate between the user & the DM system.
Also, this module helps the user use the system easily & efficiently.
- they don't know the real complexity of the process.
When the user specifies a query, this module interacts with the DM system. Thus, displays the result in an easily understandable manner

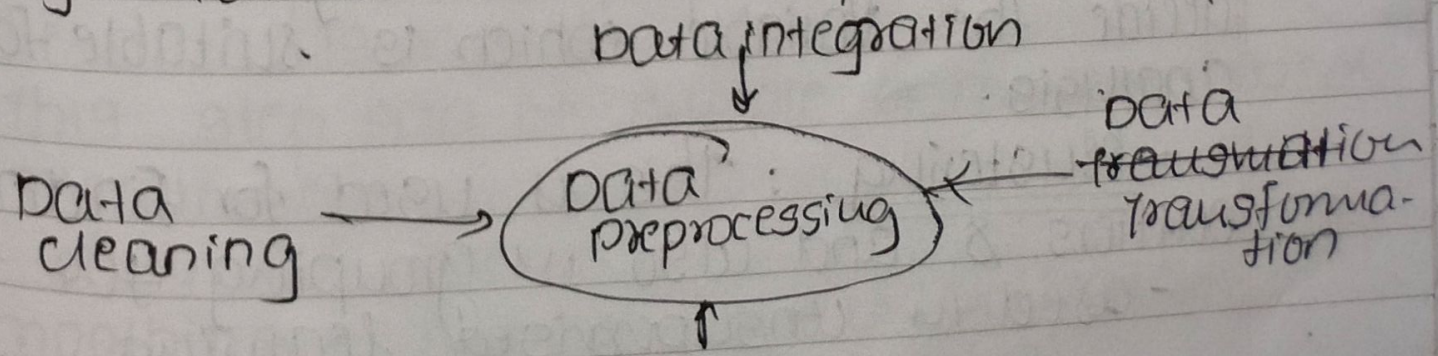
f. Knowledge Base:-
- In whole DM process, the knowledge base is beneficial.
- we use it to guiding the search for the result patterns.
- The KB might even contains user beliefs and data from user experiences.
- that can be useful in the process of DM.
- the pattern evaluation module interacts with the KB.

DM in business intelligence
Business intelligence & DM differ in a few core ways. Namely, in purpose, volume & results.
The purpose of BI is to convert data into useful information for executive.
BI tracks key performance indicators & presents data in a way that encourages

- It is also an important step in DM as we cannot work with raw data.
- The quality of the data should be checked before applying ML or DM algor.

Preprocessing of data is mainly to check the data quality.
- The quality can be check by following
- Accuracy:- To check whether the data entered is correct or not.
- Completeness:- To check whether the data is available or not recorded.
- Consistency:- To check whether the same data is kept in all the places that do or do not match.
- Timeliness:- The data should be updated correctly.
- Believability:- The data should be trustable.
- Interpretability:- The understandability of the data.

Major tasks in Data Preprocessing:



Data cleaning:- is the process to remove incorrect data, incomplete data & inaccurate data from the datasets, and it also replaces the missing values.

* noisy:- generally means random error or containing unnecessary data pts.

some methods to handle noisy data.

- Binning- this method is to smooth or handle noisy data.

- there are 3 methods for smoothing by bin mean method-

- smoothing by ~~mee~~ bin mean method.
- the values in the bin are replaced by the mean value of the bin.
- smoothing by bin boundary:- the using minimum and maximum values of the bin values are taken and the values are replaced by the closest boundary value
- ~~something~~ smoothing by the bin median- the values in the bin are replaced by the median.

regression:-

this is used to smooth the data & will help to handle data when unnecessary data is present.

- for analysis, purpose regression helps to decide the variable which is suitable for our analysis.

- clustering : this is used for finding the outliers & and also in grouping the data
- used in unsupervised learning.