4) What are Data Warehousing (DW) and Business Intelligence? Describe its components and its uses. How does DW the same or different with respect to Business Intelligence? Define and explain ETL operations.

Data Warehousing focuses on using the accumulated data from databases to provide information for effective decision making. It involves in depth analysis of the available historical data. Relational databases are a key technology underlying data warehousing applications. Typically, data warehouses have millions of rows.

Business Intelligence focuses on supporting decision making through data analysis and reporting.

Data warehouses have the following components:

Warehouse Loading Tools: They are programs that extract data from transaction processing systems, process it and load it into the warehouse. It involves filtering the data, reformatting it and loading bulk data into the data warehouse.

A database: Warehouse databases are relational databases capable of storing large amount of data. They must be able to load data at high speeds and allow complex queries to run on them for business analysis.

Data analysis tools: They are tools that are used to perform statistical analysis of data. In general, results are displayed in graphical format for understanding.

Business intelligence tools are used to transform the large amount of data in the data warehouses or data marts into useful information. They are used for easy interpretation of large the large volume of data for business analysis and decision making.

ETL is the process of extracting data from one database and placing it into another database. It involves three operations

Extract: This stage involves pulling data from different source systems. This is the data which is to be processed and loaded into the target system. There ae several data source formats being relational or non-relational database structures. An important part of extraction involves validating the data which

has been pulled from the source system to ensure it has the correct values. In case the data validation process fails, data needs to be sent back to the source system for altering the data.

Transform: In this stage a series of rules are applied to the data which has been extracted so that the processed data is ready to be loaded to the end target. This might involve applying a few business rules (calculations, derivations), filtering the data, joining the data extracted from multiple sources, applying data validation, sorting the data etc.

Load:

Load phase loads the data to the end target which might be a file or a database. Some data warehouses might update existing information. Updating of the data is done on a frequent basis. The quality of the ETL process depends on effective working of all the three phases.

5) What is unstructured database (like Hadoop, Casandra, No-SQL) and how is it different from the traditional DB and its application?

Unstructured data refers to data that does not have a predefined data model. Data in traditional databases are confined in records.

Unstructured data is data that comes from a variety of sources such as emails, text documents, videos photos etc. Unstructured data are complex and voluminous. Hence handling queries for such data in traditional databases is not efficient. Traditional relational databases can't categorize unstructured data. They are structured to handle categorized data. Traditional databases are designed to handle steady data rather than unstructured databases which have the ability to handle growing data. Even if traditional databases are used to handle growing data, it will be expensive. There will be a need to update the software and hardware requirements to process unstructured data which will be expensive.

Temporary files store the contents of the deleted files temporarily.

After deleting data from the files, all indexes have to be rebuilt. New files have to be created. The data

from the old files need to be copied to the new file. Now the old file has to be deleted.

6) Suppose we are commissioned to design a schema for a new map and direction-finding site, Mondo Maps. Like MapQuest and Google Maps, our site needs to display route and map information. Underneath it lies a database of cities, states, roads, and landmarks, in a two dimensional plane (with longitude and latitude specifying a coordinate).

• States have abbreviations (unique) and names, and each state has a unique boundary.

• Cities are unique within states and have names, and each city has a single boundary.

• A boundary corresponds to a city or state, and it has an ID and a polygon. (Assume there is a special polygon data type.)

• Roads have IDs and names, and are made up of multiple segments. Assume that a road is associated with a single city.

• A road segment has a start and an end coordinates, as well as directionality (one-way or twoway).

• A landmark has a single coordinate, a name, and a type. Assume landmarks are associated with cities, and that landmark names are globally unique.

7) Provide definition and example of the following normal forms: 1NF, 2NF, 3NF and 3.5NF. Give two examples of each, a) when would you want to use normalization; b) when not to use normalization

Normalization is the process of successfully reducing relation with anomalies and create smaller and well-structured relations.

First Normal Form

A relation is in first normal form if it satisfies the following criteria

a. There should not be any repeating groups in the relation (multivalued attributes should not be present in a particular column).

b. A primary key is defined in the relation which uniquely identifies each of the row in the table.

Student ID Student Name Course Year of Enrollment

1000 John History

Science

2015

Student ID Student Name Course Year of Enrollment

1000 John History 2015

1000 John Science 2015

Second Normal Form

A relation is said to be in second normal form if no partial dependency exists. A partial dependency is a condition in which a non-key attribute is functionally dependent on part of the primary key (primary key here is a composite primary key).

Order ID Date Customer

ID

Customer

Name

Customer

Address

Product

ID

Product

Descri.

Product

Finish

Prod. St.

Price

Partial Dependencies Removed

Order ID Product ID

ProductID Product Description Product Finish Product Standard Price

OrderID Order Date Customer ID Customer Name Customer Address

Third Normal Form

A relation is said to be in third normal form if no transitive dependencies exist in the relation. A transitive dependency is a functional dependency between the primary key and one or more non key attributes that are dependent on the primary key via another non key attribute.

In 2 NF

Order ID Product ID

ProductID Product Description Product Finish Product Standard Price

OrderID Order Date Customer ID Customer Name Customer Address

In 3NF

Order ID	Product ID

ProductID	Product Description	Product Finish	Product Standard Price

Order ID	Order Date	CustomerID

Customer ID	Customer Name	CustomerAddress

Boyce-Codd Normal Form (3.5 NF)

A table is in BCNF if every determinant could be a primary key (candidate key).

Student	Subject	Teacher
Smith	Math	Dr.White
Smith	English	Mr.Brown
Jones	Math	Dr.White
Jones	English	Dr.Brown
Doe	Math	Dr.Green

In 3.5 NF

Student	Teacher
Smith	Dr.White
Smith	Dr.Brown
Jones	Dr.White
Jones	Dr.Brown
Doe	Dr.Green

Teacher	Subject
Dr.White	Math
Dr.Brown	Engllish
Dr.Green	Math
Teacher	Subject
Dr.White	Math

7 a) Normalization – Reasons

- Reducing anomalies
- When old systems are reverse engineered to avoid redundancies
- Quality check for the relations

7 b) Not to use Normalization

- All the required data is present in the parent table, hence the data can be queried fast. There is no need to normalize such a table.
- Searching for data is difficult in many normalized tables than searching for one single table.

8) Describe the process and implementation of a database performance tuning

- We can use indexes to make database querying more efficient. If it is known that certain records will be searched more than others, it is better to put an index on that record.
- Query analyzers or execution in the databases can be used to understand how the queries function and it can help to look for problems in the database design. It can point out troublesome commands that can cause problems in execution
- Database replication allows several databases to contain the same information. This can help in effective querying since the copies of the database can take some burden of the querying carried out. Database replication helps keep all databases synchronized.
- Partitioning is a technique used to store different records in different physical locations. The separate partitions act independently. Backup of individual partitions can be done separately to improve performance.