What is DWH?

Data Huge Storage.

What are used for DWHs?

Only Databases used for DWHs.

Can you define some warehouses?

Teradata, Oracle, SQL Server , DB2 etc...

How do you differentiate OLTP DB and OLAP database if they are in same size?

See the model of the database, if it is normalized model then OLTP, otherwise OLAP.

Differences between OLTP to DWH DB?

OLTP DB:

Structure / Model: Normalized structure

More tables available, Small sized tables.
No duplicates [no redundancy]
Suitable for Select, Insert, Update, Delete etc... operations [Volatality]
To work with a specific business operation, multiple tables requred,
hence multiple joins required.
Access frequency is high

Ex: Trains tickets booking store database

Sizes in real-time: GBs-TB only [<1 TB]

OLAP DB:

Structure / Model: De-Normalized structure

Less tables available [but more data volumed tables]
More duplicates [redundant data], hence we are able to maintain History
Suitable for Retrieval or Analysis of Data [Non-Volatality]
To work with a specific business operations, less tables required, hence less joins sufficient.
Access frequency is low
Ex: Train tickets historical bookings data store

Sizes in real-time: TBs-PBs.

BOA Production DWH size: 3 PB
DBS Bank: 10 PB

What are the memory size calculations?

1 Byte= 8 Bits
1 KB= 1024 Bytes
1 MB=1024 KB
1 GB=1024 MB
1 TB=1024 GB
1 PB=1024 TB
etc...

What are the protocols we need to follow to construct a DWH?

a) Characteristics / Principles [Inmon, Kimball, Kelly etc...]
b) Approach [Top down Inmon approach / Bottom up Kimball approach]
c) Life Cycle [Agile Waterfall, Iterative Incremental etc...]

Which DWH characteristics /principles companies follow?

Inmon Characteristics, so we called him with gratitude as "Father of DWH".

He suggested four principles to construct a DWH strongly

a) Subject Oriented

Keep the data in subject area based or business functionality based. So that analysis will be easier and respective team can go for it.
These technically called as "Datamarts".

Ex:
DWH Example: ICICI DWH
Datamart example: Savings account datamart, Current account datamart

b) Integrated

While loading data from different systems into DWH, we should follow some standard rules or protocols.

c) Non-Volatile

Do not change the data and always accumulate [increment] data, so that history maintained, and support detailed analysis.

d) Time varient

If you store data based on timeframes, then it is easier for the below types of analysis.

a) Current understanding
b) Past analysis
c) Future prediction

Which approach a company uses and how many approaches available?

There are two types of approaches

a) Top-Down approach: DWH first----> Datamart [Inmon approach]
b) Bottom-up approach: Datamarts----> DWH [Kimball approach]

Most of the companies following KIMBALL approach because it is cost / resource / memory effective.

Which life cycles are used by DWH projects and explain steps?

Mostly organizations using "Agile methodology".

Agile: Agility --> Quickly Changing

--Small teams [5-8 members]
--Fastest development and deployment approach
--Cross Functional Teams available
- **Requirements Gathering**
- **Analysis**
- **Design**
- **Code**
- **Test**
- **Peer-Review**
- **Deploy**

--Daily Scrum meeting [15 minutes and Standup meeting]

DWH Projects require 3 Life cycles

a) ETL Life cycle [above steps]
b) Semantic Life Cycle [above steps]
c) Reporting Life Cycle [above steps]

DWH tables load?

Few tables loaded daily, few tables weekly, few tables monthly....
This is called "granularity". The level of data you are maintaining in DWH projects.

Low Grain and Recommended: "Day" [So that all upper levels of analysis we can do easily]

What is Hierarchy, where we create hierarchy and Explain Date Hierarchy?
Hierarchy contain multiple levels and members. Minimum two levels required.
We usually create on textual columns.

Ex: Location hierarchy
Continent-->Country-->States-->Districts-->Mandals-->Villages-->Wards

Date hierarchy:

Day-->Week-->Fortnight-->Month-->Quarter--> Semister--> Year -->Quadyear (leap year)--> Decade-->Century etc...

Time Hierarchy:

Milliseconds-->Seconds-->Minutes-->Hours

What is Datetime and timestamp?

It is the combination of date and time value. Based on the milliseconds we would recognize the timestamp type.

YYYY-MM-DD HH:MI:SS.NNNNNN

As we have 6 Ns, it is timestmap (6).

What is OLAP and How many OLAPs available?

Online Analytical processing. There are 4 types of OLAP

a) ROLAP (Relational OLAP):

Data and Aggregated information in relational format.

Ex:Location wise total business value.

b) MOLAP (Multidimensional OLAP):

Data and aggreated information in multidimensional format (cube area)

c) HOLAP (Hybrid OLAP):

Data in relational format and aggregated data in multidimensional format

d) DOLAP (Desktop OLAP):

Data and aggregated data in the form of documents [excel, foxpro,lotus etc...]

In real-time, DWH and BI projects mostly use MOLAP

Small and medium analytical applications in ROLAP and DOLAP (HR team, Executives, Managers etc...)

What is Modern DWH ?

Here the warehouse support all types of data [structured, semi structured, and unstructured] and possible for all storages [cloud, on-premises].

Companies using along with regular warehouse, Data lakes for this purpose.

Data Lake: Which stores all types of data, structured data in "relational format", and semi structured and unstructured data in "file stream format"

Popular lakes are a) Azure Data Lake b) Hadoop Data Lake

What is Data Insight in Modern DWH?

Simply insight talks about indetailed data presentation.

Quick Insights are automatically generated visuals in Power BI.

What are the famous big data storages in Azure?

Blob storage, Data Lake, AZURE SQL Data warehouse, Snow Flake etc.

How many places Databases and Datawarehouses maintained in the organizations?
Now a days in two places

a) On-Premises

- **SQL Database**
- **Datawarehouse**
- **Analysis Services Databases [tabular model and multidimensional model]**

b) Cloud [Azure]

- **SQL Database**
- **Datawarehouse**
- **Analysis Services Databases [Tabular Model]**
- **Data Lakes**
- **Blob storage [Binary Large Objects]**

a) KIMBALL approach: Bottom-up approach

The advantages of this approach are:

Faster and easier implementation of manageable pieces
Favorable return on investment and proof of concept _ Less risk of failure
Inherently incremental; can schedule important data marts first
Allows project team to learn and grow

The disadvantages are:

Takes longer to build even with an iterative method
High exposure/risk to failure
Needs high level of cross-functional skills
High outlay without proof of concept

b) INMON approach: Top-down approach

The advantages of this approach are:

A truly corporate effort, an enterprise view of data
Inherently architected not a union of disparate data marts
Single, central storage of data about the content centralized rules
and control
May see quick results if implemented with iterations

The disadvantages are:

Each data mart has its own narrow view of data _ Permeates redundant data in every data mart
Perpetuates inconsistent and irreconcilable data
Proliferates unmanageable interfaces