Data Warehousing Interview Questions and Answers

What is Data Warehousing?
A data warehouse is the main repository of an organization's historical data, its corporate memory. It contains the raw material for management's decision support system. The critical factor leading to the use of a data warehouse is that a data analyst can perform complex queries and analysis, such as data mining, on the information without slowing down the operational systems (Ref:Wikipedia). Data warehousing collection of data designed to support management decision making. Data warehouses contain a wide variety of data that present a coherent picture of business conditions at a single point in time. It is a repository of integrated information, available for queries and analysis.

What are fundamental stages of Data Warehousing?
Offline Operational Databases - Data warehouses in this initial stage are developed by simply copying the database of an operational system to an off-line server where the processing load of reporting does not impact on the operational system's performance.
Offline Data Warehouse - Data warehouses in this stage of evolution are updated on a regular time cycle (usually daily, weekly or monthly) from the operational systems and the data is stored in an integrated reporting-oriented data structure
Real Time Data Warehouse - Data warehouses at this stage are updated on a transaction or event basis, every time an operational system performs a transaction (e.g. an order or a delivery or a booking etc.)
Integrated Data Warehouse - Data warehouses at this stage are used to generate activity or transactions that are passed back into the operational systems for use in the daily activity of the organization. (Reference Wikipedia)

What is Dimensional Modeling?
Dimensional data model concept involves two types of tables and it is different from the 3rd normal form. This concepts uses *Facts* table which contains the measurements of the business and *Dimension* table which contains the context(dimension of calculation) of the measurements.

What is Fact table?
Fact table contains measurements of business process. Fact table contains the foreign keys for the dimension tables. Example, if you are business process is "paper production", "average production of paper

by one machine" or "weekly production of paper" will be considered as measurement of business process.

What is Dimension table?
Dimensional table contains textual attributes of measurements stored in the facts tables. Dimensional table is a collection of hierarchies, categories and logic which can be used for user to traverse in hierarchy nodes.

What are the Different methods of loading Dimension tables?
There are two different ways to load data in dimension tables.

Conventional (Slow) :
All the constraints and keys are validated against the data before, it is loaded, this way data integrity is maintained.

Direct (Fast) :
All the constraints and keys are disabled before the data is loaded. Once data is loaded, it is validated against all the constraints and keys. If data is found invalid or dirty it is not included in index and all future processes are skipped on this data.

What is OLTP?
OLTP is abbreviation of On-Line Transaction Processing. This system is an application that modifies data the instance it receives and has a large number of concurrent users.

What is OLAP?
OLAP is abbreviation of Online Analytical Processing. This system is an application that collects, manages, processes and presents multidimensional data for analysis and management purposes.

What is the difference between OLTP and OLAP?

Data Source
OLTP: Operational data is from original data source of the data
OLAP: Consolidation data is from various source.

Process Goal
OLTP: Snapshot of business processes which does fundamental business tasks
OLAP: Multi-dimensional views of business activities of planning and decision making

Queries and Process Scripts
OLTP: Simple quick running queries ran by users.
OLAP: Complex long running queries by system to update the aggregated data.

Database Design
OLTP: Normalized small database. Speed will be not an issue due to smaller database and normalization will not degrade performance. This adopts entity relationship(ER) model and an application-oriented database design.
OLAP: De-normalized large database. Speed is issue due to larger database and de-normalizing will improve performance as there will be lesser tables to scan while performing tasks. This adopts star, snowflake or fact constellation mode of subject-oriented database design.

Back up and System Administration
OLTP: Regular Database backup and system administration can do the job.
OLAP: Reloading the OLTP data is good considered as good backup option.

What are normalization forms?
Please visit here http://blog.sqlauthority.com/2007/04/15/sql-server-interview-questions/

Describes the foreign key columns in fact table and dimension table?
Foreign keys of dimension tables are primary keys of entity tables.
Foreign keys of facts tables are primary keys of Dimension tables.

What is Data Mining?
Data Mining is the process of analyzing data from different perspectives and summarizing it into useful information.

What is the difference between view and materialized view?
A ***view*** takes the output of a query and makes it appear like a virtual table and it can be used in place of tables.
A ***materialized view*** provides indirect access to table data by storing the results of a query in a separate schema object.

What is ER Diagram?
Entity Relationship Diagrams are a major data modelling tool and will

help organize the data in your project into entities and define the relationships between the entities. This process has proved to enable the analyst to produce a good database structure so that the data can be stored and retrieved in a most efficient manner.

An entity-relationship (ER) diagram is a specialized graphic that illustrates the interrelationships between entities in a database. A type of diagram used in data modeling for relational data bases. These diagrams show the structure of each table and the links between tables.

What is ODS?
ODS is abbreviation of Operational Data Store. A database structure that is a repository for near real-time operational data rather than long term trend data. The ODS may further become the enterprise shared operational database, allowing operational systems that are being re-engineered to use the ODS as there operation databases.

What is ETL?
ETL is abbreviation of extract, transform, and load. ETL is software that enables businesses to consolidate their disparate data while moving it from place to place, and it doesn't really matter that that data is in different forms or formats. The data can come from any source.ETL is powerful enough to handle such data disparities. First, the extract function reads data from a specified source database and extracts a desired subset of data. Next, the transform function works with the acquired data - using rules orlookup tables, or creating combinations with other data - to convert it to the desired state. Finally, the load function is used to write the resulting data to a target database.

What is VLDB?
VLDB is abbreviation of Very Large DataBase. A one terabyte database would normally be considered to be a VLDB. Typically, these are decision support systems or transaction processing applications serving large numbers of users.

Is OLTP database is design optimal for Data Warehouse?
No. OLTP database tables are normalized and it will add additional time to queries to return results. Additionally OLTP database is smaller and it does not contain longer period (many years) data, which needs to be analyzed. A OLTP system is basically ER model and not Dimensional Model. If a complex query is executed on a OLTP system, it may cause a heavy overhead on the OLTP server that will affect the

normal business processes.

If de-normalized is improves data warehouse processes, why fact table is in normal form?
Foreign keys of facts tables are primary keys of Dimension tables. It is clear that fact table contains columns which are primary key to other table that itself make normal form table.

What are lookup tables?
A lookup table is the table placed on the target table based upon the primary key of the target, it just updates the table by allowing only modified (new or updated) records based on thelookup condition.

What are Aggregate tables?
Aggregate table contains the summary of existing warehouse data which is grouped to certain levels of dimensions. It is always easy to retrieve data from aggregated tables than visiting original table which has million records. Aggregate tables reduces the load in the database server and increases the performance of the query and can retrieve the result quickly.

What is real time data-warehousing?
Data warehousing captures business activity data. Real-time data warehousing captures business activity data as it occurs. As soon as the business activity is complete and there is data about it, the completed activity data flows into the data warehouse and becomes available instantly.

What are conformed dimensions?
Conformed dimensions mean the exact same thing with every possible fact table to which they are joined. They are common to the cubes.

What is conformed fact?
Conformed dimensions are the dimensions which can be used across multiple Data Marts in combination with multiple facts tables accordingly.

How do you load the time dimension?
Time dimensions are usually loaded by a program that loops through all possible dates that may appear in the data. 100 years may be represented in a time dimension, with one row per day.

What is a level of Granularity of a fact table?
Level of granularity means level of detail that you put into the fact

table in a data warehouse. Level of granularity would mean what detail are you willing to put for each transactional fact.

What are non-additive facts?
Non-additive facts are facts that cannot be summed up for any of the dimensions present in the fact table. However they are not considered as useless. If there is changes in dimensions the same facts can be useful.

What is factless facts table?
A fact table which does not contain numeric fact columns it is called factless facts table.

What are slowly changing dimensions (SCD)?
SCD is abbreviation of Slowly changing dimensions. SCD applies to cases where the attribute for a record varies over time. There are three different types of SCD.

1) SCD1 : The new record replaces the original record. Only one record exist in database - current data.

2) SCD2 : A new record is added into the customer dimension table. Two records exist in database - current data and previous history data.

3) SCD3 : The original data is modified to include new data. One record exist in database - new information are attached with old information in same row.

What is hybrid slowly changing dimension?
Hybrid SCDs are combination of both SCD 1 and SCD 2. It may happen that in a table, some columns are important and we need to track changes for them i.e capture the historical data for them whereas in some columns even if the data changes, we don't care.

What is BUS Schema?
BUS Schema is composed of a master suite of confirmed dimension and standardized definition if facts.

What is a Star Schema?
Star schema is a type of organizing the tables such that we can retrieve the result from the database quickly in the warehouse environment.

What Snow Flake Schema?
Snowflake Schema, each dimension has a primary dimension table, to

which one or more additional dimensions can join. The primary dimension table is the only table that can join to the fact table.

Differences between star and snowflake schema?
Star schema - A single fact table with N number of Dimension, all dimensions will be linked directly with a fact table. This schema is de-normalized and results in simple join and less complex query as well as faster results.

Snow schema - Any dimensions with extended dimensions are know as snowflake schema, dimensions maybe interlinked or may have one to many relationship with other tables. This schema is normalized and results in complex join and very complex query as well as slower results.

What is Difference between ER Modeling and Dimensional Modeling?
ER modeling is used for normalizing the OLTP database design. Dimensional modeling is used for de-normalizing the ROLAP/MOLAP design.

What is degenerate dimension table?
If a table contains the values, which are neither dimension nor measures is called degenerate dimensions.

Why is Data Modeling Important?
Data modeling is probably the most labor intensive and time consuming part of the development process. The goal of the data model is to make sure that the all data objects required by the database are completely and accurately represented. Because the data model uses easily understood notations and natural language , it can be reviewed and verified as correct by the end-users.

In computer science, data modeling is the process of creating a data model by applying a data model theory to create a data model instance. A data model theory is a formal data model description. When data modelling, we are structuring and organizing data. These data structures are then typically implemented in a database management system. In addition to defining and organizing the data, data modeling will impose (implicitly or explicitly) constraints or limitations on the data placed within the structure.

Managing large quantities of structured and unstructured data is a primary function of information systems. Data models describe

structured data for storage in data management systems such as relational databases. They typically do not describe unstructured data, such as word processing documents, email messages, pictures, digital audio, and video. (Reference : Wikipedia)

What is surrogate key?
Surrogate key is a substitution for the natural primary key. It is just a unique identifier or number for each row that can be used for the primary key to the table. The only requirement for a surrogate primary key is that it is unique for each row in the table. It is useful because the natural primary key can change and this makes updates more difficult.Surrogated keys are always integer or numeric.

What is Data Mart?
A data mart (DM) is a specialized version of a data warehouse (DW). Like data warehouses, data marts contain a snapshot of operational data that helps business people to strategize based on analyses of past trends and experiences. The key difference is that the creation of a data mart is predicated on a specific, predefined need for a certain grouping and configuration of select data. A data mart configuration emphasizes easy access to relevant information (Reference : Wiki). Data Marts are designed to help manager make strategic decisions about their business.

What is the difference between OLAP and data warehouse?
Datawarehouse is the place where the data is stored for analyzing where as OLAP is the process of analyzing the data,managing aggregations, partitioning information into cubes for in depth visualization.

What is a Cube and Linked Cube with reference to data warehouse?
Cubes are logical representation of multidimensional data.The edge of the cube contains dimension members and the body of the cube contains data values. The linking in cube ensures that the data in the cubes remain consistent.

What is junk dimension?
A number of very small dimensions might be lumped together to form a single dimension, a junk dimension - the attributes are not closely related. Grouping of Random flags and text Attributes in a dimension and moving them to a separate sub dimension is known as junk dimension.

What is snapshot with reference to data warehouse?
You can disconnect the report from the catalog to which it is attached by saving the report with a snapshot of the data.

What is active data warehousing?
An active data warehouse provides information that enables decision-makers within an organization to manage customer relationships nimbly, efficiently and proactively.

What is the difference between data warehousing and business intelligence?
Data warehousing deals with all aspects of managing the development, implementation and operation of a data warehouse or data mart including meta data management, data acquisition, data cleansing, data transformation, storage management, data distribution, data archiving, operational reporting, analytical reporting, security management, backup/recovery planning, etc. Business intelligence, on the other hand, is a set of software tools that enable an organization to analyze measurable aspects of their business such as sales performance, profitability, operational efficiency, effectiveness of marketing campaigns, market penetration among certain customer groups, cost trends, anomalies and exceptions, etc. Typically, the term "business intelligence" is used to encompass OLAP, data visualization, data mining and query/reporting tools. (Reference : Les Barbusinski)

Explain paradigm of Bill Inmon and Ralph Kimball.

Bill Inmon's paradigm: Data warehouse is one part of the overall business intelligence system. An enterprise has one data warehouse, and data marts source their information from the data warehouse. In the data warehouse, information is stored in 3rd normal form.

Ralph Kimball's paradigm: Data warehouse is the conglomerate of all data marts within the enterprise. Information is always stored in the dimensional model.