**Part 1**[**Part 2**](http://careerride.com/Data-warehousing-part-2.aspx)[**Part 3**](http://careerride.com/Data-warehousing-part-3.aspx)[**Part 4**](http://careerride.com/Data-WareHousing-Interview-Questions-4.aspx)[**Part 5**](http://careerride.com/Data-WareHousing-Interview-Questions-5.aspx)[**Part 6**](http://careerride.com/Data-WareHousing-Interview-Questions-6.aspx)[**Part 7**](http://careerride.com/Data-WareHousing-Interview-Questions-7.aspx)

**Data Warehousing Interview questions and answers- posted on August 07, 2008 at 17:10 pm by Rajmeet Ghai**

**1.**[**What is Data warehousing?**](http://careerride.com/Data-warehousing-defined.aspx)

**Answer -**A data warehouse can be considered as a storage area where interest specific or relevant data........

**2.**[**What are fact tables and dimension tables?**](http://careerride.com/Data-warehousing-fact-dimension-tables.aspx)

**Answer -**As mentioned, data in a warehouse comes from the transactions. Fact table in a data warehouse consists........

**3.**[**What is ETL process in data warehousing?**](http://careerride.com/Data-warehousing-ETL-process.aspx)

**Answer -**ETL is Extract Transform Load. It is a process of fetching.........

**4.**[**Explain the difference between data mining and data warehousing.**](http://careerride.com/Data-warehousing-data-mining.aspx)

**Answer -**Data warehousing is merely extracting data from different sources, cleaning the........

**5.**[**What is an OLTP system and OLAP system?**](http://careerride.com/Data-warehousing-OLTP-OLAP.aspx)

**Answer -**OLTP: Online Transaction and Processing helps and manages applications based........

**6.**[**What are cubes?**](http://careerride.com/Data-warehousing-cubes.aspx)

**Answer -**A data cube stores data in a summarized version which helps in a faster analysis of data..........

**7.**[**What is snow flake scheme design in database?**](http://careerride.com/Data-warehousing-snow-flake.aspx)

**Answer -**A snowflake Schema in its simplest form is an arrangement of fact tables.........

**8.**[**What is analysis service?**](http://careerride.com/Data-warehousing-analysis-service.aspx)

**Answer -**Analysis service provides a combined view of the data used in OLAP.........

**9.**[**Explain sequence clustering algorithm.**](http://careerride.com/Data-warehousing-sequence-clustering.aspx)

**Answer -**Sequence clustering algorithm collects similar or related paths, sequences of data.........

**10.**[**Explain discrete and continuous data in data mining.**](http://careerride.com/Data-warehousing-discrete-continuous.aspx)

**Answer -**Discreet data can be considered as defined or finite data..........

**11.**[**Explain time series algorithm in data mining.**](http://careerride.com/Data-warehousing-time-series-algorithm.aspx)

**Answer -**Time series algorithm can be used to predict continuous values of data.......

**12.**[**What is XMLA?**](http://careerride.com/Data-warehousing-XMLA.aspx)

**Answer -**XMLA is XML for Analysis which can be considered as a standard for accessing data in OLAP.......

**13.**[**Explain the difference between Data warehousing and Business Intelligence.**](http://careerride.com/Data-warehousing-Business-Intelligence.aspx)

**Answer -**Data Warehousing helps you store the data while business intelligence.....

**14.**[**What is Dimensional Modeling?**](http://careerride.com/Data-warehousing-Dimensional-Modeling.aspx)

**Answer -**Dimensional modeling is often used in Data warehousing.......

**15.**[**What is surrogate key? Explain it with an example.**](http://careerride.com/Data-warehousing-surrogate-key.aspx)

**Answer -**Data warehouses commonly use a surrogate key.......

**16.**[**What is the purpose of Factless Fact Table?**](http://careerride.com/Data-warehousing-Factless-Fact-Table.aspx)

**Answer -**Fact less tables are so called because they simply contain........

**17.**[**What is a level of Granularity of a fact table?**](http://careerride.com/Data-warehousing-Granularity-fact-table.aspx)

**Answer -**A fact table is usually designed at a low level of Granularity...........

**18.**[**Explain the difference between star and snowflake schemas.**](http://careerride.com/Data-warehousing-star-snowflake.aspx)

**Answer -**A snow flake schema design is usually more complex than a start schema.....

**19.**[**What is the difference between view and materialized view?**](http://careerride.com/Data-warehousing-materialized-view.aspx)

**Answer -**A view is created by combining data from different tables........

**20.**[**What is a Cube and Linked Cube with reference to data warehouse?**](http://careerride.com/Data-warehousing-Cube-Linked-Cube.aspx)

**Answer -**A data cube stores data in a summarized version which helps in a faster analysis............

**21.**[**What is junk dimension?**](http://careerride.com/Data-warehousing-junk-dimension.aspx)

**Answer -**In scenarios where certain data may not be appropriate to store in....

**22.**[**What are fundamental stages of Data Warehousing?**](http://careerride.com/Data-warehousing-fundamental-stages.aspx)

**Answer -**Stages of a data warehouse helps to find and understand........

**Data Warehousing Interview - posted on May 11, 2009 at 14:40 pm by Vidya Sagar**

**23.**[**What is Virtual Data Warehousing?**](http://careerride.com/Data-warehousing-what-is-virtual-data.aspx)

The aggregate view of complete data inventory is provided by Virtual Warehousing........

**24.**[**What is active data warehousing?**](http://careerride.com/Data-warehousing-what-is-active-data.aspx)

The transactional data captured and reposited in the Active Data Warehouse......

**25.**[**List down differences between dependent data warehouse and independent data warehouse.**](http://careerride.com/Data-warehousing-dependent-data-independent-data.aspx)

Dependent data ware house are build........

**26.**[**What is data modeling and data mining? What is this used for?**](http://careerride.com/Data-warehousing-data-modeling-and-data-mining.aspx)

Designing a model for data or database is called data modelling........

**27.**[**Difference between ER Modeling and Dimensional Modeling.**](http://careerride.com/Data-warehousing-ER-modeling-and-dimensional-modeling.aspx)

Dimensional modelling is very flexible for the user perspective........

**28.**[**What is snapshot with reference to data warehouse?**](http://careerride.com/Data-warehousing-what-is-snapshot.aspx)

A snapshot of data warehouse is a persisted report from the catalogue.........

**29.**[**What is degenerate dimension table?**](http://careerride.com/Data-warehousing-what-is-degenerate-dimension-table.aspx)

The dimensions that are persisted in the fact table is called dimension table.........

**30.**[**What is Data Mart?**](http://careerride.com/Data-warehousing-what-is-data-mart.aspx)

Data Mart is a data repository which is served to a community of people.......

**31.**[**What is the difference between metadata and data dictionary?**](http://careerride.com/Data-warehousing-metadata-and-data-dictionary.aspx)

Metadata describes about data. It is ‘data about data’. It has information about how and when......

**32.**[**Describe the various methods of loading Dimension tables.**](http://careerride.com/Data-warehousing-methods-of-loading-dimension-tables.aspx)

The following are the methods of loading dimension tables.......

**33.**[**What is the difference between OLAP and data warehouse?**](http://careerride.com/Data-warehousing-OLAP-and-data-warehouse.aspx)

The following are the differences between OLAP and data warehousing......

**34.**[**Describe the foreign key columns in fact table and dimension table.**](http://careerride.com/Data-warehousing-foreign-key-columns.aspx)

The primary keys of entity tables are the foreign keys of dimension tables.......

**35.**[**What is cube grouping?**](http://careerride.com/Data-warehousing-what-is-cube-grouping.aspx)

A transformer built set of similar cubes is known as cube grouping. A single level in one dimension......

**36.**[**Define the term slowly changing dimensions (SCD).**](http://careerride.com/Data-warehousing-slowly-changing-dimensions-(SCD).aspx)

Slowly changing dimension target operator is one of the SQL warehousing operators......

**37.**[**What is a Star Schema?**](http://careerride.com/Data-warehousing-what-is-a-star-schema.aspx)

The simplest data warehousing schema is star schema.......

**38.**[**Differences between star and snowflake schema.**](http://careerride.com/Data-warehousing-star-and-snowflake-schema.aspx)

Star Schema: A de-normalized technique in which one fact table is associated with several dimension tables.......

**39.**[**Explain the use of lookup tables and Aggregate tables.**](http://careerride.com/Data-warehousing-lookup-tables-and-Aggregate-tables.aspx)

At the time of updating the data warehouse, a lookup table is used.......

**40.**[**What is real time data-warehousing?**](http://careerride.com/Data-warehousing-real-time.aspx)

The combination of real-time activity and data warehousing is called real time warehousing.......

**41.**[**What is conformed fact? What is conformed dimensions use for?**](http://careerride.com/Data-warehousing-conformed-fact-dimensions.aspx)

Allowing having same names in different tables is allowed by Conformed facts.......

**42.**[**Define non-additive facts.**](http://careerride.com/Data-warehousing-non-additive-facts.aspx)

The facts that can not be summed up for the dimensions present in the fact table are called non-additive facts.......

**43.**[**Define BUS Schema.**](http://careerride.com/Data-warehousing-BUS-schema.aspx)

A BUS schema is to identify the common dimensions across business processes......

**44.**[**List out difference between SAS tool and other tools.**](http://careerride.com/Data-warehousing-SAS-tool.aspx)

The differences between SAS and other tools are......

**45.**[**Why is SAS so popular?**](http://careerride.com/Data-warehousing-why-is-SAS-so-popular.aspx)

Statistical Analysis System is an integration of various software products which allows the developers to perform.......

**46.**[**What is data cleaning? How can we do that?**](http://careerride.com/Data-warehousing-what-is-data-cleaning.aspx)

Data cleaning is also known as data scrubbing. Data cleaning is a process which ensures the set of data is correct and accurate......

**47.**[**Explain in brief about critical column.**](http://careerride.com/Data-warehousing-critical-column.aspx)

A column (usually granular) is called as critical column which changes the values over a period of time.......

**48.**[**What is data cube technology used for?**](http://careerride.com/Data-warehousing-what-is-data-cube-technology.aspx)

Data cube is a multi-dimensional structure. Data cube is a data abstraction to view aggregated data from a number of perspectives.........

### 56. What is the purpose of cluster analysis in Data Warehousing?

Cluster analysis is used to define the object without giving the class label. It analyzes all the data that is present in the data warehouse and compare the cluster with the cluster that is already running. It performs the task of assigning some set of objects into the groups are also known as clusters. It is used to perform the data mining job using the technique like statistical data analysis. It includes all the information and knowledge around many fields like machine learning, pattern recognition, image analysis and bio-informatics. Cluster analysis performs the iterative process of knowledge discovery and includes trials and failures. It is used with the pre-processing and other parameters as a result to achieve the properties that are desired to be used.

### 57. What are the different models used in cluster analysis?

There are many algorithms that can be used to analyze the database to check the maintenance of all the data sets that are already present. The different types of cluster models include as follows:  
• Connectivity models: these are the models that connect one cluster to another cluster. This includes the example of hierarchical clustering that is based on the distance connectivity of one model to another model.   
• Centroid models: these are the models that are used to find the clusters using the single mean vector. It includes the example of k-means algorithm.  
• Distribution models: it includes the specification of the models that are statistically distributed for example multivariate normal distribution model.  
• Density models: deals with the clusters that are densely connected with one another in the regions having the data space.   
• Group models: specifies the model that doesn’t provide the refined model for the output and just gives the grouping information.

### 58. What is the difference between agglomerative and divisive Hierarchical Clustering?

• Agglomerative Hierarchical clustering method allows the clusters to be read from bottom to top and it follows this approach so that the program always reads from the sub-component first then moves to the parent. Whereas, divisive uses top-bottom approach in which the parent is visited first then the child.   
• Agglomerative hierarchical method consists of objects in which each object creates its own clusters and these clusters are grouped together to create a large cluster. It defines a process of merging that carries on till all the single clusters are merged together into a complete big cluster that will consists of all the objects of child clusters. Whereas, in divisive the parent cluster is divided into smaller cluster and it keeps on dividing till each cluster has a single object to represent.

### 59. Why is chameleon method used in data warehouse?

Chameleon is a hierarchical clustering algorithm that overcomes the limitations of the existing models and the methods present in the data warehousing. This method operates on the sparse graph having nodes that represent the data items and edges represent the weights of the data items. The representation of it allows large data set to be created and operated on successfully. The method finds the clusters that are used in the data set using the two phase algorithm. The first phase consists of the graph partitioning that allows the clustering of the data items into large number of sub-clusters. Second phases use an agglomerative hierarchical clustering algorithm to search for the clusters that are genuine and can be combined together with the sub-clusters that are produced.

### 60. What are the key features of chameleon that separates it from other algorithms?

The key features that are in the chameleon are:  
• The chameleon method determines the pair of similar sub-clusters that can be connected with other clusters. It also finds the closeness of the clusters from one another.   
• The chameleon with the above property overcomes the limitation that is present in agglomerative hierarchical model.   
• It uses different methods to take the internal characteristics of the clusters and matches with those which are already present.   
• It doesn't depend on static model that is supplied by the user and uses automated functions to perform the merging of the clusters that are already associated in the cluster.

### 61. Why is Business Intelligence (BI) important in data warehousing?

Business intelligence is required everywhere and it is most importantly required in data warehousing due to the technologies that are getting used in it. It allows the applications and the practices to be performed according to the business rules and regulations. Business intelligence consists of collection, integration, analysis and the presentation of the information that is used to make business related decisions. This is also used to build business related system like decision support system that provide historical, current and predictive data that can be used in the favor of setting up the environment of the database system. It uses computer based techniques to extract the information from many resources and analyze it according to the industry. It is used to provide nearly accurate decision making system.

### 62. What is the use of dimensional modeling in data warehousing?

Dimensional modeling is a set of techniques that is used in designing the overall structure of data warehousing. It doesn't involve relational database but at the logical level uses the physical form of the database. It is used to support user queries and to increase the performance and understanding of a particular database concept. It uses facts and dimensions to support the measures and the context of the database. The facts define the values that can be aggregated and dimensions represent the group of hierarchies and the descriptors that define the facts in return. This type of models is built by business process model and consists inside the process area. This process area consists of the same design and operation details as others.

### 63. What are the steps involved in creating dimensional modeling process?

The business process of the dimensional modeling includes:  
• Choose The Business Process: in this 4-step design method is followed that helps to provide the usability of the dimensional model. This allows the business process to be more systematic in representation and more helpful in explaining it as well. It includes the use of Business Process Modelling Notation (BPMN) or Unified Modelling Language (UML).  
• Declaring The Grain: after choosing the business process the declaration of the model comes that consists of grains. The grain of the model provides the accurate description of the dimensional model and allows the focus should be shifted there.  
• Identify The Dimensions: in this phase the identification of the dimension takes place in the dimensional model. The dimensions are defined in the grain that is defined in the declaration part above. Dimensions acts as a foundation of the fact table where the data gets collected that comes under the fact.   
• Identify The Facts: defining the dimensions provides a way to create a table in which the fact data can be stored. These facts are populated on the basis of the numerical figures and facts.

### 64. Why is dimensional normalization not required?

Dimensional normalization allows the database related problems to be solved. It is used to remove the redundant attributes that are used as de-normalized dimensions. Dimensions consist of sub-dimensions that are joined together in one. The dimensional normalization is not used due to the fact that it makes:  
• Data structure more complex and due to which the performance can be degraded as it requires lots of joining of tables and keep the relations intact.   
• The space is not utilized properly and use of more space is required.   
• The query performance suffers when aggregating or retrieving many dimensional values. This requires proper analysis and making of operational reports are necessary.

### 65. Why facts table is useful in representing the data?

Fact table allows the measurement and the values of the facts of the data to be contained inside the table. This table consists of the foreign keys and primary keys of the dimension tables. It is located in between the star schema or snowflake schema. It provides values that are additive and independent variables through which the dimensional attributes are analyzed. This table consists of the grains, which consist of atomic level of data and through which the facts in the tables are defined. Each record defines the independent facts that provide higher level of data to be given to the user. It is useful in representing the data due to easy storage and less memory to be taken to the facts of the data that are associated with it.

### 66. What are the stages that are required in Data warehousing?

There are four different kinds of stages that are required in data warehousing and they are:  
  
• Offline Operational Databases: this is the top most and initial stage that allows the database to be viewed offline without going to online. This copy the database to the operational system and an offline server that processes the load of the online and offline and allow the performance to be balanced.   
• Offline Data Warehouse: this is the second stage where the updation of the time cycle that is regular takes place. The settings are given through which the data can be set like daily, weekly, monthly and yearly. This data is taken from the operational system. The data is stored in the report oriented data structure.   
• Real Time Data Warehouse: this allows the transaction update on the event basis. It means on an event an updation occurs. The transaction is performed in the operational system as well.   
• Integrated Data Warehouse: this is the final stage and it is used to generate activity or transactions. After generation they are again put back to the operational system to be used by the user on the daily basis.

### 67. What are the different ways to represent the loading of Dimension tables?

The loading of the dimension table can be represented in two different ways:  
• Conventional (Slow): in this way the constraints and keys goes for validations that are done against the data. These validations are performed before the loading of the data takes place in the data warehousing. This ensures the data integrity that is maintained throughout the loading phase. As the name suggests it is slower method that has to be implemented. This way deals in too much complexity in creating the database.   
• Direct (Fast): in this all the constraints and the keys that are used with the data are disabled. The validations are performed after the data is loaded in data warehousing. These validations include everything that are applied on constraints and keys. Data are checked for invalid bit or dirty bit, if anything like this is found then the data is not included in the index and all the processes that require the data are skipped.

### 68. What are the different types of fact tables used in data warehousing?

There are three types of fact tables that are used and the characterization is given below:  
  
• Transactional table: is the most basic table of data warehousing. It consists of grain that has the association with the transactional fact table and it is defines to use one row per line in a transactional. It consists of the most detailed data that have many number of dimensions associated with it.   
  
• Periodic snapshots: is used to take the snapshot of the currently working data set. It depends on the period of time when the snapshots are taken. It increases the performance. This table is dependent on the transactional table as it has to be synchronized to keep all the detailed data that is mentioned in transactional table. The snapshot can be looked if any problem take place in the database and it helps in recovering the state that is lost.   
  
• Accumulating snapshots: this type of table is used to show the activity of the process that is defined from the beginning till end. For example, order process. The snapshots consist of multiple date columns that represent a timeline of the process. It keeps and entry of the associate date dimension representing the unknown date.

### 69. What are the steps involved in designing a fact table?

The fact table allows the data to be represented in the detailed manner. It has an association with the dimensional table. The steps that are required to design the fact table includes:  
• Identify the business process for analysis so that all the process can be defined and used with the complete details provided by the business process.   
• Identify the measures, constraints, facts and keys before designing the fact table. The questions have to be asked during the business process regarding the table that has to be created and the purpose of it.   
• Identify the dimensions for facts like product dimensions, location, time, etc. This phase also include the analysis of the components that are required to be involved.   
• List of the columns takes place that describe the dimension and the lowest level of detail is to be found out to be used in the fact table.

### 70. What is the difference between OLTP and OLAP?

• OLTP is known as online transaction processing and it refers to the class of systems that manages the transaction oriented applications used for the data entry and processing of the transaction. Whereas, OLAP stands for online analytical processing that defines an approach to give the reply to multi-dimensional analytical queries. It is a part of business intelligence. It allows the modification to be done of the traditional database terms.   
• OLTP uses the original data source to be taken in the database. It doesn't create any copy or uses any virtual data. OLAP uses many data sources that are taken from many places and gets stored in the database.   
• OLTP uses the business process snapshots that will handle the recovery process if any that has to be done in future. Whereas, OLAP uses the snapshots of multi-dimensional views of business activities of planning and decision making that is required to be used in the database.   
• OLTP uses the normalization database that will slow down the system due the size of the database and the normalization also degrades the performance. Whereas, OLAP uses the de-normalized process and uses the large database so the speed also increases and it improves the overall performance of the system as well.

71. What is the function of ETL when used in data warehousing?

ETL is used for extract, transform, and load and it is used to enable the business models to consolidate their data that is moving from place to place and it allows the saving of the data in different forms and different formats. The data can be used from any source and can be included by any source that is defined. It is a powerful method to handle the data disparities and other problems that are related to the data. There are few steps that are involved in using the function of ETL and it is as follows:  
• The extract function reads the data from the source that is given and stores the extracted information in the database of the path that is given.   
• The transformation of the data occurs that is done through some rules keeping the lookup tables to see the combination of the data that is getting formed and the transformed data are arranged and kept in order of the choice that is made while transforming it.   
• The load function is used load the data to the database or the user location where it needs to be located and then the resulting data is to be loaded to the target location.

72. What are the different types of SCD's used in data warehousing?

SCD (Slowly changing dimensions) that provides different attributes that are used for the record that varies over time and doesn't remain stable. There are three types of SCDs are used in data warehousing. These are defined as:   
• SCD1: this is the record that is used to replace the original record even there is only one record exists in the database. The current data will be replaced and the new data will take its place.   
• SCD2: this is the new record file that is added to the dimension table. This record exists in the database with the current data and previous data that is stored in the history.   
• SCD3: this uses the original data that is modified to the new data. This consists of two record one record that exist in the database and another record that will replace the old database record with the new information.

73. What are the different types of schemas used in Data warehousing?

Schema defines the type of structure that is used by the database to hold on some data that can be related or can be different. There are three types of schema that exists in the database as follows:  
  
• BUS schema: this is the schema that is composed of the master file that is totally confirmed with the dimension and with standardized definition including the facts.   
  
• Star schema: this is the schema that defines the organization type of the tables and it is used to retrieve the result from the database quickly in a controlled environment.   
• Snow flake schema: this schema is used to show the primary dimension table, that includes one or more dimensions that can be joined. Primary dimension table only allows the joining of the fact table.

74. What is the function of surrogate key in data warehousing?

Surrogate key is the key that can be substituted with the primary key. It is a unique key that uniquely identifies the each row that is used. This also allows the unique search of the primary key to the table. It is used in the database to uniquely identify an entity in a model or an object that is present in the database. This key is not derived from any of the data entity and not even from the application data. These are represented either in the form of numbers or integers. This kind of keys is generated from the system and it is not at all visible to the outside world. This key can be used as a primary key for a given database and there are few differences between primary key and surrogate key. There remains many to one relationship between primary and surrogate keys.

75. What is the difference between data warehouse and operational systems?

• Operational systems are optimized to preserve the data integrity of the system, whereas data warehouse are optimized to speed up the process of data analysis.   
• Operational system increases the speed of the business transactions through the use of normalization of the database and using the entity relationship models, whereas data warehouse uses de-normalization and dimension based model to speed the data retrieval.   
• Operational system uses relational databases to maintain the relationship between the tables. It also consists of insert and update process that takes very less time hence increment in the performance of the system to create the transaction. Whereas, data warehouse store the same data multiple times to keep the aggregation of the data and gather the data from the operational systems.