Extract, Transform and Load Term Test Notes (Theory)

**Functions a data quality tool performs:**

1. Error Discovery Features:

* Identify duplicate records
* Identify values that are outside the domain
* Find inconsistent data
* Check range of allowable values
* Monitor trends in data
* Report user on data quality

1. Error Correction Features:

* Normalize inconsistent data
* Remove duplicate records
* Improve merging of data
* Provide measurement of data quality
* Prevent wrong data entry

**Explain security measures in a Data warehouse environment:**

Provided using 3 mechanisms

1. User Privileges:

* Role based security: Grouping of users with common requirements for database access. Access privilege based on role/user position.

1. Password Protection:

* Users need passwords to enter into the DW environment
* Patterns and expiry dates of these passwords must be checked by the security administration
* Check on unauthorized access with multiple wrong password entries resulting in suspension of such users until the DW administration validates him/her

1. Security Tools:

* Basic third party security tools such as anti-virus or firewalls are installed to prevent any sort of unauthorized access: human or malware

**ROLAP vs MOLAP:**

| **BASIS FOR COMPARISON** | **ROLAP** | **MOLAP** |
| --- | --- | --- |
| Full Form | ROLAP stands for Relational Online Analytical Processing. | MOLAP stands for Multidimensional Online Analytical Processing. |
| Storage & Fetched | Data is stored and fetched from the main data warehouse. | Data is Stored and fetched from the Proprietary database MDDBs. |
| Data Form | Data is stored in the form of relational tables. | Data is Stored in the large multidimensional array made of data cubes. |
| Data volumes | Large data volumes. | Limited summaries data is kept in MDDBs. |
| Technology | Uses Complex SQL queries to fetch data from the main warehouse. | MOLAP engine created a recalculated and prefabricated data cubes for multidimensional data views. Sparse matrix technology is used to manage data sparsity. |
| View | ROLAP creates a multidimensional view of data dynamically. | MOLAP already stores the static multidimensional view of data in MDDBs. |
| Access | Slow access. | Faster access. |

**What is data velocity? Explain the cyclicity of data:**

* Speed at which data passes from point of capture to point of use
* Average of time-taken for data to enter the system to when it is used.
* Includes the editing, preprocessing and ETL of data so that the data can be stored in the warehouse from where the user can make use of it
* Main factor affecting data velocity : Data size
* Cyclicity: Time between/taken to change of data in the OS and reflecting that change in the DW.

**What is data cleaning? Mention reasons for dirty data. List steps in data cleaning**

* Detecting and removing errors and inconsistency in data. This could happen because of misspelled or missing values during the data entry process.
* Reasons for dirty data:

1. Dummy values
2. Absent data
3. Multipurpose fields
4. Contradicting data
5. Inappropriate use of address lines
6. Reused Primary keys
7. Data integration problems

* Steps in data-cleaning:

1. Parsing: Individual data elements are located and identified in the source system and are isolated in target files (name to first name, last name etc.)
2. Correcting: data entries are corrected using algorithms and secondary data sources
3. Standardizing: Conversion routines are used to convert the data into a standard/consistent format
4. Matching: Remove duplicates and match original data with new target and standardized elements
5. Merging: Merging of records into one representation by analyzing and identifying matching records
6. Data Staging: data is staged onto the warehouse meaning loaded but with no end user access

**Explain roles and responsibilities with data quality framework**

* Data consumer: Users, that use the DW for queries, reports, etc.
* Data producer: Maintain the quality of data input from sources
* Data expert: Identify pollution in source system
* Data policy administrator: Resolve the data corruption that occurs when data is loaded to DW
* Data integrity specialist: Make sure data in source system conform to the business rules
* Data correction authority: Apply data cleaning techniques
* Data consistency expert: Responsible for synchronizing the data within the DW

**Explain different levels of testing in data warehouse**

* Unit testing: Each development unit is tested on its own by the developer of that particular module
* Integration testing: Different components of DW are tested together to make sure they work together.
* System testing: Entire DW application is tested as a single unit.
* Performance testing: Test whether the ETL process completes within load window, and what is the time taken. Test the time taken to refresh standard reports and complex reports

**Explain Data extraction process in ETL with Figures**

**Explain data transformation in ETL**

The transformation process deals with rectifying the inconsistency. It takes the following steps:

* Map the input data from source system to data to DW repository
* Clean data, fill NA values
* Remove duplicate records, Perform merging and splitting. Sort the records.
* De-normalize the extracted data according to dimension model of DW
* Convert to appropriate data types. Perform summarization

Transformation tasks that are commonly performed on extracted data are:

1. Format Revision: These include changes to data types and lengths of individual data fields
2. Decoding of fields: Sometimes same data items may be described by different field values.
3. Splitting of fields: Splitting name, address, etc. into smaller target fields
4. Character set Conversion: Convert textual data into standard character set limit
5. Conversion of Unit: Companies can have branches all around the globe, price must be displayed in different currencies, and other units depending on the region
6. Date and Time Conversion: Must be stored in standard format
7. De-Duplication: Remove duplicate customer information that might be stored in different files

**Explain the Operational Data Source (ODS) with respect to nature of data, underlying technology, profile records, classes.**

* Nature of Data: Very Limited amount of historic data (compared to DW that stores 5-10 years old data, ODS only has a month’s data)
* Underlying technology: Part of it is designed using relation technology and rest is designed using multidimensional technology.
* Profile records: Formed from many observations about one entity. Once information is captured in the Profile record, it can be easily and quickly accessed.

Classes of ODS:

* Class I: Takes milliseconds for data to arrive in ODS. Time is transparent to user. Example : Airline registration system
* Class II: Might take several hours. Example: Name/Address change
* Class III: Overnight gap or longer. Example: Applying Sales Transactions
* Class IV: Can take months or years. Example: A survey to monitor customer’s buying habits

**Write a note on Slowly Changing Dimensions (SCD)**

**Differentiate between OLAP and OLTP**

|  |  |
| --- | --- |
| **OLAP** | **OLTP** |
| Online Analytical Processing | Online transaction processing |
| Type of software tools that are used for data analysis for business decisions. | Provides transaction-oriented applications in a 3-tier architecture. |
| Consists of historical data from various Databases. | Consists only operational current data. |
| It reveals a snapshot of present business tasks. | It provides a multi-dimensional view of different business tasks. |
| Large amount of data is stored typically in TB, PB | The size of the data is relatively small as the historical data is archived. For ex MB, GB |
| Relatively slow, as the amount of data involved is large. Queries may take hours. | Very Fast as the queries operate on 5% of the data |
| It only need backup from time to time as compared to OLTP. | Backup and recovery process is maintained religiously |
| Only read and rarely write operation. | Both read and write operations. |
| Any type of Data warehouse system is an OLAP system |  |
| Example: Netflix movie recommendation system. | ATM center is an OLTP application |