SUBJECT: EXTRACT, TRANSFORM, LOAD (ETL) PROCESS {Notes for Semester Exam}

1. OLTP vs OLAP

OLTP	OLAP	Parameter
Online transactional system	Online Analytical Processing	Basic
Inserts, updates and deletes information from the database	Extracts data from this database for analysis and decision making	Focus/Job
Original source of data	Different OLTP's become source of data for OLAP	Data
Short transactions	Long transactions	Transaction
Lesser processing time	Longer processing time	Time
ATM	Sales report, marketing management, financial budgeting	Example

2. ROLAP vs MOLAP

ROLAP	MOLAP
Relational online analytical processing	Multidimensional online analytical processing
Data stored in the form of tables	Data stored in the form of multidimensional arrays
Data stored and fetched from main data warehouse	Data stored and fetched from proprietary database
Can handle very large volumes of data	Has limitations to amount of data it can handle
Limitations on complex analysis functions	Has large library of complex functions which can be used for analysis

3. Dimensional modelling vs ER modelling

Dimensional Model	ER model
Supports ad hoc queries for business analysts and complex analysis (DW and multidimensional database) - Supports OLAP	Supports OLTP and ODS (operational data source)
It is asymmetric	Symmetric – all tables look the same
Easy visualisation – the data cubes can be rotated to see different views of the data	It is a one dimension model
It permits redundancy	Removes the redundancy in the data
Denormalised	Normalised
It is extensible to accommodate new data elements without a change in its application	If the model is modified then even the application is modified

4. Roles and responsibilities with data quality framework

Roles	Responsibilities
Data consumer	End users who pose queries and do analysis and submit reports based on the data in the DW
Data producer	Charged with maintaining the quality of data input from source systems
Data expert	Identifies pollution in the source system
Data policy administrator	Responsible for resolving data corruption as data is transferred and moved into the data warehouse
Data consistency expert	Responsible for synchronising the data within the DW repository
Data correction authority	Apply data cleaning techniques
Data integrity specialist	Responsible to make sure that the data in source systems conform to the business rules

5. Explain extraction process in ETL

- Data flows from data sources and pauses at staging area.
- After transformation and integration, data is made ready for loading into the data warehouse.
- Effective data extraction strategies include:

1. Identifying the applications and systems from which the data is extracted.
2. For each identified data source, determine the way to extract data. Be it manually or by using tools.
3. Determine the extraction frequency.
4. Estimate the time window for the process for each data source.

- Different data extraction techniques are:
  1. Immediate Data extraction
  2. Deferred Data extraction
- Immediate Data Extraction:
  -> Data extraction is real time
  -> Occurs as transactions happen at source databases and files
  1. Capture through transactional logs (CTTL)
     -> Makes use of transaction logs of DBMS
     -> Reads transaction logs and selects all completed transactions
     -> Logging is already done
     -> No extra overhead incurred in operational system
  2. Capture through database triggers (CTDT)
     -> Triggers are inbuilt procedures
     -> Stored on database
     -> Fixed when certain predefined events occurs
     -> Triggers can be created for all events for which data needs to be captured
     -> Output is written on a separate file that will be used to extract data
  3. Capture in source application (CISA)
     -> Source application is used to capture data for data warehouse
     -> All relevant applications that write to source files are modified to write all adds, updates and deletes to both the source files and the database tables
- Deferred Data Extraction:
  -> Does not take place in real time
  -> Done at a later point of time
  1. Capture based on date and timestamp
     -> When a record is created or recorded, it is marked with a timestamp in the source system that will be used for selecting the record for data extraction
     -> Timestamp shows the date and time at which the source record was created
  2. Capture based on comparing files
     -> Also known as snapshot differential technique
     -> Compares two snapshot of the source data
     -> It forces the keeping of prior copies of all relevant data

➔ **DO THE DIAGRAM FOR THIS QUESTION.**

6. Explain the data transformation tasks in ETL

- Transformation process deals with rectifying the inconsistency
- Improving the quality of data becomes an important task within the data transformation process
- Takes the following course:
  1. Map the input data from source system to the data for data warehouse repository
  2. Clean the data – fill missing values with some default values
  3. Remove duplicates, perform merging and splitting of fields, sort the data
  4. De-normalize the extracted data according to the dimensional model of the data warehouse
  5. Convert to appropriate data types. Perform aggregations and summarisations.
- Transformation tasks that are often performed on extracted data are as follows:
  1. Format revision – Changes in data types and length of individual fields
  2. Decoding fields – When data comes from different sources, then same data may be stored by different field values
  3. Splitting fields – Entire name and address were stored in the same text field. Need to split according to first name, middle name, last name, flat number, building name, street name etc.
  4. Character set conversion – Done to textual data; converts the character set to an agreed standard character set

5. Conversion of units – Same company have global branches; sales represented in different currencies; need to convert to one common unit of measurement before moving the data into the data warehouse
6. Date and time conversion – Needs to be represented in a standard format
7. De-duplication – Customer data may be stored in different files; special attention needs to be given to these records; find such duplicates and remove them before storing them in data warehouse

7. Explain the role of metadata in ETL environment and describe the classification of metadata

- Metadata is data about data
- It is like data dictionary in database management system
- Used for building, maintaining, managing and using the data warehouse
- Provides users a roadmap about information in the data warehouse
- Functions of metadata:
    1. Acts as glue, connecting parts of data warehouse
    2. Provides information about the contents of the data to its users
    3. Enables users to search for data in their own business terms
- Without metadata, users will not understand the meaning of the data – where, how, why, what data exists within the organisation
- Classified into 3 main groups:
    1. Operational metadata:
        -> Data for DW comes from different operational systems which have different field lengths and data types
        -> In selecting data, it may either have to be split or combine certain parts of the record from different fields
        -> Operational metadata solves this problem by containing all the information about the data from operational data source
    2. Extraction and transformational metadata:
        -> Contains data about data extracted from source systems and various transformation techniques that were applied to that data before storing it in DW
        -> Maps every individual data element from its source system to DW
    3. End user metadata:
        -> Map of DW to enable users to find data in DW
        -> Uses a cryptic code which is an abbreviation for the actual description of the data field for users to understand. Ex: Cname is the cryptic code for Customer Name.

8. Framework of Multidimensional structure allows:

- Visualisation of DW schemas in terms of a multidimensional model which is used as a reference for querying
- Executing textual and graphical queries against available multidimensional schemas and views
- Specification of views
- Visualisation of result set of query execution
- Rectangles – software modules (for diagram)
- Cylinders – data repository (for diagram)
- **REFER DIAGRAM FOR THIS QUESTION**

9. Architecture of a data warehouse

- Data in data warehouse comes from the operational sources of the organisation and external sources, collectively called as the source systems
- This data is stored in the data staging area where this data is cleaned, transformed, combined etc. to prepare data for the warehouse
- The three different kind of systems required are:
    i. Source systems
    ii. Data staging area
    iii. Presentation servers
- Operational sources:
    i. Data comes from the mainframe systems of the organisations
    ii. Data is also supplied by relational DBMS like Oracle etc.

- Load manager:
    i. Performs all operations related to extraction and loading of data in data warehouse
    ii. Performs operations like simple transformations to get the data ready for entry into the data warehouse
- Warehouse manager:
    i. Performs operations related to management of data into the data warehouse
    ii. Performs operations like:
        1. Analysis of data to ensure data consistency
        2. Create indexes and vies on the base table
        3. Denormalisation
        4. Generation of aggregation
        5. Backing up and archiving of data
- Query manager:
    i. Performs operations related to management of user queries
    ii. Constructed using vendor end-user access tools and custom built programmes
- Detailed data:
    i. Stores detailed data in the database schema
    ii. Detailed data in regularly entered into the data warehouse to supplement the aggregated data
- Lightly and highly summarised data:
    i. Stores predefined L and H summarised data which is generated by the warehouse manager
    ii. Main function is to speed up the query performance
- Archive and Backup:
    i. The different kinds of data is stored for the purpose of archiving and backup
    ii. Transferred into storage archives as optical disks, magnetic tapes etc.
- Metadata:
    i. Extraction and loading process – maps data sources to a common view of information within the data warehouse
    ii. Warehouse management process – used to automate the production of summary tables
    iii. Query management process – used to direct query to the most appropriate data source
- End user access tools:
    i. Query and reporting tools
    ii. Application development tools
    iii. Data mining tools

- **DO DIAGRAM FOR THIS QUESTION**

10. Explain the backroom and front room of a data warehouse

- Front and back room are physically, logically and administratively different
- Raw data coming from source system is written into the disc with minimal restructuring
- Structured data from structured source systems is written into flat files and relational tables
- This allows the original extract to be simple, and fast and has greater flexibility to restart if the extract faces an interruption
- The data quality required by the source systems as compared to the data quality acceptable for data warehouse is different in most cases and hence, discrete steps like below are done:
    i. Checking for valid values
    ii. Ensuring consistency
    iii. Checking whether complex business rules and procedures have been followed
- Data conformation is required when 2 or more data sources are merged into the data warehouse
- Thus whole point of the backroom is to make the data ready for querying
- Final step of backroom is to physically structure the data into a set of schemas

- **DO DIAGRAM FOR THIS QUESTION**

11. Data velocity and cyclicity of data

- Data velocity:
    i. Speed with which the data passes from the initial capture to the point of use
    ii. Calculated as an average of the total time elapsed between the entry of data into the system and the point where the data is used by the user

iii. The main factor which affects the data velocity is the integration process – more the data to be integrated, lesser will be the velocity
iv. It includes the time needed for:
1. Editing the data
2. Extraction, transforming and loading the data
3. Passing the data into the appropriate application

- Data cyclicity:
  i. It is the time taken between a change in the data in the operational system and its reflection in the data warehouse
  ii. For example:
  1. If a customer name was edited on 12th feb 2017 at 2pm and the change was reflected in the data warehouse on 13th feb 2017 10am, then the cyclicity of the data is 20 hours.

12. What is data cleaning and explain the steps involved in data cleaning along with, mention reasons for dirty data.

- Data cleaning also known as data scrubbing
- Deals with detecting and removing errors, finding inconsistencies, and hence, improving the quality of the data
- Reasons for dirty data are as follows:
  i. Dummy values
  ii. Duplicate values
  iii. Absence of values
  iv. Inappropriate use of address lines
  v. Contradicting data
  vi. Reused primary keys
  vii. Data integration problems
- Steps in data cleaning are described as follows:
  i. Parsing
  1. Individual data elements are identified in the source systems and then isolated in the target files
  2. For example: name parsed into first, middle and last name
  ii. Correcting
  1. Individual data elements are corrected using data algorithms and secondary data sources
  iii. Standardising
  1. Data is transformed into a consistent format using standard or custom business rules
  iv. Matching
  1. Involves removal of duplicate data by going through parsed, corrected and standardised data
  2. For example: matching similar names and addresses
  v. Consolidating
  1. Involves merging of records into one representation by analysing and identifying the relationship between the matched records
  vi. Data staging
  1. Interim step between data extraction and the remaining steps
  2. Using different process like FTP sessions, flat files etc., data is accumulated from asynchronous sources
- After a certain predefined interval, data is loaded into the warehouse after the transformation process

13. Explain different mechanisms to manage flow of data from data warehouse to near line storage

- Manual transfer:
  i. The administrator manually moves the data from one medium to another
  ii. He places a DW monitor which keeps the track of the frequency of the usage of the data
  iii. The data which is not used frequently is moved from data warehouse to near line storage
  iv. Uses minimal technology
  v. Provides flexibility
- Hierarchical storage management (HSM):
  i. Free from manual interference

ii. Fully automated
iii. Moves the entire set of data from the data warehouse to the near line storage

- Cross-media storage management (CMSM):
  i. Fully automated
  ii. Works on the row level of granularity of data
  iii. If user poses a query, that requires storing the data in DW, then data is fetched by the system and proceeds with the execution
  iv. Otherwise if the data is present in the near line storage, then the system collects the data from there and then proceeds with the execution of the query
  v. DW monitor monitors what data is being used by the queries posed by the end users
  vi. Identifies the data which is not being frequently used (at a row/record level) by which the data can be more finely tuned with the data warehouse

14. Explain the various sources of pollution of data

- System conversion:
  i. System conversion and migrations are prominent reasons for data pollution
  ii. Conversion from flat files, to hierarchical database to relational database
- Data ageing:
  i. Older values loose their significance with time
  ii. For example: product code as a part of the products table is not needed anymore and hence, newer applications may want to remove it
- Heterogeneous system integration:
  i. Heterogeneous and dissimilar source systems lead to errors and inconsistency in data
  ii. For example: a table involving flat files, hierarchical and relational databases as its source system
- Poor database design:
  i. Database which cannot support the verification of the data which is entered and is not robust enough to handle large volumes of data can pollute the data
- Incomplete information at data entry:
  i. Some fields have zero's or null values leads to corruption of data
  ii. For example: phone number field may have 1's or 2's filled into them
- Frauds:
  i. Deliberate attempts to enter incorrect data
  ii. For example: fields representing the quantity of a particular product sold, may be incorrect
- Lack of policies:
  i. If there are no prevention rules made by the company to cater to entering of incorrect and invalid data, then there are high chances of pollution of data
  ii. For example: if there is no check to see whether or not the user entering the data is within the domain range, then the user may intentionally or unintentionally fill incorrect values

15. Explain security mechanism in the data warehouse environment

- User privileges:
  i. Prefers a ROLE based security
  ii. People with common requirements are grouped together and given certain privileges
  iii. Privileges can be assigned to an individual as well as a group
  iv. For example: john is an end user. Certain privileges are given to the end users under the same role, hence it being available to john as well. But if any extra privilege has been given to john, like to access a certain dimensional table, then only john can access it, and not the other end users of his group.
- Password protection:
  i. Users need password to access the data warehouse
  ii. Data administrator needs to set up acceptable patterns and the expiry periods of these passwords
  iii. A record needs to be maintained for wrong password attempts and after a certain threshold the user must be suspended temporarily from accessing the data warehouse till the time data administrator reinstates the user
- Security tools:
  i. The security provided by the DBMS is the primary security tool
  ii. Third party security systems are involved to govern the security of the data warehouse

16. Explain the concept of data granularity

- Granularity refers to the level of detailing of data; that is the level of summarisation of data
- In technical terms, the granularity of data is inversely proportional to the level of detailing in the data
- Basically, higher the level of detail, lesser is the granularity; and more the granularity, then lesser is the detail of the data
- For example:
    i. Lesser granularity – more detail
        1. Details about a specific phone call made by the customer in the particular month
    ii. More granularity – less detail
        1. Details about all the phone calls made by the customer in a particular month
- In data warehouse, data is kept at different levels of granularity; and the user can go to a specific level of detail according to the query posed and satisfy the query
- Lower the level of granularity means that more volumes of data needs to stored and vice versa
- The choice of granularity calls for a trade-off between the volume of data and the level of query detail

17. Features of a data warehouse

- Subject oriented:
    i. Data warehouse is designed in way to help analysts find out information from the data
    ii. For example:
        1. A data warehouse can be built which concentrates on transactions, loans etc.
        2. Thus, a query can be posed for finding the customer who was granted a loan of the largest amount or find the customer who was granted a home loan
        3. In this way, since the data warehouse can defined having the subject matter – loans – makes data warehouse subject oriented
- Non-volatile:
    i. It means that once the data has been stored in the data warehouse, it cannot be removed or changed as the purpose of the data warehouse is to analyse the data
- Integrated:
    i. A data warehouse is constructed using many heterogeneous source systems like flat files, relational databases, hierarchical databases etc.
    ii. The data collected is cleaned and then data integration techniques are used to ensure consistency in the data stored
- Time variant:
    i. Data warehouse also maintains historical data
    ii. For example: an employee data also contains the details of his past jobs (historical data)
    iii. Hence, all data in the data warehouse is identified with a particular time period

18. Write a note on loading in a temporal and non-temporal data mart

- Loading a temporal data mart:
    i. Complete refresh:
        1. Data marts are loaded by reloading the entire table every time
        2. This is done by truncating the table and then loading the data again
        3. Benefit – this technique captures everything present in transaction repository and reduces the back requirements
    ii. Cumulative refresh:
        1. In this technique, reloading is executed only for the facts from the transactional repository which has appeared since the last load process
        2. Speeds up the load process considerably for large data marts
- Loading a non-temporal data mart:
    i. Loading from transactional repository:
        1. When data is loaded from here, then we are directly going to the source system
        2. Here we build a current data mart without relying on the temporal data mart which has been loaded first
    ii. Loading from a temporal data mart:
        1. Most attractive method
        2. Takes full advantage of the work already done by the load of the other data mart

3. As long as the dependency on the load time for the temporal data mart is not an issue, this technique is always recommended

19. Explain the different levels of testing a data warehouse

- Unit testing:
  i. Also called as the white-box testing
  ii. Each development unit is tested on its own by the developer of that particular module
- Integration testing
  i. Different modules form a component in the data warehouse application
  ii. It is tested to ensure that they work properly together
- System and acceptance testing
  i. The entire data warehouse application is tested as a single unit
  ii. Acceptance testing – the users conduct their own tests on the system
- Performance testing:
  i. Most important as the system may satisfy the above tests but may fail in the performance level test in the end
  ii. Test whether the ETL process is completed within the load window
  iii. Check the time taken for updating and processing the reject records
  iv. Check the time taken to refresh standard reports
  v. Check the time taken to refresh complex reports

20. Describe the life cycle of Data Science

- Discovery:
  i. Collecting data from all kinds of internal and external sources
  ii. Can be in any form like files, sheets, etc.
- Data preparation:
  i. Raw data can have a lot of inconsistencies, missing values, abrupt values, incorrect values, blank columns etc. which need to be cleaned
  ii. Hence it is important to clean and process data prior to modelling
- Model planning:
  i. Here we determine the technique to draw the relationship between the variables
  ii. Use visualisation techniques like plotting histograms, boxplots, bar charts to get a fair idea of the distribution of data
- Model building:
  i. Divide the data into training and testing sets
  ii. Analyse various techniques like classification, clustering, association etc. to build a model
- Operationalize:
  i. Make the final report, technical document, briefing code etc.
- Communicate result:
  i. Identify all the key findings and present it to the stakeholders or board members of the organisation and determine whether the results were a success or a failure

21. Define data warehouse and the benefits of the data warehouse

- Potential high returns on investment:
  i. It delivers enhanced business intelligence
  ii. Implementation of DW requires huge investment but it helps the company make the right strategic decision which is very useful in their marketing and sales segment
- Competitive advantage:
  i. Unknown and unavailable data is present in DW thus helping the analysts to take the right decisions for the organisation and have a competitive advantage
- Saves time:
  i. As data in DW is present in integrated format from multiple sources, there is no need to retrieve data from other sources
- Better enterprise intelligence:
  i. Improves customer service and productivity
- High quality data:
  i. Data in DW is cleaned and transferred in a specific/desired format, so data quality is high

22. List the functions of a data quality tool

- Error discovery feature
  i. Identify duplicate record
  ii. Identify values that are outside the domain range
  iii. Find inconsistent data
  iv. Monitor trends in data quality over time
  v. Report to the users about the data quality
- Error correcting feature
  i. Normalise inconsistent data
  ii. Improve the merging of data from dis-similar data sources
  iii. Prevent entry of data values which are outside the domain range
  iv. Provide measurements of data quality

23. Write a detailed note on SCD (slowly changing dimensions)

- Term associated with managing issues arising from changing an attribute of a dimension table
- Has 3 types:
  i. TYPE 1 – overwrite a dimension record
  ii. TYPE 2 – add a new dimension record
  iii. TYPE 3 – create new field in the dimension record
- TYPE 1:
  i. Correction of error
  ii. Doesn't require to store the incorrect value (discarded)
  iii. No other changes are made in the dimension record
  iv. Key of the dimension record is not affected
  v. For example: incorrect name "John Michel" changed to correct name "John Michael"
- TYPE 2
  i. Preservation of historical values
  ii. A new dimension record with the changed attribute is added with its own new key
  iii. No change in the key of the original record
  iv. Identity (ID) of the individual remains same
  v. For example: marital status of "John Michael" has changed from single to married on 26th march 2006, so orders before 26th march 2006 should have the marital status as "single" and after that date, orders should have marital status as "married"
- TYPE 3
  i. Tentative soft revisions
  ii. New field is added in the same dimension record
  iii. Used to compare performance across transition
  iv. Used when history needs to be tracked with both old and new value of the same attribute
  v. An "old" field is added in the dimension table where the existing value is pushed into from the current field
  vi. For example: addition of the field "old location". Change of location from delhi to Mumbai so delhi is pushed into the old location field and Mumbai is written into the current location field
- **DO DIAGRAM FOR THIS QUESTION**

24. Explain operational data store (ODS)

- Nature of data:
  i. ODS contains very limited amount of historical data as compared to the DW
  ii. In contrast with DW which contains 5-10 years of historical data, ODS contains barely 1-2 months of historical data
- Underlying technology:
  i. Hybrid approach - where a part is designed using relational technology and the rest is designed using multidimensional technology
- Profile records:
  i. Formed from many observations about the entity
  ii. Creates a summary of the multiple observations/occurrences of the data
  iii. Captures massive amount of data in a very concise manner

iv. Once the information is captured in the profile records, it is easily and quickly accessed whenever required

- Classes of ODS:
  i. Depends on how fast the data arrives into the ODS
  ii. CLASS 1
    1. Takes a few milliseconds
    2. Time elapsed is transparent to the user
    3. For example: airline reservation system

  iii. CLASS 2
    1. Takes several hours to arrive in the ODS
    2. Users can see the significant time gap between the occurrence of transaction and arrival of the same
    3. For example: updating name and address change of the customer
  iv. CLASS 3
    1. Takes overnight or longer for the data to arrive
    2. For example: used in sales transaction
  v. CLASS 4
    1. Time gap between the occurrence of transaction and its arrival into the ODS is in few months or even a year or so
    2. Source of data can be a DW or external
    3. Maybe created from the output of special reports or projects/analysis
    4. For example: survey of customer buying habits

- **DO DIAGRAM FOR THIS QUESTION**

25. Write a note on ODBC in ETL process

- Enables users to access databases on their windows applications
- Original intention was to make the application portable
- So if the underlying database changed from DB2 to oracle, then the application layer need not be recorded or compiled to accommodate the change
- Simply change the ODBC driver which is transparent to the applications
- ODBC can also be used to access flat files
- ODBC manager:
  i. Maintains connection between the application and the ODBC driver
  ii. Accepts SQL from the ETL application and routes it to the appropriate ODBC driver
- ODBC driver:
  i. Real workhouse of the ODBC environment
  ii. Translates the ODBC SQL to native SQL of the underlying database
  iii. Disadvantage:
    1. Comes at performance cost
    2. Adds several layers of processing and passing of data
    3. For the ETL process which uses data via ODBC, 2 new layers are added between the application and the underlying database

- **DO DIAGRAM FOR THIS QUESTION**