Answers the following questions :-

1. **Define ETL and explain its importance in data management.**

   Ans : ETL [Extract Transform Load] – It has 3 major functions

   i) Extracting the data from a particular source.
   ii) Transforming the data into particular formats necessary structured way.
   iii) Loading the data onto a specified target.

   Using ETL tools helps a lot in data management, as it combines data and does analysis with respect to its format and gives the output based on the insights. Which makes a lot of work in a simple and faster way.

2. **Describe a scenario where ETL could be beneficial in a business setting.**

   Ans : Using ETL in business could be a very smart way to compress the load work. As it has many functions and capabilities to arrange , analyze the data to get the information what we need.

3. **What challenges might a data analyst face during the transformation phase of ETL and how can they be addressed?**

   Ans : During Transformation , a data analyst might face some challenges regarding the errors of computing the

data , such as some of the missing values , choosing wrong file extension , not correctly mapping one to another ,etc.

Those challenges can be addressed by filling the missing values, implementing data validation using most of the ETL tools , with respect to the file extensions.

4. Explain the concept of data warehousing and its relationship with ETL processes.

Ans : Data warehouse is a system that aggregates data from different sources into a single, central, consistent data store to support data analysis, data mining, AI and ML.

Whereas, ETL processes helps to make sure that the data is clean ,integrate and structured to store in Data warehouse from multiple sources.

5. Define a database and a data warehouse.

Ans : Database :

A place where data is been stored.

Data warehouse :

It is a relational database which is deigned for query and analysis rather than transactional processing.

6. How do the purposes of a database and a data warehouse differ in a business environment?

Ans : In business perspective, data base is not so prominent because it only stores the data , irrespective of order and used for OLTP(Online Transaction Processing).

Whereas, Data warehouse collects the data and stores in an order for some of the analysis and used for OLAP(Online Analytical Processing)

7. **Can you illustrate with an example when you would use a database versus a data warehouse?**

Ans : For example, if we consider a retail shop,

Database is used for managing Daily Transactions and Operations such as Sales Transactions, managing the information such as current stock items.

Whereas, Data warehouse is used to improve the Decision Making and getting accurate data . It is helpful analyzing performances of such as Sales Analysis, also extracts meaningfull insights.

8. **List 5 Popular Data Warehouse, ETL Tools and Database.**

Ans : Data Warehouse : Big query, PostgreSQL, DB2, Snowflake , Redshift,etc.

ETL Tools : Pentaho , IBM Datastoragsqle, Stitch, Oracle, Talend,etc.

Database: Oracle, MySQL, Microsoft SQL, Microsoft SQL Server, DB2 , PostgreSQL, etc.

9. . <u>Who is Data Analyst, Business Analyst and Data scientist?</u>

Ans : Data Analyst :

A person who analyses the data precisely and provides needful insights.

Business Analyst:

A person who focuses more on understanding and improving the business processes.

Data Scientist:

A person who uses analytical, programming as well as statistical skills in an advanced way. The person also creates models , algorithms and also predict future .

10. <u>llustrate with an example how data visualization can assist in business decision-making .</u>

<u>Ans : Let us consider a retail shop which uses data visualization for it's business decision making by using some tools Power BI, Excel ,etc. to create visual representation of the collected data. It also may be used to plot some charts lie bar chart, line chart, pie chart,etc</u>

<u>For better allocation marketing and strategic planning sales period.</u>

11. Practical

I am going to explain about which dataset I have chosen and what are the transformations I did to the datasetUsing Pentaho (which is a freely available software to use many of the ETL tools).

Step 1:

Downloading a dataset from Kaggle website make sure that you would organize those files in a specific folder , wher you can access easily , in my case I have downloaded a dataset containing the full information about different cars.

I didn't find the dataset again after downloading it from the website, so I have uploaded in the drive and shared here below.

https://drive.google.com/drive/folders/1ADqvdfrR5JTclPtiH8OryfTh82d0nElT?usp=drive_link

Step 2:

In Pentaho, A dataset should be taken as input initially , it should be done with respect to its extension also. Accessing the right file with right extension plays a major role, so be careful in mentioning / browsing the path and give a name.

In my case both the datasets were in the form of text , so I took two different csv file input in the design section, because I had two datasets.

Step 3:

To sort a particular dataset by rows, we use sort rows , which we take the input thorugh pipeline and sort it in ascending order .

I mapped each dataset to a different sort rows through connecting them via piplines, weher pipelines shpw the flow data. It is very important to map the pipelines very well.

Step 4:

In order to merge two datasets and sort them together , we use sorted merge, generally it can sort any number of datasets into a single dataset.

I mapped both the sorted datasets to the sorted merge, in order to get every data sorted in ascending in one file.

Step 5:

If we want to remove some duplicate data , we use unique rows, which helps to display unique data and remove duplicate ones.

I had mapped the sorted merge to unique rows , and also I had selected the company column to get only unique car company details.

Step 6:

If we want to remove some rows containing specific values , we use filter rows .

I mapped the unique rows to filter rows to remove only Electric cars and CVT in the column Transmission Type, so that I can view all the cars except only electric and cvt, by adding the logical conditions like(AND,OR and NOT) and selecting the column name and entering the values we want to filter out is the major part.

Step 7:

To view particular sub data consisting of minimum columns we want , we use select values. So that we can able to get the only columns we want.

I mapped filter rows to select values, to get company ,model, horsepower, torque, transmission type, drivetrain, fuel economy, number of doors, price, model year range , engine type. So iam getting every data which I want except body type and number of cylinders which I didn't select and write yes.

Step 8:

Finally , if we want that particular data to be stored as a text file , we use text file output.

I mapped select values to text file output, to get the data which is been required by me after the transformation will be stored in the folder where I specified it's path.

So , This is the end of transformation what I did to the datasets .

My transformation is shared below:

https://drive.google.com/drive/folders/1zHoHm9NilUbRVXyiuMMzBZbrBMY4-QtM?usp=drive_link

Screenshot of My final completed transformation by running It: