1. **What is a surrogate key? How do you generate surrogate keys in the ETL process?**

   A surrogate key is a unique row identifier on type 2 dimension.
   It is required for identifying unique rows on a type 2 dimension table.
   Type 2 dimensions are built to maintain a history of the changes occurring on dimension attributes. Each natural key can have duplicate entries on type 2 dimension, as the attribute values keep on changing. In such cases, a new row is inserted with new attribute values, and the surrogate key is incremented on the new row.

   A surrogate key is a random number that is generated by incrementing the previous maximum value of the surrogate key on the dimension table.

2. **What are the possible error scenarios in the DW load process?**
   - Referential validation failures can be caused by missing dimension IDs. Fact tables are connected to a dimension table by a foreign key. If a dimension ID in a fact table is not present in the DImension table, then the load to the fact table will fail with a referential constraint error.
   - Late arriving dimensions will cause referential integrity failures.
   - Data quality rejections - NULL values on PKs or any key columns

3. **What are the validation and data cleansing rules that you can apply on string fields, integer fields and date fields?**
- String fields should be trimmed of extra spaces, non-ascii characters.
- Integer fields should be rounded to specific decimal positions as defined in the target.
- Date fields should be converted into proper date and datetime formats and time zone formats.

4. **Give me the SQL functions for converting the Unix datetime field into MM/DD/YYYY?**
   The Unix time is the number of seconds that have elapsed since the Unix epoch, that is, the time 00:00:00 UTC on 1 January 1970.

   Select

   date_format(to_char(unix_to_date(1588996715),'DD-MM-YYYY HH:MM:SS'),'%d-%m-%Y)

5. **How do you improve the performance of the DW load process?**
   -Perform referential integrity validations before loading in the ETL platform.
   Referential integrity checks will take time during the loading process.
   -Generate the surrogate key in the ETL platform.
   The performance of the load process can be improved substantially by reducing the effort of surrogate key generation away from the database into the ETL platform.
   Write commits on the target table should be done in bulk to save time.

6. **What is data partitioning? And, what is the use of partitioning data on processing platforms?**

   Data partitioning is a data processing technique used for increasing parallelism. ETL platforms store and process huge volumes of data. Transforming these huge volumes of data into target structure involves heavy denormalisation and aggregation logic.

These operations can be done faster by processing data in parallel threads.

1. Pipeline parallelism:
   In pipeline parallelism, multiple tasks are performed in an ETL flow in parallel.

2. Data parallelism:
   In data parallelism, data is stored in multiple chunks and processed in parallel. After all the parallel threads are completed, all their results can be collected to get the final result.

7. What is the difference between Data Lakes and Data Warehouses?

Just like a data warehouse, a data lake is a centralized repository that stores data collected from different data sources in the organizational ecosystem; the only difference is that it stores both structured and unstructured data from these data sources. Data can be dumped into this repository in its raw format without any preprocessing.

Both data warehouses and data lakes are used by an organisation for different purposes. A data warehouse is mainly used for reporting various financial data. Data lakes, on the other hand, are used as data repositories by organisations, where they can load their entire data, whether it is structured, semi-structured or unstructured. For reporting, the data from data lakes can be transformed and loaded into data warehouses in a structured format.

8. What is the difference between ETL and ELT?

The process of ETL extracts the data, transforms the entire data and then loads the data into data warehouses. In ELT process, the data is extracted from data sources and is first loaded into a central repository. Whenever the data is required for analysis, the required data is transformed and cleaned.

Data transformation and loading process is faster in ELT than ETL; however, the analysis speed is faster in ETL than ELT. This is because ETL loads the structured data into a central repository, which can be directly used for analysis. On the other hand, in the case of ELT, the data is first transformed when analysis has to be done.