SQL Data Engineer Interview

Q&A

Q1. What Is the difference between using WHERE versus HAVING?

WHERE: is used to filter based on condition(s) before the aggregation takes place. HAVING: specifies a search condition for a group or an aggregate function used in SELECT statement. When GROUP BY is not used, HAVING behaves like a WHERE clause.

Because WHERE clause is evaluated before groups are formed, it evaluates for per row, while HAVING clause is used in the column operation and is applied to aggregate rows or groups according to given conditions.

Because HAVING clause is processed after the rows have been grouped, it can contain aggregate functions like sum(), count(), while WHERE will give an error.

Q2: If I want to track historical changes for a given dimensional data, what type of Slowly Changing Dimension should we be implementing

Type 2 is the most common method of tracking change in data

warehouses.It allows to maintain the history for the entire record and you can easily perform change-over-time analysis.

Q3. Explain the difference between an inner join and outer join using an example.

I created two tables: Table 1 and Table 2 and joined them: inner and left outer join. See:

https://nbviewer.jupyter.org/github/kyramichel/SQL/blob/main/SQL%20Language1.ipynb

Explanation: JOIN clauses are used to combine rows from tables based on a related column between them.

Inner Join returns results by combining rows from the tables (here 2 rows).

The (left) outer join returns all rows for the table1. When a match isn't found, then a NULL is placed. (here 3 rows)

Q4. DISTINCT vs GROUP BY: when would you use each + what is the commonality

DISTINCT is used when you just want to remove duplicates. GROUP BY implicitly does a DISTINCT over the values of the column you're grouping by. GROUP BY lets you use aggregate functions, like SUM(), COUNT(), AVG(). There is the possibility for subtle differences in their execution too.

GROUP BY removes duplicates without actually selecting the field.. Using DISTINCT with multiple columns, the result set won't be

grouped as it does with GROUP BY.

Q5. What is Primary Key and Foreign key in a table (answer to this question exposes data modeling expertise)

A primary key ensureS data in the specific column is unique.
A foreign key is a column(s) that provides a link between data in 2 tables.

While there is only 1 primary key allowed in a table, there could be more than 1 foreign keys in a table.

A primary key does not allow NULL values, while a foreign key can contain NULLs. A primary key is a combination of UNIQUE and Not Null constraints.

A primary key value cannot be deleted from the parent table.