Database performance testing in an ETL context

In previous lessons, you learned about database optimization as part of the database building process. But it's also an important consideration when it comes to ensuring your ETL and pipeline processes are functioning properly. In this reading, you are going to return to database performance testing in a new context: ETL processes.

How database performance affects your pipeline

Database performance is the rate that a database system is able to provide information to users. Optimizing how quickly the database can perform tasks for users helps your team get what they need from the system and draw insights from the data that much faster.

Your database systems are a key part of your ETL pipeline– these include where the data in your pipeline comes from and where it goes. The ETL or pipeline is a user itself, making requests of the database that it has to fulfill while managing the load of other users and transactions. So database performance is not just key to making sure the database itself can manage your organization's needs– it's also important for the automated BI tools you set up to interact with the database.

Key factors in performance testing

Earlier, you learned about some database performance considerations you can check for when a database starts slowing down. Here is a quick checklist of those considerations:

- Queries need to be optimized
- The database needs to be fully indexed
- Data should be defragmented
- There must be enough CPU and memory for the system to process requests

You also learned about the five factors of database performance: workload, throughput, resources, optimization, and contention. These factors all influence how well a database is performing, and it can be part of a BI professional's job to monitor these factors and make improvements to the system as needed.

These general performance tests are really important– that's how you know your database can handle data requests for your organization without any problems! But when it comes to database performance testing while considering your ETL process, there is another important check you should make: testing the table, column, row counts, and Query Execution Plan.

Testing the row and table counts allows you to make sure that the data count matches between the target and source databases. If there are any mismatches, that could mean that there is a potential bug within the ETL system. A bug in the system could cause crashes or errors in the data, so checking the number of tables, columns, and rows of the data in the destination database against the source data can be a useful way to prevent that.

Key takeaways

As a BI professional, you need to know that your database can meet your organization’s needs. Performance testing is a key part of the process. Not only is performance testing useful during database building itself, but it’s also important for ensuring that your pipelines are working properly as well. Remembering to include performance testing as a way to check your pipelines will help you maintain the automated processes that make data accessible to users!