Exercise - work with approximate execution

5 minutes

Approximate execution using HyperLogLog functions

As Tailwind Traders starts to work with large data sets, they struggle with slow running queries that typically run quickly. For instance, obtaining a distinct count of all customers in the early stages of data exploration slows down the process. How can they speed up these queries?

You decide to use approximate execution using HyperLogLog accuracy to reduce query latency in exchange for a small reduction in accuracy. This tradeoff works for Tailwind Trader's situation where they just need to get a feel for the data.

To understand their requirements, let's first execute a distinct count over the large Sale_Heap table to find the count of distinct customers.

In the query window, replace the script with the following code:

```sql
SELECT COUNT(DISTINCT CustomerId) from wwi_perf.Sale_Heap
```

1. Select Run from the toolbar menu to execute the SQL command.

The query takes up to 20 seconds to execute. That is expected, since distinct counts are one of the most difficult to optimize types of queries.

The result should be 1,000,000.

2. In the query window, replace the script with the following to use the HyperLogLog approach:

SQL

```
SELECT APPROX_COUNT_DISTINCT(CustomerId) from wwi_perf.Sale_Heap
```

3. Select Run from the toolbar menu to execute the SQL command.

The query takes about half the time to execute. The result isn't quite the same, for example, it may be 1,001,619.

APPROX_COUNT_DISTINCT returns a result with a 2% accuracy of true cardinality on average.

This means, if COUNT (DISTINCT) returns 1,000,000, HyperLogLog will return a value in the range of 999,736 to 1,016,234.