Copy activity performance slowness for parquet file source

Last updated by | Veena Pachauri | Mar 8, 2023 at 11:34 PM PST

When copying the parquet files from ADLS Gen2 to azure mysqldatabase, the pipeline execution is slow

2019年11月13日
10:50 AM

When loading the data from ADLS Gen2 to azure mysqldatabase, the pipeline execution is slow, here the file format at source side is parquet. Customer configured ADF pipelines with high configuration for both IR and for sink(mysqldatabase) DB, still the performance is not good.

Troubleshooting steps:

- Suggesting customer to increase DIU's and degree of copy parallelism and also requesting them to increase the cores to 256 in integration runtime will not work.
- In the **Copy scenario**: From file store to non-file store. **Parallel copy behavior**: When copying data into Azure SQL Database or Azure Cosmos DB, default parallel copy also depend on the sink tier (number of DTUs/RUs). When copying data into Azure Table, default parallel copy is 4. Please refer the link https://docs.microsoft.com/en-us/azure/data-factory/copy-activity-performance-features
- Even though we increase the DIU's to 256, based on this copy activity(file-store to non-file store) it will work as single thread and default runs with 4 DIU's.

Performance improvement recommendations:

- Since the source file format is .parquet, by-design, one Parquet file for one thread. No easy way to scale out.
- It's always good to recommend customers to split the files into many smaller parquet files, they would get better degree of parallelism, meaning better throughput.

Best practices to follow when migrating the large databases from onprem to Azure databases

- Please follow the below link as a best practice when planning to migrate large databases from onprem to Azure Databases for MYSQL. Please refer the link https://docs.microsoft.com/en-us/azure/data-factory/copy-activity-performance-features

How good have you found this content?