USE Case 1 : Data Transformation / ETL

Database Load (ETL)

Description: ETL (Extract, Transform, Load) is the ability to load raw data from disk into a database to make it available for direct use by users or for subsequent processing (aggregating, combining, etc.).

Benchmark Results

Cloud-Edge loaded 2.1 Tb of raw data[2] from disk in 35 minutes and 39 seconds. This works out to an average load of 1.067 Gb/second. Note that the I/O bandwidth available for the raw data was virtually identical to our load throughput. That is, the Cloud-Edge engine running on the DL980 loaded data as fast as the I/O could provide it, only utilizing the CPUs at less than one-third of their capacity.

We believe this benchmark compares favorably with the Syncsort and Vertica ETL World Record, which was able to load 5.4 Tb in 57 minutes 21.51 seconds. While this performance is impressive, it was done using sixteen (16) nodes and two (2) HP BladeSystem c7000 enclosures. As noted above, we are confident Cloud-Edge can improve on this record on a single node given additional I/O bandwidth for the raw data.

Cloud-Edge Advantage

The Cloud-Edge ETL process is extensively parallelized (multi-threaded), enabling it to make efficient use of both the available I/O as well as the CPUs/cores for each node instance.

Additional features of Cloud-Edge ETL include:

•The ability to apply complex filtering to the raw input data.

•Optional "rejected records file" to store info for input records that did not pass applied filters.

•An unlimited number of raw input files can be combined to create a single Cloud-Edge database table.

•Unlimited number of calculated columns can be created on import, based on the raw data file.

•Field parsing of input columns, allowing multiple output columns to be created from a single raw input field.

•The Cloud-Edge .Net Data Provider (with custom Cloud-Edge extensions for SSIS) allows SSIS packages to be quickly developed and deployed, which fully leverage the power and speed of the ETL engine.

Business Advantage

Cloud-Edge's ability to quickly load data together with the power of the DL980 makes it possible to combine technologies on a single machine. For example, an OLTP database can be used to process transactions, which can then be quickly loaded into Cloud-Edge for subsequent analysis and processing. Quick ETL speeds are especially critical in business systems with a high requirement on fast availability of analytic data. Cloud-Edge leadership in ETL speeds enable business to meet the most demanding targets for data availability and to do so with fewer nodes of investment and maintenance.

[2] Four (4) scale factor 500 instances of the TPC-H data were used for this test. Each instance was created using the dbgen utility onto a separate RAID5 LUN on an HP P2000 Storage Area Network (SAN) connected to the DL980 via two (2) Brocade 8/24 switches.