

Aster Data® nCluster®: Compression

Organizations are seeing a phenomenal increase in data. This data growth is being driven by new-generation applications and devices, online user activities that generate data at very high rates, and government regulations that force organizations to retain data for longer periods. While such high-volume data offers the potential to analyze and extract rich business insights, managing it is a significant challenge.

Aster Data nCluster delivers the first analytic platform, a massively parallel (MPP) row and column database with an integrated analytics engine. It is the first MPP database that is a hybrid row and column store and allows applications to be fully embedded within the database engine to enable ultra-fast, deep analysis of massive data sets. Aster Data nCluster provides a powerful suite of compression capabilities that not only help reduce storage costs, but also improve system performance across both a row and column store:

- **Reduction in Storage Costs and Overheads** – The size of data in a data warehouse has a direct impact on storage costs across all environments (production, test, development, backup, etc.). A larger storage infrastructure also results in much higher data center overheads such as cooling, floor space, etc. nCluster provides the capability to compress data and reduce these costs.
- **Improvement in Performance** – Compression of table data improves disk I/O performance as effectively more data can be read per I/O. For I/O bound workloads that are very common in data warehousing applications, this provides performance improvements. In addition, nCluster provides network compression capabilities that significantly enhance query performance. Similarly, compressed backups require less data to be transferred over the network, improving backup as well as recovery performance.

Multi-Level Table Compression

nCluster provides the ability to compress table data to minimize the storage footprint for primary partitions as well as replica partitions used for high availability. The administrator can choose from three levels of compression: HIGH, MEDIUM, and LOW. A higher level of compression provides larger reduction in size and would usually require more CPU cycles. Table compression is completely transparent to applications—no changes are required on the application side.

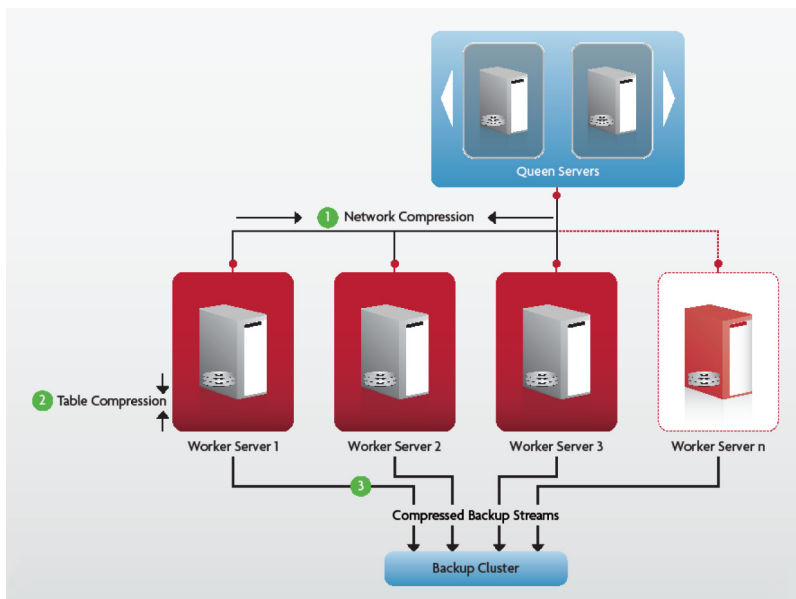


Figure 1: Aster nCluster Compression.

Quick Overview

Aster Data nCluster Compression provides a powerful set of data compression capabilities that not only reduce storage costs, but also improve performance.

Highlights

- Multi-Level Table Compression reduces storage costs and provides information lifecycle management (ILM) capabilities
- Backup Compression reduces backup size and provides fast network transfer during backup and recovery
- Network Compression for intra-cluster data transport makes queries run fast

The availability of different compression levels provides flexibility to choose the right level of compression, according to the balance required between storage savings and the CPU overhead of compression. It should be noted that *nCluster* leverages all process ore cores on every server to compress data in a massively parallel way—the CPU overhead is shared across a very large number of processor cores. Thus, the overall throughput achieved during compression is very high. For I/O bound workloads, the I/O benefits more than compensate for CPU overhead and provide overall performance benefits.

nCluster can typically compress data by a factor varying from 3X to 12X, depending on the nature of data and the compression level. Higher compression ratios are also possible depending on the data distribution.

Compression can be enabled selectively at a table level as well as a table partition (child table) level. *nCluster* physically partitions the data across servers, and administrators can create further table partitions by creating child tables. *nCluster* allows for an arbitrarily deep hierarchy of parent-child relationships, which provides a very flexible data partitioning scheme. Compression can be independently enabled at any level in the hierarchy. For example, an *Orders* table may have child tables for each quarter. In such a case, a child table *Orders_Q1_2005* may have compression set to HIGH while another child table *Orders_Q3_2008* may have it set at MEDIUM. This granular control over compression scope enables information lifecycle management (ILM) capabilities within *nCluster*—infrequently accessed data can be compressed at higher compression levels while frequently accessed data can be stored using lower levels.

Backup Compression

nCluster allows data to be compressed during the online backup process, resulting in a smaller backup size. Data is compressed on the Worker servers of *nCluster* and then streamed to the destination. *nCluster* Backup Compression does not require materialization of data on the Worker servers, eliminating any need for disk space overhead for the process of compression.

Compression at the Worker servers not only reduces storage needs, it also makes backup data transfer over the network more efficient. Similarly during restores, data is transferred over the network in the compressed format and decompressed by the Worker servers, providing faster recovery. For backups too, compression ratios are heavily data dependent and can typically be in the range of 3X to 12X for different tables, depending on the nature of data.

Network Compression

Limited bandwidth of inter-server networks has traditionally been the limiting factor for performance and scalability of massively parallel processing (MPP) databases. Queries that require data from different servers to be analyzed together by operations such as joins and aggregations often result in shuffling large amounts of data across servers, impacting system performance. *nCluster* includes patent-pending algorithms that address the network bottleneck issue. It provides intelligent data placement and dual-stage query planning capabilities that minimize inter-server data transfer. When such transfer is necessary, it compresses the data on the source server and decompresses it on the destination server so the network is efficiently used. Given that network speeds are lower than the data compression/decompression speeds of processors, network compression provides significant improvement in overall system performance.

About Aster Data

Aster Data is a market leader in data management and advanced analytics for diverse and big data, enabling the powerful combination of cost-effective storage and ultra-fast analysis of relational and non-relational data. Aster Data *nCluster* is an analytic platform that incorporates a massively parallel processing (MPP) hybrid row and column database with an integrated analytics engine, allowing application logic to execute with data to deliver breakthrough performance and scalability. Aster Data's solution utilizes Aster Data's patent-pending SQL-MapReduce® to parallelize processing of data and applications and deliver rich analytic insights at scale. Companies including Barnes & Noble, Intuit, LinkedIn, Akamai, Full Tilt Poker, and MySpace use Aster Data to deliver applications such as deep clickstream analysis, recommendation and personalization analytics, real-time fraud detection, and churn analysis.