Paper BI14-2014

SAS® In-Memory Analytics: Beyond Foundation SAS

Tho Nguyen, Teradata Corporation, Raleigh, North Carolina

John Cunningham, Teradata Corporation, Danville, California

ABSTRACT

In-memory analytics help to analyze large volumes of data in-memory with SAS® High-Performance Analytics (HPA) products and SAS® Visual Analytics. This latest innovation provides an entirely new approach to tackle big data by using an in-memory analytics engine to deliver super-fast responses to complex analytical problems. It is a set of products beyond SAS Foundation technologies to explore and develop data models using all of your data. Jointly developed with SAS, the Teradata Appliance for SAS High-Performance Analytics, Model 720 eliminates the need to copy data to a separate appliance with dedicated SAS nodes for in-memory processing. This paper explains the in-memory analytics process and how the Teradata Appliance for SAS extends the analytic capabilities of the Teradata environment.

INTRODUCTION

Data is a strategic asset and organizations are collecting more data than ever before. It is presenting great opportunities but also challenges to analyze ALL of the data in a timely manner. In this volatile and competitive economy, there has never been a bigger need for proactive and agile strategies to overcome these challenges by applying the analytics directly to the data rather than moving data around. In addition, trends in analytics and data management, along with heightened regulatory and governance burdens, demand new, innovative approach that can quickly transform massive volumes of data into meaningful insights.

SAS and Teradata are addressing these challenges by taking advantage of the parallel processing capabilities of the Teradata database and SAS® in-memory analytics technologies. SAS in-memory processing excels at analyzing big data using parallel algorithms to visualize and develop data models. Data models that used to take weeks and days to process can now be executed in hours and minutes, transforming data into value for sustaining competitive advantage.

This paper will cover the following topics:

- In-memory analytics process
- Leveraging in-memory analytics with Teradata
- Teradata Appliance for SAS® High-Performance Analytics

IN-MEMORY ANALYTICS PROCESS

The SAS in-memory environment leverages Teradata's MPP (Massively Parallel Processing) architecture which is ideal for retaining, preparing and partitioning large data sets for big data analytics. It is capable of high data consumption rates through parallelized data movement which means completing any task at a fraction of the time. This latest innovation provides an entirely new approach to tackle big data by using an in-memory analytics engine to deliver super-fast responses to complex analytical problems. It is a set of products beyond SAS Foundation technologies to explore and develop data models using all of your data.

The SAS Foundation software is located on a user's workstation or on a SAS server. When it runs a SAS program containing High-Performance procedures or analytics, it initially connects to the Teradata database containing the source data, and then it instigates a parallel computing job on the SAS processing nodes. One of the SAS nodes is designated to be the controlling root node and the other nodes are worker nodes.

The SAS client coordinates with the root node, and the root node in turn directs with the corresponding processes on the worker nodes. The worker processes are multi-threaded to take advantage of the large number of CPUs. Therefore, once an in-memory analytics process runs on the appliance, all of the nodes are dedicated to that specific task. Analysis can be executed in minutes or seconds using this approach.

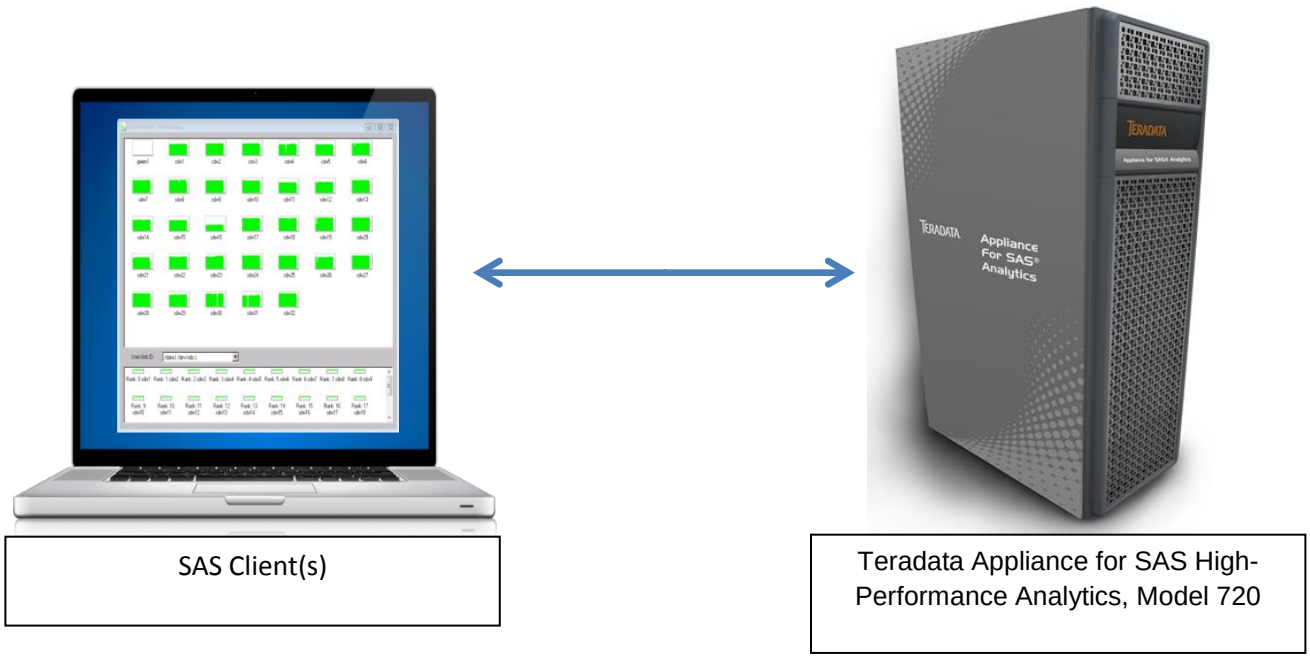

Figure 1: In-memory processing

When all of the processes are running for an in-memory task, the root node submits a SQL query to Teradata that causes the SAS Embedded Process (EP) table function to read data from the database and send it to a SAS in-memory worker. The Teradata database is multi-threaded and utilizes parallel processing in MPP environments. It can take each SQL request and splits into thousands of subsets, and have each database worker thread process the subsets in parallel, passing to the parallel processing of the in-memory appliance, Model 720. This parallelized process is extremely fast, fully leverages the optimized BYNET interconnect fabric, likely >50-100x faster than other traditional data access models.

After the data is transferred to memory and while the SAS in-memory job is active, there is no activity in the Teradata database. Thus, there is no performance impact to the Teradata database as data is only lifted into memory when requested. SAS software coordinates the analytical processing between the SAS client that is running the procedure, the SAS HPA root node, and the SAS worker nodes. All of the nodes in the Teradata Appliance for SAS are designated to compute the analytical tasks.

When the SAS HPA in-memory processing is complete, results can be written back to Teradata into a permanent client for additional analysis, depending on the type of procedure and the procedure options that are selected.

TERADATA APPLIANCE FOR SAS HIGH-PERFORMANCE ANALYTICS, MODEL 720

The Teradata® Appliance for SAS High-Performance Analytics, Model 720 supports SAS High-Performance Analytics Products and SAS® Visual Analytics, integrating SAS in-memory capabilities with the industry leading data warehouse platform, for data model development and data visualization. Jointly developed with SAS, the Teradata Appliance for SAS High-Performance Analytics, Model 720 eliminates the need to copy data to a separate appliance with dedicated SAS nodes for in-memory processing.

There are a number of SAS products that seamlessly integrate with the Model 720.

- **SAS Visual Analytics** - Explore massive volumes of data quickly to visualize and uncover patterns and trends for further analysis
- **SAS High-Performance Analytics Products**
  - **SAS High-Performance Statistics**: Enables use of predictive models for faster and more effective decision-making.
  - **SAS High-Performance Data Mining**: Develops predictive models using thousands of variables to produce more accurate and timely insights.
  - **SAS High-Performance Text Mining**: Explores all your data, including textual information, to gain rich new knowledge from previously unknown themes and connections.
  - **SAS High-Performance Forecasting**: Generates models for faster high-value and time-sensitive decision making, using thousands or even millions of granular-level forecasts.
  - **SAS High-Performance Econometrics**: Provides econometric modeling facility, such as the number and severity of events, using big data.
  - **SAS High-Performance Optimization**: Performs more frequent modeling iterations and uses sophisticated analytics to get answers to questions you never had time to ask.

By leveraging analytical features, including statistics, data mining, text mining, forecasting econometrics and optimization, organizations can quickly identify and add important variables. More data model iterations can be performed to gain understanding and make decisions with confidence.

The Teradata Appliance for SAS High-Performance Analytics readily extends the entire Teradata Platform Family as shown in Figure 2, providing ultra-high speed SAS In-Memory Analytics against Teradata Data Warehouses. The appliance features clustered servers, each with dual Intel® eight core Sandy Bridge processors, SUSE® Linux operating system, 128-256GB of RAM, and enterprise class Infiniband networking infrastructure—into a power-efficient system. The appliance connects directly to Teradata BYNET, ensuring unsurpassed data access speeds, 50-250x faster than traditional ODBC, and superior analytic processing. Best of all, the solution is supported by the most trusted name in data warehousing—Teradata.

Figure 2: Teradata Platform Family Connects with Model 720

The Teradata Appliance for SAS High-Performance Analytics, Model 720 enables advanced analytics with incredibly fast parallel processing, scalability to process massive volumes of data, and rich in-memory analytics capabilities. This environment provides a set of in-memory analytics algorithms that

leverages the database's speed, while eliminating time-consuming and costly data analysis. This Teradata appliance includes analytical capabilities spanning data visualization and data model development executed in a highly scalable, in-memory processing architecture. It will let customers explore massive volumes of data with SAS Visual Analytics and develop analytical models using complete data—not just a subset—with SAS High-Performance Analytics products to get accurate and timely insights and make well-informed decisions. Often faced with hundreds of candidate variables, this offering helps to determine unimportant variables, describe important relationships, and identify the important factors for subsequent models and data exploration.

The Teradata Appliance for SAS High-Performance Analytics is easy to manage, you can free up your DBA resources to do other valuable tasks. With virtualized CPU, memory, and storage all designed to work together as a unit, you get automated management of physical disk space so your DBAs never have to worry about data placement or data reorganization.

With this Teradata appliance, companies can start with a small configuration, and then expand as needed driven by the ongoing analytic needs of the business.

CONCLUSION

In-memory processing is a new and innovative approach to analyzing large volumes of data in an era of big data analytics. By leveraging the power of the Teradata data warehouse and MPP architecture, SAS complex computations can be executed in-memory. With a dedicated appliance designed for SAS In-memory analytics, SAS users can

- Ask and answer new business questions in near-real time and provide new information for the daily operations.
- Improve productivity and expedite decision making for competitive advantage.
- Increase orders-of-magnitude performance gains delivered by leveraging Teradata Database's Massively Parallel Processing architecture and executing SAS in-memory analytics.

RECOMMENDED READING

www.sas.com/teradata

www.teradata.com/sas

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: John Cunningham
Enterprise: Teradata
E-mail: john.cunningham@teradata.com
Web: www.teradata.com/sas

Name: Tho Nguyen
Enterprise: Teradata
E-mail: tho.nguyen@teradata.com