THE POWER TO KNOW®

In-Memory Analytics for Big Data

Game-changing technology for faster, better insights

WHITE PAPER

Table of Contents

Introduction: A New Breed of Analytics

In today's era of big data, organizations depend on increasingly sophisticated analysis of ever-growing volumes and varieties of data. They count on having reliable access to massive volumes of data, and they are performing advanced analytics that traditional relational technology is unsuited for and even incapable of. While other vendors are recycling relational and OLAP technology from 20 years ago and marketing it as new and improved, SAS has taken a different approach.

SAS® LASR™ Analytic Server is the world's first system specifically engineered to address diverse advanced analytic use cases. It is a "read-mostly," stateless, distributed in-memory server that provides secure, multiuser, concurrent access to data in a distributed computing environment.

Built upon core design principles set forth by Jim Goodnight, CEO of SAS, SAS LASR Analytic Server is a direct-access, NoSQL, NoMDX server that is engineered for maximum analytic performance through multithreading and distributed computing. Built on industry-standard hardware and distributed parallel architectures for big data, SAS LASR Analytic Server is a new breed of distributed in-memory analytic architecture for the next generation of high-performance analytics.

SAS LASR Analytic Server's unmatched performance and scalability provide the ability to answer questions that previously were computationally infeasible. It enables users to explore data and analyze a variety of big-data problems, including areas such as risk management, customer intelligence, revenue optimization, and assortment and merchandise planning.

SAS® In-Memory Overview

SAS provides a distributed in-memory computing environment configured for the demands of interactive, advanced analytic workloads. It is specifically engineered to address the computational complexity of large-scale exploratory data analysis and visualization as well as predictive analytics and data mining.

Performing near-instantaneous query and reporting on sums and counts of billions of records has become one of the most common demonstrations from some vendors of their high-performance analytic capabilities. What you won't find is other vendors doing this on hundreds or thousands of variables with the types of distribution analysis and predictive modeling required by today's more sophisticated analytic community.

For example, take something as simple as a box plot. It is a convenient way of graphically depicting groups of numerical data through summary calculations – minimum, maximum, upper and lower quartile, and median. Box plots display different populations without making assumptions about the underlying distribution (see Figure 1).

When it comes to the most common descriptive statistics calculations, SQL-based solutions have a number of limitations, including column limits, storage constraints and limited data type support. In addition, the iterative nature of exploratory data analysis and data mining operations, such as variable selection, dimension reduction, visualization, complex analytic data transformations and model training, require multiple concurrent passes through the data – operations that SQL and relational technology are not well-suited for. By creating an in-memory engine that is designed to speed the tasks of data exploration, predictive modeling, forecasting and optimization, SAS bypasses these issues that relational technologies face.

Figure 1: SAS LASR Analytic Server enables you to produce box plots based on huge numbers of variables in just seconds.

Take the next example (Figure 2). This is a simple heat map overlaid with a regression model. Most vendors would send data back to the front-end reporting tools to serially perform complex calculations. But when huge amounts of computations are needed to analyze and produce information, bottlenecks can occur. SAS' in-memory technology performs the calculations on the server – on the fly – and in parallel. As a result, computations are very fast because you are not moving large amounts of data elsewhere for processing, and you have many computational blades to take advantage of. With SAS, processing can take place on the analytic server with the thin results sent back to the client for presentation, rather than for computation.

Figure 2: A correlation map overlaid with a regression model that uses millions of variables can be produced very quickly using SAS in-memory technology.

SAS® In-Memory Architectures

SAS LASR Analytic Server is new thin-layer technology that enables SAS to run within distributed computing environments such as Hadoop, or alongside distributed relational databases such as Teradata and Greenplum. It provides applications with the responsiveness and extremely high throughput required by large analytic workloads and analytic-intensive applications. Applications access the SAS LASR Analytic Server using direct SAS connections and standard interfaces. The following is an overview of the two SAS architecture options.

SAS® LASR™ Analytic Server Running Within Hadoop

It is easy to get caught up in the big data craze by focusing on data measured in the hundreds of terabytes and tens of petabytes. In the past, it would have cost millions of dollars to store even a few terabytes of data. However, Hadoop has changed that game. Hadoop can aid in the storage of the data as well as in exploring and analyzing it by allowing businesses to deploy distributed applications running on thousands of nodes and sifting through petabytes of data. Now, instead of using specialized hardware and software to scale applications, Hadoop enables you to use industry-standard commodity hardware and open-source software. For SAS solutions, Hadoop provides an open, simple and robust architecture that addresses the needs for fault tolerance, redundancy and scalability.

Within each Hadoop node, a thin-layer SAS process takes incoming SAS commands and returns results (see Figure 3). It's that simple. You don't need to install SAS on every node because the Hadoop system takes care of distributing the processing, managing memory, controlling the job and managing the workload.

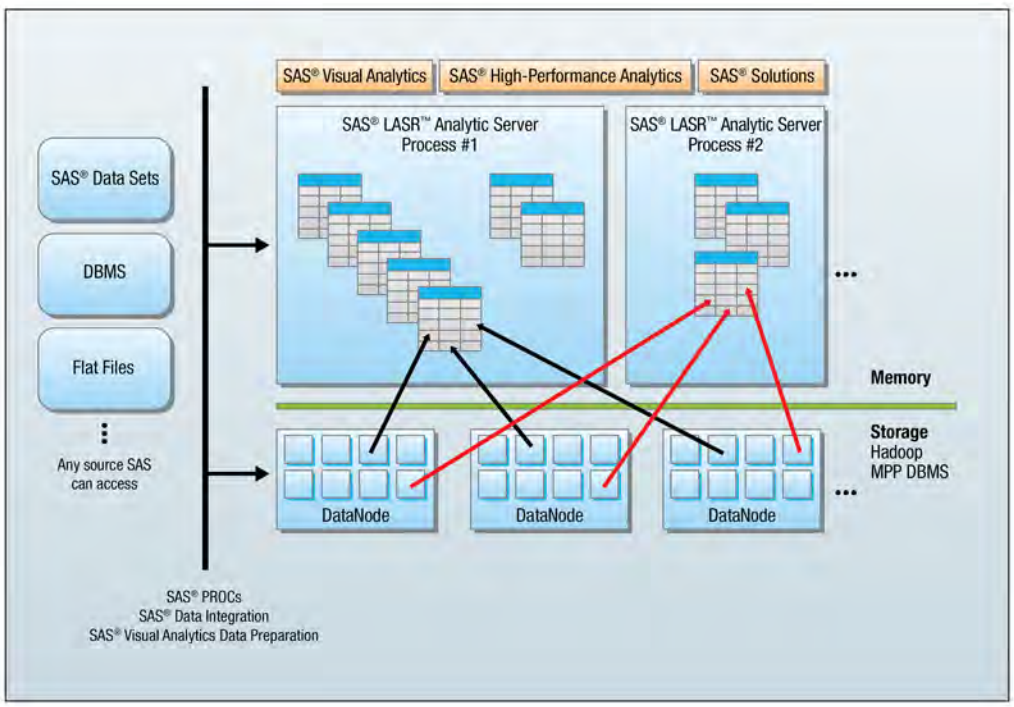


Figure 3: SAS provides two options for loading data into the SAS LASR Analytic Server. Depending on your needs and the size of your data, you can either take data from the Hadoop storage layer and load it into the server (as shown in the bottom portion of the diagram) or you can bypass the Hadoop storage layer and load data directly into the SAS LASR Analytic Server (as shown on the left side of the diagram).

Drilling a little deeper (Figure 4), you can see that SAS LASR Analytic Server operates directly on the Hadoop Distributed File System (HDFS), bypassing traditional MapReduce tasks. MapReduce is a Hadoop-based method for performing data manipulation and simple summary analysis tasks in the Hadoop environment. However, it lacks important internode communication capabilities that are critical for the high-end analytic work performed with SAS. In addition, the MapReduce paradigm does not provide the in-memory environment that is critical for analytical tasks that make multiple passes through data. On the other hand, SAS LASR Analytic Server optimizes data homogeneity in-memory for fast, steady processing even under concurrent, multiple-user access.

Note that SAS/ACCESS® Interface to Hadoop and new capabilities in Base SAS and the SAS Data Integration Studio component of SAS Enterprise Data Integration Server do use Hadoop's Hive and Pig languages along with MapReduce to access data stored in HDFS, thus providing several options for integrating SAS with Hadoop.

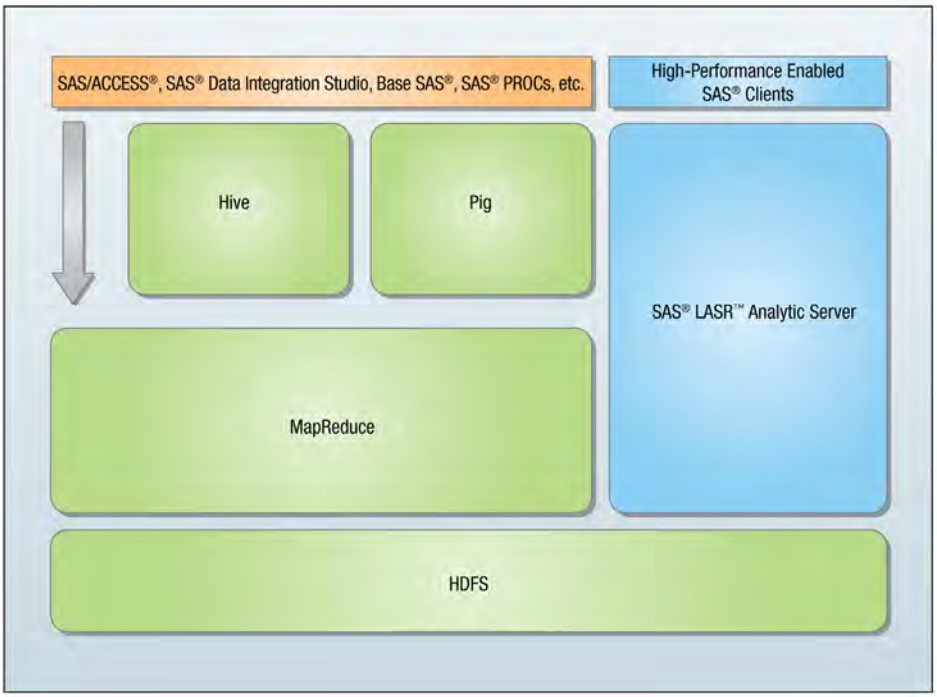


Figure 4: SAS provides two options for accessing and operating on data stored in Hadoop's HDFS, which is the primary storage system used by Hadoop applications.

SAS® LASR™ Analytic Server Running Alongside Greenplum or Teradata

Data warehousing, first popularized in the mid-1990s, is now a mainstream technology that has moved from back-office, query-and-reporting style analysis to become an aid in operational decision making. The hardware architecture and enterprise-class features of Teradata and Greenplum data warehouse appliances make them excellent vehicles to deliver SAS in-memory technology.

During the last several years, SAS has worked closely with RDBMS partners such as Teradata and Greenplum so that SAS can now run alongside their distributed computing architecture, bringing SAS processing to the data rather than bringing data to the SAS process. Massively parallel shared-nothing relational databases, such as Teradata and Greenplum, are good choices for SAS in-memory technology. They provide the data distribution, process management and memory management needed to run analytics alongside relational processing.

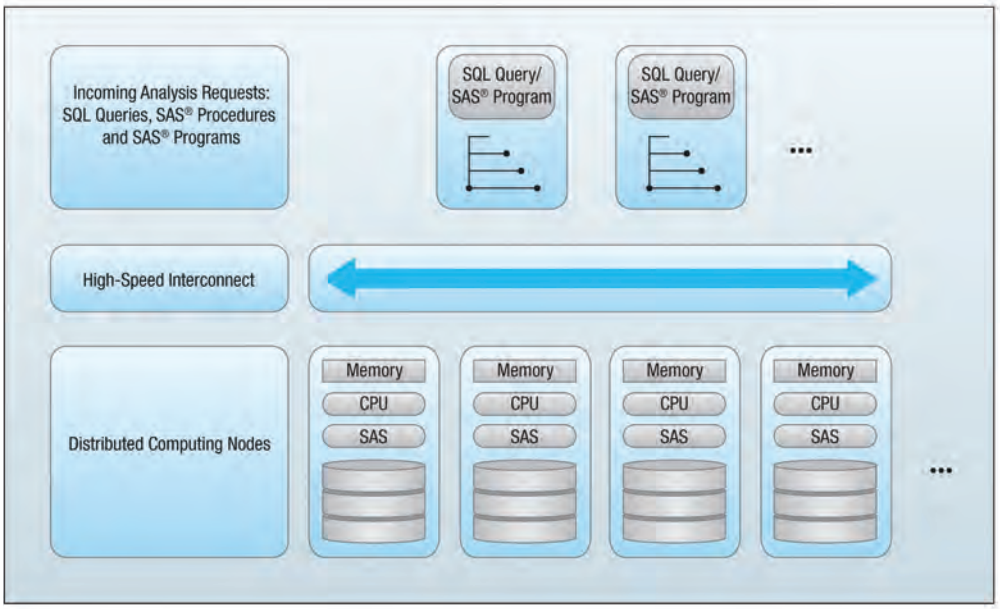


Figure 5: Greenplum and Teradata, like most MPP data warehouse appliances, use a shared-nothing MPP architecture. On each node, SAS' thin-layer, in-memory technology is used to communicate with distributed data.

Conclusion: Boosted Performance for Faster, More Precise Insights

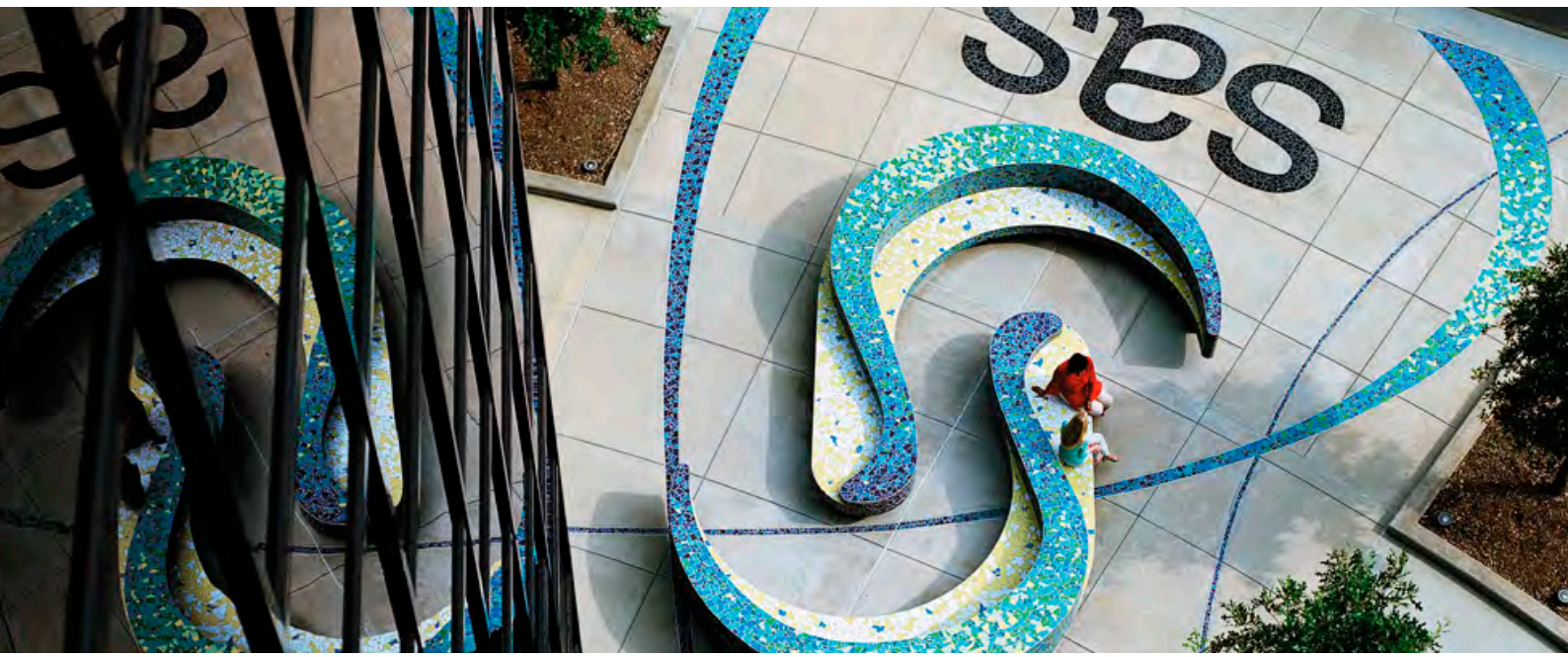
Data-intensive business applications use large, wide and deep data sets. They can be both memory-intensive and CPU-intensive, while requiring little or no persistence or having big-data persistence needs.

Traditional relational technology is not suited for the varied computing demands of real-time advanced analytics, but SAS LASR Analytic Server provides a new distributed, in-memory analytic computing environment specifically engineered for the next generation of real-time advanced analytics. It enables users to explore data and analyze a variety of big-data problems, including areas such as risk management, customer intelligence, revenue optimization, and assortment and merchandise planning.

Built using industry-standard hardware, market-leading advanced analytics software and in-memory technology, SAS provides an optimized system that delivers highly precise answers to all your business questions with unmatched speed, intelligence and simplicity while providing enterprise-class manageability and security.

About SAS

SAS is the leader in business analytics software and services, and the largest independent vendor in the business intelligence market. Through innovative solutions, SAS helps customers at more than 55,000 sites improve performance and deliver value by making better decisions faster. Since 1976, SAS has been giving customers around the world THE POWER TO KNOW®. For more information on SAS® Business Analytics software and services, visit **sas.com**.



SAS THE POWER TO KNOW®

SAS Institute Inc. World Headquarters +1 919 677 8000

To contact your local SAS office, please visit: **sas.com/offices**