The Bloor Group

SAS and The Hadoop Ecosystem

VENDOR PROFILE

Our research indicates that, at the current rate of adoption, Hadoop and its ecosystem will become dominant in the area of analytics and BI applications for the enterprise environment. This could constitute as much as one third of enterprise IT in the near future. Despite the extensive and growing stack of capabilities in the Hadoop ecosystem, there are still significant gaps. Software companies like SAS are filling many of those gaps and dramatically improving its overall functionality for data management, data visualization and analytics activities. This paper will discuss the qualities of SAS software and how it can enhance a Hadoop implementation.

The SAS Big Data Architecture

The SAS brand is inextricably tied to analytics. Incorporated in 1976, SAS has long been an industry leader in the analytics and data management market. Due to the current trend toward Hadoop platforms dominating the realm of analytics and BI, SAS and Hadoop would seem like competitors. However, SAS has evolved instead to be a powerful complement to the Hadoop ecosystem. In recent years, SAS has heavily modified their products or launched new ones to augment the Hadoop ecosystem and expand the reach of SAS customers into the rich data sets that reside in Hadoop clusters.

To their suite of analytics products, SAS has added parallelized algorithms and several techniques to accommodate cluster or distributed computing needs. SAS has embraced the philosophy of minimizing data movement by pushing the SAS compute engine out onto the Hadoop cluster. The new technology reads data on the cluster into an in-memory store on the cluster itself, where the SAS compute engine can then do multiple parallel operations on it.

Leading-edge parallel algorithms have been developed to work against this in-memory data set on the cluster. The in-memory structure allows the algorithms to operate on the data at impressive speeds and return results in very short time frames. Modern data visualization and machine learning algorithms have been developed, including data mining, predictive analysis, text mining, forecasting and optimization.

The SAS suite of Hadoop-related products includes SAS Data Management offerings, SAS In-Memory Statistics, SAS Visual Analytics and SAS Visual Statistics and the suite of SAS High-Performance Analytics products (High-Performance Statistics, High-Performance Data Mining High-Performance Text Mining, High-Performance Econometrics and High-Performance Optimization). These sets of products were created specifically to meet the different needs of data scientists, statisticians and business analysts working in Hadoop environments. SAS In-Memory Statistics offers an interactive programming environment, wherein multiple users can explore data, design data preparation workflows, design analytic models, test them iteratively directly on the cluster, compare them, score them and deploy them when complete. SAS Visual Analytics and SAS Visual Statistics together offer an interactive, drag and drop environment for visual data discovery, interactive reporting and predictive analytics. The SAS High-Performance Analytics products enable distributed processing for sophisticated analytics on distributed data in Hadoop for Statistics, Data Mining and Machine Learning, Text Mining, Optimization and Econometrics. They can be accessed through an interactive programming interface and also are tightly integrated with graphical analytical workbenches, such as SAS Enterprise Miner and the new SAS Factory Miner.

At no point in the process is data removed from the Hadoop cluster. In fact, if it is advantageous to combine other data sets, such as EDW data sets, SAS operators pull that data onto the Hadoop cluster so that the SAS data processes can make use of the cluster's computational power and massive storage capacity.

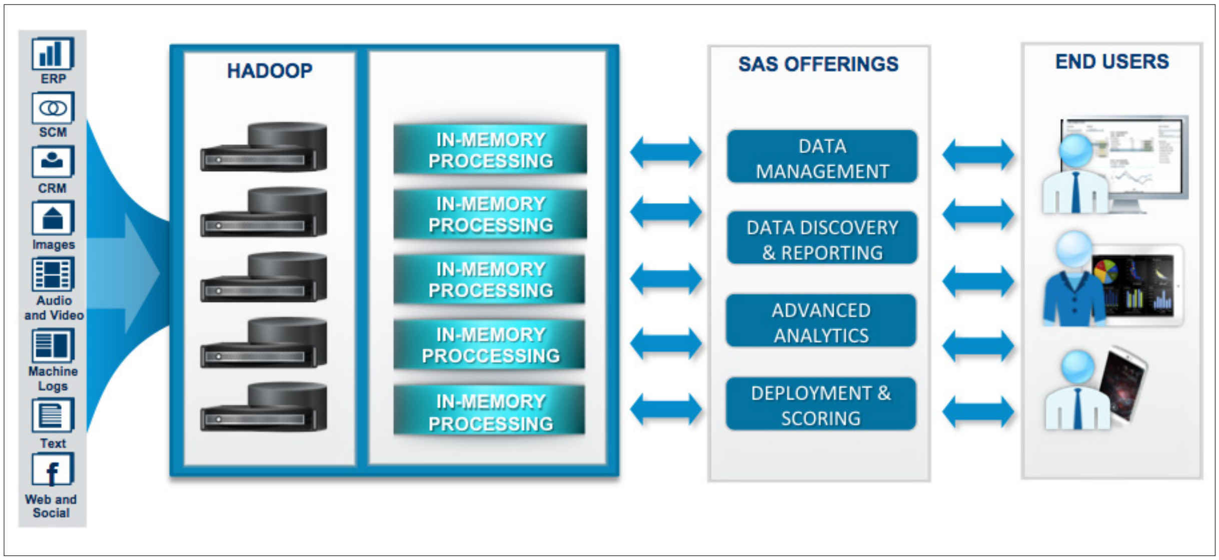

Figure 1. SAS In-Memory Analytics and Data Management for Hadoop in Overview

In our view, the four main advantages that SAS provides in combination with Hadoop are:

1. Sophisticated parallel optimized analytic algorithms
2. In-memory processing of data for machine learning and data integration
3. Data management and data quality processing for all data in Hadoop and associated systems
4. Interactive visualization and exploration of data both in Hadoop and in combination with other systems, such as data warehouses

SAS Addresses Specific Weaknesses in Hadoop

SAS analytics and data management tools use the cluster storage, interact with the Hadoop Distributed File System, Hive and/or Impala and HCatalog and use the cluster's computational power. It does not require any specific Hadoop distribution, or even require Hadoop at all, but it is designed to function well either beside a Hadoop cluster or within a Hadoop cluster, as illustrated in Figure 1. Still, one might logically ask why it would make sense to use SAS with Hadoop.

Several weaknesses in the Hadoop platform have been identified as its adoption rate has increased. These weaknesses have stalled Hadoop projects or prevented Hadoop adoption in many cases. SAS has specifically sought to address these weaknesses.

The shortage of skilled MapReduce coders in the current marketplace is well known. SAS addresses this problem with graphical drag-and-drop interfaces that allow the definition of data preparation and analytics workflows. These graphical workflows can be designed by non-programmers and can use the MapReduce framework to profile, prepare, transform, and cleanse data in parallel across the cluster.

SAS also addresses the shortage of Hadoop skills by providing a wide range of pre-built analytic, data quality and data preparation procedures. SAS has over the last years and continues to engineer these procedures to take full advantage of massively parallel Hadoop environments. Even the sophisticated SAS machine learning algorithms run smoothly in Hadoop, which addresses the shortage in the Hadoop ecosystem of mature, capable, parallel algorithms.

MapReduce is very batch oriented, and in many ways, not appropriate for iterative, multi-step analytics algorithms. In particular, its strict paradigm of doing a shuffle and write to disk between each step in a process would cause multiple intermediate files to be created. This is highly inefficient. By pulling the Hadoop data into an in-memory format, SAS In-Memory Statistics and SAS Visual Analytics, for example, provide algorithms that can apply to multiple steps without touching disk. This vastly increases the productivity of data scientists and business analysts.

One of the difficulties associated with the Hadoop data lake architecture is gaining an initial understanding of the content, combinations and potential correlations of all the many types of data stored there. SAS provides an interactive environment for analytic exploration and visualization. The SAS interface allows both visual and SQL-style interactive querying of the data without any requirement to write code. The interface generates Hive QL, a native querying language for Hadoop. While no coding is required, power users may enter Hive QL directly if they wish.

The shortcomings of Hadoop in the areas of data security and management are also well known. SAS has a federation capability that helps mitigate this weakness. By creating a virtual data layer, role-based access, data masking and many other security measures that can be implemented between the data and the users. This layer can also be a virtual integration environment that combines data from the Hadoop cluster and other data sets, such as the data warehouse.

This federation capability also simplifies and augments the Hadoop interactive experience by abstracting away much of the complexity of the Hadoop data environment. SAS has a well-established user community that can now use this familiar environment to leverage the power of Hadoop. The wide variety of Hadoop data sets can become simply another data source for SAS. This, too, helps address the shortage of Hadoop-related skills.

One inherent weakness of the Hadoop data lake concept is that data is often stored without regard to its usefulness or quality. The old adage of "garbage in, garbage out" still holds true in the modern world of massive, widely-varied data sources. Several of the Hadoop specific operators in SAS are designed for fast parallel data access and for performing data profiling, data quality and data integration tasks directly on the Hadoop cluster. This provides quality, vetted data for analytics, improving the eventual accuracy of the analyses done on that data.

While addressing these weaknesses of Hadoop, SAS has also sensibly exploited the fundamental strengths of the Hadoop platform. SAS pushes the analytic processing to the data rather than trying to move the data elsewhere for processing. When other data sets are needed, copies of that data are pulled into the Hadoop cluster to take advantage of both the Hadoop cluster's massive storage capacity and the cluster's sheer parallel compute power. Additionally, to meet customer demands, SAS has aligned its Hadoop strategy and roadmap to support variety of Hadoop distribution partners like Cloudera, Hortonworks, MapR, IBM BigInsights, and Pivotal.

SAS and Hadoop Use Cases

The intersection of capabilities provided by the combination of the Hadoop ecosystem and SAS offerings lends itself logically to specific business use cases. This software is ideal for analytics intensive workloads, not just data or compute intensive workloads. It is well-suited to the data lake architectural concept, meaning "a repository for data which supports and supplements a data warehouse." Hadoop alone is not capable of providing the optimized performance associated with a data warehouse. It requires complementary data warehouse and sophisticated analytics capabilities alongside it.

In cooperation with the Hadoop platform, SAS provides several unique capabilities. With the two platforms operating in tandem, users have the ability to leverage diverse data sets and evaluate multiple analytic scenarios to zero in on the best fit for the job at hand. Situations where we see this combination of SAS and Hadoop software as particularly advantageous include:

- Analytic sandbox implementations, where new data may be combined with existing data – possibly data warehouse data – then understood and analyzed to see if it yields useful new insights. The SAS software is particularly well-suited to support in an easy, integrated and comprehensive way the many different activities that data scientists and business analysts have to apply to extract actionable insights from data in Hadoop – ranging from data wrangling and data cleansing, to discovery and exploratory data analysis, the application of machine learning for model development and the deployment of models for operationalized analytics – all directly in Hadoop environments.
- Self-service large scale business intelligence type implementations, where non-programmers need to explore large and diverse data sets via interactive querying in a visual environment. This technology can also be used to offload many BI workloads from overworked data warehouses.
- Active archive types of implementations. In these cases, data that may no longer be fresh, but may still contain long term pattern insights, can be archived away from the expensive storage of the data warehouse and be accessible for long term analysis.

These are the general architectural use cases that we believe this combination of software platforms is well-suited for. These general use cases can be put to work to solve a variety of business problems, including insurance underwriting, risk mitigation, fraud detection, customer behavior analytics, location based marketing, cyber security, recommender systems, bandwidth allocation, network quality analysis and a wide variety of other business problems.

Organizations who are already invested in SAS – or organizations looking to add a robust Hadoop component – would do well to consider the SAS suite of Hadoop-related solutions.

To learn more about SAS and the research used in this paper, please visit www.sas.com/bloorreport.

About The Bloor Group

The Bloor Group is a consulting, research and technology analysis firm that focuses on open research and the use of modern media to gather knowledge and disseminate it to IT users. Visit both www.TheBloorGroup.com and www.InsideAnalysis.com for more information. The Bloor Group is the sole copyright holder of this publication.

Austin, TX 78720 | 512-524–368