Bringing the Power of SAS® to Hadoop
Combine SAS® World-Class Analytic Strength with Hadoop’s Low-Cost, High-Performance Data Storage and Processing to Get Better Answers, Faster
Table of Contents

Introduction ................................................................. 1
Chief Merits of Hadoop .................................................... 2
How Are Early Adopters Using Hadoop? ................................. 3
How Organizations Are Using Hadoop: The Technical Perspective ... 3
  1) Staging Area for Data Warehouse and Analytics Store ............. 4
  2) Initial Discovery and Analysis ........................................ 4
  3) Storage and Analysis of Unstructured and Semistructured Content .... 4
  4) Making Total Data Available for Analysis ............................ 4
  5) Low-Cost Storage of Large Data Volumes ............................ 4
How Organizations Are Using Hadoop: The Business Perspective .... 5
What Is SAS Doing with Hadoop? ......................................... 5
  Bringing the Power of SAS® Analytics to Hadoop with Easy Data Access ...... 6
  Benefits of SAS/ACCESS® Interface to Hadoop™ ...................... 6
  Technical Highlights of SAS/ACCESS® Integration ................... 6
  Taking Advantage of Hadoop’s Distributed Processing ................. 6
  Benefits of Base SAS® Integration with Hadoop Distributed Processing .... 7
  Technical Highlights of Base SAS® Integration with Hadoop ........... 7
  Bringing Governance to Hadoop Implementation with SAS® Information Management ....... 7
  Benefits of SAS® Information Management for Hadoop ............... 8
  Technical Highlights of SAS® Information Management for Hadoop .... 8
Closing Thoughts ............................................................ 8
Introduction

In 2000, a gigabyte of storage cost about $16. A terabyte data warehouse was rare. The largest data warehouse held about 100 terabytes of data. Today, you can buy a gigabyte of storage for about six cents and a terabyte for less than $100. Data stores of 100TB are almost passé. In three years, terabyte could be the new megabyte.

The plummeting cost of storage created tantalizing opportunities to do more with masses of data that had previously been unused – and to use enormous data volumes to deliver more time-sensitive insights. However, there was still the issue of processor cost. A 1TB, massively parallel processing (MPP) appliance could run $100,000 to $200,000. The total cost for an implementation could be millions. In contrast, a terabyte of processing capacity on a cluster of commodity servers can be had for $2,000 to $5,000.

For certain types of high-volume data and complex analyses, wouldn’t it be great to have the option to distribute processing across hundreds or thousands of low-cost servers?

That is the value proposition of Apache Hadoop. A project of the Apache Software Foundation, Hadoop is a software framework for running applications on large clusters of commodity hardware. It was inspired by technical papers and development emerging from Google, Yahoo and a growing community of contributors.

Hadoop provides a distributed file system (HDFS) that can store data across hundreds or thousands of servers and a computational approach called MapReduce, which divides an application into smaller components of work and distributes it across those machines, processing the work near the data.

The largest clusters in production as of early 2012 contain about 4,000 nodes with approximately 15 petabytes of storage in each cluster. Yahoo runs thousands of clusters with more than 42,000 Hadoop nodes, storing more than 200PB of data. With the next-generation MapReduce algorithm, Hadoop clusters can scale to between 6,000 and 10,000 nodes and handle more than 100,000 concurrent tasks and 10,000 concurrent jobs.

In a performance test, a 1,400-node cluster sorted a terabyte of data in 62 seconds; a 3,400-node cluster sorted 100TB in 173 minutes. To put it in context, 1TB contains 2,000 hours of CD-quality music; 10TB could store the entire Library of Congress print collection.

You get the idea. Hadoop handles big data. It does it fast. As part of a big data strategy, it redefines the possible when it comes to analyzing large volumes of data, especially semistructured and unstructured data (text).
Companies such as Cloudera, Hortonworks and MapR, founded by Hadoop contributors, provide validated software builds and enterprise support contracts for organizations that aren’t comfortable with unsupported open-source software. So it’s no surprise that many US Fortune 500® companies have either completed or are planning a project involving Hadoop.

Chief Merits of Hadoop

There are several reasons to embrace Hadoop for certain applications:

• **Hadoop has a large community of supporters.** In 2009, Yahoo made available the source code for its production version of Hadoop, which allows for innovation by partners and customers. Yahoo developers continue to improve the software, and then release the patch source code for the benefit of the developer community. Many contributors share in a collaborative development process that has spawned complementary tools, such as the intuitive Pig code-generation framework for MapReduce and the Hive interface that supports SQL (structured query language) queries of Hadoop data.

• **Hadoop operates in a “shared nothing” mode.** All Hadoop data is stored locally on each node rather than on networked storage. Processing is distributed across an array of commodity hardware with independent CPUs and memory. The system is intelligent in the sense that the MapReduce scheduler optimizes for the processing to happen on the same node storing the associated data (or located on the same leaf Ethernet switch).

  Hadoop is good for processing that is independent, such as sums and counts that don’t require internode communications. It is not as well-suited for processing that requires iteration or communication among processing nodes, such as predictive analytics.

• **Hadoop is fault-tolerant.** Hardware failure is expected and is mitigated by data replication and speculative processing. If capacity is available, Hadoop will start multiple copies of the same task for the same block of data. Results are accepted from the task that finishes first, while the other task is canceled and the results discarded. Speculative processing enables Hadoop to work around slow nodes and eliminates the need to restart a process if a failure occurs during a long-running calculation.

  Hadoop’s fault tolerance is based on the fact that the Hadoop cluster can be configured to store data on multiple worker nodes. At this point, HDFS has not provided redundancy at the name node level – something that should be considered in your Hadoop infrastructure design.

• **Hadoop does not require a predefined schema.** A key benefit of Hadoop is the ability to upload unstructured files without having to “schematize” them first. You can dump any type of data into Hadoop and allow the consuming programs to determine and apply structure when necessary.
• **Hadoop is distributed.** Bringing a high-tech twist to the adage, “Many hands make light work,” data is stored on local disks of a distributed cluster of servers. Optimized for batch runs, processing is pushed to the data rather than vice versa for faster processing speeds.

• **Hadoop runs on commodity hardware.** Based on the average cost per terabyte of compute capacity of a prepackaged system, Hadoop is easily 10 times cheaper for comparable computing capacity, compared to higher cost specialized hardware.

On the other hand, Hadoop also has some limitations, at least in its present state:

• It is difficult to deploy and difficult to use, requiring custom code.
• It is primarily batch-oriented, which limits its performance potential.
• It provides only rudimentary analytic capabilities.
• It exists in an immature ecosystem in terms of tools, security and support.
• Skilled Hadoop professionals are in short supply.

SAS addresses these shortcomings and enables analysts to work with Hadoop from within the familiar, powerful SAS environment. More on that later.

How Are Early Adopters Using Hadoop?

By now many of us have heard how Yahoo uses Hadoop to optimize the user experience by delivering personalized ads and content to users. We know that Hadoop is well-suited for analysis or processing that can be distributed in a parallel fashion on multiple nodes. We know it’s good for managing big data. And we also know that Hadoop is ill-suited for transactional use cases because it lacks the ACID (atomic, consistent, isolated and durable) set of properties to ensure reliable processing of database transactions.

Granted, Hadoop is still maturing. Many related projects are gaining traction and may expand Hadoop’s potential. Patterns are already starting to emerge among the early adopters. Let’s take a look at these trends from the technical and business perspectives.

How Organizations Are Using Hadoop: The Technical Perspective

Hadoop does not replace enterprise data warehouses, data marts and other conventional data stores. It supplements those enterprise data architectures by providing an efficient and cost-effective way to store, process, manage and analyze the daily flood of semistructured and unstructured data. Organizations tend to use Hadoop in five ways:

A Hadoop system is self-healing in the sense that it automatically routes around failure. If a node fails, its workload and data are transparently shifted somewhere else.

Hadoop is both a data storage mechanism using the Hadoop Distributed File System (HDFS) and a parallel and distributed programming model based on MapReduce.
1) Staging Area for Data Warehouse and Analytics Store

One of the most prevalent uses of Hadoop is as a vehicle to load data into a traditional data warehouse for such activities as data mining, online analytical processing (OLAP) and reporting – or for loading data into an analytical store for advanced analytics. Organizations can dump large amounts of data into a Hadoop cluster, use a visual analytics tool to make sense out of the data, aggregate it if desired, and export the data or aggregate values into the warehouse or the data source used for analysis. The distributed processing capability of Hadoop can then be used to facilitate the extract-transform-load (ETL) processes for getting the data into the warehouse.

2) Initial Discovery and Analysis

Hadoop does well in supporting ad hoc data exploration in a sandbox environment. This involves moving data temporarily into Hadoop from various sources, potentially even the enterprise data warehouse, and using a visual analytics tool to ascertain what data is available, what value it might contain and how it can be used.

3) Storage and Analysis of Unstructured and Semistructured Content

Hadoop can be used as a storage mechanism for capturing unstructured or semistructured content and then using MapReduce to parse it, tokenize it and join it with other data. This capability explains why Hadoop was created and first adopted by Google, Yahoo and other large Web properties to support their search engines and customized placement of ads and content.

4) Making Total Data Available for Analysis

In the past, organizations were limited by how much data could be stored and what type of analytics could be applied to that data. Often, analysts could only analyze a sample subset of the data that attempted to simulate a larger data population, even when using all the data would have yielded a more accurate result.

The distributed parallel processing capacity of Hadoop can overcome the bandwidth and coordination issues associated with processing billions of records that previously might not have been saved. Because all of the data can now be captured, it can be used in high-performance analytics to search for anomalies or patterns in millions of records.

5) Low-Cost Storage of Large Data Volumes

The modest cost of commodity hardware makes Hadoop useful for simply storing large amounts of data, such as transactional data, social media data, sensor data, scientific data and emails. Hadoop can provide low-cost storage for information that is not presently critical but may become valuable later when it can be pulled back into the primary analytics data store.

For many organizations, Hadoop is used as a sandbox to dump data of uncertain value that is too unwieldy to store in a traditional relational database. The data can be loaded into Hadoop, and then aggregated, filtered and structured to see if there is any value to be gleaned from it.

Prime Data Types for Hadoop

- Radio Frequency Identification (RFID) data.
- Environmental, security monitoring and other sensor data.
- Life sciences and manufacturing process control data.
- Telecommunications device data.
- Social media and social network history.
- Image and video archives.
- Large-scale e-commerce data.
- Unstructured data (text).
How Organizations Are Using Hadoop: The Business Perspective

Hadoop can be useful in a variety of ways and in virtually every vertical market. It is gaining ground anywhere organizations need to store, process and analyze large volumes of data. For example, the technology behind Hadoop powers the world’s most successful search engines (e.g., Google and Yahoo). Each time you launch a Yahoo Search Webmap query, you are triggering a Hadoop application that spans more than 10,000 Linux servers.

In addition, Hadoop is finding a niche for a variety of other business purposes, including:

- **Behavioral analysis** of key business entities to assess customer churn, propensity to respond and other important business concerns.
- **Target marketing** to optimize advertising strategies and determine which offers should be made to each customer or prospect.
- **Marketing analysis**, such as evaluating campaign ROI and effectiveness after the fact or performing clustering, classification, market basket analysis and abandoned cart analysis.
- **Root-cause analysis** from in-depth investigation of user sessions, network logs, machine sensor data and more to determine the source of failures, issues and defects.
- **Sentiment analysis** of customers or prospects based on unstructured and semistructured data from emails, customer contact records, claims data, social media and more.
- **Fraud analysis**, using clickstream or Web analysis, mining and other techniques to detect potentially fraudulent behavior.
- **Risk mitigation** based on predictive analysis of market trends and possibilities to mitigate risks to financial position, assets, compliance, return on capital or other measures of business health.

What Is SAS Doing with Hadoop?

SAS and Hadoop are natural complements. Although there are a lot of technical details involving the various Apache subprojects and Hadoop-based capabilities, SAS support for Hadoop can be boiled down to two simple ideas:

- Hadoop data can now be leveraged using SAS.
- The power of SAS Analytics has been extended to Hadoop.

Just as with other data sources, data stored in Hadoop can be consumed across the SAS software stack in a transparent fashion. This means that analytic tools already in place (such as SAS® Enterprise Miner™), tools relating to data management (such as SAS Data Integration Studio), and foundation tools (such as Base SAS) can be used to work with Hadoop data. What you have can become even more valuable with the addition of SAS capabilities.

In 2010, Facebook claimed to have the largest Hadoop cluster in the world, with 21PB of storage, which had grown another 50 percent by mid-2011. By some estimates, half the world’s data will be processed by Hadoop within five years.
SAS has developed a number of initiatives to enable SAS users to access, load, process, visualize and analyze data stored in Hadoop. These efforts are focused on three key areas:

- Bringing the power of SAS Analytics to Hadoop with easy data access.
- Taking advantage of Hadoop’s distributed processing from within SAS.
- Bringing governance to Hadoop implementations using SAS Information Management.

**Bringing the Power of SAS® Analytics to Hadoop with Easy Data Access**

The SAS/ACCESS® Interface to Hadoop™ provides seamless and transparent data access to Hadoop, enabling Hadoop users to tap into the power of SAS. SAS extends support for the complete analytics life cycle to Hadoop, including discovery, data preparation, modeling, deployment, etc. Of particular importance are the abilities to:

- Integrate data stored in Hadoop with data from other sources.
- Apply SAS text mining, rich predictive analytics and business intelligence to data stored in Hadoop to gain and share new insights.
- Use SAS Metadata Server to create and manage metadata relating to data stored in Hadoop.

SAS/ACCESS makes Hive-based tables appear native to SAS, providing seamless connectivity to Hadoop from any SAS product or capability. Development of analytics or data processes can be done using familiar SAS tools, while the run-time execution is optimized by performing the capability in the Hadoop or SAS environment as appropriate.

**Benefits of SAS/ACCESS® Interface to Hadoop™**

- “Write once, run anywhere” can be extended to Hadoop.
- Integrate data stored in Hadoop with data from other sources.
- Directly access data, easily and securely, with native interfaces.
- Seamlessly leverage Hadoop data for existing SAS capabilities.
- Gain faster performance by minimizing data movement.
- Support both technical and business users.

**Technical Highlights of SAS/ACCESS® Integration**

- LIBNAME statement makes Hive tables look just like SAS data sets.
- PROC SQL provides the ability to perform explicit HiveQL commands into Hadoop.

**Taking Advantage of Hadoop’s Distributed Processing from Within SAS®**

SAS provides the ability to execute Hadoop functionality by enabling MapReduce programming, scripting support and the execution of HDFS commands from within the SAS environment. This complements the capability that SAS/ACCESS provides for Hive by extending support for Pig, MapReduce and HDFS commands.
Running SAS within Hadoop distributed computing leads to the responsiveness and high throughput necessary for large analytic workloads. This is accomplished by running a thin-layer SAS process on each Hadoop node, which processes SAS commands. Hadoop takes care of distributing the processing, managing memory, controlling job flow, etc.

Benefits of Base SAS® Integration with Hadoop Distributed Processing

- Combine world-class analytics with a strong distributed processing and commodity-based storage infrastructure.
- Access your Hadoop data in SAS, the language you already know.

Technical Highlights of Base SAS® Integration with Hadoop

- Submit Hadoop programming language, scripting and file commands from the SAS execution environment, including support for Pig, MapReduce and HDFS commands.
- Support external file references, so Hadoop files can be conveniently referenced from any SAS component.
- Parameters necessary to process the file, such as delimiters, are externalized, which makes it convenient to work with a Hadoop file.
- SAS procedures (including PROC FREQ, PROC RANK, PROC REPORT, PROC SORT, PROC SUMMARY, PROC MEANS and PROC TABULATE) are supported.

Bringing Governance to Hadoop Implementation with SAS® Information Management

One of the issues plaguing Hadoop is the lack of, or relative immaturity of, tools for developing and managing Hadoop deployments. SAS data management and analytics management offerings can help organizations derive value from Hadoop more quickly with fewer resources. For example, Hadoop integration with SAS Information Management provides the ability to:

- Use an intuitive, graphical SAS user interface to develop Hadoop capabilities that use Pig, Hive, MapReduce and HDFS commands inline.
- Use standard SAS Data Integration Studio transformations to work with Hadoop data and build job flows.
- Streamline code generation with visual editing and a built-in syntax checker for Pig and MapReduce code.
- Submit Hive queries through the PROC SQL capability that can also be surfaced through Base SAS and other SAS components.
- Use SAS to create user-defined functions (UDFs) that can be deployed within HDFS, including the ability to use SAS Enterprise Miner to take analytical scoring code and produce a UDF that can be deployed within HDFS and then accessed by Hive, Pig or MapReduce code.
- Apply DataFlux® data quality capabilities to data coming in or out of Hadoop.
- Register and manage Hadoop files via the SAS Management Console for easy inclusion of Hadoop within a SAS capability – business intelligence, data integration, SAS/STAT®, etc.
- Use SAS data lineage and security capabilities to more effectively manage a Hadoop environment and gain a complete view into changes made to Hadoop data.
- Support capabilities provided by SAS MPP partners such as EMC Greenplum and Teradata.

Benefits of SAS® Information Management for Hadoop
- Extend “write once, run anywhere” to the Hadoop environment.
- Extend Hadoop access to any SAS component.
- Use SAS visual tools to simplify working with Hadoop data.
- Provide full visibility into Hadoop data access and usage.
- Apply industry-leading data quality capabilities to Hadoop data.
- Extend the power of Hadoop by embedding SAS analytics and data management logic as UDFs.

Technical Highlights of SAS® Information Management for Hadoop
- From within SAS Data Integration Studio, build job flows that use Pig, MapReduce and HDFS commands, Hive queries, and UDFs.
- Streamline code generation with visual editing and a built-in syntax checker for Pig and MapReduce code.
- Apply SAS data quality, data lineage, security and data management capabilities to work with Hadoop data.

SAS® Data Integration Studio users can load and unload data to and from Hadoop, perform ELT-like processing with the Hive query language and ETL-like processing with the Pig language for MapReduce code generation.

Closing Thoughts
The SAS approach combines the power of world-class analytics with Hadoop’s ability to use commodity-based storage and to perform distributed processing. This integration provides benefits to organizations looking to get the most from their big data assets:

- SAS both simplifies and augments Hadoop. The ability to abstract the complexity of Hadoop by making it function as just another data source brings the power of SAS and its well-established community to Hadoop implementations. SAS graphical tools enable users to access, process and manage Hadoop data and processes from within the familiar SAS environment. This is critical, given the skills shortage and the complexity involved with Hadoop. SAS augments Hadoop with world-class analytics, along with metadata, security and lineage capabilities, which helps ensure that Hadoop will be ready for enterprise expectations.
• **SAS delivers optimal solutions for each organization’s specific mix of technologies.** SAS Data Integration supports Hadoop alongside other data storage and processing technologies. This integration offers greater flexibility than other vendor-specific products that only use Hadoop as a vehicle for landing more information on certain database or hardware platforms.

• **SAS brings much-needed governance to Hadoop.** SAS provides a robust, comprehensive, information management lifecycle approach to Hadoop that includes data management and analytics management support. This is a huge advantage over other products that focus primarily on moving data in and out of Hadoop.

SAS support for Hadoop is part of our broader big data strategy that includes information management for big data and high-performance analytics, including grid, in-database and in-memory computing.

The ability to work with Hadoop and SAS enables customers to combine world-class analytics with a strong distributed processing and commodity-based storage infrastructure.
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.