



Bringing the Power of SAS® to Hadoop

Combine SAS® World-Class Analytic Strength with Hadoop's Low-Cost, Distributed Data Storage to Uncover Hidden Opportunities

Contents

Introduction.....	1
What Is Hadoop and What Does It Do?.....	1
The Benefits of Hadoop	1
Its Current Limitations	1
Technical Use-Case Scenarios.....	2
Low-Cost Storage and Active Archive.....	2
Staging Area for a Data Warehouse and Analytics Store	2
Data Lake	2
Sandbox for Discovery and Analysis	2
Prime Business Applications for Hadoop	2
What Can You Do With SAS® and Hadoop?.....	3
Access and Manage Hadoop Data.....	3
Interactively Explore and Visualize Hadoop Data	4
Analyze and Model Using Modern Statistical and Machine-Learning Methods	4
Deploy and Integrate	4
For the Data Scientist	4
Closing Thoughts.....	5

Introduction

Hadoop. That cute little yellow elephant icon is all over the place. And the distributed, open-source framework behind it has caught the attention of many organizations searching for better ways to store and process large volumes and varieties of data. From executives and business analysts to data stewards, data scientists and analytic professionals – in many circles, it seems like Hadoop is the most popular kid (or animal) on the block.

What Is Hadoop and What Does It Do?

Apache Hadoop is an open-source software framework for storage and large-scale data processing on clusters of commodity hardware. As an Apache top-level project, Hadoop is being built and used by a global community of contributors and users. The Hadoop framework is composed of the following modules:

- Hadoop Common – contains libraries and utilities needed by other Hadoop modules.
- Hadoop Distributed File System (HDFS) – a distributed file system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster.
- Hadoop YARN – a resource-management platform responsible for managing compute resources in clusters and using them to schedule users' applications.

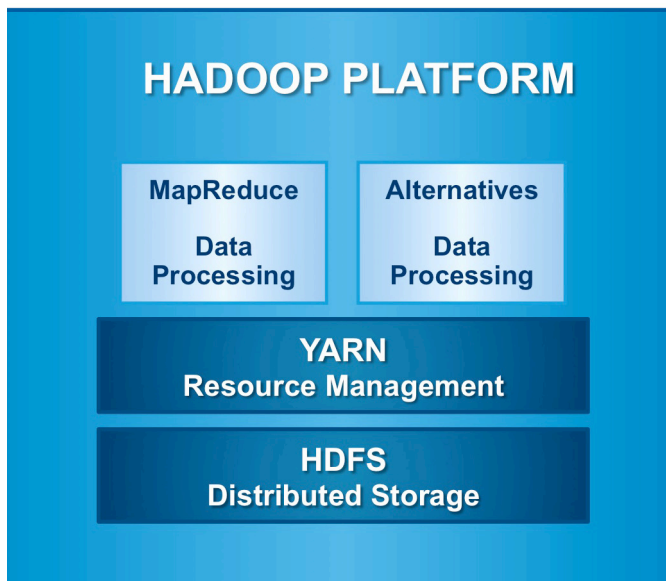


Figure 1: The Hadoop platform

- Hadoop MapReduce – a programming model for large-scale data processing. It divides applications into smaller components and distributes them across numerous machines.
- Other open-source components that augment HDFS, MapReduce and YARN are Pig, Hive, Hbase, etc.

Hadoop clusters can easily and economically scale to thousands of nodes to store and process structured and unstructured data. So it's no surprise that many organizations are either planning a project or expanding their Hadoop footprint. TDWI reports that 51 percent of organizations it surveyed expect to have a Hadoop implementation in place by 2016.¹ So now is the time to start thinking about integrating this powerful ecosystem with your data management, data discovery and advanced analytics solutions.

The Benefits of Hadoop

There are several reasons that 88 percent of organizations consider Hadoop an opportunity.²

- It's inexpensive. Hadoop uses lower-cost commodity hardware to reliably store large quantities of data.
- Hadoop provides flexibility to scale out by simply adding more nodes.
- You can upload unstructured data without having to "schematize" it first. Dump any type of data into Hadoop and apply structure as needed for consuming applications.
- If capacity is available, Hadoop will start multiple copies of the same task for the same block of data. If a node goes down, jobs are automatically redirected to other working servers.

Its Current Limitations

- Management and high-availability capabilities for rationalizing Hadoop clusters with data center infrastructure are only now starting to emerge.
- Data security is fragmented, but new tools and technologies are surfacing.
- MapReduce is very batch-oriented and not suitable for iterative, multi-step analytics processing.
- The Hadoop ecosystem does not have easy-to-use, full-feature tools for data integration, data cleansing, governance and metadata. Especially lacking are tools for data quality and standardization.
- Skilled professionals with specialized Hadoop skills are in short supply and at a premium.

Hadoop definitely provides economical data storage. But the next step is to manage the data and use analytics to quickly identify previously unknown insights. Enter SAS. More on that later.

¹ Philip Russom, TDWI Best Practices Report: Integrating Hadoop into Business Intelligence and Data Warehousing, Second Quarter, 2013.

² Ibid.

MapReduce is file intensive. And because the nodes don't intercommunicate except through sorts and shuffles, iterative algorithms require multiple map-shuffle/sort-reduce phases to complete. This creates multiple files between MapReduce phases and is very inefficient for advanced analytic computing.

Technical Use-Case Scenarios

Low-Cost Storage and Active Archive

The modest cost of commodity hardware makes Hadoop useful for storing big data such as transactional, social media, sensor, machine, scientific, click streams, etc. and combining those with public data sources. IT has the flexibility to adjust configurations up or down to meet new data collection demands. Hadoop also provides low-cost storage for information that is not currently critical but could become useful later for business analytics. There is value in keeping archived data and revisiting it when context changes or new constraints need to be evaluated.

Staging Area for a Data Warehouse and Analytics Store

One of the most prevalent uses of Hadoop is to stage large amounts of raw data for subsequent loading into an enterprise data warehouse (EDW) or an analytical store for high-value activities such as advanced analytics, query and reporting, etc. Organizations are looking at Hadoop to handle new types of data (e.g., unstructured) at a granular level, as well as to offload some historical data from their EDWs. Visual analytics tools can help make sense of the data, refine and aggregate it, and export it into an EDW for analysis. This could help contain EDW costs. Or, enable more or new data to be analyzed at a lower cost to support a business need before moving the data to other repositories.

Data Lake

Hadoop is often used to create a data lake, which is a way of storing large amounts of data without the constraints introduced by schemas commonly found in the SQL-based world. It is used as a low-cost compute-cycle platform that supports processing of ETL and data quality jobs in parallel using hand-coded or commercial data management technologies. Moving big data to a data integration server can become infeasible because of bandwidth issues or inability to meet batch-processing windows. Instead, relevant data management jobs are processed in Hadoop without the cost of expanding current EDW infrastructures. Refined results can then be passed to other systems (e.g., EDWs, analytic marts) as needed for reuse.

Sandbox for Discovery and Analysis

Because Hadoop was designed to deal with volumes of data in a variety of shapes and forms, it will be an enabler for analytic use cases. Big data analytics on Hadoop will help run current business more efficiently, uncover new opportunities and derive next-level competitive advantage. Hadoop serves as a data platform for ad hoc data discovery and analytics in a sandbox environment. This might involve analyzing provisional data (i.e., used once) or moving data temporarily into Hadoop from various sources and using in-memory analytic tools to ascertain what relationships, insights and value the data holds. And then sharing it. You can develop analytical (e.g., predictive, text mining) models at a greater level of data granularity or full-scale analysis can be performed. The sandbox setup provides a quick and perfect opportunity to innovate with minimal investment.

Prime Business Applications for Hadoop

Hadoop is providing a data storage and analytical processing environment for a variety of business uses, including:

- Financial services: Insurance underwriting, fraud detection, risk mitigation and customer behavior analytics.
- Retail: Location-based marketing, personalized recommendations and website optimization.
- Telecommunications: Bandwidth allocation, network quality analysis and call detail records analysis.
- Health and life sciences: Genomics data in medical trials and prescription adherence.

- Manufacturing: Logistics and root cause for production failover.
- Oil and gas and other utilities: Predict asset failures, improve asset utilization and monitor equipment safety.
- Government: Sentiment analysis, fraud detection and smart city initiatives.

What Can You Do With SAS® and Hadoop?

SAS support for Hadoop spans the entire analytic life cycle and centers on a singular goal – helping you know more, faster, so you can make better decisions with more confidence. Our unique in-memory processing brings analytics as close to the data as possible to avoid time-consuming data movement. With SAS, you're covered. You can access and integrate data from Hadoop, push SAS processing to the Hadoop cluster via MapReduce, or lift data from HDFS into memory and perform distributed data processing, exploratory analysis, analytic calculations and more – all interactively. From data access and management to exploration, modeling and deployment, let's take a look at what SAS offers for Hadoop.

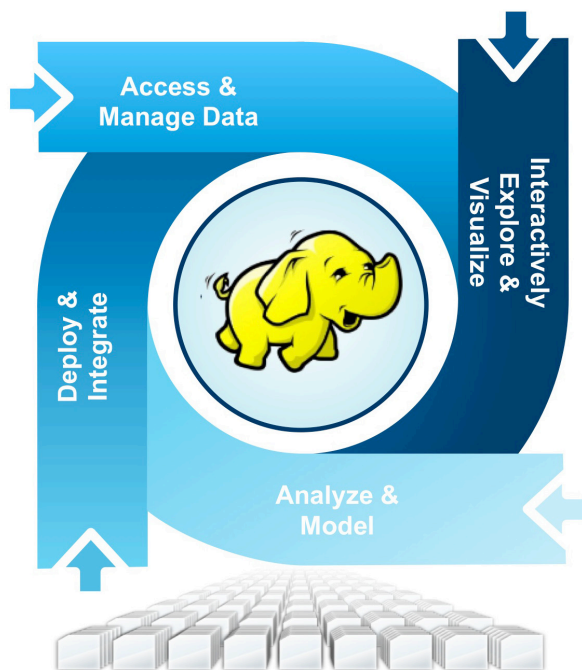


Figure 2: SAS® data-to-decision life cycle

Access and Manage Hadoop Data

- **Easily access data stored in Hadoop.** SAS/ACCESS® software modules are out-of-the box solutions that provide data connectivity and integration between SAS and external data. **SAS/ACCESS to Hadoop** (for Hive and Hive Server2) provides fast, efficient access to data stored in Hadoop via HiveQL. You can access Hive tables as if they were native SAS data sets and then analyze them using SAS. Cludera Impala is an open-source massively parallel processing (MPP) query engine that runs natively on Apache Hadoop. It enables users to issue SQL queries to data stored in HDFS and Apache Hbase without requiring data movement or transformation. Just like any other **SAS/ACCESS engine**, **SAS/ACCESS to Impala** lets you run SAS procedures against data stored in Impala and returns the results to SAS.
- **Read/write and integrate data to and from Hadoop.** **SAS Data Loader for Hadoop** enables business users to profile, transform, cleanse and move data on Hadoop without requiring specialized skills or help from IT. The intuitive, wizard-based web user interface provides self-service access to Hadoop data. SAS programmers can use the **HADOOP procedure (Base SAS)** to submit HDFS commands, MapReduce programs and Pig language code against Hadoop data. And, being able to push down common SAS statistics procedures to Hadoop data from SAS makes it easy for SAS users to get started with Hadoop. Base SAS also lets you easily store large SAS data sets on Hadoop clusters and provides security features such as encryption and password protection. **SAS Data Integration Studio (SAS Data Management)** provides an intuitive GUI, templates and a data transformation library for writing Hadoop programs in Pig, Hive and MapReduce, and for loading that data into the SAS® LASR™ Analytic Server for visualization.
- **Get virtual views of Hadoop data without physically moving it.** **SAS Federation Server** offers simplified data access, administration, security and performance by creating a virtual data layer without physically moving data. This frees business users from the complexities of the Hadoop environment. They can view data in Hadoop and virtually blend it with other database systems like SAP HANA, IBM DB2, Oracle or Teradata. Improved security and governance features, such as dynamic data masking, ensure that the right users have access to the right data.
- **Accelerate SAS processing and data quality functions in Hadoop.** The **SAS Data Quality Accelerator for Hadoop** and the **SAS In-Database Code Accelerator for Hadoop** (both components of the **SAS Data Loader for Hadoop**) process SAS DS2 code inside Hadoop using the MapReduce

framework and the SAS Embedded Process. This minimizes the time spent moving and manipulating data for analytic processing and cleansing. It also eliminates the need to know how to code this capability in MapReduce and improves performance and execution response times by harnessing the power of the Hadoop cluster.

Interactively Explore and Visualize Hadoop Data

- **Quickly visualize data stored in Hadoop, discover new patterns and publish reports.** To get value from vast and diverse data stored in Hadoop, organizations often start with data exploration. **SAS Visual Analytics** lets you explore all types of data stored in Hadoop in an interactive and very visual way. This easy-to-use, in-memory solution identifies trends and relationships between variables that weren't evident before so you can spot opportunities for further analysis. You can then share results quickly and easily via web reports, mobile devices or in Microsoft Office applications.

Analyze and Model Using Modern Statistical and Machine-Learning Methods

- **Uncover patterns and trends in Hadoop data with an interactive and visual environment for analytics.** **SAS Visual Statistics** provides an interactive, intuitive, drag-and-drop web browser interface for building descriptive and predictive models on data of any size – rapidly! It also takes advantage of the SAS LASR Analytic Server to persist and analyze data in-memory and deliver near instantaneous results. When combined with **SAS Visual Analytics**, users get best-in-class data visualization and exploratory predictive modeling in a common environment. Quickly discover hidden relationships between variables to accelerate the model-building and refining process for numerous scenarios.
- **Apply domain-specific high-performance analytics to data stored in Hadoop.** **SAS High-Performance Analytics products** provide in-memory capabilities so you can develop superior analytical models using all data, not just a subset. These products load data into memory in parallel and apply complex analytical algorithms to the distributed data in-memory – minimizing time to get answers. SAS high-performance products are available for statistics, data mining, text mining, econometrics and optimization. And, you can run your existing SAS analytical programs with minimal modifications in the distributed mode.

- **Automatically deploy and score analytic models in the parallel environment.** Once an analytical model has been reviewed, signed off and declared ready for production, **SAS Scoring Accelerator for Hadoop** takes the model and publishes it into scoring files or functions that are stored and automatically deployed inside HDFS. The scoring files are then used by a MapReduce function to run the scoring model. SAS Scoring Accelerator minimizes time to production for predictive models by automating model deployment in Hadoop, reduces data movement by bringing the scoring process to Hadoop data, and produces faster answers using the distributed processing in Hadoop.
- **Manage and analyze real-time data – as it enters Hadoop.** Big data isn't "big" simply because of the disk space it takes up. Big data also relates to the data velocity. Today's organizations receive streams of data, such as financial transactions or smart meter readings, that feed your big data environment. **SAS Event Stream Processing Engine** puts intelligence into this process by continuously querying data streams to examine and filter data, even if you process millions of events per second. By analyzing data as it arrives, you can detect anomalies and uncover useful data more quickly – and publish filtered data to your Hadoop cluster.

For the Data Scientist

SAS In-Memory Statistics for Hadoop was created especially for data scientists working in Hadoop environments. It provides a single, powerful interactive programming environment that is highly scalable and optimized for distributed analytics. The single interactive programming environment enables multiple users to concurrently prepare Hadoop data for analysis, transform variables, perform exploratory analysis, build and compare models, and score models – all inside the Hadoop environment.

Closing Thoughts

Deploy and Integrate

The SAS approach combines the power of world-class analytics with Hadoop's ability to use commodity-based storage and large-scale data processing. This integration provides benefits organizations looking to get the most from their big data assets. Here are a few things to consider:


- **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 Analytics, and our well-established user community, to Hadoop implementations. Users can explore, manage and analyze Hadoop data and processes from within their familiar SAS environment. This is critical, given the skills shortage and complexity involved with Hadoop.
- **Better manage your Hadoop data.** Not only can SAS access data from Hadoop faster than anyone else, in-Hadoop data quality routines and data integration help reduce data movement, keep track of data and speed up the process of getting trusted data to fuel analytics. Manage Hadoop data to promote reuse and consistency, and comply with IT policies. With SAS, you can stop the "garbage in, garbage out" cycle and get the most accurate answers from even your biggest data.
- **Get more precise results with superior analytics for Hadoop.** Big data opportunities hidden in Hadoop are best exploited with modern analytic techniques. We bring the largest, most proven variety of analytical algorithms – predictive analytics, machine learning techniques, optimization and text mining. So you can unlock the value in the structured and unstructured data stored in Hadoop. Turn new ideas into meaningful actions. Use the timely and precise insights to identify opportunities that evade your competitors.
- **Add unprecedented speed and interactivity to Hadoop for increased productivity for data scientists, analysts and**

statisticians. Our unique in-memory engine, the SAS LASR Analytic Server, takes analytics closer to Hadoop, reading the data once and holding it in-memory for multiple analyses within a session – across multiple users. This provides faster response times and enables everyone to take faster actions. Compare that to using the Map-Shuffle/Sort-Reduce pattern that writes data to disk after each phase of computation. All of that big data shuffling is extremely inefficient.

Hadoop or not, success for any project is determined by value, rather than volume, of results. Metrics built around revenue generation, margins, risk reduction and business process improvements will help pilot-based projects gain wider acceptance and garner interest from other departments.

SAS can be a valuable, integral part of successful Hadoop implementations. Our comprehensive support for the entire data-to-decision process enables organizations to supplement and grow Hadoop's potential from a case-by-case basis (i.e., pilot projects) to broad, significant use across your enterprise.

For more information, visit sas.com/hadoop.



SAS works with Apache Hadoop, Cloudera, Hortonworks, Pivotal, IBM Big Insights and MapR.

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

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. Copyright © 2014, SAS Institute Inc. All rights reserved.
105776_S130566.0914

