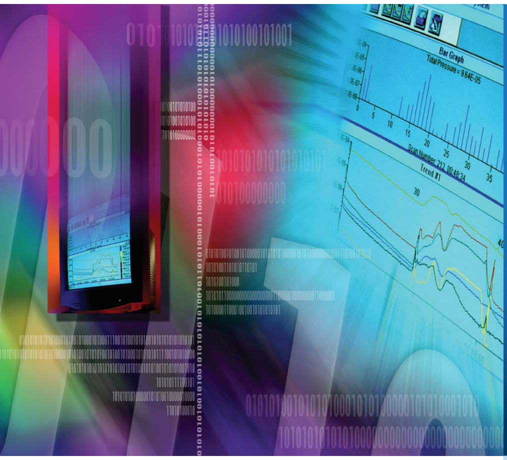
White Paper

alteryx



Everything You Need to Know about Apache Spark for Analytics

"Professionals with big data or cloud expertise continue to serve as high-priority candidates. Profitable companies today understand employing tech professionals who understand business goals and strategic priorities are more than employees, they're partners in achieving success."

Bob Melk, President, Dice

Introduction

You're a line-of-business analyst. Your job is to uncover and deliver insights that respond to a variety of business needs—which you need to do as quickly as possible. But with analytics technology changing at a torrid pace, it's hard to keep up. The proliferation of cloud data warehouses and applications, the emergence of data lakes, and the rise of the Internet of Things (IoT) are making the already complex world of analytics even more entangled. And, as analytics become the lifeblood of organizational performance, you will face increasing pressure to deliver deeper insights even faster.

The key to staying ahead of the analytic curve is to understand, prepare for, and leverage new and emerging tools and technologies, such as Apache Spark™. And, as the lines between analysts, data scientists, and IT blur even more, understanding Spark and how it can impact the way you approach analytics is crucial.

What Is Apache Spark and How Does It Work?

Spark is an open-source big data processing framework built around speed, ease of use, and sophisticated analytics. It is easy to deploy because it runs on existing Hadoop clusters and data. An in-memory, parallelized compute engine that combines SQL, streaming, and complex analytics, Spark can handle a range of data processing needs, making powerful analytics simple and scalable.

As a framework, Spark includes several components, all of which contribute to how flexibly and scalably Spark can handle data and advanced analytics. There are three (3) components that impact line-of-business analysts in particular:

- **SparkSQL** – Performs SQL queries in Hadoop File Systems (HDFS) and other data stores using the parallelized in-memory compute engine. SparkSQL allows for rapid execution of SQL queries, especially when compared with Hadoop MapReduce-based Hive.
- **Spark MLib** – Provides predictive analytics capabilities within the Spark framework.
- **SparkR** – Enables users to access Spark with a lightweight front end based on the popular R programming language.

Why Should You Care about Spark?

The same reasons that Spark is attracting the attention of information architects, developers, and data scientists are why you should care too: speed and scalability, ease of use, and rapid adoption.

Speed and Scalability

The ability to process a wide variety of data types, and do so extremely quickly, should make every line-of-business analyst sit up and take notice. According to the Apache Spark project, Spark can run programs up to 100 times faster than Hadoop MapReduce in memory, or 10 times faster on disk[1]—a significant difference when you're looking for time-critical business insights.

When working with traditional big data, where datasets are very large, or when performing many advanced or computation-intensive actions, Spark dramatically improves processing by optimizing and processing data in parallel. The result? Spark can easily keep up with the increasingly large volumes of information associated with data lakes, IoT, and cloud-based data.

[1] http://spark.apache.org

In 2014 alone, Spark "saw growth in ... open source contributors from approximately 150 to 500, a near doubling of the number of lines of code, and an increase to more than 500 active production deployments ... [and] Spark was the most active Apache Software Foundation project, being two to three times more active than any other open source big data project." [2]

Ease of Use

Designed to break down the silos created by data that's distributed across a variety of storage engines, Spark makes accessing big data easier through the use of interactive, distributed queries that can process petabytes of data.

Because of this, Spark is perfectly suited for today's increasingly sophisticated analytics, including advanced analytic methods that require scalable computing clusters. And, as a general processing engine, Spark can handle a wide variety of data processing tasks as well as data source types, including flat files, external databases, and Hive tables.

Rapid Adoption

The growing investment in Spark development by all major Hadoop vendors—including Hortonworks, Cloudera, and MapR—and an increase in certified Spark developer programs are combining to position Spark as the next big technology for analytics. What's more, because Spark is easily deployed on a Hadoop cluster and comes packaged with higher-level libraries for ease of use, organizations large and small are adopting Spark at a rapid pace. Growing investment in the technology as well as its rapid adoption rate mean that Spark will soon be a mainstream part of the analytic infrastructure, enabling organizations to be increasingly agile in their decision-making processes.

What Business Benefits Will You Get from Spark?

According to a recent study by *Harvard Business Review*, companies that believe strongly in the benefits of adopting new technologies and pursue first-mover advantage are more likely to lead in both revenue growth and market position.[3]

The correlation between the early adoption of new technologies and better business outcomes carries over to all parts of an organization, including analytic departments. The ability to readily adopt technologies—such as Spark—ensures that you can be highly responsive to decision-makers throughout your company—protecting your competitive position.

And just as the seamless adoption of new technologies results in stronger competiveness, so too does the ability to deliver deeper business insights faster.

One of the fastest scalable processing technologies currently available, Spark is uniquely designed to handle the needs of big data and advanced analytics, enabling you to quickly and easily process vast amounts of diverse data. The result? You can understand and analyze the wealth of information at your disposal faster.

Not only must you make decisions faster, but you need to base these decisions on what is likely to happen in the future instead of on what has happened in the past. Predictive analytics is quickly becoming a necessity for businesses large and small. Because Spark can deliver predictive insights at petabyte scale, you can glean these important insights—and prepare your organization to make impactful changes to stay ahead of your competition.

Alteryx and Spark: Not just for Data Scientists or Big Data Engineers

Alteryx, with its repeatable workflow architecture, intuitive interface, and tight integration with Spark, makes the benefits of Spark accessible to more than just big data engineers or data scientists. The Alteryx Analytics platform combines ease of use and sophisticated analytics so you can make impactful decisions for your business—in hours rather than days or weeks.

[2] Brief: Apache Spark Ignites The Big Data Landscape; Forrester; June 15, 2015

[3] https://hbr.org/resources/pdfs/comm/verizon/18832_HBR_Verizon_Report_IT_rev3_webview.pdf

With Alteryx, you can access, prep, and blend your data sources, then perform advanced analytics—without any coding. Alteryx also makes it easy to combine data stored within a Spark framework with other datasets (e.g., an Excel file, an Access database, or even another data warehouse), and conduct analytics on the combined dataset.

What's more, you can leverage the power of the Spark framework with the use of Alteryx in-database (In-DB) tools. Without writing any SQL code, these tools allow you to push the data processing steps into the Spark platform and retrieve only the data you need, rather than pulling the entire dataset to the processing location (typically, your desktop). When working with extremely large datasets, the In-DB processing option can yield significantly improved performance since you're limiting the movement of vast amounts of data and taking full advantage of the processing speed Spark offers. And the flexibility of Alteryx provides bidirectional functionality, enabling you to easily pull data out of the Spark platform as well as push in external data.

Once you've prepared your data, Alteryx enables you to perform advanced analytics within the same workflow. Alteryx Analytics includes more than 40 prebuilt predictive tools, built using the R language, that you can easily integrate into your analytic workflow via a drag-and-drop user interface. And with the ability to prep, blend, and analyze data in one unified workflow, you no longer waste time integrating data before conducting analysis.

After you complete your analysis, Alteryx makes it easy to package insights into any form your colleagues want to use for viewing, from reports and customizable analytic apps to native data files for all popular visualization tools, including Tableau, Qlik, and Microsoft Power BI. The result? You get insights into the hands of decision-makers faster. With a single click, you can share, iterate, and update your analytic insights, without requiring you to rework your analysis or request assistance from other departments.

What's the Net-Net of Spark?

Analytic technology is changing at a breakneck pace, and it's not easy to keep up. But if you, as a data analyst, can quickly understand as well as leverage the breadth of analytic technology available, you can not only increase your value to your organization, but also improve the quality and quantity of insights you can deliver to decision-makers across the executive suite.

Alteryx makes it easier for you to keep pace with and embrace Spark technology to provide added value to your organization—without requiring you to be a data scientist, have a Ph.D., or go through a certification process.

With the Alteryx self-service data analytics platform, you can easily prep and blend data—regardless of data size or location—and perform predictive, statistical, and spatial analytics in a single, repeatable workflow. Experience it for yourself by visiting **www.alteryx.com/trial** for your free trial today.

About Alteryx

Alteryx is the leader in self-service data analytics. *Alteryx Analytics* provides analysts with the unique ability to easily prep, blend, and analyze all of their data using a repeatable workflow, then deploy and share analytics at scale for deeper insights in hours, not weeks. Analysts love the Alteryx Analytics platform because they can connect to and cleanse data from data warehouses, cloud applications, spreadsheets, and other sources, easily join this data together, then perform analytics—predictive, statistical, and spatial—using the same intuitive user interface, without writing any code. Thousands of companies and data analysts worldwide rely on Alteryx daily. Visit www.alteryx.com or call 1-888-836-4274.

alteryx

230 Commerce, Ste. 250
Irvine, CA 92602
+1 714 516 2400
www.alteryx.com