The Business Analyst's Guide to Hadoop

Get Ready, Get Set, and Go: A Three-Step Guide to Implementing Hadoop-based Analytics

By Alteryx and Hortonworks

"...(T)here is considerable evidence that organizations are entering the Analytics 3.0 world. It's an environment that combines the best of 1.0 and 2.0—a blend of big data and traditional analytics that yields insights and offerings with speed and impact."

—Thomas H. Davenport ,
author of Competing on Analytics:
The New Science of Winning

Introduction

Rapidly emerging as a transformative technology framework for storing and processing massive amounts of structured and unstructured data, Apache Hadoop plays a central role in many organizations' strategies to exploit the analytic potential of Big Data.

However, most Big Data strategies stall or slow down to a crawl as the organization struggles with the IT challenges associated with data volume, velocity, and variety. How do we collect all the data? How should we store all that data? And where should we store it? Companies spend so much time on these technical issues that they lose sight of the most important question: How do we identify and prioritize areas where Big Data can yield the greatest business value?

In order to harness Big Data for competitive advantage, organizations must enable more than a handful of scarce IT specialists and expensive data scientists to access the information. By making Big Data usable by a broader community of business decision-makers and analysts, organizations 'humanize' Big Data, thereby extracting real business value.

If you are a business analyst tasked with analyzing Big Data, understanding Hadoop and key related concepts is critical to your success. This paper outlines three (3) basic steps to help you get started with Hadoop-based analytics and deliver value to your organization.

Before You Start: Know Key Hadoop Concepts

Here are some key Hadoop-related concepts and technologies to familiarize yourself with before you start evaluating and implementing Hadoop-based analytics in your organization:

- **Apache Hadoop** An open source project from the Apache Software Foundation that has rapidly emerged as the best way to handle massive amounts of data, a.k.a., Big Data.
- **MapReduce** A framework for writing applications that processes large amounts of structured and unstructured data in parallel batches across large clusters of machines in a very reliable and fault-tolerant manner.
- **Apache Hive** Built on the MapReduce framework, Apache Hive is a data warehouse that enables easy data summarization and ad-hoc queries via a SQL-like interface for large datasets stored in Hadoop Distributed File System (HDFS).
- **Apache Pig** A platform for processing and analyzing large data sets. Pig consists of a high-level language (Pig Latin) for expressing data analysis programs paired with the MapReduce framework for processing these programs.
- **Apache HCatalog** A table and metadata management service that provides a centralized way for data processing systems to understand the structure and location of the data stored within Apache Hadoop.
- **Hortonworks Stinger Initiative** A community-driven project to accelerate and expand the capabilities of Apache Hive. The goal of the of the initiative is to increase the performance of Hive, the defacto SQL standard for Hadoop, by 100x, enabling Hive to meet a wider set of end-user workloads.

Many organizations are augmenting internally generated data from sales and service transactions with social media chatter and external demographic data using Hadoop-based analytics to:

- Identify new customer segments
- Personalize offers
- Reduce customer churn

Businesses in asset-intensive industries, such as utilities, oil and gas, and industrial manufacturing, can reduce maintenance costs and improve asset utilization with Hadoop-based analytics. By integrating machine-generated data, internally-generated service and warranty data, and external data from asset manufacturers, and then applying predictive analytics, these businesses can move from scheduled to "as needed" maintenance intervals.

Get Ready – Understand the Value of Hadoop-based Analytics

While IT professionals who utilize Hadoop are well-versed in the defining attributes of Big Data—volume, variety, and velocity—many are unable to identify and articulate potential business value and use cases for Big Data. This is where you, the business analyst, come in. Why? Because no matter the technology underpinnings, you understand the answers your business needs—and the relevant questions to ask in order to uncover them.

One of the major advantages of Hadoop is that it overcomes the performance and scalability limitations of traditional data storage technologies while leveraging low-cost commodity hardware. As a result, organizations can perform analytics against much larger and more diverse data sets than ever before—and at much lower cost. These technology breakthroughs create unprecedented opportunities to combine Big Data with information from traditional data sources and enable skilled business analysts to discover new insights, patterns, and trends in the business.

Most organizations actively using Hadoop for Big Data analytics use one of two primary approaches, driven by the specific needs of the application:

1. Use Hadoop to refine and load data into a data warehouse

In this type of deployment, the organization pulls large data sets, which can include both structured and unstructured data, from various sources and moves them into a Hadoop data platform. Subsequently, the organization processes and distills the information into a more manageable data set that can then be loaded into a data warehouse.

As an example, a major US-based specialty department store chain gains an integrated view of customer behavior and preferences by storing and processing massive volumes of weblog data in Hortonworks Data Platform (HDP). After processing the data in HDP, the company moves the distilled information into a data warehouse for analysis. In the warehouse, the data is combined with purchase information and other data to give it context, so the analysis can show how certain actions on the website lead to purchases.

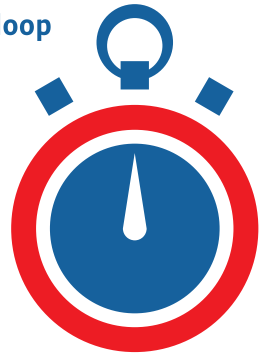
2. Leverage the Hadoop platform as the data store

The second, and more popular deployment approach, leverages the Hadoop platform as the data store for exploration and visualization, without refining and moving information into a data warehouse. In other words, Hadoop serves as a peer to the data warehouse. The company then uses business intelligence and analytic tools to directly access extremely large data sets that are unwieldy and costly in a traditional data warehouse. In this type of deployment, organizations are typically looking for patterns within the data that will yield new business opportunities and efficiencies, identify areas of potential risk, or detect fraud.

Almost daily, new applications for Hadoop-based analytics surface across a broad spectrum of industries.

- **Retailers** deploy Hadoop-based solutions for site selection, brand and sentiment analysis, market basket analysis, and loyalty program optimization.
- **Financial services organizations** leverage Hadoop for similar applications, as well as fraud detection and risk assessment.
- **Government agencies** use Hadoop-based analytics for applications related to law enforcement, public transportation, national security, health, and public safety.

Get Set – Maximize the Value of Data Stored in Hadoop



In bridging the challenges of IT with the needs of the business, you must address three key priorities. First, analytic solutions must be delivered quickly to serve time-sensitive problems and opportunities. Second, these solutions must incorporate all relevant data to ensure questions are answered with proper context. Finally, the solutions must be easy to use by a large base of consumers within the organization.

Speed Time to Value

Traditional data warehousing and business intelligence solutions often take months or even years to deploy. Even when a data warehouse is already in production, basic changes can be time-consuming and expensive. For example, the process of adding new data sources to the warehouse can involve extensive source system analysis, changes to the warehouse data model, as well as designing, testing, and maintaining ETL maps.

Using Big Data solutions from Alteryx and Hortonworks, you can overcome many of the obstacles to rapid deployment. Hortonworks Data Platform (HDP) provides a flexible data platform to store, process, and analyze data at any scale. You can store processed Big Data in a data warehouse, HDP, or in a hybrid mode. The Alteryx Strategic Analytics platform complements this flexibility by enabling you to access and blend data from Hadoop and traditional data sources, without first needing to move all the data into a common data warehouse. This query federation capability eliminates the time-consuming process of modifying the data warehouse, creating ETL maps, and running batch load processes.

Blend Data to Add Context

You gain the greatest value from Big Data when you blend data—from new as well as traditional sources, such as transactional systems and data warehouses—to provide the right context. For example, many organizations use business intelligence tools for sales and revenue analysis. Key Performance Indicators (KPIs) generated from structured data in CRM and ERP systems provide valuable insight for sales managers about the current and projected state of the business. Those KPIs can be even more valuable, however, when you augment them with customer sentiment information derived from unstructured data sources, such as product forums and social media. By providing this contextual information, you can understand why certain trends are occurring and suggest what actions might lead to better performance.

The ability to add context by blending Big Data stored in sources such as HDP with other data sources using Alteryx gives you a deeper understanding of the reason for a trend and helps predict future outcomes. Alteryx also offers a range of packaged industry-specific analytic solutions that enable organizations to overlay internal data with U.S. census data and syndicated data from dozens of providers, including Dun & Bradstreet, Experian, TomTom, and many others.

Analyze without Complexity

Deploying and using Big Data analytic solutions can be a daunting task. While Apache Hadoop is extremely powerful, it is also a very sophisticated and comprehensive framework. For many organizations, the complexity of integrating multiple Hadoop components with each other, as well as with the existing data architecture, can be a significant challenge.

A global communication service provider has deployed Alteryx to blend unstructured cellular network data with structured customer profile data, providing a geospatial view of how specific network problem spots impact customer churn. Using predictive analytics to anticipate which customers are most likely to churn, the service provider can take proactive preventative action.

Using Hadoop with traditional business intelligence tools and specialty Big Data analytics tools can be equally challenging. The traditional enterprise business intelligence platforms found in most medium and large organizations are primarily designed for highly trained IT professionals tasked with developing and maintaining high volume, standardized production reports. What's more, a vast majority of specialized Big Data analytic tools can only access Big Data sources and can only be used by data scientists with advanced training in statistics and computer science.

Together, Alteryx and Hortonworks dramatically simplify Hadoop-based analytics. Hortonworks eliminates the barriers to Hadoop adoption by providing the only 100% open source platform for Apache Hadoop that is easy to deploy and integrate. Its Hadoop Data Platform (HDP) is a pre-integrated package of essential Hadoop components that combines ease of installation, configuration, and management with the scalability and reliability required for enterprise deployments.

Alteryx represents a new generation of analytic platforms built specifically for business professionals like you who require the ability to access, analyze, and consume Big Data with agility—and without complexity. Using Alteryx, you can deliver powerful analytic solutions that include statistical modeling, predictive analysis, and spatial analysis without reliance on data scientists or IT specialists. A single point-and-click workflow enables you to access, blend, and analyze Big Data before publishing for use by business decision-makers.

GO! – Get Started with Hadoop-based Analytics Now

For those of you who are ready—now—to make the transition from reading about Hadoop to gaining hands-on exposure, Hortonworks offers a personal Apache Hadoop solution and learning platform in one convenient package. Available as a free download, the **Hortonworks Sandbox** includes a complete, self-contained virtual machine with Apache Hadoop pre-configured, along with step-by-step hands-on tutorials, demos, and videos.

If you have already deployed Hadoop, Alteryx offers several tools to help you get started leveraging Big Data in your organization. The **Alteryx Analytics Gallery** is the industry's first analytics cloud platform that delivers a consumer-oriented experience to business users. With it, you can consume, share, and publish applications through a highly intuitive, social enterprise environment. The gallery includes an extensive range of industry and special-purpose analytic applications for free public browsing.

For more guidance on putting Big Data into practice, check out **Big Data Analytics for Dummies – Alteryx Special Edition**, which demonstrates how to maximize the value from Big Data by leveraging analytic applications, as well as how to improve decision-making by combining Big Data with sophisticated predictive and spatial analytics.

About Hortonworks

Hortonworks develops, distributes and supports the only 100% open source distribution of Apache Hadoop explicitly architected, built and tested for enterprise grade deployments. Formed by the original architects, builders and operators of Hadoop, Hortonworks stewards the core and delivers the critical services required by the enterprise to reliably and effectively run Hadoop at scale.

Hortonworks

3460 West Bayshore Road
Palo Alto, CA 94303
USA: (855) 8-HORTON
Intl: (408) 916-4121
www.hortonworks.com

About Alteryx

Alteryx provides indispensable analytic solutions for enterprise companies making critical decisions about how to expand and grow. Our product, *Alteryx Strategic Analytics*, is a desktop-to-cloud Agile BI and analytics solution designed for Data Artisans and business leaders that brings together the market knowledge, location insight, and business intelligence today's organizations require. For more than a decade, Alteryx has enabled strategic planning executives to identify and seize market opportunities, outsmart their competitors, and drive more revenue. Customers like Experian Marketing Services and McDonald's rely on Alteryx daily for their most important decisions. Headquartered in Irvine, California, and with offices in Boulder and Silicon Valley, Alteryx empowers 250+ customers and 200,000+ users worldwide. Get inspired today at www.alteryx.com or call 1-888-836-4274.

alteryx

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

Conclusion

Organizations today blend large and small data sets, from internal and external sources and in structured and unstructured formats, to gain new insights. But harnessing Big Data to extract real business value requires more than simply collecting and blending the raw data. And it requires more than allowing a select group of highly trained data scientists to access the data for analysis.

By 'humanizing' Big Data and giving a broader community of business analysts and decision-makers access to this vast, untapped mine of business information, organizations can uncover trends, discover new sources of revenue, and pinpoint areas of improvement—all with the goal of improving competitive advantage.

Business analysts who can link the technological capabilities of Hadoop-based analytics with specific business applications will be well-positioned to drive the transformation to Thomas Davenport's Analytics 3.0 world in their organizations.

© 2013 Alteryx, Inc. Alteryx is a registered trademark of Alteryx, Inc. 5/13