

Apache Hadoop Basics

May 2013



© 2013 Hortonworks Inc. http://www.hortonworks.com

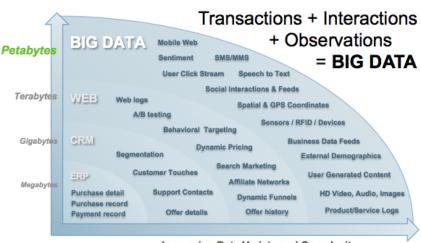


"Big Data"

A "big" shift is occurring. Today, the enterprise collects more data than ever before, from a wide variety of sources and in a wide variety of formats. Along with traditional transactional and analytics data stores, we now collect additional data across social media activity, web server log files, financial transactions and sensor data from equipment in the field. Deriving meaning from all this data using conventional database technology and analytics tools would be impractical. A new set of technologies has enabled this shift.

Now an extremely popular term, "big data" technology seeks to transform all this raw data into meaningful and actionable insights for the enterprise. In fact, big data is about more than just the "bigness" of the data. Its key characteristics, coined by industry analysts are the "Three V's," which include volume (size) as well as velocity (speed) and variety (type). While these terms are effective descriptors, big data also needs to provide value and we often find that as the intersection of transactions, interactions and observations.

Each of these data source introduces its own set of complexities to data processing; combined they can easily push a given application or data set beyond the pale of traditional data processing tools, methods or systems. In these cases, a new approach is required. Big data represents just that.



Increasing Data Variety and Complexity

At its core then, big data is the collection, management and manipulation of data using a new generation of technologies. Driven by the growing importance of data to the enterprise, big data techniques have become essential to enterprise competitiveness.



Apache Hadoop: Platform for Big Data

Apache Hadoop is a platform that offers an efficient and effective method for storing and processing massive amounts of data. Unlike traditional offerings, Hadoop was designed and built from the ground up to address the requirements and challenges of big data. Hadoop is powerful in its ability to allow businesses to stop worrying about building big-data-capable infrastructure and to focus on what really matters: extracting business value from the data.

Apache Hadoop use cases are many, and show up in many industries, including: risk, fraud and portfolio analysis in financial services; behavior analysis and personalization in retail; social network, relationship and sentiment analysis for marketing; drug interaction modeling and genome data processing in healthcare and life sciences... to name a few.

The first step in harnessing the power of Hadoop for your business is to understand the basics: what Hadoop is, where it comes from, how it can be applied to your business processes, and how to get started using it.

The Early Days of Hadoop at Yahoo!

Hadoop has its roots at Yahoo!, whose Internet search engine business required the continuous processing of large amounts of Web page data. In 2005 Eric Baldeschwieler (aka "E14" and Hortonworks CTO) challenged Owen O'Malley (Hortonworks co-founder) and several others to solve a really hard problem: store and process the data on the internet in a simple, scalable and economically feasible way. They looked at traditional storage approaches but quickly realized they just weren't going to work for the type of data (much of it unstructured) and the sheer quantity Yahoo! would have to deal with.

The team designed and prototype a new framework for Yahoo! Search. This framework, called Dreadnaught, was implemented in the C++ programming language and modeled after research published by Google describing its Google File System and a distributed processing algorithm called MapReduce.

At the same time, an open-source search engine project called Nutch, lead by Doug Cutting and Mike Carafella, had implemented similar functions using Java. By the end of 2005, Doug and Mike had the Java version working on small cluster of 20 nodes. Impressed by their progress,



Yahoo! Hired Cutting in January 2006, and merged the two teams together, choosing the Java version of MapReduce because it had some of the search functions already built out. With an expanded team including Owen O'Malley and Arun Murthy, the code base of this distributed processing system tripled by mid-year. Soon, a 200-node research cluster was up and running and the first real users were getting started. Apache Hadoop had liftoff.

The Power of Open Source

Apache Hadoop is an open source project governed by the Apache Software Foundation (ASF). As part of the ASF, development of will remain open and transparent for all future users and more importantly freely available.

Eric and team realized that with a community of like-minded individuals, Hadoop would innovate far faster. At the same time, they'd enable other organizations to realize some of the same benefits that they were starting to see from their early efforts. When organizations such as Facebook, LinkedIn, eBay, Powerset, Quantcast and others began picking up Hadoop and innovating in areas beyond the initial focus, it reinforced the fact that the choice of community driven open source was the right one.

A case in point being when a small startup (Powerset) started working on a project to support tables on HDFS inspired by Google's BigTable paper; that effort turned into what's now Apache HBase! Further... Facebook started an effort to build a SQL layer on top of MapReduce, which became Apache Hive!

Community-driven open source will always outpace the innovation of a single group of people or single company.

It has proven time and again when it comes to platform technologies like Hadoop that community-driven open source will always outpace the innovation of a single group of people or single company.

Commercial Adoption

Apache Hadoop became a foundational technology at Yahoo, underlying a wide range of business-critical applications. Companies in nearly every vertical started to adopt Hadoop. By 2010, the community had grown to thousands of users and broad enterprise momentum had been established.



After many years architecting and operating the Hadoop infrastructure at Yahoo! and contributing heavily to the open source community, E14 and 20+ Hadoop architects and engineers spun out of Yahoo! to form Hortonworks in 2011. Having seen what it could do for Yahoo, Facebook, eBay, LinkedIn and others, their singular objective is to focus on making Apache Hadoop into a platform that is easy to use and consume by the broader market of enterprise customers and partners.

Understanding Hadoop and Related Services

At its core, Apache Hadoop is a framework for scalable and reliable distributed data storage and processing. It allows for the processing of large data sets across clusters of computers using a simple programming model. It is designed to scale up from single servers to thousands of machines, aggregating the local computation and storage from each server. Rather than relying on expensive hardware, the Hadoop software detects and handles any failures that occur, allowing it to achieve high availability on top of inexpensive commodity computers, each individually prone to failure.

At the core of Apache Hadoop are the Hadoop Distributed File System, or HDFS, and Hadoop MapReduce, which provides a framework for distributed processing.

HDFS

The Hadoop Distributed File System is a scalable and reliable distributed storage system that aggregates the storage of every node in a Hadoop cluster into a single global file system. HDFS stores individual files in large blocks, allowing it to efficiently store very large or numerous files across multiple machines and access individual chunks of data in parallel, without needing to read the entire file into a single computer's memory. Reliability is achieved by replicating the data across multiple hosts, with each block of data being stored, by default, on three separate computers. If an individual node fails, the data remains available and an additional copy of any blocks it holds may be made on new machines to protect against future failures.

This approach allows HDFS to dependably store massive amounts of data. For instance, in late 2012, the Apache Hadoop clusters at Yahoo had grown to hold over 350 petabytes (PB) of data across 40,000+ servers. Once data has been loaded into HDFS, we can begin to process it with MapReduce.



MapReduce

MapReduce is the programming model that allows Hadoop to efficiently process large amounts of data. MapReduce breaks large data processing problems into multiple steps, namely a set of Maps and Reduces, that can each be worked on at the same time (in parallel) on multiple computers.

MapReduce is designed to work with of HDFS. Apache Hadoop automatically optimizes the execution of MapReduce programs so that a given Map or Reduce step is run on the HDFS node that contains locally the blocks of data required to complete the step. Explaining the MapReduce algorithm in a few words can be difficult, but we provide an example in Appendix A for the curious or technically inclined. For the rest, suffice it to say that MapReduce has proven itself in its ability to allow data processing problems that once required many hours to complete on very expensive computers to be written as programs that run in minutes on a handful of rather inexpensive machines. And, while MapReduce can require a shift in thinking on the part of developers, many problems not traditionally solved using the method are easily expressed as MapReduce programs.

Hadoop 2.0: YARN

As previously noted, Hadoop was initially adopted by many of the large web properties and was designed to meet their needs for large web-scale batch type processing. As clusters grew and adoption expanded, so did the number of ways that users wanted to interact with the data stored in Hadoop. As with any successful open-source project, the broader ecosystem of Hadoop users responded by contributing additional capabilities to the Hadoop community, with some of the most popular examples being Apache Hive for SQL-based querying, Apache Pig for scripted data processing and Apache HBase as a NoSQL database.

These additional open source projects opened the door for a much richer set of applications to be built on top of Hadoop – but they didn't really address the design limitations inherent in Hadoop; specifically, that it was designed as a single application system with MapReduce at the core (i.e. batch-oriented data processing). Today, enterprise applications want to interact with Hadoop in a host of different ways: batch, interactive, analyzing data streams as they arrive, and more. And most importantly, they need to be able to do this all simultaneously without any single application or query consuming all of the resources of the cluster to do so. Enter YARN.



YARN is a key piece of Hadoop version 2 which is currently under development in the community. It provides a resource management framework for any processing engine, including MapReduce. It allows new processing frameworks to use the power of the distributed scale of Haodop. It transforms Hadoop from a single application to a multi application data system. It is the future of Hadoop.

The Hadoop Project Ecosystem

As Apache Hadoop has matured, a number of important tools have been built to support it.

These tools may be categorized into Hadoop Data Services and Hadoop Operational Services based on the functionality they offer.

Hadoop Data Services

Hadoop Data Services are tools that allow users to more easily manipulate and process data. They include the following:

Apache Hive

Apache Hive is data warehouse infrastructure built on top of Hadoop for providing data summarization, ad-hoc query, and the analysis of large datasets. It provides a mechanism for imparting structure onto the data in Hadoop and for querying that data using a SQL-like language called HiveQL (HQL). Hive eases integration between Hadoop and various business intelligence and visualization tools.

Apache Pig

Apache Pig allows you to write complex map reduce transformations using a simple scripting language called Pig Latin. Pig Latin defines a set of transformations on a data set such as aggregate, join and sort. Pig translates the Pig Latin into Hadoop MapReduce so that it can be executed within HDFS.

Apache HCatalog

HCatalog is a table and storage management layer for Hadoop that enables users with different data processing tools – including Hive, Pig and MapReduce – to more easily read and write data. HCatalog's table abstraction presents users with a relational view of data in HDFS and ensures that users need not worry about where or in what format their data is stored.

Apache HBase

Apache HBase is a non-relational database that runs on top of HDFS. It provides fault-



tolerant storage and quick access to large quantities of sparse data. It also adds transactional capabilities to Hadoop, allowing users to conduct updates, inserts and deletes.

Apache Sqoop

Apache Sqoop is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases. Sqoop imports data from external sources either directly into HDFS or into systems like Hive and HBase. Sqoop can also be used to extract data from Hadoop and export it to external repositories such as relational databases and data warehouses.

Apache Flume

Apache Flume is a service for efficiently collecting, aggregating, and moving large amounts of log data. Flume allows log data from many different sources, such as web servers, to be easily stored in a centralized HDFS repository.

Hadoop Operational Services

Distributed computing at scale can present significant operational challenges. Several projects have emerged to aid in the operations and management of a Hadoop cluster. These include:

Apache Ambari

Apache Ambari provides an intuitive set of tools to monitor, manage and efficiently provision an Apache Hadoop cluster. Ambari simplifies the operation and hides the complexity of Hadoop, making Hadoop work as a single, cohesive data platform.

Apache Oozie

Apache Oozie is a Java web application used to schedule Apache Hadoop jobs. Oozie combines multiple jobs sequentially into one logical unit of work. It supports Hadoop jobs for MapReduce, Pig, Hive, and Sqoop.

Apache ZooKeeper

Apache Zookeeper provides operational services for a Hadoop cluster, including a distributed configuration service, a synchronization service and a naming registry. These services allow Hadoop's distributed processes to more easily coordinate with one another.

Hadoop Distributions

Apache Hadoop is open source. Any interested party can visit the Hadoop web site to download the project's source code or binaries directly from the ASF.



Downloading directly, however, presents a number of challenges to enterprises that want to get started quickly. Modern Hadoop deployments require not only MapReduce and HDFS, but many—often all—of the supporting projects already discussed, along with a variety of other open source software. Assuming the enterprise even knows which software is required, this diversity adds complexity to the process of obtaining, installing and maintaining an Apache Hadoop deployment. Each of the aforementioned projects is developed independently, with its own release cycle and versioning scheme. Not all versions of all projects are API compatible, so care must be taken to choose versions of each project that are known to work together. Achieving production-level stability for a given set of projects presents even more complexity. To ensure stability, an enterprise must attempt to determine which versions of which projects have been tested together, by whom, and under what conditions—information that is not often readily available.

How then can an enterprise ensure that its first steps with Hadoop go smoothly, that the software they download is internally compatible and completely stable, and that support will be available when needed? The answer to these questions is a Hadoop *Distribution*, which provides an integrated, pre-packaged bundle of

A Hadoop Distribution provides an integrated, pre-packaged bundle of software that includes all the required components and supporting project software from the community.

software that includes all required Hadoop components, related projects, and supporting software. Enterprise-focused Hadoop distributions integrate pre-certified and pre-tested versions of Apache Hadoop and related projects and make them available in a single easily installed and managed download.

The Hortonworks Data Platform (HDP) is one such distribution of Apache Hadoop. In addition to providing all of the essential components for an enterprise Hadoop deployment, HDP is supported by the team that created Hadoop and built and managed the largest production Hadoop deployment in the world at Yahoo!

Hadoop in the Enterprise

Now that we understand Hadoop, we can better understand the role that it plays within the enterprise analytics landscape. Today, most enterprises utilize one or more analytical applications to manage the day-to-day business. The traditional approach to analytics is relatively well understood and is depicted in Figure 1. In this approach:



- Data comes from a set of data sources; typically from enterprise applications such as ERP, CRM, or other transactional applications that power the business
- That data is extracted, transformed, and loaded into a data repository such a relational database or a data warehouse
- A set of analytical applications either packaged (e.g. SAS) or custom are then used to manipulate the data in the repository, producing reports or visualizations that deliver insights to business users

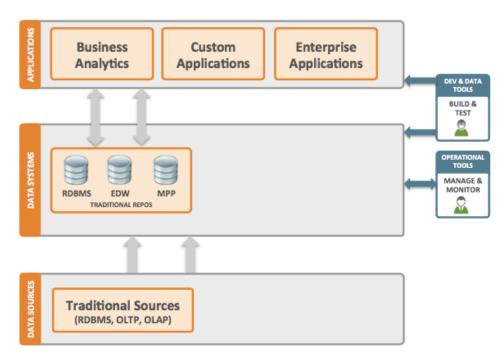


Figure 1: Traditional Enterprise Data Architecture

With the introduction of new data sources such as web logs, email, social media and sensors, enterprises are forced to think differently. The emerging data architecture most commonly seen introduces Apache Hadoop to handle these new types of data in an efficient and cost-effective manner. Hadoop does not replace the traditional data repositories used in the enterprise, but rather is a complement.



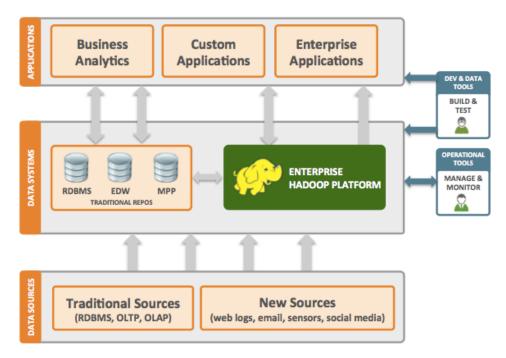
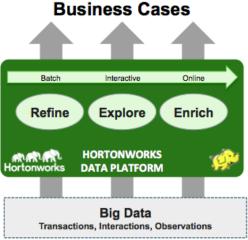


Figure 2: The Emerging Big Data Architecture

With the availability of enterprise-ready Apache Hadoop distributions like Hortonworks Data Platform, enterprises are embarking on a wide variety of Hadoop-based big data projects. While examples are many, three patterns of use have emerged:

- Refine. The capturing of a variety of data sources into a single platform where value can be added by refining the data into formats that are more easily consumed by platforms such as a data warehouse
- Explore. Interactively examining huge volumes of data to identify patters and relationships in order to unlock opportunities for business value
- Enrich. Enabling organizations to apply advanced analytics to the data they are
 collecting via log files or social media streams in order to make other applications, such
 as mobile commerce applications, more "intelligent" with respect to the experience they
 deliver





As the amount of data grows, Hadoop becomes more and more vital to enabling these activities, which power common uses cases including:

- Social/Mobile Data Analysis. Analyzing the data collected via social media and mobile applications, including Twitter, Facebook, LinkedIn, GPS services, etc.
- Log Analysis. Turning log files into useful information, for example, to troubleshoot IT systems, track compliance with government regulations and corporate security policies, support forensic data analysis, and more.
- Clickstream Analysis. Analyzing user clicks while web browsing or using software applications for such business purposes as testing software, conducting market research, optimizing employee productivity, etc.
- Sentiment Analysis. Enabling computers to determine the attitudes, preferences, etc. of
 individuals and groups by using natural language processing, computational linguistics,
 and text analytics to extract information from written materials.

Hortonworks Data Platform

Hortonworks Data Platform (HDP) is the only 100% open source data management platform for Apache Hadoop. HDP allows enterprises to capture, process and share data in any format and at full scale. Built and packaged by the core architects, builders and operators of Hadoop, HDP includes all of the necessary components to manage a cluster at scale and uncover business insights from existing and new big data sources.

Hortonworks Data Platform is the most stable and reliable Apache Hadoop distribution available. It delivers the advanced services required for enterprise deployments without compromising the cost-effective and open nature of Apache Hadoop, including:

- Data Services required to store, analyze and access data
- Operational Services required to manage and operate Hadoop
- Platform Services such as high availability and snapshots which are required to make
 Hadoop enterprise grade

The Hortonworks Approach

At Hortonworks, We believe the most effective path is to do our work within the open source community, introduce enterprise feature requirements into that public domain, and to work diligently to progress existing open source projects and incubate new projects to meet those needs. And through the application of enterprise rigor to the build, test and release process we



can provide a 100% open source distribution of Apache Hadoop that is truly enterprise grade.

Our approach to creating a Hadoop distribution follows these three key steps:

- 1. Identify and introduce enterprise requirements into the public domain

 To help us determine where to focus efforts, we spend time working within the Hadoop community and with our customers to understand the requirements for broader enterprise adoption. We combine this with our unrivaled Hadoop experience to connect the voice of the enterprise with the community forum in order to help define and float the "right" requirements into the Hadoop ecosystem.
- 2. Work with the community to advance and incubate open source projects
 Community driven open source has proven to be the fastest path for Hadoop innovation
 and we are committed to this approach. Once Enterprise requirements are identified,
 we work diligently with others in the community to address them: either within existing
 projects or by incubating new projects in the Apache Software Foundation. And we
 commit everything back into the open community, with nothing held back.
- 3. Apply Enterprise Rigor to provide the most stable and reliable distribution We play key roles in the test and release process for the open source Hadoop projects. We also take great pains to test and certify a consolidated distribution – the Hortonworks Data Platform – on large and complex clusters running across a range of operating platforms. By the time HDP sees any customer environment it has been validated at Yahoo!, which has the richest test suite for Hadoop on the planet.

Broad Deployment Options

Hortonworks Data Platform is also the only Apache Hadoop distribution available for Windows Server. HDP can be deployed in the cloud with Rackspace, Microsoft Azure and OpenStack. It can be virtualized or even delivered in an appliance with Teradata. With these options, clients are given the flexibility to deploy as need be, to migrate between on- and off-premise, and to be free from the burden of lock-in.

Interoperable & Ecosystem Friendly

Apache Hadoop has become a core component of many enterprises' data architectures, as a complement to existing data management systems. Accordingly, HDP is architected to easily interoperate so enterprises can extend existing investments in applications, tools and processes



with Hadoop. This approach encourages a thriving ecosystem of big-data-driven solutions and leaders such as Microsoft and Teradata have looked to Hortonworks to help deliver Hadoop-based applications.

The Future of Hadoop

Hadoop has "crossed the chasm" from a framework for early adopters, developers and technology enthusiasts to a strategic data platform embraced by innovative CTOs and CIOs across mainstream enterprises. These people, who want to improve the performance of their companies and unlock new business opportunities, realize that including Apache Hadoop as a deeply integrated supplement to their current data architecture offers the fastest path to reaching their goals while maximizing their existing investments.

Going forward, Hortonworks and the Apache Hadoop community will continue to focus on increasing the ease with which enterprises deploy and use Hadoop, and on increasing the platform's interoperability with the broader data ecosystem. This includes making certain it is reliable and stable and more importantly, ready for all and any enterprise workloads.

Next Steps

To learn more about Hortonworks and Hortonworks Data Platform, visit us on the Web at http://www.hortonworks.com.



Appendix A - MapReduce Example

As you might guess from the name MapReduce, the first of these steps is the Map; the next is the Reduce. To understand how Map and Reduce work, consider the following example, based on an extended explanation by <u>Steve Krenzel</u>¹.

Let's say a social media site wants to calculate every member's common friends once a day and store those results. As is typical for inputs to and outputs from MapReduce tasks, friends are stored as a key-value pair expressed in the form Member->[List of Friends]:

A -> B C D

 $B \rightarrow ACDE$

C -> A B D E

. . .

The first step in calculating the members' common friends using MapReduce is the Map, in which each of the friend lists above gets mapped to a new key-value pair. In this case, the mappers sort and group the input so that a new key is output for every combination of two members, whose value is simply a list of the first member's friends followed by a list of the second member's friends. In our example, (A B) forms a new key whose value is a list of A's friends (BCD) and B's friends (ACDE):

$$(A B) \rightarrow (B C D) (A C D E)$$

. . .

All these intermediate key-value pairs are then sent to the Reduce step, which simply outputs for each key any friends in both members' friend lists; in this case resulting in:

. . .

which says that members A and B have C and D as common friends.

Without MapReduce, this process would require a slow, serial process of comparing A's friends with B's friends, A friends with C's friends, A's friends with D's friends, etc., for every one of A's friends. And so on, for every member of the social media site. Easy when you're just starting out and have five members, but the traditional approach quickly becomes untenable when membership gets into the millions.

With MapReduce and HDFS, the map and reduce tasks are distributed across the Hadoop cluster and many tasks run simultaneously across the cluster. Because of the data locality

http://stevekrenzel.com/finding-friends-with-mapreduce



property of Hadoop, the map tasks each run on a node that contains the blocks of input data (lines in the Friends List file) that they will operate on.

Hortonworks, Hadoop and You

We encourage you to follow us, get engaged with our learning tools, or download the HDP Sandbox, a single node installation of HDP that can run right on your laptop. Hadoop has the potential to have a profound impact on the data landscape, and by understanding the basics, you can greatly reduce the complexity.

Download the Hortonworks Sandbox to get started with Hadoop today

About Hortonworks

Hortonworks is a leading commercial vendor of Apache Hadoop, the preeminent open source platform for storing, managing and analyzing big data. Our distribution, Hortonworks Data Platform powered by Apache Hadoop, provides an open and stable foundation for enterprises and a growing ecosystem to build and deploy big data solutions. Hortonworks is the trusted source for information on Hadoop, and together with the Apache community, Hortonworks is making Hadoop more robust and easier to install, manage and use. Hortonworks provides unmatched technical support, training and certification programs for enterprises, systems integrators and technology vendors.

