Hortonworks Architecting the Future of Big Data

Whitepaper

Apache Hadoop: The *Big Data* Refinery

Introduction

"Big data" has become an extremely popular term, due to the well-documented explosion in the amount of data being stored and processed by today's businesses. Big data is about more than just the "bigness" of the data, however. Data *volume* is in fact only the first of several defining characteristics of big data. Other characteristics, coined by industry analysts as the "Four V's," include *velocity*—the speed at which the data must be processed; *variety* – the different types and sources of data that must be analyzed and the complexity of each and the whole; and *variability* – referring to the inherent "fuzziness" of the data, in terms of its meaning or context.

Each of these characteristics introduces its own set of complexities to data processing. When one of these characteristics is present, a given data set may fall beyond the pale of traditional data processing tools, methods or systems. When more than one is present, a whole new way of looking at things is required.

Apache Hadoop is a data processing platform offering just that – a new way of looking at things – designed and built from the ground up for big data. The power of Apache Hadoop is its ability to allow businesses to stop worrying about building big-data-capable infrastructure and focus on what matters: extracting business value from the data.

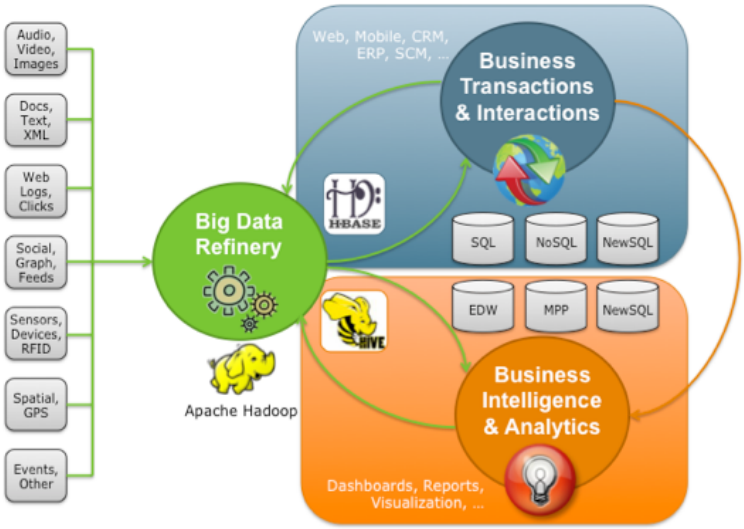
Capturing the business value of big data requires transforming raw data into usable products that offer new insights to the business. In this way, harnessing big data is very much like the capturing the value in a natural resource like petroleum, turning it into usable end products like fuel oil, gasoline and lubricants.

In the oil and gas world, this process is called refinement and the system that does the refinement is the refinery. Likewise, in the big data world, we can think of the role Apache Hadoop plays as that of a *data refinery*.

In this white paper, targeted at technical executives and practitioners new to big data, we expand upon the data refinery analogy to explore the key characteristics and capabilities of Apache Hadoop and the various tools in its ecosystem.

Enter the Data Refinery

To an outsider driving past, an oil refinery looks like a tangled mess of pipes flowing to and from various undifferentiated structures. To the uninitiated, the Apache Hadoop platform appears equally opaque. Just as a process engineer might lead a refinery tour by discussing the refinery's inputs, outputs, processes and major structures, we'll begin down the path towards understanding Hadoop with a similar tour of its components and ecosystem.



Inputs

It turns out that petroleum, the input to an oil refinery, and raw data, the input to a data refinery, have a lot in common. Both come in many different varieties, are typically found in large reservoirs, can leak if not handled correctly, and can be difficult to extract at the source.

Data within the enterprise is widely varying in source and type. Depending on your enterprise and use scenarios, your raw data may originate as log files, sensor readings, transactional data, stored social interactions, or downloaded data sets, among other possibilities.

One distinction between different types of data is whether that data is *structured*, such as database records, *unstructured*, such as documents, or *semi-structured*, such as log data stored within text files. A key distinction between today's data refinery and previous generations of business intelligence solutions is the growing concern with extracting meaning from unstructured and semi-structured data. Apache Hadoop is unique in its ability to simultaneously process and analyze complex and disparate types of data. Indeed, this has become one of the principal functions of Hadoop-based data refineries.

Outputs

A single oil refinery is capable of producing many different petroleum products. Similarly, Apache Hadoop has proven its utility in a wide range of applications including log file and clickstream analysis, business and marketing analytics, text processing, social network analysis, complex data mining, and many more. A single Hadoop deployment, fed the needed input data, can simultaneously support each of these different analyses.

Data refined by Hadoop can be presented directly to end users in the form of text-based reports, or via a visual reporting tool. It can also be loaded into relational databases and data warehouses, or otherwise made available to downstream applications.

While Hadoop is traditionally operated in a batch-oriented mode, users are increasingly looking to Hadoop to power real-time applications, where analyses are performed on a continuous basis over incoming data, and streamed to waiting applications and dashboards.

Extraction and Loading

Enterprise data is typically stored in far-flung repositories, just as petroleum may be found in reservoirs around the world. An important role played by the Apache Hadoop data refinery is, in fact, to synthesize disparate data sources of marginal value, independent of origin and type, into formats that provide higher value to the business.

In order to refine data, that data must be loaded into Hadoop. Just as oil drills and pipelines are not part of the oil refinery, extracting data from existing repositories is not, strictly speaking, a function of the data refinery. Yet, many tools exist to aid in this task. Due to Hadoop's growing popularity, many existing databases and ETL[1] tools already have the ability to add data to Hadoop.

Storage

The ability to store vast amounts of raw crude, as well as intermediate and output products, is a key component of an oil refining operation. To meet this need, an oil depot or tank farm is typically located at or near an oil refinery. In like manner, storage is a key component of a data refinery.

The Apache Hadoop storage subsystem, Hadoop Distributed File System (HDFS), is one of the two main components of the platform. The distributed nature of HDFS is its most important characteristic: files stored in HDFS are broken into blocks, which are dispersed among the nodes in the system. This allows a single file to be processed in parallel, as we'll discuss below, and also allows the system's storage capacity to be easily scaled by adding additional nodes.

An additional advantage of the HDFS approach is high availability, which is achieved by replicating each block across multiple nodes. This allows Hadoop systems to be dependably built using inexpensive, commodity grade servers. Because of the resulting low cost-per-gigabyte of HDFS-based storage, Hadoop is increasingly being used for long-term data archival.

As its name implies, HDFS is just that, a file system. For structured data sets, *Apache HBase* is often appropriate. HBase, modeled after an internal Google system called BigTable, sits on top of HDFS, turning it into a distributed non-relational database.

Processing

Apache Hadoop's robust processing capabilities are based on MapReduce, a framework for performing highly parallelized processing of huge datasets, using a large number of compute nodes. MapReduce was initially popularized by Google, where it was used to generate the company's vast search index.

MapReduce programs refine data by manipulating it in a series of "map" and "reduce" steps, just as a series of individual steps like distillation, coking and reforming are used in the process of refining oil. In MapReduce, many instances of "map" steps process individual blocks of an input file to produce one or more outputs; these outputs are passed to "reduce" steps where they are combined to produce a single end result.

Hadoop MapReduce is Hadoop's implementation of the MapReduce framework and the second main component of Apache Hadoop. Hadoop MapReduce is unique in that it distributes data processing across a large set of server nodes and dynamically routing processing work to those nodes containing the blocks of data that need to be processed.

Unlike an oil refinery, Hadoop preserves the original raw material. This means that you can continue to

[1] Extract, Transform and Load tools are used to load data into a data warehouse.

improve and optimize your refinement process against the original source data in HDFS, running and tweaking your MapReduce jobs, over and over again, until the needed results are achieved.

MapReduce programs, while powerful, can require non-traditional thinking to build. A number of Hadoop ecosystem projects have emerged seeking to provide more familiar paradigms for taking advantage of Hadoop. *Apache Pig* allows developers to write data analysis programs in a high-level language, called Pig Latin, custom made for exploring very large datasets. *Apache Hive* presents a data warehouse interface on top of Hadoop, allowing users to issue queries in HiveQL, a SQL-like query language.

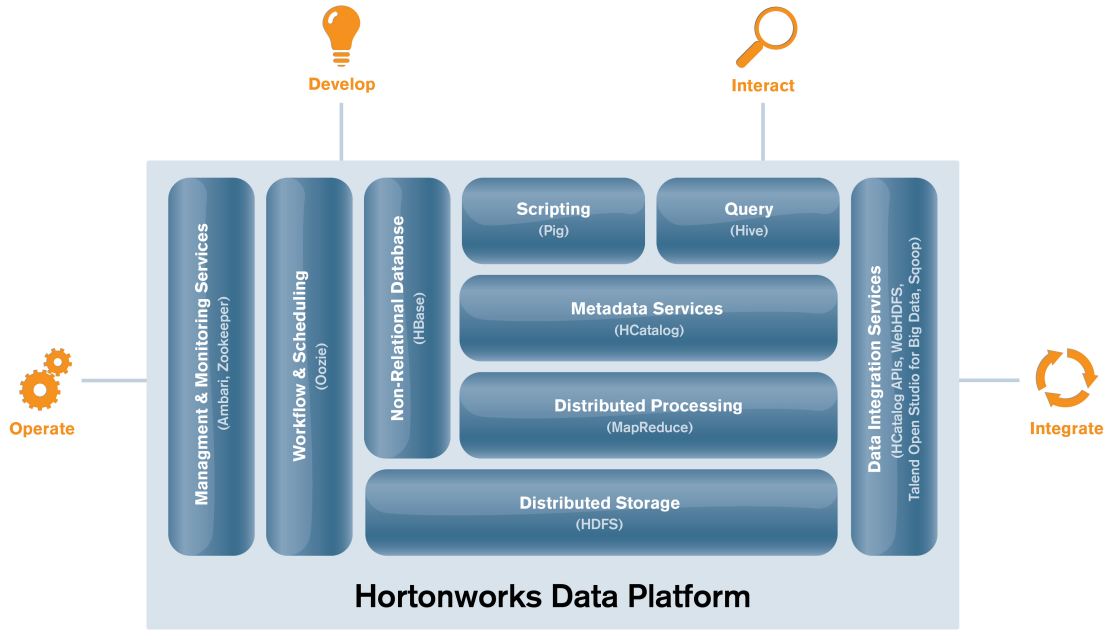
Real-world data analysis applications often require complex sequences of individual tasks. *Apache Oozie*, a workflow coordination system for managing Hadoop jobs, is integrated with the rest of the Hadoop stack and supports several job types out of the box, including MapReduce and Pig.

The Hortonworks Data Platform

As we've seen through our tour, Apache Hadoop can be the basis for an extremely powerful and flexible data refinery. And, due to the wonder that is open source software, this power and flexibility is freely available to all. For many organizations, however, the complexity of integrating the various Hadoop components with one another, and with existing data architectures, represents a hidden cost and a barrier to successful adoption.

Hortonworks was formed to eliminate the barriers to Hadoop adoption, making the Hadoop platform easier to use and consume, and making its benefits more accessible to the entire ecosystem, including end user organizations, hardware and software vendors, and consulting and systems integration providers.

To further this mission, Hortonworks offers the **Hortonworks Data Platform**, a pre-integrated package of essential Apache Hadoop components, which allows users to easily harness the power of Hadoop and maximize the value of all of their data.



Hortonworks Data Platform delivers, in a single, tightly integrated package, popular Apache Hadoop projects such as HDFS, MapReduce, Pig, Hive, HBase and Zookeeper. To this base, Hortonworks Data Platform includes additional open source technologies that make the Hadoop platform more manageable, open, and extensible. A complete set of open APIs is provided, making it easier for enterprises and ISVs to integrate and extend Apache Hadoop.

Making Hadoop accessible begins with installation and configuration. Typically a laborious task, installation and configuration of Hadoop is made all the more complex by the fact that the open source projects that make up the Hadoop platform are independently developed and frequently updated codebases, each with their own release schedules, versions and dependencies.

To ensure a consistent and stable platform for enterprise use, Hortonworks Data Platform includes only stable component versions that have been fully integrated, tested and certified as part of Hortonworks' extensive Q/A process, and are supported by the company's multi-year support and maintenance policy.

Hortonworks Data Platform supplies installation and configuration tools that make it easy to install, deploy and manage these certified components. The Hortonworks Management Center is based on *Apache Ambari*, an open source installation, configuration and management system for Hadoop, and is included in Hortonworks Data Platform. The Hortonworks Management Center provides a comprehensive web dashboard that integrates monitoring, metrics and alerting information into a unified, Hadoop-specific management console.

Important metadata management functionality is included in Hortonworks Data Platform via an open source project called *Apache HCatalog*. HCatalog provides centralized metadata services, including table and schema management, to all of the platform components. Additionally, it provides a method for deeper integration with third-party data management and analysis tools, improving interoperability.

Beyond technology, as the industry-leading distribution of Apache Hadoop, Hortonworks Data Platform is backed by a powerful ecosystem of partners, including leading software vendors, hardware vendors and systems integrators. These partnerships help ensure that your investment in Hadoop extends and complements existing IT investments and enterprise relationships.

Getting Started With Hadoop

In this white paper, we've introduced the notion of Apache Hadoop as a data refinery and illustrated the analogy with comparisons to an oil refinery. We've used this analogy as the context for an introduction to Hadoop and some of the major projects in the Hadoop ecosystem.

We've also introduced Hortonworks Data Platform; a pre-integrated distribution of Apache Hadoop designed to help you be more successful, more quickly, in your efforts to harness big data.

Extending your knowledge of Hadoop couldn't be easier. The Hortonworks web site is the place to start, offering a wealth of practical educational resources including software downloads, video tutorials and blog posts.

To go further, Hortonworks University is your expert source for Apache Hadoop training and certification. Public and private courses are available for developers, administrators and other IT professionals involved in implementing big data solutions. Training courses combine presentation material with hands-on labs that fully prepare students for real-world Hadoop scenarios. Successfully completing a Hortonworks training course entitles you to sit for the respective Hortonworks certification exam; earning Hortonworks certification identifies you as an expert in the Apache Hadoop ecosystem.

When you're ready to get started building your data refinery, visit the Hortonworks web site for additional resources, including access to the Hortonworks Virtual Sandbox, an online Hadoop environment where you can gain hands-on experience with Hortonworks Data Platform.

For additional assistance, Hortonworks provides expert technical support subscriptions covering the entire Hadoop lifecycle: from development to proof-of-concept to staging and production deployment. Both short-term and long-term support options are available.

In Conclusion

The history of the oil industry is rich with waves of incredible productivity and value creation spurred on by the introduction of new technologies. At Hortonworks we believe that we are at the beginning of an analogous wave in business, centered on Hadoop and its ability to help enterprises harness the power of data. Our company has dedicated itself to unlocking the potential of Hadoop by addressing the technical and knowledge gaps that challenge enterprises as they seek to use it.

Hortonworks was formed by the key architects and developers from the Hadoop engineering team at Yahoo. Our team led the effort to design and build every major release of Apache Hadoop from 0.1 to the current release. Our team continues to be the leading contributor to core Hadoop (MapReduce and HDFS) as well as many other important Hadoop projects, and is the major driving force behind the next generation of Apache Hadoop.

We encourage you and your enterprise to explore the new opportunities that Hadoop presents, and we would love to hear from you as you do so.

About Hortonworks

Hortonworks is accelerating the adoption of Apache Hadoop by bridging the technology and knowledge gaps that exist today. We are driving much of the design and development of both the current and future generations of Apache Hadoop and leveraging our development expertise to provide unmatched expert technical support, training and certification programs for enterprises, systems integrators and ISVs.

455 W. Maude Avenue, Suite 200
Sunnyvale, CA 94085 USA

US: 1.855.846.7866
International: 1.408.916.4121
www.hortonworks.com