Modern Data Architecture with Apache™ Hadoop®

Talend Big Data

Presented by Hortonworks® and Talend

Hortonworks

talend

Executive Summary

Apache™ Hadoop® didn't disrupt the datacenter, the data did.

Shortly after Corporate IT functions within enterprises adopted large-scale systems to manage data, the Enterprise Data Warehouse (EDW) emerged as the logical home of all enterprise data. Today, every enterprise has a data warehouse that serves to model and capture the essence of the business from their enterprise systems.

The explosion of new types of data in recent years – from inputs such as the web and connected devices, or just sheer volumes of records – has put tremendous pressure on the EDW.

In response to this disruption, an increasing number of organizations have turned to Apache Hadoop to help manage the enormous increase in data while maintaining coherence of the data warehouse.

This paper discusses Apache Hadoop and its capabilities as a data platform and how it can integrate with Talend Big Data to deliver integration projects 10 times fast than manually doing MapReduce.

Talend simplifies the integration of big data so you can respond to business demands without having to write or maintain complicated Apache Hadoop code. With Talend Big Data, you can easily integrate all your data sources for use cases including data warehouse optimization, sentiment analysis, web log analysis, predictive analytics, fraud detection or building an enterprise data lake.

An enterprise data lake provides the following core benefits to an enterprise:

New efficiencies for data architecture through a significantly lower cost of storage, and through optimization of data processing workloads such as data transformation and integration.

New opportunities for business through flexible "schema-on-read" access to all enterprise data, and through multi-use and multi-workload data processing on the same sets of data, from batch to real-time.

Apache Hadoop provides these benefits through a technology core comprised of:

Hadoop Distributed Filesystem. HDFS is a Java-based file system that provides scalable and reliable data storage that is designed to span large clusters of commodity servers.

Apache Hadoop YARN. YARN provides a pluggable architecture and resource management for data processing engines to interact with data stored in HDFS.

For an independent analysis of Hortonworks® Data Platform, download Forrester Wave™: Big Data Hadoop Solutions, Q1 2014 from Forrester Research.

The Disruption in the Data

Corporate IT functions within enterprises have been tackling data challenges at scale for many years now. The vast majority of data produced within the enterprise stems from large-scale Enterprise Resource Planning (**ERP**) **systems, Customer Relationship Management** (**CRM**) **systems, and other systems supporting a given enterprise function.** Shortly after these "systems of record" became the way to do business the Data Warehouse emerged as the logical home of data extracted from these systems to unlock "business intelligence" applications, and an industry was born. Today, every organization has data warehouses that serve to model and capture the essence of the business from their enterprise systems.

The Challenge of New Types of Data

The emergence and explosion of new types of data in recent years has put tremendous pressure on all of the data systems within the enterprise. These new types of data stem from "systems of engagement" such as websites, or from the growth in connected devices.

The data from these sources has a number of features that make it a challenge for a data warehouse:

Exponential Growth. An estimated 2.8ZB of data in 2012 is expected to grow to 40ZB by 2020. Eighty-five percent of this data growth is expected to come from new types, with machine-generated data being projected to increase 15x by 2020. (Source: IDC)

Varied Nature. The incoming data can have little or no structure, or structure that changes too frequently for reliable schema creation at time of ingest.

Value at High Volumes. The incoming data can have little or no value as individual or small groups of records. But at high volumes or with a longer historical perspective, data can be inspected for patterns and used for advanced analytic applications.

The Growth of Apache Hadoop

Challenges of capture and storage aside, the blending of existing enterprise data with the value found within these new types of data is being proven by many enterprises across many industries from retail to healthcare, from advertising to energy.

The technology that has emerged as the way to tackle the challenge and realize the value in Big Data is Apache Hadoop, whose momentum was described as "unstoppable" by Forrester Research in the Forrester Wave™: Big Data Hadoop Solutions, Q1 2014.

The maturation of Apache Hadoop in recent years has broadened its capabilities from simple data processing of large data sets to a full-fledged data platform with the necessary services for the enterprise, from security to operations management and more.

Find out more about these new types of data at Hortonworks.com

• Clickstream

• Social Media

• Server Logs

• Geolocation

• Machine and Sensor

What is Hadoop?

Apache Hadoop is an open-source technology born out of the experience of web scale consumer companies such as Yahoo, Facebook and others, who were among the first to confront the need to store and process massive quantities of digital data.

Hadoop and Your Existing Data Systems: A Modern Data Architecture

From an architectural perspective, the use of Hadoop as a complement to existing data systems is extremely compelling: an open source technology designed to run on large numbers of commodity servers. Hadoop provides a low-cost scale-out approach to data storage and processing and is proven to scale to the needs of the very largest web properties in the world.

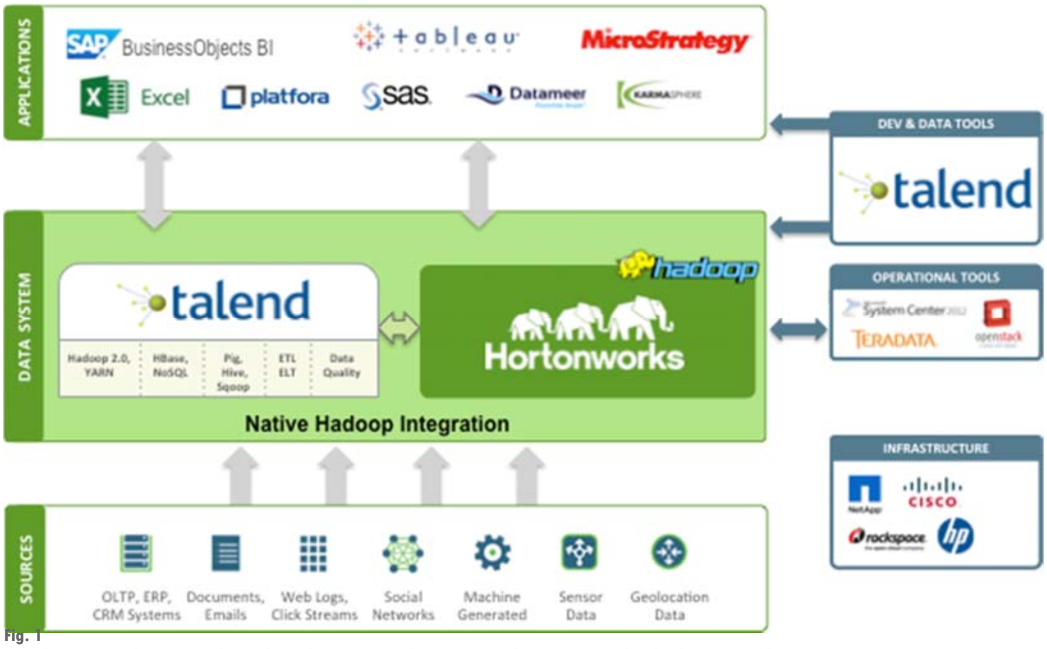

Fig. 1
A Modern Data Architecture with Apache Hadoop integrated into existing data systems with Talend Big Data solutions

Hortonworks is dedicated to enabling Hadoop as a key component of the data center, and having partnered deeply with some of the largest data warehouse vendors we have observed several key opportunities and efficiencies Hadoop brings to the enterprise.

New Opportunities for Analytics

The architecture of Hadoop offers new opportunities for data analytics:

Schema On Read. Unlike an EDW, in which data is transformed into a specified schema when it is loaded into the warehouse – requiring "Schema On Write" – Hadoop empowers users to store data in its raw form and then analysts can create the schema to suit the needs of their application at the time they choose to analyze the data, empowering "Schema On Read." This overcomes issues around the lack of structure and investing in data processing when there is questionable initial value of incoming data.

Multi-use, Multi-workload Data Processing. By supporting multiple access methods (batch, real-time, streaming, in-memory, etc.) to a common data set, Hadoop enables analysts to transform and view data in multiple ways (across various schemas) to obtain closed-loop analytics by bringing time-to-insight closer to real time than ever before.

New Efficiencies for Data Architecture

In addition to the opportunities for Big Data analytics, Hadoop offers efficiencies in a data architecture:

Lower Cost of Storage. By design, Hadoop runs on low-cost commodity servers and direct-attached storage that allows for a dramatically lower overall cost of storage. In particular when compared to high-end Storage Area Networks (SAN) from vendors such as EMC, the option of scale-out commodity compute and storage using Hadoop provides a compelling alternative—and one that allows the user to scale-out their hardware only as their data needs grow. This cost dynamic makes it possible to store, process, analyze, and access more data than ever before.

Data Warehouse Workload Optimization. The scope of tasks being executed by the EDW has grown considerably across ETL, analytics and operations. The ETL function is a relatively low-value computing workload that can be performed on in a much lower-cost manner. Many users offload this function to Hadoop, wherein data is extracted, transformed and then the results are loaded into the data warehouse.

The result: critical CPU cycles and storage space can be freed up from the data warehouse, enabling it to perform the truly high-value functions—analytics and operations—that best leverage its advanced capabilities.

Enterprise Hadoop with Hortonworks Data Platform

To realize the value in your investment in Big Data, use the blueprint for Enterprise Hadoop to integrate with your EDW and related data systems. Building a modern data architecture enables your organization to store and analyze the data most important to your business at massive scale, extract critical business insights from all types of data from any source, and ultimately improve your competitive position in the market and maximize customer loyalty and revenues. Read more at http://hortonworks.com/hdp.

Hortonworks Data Platform is the foundation for a Modern Data Architecture

Hortonworks Data Platform (HDP™) is powered by 100% Open Source Apache Hadoop. HDP provides all of the Apache Hadoop-related projects necessary to integrate Hadoop alongside an EDW as part of a Modern Data Architecture.

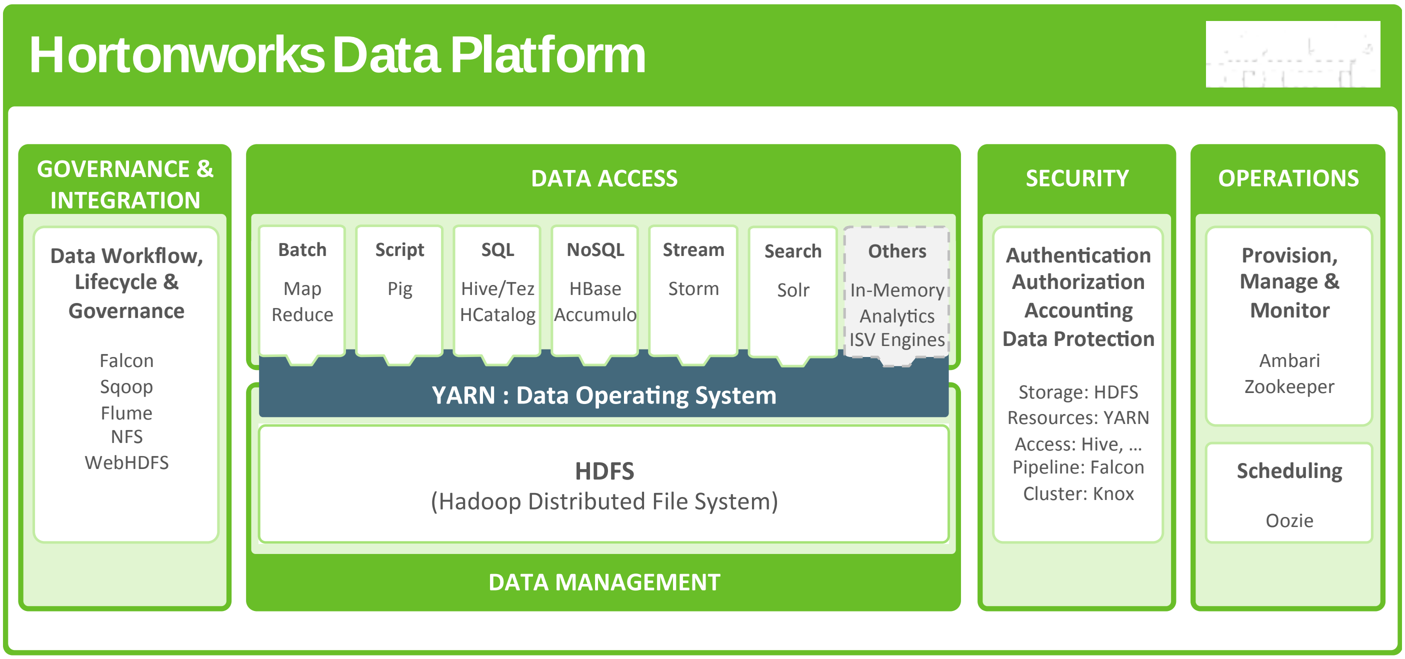

Fig. 12

Data Management: Hadoop Distributed File System (HDFS) is the core technology for the efficient scale-out storage layer, and is designed to run across low-cost commodity hardware. Apache Hadoop YARN is the prerequisite for Enterprise Hadoop as it provides the resource management and pluggable architecture for enabling a wide variety of data access methods to operate on data stored in Hadoop with predictable performance and service levels.

Data Access: Apache Hive is the most widely adopted data access technology, though there are many specialized engines. For instance, Apache Pig provides scripting capabilities, Apache Storm offers real-time processing, Apache HBase offers columnar NoSQL storage and Apache Accumulo offers cell-level access control. All of these engines can work across one set of data and resources thanks to YARN. YARN also provides flexibility for new and emerging data access methods, including search and programming frameworks such as Cascading.

Data Governance & Integration: Apache Falcon provides policy-based workflows for governance, while Apache Flume and Sqoop enable easy data ingestion, as do the NFS and WebHDFS interfaces to HDFS.

Security: Security is provided at every layer of the Hadoop stack from HDFS and YARN to Hive and the other Data Access components on up through the entire perimeter of the cluster via Apache Knox.

Operations: Apache Ambari offers the necessary interface and APIs to provision, manage and monitor Hadoop clusters and integrate with other management console software.

Deployment Options for Hadoop

HDP offers multiple deployment options:

On-premises: HDP is the only Hadoop platform that works across Linux and Windows.

Cloud: HDP can be run as part of IaaS, and also powers Rackspace's Big Data Cloud, and Microsoft's HDInsight Service, CSC and many others.

Appliance: HDP runs on commodity hardware by default, and can also be purchased as an appliance from Teradata.

Talend And Enterprise Hadoop

Talend Big Data generates native and optimized Hadoop code and can load, transform, enrich and cleanse data inside Hadoop for maximum scalability. Its easy-to-use graphical development environment speeds design, deployment and maintenance. Support is provided for simple transformations, advanced transformations and custom transformations. Talend Big Data is the only solution to natively run data quality rules on Hadoop at infinite scale to parse, cleanse and match all of your data.

Features and benefits of Talend:

- 800+ components and connectors to all data sources and applications including big data and NoSQL
- Support for ETL and ELT, real-time delivery and event-driven delivery
- YARN and Hadoop 2.0 support for better resource optimization
- Talend code generation for better scalability and portability
- Visually optimize MapReduce jobs before production for faster development
- A large collaborative community for support

Zero to Big Data in Less than 10 Minutes

The Talend Big Data Sandbox is a ready-to-run virtual environment that includes Talend Big Data Platform, Hortonworks Data Platform and big data examples. Download your free sandbox at http://www.talend.com/talend-big-data-sandbox

Case Study 1:

Maximizing Online Revenue with Talend

A global retailer with 12 billion euros in annual turnover was looking for a way to improve revenue. The firm was experiencing a high rate of shopping cart abandonment and could not quickly adjust prices based on demand, inventory and competition. In the highly competitive online retail sector, buyers can easily compare prices and the competition is just one click away.

The retailer needed to have a better understanding of consumer online activity and correlate their behavior to historical buying patterns. To do this however, required analyzing terabytes of data in real-time and the ability to act before the buyer left the website.

The retailer selected Talend Big Data and Hadoop to glue all of their applications, data silos and data formats together to gain new insight into their business and online buyer behavior.

With Talend the retailer is now able to analyze live and historical clickstream data (over 5 terabytes) and provide sub-second responses, such as advertisements or dynamic price changes, while customers are shopping online. They can predict with 90% certainty whether someone will abandon their shopping cart. Additionally they are able to reduce the amount of leftover merchandise by 20% through more thorough historical analysis and better forecasting techniques.

With Talend & Hadoop, the online retailer can predict with 90% certainty whether someone will abandon their shopping cart

About Talend

At Talend, it’s our mission to connect the data-driven enterprise, so our customers can operate in real-time with new insight about their customers, markets and business. Founded in 2006, our global team of integration experts builds on open source innovation to create enterprise-ready solutions that help unlock business value more quickly. By design, Talend integration software simplifies the development process, reduces the learning curve, and decreases total cost of ownership with a unified, open, and predictable platform. Through native support of modern big data platforms, Talend takes the complexity out of integration efforts. For more information, visit http://www.talend.com

About Hortonworks

Hortonworks develops, distributes and supports the only 100% open source Apache Hadoop data platform. Our team comprises the largest contingent of builders and architects within the Hadoop ecosystem who represent and lead the broader enterprise requirements within these communities. The Hortonworks Data Platform provides an open platform that deeply integrates with existing IT investments and upon which enterprises can build and deploy Hadoop-based applications. Hortonworks has deep relationships with the key strategic data center partners that enable our customers to unlock the broadest opportunities from Hadoop. For more information, visit http://www.hortonworks.com.