Hortonworks

A Modern Data Architecture with Apache™ Hadoop®

The Journey to a Data Lake

A Hortonworks White Paper
March 2014

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 then 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 whilst maintaining coherence of the Data Warehouse.

This paper discusses Apache Hadoop, its capabilities as a data platform and how the core of Hadoop and its surrounding ecosystem solution vendors provides the enterprise requirements to integrate alongside the Data Warehouse and other enterprise data systems as part of a modern data architecture, and as a step on the journey toward delivering 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 comprising:

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. 85% 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 high volumes and longer historical perspectives can be inspected for patterns and used for advanced analytic applications.

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

- *Clickstream*
- *Social Media*
- *Server Logs*
- *Geolocation*
- *Machine and Sensor*

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 fully-fledged data platform with the necessary services for the enterprise from Security to Operational Management and more.

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.

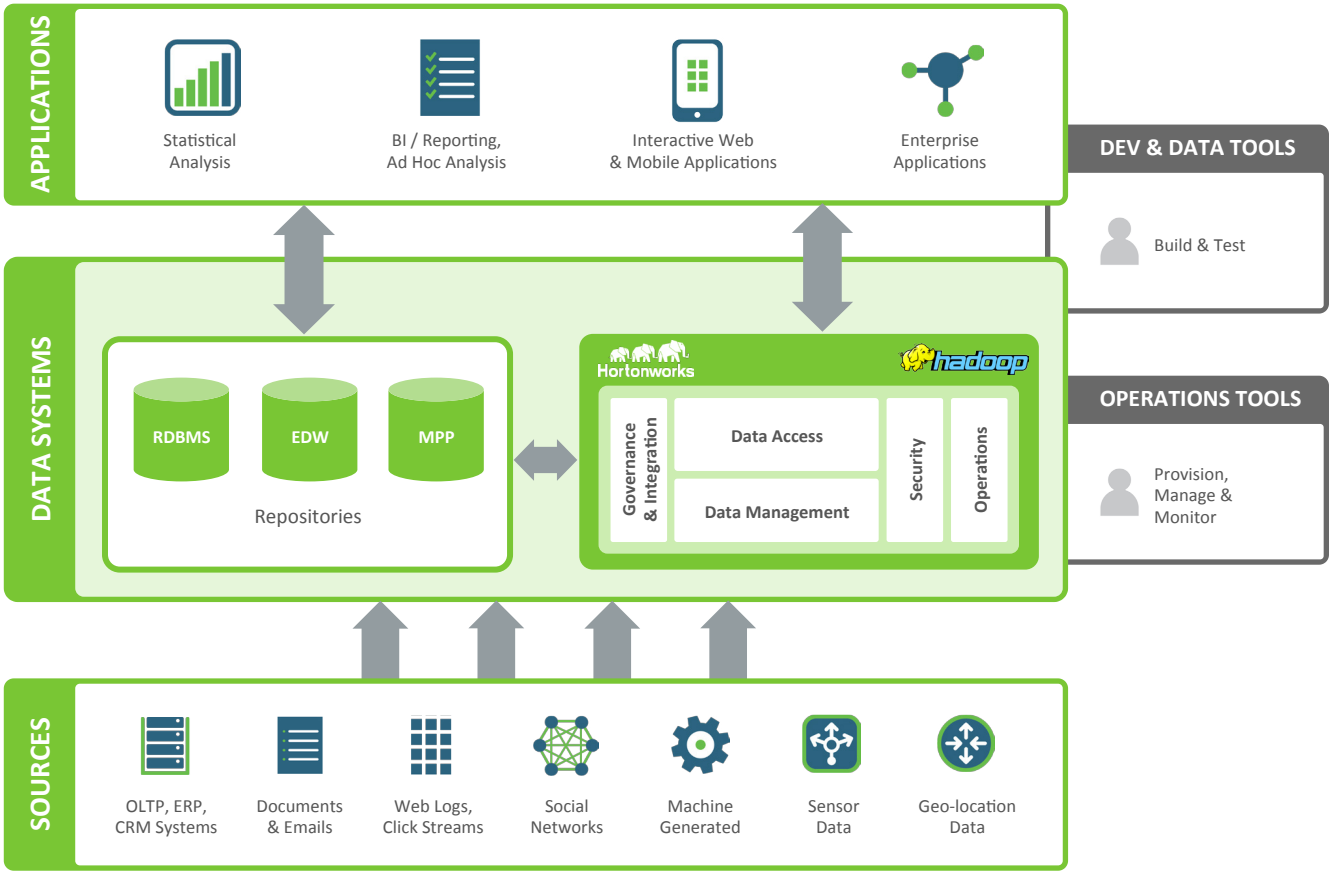

Fig. 1
A Modern Data Architecture with Apache Hadoop integrated with existing data systems

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.

For example, assume an application exists and that combines CRM data with Clickstream data to obtain a single view of a customer interaction. As new types of data become available and are relevant (e.g. server log or sentiment data) they too can be added to enrich the view of the customer. The key distinction being that at the time the data was stored, it was not necessary to declare its structure and association with any particular application.

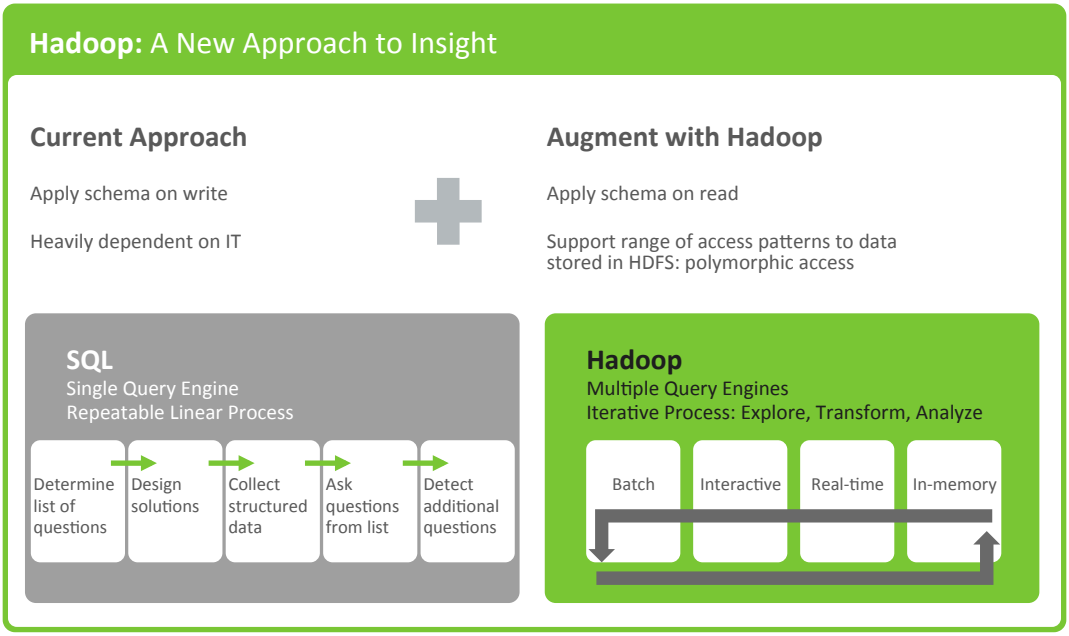

Fig. 2

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.

For example, a manufacturing plant may choose to react to incoming sensor data with real-time data processing, enable data analysts to review logs during the day with interactive processing, and run a series of batch processes overnight. Hadoop enables this scenario to happen on a single cluster of shared resources and single versions of the data.

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 which 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.

For example: in a traditional business intelligence application, it may have only been possible to leverage a single year of data after it was transformed from its original format, whereas by adding Hadoop it becomes possible to store that same 1 year of data in the data warehouse and 10 years of data, including its original format. The end results are much richer applications with far greater historical context.

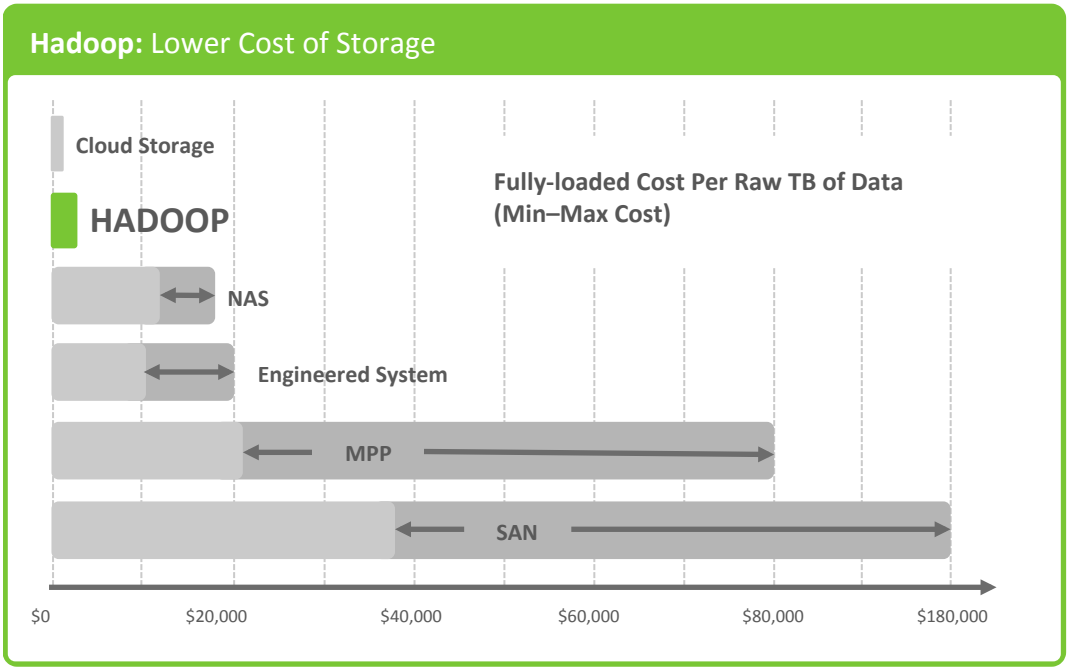

Fig. 3
Source: Juergen Urbanski, Board Member Big Data & Analytics, BITKOM

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 off-load 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.

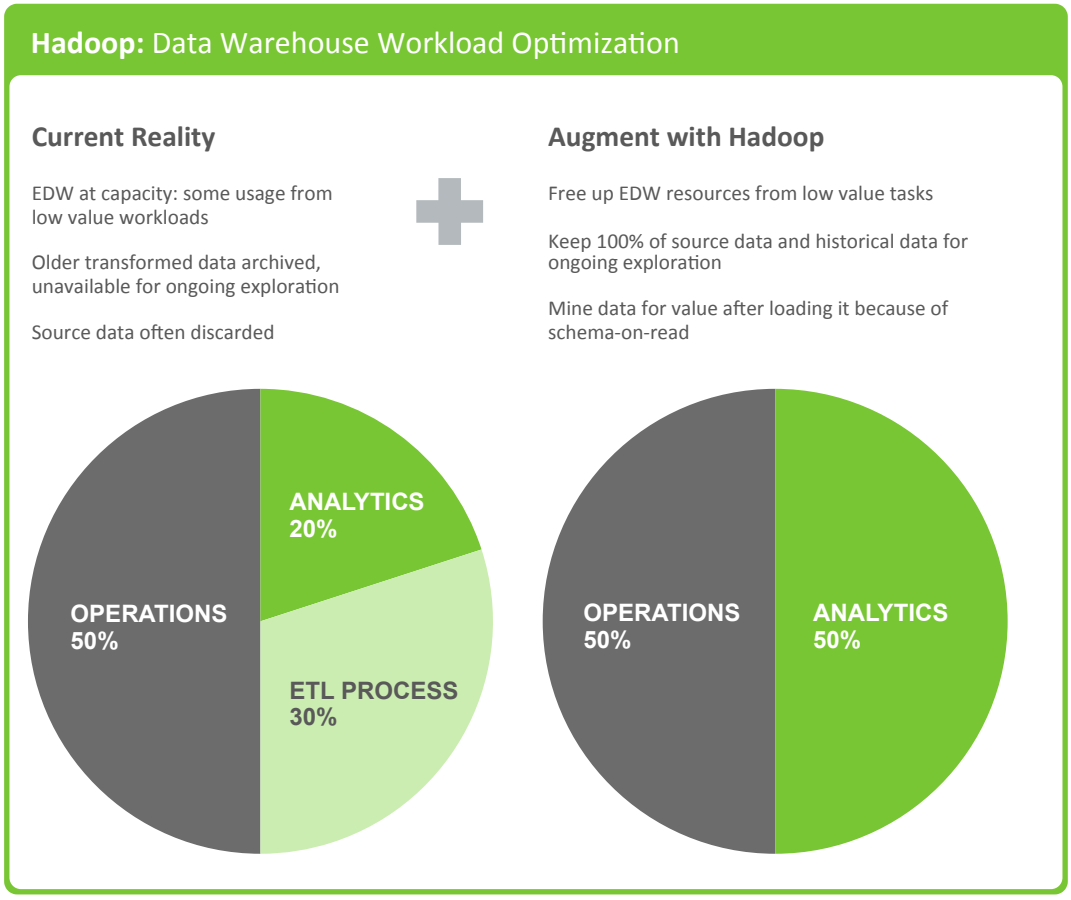

Fig. 4

A Blueprint for Enterprise Hadoop

As Apache Hadoop has become successful in its role in enterprise data architectures, the capabilities of the platform have expanded significantly in response to enterprise requirements. For example in its early days the core components to enable storage (HDFS) and compute (MapReduce) represented the key elements of a Hadoop platform. While they remain crucial today, a host of supporting projects have been contributed to the Apache Software Foundation (ASF) by both vendors and users alike that greatly expand Hadoop's capabilities into a broader enterprise data platform.

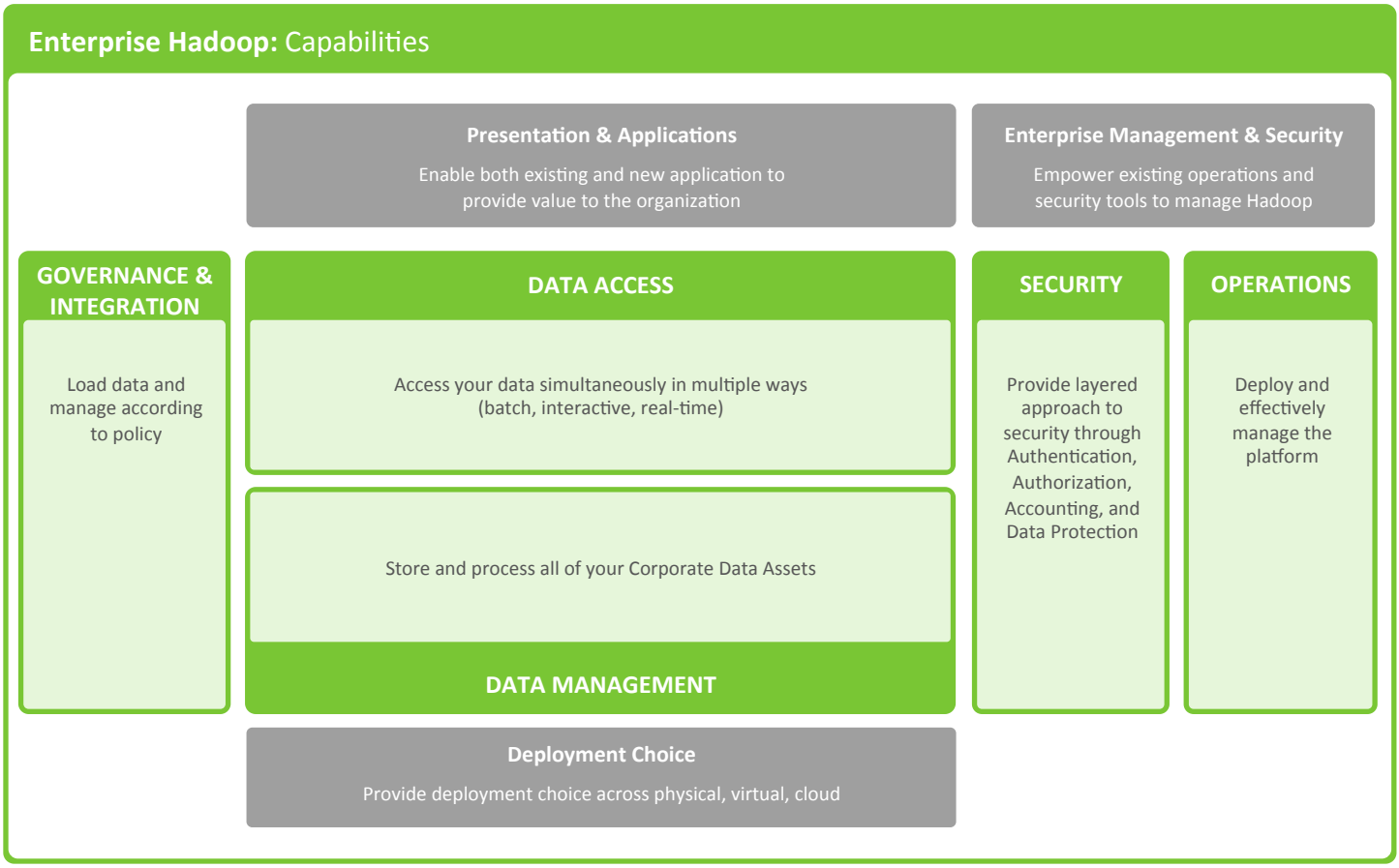

Fig. 5

These Enterprise Hadoop capabilities are aligned to the following functional areas that are a foundational requirement for any platform technology:

Data Management. Store and process vast quantities of data in a scale out storage layer.

Data Access. Access and interact with your data in a wide variety of ways – spanning batch, interactive, streaming, and real-time use cases.

Data Governance & Integration. Quickly and easily load data, and manage according to policy.

Security. Address requirements of Authentication, Authorization, Accounting and Data Protection.

Operations. Provision, manage, monitor and operate Hadoop clusters at scale.

The Apache projects that perform this set of functions are detailed in the following diagram. This set of projects and technologies represent the core of Enterprise Hadoop. Key technology powerhouses such as Microsoft, SAP, Teradata, Yahoo!, Facebook, Twitter, LinkedIn and many others are continually contributing to enhance the capabilities of the open source platform, each bringing their unique capabilities and use cases. As a result, the innovation of Enterprise Hadoop has continued to outpace all proprietary efforts.

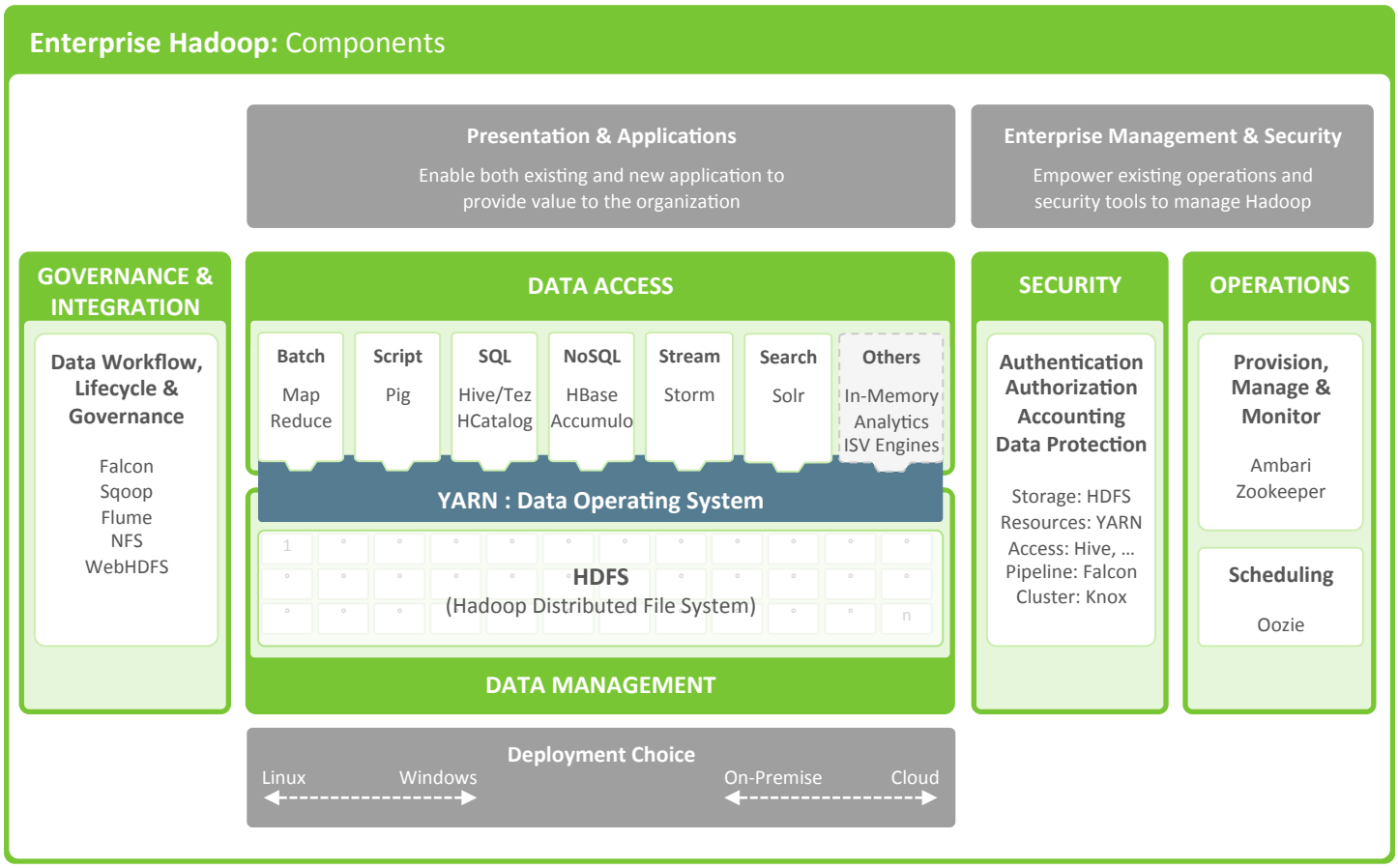

Fig. 6

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 pre-requisite 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, for instance 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.

A Thriving Ecosystem

Beyond these core components, and as a result of innovation such as YARN, Apache Hadoop has a thriving ecosystem of vendors providing additional capabilities and/or integration points. These partners contribute to and augment Hadoop with given functionality, and this combination of core and ecosystem provides compelling solutions for enterprises whatever their use case. Examples of partner integrations include:

Business Intelligence and Analytics: All of the major BI vendors offer Hadoop integration, and specialized analytics vendors offer niche solutions for specific data types and use cases.

Data Management and Tools: There are many partners offering vertical and horizontal data management solutions along side Hadoop, and there are numerous tool sets – from SDKs to full IDE experiences – for developing Hadoop solutions.

Infrastructure: While Hadoop is designed for commodity hardware, it can also run as an appliance, and be easily integrated into other storage, data and management solutions both on-premise and in the cloud.

Systems Integrators: Naturally, as a component of an enterprise data architecture, then SIs of all sizes are building skills to assist with integration and solution development.

As many of these vendors are already prevalent within an enterprise, providing similar capabilities for an EDW, risk of implementation is mitigated as teams are able to leverage existing tools and skills from EDW workloads.

There is also a thriving ecosystem of new vendors that is emerging on top of the enterprise Hadoop platform. These new companies are taking advantage of open APIs and new platform capabilities to create an entirely new generation of applications. The applications they're building leverage both existing and new types of data and are performing new types of processing and analysis that weren't technologically or financially feasible before the emergence of Hadoop. The result is that these new businesses are harnessing the massive growth in data creating opportunities for improved insight into customers, better medical research and healthcare delivery, more efficient energy exploration and production, predictive policing and much more.

Hortonworks has a deep and broad ecosystem of partners, and strategic relationships with key data center vendors:

- *HP*
- *Microsoft*
- *Rackspace*
- *Red Hat*
- *SAP*
- *Teradata*

Toward a Data Lake

Implementing Hadoop as part of an enterprise data architecture is a substantial decision for any enterprise. While Hadoop's momentum is 'unstoppable', its adoption is a journey from single instance applications to a fully-fledged data lake. This journey has been observed many times across our customer base.

New Analytic Applications

Hadoop usage most typically begins with the desire to create new analytic applications fueled by data that was not previously being captured. While the specific application will be invariably unique to an industry, or organization, there are many similarities between the types of data.

Examples of analytics applications across industries include:

INDUSTRY	USE CASE	DATA TYPE								
		Sensor	Server Logs	Text	Social	Geographic	Machine	Clickstream	Structured	Unstructured
Financial Services	New Account Risk Screens		✔	✔						
	Trading Risk		✔							
	Insurance Underwriting	✔		✔		✔				
Telecom	Call Detail Records (CDR)					✔	✔			
	Infrastructure Investment		✔				✔			
	Real-time Bandwidth Allocation		✔	✔	✔					
Retail	360° View of the Customer			✔				✔		
	Localized, Personalized Promotions					✔				
	Website Optimization							✔		
Manufacturing	Supply Chain and Logistics	✔								
	Assembly Line Quality Assurance	✔								
	Crowd-sourced Quality Assurance				✔					
Healthcare	Use Genomic Data in Medial Trials	✔							✔	
	Monitor Patient Vitals in Real-Time									
Pharmaceuticals	Recruit and Retain Patients for Drug Trials				✔			✔		
	Improve Prescription Adherence				✔	✔				✔
Oil & Gas	Unify Exploration & Production Data	✔				✔				✔
	Monitor Rig Safety in Real-Time	✔								✔
Government	ETL Offloaded Response to Federal Budgetary Pressures								✔	
	Sentiment Analysis for Government Programs				✔					

Fig. 7

Enterprise Hadoop

Read about other industry use cases

- *Healthcare*
- *Telecommunications*
- *Retail*
- *Manufacturing*
- *Financial Services*
- *Oil & Gas*
- *Advertising*
- *Government*

Increases in Scope and Scale

As Hadoop proves its value on one or more application instances, increased scale or scope of data and operations is applied. Gradually, the resulting data architecture assists an organization across many applications.

The case studies later in the paper describe the journeys taken by customers in the retail and telecom industries in pursuit of a data lake.

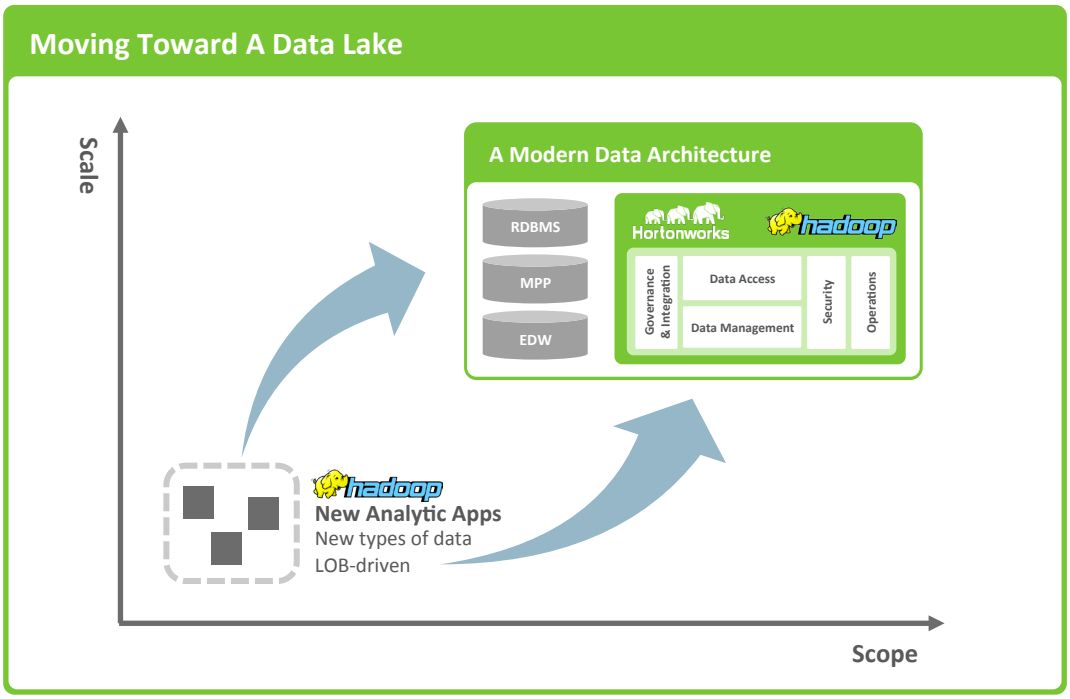

Fig. 8

Vision of a Data Lake

With the continued growth in scope and scale of analytics applications using Hadoop and other data sources, then the vision of an enterprise data lake can become a reality.

In a practical sense, a data lake is characterized by three key attributes:

Collect everything. A data lake contains all data, both raw sources over extended periods of time as well as any processed data.

Dive in anywhere. A data lake enables users across multiple business units to refine, explore and enrich data on their terms.

Flexible access. A data lake enables multiple data access patterns across a shared infrastructure: batch, interactive, online, search, in-memory and other processing engines.

The result: A data lake delivers maximum scale and insight with the lowest possible friction and cost.

As data continues to grow exponentially, then Enterprise Hadoop and EDW investments can provide a strategy for both efficiency in a modern data architecture, and opportunity in an enterprise data lake.

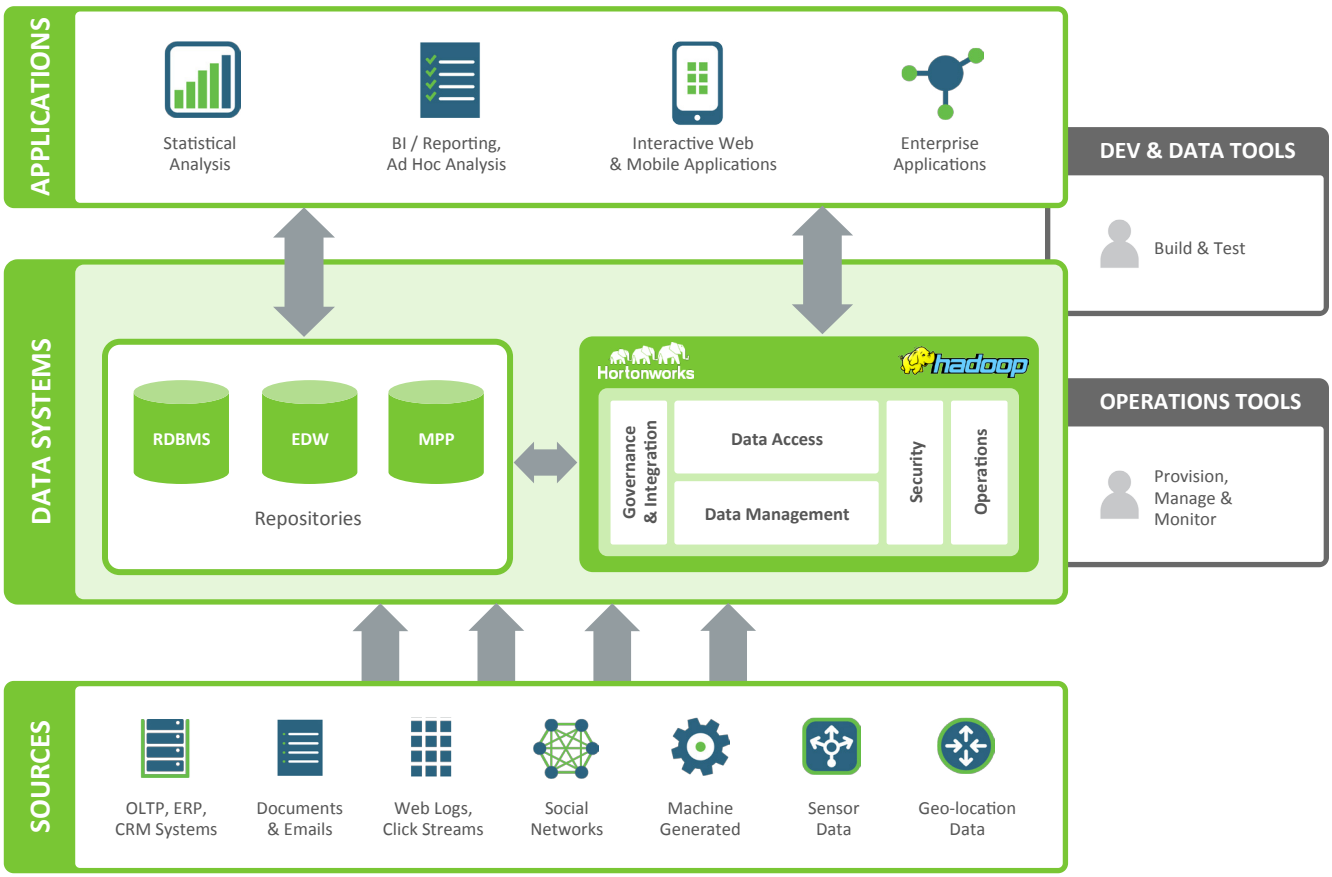

Fig. 9
A Modern Data Architecture with Apache Hadoop integrated with existing data systems

Case Study 1:

Telecom company creates a 360° view of customers

In the telecommunications industry, a single household is often comprised of different individuals who have each contracted with a particular service provider for different types of products, and who are served by different organizational entities within the same provider.

These customers communicate with the provider through various online and offline channels for sales- and service-related questions, and in doing so, expect that the service provider be aware of what's going on across these different touch points.

For one large U.S. telecommunications company, keeping up with the rapid growth in the volume and type of customer data it was receiving proved too challenging, and as a result, it lacked a unified view of the issues and concerns affecting customers. Valuable customer data was highly fragmented, both across multiple applications and across different data stores such as EDWs.

Apache Hadoop 2.0 enabled this service provider to build a unified view of the households it served across all the different data channels of transaction, interaction and observation, providing it with an unprecedented 360° view of its customers. Furthermore, Hadoop 2.0 allowed the provider to create an enterprise-wide data lake of several petabytes cost effectively, giving it the insight necessary to significantly improve customer service.

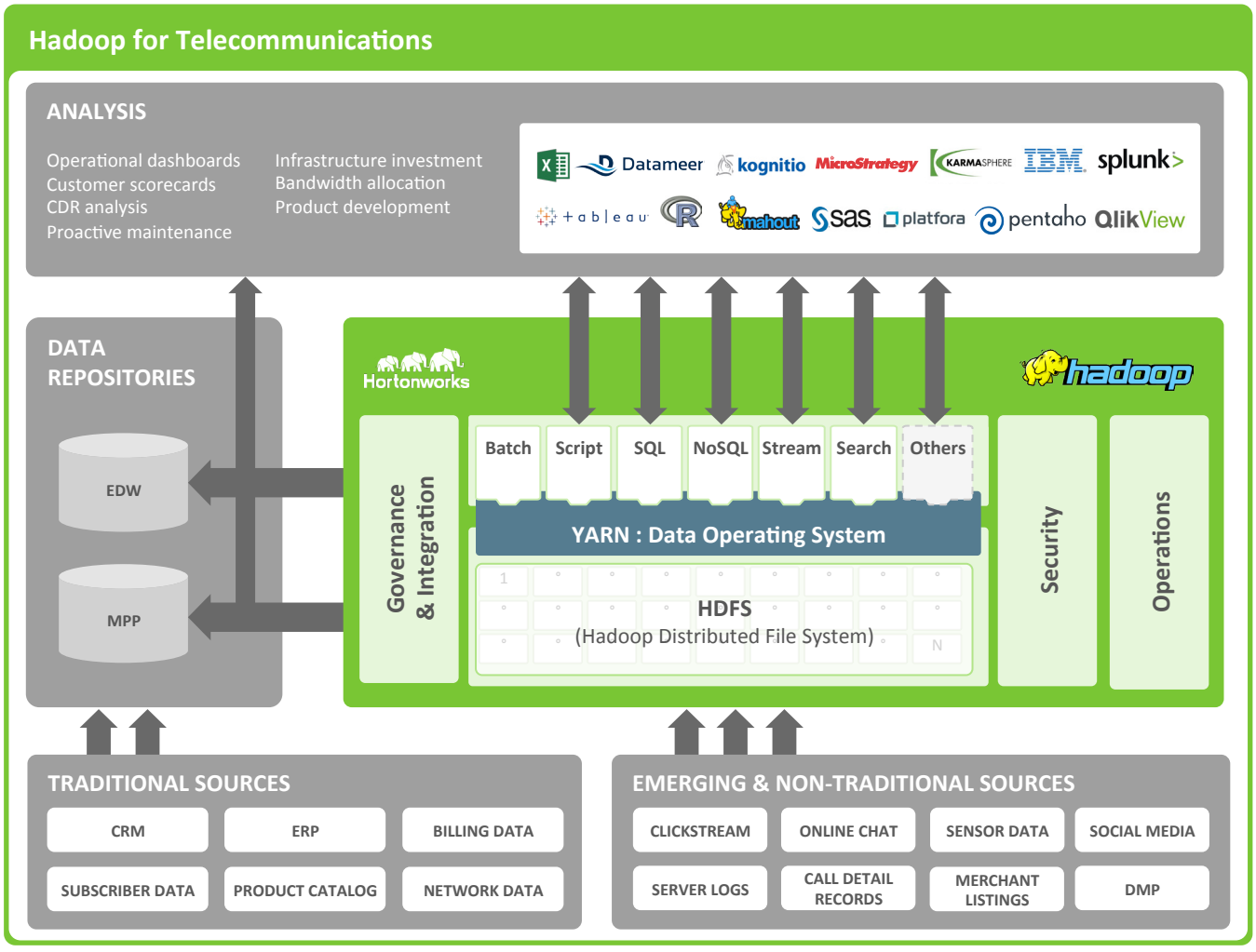

Fig. 10

Case Study 2:

Home improvement retailer improves marketing performance

For a large U.S. home improvement retailer with an annual marketing spend of more than $1 billion, improving the effectiveness of its spend and the relevance of marketing messages to individual customers was no easy feat, especially since existing solutions were ill-equipped to meet this need.

Although the retailer's 100 million customer interactions per year translated into $74 billion in annual customer purchases, data about those transactions was still stored in isolated silos, preventing the company from correlating transactional data with various marketing campaigns and online customer browsing behavior. And merging that fragmented, siloed data in a relational database structure was projected to be time-consuming, hugely expensive, and technically difficult.

What this large retailer needed was a "golden record" that unified customer data across all time periods and across all channels, including point-of-sale transactions, home delivery and website traffic, enabling sophisticated analytics whose results could then be turned into highly targeted marketing campaigns to specific customer segments.

The Hortonworks Data Platform enabled that golden record, delivering key insights that the retailer's marketing team then used to execute highly targeted campaigns to customers, including customized coupons, promotions and emails. Because Hadoop 2.0 was used to right-size it data warehouse, the company saved millions of dollars in annual costs, and to this day, the marketing team is still discovering unexpected and unique uses for its 360° view of customer buying behavior.

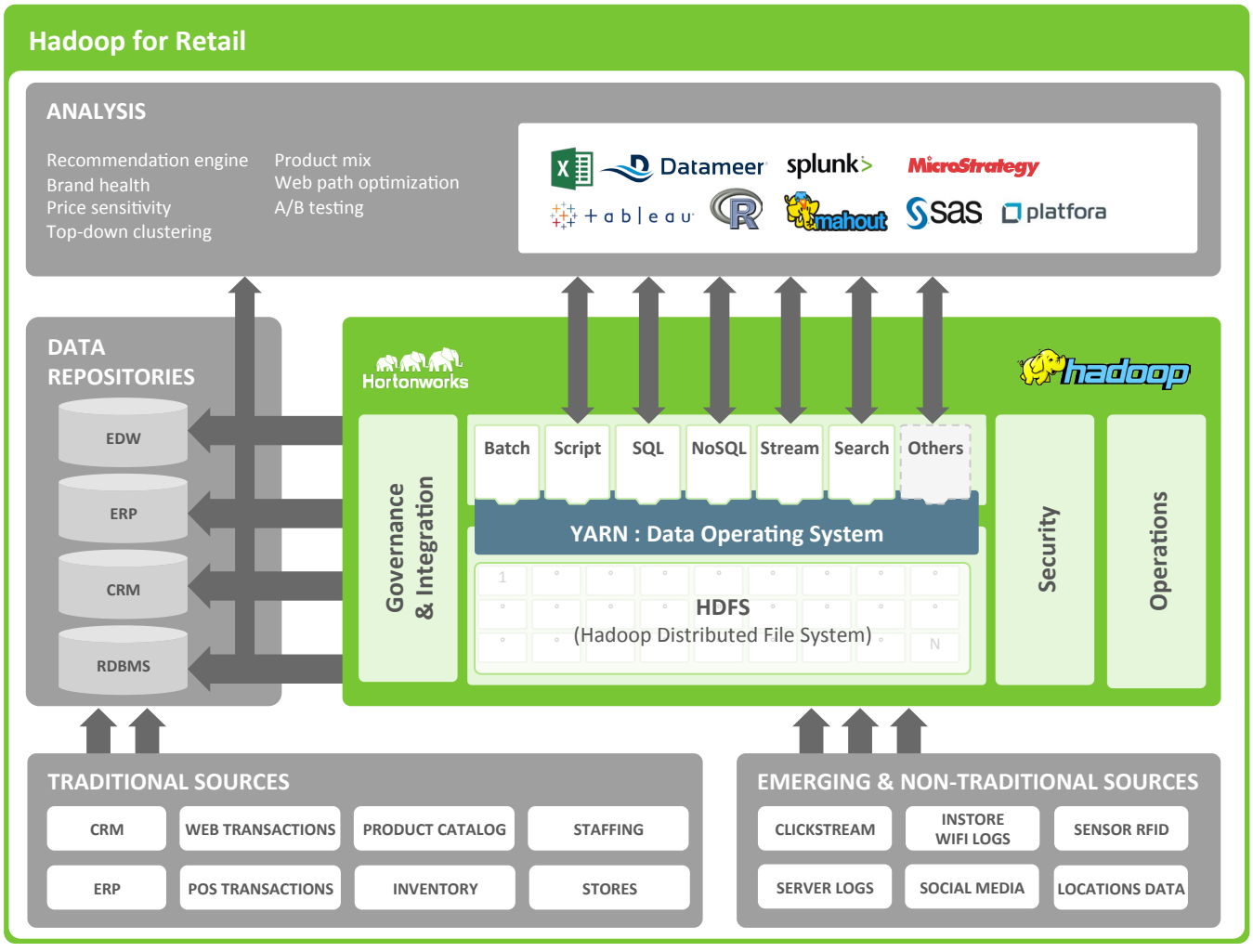

Fig. 11

Build a Modern Data Architecture with Enterprise Hadoop

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 provides Enterprise Hadoop

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.

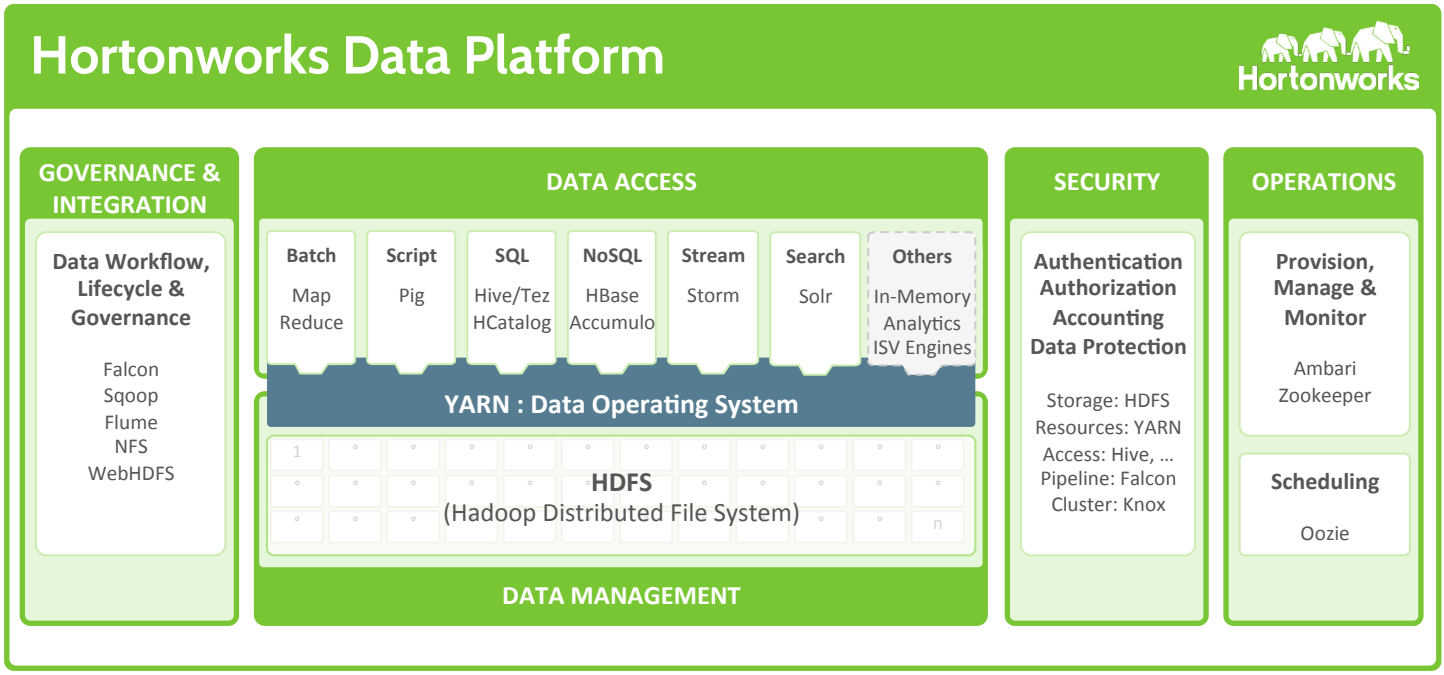

Fig. 12

HDP provides an enterprise with 3 key values:

Completely Open

HDP provides Apache Hadoop for the enterprise, developed completely in the open, and supported by the deepest technology expertise.

HDP incorporates the most current community innovation and is tested on the most mature Hadoop test suite and on thousands of nodes.

HDP is developed and supported by engineers with the deepest and broadest knowledge of Apache Hadoop.

Fundamentally Versatile

HDP is designed to meet the changing need of big data processing within a single platform while providing a comprehensive platform across governance, security and operations.

HDP supports all big data scenarios: from batch, to interactive, to real-time and streaming.

HDP offers a versatile data access layer through YARN at the core of Enterprise Hadoop that allows new processing engines to be incorporated as they become ready for enterprise consumption.

HDP provides the comprehensive enterprise capabilities of security, governance and operations for enterprise implementation of Hadoop.

Wholly Integrated

HDP is designed to run in any data center and integrates with any existing system.

HDP can be deployed in any scenario: from Linux to Windows, from On-Premise to the Cloud.

HDP is deeply integrated with key technology vendor platforms: Red Hat, Microsoft, SAP, Teradata and more.

Deployment Options for Hadoop

HDP offers multiple deployment options:

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

In-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.

Components of Enterprise Hadoop

Read more about the individual components of Enterprise Hadoop.

Data Management

hdfs
yarn

Data Access

mapreduce
pig
hive
tez
hbase
accumulo
storm
hcatalog

Data Governance & Integration

falcon
flume

Security

knox
security

Operations

ambari

Why Hortonworks for Hadoop?

Founded in 2011 by 24 engineers from the original Yahoo! Hadoop development and operations team, Hortonworks has amassed more Hadoop experience under one roof than any other organization. Our team members are active participants and leaders in Hadoop development; designing, building and testing the core of the Hadoop platform. We have years of experience in Hadoop operations and are best suited to support your mission-critical Hadoop project. Read more at http://hortonworks.com/why

Open Leadership

Hortonworks has a singular focus and commitment to drive innovation in the open exclusively via the Apache Software Foundation process.

Hortonworks is responsible for the majority of core code base advances to deliver Apache Hadoop as an enterprise data platform.

Ecosystem Endorsement

Hortonworks is focused on the deep integration of Hadoop with existing data center technologies and team capabilities.

Hortonworks has secured strategic relationships with trusted data center partners including Microsoft, SAP, Teradata, Rackspace, and many more.

Enterprise Rigor

Hortonworks has a world-class enterprise support and services organization with vast experience in the largest Hadoop deployments

Hortonworks engineers and certifies Apache Hadoop with the enterprise in mind, all tested with real-world rigor in the world's largest Hadoop clusters.

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

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 www.hortonworks.com.