Modern Data Architecture for Retail with Apache™ Hadoop® on Windows

The Journey to a Retail Data Lake

A Hortonworks and Microsoft White Paper
JUNE 2014

Executive Summary

Retailers have a long history of investing in data and analytics technologies to drive more effective marketing, merchandising and operations.

The recent explosion of new types and sources of data – from the web and connected devices, social media, suppliers and elsewhere – has created tremendous new opportunities for savvy retailers to boost sales, but also raises the bar: the race is on to harness this 'big data' to keep pace with the competition.

Retail firms that have adapted to the challenge of big data are using it to derive greater customer insights, promote increased brand engagement and loyalty, optimize pricing and promotions, streamline the entire supply chain, and develop innovative new business models. The most successful big data initiatives have integrated data from across many disparate data sources and siloes to achieve outsized results.

Many retail enterprises have turned to Apache Hadoop to collect and manage diverse volumes of unstructured and semi-structured data alongside traditional repositories like the enterprise data warehouse.

For these organizations, Hadoop is fulfilling the vision of an enterprise-wide repository for big data, frequently called a 'data lake.'

An enterprise data lake provides the following benefits to retailers:

> ***New efficiencies,*** through a significantly lower cost of storage and the optimization of data processing workloads such as data transformation and integration.
>
> ***New opportunities,*** through accelerated analytical applications, able to access all enterprise data in both batch and real-time modes.
>
> ***New insights,*** through allowing data from traditional and emerging data sources to be retained, combined and mined in new and unforeseen ways.

Apache Hadoop provides these benefits through a technology core consisting of **a scalable and flexible storage system that can accept data in any format, and an application framework that allows different types of processing workloads to interact with a common pool of stored data.**

This paper explores the role of Hadoop and its supporting ecosystem in meeting enterprise requirements for integrating big data with existing enterprise data systems as part of a modern data architecture. We pay particular attention to the capabilities of Apache Hadoop as a data platform and as an enabler of the enterprise data lake.

Disruption in the Data

Corporate IT organizations in the retail industry have been tackling data challenges at scale for many years now. Traditional sources of data in retail firms include customer relationship management (CRM) systems, enterprise resource planning (ERP) systems, inventory management systems, transactional databases, product catalogs, and other systems supporting core enterprise functions. Shortly after these "systems of record" became established, the enterprise data warehouse (EDW) emerged as a common repository for data extracted from these systems, allowing "business intelligence" applications to more easily tap into enterprise data. Today, every retailer 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.

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 value of blending existing enterprise data with these new types of data is being proven by many retailers.

The technology that has emerged as the preferred way to realize this value 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 processing of large data sets to a full-fledged data platform offering the services necessary for enterprise adoption, including security, operations 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 meet the needs of the very largest retail organizations in the world.

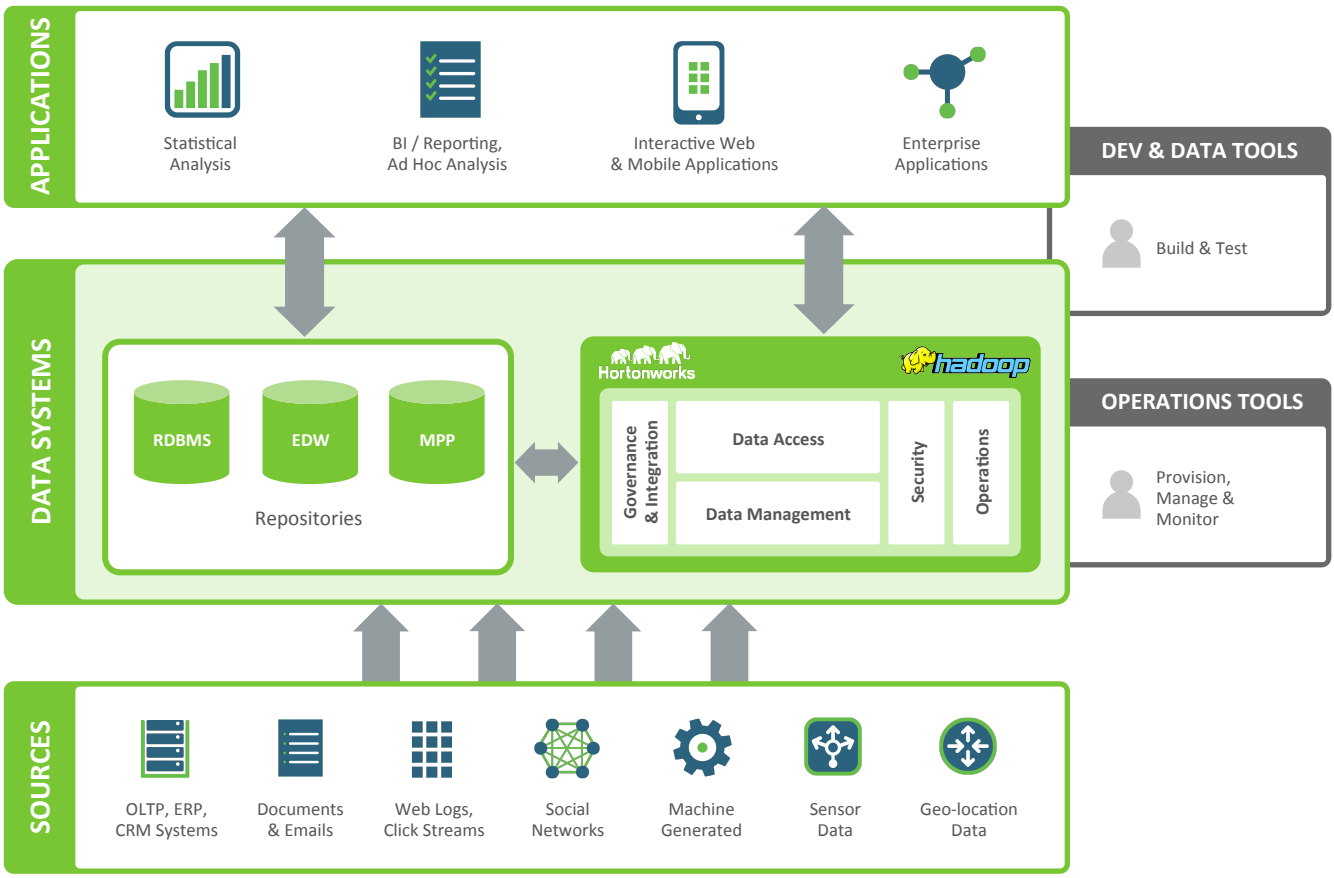

Fig. 1
A Modern Data Architecture with Apache Hadoop integrated into existing data systems

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. Analysts can then create a schema to suit the needs of their application or analysis at the time of use, enabling "schema on read." Schema on read helps organizations overcome concerns about the lack of inherent structure in new data sources, and allows them to avoid the rigidity and waste that comes from pre-processing data when its ultimate use, format or value cannot be foreseen.

For example, assume an application exists and combines CRM data with clickstream data to obtain a single view of a customer interaction. As new types of data become available and relevant (for example, 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.

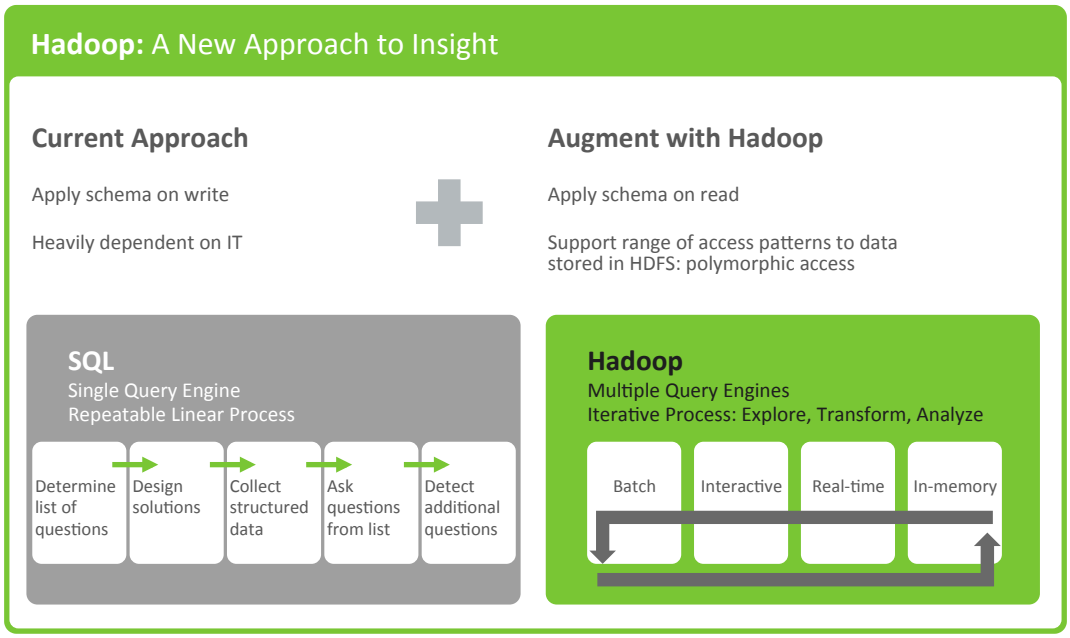

Fig. 2

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), achieving closed-loop analytics and bringing time-to-insight closer to real time than ever before.

For example, an online merchant may produce checkout recommendations in real-time based on a customer's actual shopping pattern, and run a series of batch processes overnight to produce generic product affinity recommendations. 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, allowing for a dramatically lower overall cost of storage. In particular, when compared to high-end Storage Area Networks (SANs) from vendors such as EMC, scaling-out compute and storage using Hadoop provides an extremely compelling alternative. Hadoop allows the user to reduce capital expenditures both because it runs on commodity hardware and also because it allows users to invest in "just enough" hardware to meet immediate needs, and easily expand later as needs grow. The cost dynamic Hadoop offers 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 access a single year of data after it was transformed from its original format and stored in the data warehouse. Using Hadoop, it becomes possible to retain 10 years of that data in its original format, in addition to the year stored in the EDW. This enables much richer applications with far greater historical context.

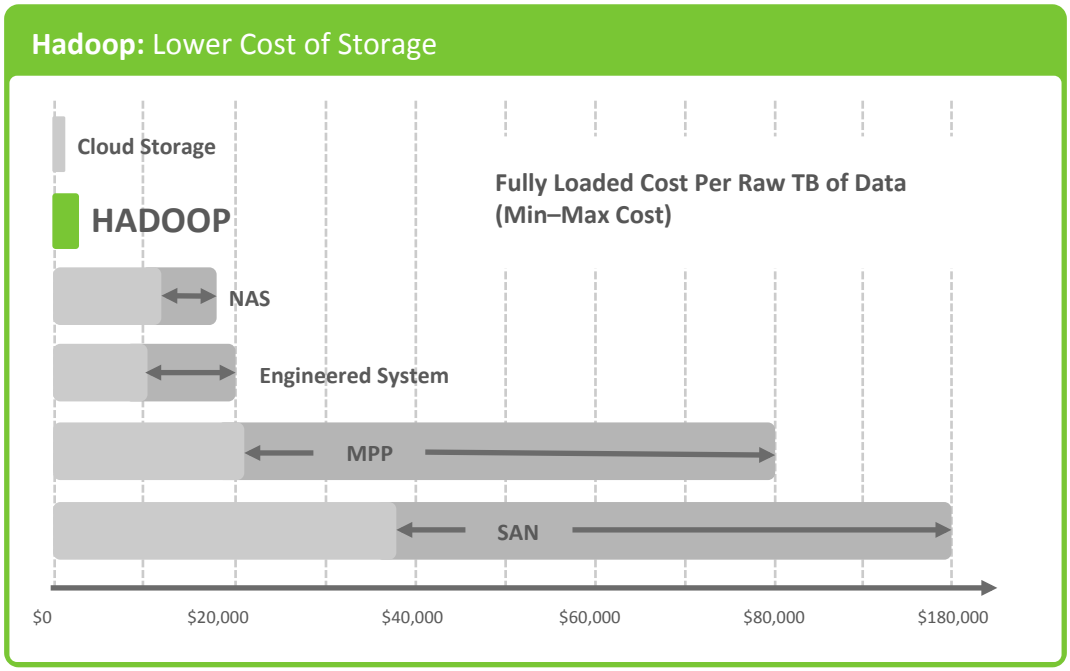

Fig. 3
Source: Juergen Urbanski, Board Member Big Data & Analytics, BITKOM

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 Hadoop for a much lower cost. Many users offload this function to Hadoop, wherein data is extracted and transformed on the Hadoop cluster and the results are loaded into the data warehouse.

This allows critical and high-cost EDW CPU cycles and storage space to be freed, allowing that system to perform the high-value analytics and operations functions that best leverage its advanced capabilities.

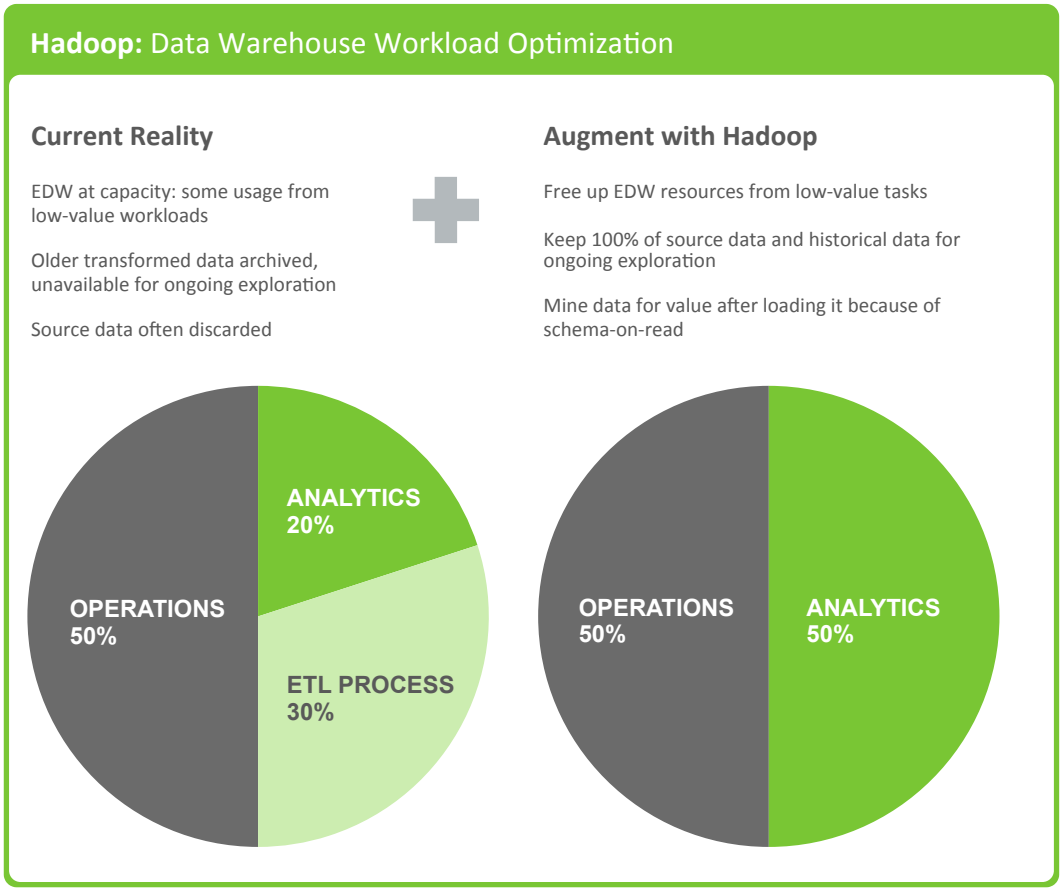

Fig. 4

A Blueprint for Enterprise Hadoop

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

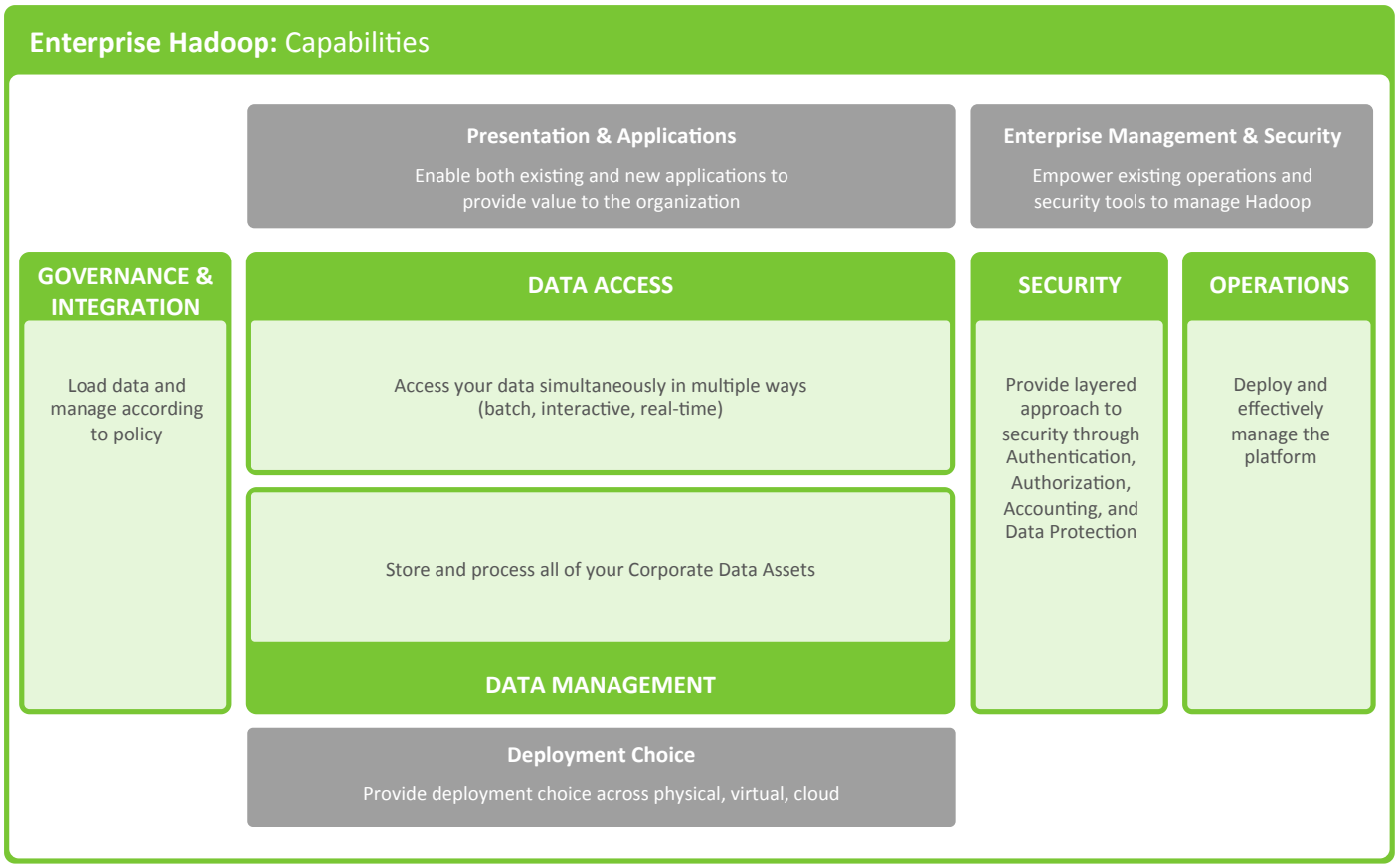

Fig. 5

Hadoop's enterprise capabilities are aligned with the following functional areas, each a foundational requirement for any data platform technology:

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

Data Access. Access and interact with 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 it 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 these 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.

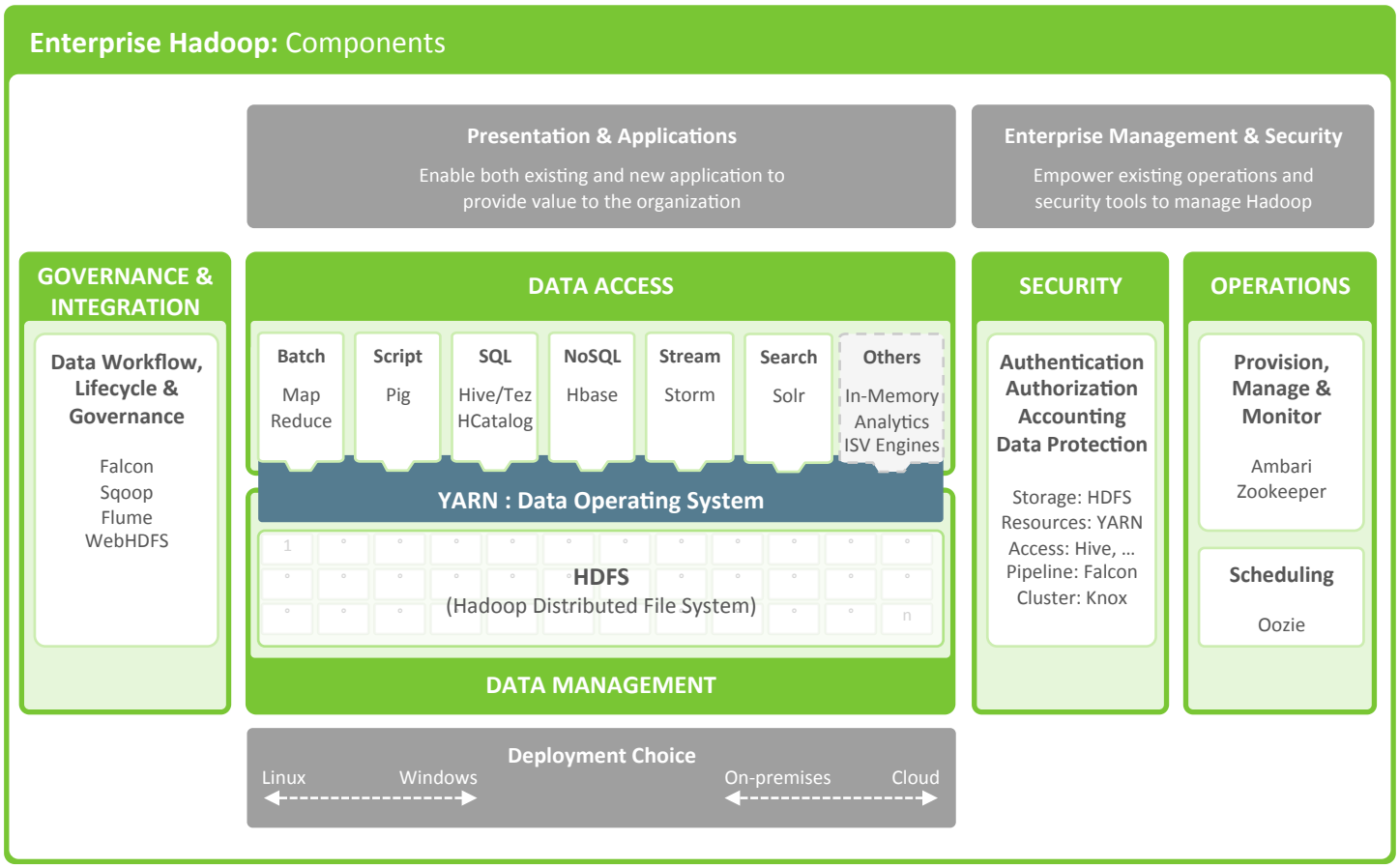

Fig. 6

Data Management: HDFS provides Hadoop's efficient scale-out storage layer, and is designed to run across low-cost commodity hardware. YARN enables Hadoop to serve broad enterprise use cases, providing the resource management and pluggable architecture required to allow a wide variety of data access methods to operate on data stored in Hadoop with predictable performance and service levels.

Data Access: Apache Hive offers direct data connections to Microsoft Excel and Power BI, and 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, and Apache HBase offers columnar NoSQL storage. 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.

A Thriving Ecosystem

Beyond these core components, and as a result of innovations 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 technology and ecosystem support 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. Microsoft integrations with Excel and Power BI allow users to take advantage of Hadoop-based data and analytics in the familiar Office 365 environment.

Data Management and Tools: There are many partners offering vertical and horizontal data management solutions alongside 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-premises and in the cloud.

Systems Integrators: Naturally, given its success as a component of a full-featured enterprise data architecture, SIs of all sizes are building skills to assist with integration and solution development.

As many of these vendors are already present within the enterprise providing similar capabilities for data warehouses, implementation risk is mitigated as existing tools and skills from EDW workloads are applied to Enterprise Hadoop.

There is also a thriving ecosystem of new vendors 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 and Microsoft have a deep and broad ecosystem of partners, and strategic relationships with key data center vendors:

- HP
- Rackspace
- Red Hat
- SAP
- Dell

Toward the Enterprise Data Lake

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

New Analytics Applications

Hadoop usage most typically begins with the desire to create new analytics 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
- Oil & Gas
- Advertising
- Government

Increases in Scope and Scale

As Hadoop proves its value in 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 this paper describe the journeys taken by retail customers in pursuit of a data lake.

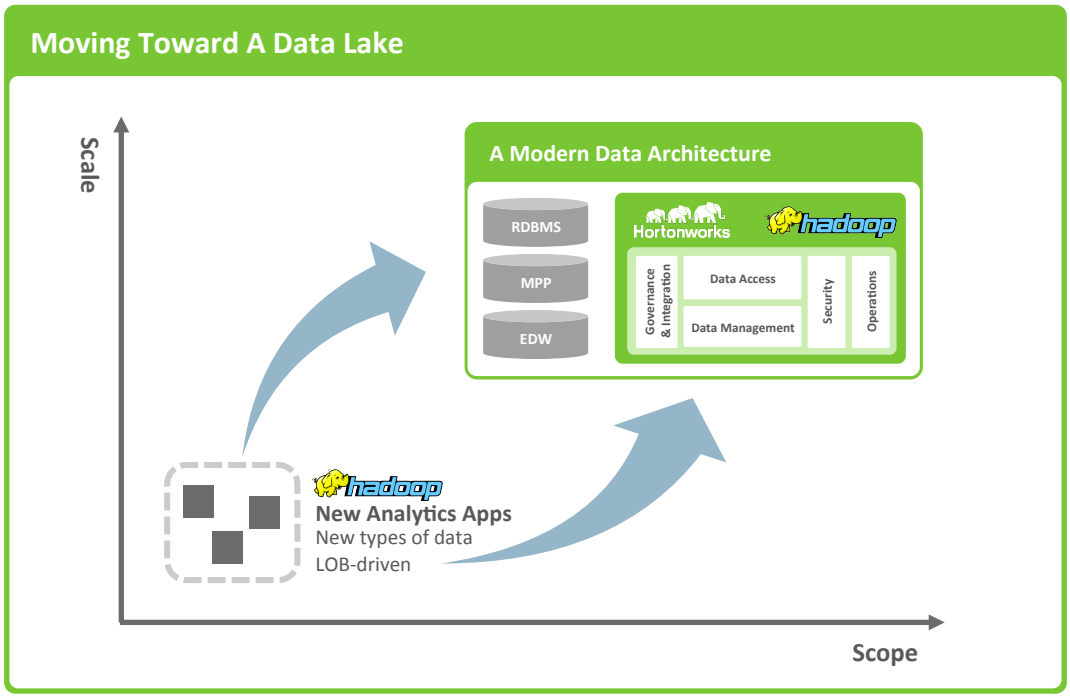

Fig. 8

Vision of a Data Lake

With continued growth in the scope and scale of analytics applications using Hadoop within the enterprise, 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, Enterprise Hadoop and EDW investments can provide a strategy for both efficiency in a modern data architecture and opportunity in an enterprise data lake.

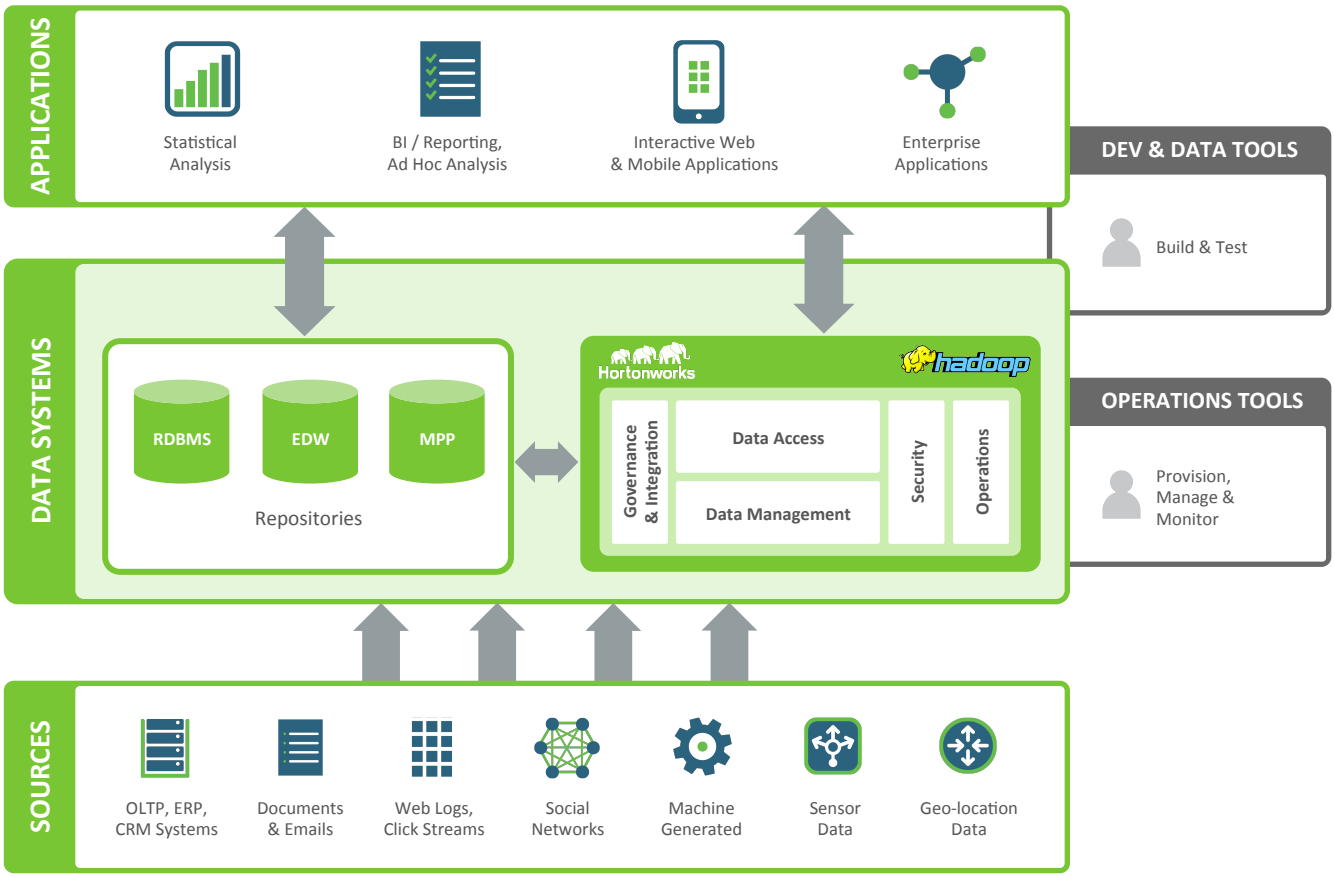

Fig. 9
A Modern Data Architecture with Apache Hadoop integrated into existing data systems

Hadoop in Retail

When Hadoop is integrated with modern retail operations, it dramatically reduces the cost of capturing, ingesting, storing and analyzing data. This enables retailers to analyze enough data to make statistically confident observations on empirical retail data, rather than rolling the dice with customer panels, small samples and focus groups, to guess what drives sales.

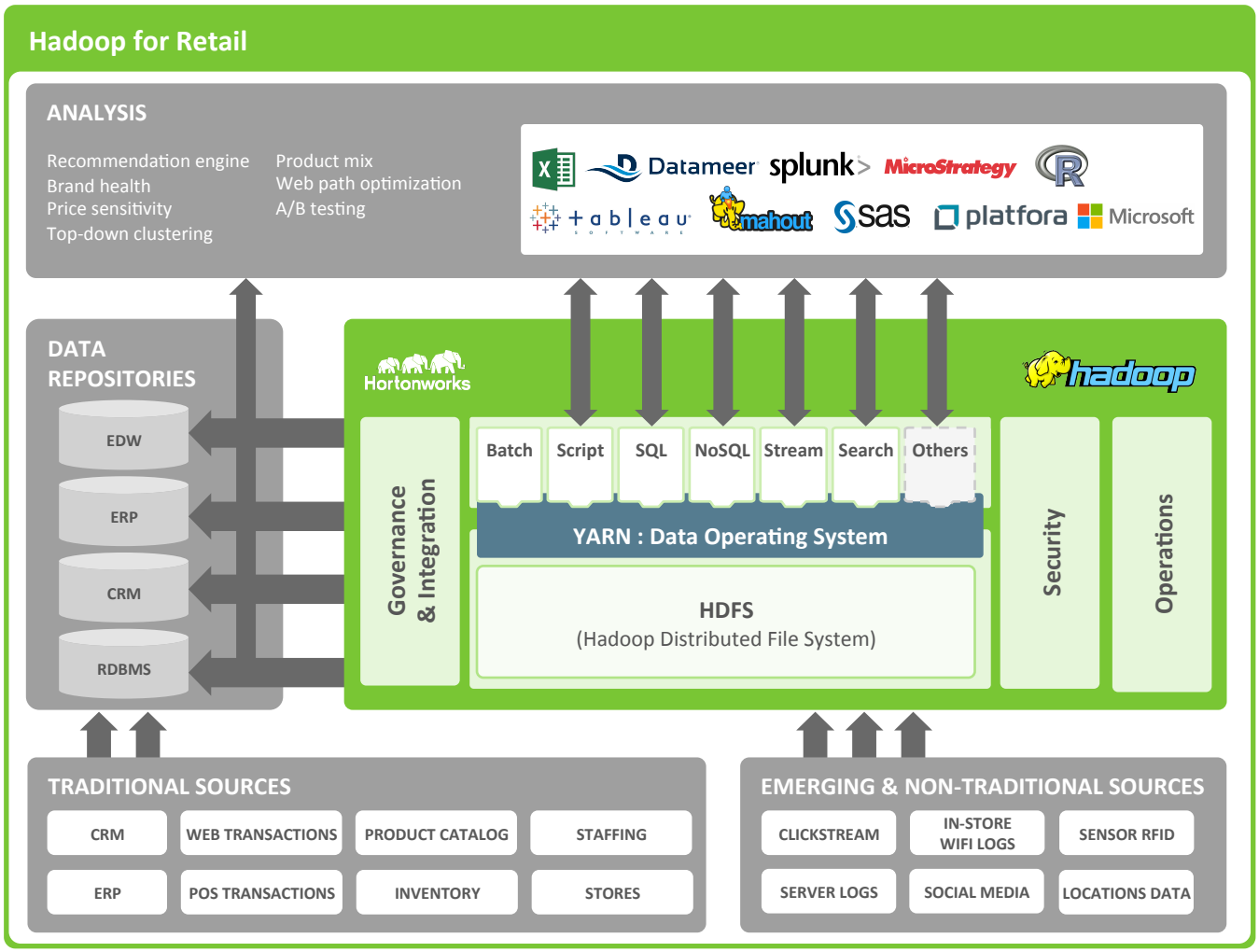

Fig. 10

Case Studies

Build a 360° View of the Customer

Retailers interact with customers across multiple channels, yet customer interaction and purchase data is often isolated in disparate silos. Few retailers can accurately correlate eventual customer purchases with marketing campaigns and online browsing behavior.

Apache Hadoop gives retailers a 360° view of customer behavior. It lets them store data longer and apply interactive and batch analytics to identify the various phases of the customer lifecycle. Better customer analytics leads to greater customer insight, which helps increase sales and loyalty, reduce inventory expenses and retain the best customers.

Analyze Brand Sentiment

Enterprises retailers lack a reliable way to track and analyze brand health. It's difficult to explore the impact of advertising, competitor moves, product launches and news stories on a brand. Canned social media dashboards are not enough, and internal brand studies can be slow, expensive and flawed.

Apache Hadoop enables quick, unbiased snapshots of brand opinions expressed in social media. Retailers can analyze sentiment from Twitter, Facebook, LinkedIn or industry-specific social media streams. With a better understanding of customer perceptions, retailers can align their communications, products and promotions with those perceptions.

Localize and Personalize Promotions

Retailers able to harness the converging forces of social, mobile and local are well positioned to win with the modern consumer. Merchants that can geo-locate their mobile subscribers can deliver localized and personalized promotions. This requires connections with both historical and real-time streaming data.

Apache Hadoop brings the data together to inexpensively localize and personalize promotions delivered to mobile devices. Retailers can develop mobile apps to notify customers about local events and sales that align with their preferences and geographic location–even down to a particular section in a specific store.

Build a Modern Data Architecture with Enterprise Hadoop and Microsoft Windows

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.

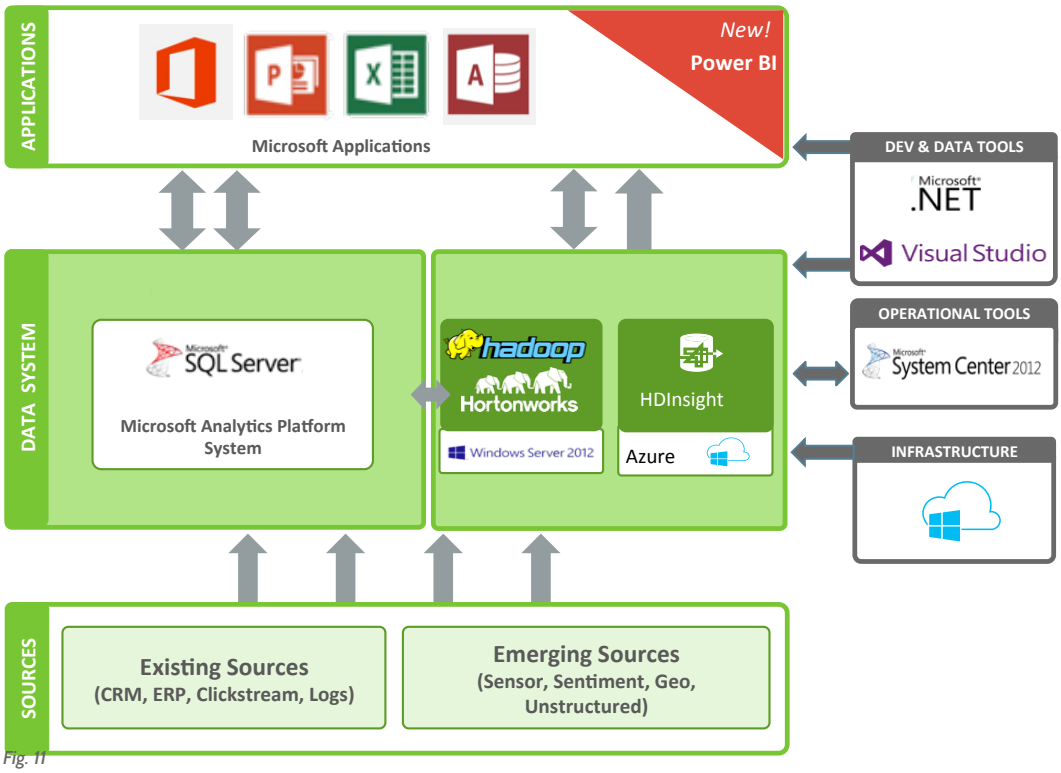

Fig. 11

Hortonworks Data Platform for Windows Provides Enterprise Hadoop

Hortonworks Data Platform (HDP) for Windows is powered by 100% Open Source Apache Hadoop. HDP for Windows provides all of the Apache Hadoop-related projects necessary to integrate Hadoop into a Windows environment as part of a modern data architecture.

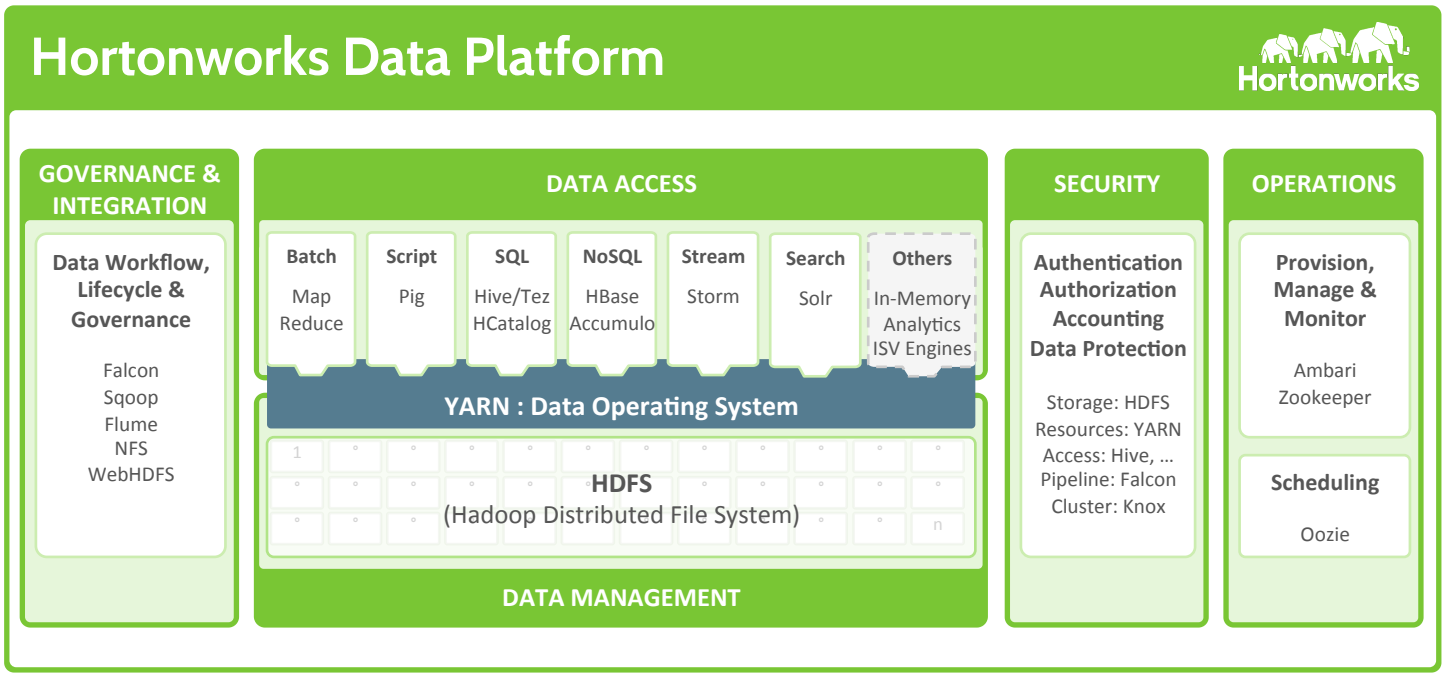

Fig. 12

Three key values:

HDP for Windows provides an enterprise with three 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 needs of big data processing within a single platform while providing a comprehensive platform across governance, security and operations.

HDP supports all big data scenarios, including batch, interactive, 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 for Windows can be deployed in any Windows environment, on-premises or in the cloud, and is fully compatible with Linux-based HDP deployments.

Developed in partnership with Microsoft, HDP for Windows is deeply integrated with Microsoft tools, technologies and applications, including .NET, Microsoft Office, Power BI, SQL Server and the Analytics Platform System.

HDP for Windows Deployment Options

HDP for Windows offers multiple deployment options:

On-premises: HDP for Windows is the only Hadoop platform that works seamlessly in the Windows environment, and is easily deployed on any commodity hardware.

Cloud: HDP for Windows can be run in the Microsoft Azure IaaS environment, and is fully compatible with the Microsoft Azure HDInsight Service.

Appliance: The Microsoft Analytics Platform System is a turnkey big data analytics appliance combining a high performance MPP SQL Server data warehouse with HDP for Windows in a single fully-integrated solution.

Components of Enterprise Hadoop
Read more about the individual components of Enterprise Hadoop.

Data Management
hdfs
yarn

Data Access
mapreduce
pig
hive
tez
hbase
storm
hcatalog

Data Governance & Integration
falcon
flume

Security
knox
security

Operations
ambari

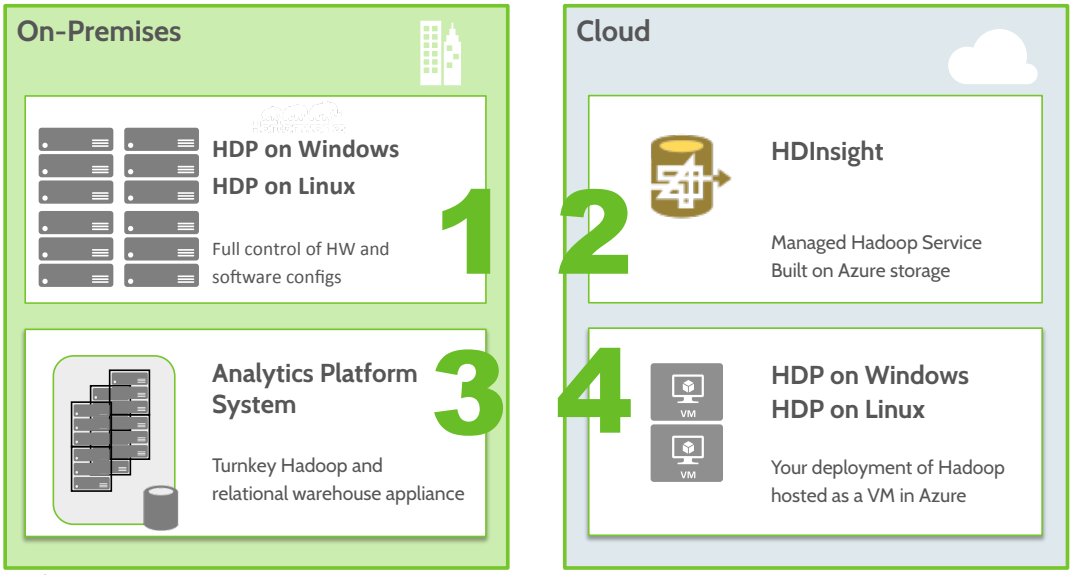

Fig. 13

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 http://www.hortonworks.com.

About Microsoft

Founded in 1975, Microsoft (Nasdaq: "MSFT") is the worldwide leader in software, services and solutions that help people and businesses realize their full potential.