Modern Data Architecture with Apache™ Hadoop®

THE HYBRID DATA WAREHOUSE

Presented by Hortonworks and Denodo

Executive Summary

Apache Hadoop didn't disrupt the data center, 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, its capabilities as a data platform and how it supports Hybrid Data Warehouse, in combination with data virtualization, to deliver a single unified and coherent logical view on all enterprise data assets in and outside the data lake, minimizing unneeded data replication.

The Hybrid Data Warehouse:

1. Extends the value in the traditional EDW by logically combining it with unstructured data from the lake and from other corporate and public sources.

2. Simplifies the offload of "cold data" from the traditional EDW to a low cost storage solution like Hadoop and offers seamless access to all data (both warehoused and offloaded to Hadoop) by applying high performance query federation.

A Hybrid Data Warehouse 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 and Denodo's data virtualization provide these benefits through a technology core comprised of:

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

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

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

Data Virtualization. Data virtualization provides a single logical data access abstraction layer across multiple data sources enabling rapid delivery of complete information to business users.

The Hybrid Data Warehouse guarantees the best query performance and introduces agility and a sound information architecture by enabling reuse of access and integration logic across applications.

Agility: Reduces development costs and time to solution. Reuse reduces maintenance costs and creates coherent as well as future-proof data architectures.

Performance: Denodo combines real-time query optimization techniques and intelligent caching to provide a high throughput architecture, making the most of Big Data computing platforms. Queries are delegated to Hadoop whenever possible to achieve the best performance.

Return on Assets: The Hybrid Data Warehouse supports more data sources (specialized NoSQL data stores, public web data, cloud apps, etc.) and can expose data to more consumers (linked data through RESTful for instance), improving the overall Return on Assets of your data solutions.

Extends the application of "schema-on-read" to specialized NoSQL operational data stores outside the Hybrid Data Warehouse like Document, Graph or Hierarchical DBs.

These benefits are provided through the use of Data Virtualization, the technology at the core of the Denodo platform.

The combination of Denodo's on-demand data access capabilities with Hortonworks' low-cost data storage and analysis platform based on Hadoop brings the best of both worlds to the Hybrid Data Warehouse.

The Disruption in the Data

Corporate IT functions within enterprises have been tackling data challenges at scale for many years. 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 2.8 ZB of data in 2012 is expected to grow to 40 ZB 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 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.

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.

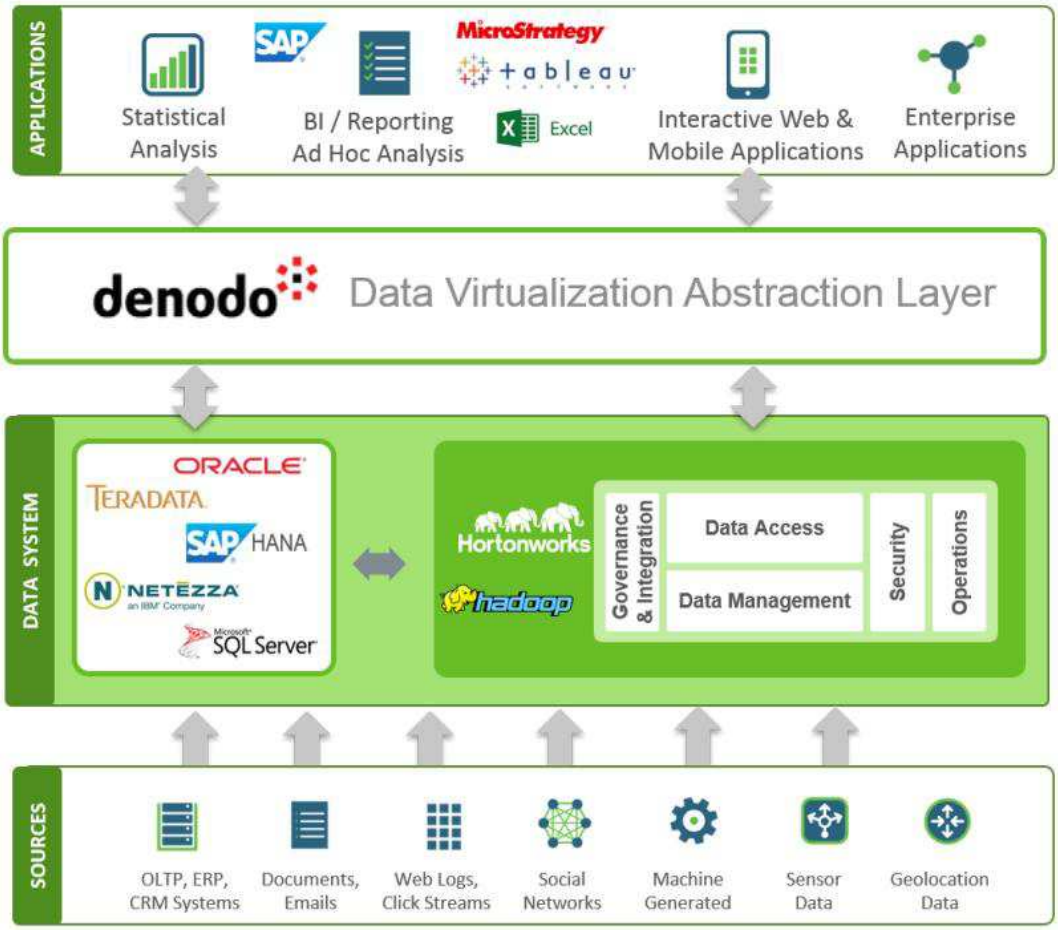


Fig. 1
The Hybrid Data Warehouse: A Modern Data Architecture combining Hortonworks' Apache Hadoop with Denodo's Data Virtualization

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

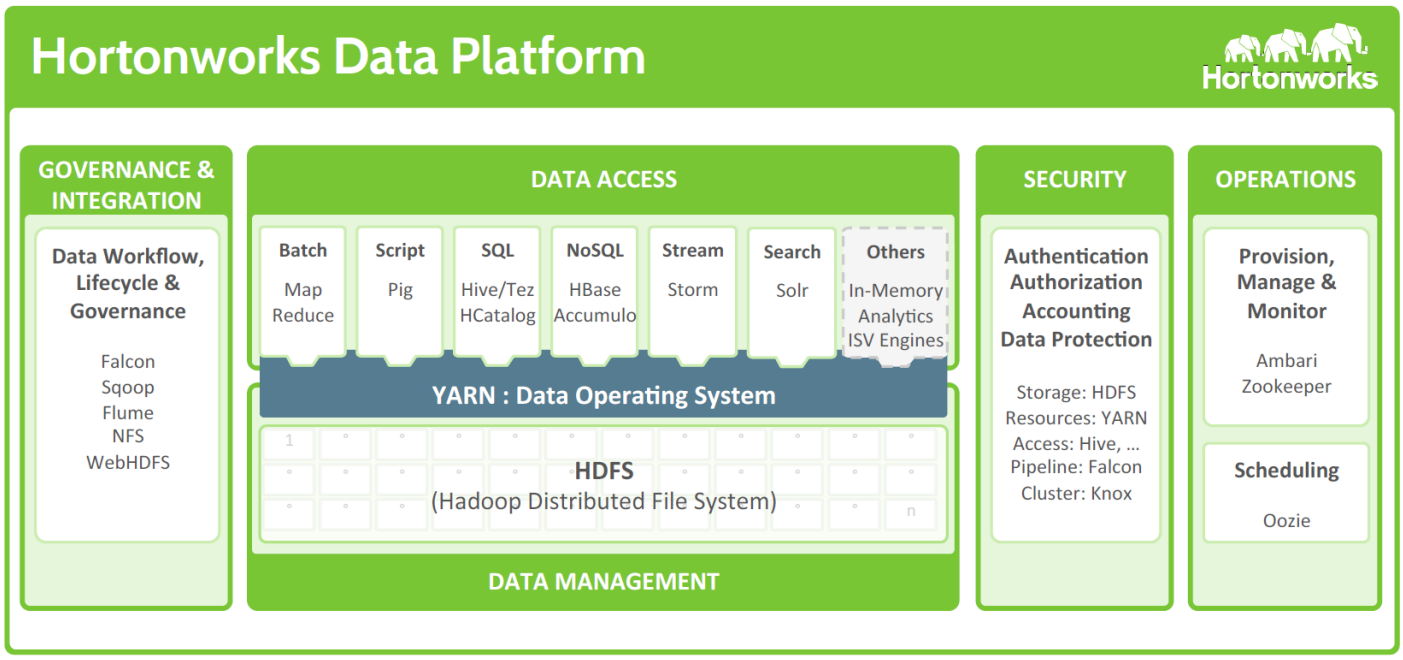
The result: critical CPU cycles and storage space can be freed up from the data warehouse, enabling it to perform the truly high-value functions—analytics and operations—that best leverage its advanced capabilities.

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.



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.

The Hybrid Data Warehouse and Enterprise Hadoop

Denodo's Data Virtualization platform together with the Hortonworks Data Platform form the Hybrid Data Warehouse. This is a modern information management platform that provides seamless access to heterogeneous data both from the point of view of its provenance, access mode and representation paradigm (corporate-public, internal-cloud, unstructured, structured, semi-structured, multi-structured, etc.). It provides the abstraction required to offer business users combined reports with better insights built on data coming from Hadoop, traditional BI repositories, public web data or Cloud app data.

For an independent analysis of the Denodo Data Virtualization Platform, download Forrester Wave™: Enterprise Data Virtualization, Q1 2015 from Forrester Research.

From an implementation perspective, Data Virtualization is middleware that sits between the data sources and the consumer applications and is in charge of encoding the access, transformation and integration logic in a declarative and explicit manner. This means that DV offers a "lazy evaluation" or "on-demand" execution where data is only retrieved from the source (and transformed and combined) on-demand and with high performance when a request comes from the application end.

In addition to its core functionality, Data Virtualization:

- Overlays a rich role-based access control model on top of unstructured and multi-structured data. Often data management systems / data stores lack this capability or do not offer the desired granularity. Data Virtualization upgrades the security capabilities of data sources to put them on the same level as relational databases in terms of security capabilities.
- Manages a unified catalogue and full data lineage and dependency model, enabling governance and self-service BI.

Data Virtualization features enable high-value use cases like:

- **Near real-time operational reporting and BI:** Data Virtualization creates virtual extensions to traditional Data Warehouses with data from operational systems, live data streams or big data from Hadoop deployments.
- **Offload of Data Warehouse data to Hadoop** and homogeneous access to both "hot" data from the Enterprise Data Warehouse and "cold" data from Hadoop.
- **Data Scientist's sandbox** where it is possible to run analytics and machine learning algorithms on data sourced in a controlled and governed manner from either the lake, other enterprise storage systems or the scientist's own data sets such as training data or predefined models.

The Hybrid Data Warehouse combines consolidation in Hadoop with on-demand data access and intelligent caching through Denodo and gives Information Architects a much more versatile toolset to deal with today's information requirements.

Case Study

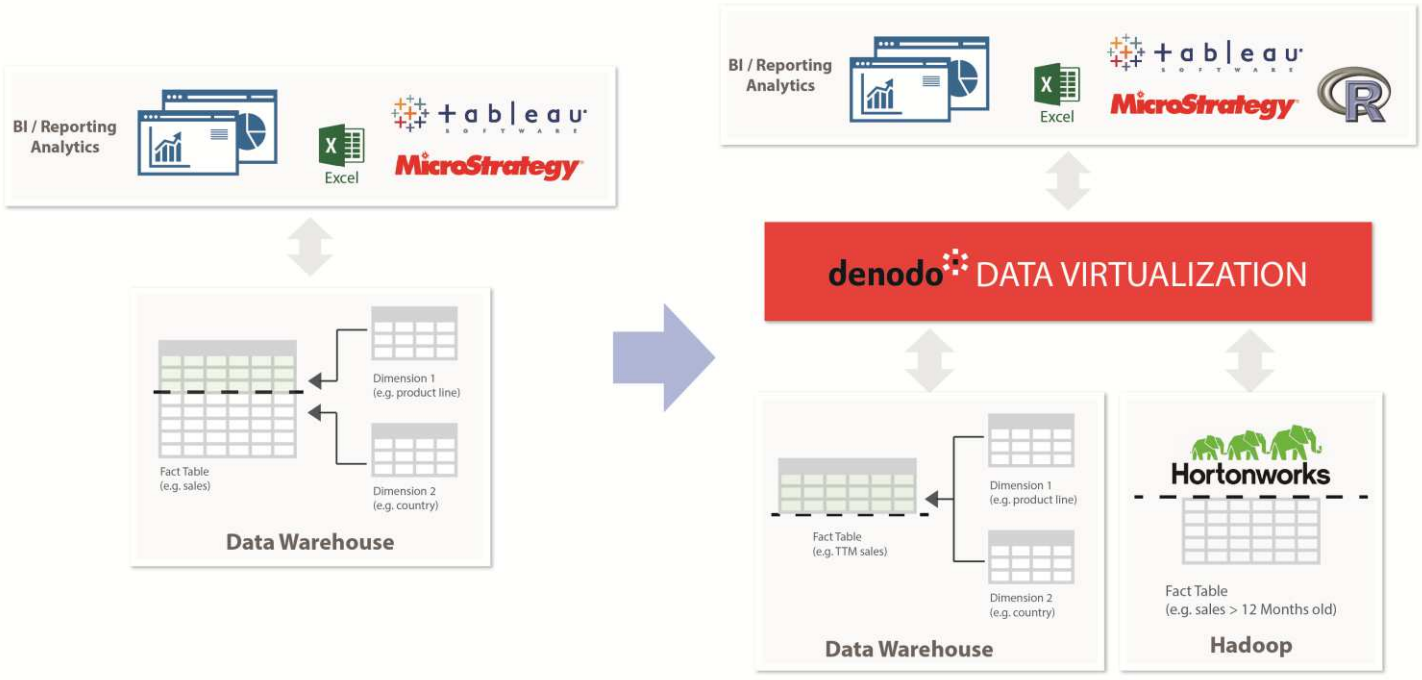
Cold Data Offload from EDW

In order to reduce storage costs and maintain query performance, organizations are forced to either remove or archive the oldest or least used data from their EDW.

The Hybrid Data Warehouse makes it possible to offload this "cold data" from the EDW to Hortonworks' Hadoop low-cost storage and at the same time, through data virtualization, make it transparent to consumer applications and processes. The whole data set appears as if it's coming from a single virtual data store, despite the fact that data is partitioned.

The result is a hybrid environment where an unlimited amount of data can be kept available for online analysis with a significant reduction of the cost of storing it using SQL technology.

Denodo Data Virtualization provides the abstraction and data federation capabilities that make it possible to optimize the execution of queries issued against the Hybrid Data Warehouse. This will involve trimming unnecessary query branches and delegating the execution of parts of the query to either the SQL EDW, Hadoop, or both, always achieving the best possible performance and allowing you to meet the client applications' Service Level Agreements (SLAs).



About Denodo

Denodo, the leader in Data Virtualization, provides a common enterprise data layer that enables organizations to better harness all of their data and deliver faster, agile information to business in every industry. Our global customers leverage the Denodo Data Virtualization platform to address a broad spectrum of use cases, including Big Data analytics and agile BI solutions using canonical business views of data. Learn more about the Denodo Data Virtualization Platform at www.denodo.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.