Combining Hortonworks with Informatica

A Reference Architecture

David Hoyle
20 August 2013

Executive Summary

Today companies collect more data than ever before. From a wide variety of sources. And in many formats.

Many companies have built large-scale environments for transactional data with analytic databases. But these databases are now being flooded with new types of data from social media, server logs, clickstreams, and machine sensors.

These new data sources all share the common Big Data characteristics of volume (size), velocity (speed), and variety (type). Up to now, this data has been too costly to store and analyze.

But with Hortonworks, you can now use this new data to gain business insights and competitive advantage.

The Hortonworks Data Platform is a hardened Hadoop distribution that allows you to store, process, and manage data at scale. It is designed to integrate with and extend existing data applications. With Hortonworks, you can retain and process more data, join new and existing data sets, and lower your data analysis costs.

Hadoop offers clear advantages in data processing power and storage to help you meet the challenges and reap the benefits associated with Big Data. But you may still lack the specialized skills needed to successfully integrate Hadoop with your existing data systems and applications. And without successful integration with existing systems, Hadoop runs the risk of becoming just another information silo.

Informatica provides a comprehensive data integration platform that enables you to integrate Hadoop with all of your existing information management systems. You can create a single unified data repository – or "data lake" – for your company.

The combined capabilities of Hortonworks and Informatica make Big Data less expensive, more accessible, and easier to integrate with your existing data systems.

In the following sections, we will show you:

- The main features of the Hortonworks Data Platform and Informatica.
- Where Hadoop and the Hortonworks Data Platform fit in with Informatica as part of an integrated data management solution.
- How you can use Hortonworks and Informatica for data exploration, advanced analytics, and data refinery.

The Hortonworks Data Platform

The Hortonworks Data Platform (HDP) is an enterprise-grade, hardened Apache Hadoop distribution that enables you to store, process, and manage large data sets.

Apache Hadoop is an open-source software framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed for high-availability and fault-tolerance, and can scale from a single server up to thousands of machines.

The Hortonworks Data Platform combines the most useful and stable versions of Apache Hadoop and its related projects in a single tested and certified package. Hortonworks offers the latest innovations from the open source community, and the testing and quality you expect from enterprise-quality software.

The Hortonworks Data Platform provides a 100% open source, enterprise-grade Hadoop distribution. We are committed to working within the 100% open source Apache Software Foundation with no commercial holdbacks. Other Hadoop vendors take a proprietary approach that can lead to closed interfaces and vendor lock-in.

The Hortonworks Data Platform is designed to integrate with your existing data applications, tools, and processes. With Hortonworks, you can refine, analyze, and gain business insights from both structured and unstructured data – quickly, easily, and economically.

Hortonworks: Key Features and Benefits

With the Hortonworks Data Platform, you can retain and process more data, join new and existing data sets, and lower the cost of data analysis. Hortonworks enables you to:

- **Retain as much data as possible.** Traditional data warehouses age, and over time will eventually store only summary data. Analyzing detailed records is critical to uncovering useful business insights.

- **Join new and existing data sets.** Enterprises can build large-scale environments for transactional data with analytic databases. But these solutions are not able to process nontraditional data such as social media data, server logs, clickstream data, machine sensor data, and geolocation data. Hortonworks enables you to incorporate both structured and unstructured data in one comprehensive data management system.

- **Archive data at low cost.** It is not always clear what portion of stored data will be of value for future analysis. Therefore, it can be difficult to justify expensive processes to capture, cleanse, and store that data. Hadoop scales easily, so you can store years of data without much incremental cost, and find deeper patterns that your competitors may miss.

- **Access all data efficiently.** Data must be readily accessible. Apache Hadoop clusters can provide a low-cost solution for storing massive data sets – with the information still readily available.

- **Apply basic data cleansing and data cataloging.** You can categorize and label all data in Hadoop with descriptive information (metadata), and enable integration with transactional databases and analytic tools. This reduces the time and effort required to integrate with other data sets.

- **Integrate with existing platforms and applications.** Hortonworks connects seamlessly with many leading analytic, data integration, and database management tools.

The Hortonworks Data Platform is the foundation for your next-generation enterprise data architecture – one that addresses both the volume and complexity of today's data.

Informatica: Data Integration for Hadoop

Informatica Corporation is the world's leading independent provider of data integration software. Companies around the globe depend on Informatica to fully leverage their information assets and gain competitive advantage with timely, relevant, and trustworthy data for their business goals.

Hadoop offers clear advantages in data processing power and storage to help you meet the challenges – and reap the benefits – associated with Big Data.

But deploying Hadoop so that it seamlessly integrates with existing information systems can be a major challenge. Many organizations lack the specialized skills and software tools needed to successfully integrate Hadoop with their existing applications and information management systems. Without successful integration with existing systems, Hadoop runs the risk of becoming yet another information silo.

Informatica provides a comprehensive data integration platform that enables you to integrate Hadoop with all of your existing information management systems and create a single unified data repository – or "data lake" – for your enterprise.

Informatica: Key Features and Benefits

- **Universal Data Access** – Companies using Hadoop often face challenges in combining their existing data systems with new types of data. With Informatica and Hadoop, you can combine transaction data from conventional applications with new types of interaction data such as social media data, server logs, clickstream data, machine/sensor data, and geolocation data.

- **Data Parsing and Exchange** – Hadoop can store a wide variety of data types. Before this data can be exchanged with other data management systems and applications, it must be parsed and transformed into a usable format. Data parsing can be complex because it is based on the type of data source, which may be hierarchical, binary, semi-structured, unstructured, or domain-specific. The Informatica platform features data parsers that enable you to extract information from any data type, and exchange that data with your other systems and applications. This eliminates the need for manual

data translation, and allows you to leverage the distributed processing power of Hadoop.

- **Processing in Hadoop** – Enterprise data system performance can be optimized by "pushing down" data transformation processing from traditional databases into Hadoop. This reduces the workload on legacy database systems while utilizing the parallel processing power of Hadoop to increase transformation processing performance and free up existing system resources. The Informatica platform features an extensive library of data transformations for Hadoop that can be used to easily refine unstructured Hadoop data for use in traditional repositories.

- **Metadata Management** – Informatica provides metadata management and auditability, enabling you to establish a standard taxonomy of metadata definitions that encompasses both business terms and technical definitions.

- **Data Quality and Data Governance** – Data quality helps ensure the accuracy of data used for analytics and reporting, as well as the trustworthiness of data used for risk mitigation and compliance. Data governance is a convergence of data quality, data management, data policies, business process management, and risk management. With data governance, you can exercise control over your data processes and methods. Informatica provides the capability to profile, cleanse, and manage data to increase data quality while effectively and securely managing data growth.

Informatica PowerCenter – Big Data Edition

Informatica PowerCenter Big Data Edition is highly scalable, high-performance enterprise data integration software that works with both emerging technologies and traditional data management infrastructures. It provides a safe on-ramp to Big Data, enabling your IT organization to integrate and analyze new sources and types of data.

With the Big Data edition, developers increase their productivity by moving away from hand coding to a no-code visual development environment. Data scientists and analysts can focus on Big Data insights—not on data integration. Your organization can use these Big Data insights to bring innovative products and services to market more quickly, and to improve the efficiency of your business operations.

Informatica PowerExchange for Hadoop

Informatica PowerExchange for Hadoop integrates Hadoop with the rest of your enterprise data systems to enable high-performance distributed computing.

Informatica PowerExchange provides native, high-performance connectivity to the Hadoop Distributed File System (HDFS). It enables your IT organization to take advantage of Hadoop's storage and processing power using your existing IT infrastructure and resources. PowerExchange for Hadoop can bring all of your data into Hadoop for data integration and processing. Fully integrated with Informatica PowerCenter, it moves data into and out of Hadoop in batch or real time using

universal connectivity to all data, including mainframe, databases, and applications, both on-premises and in the cloud.

One of the biggest challenges companies face when tackling Big Data projects is the limited number and high cost of skilled resources. Now you can quickly staff Hadoop projects with existing PowerCenter ETL/DI developers, and focus hard-to-find, expensive data scientists on breakthrough analysis. PowerCenter developers are already Hadoop data integration developers. There is no training required.

Using PowerExchange for Hadoop, your company can achieve faster time-to-value for Big Data Projects and deliver a more complete and trusted view of your data. As a result, your company can harness the power of Big Data to drive new insights and deliver competitive advantage.

Reference Architectures

Traditional Enterprise Data Architecture

Today, nearly every company already has some sort of database management system already in place. Generally, these environments are structured as follows:

- Data comes from a set of data sources – most typically from enterprise applications such as Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), and any custom applications used to gather data.

- That data is extracted, transformed, and loaded (ETL) into a data system such as a Relational Database Management System (RDBMS), an Enterprise Data Warehouse (EDW), or even a Massively Parallel Processing (MPP) system.

- A set of analytical applications – either packaged or custom – then access the data in those systems to enable users to garner insights from the data.

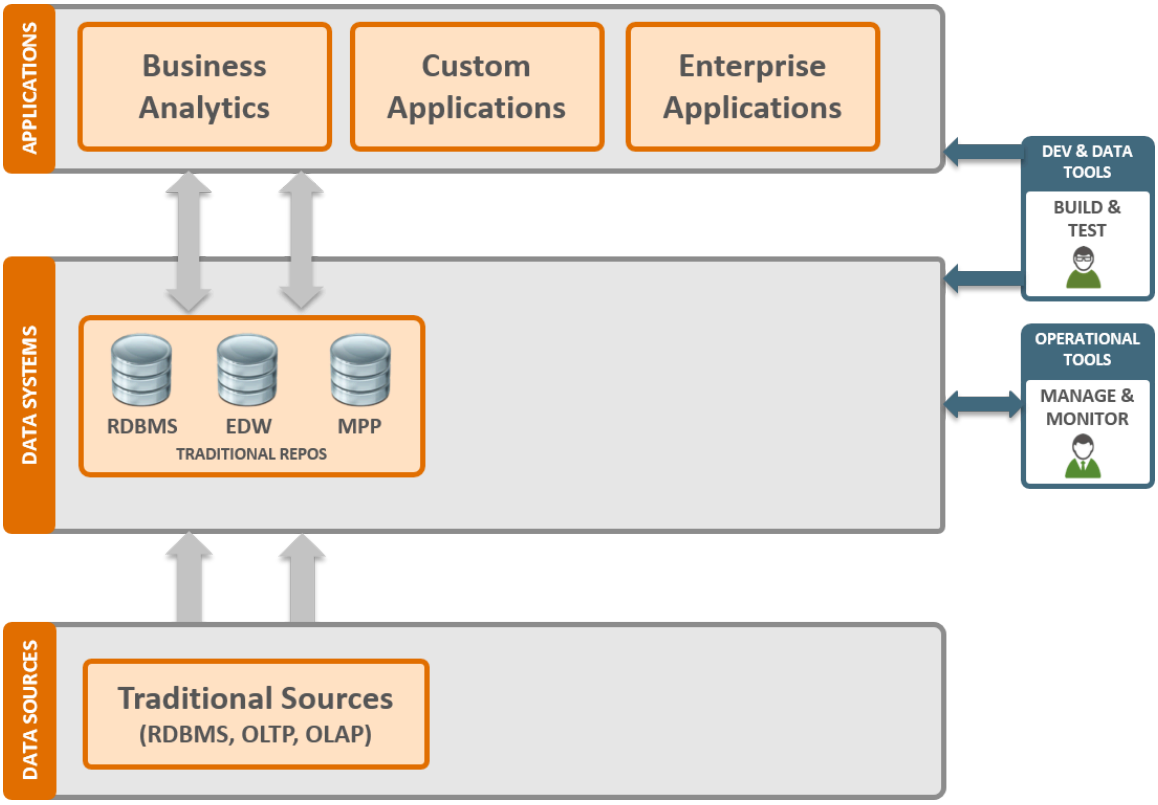

Figure 1: Traditional Database Architecture

Modern Data Architecture

In addition to traditional transactional data in analytic databases, companies must now gather, process, and analyze new unstructured data sets that are growing exponentially.

This new information can include text, images, machine-generated data, and online data from social media. It also includes data such as log files that were once thought to have relatively little value: too expensive to store and analyze. These new types of data are turning the focus from "data analytics" to "Big Data analytics" because so much insight can be gleaned from these new data sources for business advantage.

The Hortonworks Data Platform is increasingly being used to manage the massive amounts of these new types of data – as well as existing data – in an efficient and cost-effective manner.

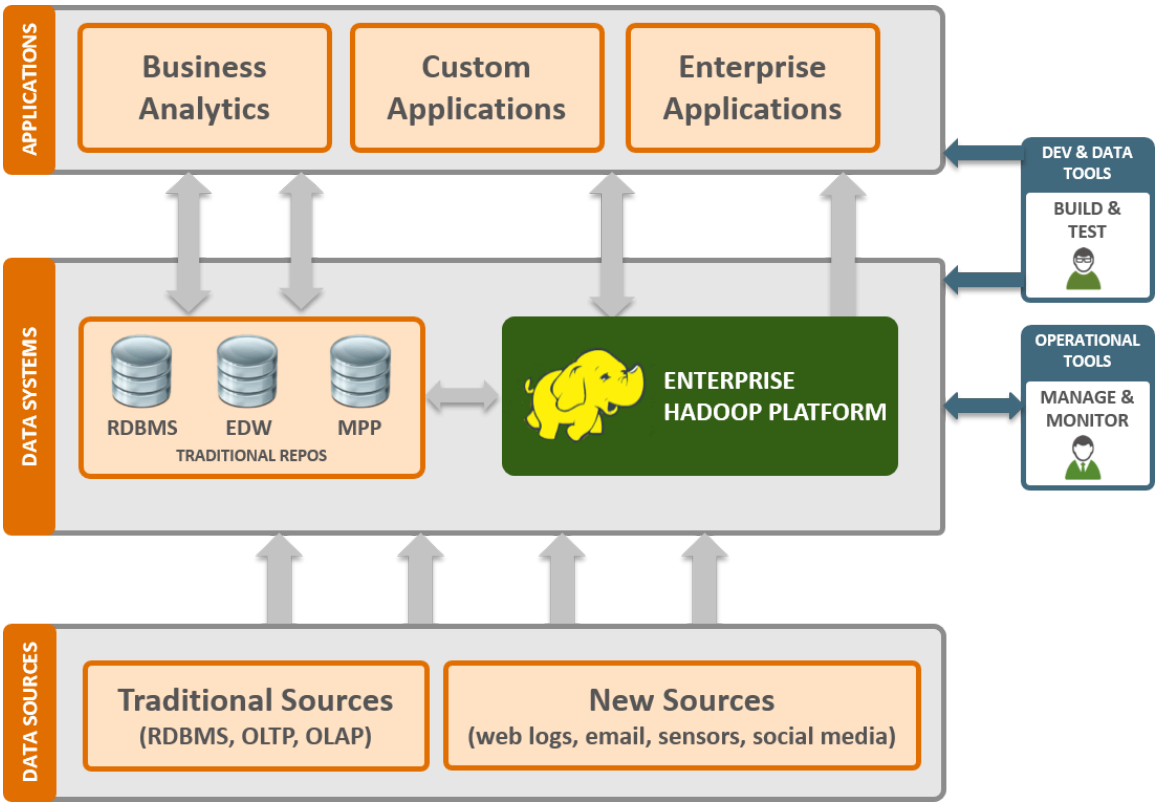

Figure 2: Modern Database Architecture

The Hortonworks Data Platform does not replace traditional data systems used for building analytic applications – the RDBMS, EDW and MPP systems – but is instead designed to integrate with and extend these systems.

By providing a framework to capture, store, and process vast quantities of both structured and unstructured data in a cost efficient and highly scalable manner, the Hortonworks platform is driving the creation of a new generation of enterprise database systems.

Informatica and the Hortonworks Data Platform

Big Data consists of Big Transaction Data, Big Interaction Data, and Big Data Processing:

- **Big Transaction Data** – Big Transaction Data is the steadily growing volume of financial, customer, product, and other information in legacy data systems such as an Enterprise Data Warehouses (EDW), Relational Database Management Systems (RDBMS), Enterprise Resource Planning (ERP) systems, Customer Relationship Management (CRM) applications, and other back-office solutions.

- **Big Interaction Data** – Big Interaction Data consists of the interactions of customers, partners, and competitors that affect your business. It is mainly web-based and external to your organization, and it is also growing at a phenomenal pace. Sources of Big Interaction Data include interactions from blog posts, discussion forums, and other social media sources, as well as clickstream data, call detail records, and information from competitor websites.

- **Big Data Processing** – A data processing platform such as Hadoop that can process the wide variety of Big Transaction and Big Interaction data types at scale.

The Informatica data integration platform lies at the convergence of Big Transaction Data, Big Interaction Data, and Big Data Processing.

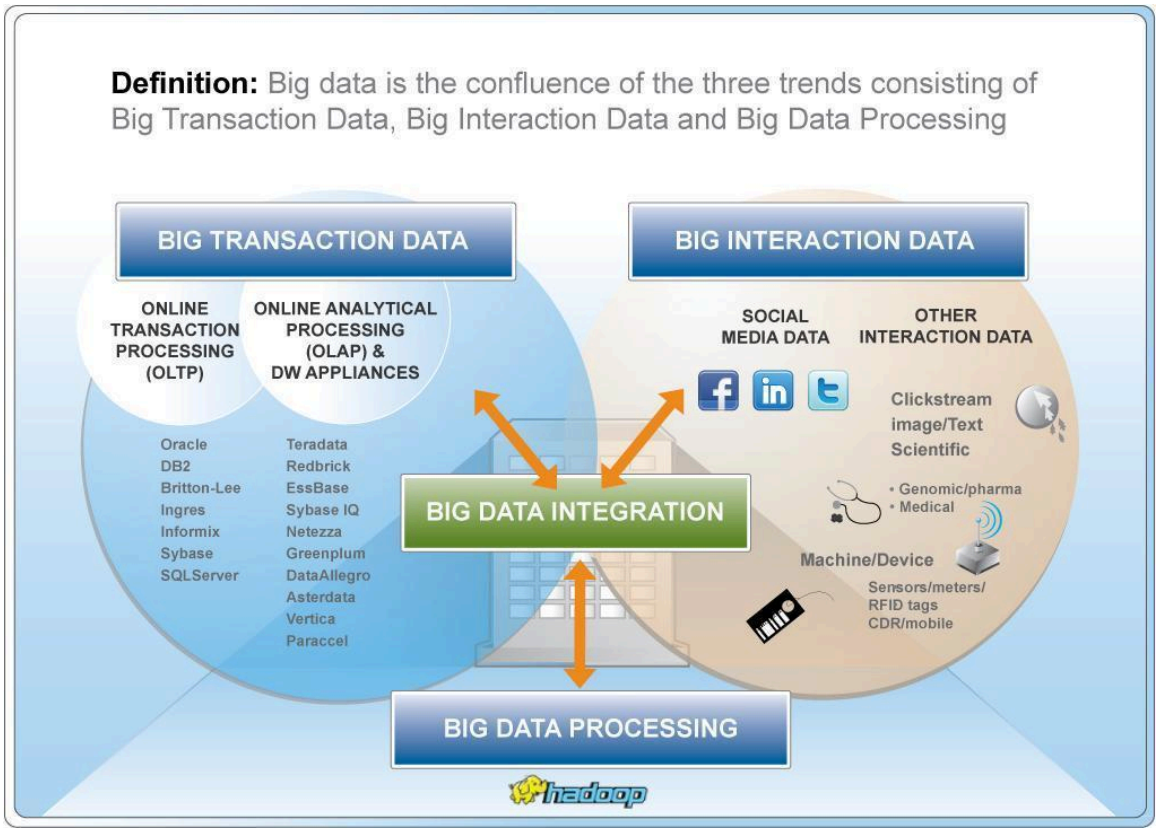

Figure 3: Big Data Integration

Informatica PowerExchange provides native, high-performance connectivity to the Hadoop Distributed File System (HDFS). PowerExchange for Hadoop can move any and all data into and out of Hadoop in batch or real time using universal connectivity to all data, including mainframe, databases, and applications, both on-premises and in the cloud.

With Informatica and Hortonworks, you can integrate Big Transaction and Big Interaction data into a single unified data repository – or "data lake". You can leverage the parallel processing power and scalability of Hadoop to refine and store new data types, and reduce the workload on legacy database systems.

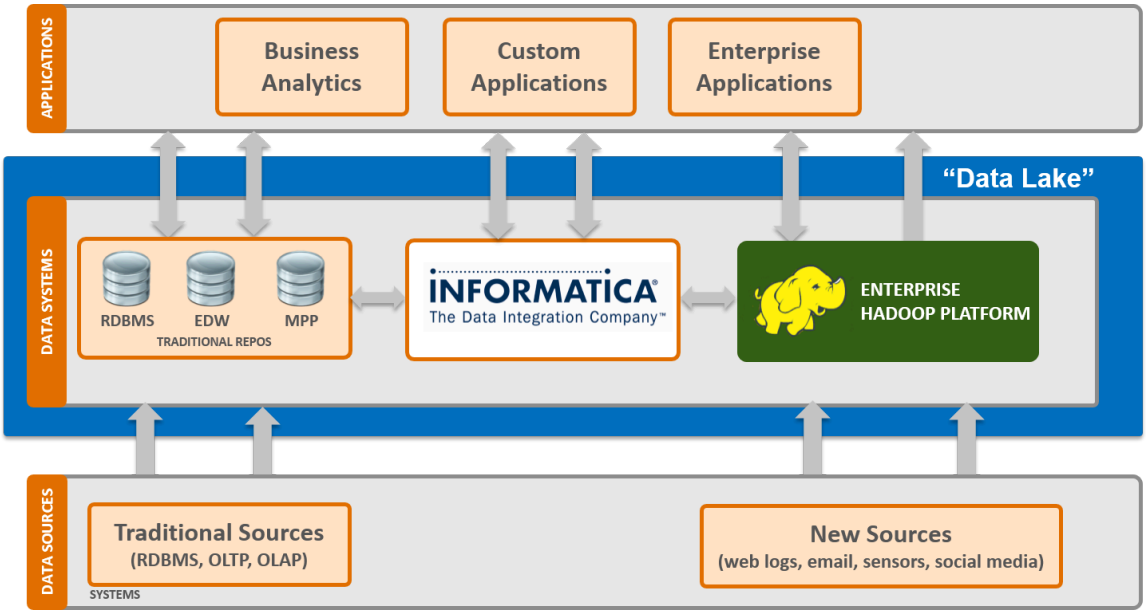

Figure 4: Big Data Lake

Use Cases

The following use case examples show you how to combine Informatica with the Hortonworks Data Platform:

- Data Refinery
- Data Exploration
- Application Enrichment

Hadoop as a Data Refinery

In the data refinery use case, organizations are using Hadoop to incorporate new unstructured data sources into their existing data systems and analytic applications. For example, an organization may have an application that provides a customer view based on their associated data in ERP and CRM systems, but would now like to use Hadoop to also incorporate website clickstream data to each customer view.

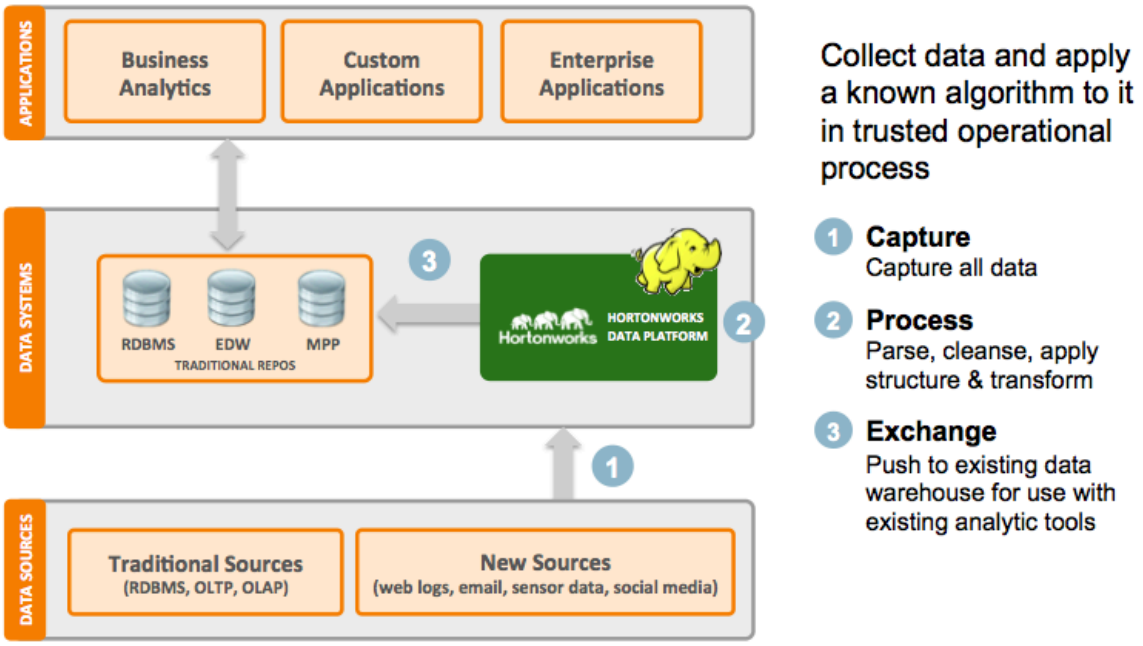

Figure 5: Hadoop as a Data Refinery

The key concept here is that Hadoop is being used to distill large quantities of data into something more manageable. The transformed data can then be loaded into existing data systems and accessed with traditional tools – but with a much richer data set. In some respects this is the simplest of all the use cases, in that it provides a clear path to value for Hadoop with very little disruption to existing data systems.

Informatica PowerExchange provides native, high-performance connectivity to the Hadoop Distributed File System (HDFS). The Informatica platform features data parsers that enable you to extract information from any data type and exchange that data with your other systems and applications, thereby eliminating the need for

manual data translation. Informatica provides the capability to profile, cleanse, and manage data to increase data quality while effectively and securely managing data growth. This enables your IT organization to take advantage of Hadoop's storage and processing power using your existing IT infrastructure and resources.

Another common use of Hadoop is to optimize enterprise data system performance by "pushing down" data transformation processing from existing databases into Hadoop. This reduces the workload on legacy database systems while using the parallel processing power of Hadoop to increase transformation processing performance and free up existing system resources. The Informatica platform features an extensive library of data transformations for Hadoop that can be used to easily refine unstructured Hadoop data for use in traditional repositories.

A Data Refinery in Practice

The refinery use case applies across market segments. In financial services we see organizations refine trade data to better understand markets, or to analyze and evaluate complex portfolios. Energy companies use Big Data to analyze consumption over geography to better predict production levels, saving millions. Retail firms (and virtually any consumer-facing company) often use data refinery to gain insight into online sentiment. Telecoms are using data refinery to extract details from call data records to optimize billing.

In any area with expensive, mission-critical equipment, we often find Hadoop being used for predictive analytics and proactive failure identification. In communications, this may be a network of cell towers. A restaurant franchise may monitor refrigerator data. Often, Hadoop is used to predict failure of these resources before it happens, thereby saving money and reducing down-time. In general, anywhere analytics are used, the refinery use case is present as it typically prepares data for use within Enterprise Data Warehouse (EDW) and Business Intelligence (BI) tools.

Data Exploration

Another common use case is data exploration. Many companies are capturing and storing a large quantity of new data types in Hadoop, and then exploring that data directly. Rather than using Hadoop as a staging area for processing and then putting the data into the EDW – as is the case in the Refinery use case – the data is left in Hadoop and then explored directly.

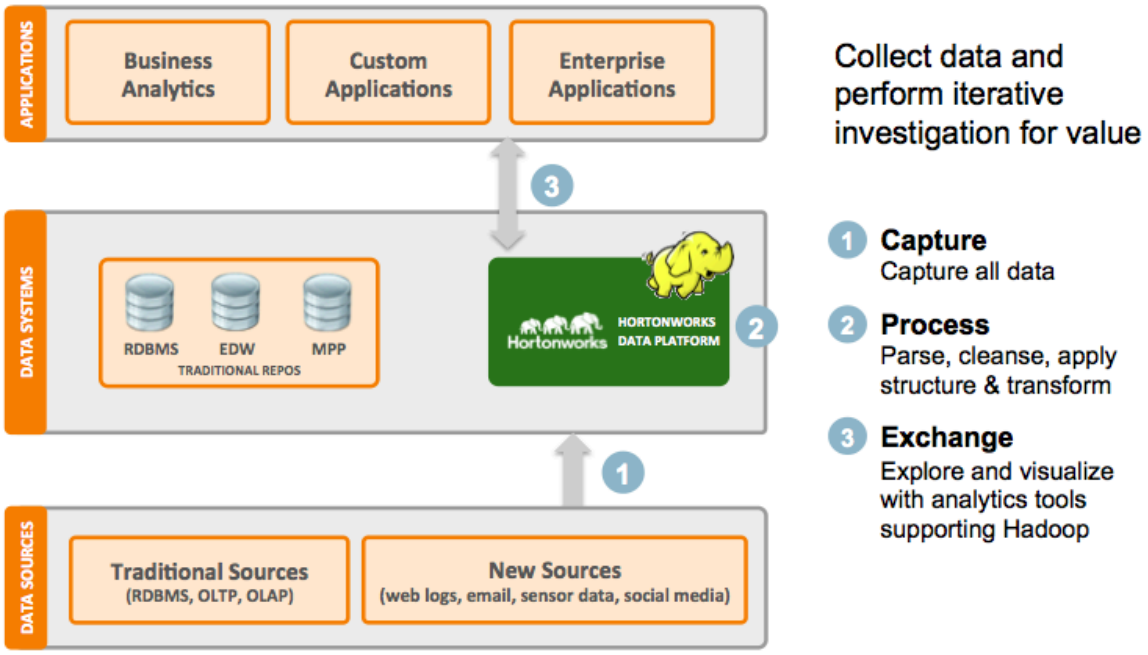

Figure 6: Hadoop and Data Exploration

The Data Exploration use case is often where Enterprises start – by capturing data that was previously discarded (web logs, social media data, etc.) and building entirely new analytic applications that leverage that data directly. One example might be a telecom using data exploration to capture huge quantities of machine data in order to predict when equipment is likely to fail. Given the huge quantities of data involved, they may not have been able to do this previously in a cost effective manner.

Data Science and Exploration

Nearly every industry can take advantage of the Data Exploration use case. In financial services, organizations can use data exploration to perform forensics, or to identify fraud. A professional sports team may use data exploration to analyze trades or their annual draft, as depicted in the movie *Moneyball*.

Many organizations are also using Hadoop to create a single comprehensive view of all of the data available for customers or products. This is proving to be of massive benefit, as it results in better customer service and increased product sales per customer as marketing programs are improved. Ultimately, data science and exploration are used to identify new insights and business opportunities to an extent that was not possible before Hadoop.

Application Enrichment

The third and final use case is application enrichment. In this scenario, data stored in Hadoop is used to impact the behavior of an application. For example, by storing all web session data, you can customize the experience for a customer when they return to the website. By storing this data in Hadoop, you can retain all session history and generate value from this accumulated data – for example, by providing a timely offer based on a customer's web history.

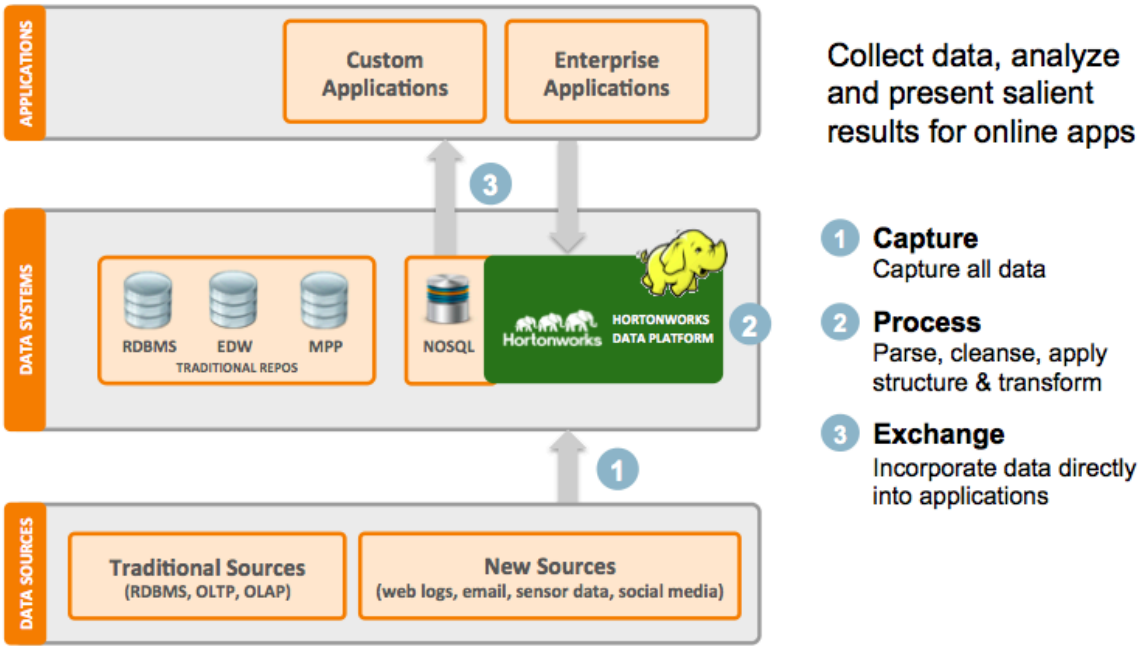

Figure 7: Application Enrichment with Hadoop

For many of the large web properties in the world – Yahoo, Facebook, and others – this use case is fundamental for their business. By customizing the user experience, they are able to differentiate themselves from their competitors. As one might expect, this is typically the last use case to be adopted – generally once organizations have become familiar with refining and exploring data in Hadoop. But at the same time, this also hints at how Hadoop usage can and will evolve over time to serve an ever greater number of applications that currently access traditional databases.

Informatica Vibe

For application enrichment, Informatica also offers the Vibe virtual data machine (VDM). Vibe enables application developers to map once and deploy anywhere:

- On premises or in the cloud.
- In databases, applications, middleware, or on a Hadoop cluster.
- In batch, request/response, or real time.

Vibe can deploy your logic regardless of the type, volume, or source of data, or the type of computing platform. Most importantly, if any of those elements change, Vibe lets you redeploy without recoding. Vibe can be embedded into applications, middleware infrastructure, and devices—wherever you need to access, aggregate, and manage data.

Enrichment: The Right Data at the Right Time to the Right Consumer

The most straightforward application enrichment use is the recommendation engines deployed by large web properties. These organizations analyze massive amounts of data to identify patterns and repeatable behavior, and then serve up the right content to the right person at the right time in order to increase sales conversation rates. In fact, this was the second use case for Hadoop at Yahoo as they realized Hadoop could help improve ad placement. This concept translates beyond the large web properties and is being used by more traditional enterprises to increase sales. Some brick and mortar organizations are even using these concepts to implement dynamic pricing in their retail outlets.

Getting Started with Hortonworks and Informatica

Here are a few links to help you get started with Hortonworks and Informatica:

- The Hortonworks Sandbox – This free download contains a stand-alone, single-node Hadoop environment, along with a set of hands-on, step-by-step tutorials.
- Informatica PowerExchange – PowerExchange enables you to retrieve all sources of enterprise data without developing custom data-access programs.