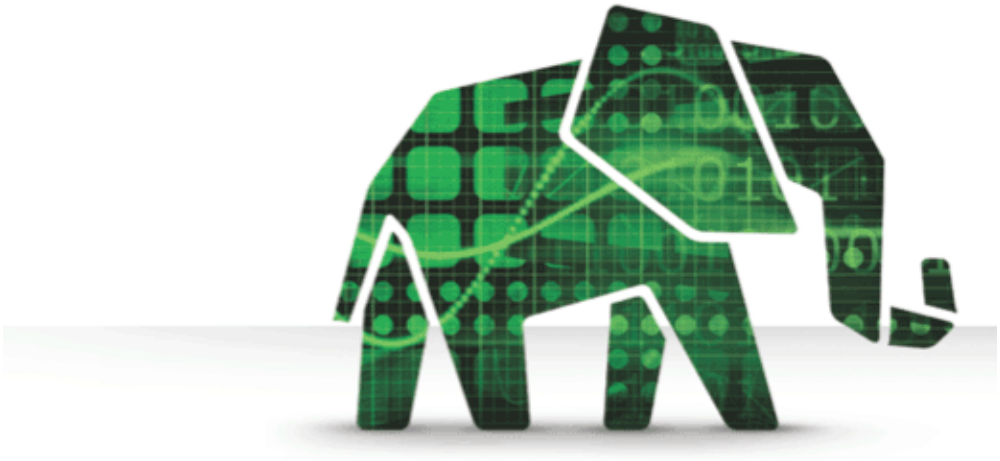
Hortonworks Enterprise Ready. Community Driven. Apache Hadoop™.

Business Value of Hadoop

...as seen through data



June 2013

http://www.hortonworks.com

Hortonworks

Big data… Hadoop Value

Hortonworks has helped many companies adopt and implement a Hadoop strategy over a number of years. While we have helped many learn about Hadoop technology, we have also learned a lot about data and how the broad enterprise can adopt and use Hadoop to create big data value.

While every organization is different their big data is often very similar. Hadoop, as a critical piece of an emerging modern data architecture, is collecting massive amounts of data across social media activity, clickstream data, web logs, financial transactions, videos, and machine/sensor data from equipment in the field.

These “new” data sources all share the common big data characteristics of volume (size), velocity (speed) and variety (type) and were sometimes thought of as low to medium value or even ‘exhaust data’: too expensive to store and analyze. And it is these types of data that is turning the conversation from “data analytics” to “big data analytics”: because so much insight can be gleaned for business advantage.

To be clear, these types of data are not strictly “new”—they have existed for some time. Text data, for example has been with us since before King Tut, but there was never very much of it (by today’s standards). With Hadoop, businesses are learning to see these types of data as inexpensive, accessible daily sources of insight and competitive advantage, all from what they were previously deleting, shredding or saving to tape.

While similar, each of these important types of big data can provide very different value. Let’s look at each.

Hortonworks

Clickstream Data

Clickstream data provides invaluable information to the Internet marketer. Analysts review the clickstream looking for which web pages visitors view, and in what order. This is the result of a succession of mouse clicks (the clickstream) that each visitor executes. Clickstream analysis can reveal how users research products and also how they complete their online purchases.

Clickstream Analysis

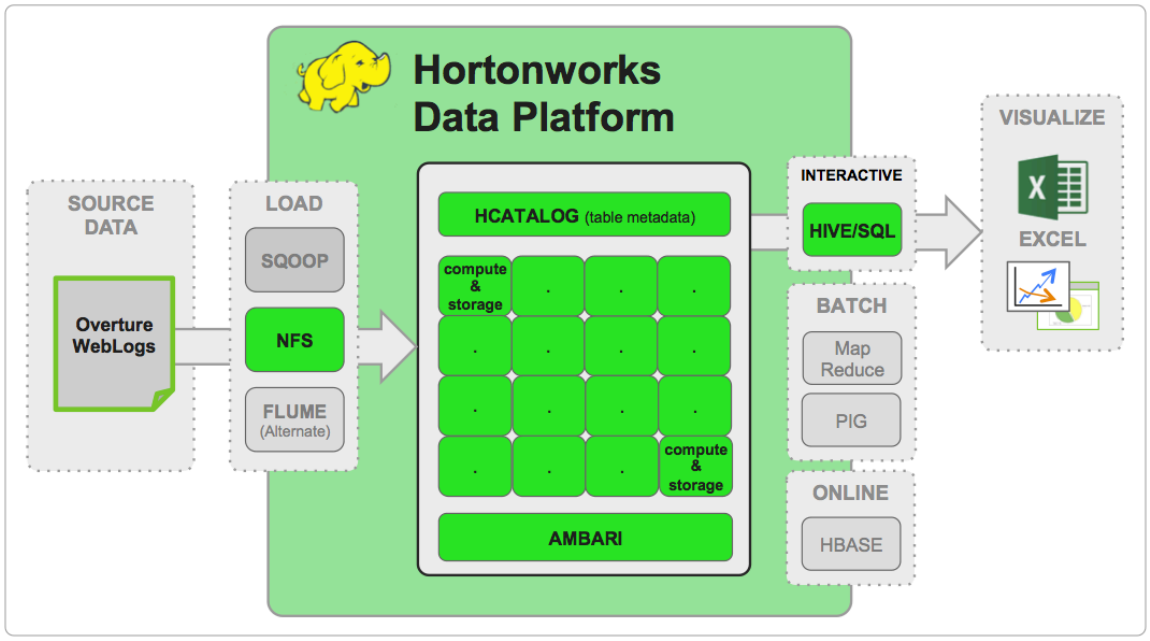
Clickstream data is often used to understand how website visitors research and consider purchasing products. With clickstream analysis, online marketers can optimize product web pages and promotional content to improve the likelihood that a visitor will learn about the products and then click the buy button. With a huge record of actual behavior patterns, web marketers can judge the effectiveness of different types of collateral and calls to action—with the confidence that their results are statistically significant and reproducible. For a particular product, a video might cause visitors to purchase more often than would a white paper. For another product, a white paper might outperform a datasheet.

Clickstream analysis sheds light on customer behavior during the actual purchase process. With patterns across millions of shopping carts, marketers can understand why clusters of customers abandon a cart at the same point in the purchase process. They can also see which products customers buy together, and then create pricing and promotional strategies to sell the product bundles that their customers define through their online behavior.

But clickstream data is not just for consumer web retailers. Any company can analyze the clickstream to see how well their website meets the needs of its customers, employees or constituents.

Hadoop Makes Your Clickstream More Valuable

Tools like Omniture and Google Analytics already help web teams analyze clickstreams, but Hadoop adds three key benefits. First, Hadoop can join clickstream data with other data sources like CRM data on customer demographics, sales data from brick-and-mortar stores, or information on advertising campaigns. This additional data often provides much more complete information than an isolated analysis of clickstream alone.



Secondly, Hadoop scales easily so that you can store years of data without much incremental cost, allowing you to perform temporal or year over year analysis on clickstream. You can save years of data on commodity machines and find deeper patterns that your competitors may miss. Storing all of the data in the Hadoop data lake makes it easy to join diverse datasets, in different ways for different purposes. And then to do it again in slightly different ways over time.

Finally, Hadoop makes website analysis easier. Without Hadoop, clickstream data is typically very difficult to process and structure. With Hadoop, even a beginning web business analyst can use Apache Hive or Apache Pig scripts to organize clickstream data by user session and then refine it to feed it to analytics or visualization tools. It is also easy to schedule recurring, periodic assessments and comparisons of behavior. Hadoop makes storing and refining the data easy, so the analyst can focus on discovery.

Sentiment Data

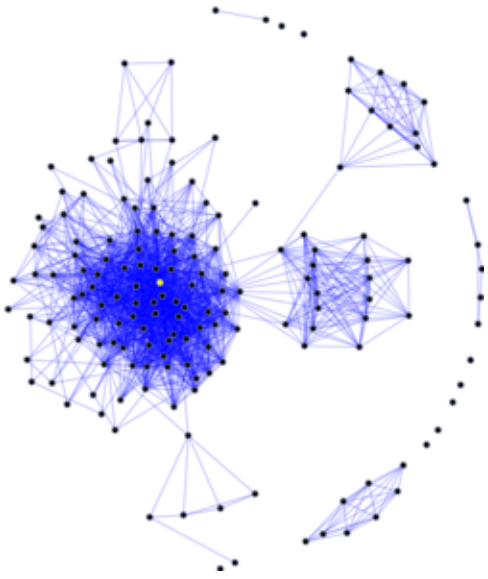
Sentiment data is unstructured data on opinions, emotions, and attitudes contained in sources like social media posts, blogs, online product reviews and customer support interactions. Organizations use sentiment analysis to understand how the public feels about something and track how those opinions change over time. An enterprise may analyze sentiment about products, services, competitors, or any other topic about which millions of people have an opinion.

Sentiment Analysis for Better Decisions

Every marketer wants to know how customers feel. They want to know how they feel today, and then track how those opinions change over time. Traditional high-touch market research tools, like customer surveys and focus groups, have three disadvantages: they are artificial, slow, infrequent and expensive. Sentiment analysis can augment or resolve some of these limitations.

- **Artificial**. Surveys and focus groups are artificial because participants are taken out of their natural environment and asked to answer hypothetical questions. Results may be biased (due to small sample size or survey methodology) and those results often take weeks to gather, tabulate and interpret.
- **Slow**. If it takes two months to complete the study and write the report, and then another month to review the results and come to a decision, consumers of the information begin to wonder whether the results are still relevant. And they may be right. Even if we have an accurate picture of what customers feel on January 1st, that may not be relevant by April 1st.
- **Expensive and Infrequent**. Researchers can create larger study groups to speed the study and reduce errors that come from a small sample size, but all of those participant incentives can become quite expensive. If one study raises new questions, answering those requires another study, with more time and expense.

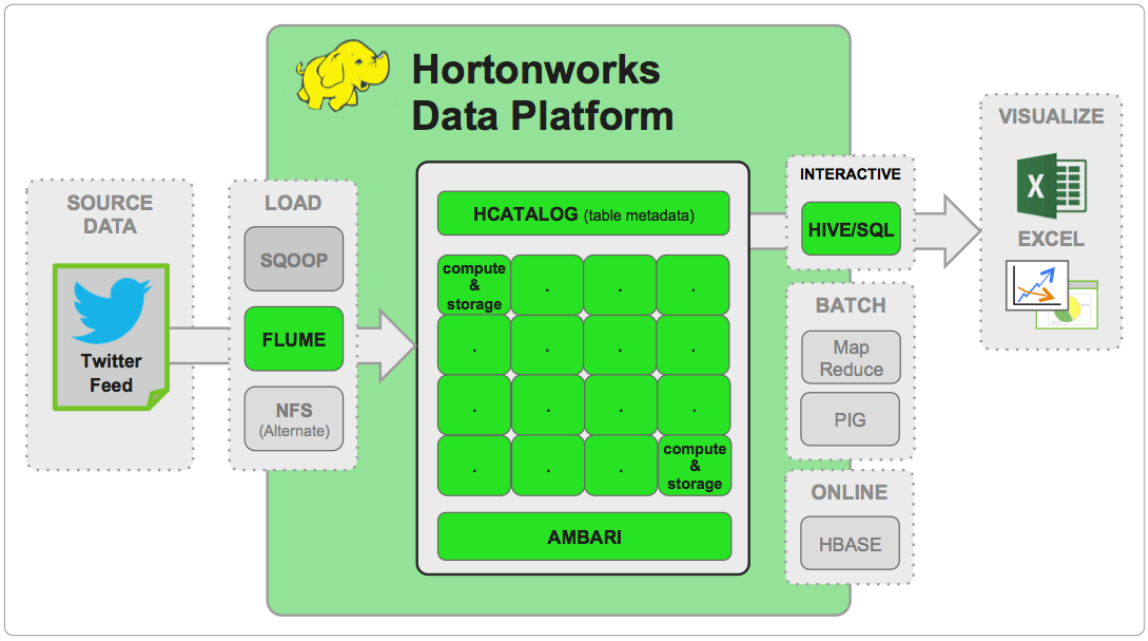
Sentiment data avoids limitations of traditional customer research methods and often augments them. Sentiment data analysis is real, immediate, repeatable and inexpensive. For example, Twitter users are not part of a study and they provide their opinions freely. They're just communicating with friends and colleagues, so they communicate candidly.



The value of online sentiment extends beyond just product marketing. It can feed a dynamic pricing schedule, to optimize prices and profits. Sentiment can be used to manage PR crises or public health emergencies. Outside of Twitter, Facebook or LinkedIn, sentiment analysis can also be performed on other "free" sources of feedback, like anonymous surveys or customer comment forms.

Hadoop Lets You to See Sentiment Today, and Store It For Later

Hadoop stores and processes huge amounts of complex, unstructured content; it is a natural fit for "messy" sentiment data. With Hadoop, social media posts can be loaded into the Hadoop Distributed Files System (HDFS) using Apache Flume for real-time streaming. Apache Pig and Apache Mahout organize the unstructured data and score sentiment with advanced machine learning methodologies.



Sentiment analysis *quantifies the qualitative* views expressed in social media. Researchers need big data to do this reliably. Ten tweets is just opinion. A million tweets tell you how most people feel about something at a given point in time.

Here's how it works. Words and phrases are assigned a polarity score of positive, neutral or negative. By scoring and aggregating millions of interactions, analysts can judge candid sentiment at scale, in real time.

After scoring sentiment, it is important to join the social data with other sources of data. The HDFS data lake makes those data joins easy and reproducible. CRM, ERP, and clickstream data can be used to attribute what was previously anonymous or semi-anonymous sentiment to a particular customer or segment of customers. All of this can be done in Hadoop, and then the results can be visualized with business intelligence tools like Microsoft Excel, Platfora, Splunk or Tableau.

Create Social Graphs

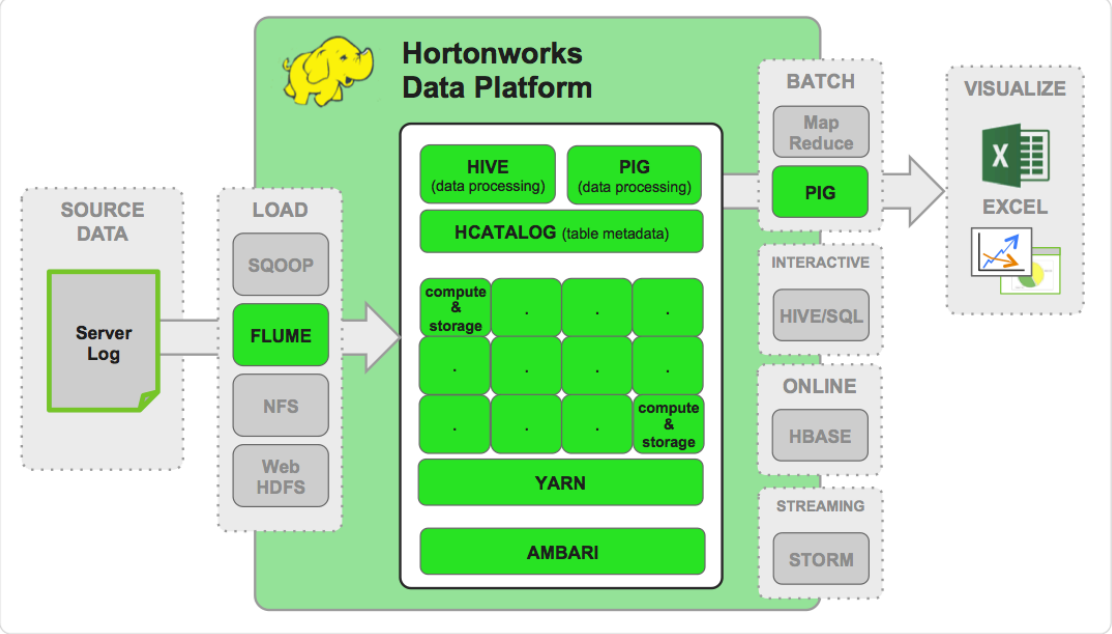
Another more complex and rather mature approach to sentiment analysis is the creation of a social graph. Often the loudest individual in a social circle is not the most influential. Separating the "loud" from the "trusted" influencers of sentiment is important. You want to sway or attract the right people to speak positively in a social circle about a brand or experience.

Again, Hadoop is a great way to not only store but also to process the enormous amount of rapidly changing information required to create and update social graphs.

Server Log Data

Large enterprises build, manage and protect their own proprietary, distributed information networks. Server logs are the computer-generated records that report data on the operations of those networks. They are like the EKG readings for the network: when there's a problem, it's one of the first places the IT team looks for a diagnosis. But no IT administrator sits down to read server logs with his morning coffee. The volume is massive, and most often those logs don't matter. However, every once in a while a handful of those logs can be very, very important.

The two most common use cases for server log data are network security breaches and network compliance audits. In both of these cases, server log information is vital for both rapid, efficient problem resolution and also longer-term forensics and resource planning.



Hadoop Helps You Protect Your Network Security

Barely a week goes by without news of a high-profile network breach by malicious individuals and groups. Enterprises and government agencies invest vast sums on antivirus and firewall software to protect their networks from malware and outside attacks, and those solutions usually work. But when security fails, Hadoop helps large organizations understand and then repair the vulnerability quickly and facilitates root cause analysis to create lasting protection.

Often, companies don't know of system vulnerabilities until they've already been exploited. So rapid detection, diagnosis and repair of the intrusion are critical. For example, in 2011, Sony's PlayStation Network was attacked *over a three-day period*. On the fourth day, Sony suspended the network, which remained offline for nearly a month. Names, birthdays and credit card numbers from nearly 25 million account holders were stolen. Sony announced the data breach six days after suspending the network. According to Sony, they did not notify the public earlier because they needed the time "to understand the scope of the breach, following several days of forensic analysis by outside experts."

Hadoop can make that type of forensic analysis faster. If an IT administrator knows that server logs are always flowing into the Hadoop data lake, to join other types of data, he can establish standard, recurring processes to flag any abnormalities. He can also prepare and test data exploration queries and transformations, for easy use when he suspects an intrusion.

Hadoop Helps Prepare for IT Compliance Audits

Of course, regulators write laws to prevent crises like the one suffered by Sony. The following compliance standards require organizations to monitor networks in real-time, ensure high levels of security for their confidential assets and provide network compliance audit reports to auditors:

- Payment Card Industry Data Security Standards (PCI DSS)
- Sarbanes Oxley (SOX)
- Health Insurance Portability and Accountability Act (HIPAA)
- Federal Information Security Management Act (FISMA)
- The Gramm-Leach-Bliley Act (GLBA)

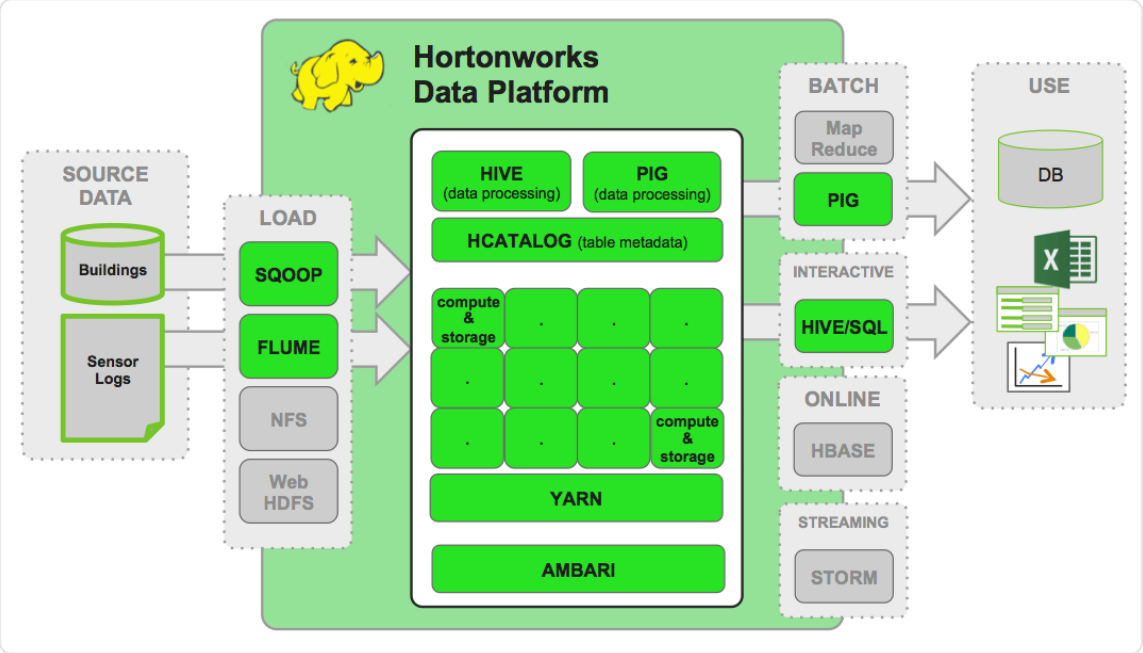
Regulatory bodies also require organizations to retain log data for long periods, allowing auditors to authenticate security incidents by checking the audit trails from the log data. Storing years of that log data in relational databases can be expensive. Hadoop can be a persistent, low cost platform to show compliance.

Sensor Data

From your refrigerator and coffee maker to the smart meters on the back of our homes, sensor data is everywhere. It is created by the machinery that runs assembly lines and the cell towers that route our phone calls. It is net new data that is increasing exponentially in the information age. We sometimes measure and report data about machines, but this does not scale well and it can be prone to errors. In some cases, it is impossible for humans to collect the data. Think of a measurement from inside a petroleum pipeline on the frozen tundra every hour of the year. No human wants that job, but sensors can do it reliably, at a very low cost, and they never sleep.

Sensors also capture data on natural systems such as:

- Meteorological patterns captured by weather balloons;
- Drilling mechanics for oil wells;
- Weather and soil data for agricultural decisions; and
- Patient vital statistics during post-operative recovery.



Sensors Deliver Big Data, Hadoop Delivers Big Data Value

Hadoop solves two big challenges that currently limit the use of sensor data: its volume and its structure.

Sensors measure and transmit small bits of data efficiently, but they are always on. So as the number of sensors increases (and time passes) the bytes or kilobytes from each sensor can soon add up to petabytes. With traditional data storage platforms, that stream of data is a liability. The cost to store the data might cause an organization to limit its collection (by restricting the number of sensors) or to limit its storage (by discarding data above a certain amount or after a particular duration). Hadoop stores this data more efficiently and the economics more favorable. Hadoop changes big sensor data from a liability to an asset.

Sensor data also tends to be flat and unstructured at time of collection and is typically generated by a mechanical, repetitive process. Apache Hive can transform the sensor readings according to its metadata (such as time, date, temperature, pressure or tilt). The data is then presented in HCatalog in a more familiar tabular format, even though the underlying data still exists in HDFS, in its original form.

Using Hadoop for Predictive Analytics and Proactive Maintenance

The ability to predict equipment failure (and respond proactively) is extraordinarily valuable, because it is far less expensive to do preventative maintenance than it is to pay for emergency repair or replacement equipment under duress. If a restaurant's refrigerator fails, the franchise loses thousands of dollars in spoiled food and a day's revenue. Fixed assets such as cellular transmission towers are difficult and expensive to replace, yet they exist to transmit data, so sensors can transmit diagnostic data that helps prolong the life of those assets. Algorithms can process massive amounts of sensor signals to identify previously invisible, subtle patterns indicating when an inexpensive repair is likely to prevent a costly replacement.

Of course, the cost benefit difference is much greater when human life is at risk. Since 2007, Children's Hospital Los Angeles has collected sensor data from its pediatric intensive care units, sampled from each patient every 30 seconds. This dataset includes more than one billion individual measurements. Doctors plan to use this data to diagnose and predict medical episodes with greater precision. According to one of the researchers, the difficulty is to find

medically useful patterns because “there are an infinite number of trivial patterns, such as people who tend to have babies are female and people over six-feet tall are over five-feet tall.”

With Hadoop, it is much easier to refine the data and explore it to find the meaningful patterns. Tools like Apache Hive and Apache Pig can be used to join various data sets together, combine that with data on health outcomes, and then refine it all into a master dataset that includes the important patterns and excludes the trivial ones.

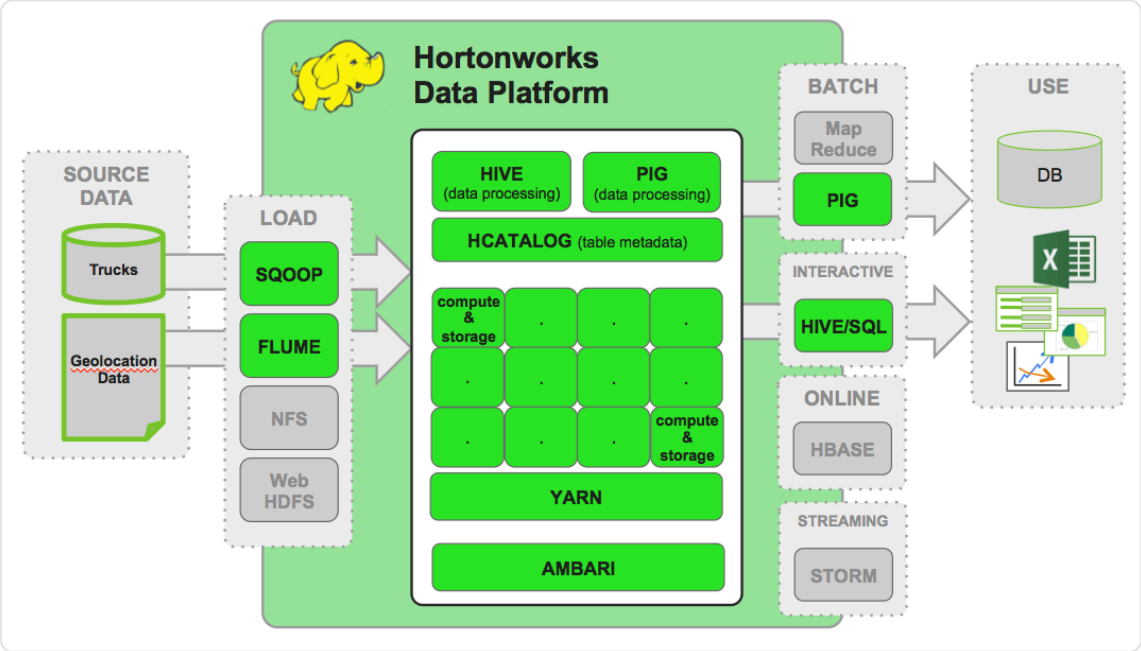
As Hadoop extends the storage and analysis of big data to processes beyond commercial Internet use cases, it can augment and assist other efforts to save lives through the prediction, identification and prevention of potential tragedy.

Location Data

Location data is a sub-variant of sensor data, since the device senses its location and transmits data on its latitude and longitude at pre-defined intervals. This is truly a new form of data, since it did not exist (outside of highly specialized military and aerospace applications) until ten years ago.

Global positioning systems (GPS) became widely available in the late 1990s. Today, smart phones can capture and transmit precise longitude and latitude at regular time intervals—the sensor is connected to the communication network in the same device. Consumer-driven businesses want to use this data to understand where potential customers congregate during certain times of the day. Local governments or chambers of commerce want to know when individuals come from abroad, so they can welcome them and suggest local tourist attractions and vendors.

Another promising use for this location data is on delivery vehicles. Companies like UPS that make residential deliveries can use fine-grain location data, taken at frequent intervals, to optimize driver routes. This leads to faster delivery times, lower fuel costs and reduced risk of accidents. Logistics companies want to know where all of its mobile assets are, at any time of day.



Big Data Helps Big Rigs Arrive On Time, Safely

Long-haul truckers provide a particularly interesting use case for location data. Sensors on big rigs can be used to measure and report both “macro-movement,” such as average speed over ten mile segments, as well as “micro-events” like sudden braking, swerving or unsafe proximity to other vehicles. This big data on location can be used for:

- Reduction in idling to save fuel
- Compliance with federal laws mandating minimum hours off duty and maximum hours on duty
- Accident prevention through detection of unsafe driving patterns.

US Xpress is one of the largest trucking companies in the United States. They developed a system called DriverTech that employs tens of thousands of sensors to stream data into a Hadoop cluster. According to an article in Computer Weekly “The stream of data is based on 900 data elements from tens of thousands of trucking systems - sensor data for tyre and petrol usage, engine operation, geo-spatial data for fleet tracking, as well as driver feedback from social media sites.”

Hadoop allows US Xpress store huge volumes of location data for many years, but just as importantly, they can join that location data with other data streams flowing into their data lake to form a more holistic dataset about what is happening with their drivers and vehicles. DriverTech saves US Xpress over $6 million per year in known savings from fuel and equipment, not to mention unknown savings from any accident or injury it may be preventing.

Unstructured Text

One Library of Congress for Every Enterprise

Free-form text in the enterprise is present in work products such web pages, emails, documents, proposals, patents and literature. Text is one of the oldest forms of data, but it has always been stored and consumed in relatively small quantities. In today's businesses almost everyone has the capacity (and the responsibility) to create free-form text, so the amount of this type of data continues to grow exponentially. The same relationship holds as with sensor data: a kilobyte here, a kilobyte there, and pretty soon you're talking about big data. Financial services, government, retail, wholesale, manufacturing and healthcare all generate significant amounts of free-form text and numerical data.

According to the US Library of Congress website, the library had 235 terabytes of data in storage in April 2011. The May 2011 McKinsey & Company report *Big data: The next frontier for innovation, competition, and productivity* looked at US 2009 companies with over one thousand employees and found that in 15 of 17 industries, the average data stored per company was greater than that in the Library of Congress. If we use the oft-quoted assumption that 80% of stored enterprise data is unstructured, then ***the average company in 14 of 17 sectors stores more data than does the US Library of Congress***. But very few organizations can use it. It's stored in unconnected data islands. It's not searchable or indexed.

Most enterprises are sitting on a gold mine of free-form text without any easy way to extract it.

Sector	Stored Terabytes per Firm (>1000 employees)	Estimate of Unstructured Terabytes per Firm (using 80% share assumption)	Library of Congress Equivalents per Firm
Securities and investment	**3,866**	**3,093**	13
Banking	**1,931**	**1,545**	7
Communications and media	**1,792**	**1,434**	6
Utilities	**1,507**	**1,206**	5
Government	**1,312**	**1,050**	4
Discrete manufacturing	**967**	**774**	3
Insurance	**870**	**696**	3
Process manufacturing	**831**	**665**	3
Resource industries	**825**	**660**	3
Transportation	**801**	**641**	3
Retail	**697**	**558**	2
Wholesale	**536**	**429**	2
Health care providers	**370**	**296**	1
Education	**319**	**255**	1
Professional services	**278**	222	<1
Construction	231	185	<1
Consumer & recreational services	150	120	<1

Source: "Big data: The next frontier for innovation, competition, and productivity". McKinsey Global Institute, May 2011.

Freeing Free-Form Text

There are three steps to setting that data free to use in your organization: extraction; summarization and analysis. The well-known "Word Count" MapReduce job – the "Hello World" of Hadoop - provides a simple example of these three steps. The word count job can count the number of times every word recurs in a huge text document.

Here are the three steps for that text processing workload:

1. ***Extraction:*** each document is split up into its individual words, and each word is counted by the *map* function
2. ***Summarization:*** the framework groups all identical words with the same key and feeds them to the same call to *reduce*
3. ***Analysis:*** for a given word, the function sums all of its input values to find the total appearances of that word

There may be some real-life enterprise use cases for counting the number of times that certain words appear in the company's email, but it is difficult to imagine that they would be particularly valuable. The real value lies in identifying particular phrases and then analyzing the occurrence of those phrases in relationship to other sources of data. The following three use case summaries show specific applications of Hadoop for analyzing free-form text.

Hadoop Text Uses for Lawyers, Insurance Underwriters, and Bankers

Text Analysis for Legal Discovery

The process of legal reasoning and argument is largely based on information extracted from a variety of documents. Lawyers are paid large hourly sums to analyze documents to build their cases. Hadoop can help make this manual review process more efficient. Firms can store the documentation in Hadoop, and then analyze that text *en masse* using processes like natural language processing or text mining. This allows legal researchers to search documents for important phrases and then use Hadoop ecosystem solutions to analyze relationships between those and other phrases. This preliminary analysis optimizes the lawyer's time reviewing the text, so she can read the parts that really matter.

Text Analysis for Insurance Underwriting

Insurance companies hold massive amounts of unstructured, text-based claim data. They can also access other structured and unstructured data sets (public and private) that they can join with claim data to improve their assessment of risk.

This helps insurance firms reduce their "moral hazard", which is essentially a data scarcity problem. Moral hazard is the situation where riskier applicants apply for insurance because they know they might need it, while safer applicants stay out of the market because they assess their risk as low.

Data analysis built on the Hadoop data lake reduces the gap between what the insurer knows and what the policyholder knows, so that each party has a more similar view of the risk. Insurance companies use more data, from more sources, for longer, so their predictive power grows over time.

Text Analysis for Application Risk Screening

Banks take thousands of loan and checking account applications daily. To assess the risk of these applications, bankers use 3rd party risk reports. Bankers can (and do) override these

recommendations to open accounts and fund loans. On average, these decisions lead to higher account charge-offs and loan defaults.

Senior managers want to correct risky behavior by sanctioning individuals, updating policies, improving training or identifying internal fraud. To do this, they need to match all banker decisions with the multiple sources of information that bankers used to make those decisions. The Hadoop data lake can store all available data that might have predictive power for improving banker decisions.

These three use specific use cases can be generalized to any industry that collects large volumes of text data that is then used to make high-value decisions.

Hadoop in the Enterprise:

An Emerging Data Architecture

Today, most enterprises utilize one or more analytical applications to manage the day-to-day business. The traditional approach to analytics is relatively well understood and is depicted in Figure 1. In this approach:

- Data comes from a set of data sources; typically from enterprise applications such as ERP, CRM, or other transactional applications that power the business
- That data is extracted, transformed, and loaded into a data repository such a relational database or a data warehouse
- A set of analytical applications – either packaged (e.g. SAS) or custom – are then used to manipulate the data in the repository, producing reports or visualizations that deliver insights to business users

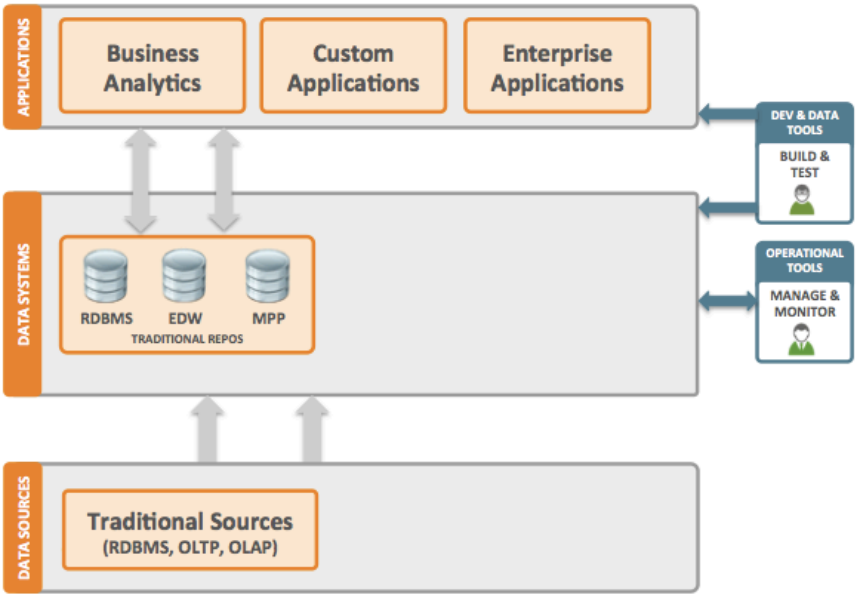


Figure 1: Traditional Enterprise Data Architecture

With the introduction of these new data sources enterprises are forced to think differently. The emerging data architecture most commonly seen introduces Apache Hadoop to handle these new types of data in an efficient and cost-effective manner. Hadoop does not replace the traditional data repositories used in the enterprise, but rather is a complement.

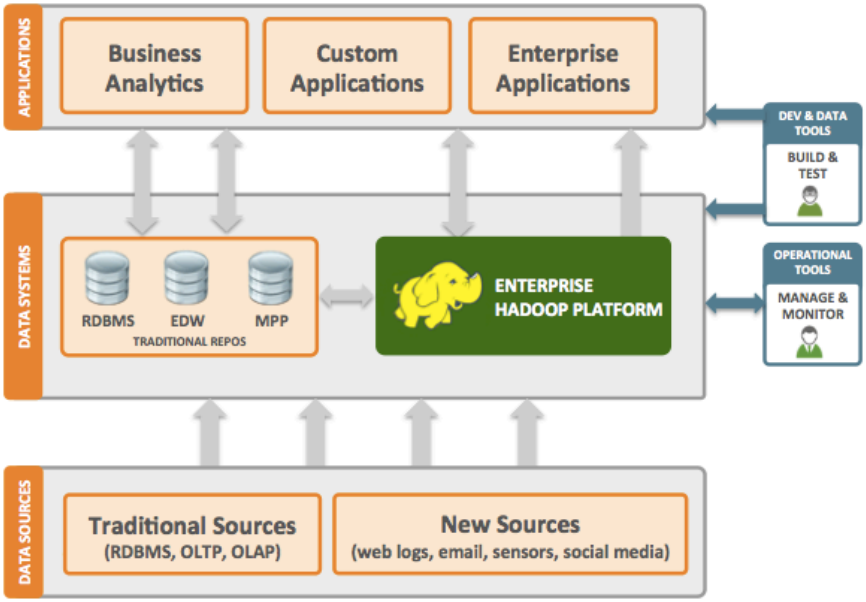
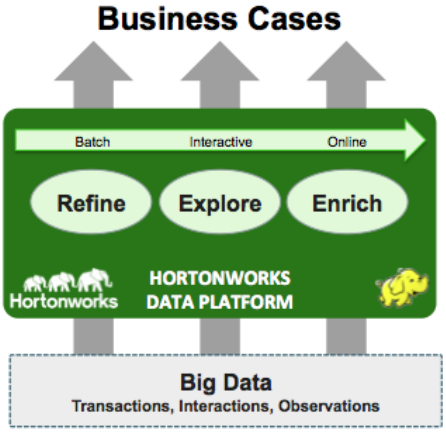


Figure 2: The Emerging Big Data Architecture



With the availability of enterprise-ready Apache Hadoop distributions like Hortonworks Data Platform, enterprises are embarking on a wide variety of Hadoop-based big data projects. While examples are many, three patterns of use have emerged:

- **Refine.** The capturing of a variety of data sources into a single platform where value can be added by refining the data into formats that are more easily consumed by platforms such as a data warehouse
- **Explore.** Interactively examining huge volumes of data to identify patters and relationships in order to unlock opportunities for business value
- **Enrich.** Enabling organizations to apply advanced analytics to the data they are collecting via log files or social media streams in order to make other applications, such as mobile commerce applications, more "intelligent" with respect to the experience they deliver

Hortonworks Data Platform

Hortonworks Data Platform (HDP) is the only 100% open source data management platform for Apache Hadoop. HDP allows enterprises to capture, process and share data in any format and at full scale. Built and packaged by the core architects, builders and operators of Hadoop, HDP includes all of the necessary components to manage a cluster at scale and uncover business insights from existing and new big data sources.

Hortonworks Data Platform is the most stable and reliable Apache Hadoop distribution available. It delivers the advanced services required for enterprise deployments without compromising the cost-effective and open nature of Apache Hadoop, including:

- **Data Services** required to store, analyze and access data
- **Operational Services** required to manage and operate Hadoop
- **Platform Services** such as high availability and snapshots which are required to make Hadoop enterprise grade

Hortonworks

Enterprise ready. Community driven. Apache Hadoop™.

At Hortonworks, we believe that Hadoop is an enterprise viable data platform and that the most effective path to its delivery is within the open community. To this end, we build, distribute and support a 100% open source distribution of Apache Hadoop that is truly enterprise grade and follow these three key principles:

1. **Identify and introduce enterprise requirements into the public domain**
2. **Work with the community to advance and incubate open source projects**
3. **Apply Enterprise Rigor to deliver the most stable and reliable distribution**

Hortonworks, Hadoop and You

We encourage you to follow us, get engaged with our learning tools, or download the HDP Sandbox, a single node installation of HDP that can run right on your laptop. Hadoop has the potential to have a profound impact on the data landscape, and by understanding the basics, you can greatly reduce the complexity.

Download the Hortonworks Sandbox to get started with Hadoop today

About Hortonworks

Hortonworks is a leading commercial vendor of Apache Hadoop, the preeminent open source platform for storing, managing and analyzing big data. Our distribution, Hortonworks Data Platform powered by Apache Hadoop, provides an open and stable foundation for enterprises and a growing ecosystem to build and deploy big data solutions. Hortonworks is the trusted source for information on Hadoop, and together with the Apache community, Hortonworks is making Hadoop more robust and easier to install, manage and use. Hortonworks provides unmatched technical support, training and certification programs for enterprises, systems integrators and technology vendors.

Hortonworks

3460 West Bayshore Rd.
Palo Alto, CA 94303 USA

US: 1.855.846.7866
International: 1.408.916.4121
www.hortonworks.com

Twitter: twitter.com/hortonworks
Facebook: facebook.com/hortonworks
LinkedIn: linkedin.com/company/hortonworks