cloudera
Ask Bigger Questions

WHITE PAPER

TEN COMMON HADOOPABLE PROBLEMS

Real-World Hadoop Use Cases

Table of Contents

Introduction

Apache Hadoop, the popular data storage and analysis platform, has generated a great deal of interest recently. Large and successful companies are using it to do powerful analyses of the data they collect. Hadoop offers two important services: It can store any kind of data from any source, inexpensively and at very large scale, and it can do very sophisticated analysis of that data easily and quickly.

Hadoop is different from older database and data warehousing systems, and those differences can be confusing to users. What data belongs in a Hadoop cluster? What kinds of questions can the system answer? Understanding how to take advantage of Hadoop requires a deeper knowledge of how others have applied it to real-world problems that they face.

This paper presents ten real-world Hadoop use cases. These are, of course, examples. One or more may apply directly to your business, but they are only a small sample of the ways that companies are using Hadoop today.

Hadoop is a high-performance distributed data storage and processing system. Its two major subsystems are HDFS, for storage, and MapReduce, for parallel data processing.

What is Hadoop?

Hadoop is a data storage and processing system. It is scalable, fault-tolerant and distributed. The software was originally developed by the world's largest internet companies to capture and analyze the data that they generate. Unlike older platforms, Hadoop is able to store any kind of data in its native format and to perform a wide variety of analyses and transformations on that data. Hadoop stores terabytes, and even petabytes, of data inexpensively. It is robust and reliable and handles hardware and system failures automatically, without losing data or interrupting data analyses.

Hadoop runs on clusters of commodity servers. Each of those servers has local CPU and storage. Each can store a few terabytes of data on its local disk.

The two critical components of the Hadoop software are:

- **The Hadoop Distributed File System, or HDFS.** HDFS is the storage system for a Hadoop cluster. When data arrives at the cluster, the HDFS software breaks it into pieces and distributes those pieces among the different servers participating in the cluster. Each server stores just a small fragment of the complete data set, and each piece of data is replicated on more than one server.
- **A distributed data processing framework called MapReduce.** Because Hadoop stores the entire dataset in small pieces across a collection of servers, analytical jobs can be distributed, in parallel, to each of the servers storing part of the data. Each server evaluates the question against its local fragment simultaneously and reports its results back for collation into a comprehensive answer. MapReduce is the plumbing that distributes the work and collects the results.

  Both HDFS and MapReduce are designed to continue to work in the face of system failures. The HDFS software continually monitors the data stored on the cluster. If a server becomes unavailable, a disk drive fails or data is damaged, whether due to hardware or software problems, HDFS automatically restores the data from one of the known good replicas stored elsewhere on the cluster. When an analysis job is running, MapReduce monitors progress of each of the servers participating in the job. If one of them is slow in returning an answer or fails before completing its work, MapReduce automatically starts another instance of that task on another server that has a copy of the data. Because of the way that HDFS and MapReduce work, Hadoop provides scalable, reliable and fault tolerant services for data storage and analysis at very low cost.

Hadoop automatically detects and recovers from hardware, software and system failures.

Hadoop stores any type of data, structured or complex, from any number of sources, in its natural format. No conversion or translation is required on ingest. Data from many sources can be combined and processed in very powerful ways, so that Hadoop can do deeper analyses than older legacy systems. Hadoop integrates cleanly with other enterprise data management systems. Moving data among existing data warehouses, newly available log or sensor feeds and Hadoop is easy. Hadoop is a powerful new tool that complements current infrastructure with new ways to store and manage data at scale.

Enterprises today must store and analyze more data from more sources than ever before. Hadoop provides a flexible, low-cost, proven platform for long-term growth.

Why use Hadoop?

Hadoop solves the hard scaling problems caused by large amounts of complex data. As the amount of data in a cluster grows, new servers can be added incrementally and inexpensively to store and analyze it. Because MapReduce takes advantage of the processing power of the servers in the cluster, a 100-node Hadoop instance can answer questions on 100 terabytes of data just as quickly as a ten-node instance can answer questions on ten terabytes.

Of course, many vendors promise scalable, high-performance data storage and analysis. Hadoop was invented to solve the problems that early internet companies like Yahoo! and Facebook faced in their own data storage and analysis. These companies and others actually use Hadoop today to store and analyze petabytes – thousands of terabytes – of data. Hadoop is not merely faster than legacy systems. In many instances, the legacy systems simply could not do these analyses.

Hadoop delivers several key advantages:

- **Store anything.** Hadoop stores data in its native format, exactly as it arrives at the cluster. Translating data on arrival so that it fits into a fixed data warehouse schema destroys information. Because Hadoop stores data without forcing that transformation, no information is lost. Downstream analyses run with no loss of fidelity. Of course it is always possible to digest, analyze and transform data, but Hadoop allows the data analyst to choose how and when to do that.
- **Control costs.** Hadoop is open source software that runs on commodity hardware. That combination means that the cost per terabyte, for both storage and processing, is much lower than on older proprietary systems. As storage and analytic requirements evolve, a Hadoop installation can, too. Adding or removing storage capacity is simple. You can dedicate new hardware to a cluster incrementally, as required, and can retire nodes from one easily, too. As new analytic techniques are developed, they are easy to apply to new and existing data with MapReduce.
- **Use with confidence.** The Hadoop community, including both developers of the platform and its users, is global, active and diverse. Companies across many industries participate, including social networking, media, financial services, telecommunications, retail, health care and others.
- **Proven at scale.** You may not have petabytes of data that you need to analyze today. Nevertheless, you can deploy Hadoop with confidence because companies like Facebook, Yahoo! and others run very large Hadoop instances managing enormous amounts of data. When you adopt a platform for data management and analysis, you are making a commitment that you will have to live with for years. The success of the biggest Web companies in the world demonstrates that Hadoop can grow as your business does.

Hadoop makes it possible to conduct the types of analysis that would be impossible or impractical using any other database or date warehouse. Hadoop lowers costs and extracts more value from data.

Large amounts of complex data demand a new approach.

Recognizing Hadoopable problems

The nature of the data that enterprises must capture, store and analyze is changing.

It's complex.

Not all data fits neatly into the rows and columns of a table. It comes from many sources in multiple formats: multimedia, images, text, real-time feeds, sensor streams and more. Data formats generally change over time as new sources come on-line. Hadoop is able to store and analyze in its native format.

There's a lot of it.

Many companies are forced to discard valuable data because the cost of storing it is simply too high. New data sources make this problem much worse: people and machines are generating more data today than ever before. The EMC Digital Universe data growth study2 predicts nearly 45-fold annual data growth by 2020. Hadoop's innovative architecture, using low-cost commodity servers for storage and processing, stores large amounts of data inexpensively.

It demands new analytics.

Simple numerical summaries – average, minimum, sum – were sufficient for the business problems of the 1980s and 1990s. Large amounts of complex data, though, require new techniques. Recognizing customer preferences requires analysis of purchase history, but also a close examination of browsing behavior and products viewed, comments and reviews logged on a web site, and even complaints and issues raised with customer support staff. Predicting behavior demands that customers be grouped by their preferences, so that behavior of one individual in the group can be used to predict the behavior of others. The algorithms involved include natural language processing, pattern recognition, machine learning and more. These techniques run very well on Hadoop.

1 Alphabetical list of institutions using Hadoop for educational or production uses, PoweredBy, Hadoop Wiki, http://wiki.apache.org/hadoop/PoweredBy

2 "A Digital Universe Decade – Are You Ready?," EMC 2010 Digital Universe Study, IDC, April 26, 2010, http://www.securityweek.com/content/emc-digital-universe-data-growth-study-projects-nearly-45-foldannual-data-growth-202

The bank used the Hadoop cluster to construct a new and more accurate score of the risk in its customer portfolios.

1. Risk Modeling

How can banks better understand customers and markets?

The Summary

A large bank took separate data warehouses from multiple departments and combined them into a single global repository in Hadoop for analysis. The bank used the Hadoop cluster to construct a new and more accurate score of the risk in its customer portfolios. The more accurate score allowed the bank to manage its exposure better and to offer each customer better products and advice. Hadoop increased revenue and improved customer satisfaction.

The Challenge

A very large bank with several consumer lines of business needed to analyze customer activity across multiple products to predict credit risk with greater accuracy.

Over the years, the bank had acquired a number of regional banks. Each of those banks had a checking and savings business, a home mortgage business, credit card offerings and other financial products. Those applications generally ran in separate silos—each used its own database and application software. As a result, over the years the bank had built up a large number of independent systems that could not share data easily.

With the economic downturn of 2008, the bank had significant exposure in its mortgage business to defaults by its borrowers. Understanding that risk required the bank to build a comprehensive picture of its customers. A customer whose direct deposits to checking had stopped, and who was buying more on credit cards, was likely to have lost a job recently. That customer was at higher risk of default on outstanding loans as a result.

The Solution

The bank set up a single Hadoop cluster containing more than a petabyte of data collected from multiple enterprise data warehouses. With all of the information in one place, the bank added new sources of data, including customer call center recordings, chat sessions, emails to the customer service desk and others. Pattern matching techniques recognize the same customer across the different sources, even when there were some discrepancies in the identifying information stored. The bank applied techniques like text processing, sentiment analysis, graph creation, and automatic pattern matching to combine, digest and analyze the data.

The result of this analysis is a very clear picture of a customer's financial situation, this risk of default or late payment and his satisfaction with the bank and its services. The bank has demonstrated not just a reduction of cost from the existing system, but improved revenue from better risk management and customer retention.

While this application was specific to retail banking services, the techniques described—the collection and combination of structured and complex data from multiple silos, and a powerful tool of analytics that combine the data and look for patterns – apply broadly. A company with several lines of business often has only a fragmentary, incomplete picture of its customers, and can improve revenues and customer satisfaction by creating a single global view from those pieces.

By combining coverage maps with customer account data, the company could see how gaps in coverage affected churn.

2. Customer Churn Analysis

Why do companies really lose customers?

The Summary

A large telecommunications provider analyzed call logs and complex data from multiple sources. It used sophisticated predictive models across that data to predict the likelihood that any particular customer would leave. Hadoop helped the telecommunications company build more valuable customer relationships and reduce churn.

The Challenge

A large mobile carrier needed to analyze multiple data sources to understand how and why customers decided to terminate their service contracts. Were customers actually leaving, or were they merely trading one service plan for another? Were they leaving the company entirely and moving to a competitor? Were pricing, coverage gaps, or device issues a factor? What other issues were important, and how could the provider improve satisfaction and retain customers?

The Solution

The company used Hadoop to combine traditional transactional and event data with social network data. By examining call logs to see who spoke with whom, creating a graph of that social network, and analyzing it, the company was able to show that if people in the customer's social network were leaving, then the customer was more likely to depart, too.

By combining coverage maps with customer account data, the company could see how gaps in coverage affected churn. Adding information about how often customers use their handsets, how frequently they replace them and market data about the introduction of new devices by handset manufactures, allowed the company to predict whether a particular customer was likely to change plans or providers. Combining data in this way gave the provider a much better measure of the risk that a customer would leave and improved planning for new products and network investments to improve customer satisfaction.

The company's custom-built early systems could not scale up as the customer base grew. The company moved both its data storage and also its analysis system to Hadoop to take advantage of its low-cost storage and fast analytic support.

3. Recommendation Engine

How can companies predict customer preferences?

The Summary

A leading online dating service uses sophisticated analyses to measure the compatibility between individual members, so that it can suggest good matches for a potential relationship. Hadoop helped customers find romance.

The Challenge

When users sign up for the dating service, they fill in surveys that describe themselves and what they look for in a romantic partner. The company combined that information with demographic and web activity to build a comprehensive picture of its customers. The data included a mix of complex and structured information, and the scoring and matching algorithms that used it were complex.

Customers naturally wanted better recommendations over time, so the analytical system had to evolve continually with new techniques for assessing romantic fit. As the company added new subscribers, the amount of data it managed grew, and the difficulty of comparing every possible pair of romantic partners in the network grew even faster.

The Solution

The company's custom-built early systems could not scale up as the customer base grew. The company moved both its data storage and also its analysis system to Hadoop to take advantage of its low-cost storage and fast analytic support.

Hadoop allowed the company to incorporate more data over time, which has improved the "compatibility" scores that customers see. Sophisticated techniques like collaborative filtering collect more accurate information about user preferences. User behavior on the web site – what profiles a customer visits, how long she looks at a particular page – gives the company a clearer picture of the customer's preferences. Combining this information and using complex compatibility models, the company is able to deliver better recommendations, and to improve them over time.

The models that the company builds and uses to score compatibility are large and expensive to run. Every user naturally wants to be matched with as many potential partners as possible. That demands that the models be run on many pairs of customers. Hadoop's built-in parallelism and incremental scalability mean that the company can size its system to meet the needs of its customer base, and that it can grow easily as new customers join. Recommendation engines like this are useful to many other businesses, too.

Making content-to-human matches

A large content publisher and aggregator uses Hadoop to determine the most relevant content for each visitor. The publisher creates personalized recommendations continually throughout the day, as new stories arrive and as user interest in specific stories peaks and declines. Every user sees the stories most likely to be of interest, and each user sees different stories over the course of a day.

Many online retailers, and even manufacturers, rely on Hadoop to store and digest user purchase behavior and to produce recommendations for products that a visitor might buy. Social networking systems provide a "people you may know" feature to recommend friends and acquaintances to users.

In each of these instances, Hadoop combines log, transaction and other data to produce recommendations. Those recommendations increase sales and improve satisfaction by showing visitors content and people most likely to appeal to them.

The model uses large amounts of historical data on user behavior to cluster ads and users, and to deduct preferences. Hadoop delivers much better-targetted advertisements by steadily refining those models and delivering better ads.

4. Ad Targeting

How can companies increase campaign efficiency?

The Summary

Two leading advertising networks use Hadoop to choose the best ad to show to any given user.

The Challenge

Advertisement targeting is a special kind of recommendation engine. It selects ads best suited to a particular visitor. There is, though, an additional twist: each advertiser is willing to pay a certain amount to have its ad seen. Advertising networks auction ad space, and advertisers want their ads shown to the people most likely to buy their products. This creates a complex optimization challenge.

Ad targeting systems must understand user preferences and behavior, estimate how interested a given user will be in the different ads available for display, and choose the one that maximizes revenue to both the advertiser and the advertising network.

The data managed by these systems is simple and structured. The ad exchanges, however, provide services to a large number of advertisers, deliver advertisements on a wide variety of Web properties and must scale to millions of end users browsing the web and loading pages that must include advertising. The data volume is enormous.

Optimization requires examining both the relevance of a given advertisement to a particular user, and the collection of bids by different advertisers who want to reach that visitor. The analytics required to make the correct choice are complex, and running them on the large dataset requires a large-scale, parallel system.

The Solution

One advertising exchange uses Hadoop to collect the stream of user activity coming off of its servers. The system captures that data on the cluster, and runs analyses continually to determine how successful the system has been at displaying ads that appealed to users. Business analysts at the exchange are able to see reports on the performance of individual ads, and to adjust the system to improve relevance and increase revenues immediately.

A second exchange builds sophisticated models of user behavior in order to choose the right ad for a given visitor in real time. The model uses large amounts of historical data on user behavior to cluster ads and users, and to deduce preferences. Hadoop delivers much better-targeted advertisements by steadily refining those models and delivering better ads.

The retailer loaded 20 years of sales transactions history into a Hadoop cluster. It built analytic applications on the SQL system for Hadoop, called Hive, to perform the same analyses that it had done in its data warehouse system—but over much larger quantities of data, at much lower cost.

5. Point-of-Sale Transaction Analysis

How do retailers target promotions guaranteed to make you buy?

The Summary

A large retailer doing Point-of-Sale transactional analysis needed to combine larger quantities of PoS transaction analysis data with new and interesting data sources to forecast demand and improve the return that it got on its promotional campaigns. The retailer built a Hadoop cluster to understand its customers better and increased its revenues.

The Challenge

Retail analytics has been a core part of the data warehousing industry and has helped to drive its growth. Today, retailers are able to collect much more data about their customers, both in stores and online. Many want to combine this new information with recent and historical sales data from PoS systems to increase sales and improve margins.

Legacy data warehousing systems are an expensive place to store complex data from new sources. They do not, generally, support the kind of sophisticated analyses—sentiment, language processing and others—that apply to this new data.

The Solution

The retailer loaded 20 years of sales transactions history into a Hadoop cluster. It built analytic applications on the SQL system for Hadoop, called Hive, to perform the same analyses that it had done in its data warehouse system—but over much larger quantities of data, and at much lower cost.

The company is also exploring new techniques to analyze the Point-of-Sale data in new ways using new algorithms and the Hadoop MapReduce interface. Integration of novel data sources, like news and online comments from Twitter and elsewhere, is underway.

The company could never have done this new analysis with its legacy data infrastructure. It would have been too expensive to store so much historical data, and the new data is complex and needs considerable preparation to allow it to be combined with the PoS transactions. Hadoop solves both problems, and runs much more sophisticated analyses than were possible in the older system.

Hadoop was able to store the data from the sensors inexpensively, so that the power company could afford to keep long-term historical data around for forensic analysis.

6. Analyzing Network Data to Predict Failure

How can organizations use machine generated data to identify potential trouble?

The Summary

A very large public power company combined sensor data from the smart grid with a map of the network to predict which generators in the grid were likely to fail, and how that failure would affect the network as a whole.

The Challenge

Utilities run big, expensive and complicated systems to generate power. Each of the generators includes sophisticated sensors that monitor voltage, current, frequency and other important operating characteristics. Operating a single generator means paying careful attention to all of the data streaming off of the sensors attached to it.

Utilities operate many of these generators spread across multiple locations. The locations are connected to one another, and then each utility is connected to the public power grid. Monitoring the health of the entire grid requires capture and analysis of data from every utility, and even from every generator, in the grid.

The volume of data is enormous. A clear picture of the health of the grid depends on both real-time and after-the-fact forensic analysis of all of it. Spotting facilities at risk of failure early, and doing preventive maintenance or separating them from the grid, is critical to preventing costly outages.

The Solution

The power company built a Hadoop cluster to capture and store the data streaming off of all of the sensors in the network. It built a continuous analysis system that watched the performance of individual generators, looking for fluctuations that might suggest trouble. It also watched for problems among generators—differences in phase or voltage that might cause trouble on the grid as a whole.

Hadoop was able to store the data from the sensors inexpensively, so that the power company could afford to keep long-term historical data around for forensic analysis. As a result, the power company can see, and react to, long-term trends and emerging problems in the grid that are not apparent in the instantaneous performance of any particular generator.

While this was a highly specialized project, it has an analog in data centers managing IT infrastructure grids.

In a large data center with thousands of servers, understanding what the systems and applications are actually doing is difficult. Existing tools often don't scale. IT infrastructure can capture system-level logs that describe the behavior of individual servers, routers, storage systems and more. Higher-level applications generally produce logs that describe the health and activity of application servers, web servers, databases and other services. Large data centers produce an enormous amount of this data. Understanding the relationships among applications and devices is hard.

Combining all of that data into a single repository, and analyzing it together, can help IT organizations better understand their infrastructure and improve efficiencies across the network. Hadoop can store and analyze log data, and builds a higher-level picture of the health of the data center as a whole.

Computers and online systems create new opportunities for criminals to act swiftly, efficiently and anonymously. Online businesses use Hadoop to monitor and combat criminal behavior.

7. Threat Analysis

How can companies detect threats and fraudulent activity?

The Summary

Businesses have struggled with theft, fraud and abuse since long before computers existed. Computers and on-line systems create new opportunities for criminals to act swiftly, efficiently and anonymously. On-line businesses use Hadoop to monitor and combat criminal behavior.

The Challenge

Online criminals write viruses and malware to take over individual computers and steal valuable data. They buy and sell using fraudulent identities and use scams to steal money or goods. They lure victims into scams by sending email or other spam over networks. In "pay-per-click" systems like online advertising, they use networks of compromised computers to automate fraudulent activity, bilking money from advertisers or ad networks.

Online businesses must capture, store and analyze both the content and the pattern of messages that flow through the network to tell the difference between a legitimate transaction and fraudulent activity by criminals.

The Solution

One of the largest users of Hadoop, and in particular of HBase, is a global developer of software and services to protect against computer viruses. Many detection systems compute a "signature" for a virus or other malware, and use that signature to spot instances of the virus in the wild. Over the decades, the company has built up an enormous library of malware indexed by signatures. HBase provides an inexpensive and high-performance storage system for this data.

The vendor uses MapReduce to compare instances of malware to one another, and to build higher-level models of the threats that the different pieces of malware pose. The ability to examine all the data comprehensively allows the company to build much more robust tools for detecting known and emerging threats.

A large online email provider has a Hadoop cluster that provides a similar service. Instead of detecting viruses, though, the system recognizes spam messages. Email flowing through the system is examined automatically. New spam messages are properly flagged, and the system detects and reacts to new attacks as criminals create them.

Sites that sell goods and services over the internet are particularly vulnerable to fraud and theft. Many use web logs to monitor user behavior on the site. By tracking that activity, tracking IP addresses and using knowledge of the location of individual visitors, these sites are able to recognize and prevent fraudulent activity.

The same techniques work for online advertisers battling click fraud. Recognizing patterns of activity by individuals permits the ad networks to detect and reject fraudulent activity.

Hadoop is a powerful platform for dealing with fraudulent and criminal activity like this. It is flexible enough to store all of the data—message content, relationships among people and computers, patterns of activity—that matters. It is powerful enough to run sophisticated detection and prevention algorithms and to create complex models from historical data to monitor real-time activity.

The bank built a Hadoop cluster that runs alongside its existing trading systems. The system continually monitors activity and builds connections among individuals and organizations that trade with one another.

8. Trade Surveillance

How can a bank spot the rogue trader?

The Summary

A large investment bank combines data about the parties that participate in a trade with the complex data that describes relationships among those parties and how they interact with one another. The combination allows the bank to recognize unusual trading activity and to flag it for human review. Hadoop allows the bank to spot and prevent suspect trading activity.

The Challenge

The bank already captured trading activity and used that data to assess, predict, and manage risk for both regulatory and non-regulatory purposes. The very large volume of data, however, made it difficult to monitor trades for compliance, and virtually impossible to catch "rogue" traders, who engage in trades that violate policies or expose the bank to too much risk.

The risk is enormous. At Barings Bank in 1995, a single trader named Nick Leeson made an escalating series of money-losing trades in an attempt to cover losses from earlier ones. The final cost to the bank, at nearly $1.3 Billion, forced Barings out of business.

The Solution

The bank built a Hadoop cluster that runs alongside its existing trading systems. The Hadoop cluster gets copies of all of the trading data, but also holds information about parties in the trade. The system continually monitors activity and builds connections among individuals and organizations that trade with one another.

The bank has built a powerful suite of novel algorithms using statistical and other techniques to monitor human and automated, or programmed, trading. The system has developed a very good picture of normal trading activity, and can watch for unusual patterns that may reflect rogue trading. The bank uses this system now to detect a variety of illegal activity, including money laundering, insider trading, front-running, intra-day manipulation, marking to close and more. Fast detection allows the bank to protect itself from considerable losses.

The Hadoop cluster also helps the bank comply with financial industry regulations. Hadoop provides cost effective, scalable, reliable storage so that the bank can retain records and deliver reports on activities for years, as required by law.

A major online retailer meets the challenge of delivering good search results by building its indexing infrastructure on Hadoop.

9. SEARCH QUALITY

What's in your search?

The Summary

A leading online commerce company uses Hadoop to analyze and index its data and to deliver more relevant, useful search results to its customers.

The Challenge

Good search tools have been a boon to web users. As the amount of data available online has grown, organizing it has become increasingly difficult. Users today are more likely to search for information with keywords than to browse through folders looking for what they need.

Good search tools are hard to build. They must store massive amounts of information, much of it complex text or multimedia files. They must be able to process those files to extract keywords or other attributes for searches. The amount of data and its complexity demand a scalable and flexible platform for indexing.

Besides the difficulty of handling the data, a good search engine must be able to assess the intent and interests of the user when a search query arrives. The word "chips" in a query may refer to fried food or to electronic components. Delivering meaningful results requires that the system make a good guess between the two. Looking at the user's recent activity and history can help.

The Solution

A major online retailer meets the challenge of delivering good search results by building its indexing infrastructure on Hadoop. The platform has scaled easily to the data volume required. Just as importantly, it runs complicated indexing algorithms, and its distributed, parallel architecture lets the retailer index very large amounts of information quickly.

Search is particularly important to this retailer, because its revenues depend on users finding, and buying, products. If the search system delivers results that do not interest the user, there is no sale. Revenue depends heavily on delivering search results that appeal to the user.

The retailer uses information about individuals, their history and their preferences when it builds its search index. The system also tracks user behavior in response to searches — which results were clicked, and which were ignored? That data is used to refine the results shown to others who run similar queries.

When a query arrives from a particular user, the results delivered depend not just on the content, but also on who the user is and what he or she has purchased in the past. This targeted search capability has made a direct and measurable contribution to revenues—the retailer can tell how much more product it is selling with targeted search. The investment in Hadoop translates directly into higher revenue.

Hadoop can store all types of data and makes it easy for analysts to pose questions, develop hypotheses and explore the data for meaningful relationships and value.

10. DATA SANDBOX

What can you do with new data?

The Summary

Companies—even those with established enterprise data warehouses—often need a cost-effective, flexible way to store, explore and analyze new types of complex data. Many companies have created "data sandboxes" using Hadoop, where users can play with data, decide what to do with it and determine whether it should be added to the data warehouse. Analysts can look for new relationships in the data, mine it and use techniques like machine learning and natural language processing on it to get new insights.

The Challenge

The variety, complexity and volume of data available to enterprises today are changing the way that those enterprises think. There is value and insight locked up in complex data. The best companies in the world are unlocking that value by building new analytical capacity.

With shifting and complex data and emerging analytics, enterprises need a new platform to store and explore the information they collect. Cost, scalability and flexibility are all forcing a move away from the single-architecture data warehouse of the past, toward a more flexible and comprehensive data management infrastructure.

The Solution

Hadoop makes an exceptional staging area for an enterprise data warehouse. It provides a place for users to capture and store new data sets or data sets that have not yet been placed in the enterprise data warehouse. Analysts can combine and compose data across several lines of business and explore it.

Small and large companies across a range of industries take advantage of Hadoop for exactly this purpose. They create a sort of "data sandbox" that encourages the exploration of new and old data.

A large online auction and shopping website, for example, uses Hadoop as a complement to its data warehouse. Reduced cost and the ability to combine different sources allow the company to identify new and promising analyses that may eventually be done in the data warehouse. Hadoop can store all types of data and makes it easy for analysts to pose questions, develop hypotheses and explore the data for meaningful relationships and value.

Why Cloudera?

Cloudera is the leading provider of Hadoop-based software and services. Our open source software offering, Cloudera's Distribution for Apache Hadoop (CDH), is the industry's most popular means of deploying Hadoop. 100% Apache licensed and free for download, CDH is a platform for data management and combines the leading Hadoop software and related projects and provides them as an integrative whole with common packaging, patching, and documentation. CDH gives Hadoop users unprecedented stability, predictability, and functionality. Cloudera's proprietary offering, Cloudera Enterprise, is a cost-effective way to perform large-scale data storage and analysis and includes the management applications, platform and support necessary to use Hadoop in a production environment.

Our founders have played key roles in the development of the Hadoop framework while serving in leadership roles at Facebook and Yahoo—companies that manage some of the world's largest Hadoop implementations. The company has engineers dedicated to Hadoop and related open source projects, and contributes Hadoop enhancements and fixes back to the open source community.

Cloudera's professional services team is experienced at delivering high value services to thousands of users supporting hundreds of implementations over a range of industries. Our customers use Cloudera's products and services to store, manage, and analyze data on large Hadoop implementations.

See how Hadoop can help your business unlock valuable insights by downloading CDH for free at www.cloudera.com/downloads.

Summary

Hadoop gives companies the power to store and analyze information more cheaply than ever before. Its power and flexibility make it the perfect complement to existing data warehousing infrastructure. Across a variety of industries, Hadoop is solving hard business problems: reducing risk, keeping customers happier and driving revenue.

Cloudera makes Hadoop easy to use by offering products and services that complement Apache Hadoop. Cloudera offers specialized applications, comprehensive training, architectural and implementation services and technical support.

To learn more about Cloudera products and services, call 1-888-789-1488 or visit www.cloudera.com.