cloudera®

WHITE PAPER

Five Common Hadoopable Problems

Real-World Apache Hadoop™ Use Cases

Version: 102

cloudera®

Table of Contents

cloudera®

Introduction

Apache Hadoop has evolved into the standard platform solution for data storage and analysis. Large, successful companies are increasingly adopting Hadoop to perform powerful analyses of their ever-growing business data.

Two key aspects of Hadoop have driven its rapid adoption by companies hungry for improved insights into the data they collect:

- Hadoop can store data of any type and from any source—inexpensively and at very large scale.
- Hadoop enables the sophisticated analysis of even very large data sets, easily and quickly.

However, Hadoop concepts are unfamiliar to many people with a background in traditional database and data warehousing systems, and its business value is often underappreciated.

It is the goal of this paper to provide concrete examples of organizations that have used the power of Hadoop to advance the state of their data analytics and drive business impact. Each use case presented here brings specific details from real-world users.

One or more of the examples detailed in this paper may apply directly to your business. However, this is only a small sample of how Hadoop can have an impact for you. For additional examples and more details, please visit the Cloudera customer references page at: http://www.cloudera.com/customers.html

"We're processing five times the data in a third of the time. The business sponsors don't know that we moved to Hadoop and they don't care. All they know is that they're now working with today's data instead of yesterday's."

Phillip Radley, Chief Data Architect, BT

Why use Hadoop?

Hadoop solves the difficult scaling problems caused by processing and analyzing large amounts of complex data. As the amount of data in a cluster grows, new servers can be added incrementally for it to be captured and analyzed.

Today, companies such as Yahoo! and Facebook use Hadoop to store and analyze petabytes of data. Hadoop performs better at scale than traditional systems, and in many instances it can perform analyses that traditional systems simply cannot.

"Our Hadoop infrastructure has become a real transformational change. Deploying Cloudera allows us to process orders of magnitude more information through our systems"

Jeff Hassemer, VP of Product Strategy, Experian

Hadoop delivers significant benefits. With it you can:

Store anything. Translating data as it is ingested so it fits into a fixed data warehouse schema is time consuming and sometimes problematic. Hadoop can store data in its native format, exactly as it arrives at the cluster. Hadoop can store data it it's original fidelity to ensure all insights receive the best level of granularity.

Control costs. Hadoop is open source software that runs on industry-standard hardware. The combination provides lower cost per terabyte than traditional data warehouse environments. As storage and analytic requirements evolve, a Hadoop installation can, too. The platform scales and handles resources independently allowing for you to tailor the resources to your workload (Compute, Storage, Memory). Scaling is incremental and granular. Nodes can easily be retired or repurposed especially in cloud environments.

Operate with confidence. Hadoop has been widely adopted across a wide range of industries to drive better analytics and business outcomes. The Hadoop community, including open source contributors and Hadoop users, is global, active, and diverse. Cloudera works to make the great innovations from the community meet the requirements of the enterprise.

Deploy and scale seamlessly. Hadoop is a proven foundation for the analysis of large-scale data sets. But whatever the size of your data, you can deploy Hadoop with confidence. When you adopt a platform for data management and analysis, you are making a commitment you'll live with for years. The global success of Hadoop in companies of all types and sizes demonstrates that Hadoop can meet your needs today – and tomorrow.

cloudera

What is Hadoop?

Hadoop is an open 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 huge amounts of data they generate.

Hadoop runs on groups of servers that work together as part of a Hadoop cluster. Each server can perform tasks and store data. While each server may lack the processor power and storage capacity necessary to process huge data sets, Hadoop allows the servers to work together to store and process truly massive amounts of data.

Hadoop consists of a number of components that work together to help your organization:

STORE. Hadoop's scalable, flexible storage architecture (based on the Hadoop Distributed File System, HDFS) allows organizations to store and analyze unlimited amounts and types of data—all in a scalable open source platform on industry-standard hardware. With Hadoop you can store data in virtually any format including a distributed data store (Apache HBase), in a columnar database (Apache Kudu), and even use fully unstructured data from object stores.

PROCESS. Hadoop helps you quickly move data into and out of relational systems through bulk load processing (Apache Sqoop) or streaming (Apache Flume, Apache Kafka). You can transform complex data, at scale, using multiple data access options (Apache Hive, Apache Pig) for batch (MR2) or fast in-memory (Apache Spark) processing. Process streaming data as it arrives in your Hadoop cluster via Spark Streaming.

DISCOVER. Use Hadoop to interact with full-fidelity data. Analysis can be performed on the fly with Apache Impala (Incubating), an analytic database that provides high-performance SQL functionality and compatibility with leading Business Intelligence tools. Apache Solr provides full-text search supporting batch, real-time and on-demand indexing (and re-indexing) of data of any type.

MODEL. With Hadoop, analysts and data scientists have the flexibility to develop and iterate on advanced models using a mix of partner technologies and open source frameworks such as Apache Spark. Pre-built tools and libraries for machine learning, statistical modeling, and time- series analysis (among others) can speed the development of impactful models and provide rapid results.

SERVE. Hadoop's distributed and scalable architecture allows you to serve data as rapidly and efficiently as you ingest it, while also developing online applications that deliver data to more users and applications. An efficient random read/write architecture supplies the "fast data" needed for online services, real-time monitoring, and "Internet of Things" applications.

> "As we've moved to Cloudera's platform, powered by Apache Hadoop, we have been able to eliminate the challenges we faced with our legacy environments. We can acquire the data quickly and bring it directly to the analysts across the company very quickly. We can also avoid the pain of building extract, transform, load (ETL) processes, and structuring all the data in potentially multiple environments. Now, everyone across multiple business units can have access to the same datasets, and analyze them with the tools and skills they already use."
>
> Scott Salter, Vice President,
> Enterprise Data Services, Cox Automotive

Recognizing Hadoopable Problems

Complex data demands a new approach. The nature of the data that enterprises must capture, store, and analyze is changing— as is the importance of the insights it can provide.

Hadoopable problems are 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 format requirements change over time as new sources are developed. Hadoop is able to store data in its native format while offering analytical access.

There's a lot of data in a Hadoopable problem

Many companies are forced to discard valuable data because the cost of storing it is too high. New data sources compound the problem: people and machines are generating more data than ever before. Hadoop's innovative architecture and use of low-cost industry-standard hardware for storage and processing helps you manage large amounts of data efficiently.

Hadoopable problems demand new approaches

Simple numerical summaries—average, minimum, sum—were sufficient for business problems in the 1980s and 1990s. But the large, complex data problems of today's businesses require new techniques. The algorithms involved include natural language processes, pattern recognition, machine learning and other advanced methods.

cloudera

Hadoopable Problem #1: Detecting Cybersecurity Threats

The Hadoopable Problem:

Cybersecurity is the most important priority for IT organizations in 2016[1]. Yet the detection and prevention of cybersecurity threats remains challenging, in part due to the complexity and ever-changing nature of these threats and the diversity of companies they threaten.

Today the data used to detect and predict threats comes from a huge range of sources including log files, email traffic, network traffic monitors, usage reports, and even physical sensors such as alarm or badge systems.

Businesses, particularly online businesses, must capture, store, and analyze both the content and the pattern of data as it flows through their systems to differentiate between legitimate activity and a true threat to their business. Because threats can have immediate impact, rapid and efficient analysis is vital.

> "Our business is very much about data, and being able to use that information – both learned and applied – and turn it back into better protection for our customers makes all the difference."
>
> Robert Scudiere, Director of Engineering, Dell SecureWorks

The Solution:

One of the largest users of Hadoop, and in particular Apache HBase (a distributed non-relational database), is a global developer of software and services to protect against computer viruses. The company analyzes viruses and computes a unique "signature" for each that allows it to quickly recognize instances of individual malware. HBase provides an inexpensive and high-performance storage system for their enormous library of signatures.

MapReduce is used to compare instances of malware and to build higher-level models of the threats they pose. The ability to examine all the data comprehensively allows the company to build robust tools for detecting known and emerging threats.

A large online email provider has a Hadoop cluster that it uses to recognize and reject spam messages. Email flowing through the system is examined automatically and compared to known patterns and behaviors. 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 better able to recognize and prevent fraudulent activity.

The same techniques work for online advertisers battling "click fraud" in which advertising revenues are artificially enhanced through automated activity. Recognizing authentic patterns of activity by known individuals permits the ad networks and advertisers to detect and reject this fraudulent activity. Machine learning is particularly helpful for this use case, as click fraud evolves rapidly and fraud prevention must evolve in parallel.

The Bottom Line:

Hadoop is a powerful platform for dealing with cybersecurity threats, fraud, and criminal activity. It is flexible enough to store all the data that matters: content, logs, relationships between people or systems, and patterns of activity. It is powerful enough to run sophisticated detection, analysis, and prevention algorithms and to create complex models from historical data to monitor real-time activity. Hadoop is also flexible enough to change as threats evolve, and to store and analyze new data sources as they become available. With Hadoop, companies can better analyze and prevent cybersecurity threats efficiently and effectively.

[1] ESG Research Report: 2016 IT Spending Intentions Survey

cloudera

Hadoopable Problem #2: Reducing Customer Churn

The Hadoopable Problem:

For wireless telecommunications providers, digital media companies, and other service organizations, customer acquisition costs are high and the impact of losing customers, dramatic. Customer loss has a direct impact on profitability and performance.

Companies like these have taken extraordinary measures to understand their customers— both how to improve overall satisfaction and how to stem attrition. The discipline of Retention Processing (operations undertaken to reduce churn) has been on the leading edge of relational database technologies for years.

But both the processing necessary to provide impactful, real-time analysis, and the amount of data required, has caused researchers to scale back analysis, rather than expand it.

> "We need to reinvent the playbook for IT to make sure we can handle the changing landscapes of the digital economy and the changing preferences of our customers"
>
> John Swieringa, Executive VP of Operations, Dish Network

The Solution:

To analyze multiple data sources and determine why users might terminate contracts, a large mobile carrier turned to Hadoop. With Hadoop they could combine traditional transactional and event data with other data sources, including information from social networks.

The company combined traditional transaction data such as call logs to develop models for inter-customer communications (who called whom) and created a graph of their users social networks. An analysis of customer loss combined with social networking details, indicated that groups of customers left together—the loss of one member of a social network was highly predictive of the loss of other members.

By combining market data on new equipment release dates and adoption (in this case, mobile phones) with known patterns of customer churn, correlation was discovered between the arrival of new phones and customer departure. New phones and discounts was were directly related to a customer's choice of a new service plan or provider.

Analysis of coverage maps and customer churn also provided important insights. Customers were more likely to choose other providers if they were present in areas of low or no coverage.

Ultimately, by combining internal and external data sources and analyzing them in new ways, the provider was able to identify and resolve likely causes of customer churn and reduce departures accordingly.

The Bottom Line:

Data analysis is key to determining customer preference and building customer lifetime value. However, to successfully analyze complex problems like customer churn, data from many sources is required. By combining these data sources using Hadoop, it's possible to create models that tie together market forces, customer preferences, and company operations into a holistic view of customer retention that can positively impact profitability and company performance.

cloudera

Hadoopable Problem #3: Targeting Advertising

The Hadoopable Problem:

Online advertisement targeting is big business—more than $60B will be spent in 2016 on online ads[2]. At its core, ad targeting is a specialized type of recommendation engine that identifies users, determines their preferences, and delivers the ads best suited to each user. In practice, ad targeting is incredibly complicated—involving paid placement by advertisers who are constantly working to increase ad views and drive viewer engagement. Ad networks auction ad space to advertisers who want their ad shown to the people most likely to buy their products. Ad placement can become a very complex optimization problem with conflicting priorities and complex models.

> "Marketers rely on real-time information and insights to deliver timely, relevant interactions. Cloudera Enterprise provides us with the tools and support we need to power real-time analytics to engage with connected customers across the channels they use most."
>
> Tim Horoho, Vice President of Infrastructure, ExactTarget

The Solution:

Ad targeting systems must identify users, understand their preferences and behavior, estimate how interested a given user will be in the various ads available for display, and then choose the ad that maximizes revenue to both the ad's owner and the advertising network.

The data used to determine user preference is frequently structured and simple, but increasingly comes from a variety of sources—both with the ad network and from external data providers. Even the simplest series of interactions with online properties, such as clicking through an online retailer's website, generates considerable data. Reproduced thousands or millions of time, the amount of data collected is staggering. Reconciling and transforming this data can be difficult, particularly with the enormous volume of data generated.

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 available requires a large-scale, massively parallel system.

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 generate reports on the performance of individual ads and to adjust the system to improve relevance and drive immediate increases in revenue.

Another ad exchange has focused their efforts on building sophisticated models of user behavior in order to choose the right ad for each viewer in real time. The model uses large amounts of historical and tracking data about each user to cluster ads and users, and to deduce preferences. By leveraging the power of Hadoop to analyze historical user data, the exchange delivers much better-targeted advertisements and can steadily refine its models and deliver increasingly better ads.

The Bottom Line:

Hadoop is a powerful platform for combining data from various sources into a unified view. It is flexible enough to store all the data that matters—content, user activity logs, relationships between people or systems, and patterns of activity. Hadoop is also flexible enough to store and quickly analyze new data sources as they become available. Using the platform, organizations can run complex analysis to understand user behavior, create targeted ads using advanced analytics, and apply machine learning to optimize ad yields. With Hadoop, companies can better analyze ad effectiveness and maximize revenue to both the ad owner and the advertising network.

[2] PWC Report: Global Entertainment and Media Outlook 2015-2019

cloudera®

Hadoopable Problem #4: Delivering Search Results

The Hadoopable Problem:

Good search tools have been a boon to web users and the companies they visit online. But as the amount of data online has grown, organizing 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 these files to extract keywords or other attributes to use in 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 interest of the user when a search query arrives. The word "chip," for example, can refer to food or to electronic components—context is vital to delivering useful search results. Delivering meaningful results is dependent on the analysis of user preferences, recent browsing history, and a number of other data sources.

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 Hadoop's distributed, parallel architecture lets the retailer index very large amounts of information quickly. Search is particularly important to retailers because revenues are highly dependent on search success. If the search system delivers results that are not of interest to the buyer, there is no sale and often the user might choose another retailer. Search relevance therefore drives both profitability and customer retention.

In addition to information about individuals, their history, and their preferences used when building search indexes, effective search engines track user behavior in response to searches themselves (which results were clicked, and which were ignored) and uses it to refine search results for subsequent users and future search results.

The Bottom Line:

Online search is a problem of massive data sets, rapid analysis, and meaningful results. By combining the storage capacity of Hadoop with the ability to analyze data in parallel, regardless of format, large amounts of data can be economically processed to provide results that take into account user behavior, preference and history, and match query results to user needs.

For businesses like online retailers, accurate and meaningful search results drive profitability and customer retention—helping build engaged and satisfied customers.

cloudera®

Hadoopable Problem #5: Predicting Utility Outages

The Hadoopable Problem:

Energy utilities run large, complicated, expensive systems to generate and distribute power. Each generator, regardless of type, contains sophisticated sensors that monitor output: voltage, current, frequency, and other important operating parameters. Operating a single generator requires paying close attention to the energy source, energy production, and all the data constantly streaming from the sensors to which it is attached.

Utilities operate many generators that are spread across multiple locations. The locations are connected to one another, and each utility is connected to a public power grid.

Monitoring the health of the entire grid requires capture and analysis of data from multiple energy providers, every generator, and the grid itself.

Failures in energy generation often result in cascading outages as the larger generation and distribution network struggles to contain a surge, or supply a shortfall. One small problem can manifest as a utility outage for tens of thousands of consumers and can be extraordinarily costly for the utility operators.

The volume of data required to predict and prevent outages is enormous. A clear picture of the health of the grid depends on both real-time and after-the-fact forensic analysis of huge amounts of collected data in a variety of formats. Spotting facilities or grid components at risk of failure before they fail, and doing preventative maintenance or separating them from the grid before they impact service delivery, is critical to preventing costly outages.

> "The obvious focus for utility companies is on their physical infrastructure. How do I manage my flow of electricity? How do I make sure the lights are on? Managing the data that gets collected from that infrastructure is secondary to them."
>
> Eric Chang, Technical Lead for Data Services, Opower

The Solution:

One power company uses Hadoop clusters to capture and store the data streaming off of sensors in both energy production and the grid. It built a continuous analysis system that watched 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 in usable form 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.

The Bottom Line:

While the use of Hadoop in power generation and outage prevention is highly specialized, it has an analog in other complex systems and infrastructure grids. In particular, the same tools and approaches are often used in large-scale data centers or other applications that deliver utility-type services.

In a large data center with thousands of servers, understanding what the systems and applications are actually doing can be difficult. Existing tools don't often scale and operational data doesn't conform to simple formats. IT infrastructure can capture system-level logs that describe the behavior of individual servers, routers, storage systems, and more. Higher-level applications also generally produce logs that describe the health and activity of application servers. Combining all the 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 in virtually any format and can build a higher-level picture of the health of the data center as a whole.

cloudera®

Summary and Suggestions

Hadoop gives companies the power to store and analyze information quickly, efficiently, and at lower cost than ever before.

Its power and flexibility make it the perfect solution to problems that involve large, complex data sets, and that demand new approaches to processing and analysis.

Hadoop is also a powerful complement to existing data warehousing infrastructure.

Across a variety of industries and use cases, 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 enhance Apache Hadoop. Cloudera offers specialized applications, comprehensive training, architectural and implementation services, and technical support.

- Learn how companies like yours have unlocked the power of their data using Hadoop at www.cloudera.com/customers
- See how Hadoop can help your business unlock valuable insights by downloading the Cloudera Distribution for Apache Hadoop (CDH) for free at www.cloudera.com/downloads

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 integrated whole with common packaging, patching, and documentation. CDH gives Hadoop users unprecedented stability, predictability, and functionality.

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.

Cloudera's founders have played a leading role in the development of the Hadoop framework. The company continues to have engineers dedicated to the enhancement of Hadoop and related open source projects, and consistently contributes enhancements and/or fixes back to the open source community.

Cloudera's professional services team is experienced at delivering high-value services to thousands of users supporting implementations over a range of industries and use cases.

About Cloudera

Cloudera delivers the modern platform for data management and analytics. The world's leading organizations trust Cloudera to help solve their most challenging business problems with Cloudera Enterprise—the fastest, easiest, and most secure data platform built on Apache Hadoop. Our customers can efficiently capture, store, process, and analyze vast amounts of data—empowering them to use advanced analytics to drive business decisions quickly, flexibly, and at lower cost than has been possible before. To ensure our customers are successful, we offer comprehensive support, training, and professional services. Learn more at cloudera.com.

cloudera.com

1-888-789-1488 or 1-650-362-0488
Cloudera, Inc. 1001 Page Mill Road, Palo Alto, CA 94304, USA