

Search the book:

[Hadoop Illuminated](#) > **Hadoop Use Cases and Case Studies**



Chapter 10. Hadoop Use Cases and Case Studies

Table of Contents

- [10.1. Politics](#)
 - [2012 US Presidential Election](#)
- [10.2. Data Storage](#)
 - [NetApp](#)
- [10.3. Financial Services](#)
 - [Dodd-Frank Compliance at a bank](#)
- [10.4. Health Care](#)
 - [Storing and processing Medical Records](#)
 - [Monitoring patient vitals at Los Angeles Children's Hospital](#)
- [10.5. Human Sciences](#)
 - [NextBio](#)
- [10.6. Telecoms](#)
 - [China Mobil Guangdong](#)
 - [Nokia](#)
- [10.7. Travel](#)
 - [Orbitz](#)
- [10.8. Energy](#)
 - [Seismic Data at Chevron](#)
 - [OPower](#)
- [10.9. Logistics](#)
 - [Trucking data @ US Xpress](#)
- [10.10. Retail](#)
 - [Etsy](#)
 - [Sears](#)
- [10.11. Software / Software As Service \(SAS\) / Platforms / Cloud](#)
 - [SalesForce](#)
 - [Ancestry](#)
- [10.12. Imaging / Videos](#)
 - [SkyBox](#)
 - [Comcast](#)
- [10.13. Online Publishing , Personalized Content](#)
 - [Gravity](#)

This is a collection of some use cases of Hadoop. This is not meant to be an exhaustive list, but a sample to give you some ideas.

A pretty extensive list is available at the [Powered By Hadoop site](#)

10.1. Politics

2012 US Presidential Election

[How Big Data help Obama win re-election](#) - by Michael Lynch, the founder of [Autonomy](#) ([cached copy](#))

10.2. Data Storage

NetApp

NetApp collects diagnostic data from its storage systems deployed at customer sites. This data is used to analyze the health of NetApp systems.

Problem: NetApp collects over 600,000 data transactions weekly, consisting of unstructured logs and system diagnostic information. Traditional data storage systems proved inadequate to capture and process this data.

Solution: A Cloudera Hadoop system captures the data and allows parallel processing of data.

Hadoop Vendor: Cloudera

Cluster/Data size: 30+ nodes; 7TB of data / month

Links:

[Cloudera case study](#) ([cached copy](#)) (Published Sep 2012)

10.3. Financial Services

Dodd-Frank Compliance at a bank

A leading retail bank is using Cloudera and Datameer to validate data accuracy and quality to comply with regulations like Dodd-Frank

Problem: The previous solution using Teradata and IBM Netezza was time consuming and complex, and the data mart approach didn't provide the data completeness required for determining overall data quality.

Solution: A Cloudera + Datameer platform allows analyzing trillions of records which currently result in approximately one terabyte per month of reports. The results are reported through a data quality dashboard.

Hadoop Vendor: Cloudera + Datameer

Cluster/Data size: 20+ nodes; 1TB of data / month

Links:

[Cloudera case study](#) ([cached copy](#)) (Published Nov 2012)

10.4. Health Care

Storing and processing Medical Records

Problem: A health IT company instituted a policy of saving seven years of historical claims and remit data, but its in-house database systems had trouble meeting the data retention requirement while processing millions of claims every day

Solution:

A Hadoop system allows archiving seven years' claims and remit data, which requires complex processing to get into a normalized format, logging terabytes of data generated from transactional systems daily, and storing them in CDH for analytical purposes

Hadoop vendor:

Cloudera

Cluster/Data size: 10+ nodes pilot; 1TB of data / day

Links:

[Cloudera case study](#) ([cached copy](#)) (Published Oct 2012)

Monitoring patient vitals at Los Angeles Children's Hospital

Researchers at LA Children's Hospital is using Hadoop to capture and analyze medical sensor data.

Problem: Collecting lots (billions) of data points from sensors / machines attached to the patients. This data was periodically purged before because storing this large volume of data on expensive storage was cost-prohibitive.

Solution: Continuously streaming data from sensors/machines is collected and stored in HDFS. HDFS provides scalable data storage at reasonable cost.

Hadoop Vendor: Unknown

Cluster/Data size: ???

Links:

[video](#)

[silicon angle story](#) (Published June 2013)

10.5. Human Sciences

NextBio

NextBio is using Hadoop MapReduce and HBase to process massive amounts of human genome data.

Problem:

Processing multi-terabyte data sets wasn't feasible using traditional databases like MySQL.

Solution:

NextBio uses Hadoop map reduce to process genome data in batches and it uses HBase as a scalable data store

Hadoop vendor:

Intel

Links:

[NextBio](#)

[Intel case study](#) ([cached copy](#)) (Published Feb 2013)

[Information Week article](#) ([May 2012](#)) ([cached copy](#))

10.6. Telecoms

China Mobil Guangdong

Problem: Storing billions of mobile call records and providing real time access to the call records and billing information to customers.

Traditional storage/database systems couldn't scale to the loads and provide a cost effective solution

Solution: HBase is used to store billions of rows of call record details. 30TB of data is added monthly

Hadoop vendor: Intel

Hadoop cluster size: 100+ nodes

Links:

[China Mobil Guangdong](#)

[Intel case study](#) ([cached copy](#)) (Published Feb 2013)

[Intel APAC presentation](#)

Nokia

Nokia collects and analyzes vast amounts of data from mobile phones

Problem:

- (1) Dealing with 100TB of structured data and 500TB+ of semi-structured data
- (2) 10s of PB across Nokia, 1TB / day

Solution: HDFS data warehouse allows storing all the semi/multi structured data and offers processing data at peta byte scale

Hadoop Vendor: Cloudera

Cluster/Data size:

- (1) 500TB of data
- (2) 10s of PB across Nokia, 1TB / day

Links:

(1) [Cloudera case study](#) ([cached copy](#)) (Published Apr 2012)

(2) [strata NY 2012 presentation slides](#) ([cached copy](#))

[Strata NY 2012 presentation](#)

10.7. Travel

Orbitz

Problem: Orbitz generates tremendous amounts of log data. The raw logs are only stored for a few days because of costly data warehousing. Orbitz needed an effective way to store and process this data, plus they needed to improve their hotel rankings.

Solution: A Hadoop cluster provided a very cost effective way to store vast amounts of raw logs. Data is cleaned and analyzed and machine learning algorithms are run.

Hadoop Vendor: ?

Cluster/Data size: ?

Links:

[Orbitz presentation](#) (Published 2010)

[Datanami article](#)

10.8. Energy

Seismic Data at Chevron

Problem: Chevron analyzes vast amounts of seismic data to find potential oil reserves.

Solution: Hadoop offers the storage capacity and processing power to analyze this data.

Hadoop Vendor: IBM Big Insights

Cluster/Data size: ?

Links:

[Presentation](#) ([cached copy](#)) (Published June 2012)

OPower

OPower works with utility companies to provide engaging, relevant, and personalized content about home energy use to millions of households.

Problem: Collecting and analyzing massive amounts of data and deriving insights into customers' energy usage.

Solution: Hadoop provides a single storage for all the massive data and machine learning algorithms are run on the data.

Hadoop Vendor: ?

Cluster/Data size: ?

Links:

[presentation](#) ([cached copy](#)) (Published Oct 2012)

[Strata NY 2012](#)

[Strata 2013](#)
[OPower.com](#)

10.9. Logistics

Trucking data @ US Xpress

US Xpress - one of the largest trucking companies in US - is using Hadoop to store sensor data from their trucks. The intelligence they mine out of this, saves them \$6 million / year in fuel cost alone.

Problem: Collecting and storing 100s of data points from thousands of trucks, plus lots of geo data.

Solution: Hadoop allows storing enormous amount of sensor data. Also Hadoop allows querying / joining this data with other data sets.

Hadoop Vendor: ?

Cluster/Data size: ?

Links:

[Computer Weekly article](#) (Published May 2012)

[Hortonworks white paper on 'Business Value of Hadoop'](#) ([cached copy](#)) (Published July 2013)

[USXpress.com](#)

10.10. Retail

Etsy

[Etsy](#) is an online market place for handmade stuff.

Problem: Analyzing large volume of log data, without taxing the databases

Solution: Etsy uses Hadoop to analyze large volumes of log data to calculate user behaviour, search recommendations...etc

Hadoop Vendor: Amazon Elastic Map Reduce (EMR)

Cluster/Data size: varies

Links:

[Hadoop at Etsy](#) (March 2013)

[gigaom article](#) (Nov 2011)

[pdf](#) ([cached copy](#)) (Nov 2011)

Sears

[Sears](#) is a department store (online and brick and mortar).

Problem: Sears' process for analyzing marketing campaigns for loyalty club members

used to take six weeks on mainframe, Teradata, and SAS servers. The old models made use of 10% of available data

Solution: The new process running on Hadoop can be completed weekly. For certain online and mobile commerce scenarios, Sears can now perform daily analyses. Targeting is more granular, in some cases down to the individual customer. New process can use 100% of available data.

Hadoop Vendor: ?

Cluster/Data size: ?

Links:

<http://www.informationweek.com/global-cio/interviews/why-sears-is-going-all-in-on-hadoop/240009717> (Oct 2012)

<http://www.metascale.com/resources/blogs/187-big-data-case-study-hadoop-first-usage-in-production-at-sears-holdings> (Aug 2013)

10.11. Software / Software As Service (SAS) / Platforms / Cloud

SalesForce

Problem: Analyzing data that is generated at a rate of multiple terabytes / day

Solution: SalesForce uses Hadoop to compute Product Metrics, Customer Behavior, Monitoring ..etc

Hadoop Vendor: Apache Hadoop

Cluster/Data size: ?

Links:

» [SalesForce](#)

» [presentation](#) (June 2012)

Ancestry

Problem:

Ancestry users have created more than 47 million family trees containing more than 5 billion profiles of relatives. Added to the current mass archive, the new flood of gene-sequencing data generated by Ancestry's recently-introduced DNA testing product will present Big Data challenges.

Ancestry manages 11 billion records (4 petabytes) of searchable structured and unstructured data consisting of birth, death, census, military, immigration and other records.

Solution:

Using HBase to manage large searchable datastore. Using Hadoop to scale geneology algorithms.

Hadoop Vendor: ?

Cluster/Data size: ?

Links:

- » [talk at Strata 2013](#) (Oct 2013)
- » [Ancestry.com](#)

10.12. Imaging / Videos

SkyBox

SkyBox is developing a low cost imaging satellite system and web-accessible big data processing platform that will capture video or images of any location on Earth

Problem:

Analyzing really large volumes image data downloaded from the satellites

Solution:

Skybox uses Hadoop to process images in parallel. Their image processing algorithms are in C/C++. Their proprietary framework 'BusBoy' allows using native code from Hadoop MapReduce Java framework.

Hadoop Vendor: Cloudera and Amazon EC2

Cluster/Data size: ?

Links:

- » [Case Study @ Cloudera](#) (Oct 2012),
- » [TechTarget article](#) (Aug 2013)
- » [SkyBox](#)

Comcast

Comcast provides video and bandwidth to lots of US customers.

Problem:

Analyzing large volumes of data generated by video players and monitoring performance issues in real time

Solution:

Comcast uses a Hadoop infrastructure to capture and analyze large volumes of 'dial-home' data generated by multitude of video players. They do both analysis in (near) real time and in batch mode

Hadoop Vendor: Cloudera

Cluster/Data size: ?

Links:

- » [Talk at Strata 2013](#) (Oct 2013),
- » [Comcast](#)

10.13. Online Publishing , Personalized Content

Gravity

Gravity's mission is to personalize the internet by generating interest graphs that help websites deliver customized content to every site visitor.

Problem:

Building user profiles from large volumes of data

Solution:




Gravity uses Hadoop to analyze data and build profile and targets content for users. With improved targeting the click rates have gone up 300-400% and users are staying on the site longer.

Hadoop Vendor: Cloudera

Cluster/Data size: ?

Links:

- » [Cloudera case study](#)
- » [Gravity](#)

| | | |
|---|---|---|
|  | |  |
| Chapter 9. Introduction To MapReduce |  | Chapter 11. Hadoop Distributions |

This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License](#).

