Envisioning the Win in a Hadoop Environment

Charles Tay

Seattle Pacific University

August 25, 2019

Abstract

Paving the way to a successful big data implementation requires a combination of good leadership, sustainable investment, prepared team, and most importantly, vision. This paper studies the use cases of revolutionary Big Data technologies, primarily focusing on Hadoop, in large multi-national corporations. Building on the results of this study, the paper examines the reasons for implementation as well as non-implementation of Hadoop and other Big Data technologies in organizations. Drawing lessons learnt from the successes and failures of real-world implementations, the paper concludes with recommendations for its audience.

Keywords: big data, hadoop, use cases, limitations, recommendations

Envisioning the Win

Given the pace of innovation and adoption of big data technologies worldwide, organizations across a multitude of industries face an unprecedented disruption. With added pressures to handle large volume and variety of information in the shortest amount of time (Rehman et al., 2016), companies now look to Hadoop and similar frameworks as a viable solution. For these companies, it is paramount to envision how a successful implementation might look like and avoid potential pitfalls in the management of Big Data technologies.

What is Hadoop?

Since this paper primarily focuses on the use cases of Hadoop, it is important to establish a common understanding of the technology. Hadoop consists of three core components: a distributed file system known as HDFS, a parallel programming framework known as MapReduce, and a resource management system known as YARN (Rodda & VijayaKumari, 2018). Linux and Windows are the supported Operating Systems for Hadoop, although BSD, Mac OS/X, and OpenSolaris are known to work as well.

Hadoop Distributed File System (HDFS)

Hadoop consists of an open-source, Java-based implementation of a clustered file system called Hadoop Distributed File System (HDFS), which allows users to do cost-efficient, reliable, and scalable distributed computing. The HDFS architecture is highly fault-tolerant and designed to be deployed on low-cost hardware. HDFS can store a large number of files. As such, internet companies, such as Alibaba, Sohu, and Amazon, use HDFS to store large amounts of data, and use big data tools to obtain useful information (Liu, 2019).

Hadoop MapReduce

Hadoop is focused on the storage and distributed processing of large data sets across clusters of computers using a MapReduce programming model: Hadoop MapReduce. With MapReduce, the input file set is broken up into smaller pieces, which are processed independently of each other (the "map" part). The results of these independent processes are then collected and processed as groups (the "reduce" part) until the task is done. A Hadoop MapReduce cluster can host a variety of big data applications running concurrently (Malik et al., 2018).

Hadoop YARN

The third core component, Hadoop YARN framework, allows users to do job scheduling and cluster resource management, meaning users can submit and kill applications through the Hadoop REST API. There are also web User Interfaces (UIs) for monitoring your Hadoop cluster. In Hadoop, the combination of files and classes needed to run a MapReduce program is called a job. Users can submit jobs to a JobTracker from the command line. These jobs contain the "tasks" that execute the individual map and reduce steps (Jin, Hao, Wang, & Yue 2019).

Common Tools Used in Hadoop

The Apache foundation host a list of Hadoop-related projects:

- **Ambari:** A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters. Ambari includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig, and Sqoop.
- **Cassandra:** Cassandra is a scalable multi-master database with no single points of failure.
- **HBase:** A scalable, distributed database, HBase supports structured data storage for large tables.
- **Hive:** Hive is a data warehouse infrastructure that provides data summaries and ad-hoc querying.
- **Pig:** This is a high-level data flow language and execution framework for parallel computation.
- **Spark:** A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications.
- **Tez:** Tez is a generalized data flow programming framework built on Hadoop YARN that provides a powerful and flexible engine to execute an arbitrary DAG of tasks to process data for both batch and interactive use-cases.

Figure 1. Commonly used tools in Hadoop.

What is Hadoop good for?

In general, Hadoop is great for MapReduce data analysis on huge amounts of data. Its specific use cases include: data searching, data analysis, data reporting, large-scale indexing of files (e.g., log files or data from web crawlers), and other data processing tasks.

Processing Large Volumes of Data

By "large" data volumes, the industry is referring to a size of at least terabytes or petabytes of data. For companies that operate on not-so-large (i.e. gigabytes) data sets, Hadoop is not recommended since there are plenty of RDBMs and NoSQL database systems available with a much lower cost of implementation and maintenance (Malik et al., 2018). Companies with small data sets right now might consider Hadoop if they are expecting their data size to expand rapidly in the near future due to various factors. In such a case, careful planning should be made especially if these companies would like all the raw data to always be available for flexible data processing.

Storing a Diverse Set of Data

Hadoop can store and process data of various sizes, types, and even different versions of a particular data format across varying time periods. A user can change how they process and analyze Hadoop data at any time. This flexibility allows companies to engage in innovative development, while still processing huge amounts of data, rather than focusing in tedious data migrations.

When is Hadoop NOT a desirable option?

Real-Time Data Analysis

Hadoop works by the batch, processing long-running jobs over huge data sets (Jin, Hao, Wang, & Yue, 2019). Since these jobs take much more time to process than a relational database query on some tables, Hadoop is not ideal for real-time data analysis.

Nonetheless, an alternative solution exists. For companies looking to run real-time data analysis on their Hadoop database, storing the data in HDFS and using the Spark framework allows processing to be done in real-time by using in-memory data. This results in a 100x speed-up.

Small Sets of Data

The core function of Hadoop is the storage and processing of big data, especially the processing of big datasets. However, in practice, companies may have numerous small files, and Hadoop has many flaws when dealing with these small files (Liu, 2019).

To be stored in the Hadoop file system, each resource small file needs to occupy a block of data. With the increase of resource small files, many defects of the processing of HDFS appear, as follows. Liu details these defects and their causes in his paper: “Storage-Optimization Method for Massive Small Files of Agricultural Resources Based on Hadoop” (2019). All these defects affect HDFS’s ability to provide data storage services for a cloud storage system that can generate large amounts of small files.

Relational Database System

Due to slow response times as mentioned earlier on, Hadoop should not be used for a relational database. A possible solution for this issue is to use the Hive SQL engine to provide data

summaries and support ad-hoc querying. Hive allows users to project some structure onto a Hadoop data and then query using a SQL-like language called HiveQL.

Graph-based and Non-Parallel Data Processing

MapReduce works very well in situations where variables are processed independently. However, when a user needs to process variables jointly, and sometimes with multiple correlations between the variables, this model does not work. Hadoop is not suitable for graph-based data processing which is a complex network of data depending on other data (Jeong & Cha, 2019). To mitigate this problem, users could use the Apache Tez framework for graph-based approach to process data using YARN instead of the more linear MapReduce workflow. Given the industry's increasing need for parallel data processing, Hadoop solutions are being developed at a fast pace (Zhao et al., 2019).

Real World Case Studies

Big Data has quickly become an established fact for Fortune 1000 firms (Bean, 2016). Here are some statistics that bolster that claim:

- 63% of firms report having Big Data in production in 2015, up from just 5% in 2012
- 63% of firms reported that they expect to invest more than $10 million in Big Data by 2017, up from 24% in 2012
- 54% of firms say they have appointed a Chief Data Officer, up from 12% in 2012
- 70% of firms report that Big Data is of critical importance to their firms, up from 21% in 2012
- At the top end of the investment scale, 27% of firms say they will invest greater than $50 million in Big Data by 2017, up from 5% of firms that invested this amount in 2015

Figure 2. Big data industry statistics. Source: HBR (Bean, 2016)

Success Stories in Hadoop

Expedia: Hadoop Provides Scalability

In a 2017 conference, Mark Okerstrom, CEO of Expedia, states that the company receives 600 million visits on its website and 55 million phone calls every month. As such, they maintain several critical, high-volumes applications on AWS, such as the Global Deals Engine (GDE) that delivers deals to its online partners and allows Expedia to create custom websites and applications using Expedia APIs and product inventory tools.

Expedia provisions Hadoop clusters using Amazon Elastic Map Reduce (Amazon EMR) to analyze and process streams of data coming from Expedia's global network of websites, primarily clickstream, user interaction, and supply data, which is stored on Amazon Simple Storage Service (Amazon S3). Expedia processes approximately 240 requests every second.

Hadoop and AWS allow Expedia to add a new cluster to manage GDE and other high-volume applications without worrying about the infrastructure. This allows the company to scale and use the infrastructure efficiently.

British Airways: Cost Reduction in Data Storage

British Airways deployed Hadoop in April 2015 as a data archive for legal cases. Previously theses were stored on an enterprise data warehouse which was costly for the airline.

Alan Spanos, who leads the airline's data analytics team at the company in 2015, states that since deploying Hortonworks 2.2 HDP, British Airways has gained ROI within a year, and is able to deliver 75% more free space for new projects, translating directly into cost reductions for the airline.

The airline use Puppet and YARN, and hopes to build a core layer of applications including Kerberos for security authentication, Ambari, Falcon and Zookeeper for deployment and management as well as Oozie and Control-M for scheduling as part of its Hadoop architecture.

Although this was intended to be a one-off archiving project, the implementation team was interested in doing further exploratory work. The vision that they had was to do more stuff on Hadoop and leverage data-as-a-service, extensively, going forward.

Sears Holdings: From 8 Weeks to 1 Week

In the early 2010s, Sears Holdings concluded that it needed to generate greater value from the huge amounts of customer, product, and promotion data it collected from its Sears, Craftsman, and Lands' End brands. In the past, Sears required about eight weeks to generate personalized promotions, at which point many of these promotions were no longer optimal for the company. The reason for the lengthy duration is because the data required for these large-scale analyses were both voluminous and extremely fragmented, i.e. housed in many databases and "data warehouses" maintained by the various brands (McAfee & Brynjolfsson 2012).

Sears turned to popular technologies and practices of big data for a faster and cheaper way to do data analytics. Sears' approach started with a Hadoop cluster that it uses to store incoming data from all its brands and to hold data from existing data warehouses. The company then conducted analyses on the cluster directly, avoiding the time-consuming complexities of extracting data from different sources and combining them so that they can be analyzed.

The result of this change is a shorter period of time needed to produce a comprehensive set of promotions (from eight weeks to just one). These promotions are also of better quality, because

they are more timely, granular, and customized. Sears's Hadoop cluster was able to store and process several petabytes of data at a fraction of the cost of a comparable standard data warehouse.

The case studies of Expedia, British Airways, and Sears illustrate the power of big data technologies such as Hadoop, which allows more accurate predictions, drives better decisions as well as precise interventions, and with seemingly limitless scalability.

Now that we have seen some Hadoop successes, let us examine the times when Hadoop failed to live up to its expectations.

Times When Hadoop Failed Expectations

In 2017, data publication, Datanami, published an article that features accounts from companies who have used Hadoop. Among those interviewed was the CEO of Snowflake Computing, Bob Muglia. According to him, "the number of customers who have actually successfully tamed Hadoop is probably less than 20 and it might be less than 10… that's just nuts given how long that product, that technology has been in the market and how much general industry energy has gone into it" (Woodie, 2017).

Facebook: Hadoop Developed by Yahoo is Not Enough

Hadoop is the primary tool that Facebook uses, not only for analysis, but as an engine to power many features of the Facebook platform, including messaging (Lampitt, 2012). Facebook is credited for the development of Hive, an open source project and the most widely used access layer within the company to query Hadoop using a subset of SQL. Business analysts at Facebook rely heavily on Hive which allows them to "use Hadoop with standard business intelligence tools, as well as their homegrown, closed source, end-user tool, HiPal. HiPal is a graphical tool that talks

to Hive and enables data discovery, query authoring, charting, and dashboard creation" (Lampitt, 2012).

In terms of raw Hadoop capacity, Facebook reached its upper limit in 2012 when it declared itself as the owner of the world's largest Hadoop cluster, a size of 100 petabytes. However, this was not big enough. To increase its Hadoop capacity, Facebook launched the Prism project, which also supports geographically distributed Hadoop data stores.

The problem Facebook encountered was that Hadoop must confine data to one physical data center location. Although Hadoop is a batch processing system, it is tightly coupled, and will not tolerate more than a few milliseconds delay among servers in a Hadoop cluster. With Prism, a logical abstraction layer is added so that a Hadoop cluster can run across multiple data centers, effectively removing limits on capacity.

According to Bobby Johnson, who helped run Facebook's Hadoop cluster before co-founding behavioral analytics company Interana, "Hadoop's strengths lie in serving as a cheap storage repository and for processing ETL batch workloads… But it's ill-suited for running interactive, user-facing applications… Getting answers out of Facebook's Hadoop environment was an exercise in patience and frustration… After years of banging our heads against it at Facebook, it was never great at it… It's really hard to dig into and actually get real answer from…You really have to understand how this thing works to get what you want" (Woodie, 2017).

Walmart, UPS, and P&G: Success Does Not Happen Overnight

While Yahoo spent years to develop what we know as Hadoop, industry giants who reported great returns on their big data strategies, such as Walmart, UPS, and P&G, did not experience overnight success either.

Walmart is not only an industry leader in global ecommerce and brick-and-mortar retail, they are also a leader in the use of Hadoop-based technologies to implement their new data-driven approach to business. According to *Real World Hadoop* author, Ellen Friedman, Walmart's approach is a great example of Hadoop-based technologies used successfully in production (Friedman 2015).

Jeremy King, Walmart's global ecommerce chief technology officer and senior vice president, described how this transformation began in 2012 with their first Hadoop cluster. Now they work with tens of petabytes of detailed, highly valuable transactional data for their 245 million customers who shop online or in thousands of stores. Therefore, it is easy to see that Walmart's success happened over years of perfection. Like Facebook, Walmart had to generate many internal solutions to overcome the limitations of Hadoop. Sometimes, this means bringing in other technologies such as NoSql.

Gartner's Report: A Lack of Intent

According to a Gartner's report, "only 26 percent of respondents claim to be either deploying, piloting or experimenting with Hadoop, while 11 percent plan to invest within 12 months and seven percent are planning investment in 24 months. Responses pointed to two interesting reasons for the lack of intent. First, several responded that Hadoop was simply not a priority. The second was that Hadoop was overkill for the problems the business faced, implying the opportunity costs of implementing Hadoop were too high relative to the expected benefit" (2015).

The following subsections include a list of other possible reasons for this lack of intent.

Lack of Expertise

Johnson mentioned that "Hadoop is great if you're a data scientist who knows how to code in MapReduce or Pig… The Hadoop community has so far failed to account for the poor performance and high complexity of Hadoop. The Hadoop ecosystem is still basically in the hands of a small number of experts" (Woodie, 2017). The lack of expertise within an organization limits the knowledge of how much potential is untapped with a Hadoop implementation.

Time and Resource Intensive

From the deployment of the first node, designing a data architecture framework, and obtaining the data for storage, to querying it requires much effort. Instead of provisioning data in a traditional database management system in weeks, Hadoop could take months or even years.

Other Cost-effective Platforms Exist

The center of the big data universe has moved away from Hadoop to the cloud, where companies can store data in an object storage system like Amazon's S3, Microsoft Azure Blob Storage, and Google Cloud Storage; these cloud platforms are five times cheaper than storing it on HDFS.

Overkill for Common Business Use Cases

Last but not least, unless a company has a large volume of unstructured data like photos, videos, or sound files that it wants to analyze, a relational data warehouse will almost always outperform a Hadoop-based warehouse.

Concluding Thoughts

Prepare for Cultural Change

Most challenges companies struggle as they operationalize Big Data are related to people, not technology (Bean, 2016). Issues like organizational alignment, business process and adoption, and change management, should be considered before implementing a revolutionary technology such as Hadoop, as a company cannot successfully adopt Big Data without a cultural change.

Recruit the Right Talent

As data become cheaper, the complements to data become more valuable; one of which is data professionals. McAfee & Brynjolfsson believe that data workers should possess knowledge of Statistics and skillsets in Visualization tools. “The best data scientists are also comfortable speaking the language of business and helping leaders reformulate their challenges in ways that big data can tackle. Not surprisingly, people with these skills are hard to find and in great demand” (McAfee & Brynjolfsson 2012).

Our Sear’s case study took place in the early 2010s. Since skills and knowledge related to new data technologies were so rare during that time period, Sears started their transition by contracting some of the work to a company called Cloudera, giving its existing team of IT and analytics professionals time to become comfortable with the new tools and approaches.

Develop the Right Metrics

While most of the Fortune 1000 firms report implementing Big Data capabilities, few firms have shown how they will derive business value over time from these substantial investments (Bean, 2016). Metrics drive improvements and help organizations focus their people and resources

on what is important. The Six Sigma approach considers good metrics to reflect and support the various strategies for all aspects of the organization, including finance, marketing, competition, standards, or customer requirements and expectations. Our case studies have highlighted numerous Big Data accomplishments that involve operational cost savings or allowing the analysis of larger and more diverse sets of data. While all these are excellent success benchmarks, continuous improvements rely heavily on innovation, an often-overlooked metric.

References

Bean, R. (2016). Just Using Big Data Isn't Enough Anymore. *Harvard Business Review.* Retrieved from: https://hbr.org/2016/02/just-using-big-data-isnt-enough-anymore

Bedeley, R., & Iyer, L. (2014). Big Data Opportunities and Challenges: The Case of Banking Industry. *Association for Information Systems (AIS)*. Retrieved from: https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1001&context=sais2014

Chen, H., Chiang, R., & Storey, V. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly, 36*(4), 1165. doi: 10.2307/41703503

Friedman, E. (2015). Walmart: Harvesting Value from Big Data with Hadoop & NoSQL. MAPR [Blog Post]. Retrieved from: https://mapr.com/blog/walmart-harvesting-value-big-data-hadoop-nosql/

Gartner (2015). Gartner Survey Highlights Challenges to Hadoop Adoption. Retrieved from: https://www.gartner.com/en/newsroom/press-releases/2015-05-13-gartner-survey-highlights-challenges-to-hadoop-adoption

Jin, P., Hao, X., Wang, X., & Yue, L. (2019). Energy-Efficient Task Scheduling for CPU-Intensive Streaming Jobs on Hadoop. *IEEE Transactions on Parallel & Distributed Systems, 30*(6), 1298–1311. Retrieved from: https://doi-org.ezproxy.spu.edu/10.1109/TPDS.2018.2881176

Jeong, H., & Cha, K. (2019). An Efficient MapReduce-Based Parallel Processing Framework for User-Based Collaborative Filtering. *Symmetry (20738994), 11*(6), 748. Retrieved from: https://doi-org.ezproxy.spu.edu/10.3390/sym11060748

Lampitt, A. (2012). Facebook pushes the limits of Hadoop. InfoWorld. Retrieved from: https://www.infoworld.com/article/2616022/facebook-pushes-the-limits-of-hadoop.html

Liu, J. (2019). Storage-Optimization Method for Massive Small Files of Agricultural Resources Based on Hadoop. *Journal of Advanced Computational Intelligence & Intelligent Informatics, 23*(4), 634–640. Retrieved from: http://search.ebscohost.com.ezproxy.spu.edu/login.aspx?direct=true&AuthType=ip&db=a9h&AN=137628199&site=ehost-live

Lycett, M. (2013). 'Datafication': making sense of (big) data in a complex world. *European Journal of Information Systems (EJIS), 22*(4), 381. Retrieved from: https://link.springer.com/article/10.1057/ejis.2013.10

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution. *Harvard Business Review.* Retrieved from: https://hbr.org/2012/10/big-data-the-management-revolution

Malik, M., Neshatpour, K., Rafatirad, S., Joshi, R. V., Mohsenin, T., Ghasemzadeh, H., & Homayoun, H. (2019). Big vs little core for energy-efficient Hadoop computing. *Journal*

of Parallel & Distributed Computing, 129, 110–124. Retrieved from: https://doi-org.ezproxy.spu.edu/10.1016/j.jpdc.2018.02.017

Mendes, A., Gueera, H., Gomes, L., Oliveira, A., & Cavique, L. (2016). Big Data in SATA Airline finding new solutions for old problems. *International Journal of Computer Science and Information Security (IJCSIS), 14*(8). Retrieved from: https://repositorioaberto.uab.pt/bitstream/10400.2/7650/1/Big_Data_in_SATA_Airline_finding_new_sol.pdf

Müller, O., Fay, M., & Brocke, J.V. (2018). The Effect of Big Data and Analytics on Firm Performance: An Econometric Analysis Considering Industry Characteristics. *Journal of Management Information Systems (JMIS), 35,* 488-509.

Rehman, M., Liew, C., Abbas, A. et al. (2016). Big Data Reduction Methods: A Survey. *Data Science & Engineering, 1*(4), 265-284. Retrieved from: https://doi.org/10.1007/s41019-016-0022-0

Rodda, J., & VijayaKumari, R. (2018). Big Data, Technologies and Trends A Study. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 3*(7). ISSN : 2456-3307

Woodie, A. (2017). Hadoop Has Failed Us, Tech Experts Say. Datanami. Retrieved from: https://www.datanami.com/2017/03/13/hadoop-failed-us-tech-experts-say/

Woodie, A. (2018). Is Hadoop Officially Dead? Datanami. Retrieved from: https://www.datanami.com/2018/10/18/is-hadoop-officially-dead/

Zhao, X., Zhang, J., Qin, X., Cai, J., & Ma, Y. (2019). Parallel mining of contextual outlier using sparse subspace. *Expert Systems with Applications, 126,* 158–170. Retrieved from: https://doi-org.ezproxy.spu.edu/10.1016/j.eswa.2019.02.020