A business perspective on Cloudera Distribution Including Apache Hadoop

Gloria Mokberi
Information and Communication
Faculty of engineering, LTH
Lund, Sweden
gl6645mo-s@student.lu.se

Matilda Andersson
Computer Science
Faculty of engineering, LTH
Lund, Sweden
ma4748an-s@student.lu.se

***Abstract*—The interest in big data is rapidly increasing and with it, also the demands on accompanying tools. Apache Hadoop is an open-source framework that simplifies data processing and this report describes and analyses the distribution offered by Cloudera. Initially, the ecosystem encircling the Apache Hadoop project is examined and the Cloudera distribution's role in this. An overview of the financial driving forces and the open-core business model is given, followed by the legal consequences of the Apache 2.0 license and Cloudera's intellectual property rights. Lastly, ethical dilemmas related to big data like dual-use, privacy questions, and algorithmic biases are discussed. The report concludes that, even though there are many concerns regarding data analysis, the societal benefits are far greater than the downsides.**

***Keywords*—big data, Apache Hadoop, cloud computing, open-source, Cloudera**

I. Introduction

During the last two decades, the amount of data available in the world has rapidly been increasing and is, in addition, expected to be doubling every other two years [1]. More and more companies are also becoming aware of the business insights possible to gain from a well-conducted data analysis and therefore, are using data as a foundation for their business decisions [2]. With this growing interest in big data, the demands on the hardware and infrastructure enabling the collection, storing, and processing is also increasing [3]. This allows for both open-source projects and profitable companies to enter the market with a solution. This report will focus on one of these offered solutions, namely the 'Cloudera distribution including Apache Hadoop'.

A. Big data

According to a number of researchers [4] [5], the term big data can be defined using the 3V's [4]:

- **Variety**: The data can be of multiple categories like raw or structured and, can be sourced from various places like audio, documents, internal business-related data, or websites.
- **Volume:** The size of the data referred to is usually greater than terabytes or petabytes [5].
- **Velocity:** Refers to the speed of which data flows at various stages of the processing.

B. Apache Hadoop

Apache Hadoop is an open-source software framework that enables massive scale data processing and storage to be done over a cluster of distributed computers rather than having the data processing being confined to a single server or computer. The core of Apache Hadoop is made up of two main components, the first component is a storage system known as Hadoop Distributed File System and the second component, being a software framework named Mapreduce, is responsible for the processing of data. The Apache Hadoop software being open-source results in multiple augmented versions of the original, with the most commonly used distribution being the one provided by Cloudera [6], named Cloudera Distribution Including Apache Hadoop, or CDH in short. The CDH includes various Apache-licensed open-source software components that have been reduced into one commercially packaged tool for storing, processing, modeling, and analyzing large amounts of data.

II. The CDH Ecosystem

The CDH acts as a niche player in a business and software ecosystem where the Apache Foundation is the platform leader and the following subsections give a high-level description of this ecosystem.

A. The Apache Hadoop ecosystem

The nonprofit Apache Software foundation provides and governs numerous open-source projects, of which a selection of software projects makes up the foundation of the CDH distribution that, in turn, is offered as a 100% open-source product. The software ecosystem surrounding the CDH software has its focal point at the platform leader [7], the Apache foundation, as they are overseeing the essential parts that make up the Cloudera Hadoop distribution. Consequently, this results in the Apache foundation influencing Cloudera and other actors on the market who also come in contact with the various Apache projects. The actors that are primary users of the Apache projects will take on the role of niche players as they use technology from the platform leader to construct a niche product offered to different sections of the market. In the ecosystem encircling the Apache Hadoop project, we find niche players including Cloudera, Hortonworks Data

Platform, MapR, and Greenplum HD who all use similar Apache projects to construct an open-source distribution with Hadoop as a base. What differs them are their business models, the modules and features included in their software and their primary markets, but the software ecosystem as a whole enables data storage and data processing to be made in a more resource-efficient way [6].

B. Open source components and user data

The Hadoop distribution offered by Cloudera is made up of a plurality of different Apache projects and a couple of self-produced modules, resulting in Cloudera being heavily reliant on the Apache foundation and directly affected by any changes to the projects introduced by the foundation. The main Apache modules included in the CDH are Apache Hadoop, Apache Spark, Apache Impala, Apache Kudu, and Apache HBase [8] and have, together with other open-source Apache modules, been interconnected by Cloudera in order to create their open-source software for handling substantial amounts of data. As the CDH performs particularly well for large-scale data storing and processing they mainly attract enterprise companies and bigger medium-sized businesses like Apple, Cisco, Citibank, and Dell among others. These companies also become a part of the business ecosystem as they are users of the CDH, and, potential users of other non-free proprietary Cloudera products. The reader should be aware of the existence of auxiliary software products and services offered by Cloudera as the CDH in itself is merely a software tool and does not initiate a transfer of user data. However, the supplementary offering of data storage and cloud-based data analysis gives a clear picture of the additional value offered by the transfer of user data. The origin of this data depends completely on where the company using the Cloudera services have obtained it.

C. Common interests

As previously mentioned, various companies and organizations have incorporated the open-source project Apache Hadoop as a basis for their products and have then built augmented products with an additional value offering around this open-source module. Being an open-source project, contributions to these modules can be made by anyone following the given contribution rules [9], allowing for updates admitted by the Apache Foundation being incorporated in all products and distributions containing the Apache Hadoop. Consequently, Cloudera and other organizations largely dependent on the Apache Hadoop, allocate resources and time for contributing and committing to the further development of the project, with Cloudera being the largest contributor. There are many organizations and institutions that directly, or indirectly through Hadoop distributions, are using the Apache Hadoop framework. A few organizations, Cloudera included, are packaging the Apache Hadoop framework together with other modules and offer this augmented product to the market. The users of these distributions are not directly in contact with the original Apache Hadoop framework, but instead uses the distribution

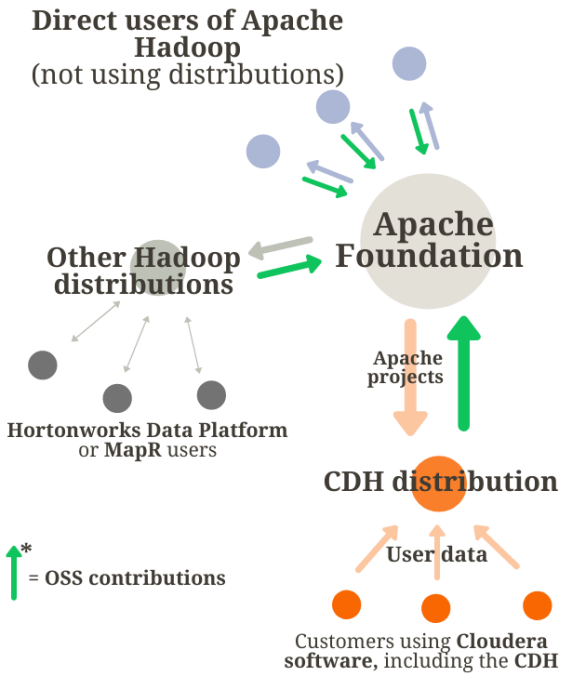

Fig. 1. Business Ecosystem surrounding Apache Hadoop.

offered by Cloudera or other niche players. As the interest in big data is growing bigger and bigger, the ecosystem surrounding the Apache Hadoop framework touches and influences an expanding range of sectors including, but not limited to, customers retail, government, financial service, healthcare, life sciences, digital media, advertising, networking, and telephony enterprises.

III. CDH from a business perspective

Cloudera's business model follows an open core model, meaning they monetize commercially produced open-source software. Thus offering a feature-limited version of their service for free download, while the full version is commercial and is offered as a proprietary software.

A. The CDH Business Model

Cloudera offers a modern and integrated platform for big data that is scalable and flexible, which makes it easy to manage rapidly increasing volumes as well as varieties of data in the customer's enterprise. The service enables the customer to deploy and manage Apache Hadoop and related OSS projects, manipulate and analyze gathered data, whether it is structured or unstructured [6].

Cloudera offers five primary value propositions: innovation, convenience, performance, risk reduction, and brand/status. The benefits that Cloudera offers is a centralized administration tool, unified batch processing, real-time data insights, interactive SQL and, role-based access control. Additionally, their solutions can be integrated with numerous hardware and software solutions, resulting in a very rich ecosystem and making them very adept to deliver a wide spectrum of different

services [8] [6]. Since the Apache Hadoop is an open-source project, its role in Cloudera's proprietary business model is being used as a foundation in their service, making them one of Cloudera's key resources. Thus, Cloudera's main source of revenue comes from their customers' subscription to the extended modules they have developed and integrated with the Apache Hadoop project, which offers several versatile features. Each module offers its own spectrum of services in the ecosystem, making the development of these extended modules one of Cloudera's main cost structures. The expenses also include costs for marketing and sales, customer support together with server and hardware related expenses [6].

B. Business Aspects

The usages of big data are many and therefore Cloudera's services are useful in a variety of different fields and industries, giving them a huge customer segment. A few examples of fields that hold investors are banking and financial services, retail, healthcare, media, and entertainment. The different aspects of benefits that Cloudera provides depend on the field it is used in. Common denominators amongst their customers are that they utilize Cloudera for the benefits of either monetary profit, preventing fraudulent activities, or to improve customer care, in one way or another [10]. Below is a more detailed analysis of the different fields.

C. The benefit of using big data

The majority of the enterprises that exist today collect and store large amounts of data on their customers, such as demographics, previous searches and previous purchases, geographic location, personal information, social networks, and keeping a record of past transactions etc. With access to this type of information, enterprises have access to a goldmine of potential benefits if the data is analyzed. This is where Cloudera comes into the picture, even though they do not collect the data they do offer cloud storage and analytics of this data. Using data such as sales data, industry trends, and market indicators, enterprises can better understand customers' behaviors and habits. This is for example what recommendation systems are based on. Based on this data the system can recommend other similar or complementary products and thus optimizing the sales in the retail industry [11]. Understanding how customers engage and respond to marketing campaigns gives the business useful contextualized data that helps to guide future promotions and strategical decisions and to enhance marketing strategy. This advantage can help enterprises stand out from their competitors, as well as guiding them in how they can optimize their customer service. The ability to turn information into action to create experiences that are more personal, responsive, and accurate is greater the bigger the collection of data about a user's taste and preferences are. Especially in the competitive area of the service industry, it is significantly beneficial for enterprises to know their customer's preferences in order to offer more customized services. For example, the insight into historical data can help concerned parties to learn more about their customer's behavior and preferences, identify existing trends, and predict future outcomes. Allowing retailers to properly appeal to a particular group, highlight their best selling items, and avoid underperforming ones [11]. Resulting in enhanced service quality, increased customer satisfaction, and gained profit. Since each customer has their own preferences, customization is key [10].

All this overwhelming amount of data needs to be gathered somewhere, and there will be more to come in the future. It is estimated that so far, every person on the planet creates 1.7 megabytes of new information every second [10]. Even though this vast amount of data is gathered it has been reported that only 0.5% of all collected data has been utilized or analyzed [10]. Meaning, finding innovative ways to utilize this existing data is a major challenge for many enterprises. The solution is storing data in an infrastructure that can split data across multiple physical servers, also known as cloud computing. Cloud computing systems such as Cloudera offers a cost-efficient solution to this issue. They facilitate the large storage and processing requirements and needs that their clients have, by offering them a solution where they only pay for the amounts of data that they actually use. Meaning that the enterprises do not have to invest in large quantities of infrastructure [10].

IV. Ethical aspects

The field of data warehousing and machine learning that Cloudera is a part of concerns several ethical dilemmas like dual-use, algorithmic biases, and privacy questions. Below we go through a few of the occurring ones when dealing with the storage and analysis of big data.

A. Dual-use and shopping addiction

As beneficial and effective as the big data industry has been from several perspectives, the emphasis is mostly on positive mechanisms of implementing big data analytics, with little attention paid to the detrimental consequences. For example, retail often uses recommendation systems where the website will recommend other items to consumers as replacements or complimentary products. Usually, this is a good feature for consumers who do not have problems with, for example, shopping addiction. It is considered a behavioral addiction that has to do with the individual's lack of impulse control. Thus this kind of feature empowers these people to spend money they might not have, and in some sense unconsciously, avails their addictive behavior [11]. One could argue that this is an example where Cloudera ends up in a dual-use dilemma. In the article of Software engineering in a digital world, A. Rashid et al. [12] defines the dual-use dilemma as "The dual-use dilemma arises as a consequence of the fact that the same piece of scientific research sometimes has the potential to be used for harm as well as for good". Cloudera might not directly have a partaking in the mentioned problem occurring, as the recommendation system per se is not offered or built by Cloudera. Nonetheless, the software Cloudera offers does play a part in the building of these systems, and thus their software

that is intended to be used for something good is indirectly part of this negative effect. This particular example is one of many probable dual-use dilemmas they are a part of.

B. Biases

It is seen in many cases that decision making via mathematical models driven by big data has led to unfair outcomes as a result of reinforced discrimination. Thus making Cloudera, which offers analytical tools in their modules, a target for this ethical dilemma. As algorithms absorb conscious or unconscious prejudices on the part of the developers, or by creeping in through undetected errors, the outcomes of a biased algorithm would give distorted outcomes. Likely in a way that is offensive and discriminating to individuals who are affected. Biased algorithms may also appear from details in data sets that are unrecognized when training data. Nonetheless, if such a situation persists for long enough the business could incur damage to its reputation and potentially ending up in trouble with the law as a result of discrimination. Therefore, it is important for developers to build transparency into their algorithms to avoid such harms, and for businesses to adhere to responsibilities for their acquirements, something Cloudera claims that they do [8] [10].

C. Privacy, transparency and integrity

Privacy is a huge concern when it comes to gathering and storing Big Data, thus making this one of the main ethical dilemmas in this area. Studies show that there is an increasing concern amongst consumers regarding their privacy on the internet. How aware are consumers to which extent information is being gathered about them and how it is being used? There is constantly a hidden and systematic aggregation of data about individuals, and there can be a fine line between using this data as an act of something good and it interfering with the individual's privacy. For example, when personal data gets cross-matched and individuals get targeted there is an invasion of the individual's integrity without them knowing. Although Cloudera is only a "container" for all this big data that is being gathered, one could still argue that they are indirectly involved since they are providing the software that promotes this interference of privacy and integrity. To reinforce transparency and protection of consumer privacy there are data regulation laws that regulate the obligations that businesses have when gathering personal data about consumers, one of them being the obligation of informing the individual when their data is being gathered [13]. In addition to this, Cloudera provides a centralized framework that complies with the different data protection rules that are applicable, so that their direct customers do not have to worry about the security and governance of the data they are responsible for, nor whether the correct legislation and regulations are being followed [8].

V. Legal aspects

Cloudera owns a number of patents and intellectual properties that are presented in this section together with a review of the Apache 2.0 license affecting the CDH.

A. Navigating an open-source world

The use of open-source software can quickly go from being like a beautiful summer day to a war-like thunderstorm if licenses and regulations are not followed correctly. Unlike proprietary programs, open-source projects do not follow traditional contractual rights but bind to standardized licensing agreements that define the permitted use and varying degrees of freedom. The Apache license is one of the most widely used licenses today and the projects released by the Apache foundation are all governed by various editions of this license. This license is permissive and does not require eventual modifications or derivative work to be released under the same license. The Apache Hadoop project follows the Apache 2.0 regulations and allows for the freedom to use, modify, and redistribute the software without concern for royalties [14]. Consequently, the Hadoop distribution offered by Cloudera needs not to be released under the same license, but nonetheless, Cloudera has chosen to keep the Apache 2.0 license for their CDH distribution, allowing for a completely free and open product.

B. Cloudera and intellectual property rights

While the CDH is distributed as open-source software, there are still other aspects of the company that falls under other regulations. In addition to the CDH, Cloudera is offering a number of other software products and services that do not fall under the open-source legislation, allowing the organization to develop products eligible for patents. As a matter of fact, Cloudera stands as the owner for a number of US, Canadian, Korean, Australian and Japanese registered patents [15]. One Cloudera owned patent affecting the CDH concerns the optimization of SQL-like queries in Hadoop [16]. More specifically, this patent allows Cloudera to exclusively use their new technology that improves the performance and speed of interactive queries. The patented software solution can give Cloudera an advantage by offering a Hadoop distribution which is superior to other alternatives and in turn, attract more potential customers paying for their other products.

In addition to patents, Cloudera has a number of intellectual property rights, holding the claim to protection of various forms of creations of the mind. For instance, the following assets: the name Cloudera and their logo, product or service names, and slogans connected to Cloudera cannot be used by other organizations since they are protected by copyrights and trademarks [8]. However, the name "Hadoop" and its logo are trademarked by the Apache foundation that also holds the copyright on some code sections in the Hadoop codebase. This results in Cloudera not being able to use these resources directly without referring to their rightful owner.

C. Legal aspects of data storage in the cloud

As the CDH product can be seen as a tool enabling its users to store and analyze data the legal questions regarding data collection and storage arise [17] [18]. No source could be found that proves whether Cloudera is collecting data of their direct customers or not. However, they offer the service of

cloud storage, by storing their customers' data on their own servers. With cloud computing being a global phenomenon, the geographical locations are blurred and it is not always clear where data is stored. Cloudera states that they may process data in multiple countries but always participates in, and has certified its compliance with the European and U.S data privacy laws [19]. However, they also state: "In certain situations, Cloudera may be required to disclose personal data in response to lawful requests by public authorities [...] We will retain personal data we process on behalf of our Clients for as long as needed. Cloudera will retain this data as necessary to comply with our legal obligations, resolve disputes, and enforce our agreements." [19]. These statements demonstrate the concerns about data privacy. In the cloud, the law becomes fuzzy, and although Cloudera states their compliance with local laws, a third party can never be completely sure of their data's destiny.

VI. Summary

In conclusion, the CDH is an open-source Apache Hadoop distribution provided by the enterprise software company Cloudera Inc. The CDH is designed to solve problems related to big data, such as issues regarding storage and analysis of massive amounts of data. The CDH consists of various software modules, where a majority of these are open-source projects provided- and governed by the Apache Foundation. Furthermore, these projects are licensed under the Apache 2.0 regulations leading the CDH distribution to be released under the same legislation. As the CDH consists of open-source modules and thus, has to remain free to use, Cloudera Inc. incorporates an open-core business model, offering proprietary software and services in connection to the CDH.

Today, with data being a widespread phenomenon used by many organizations and companies, the software and services that improve and simplify the big data process will only increase in value. This as the benefits of analyzing big data are growing larger. However, there are two sides to this big data coin, and ethical and legal questions arise regarding e.g. privacy and the consequences of insights obtained from data analysis. Who has access to uploaded data and what can happen with sensitive information stored on remote servers? Is it ethically defensible for a company to tempt shopping addicts with personalized offers or how do biases in the data affect disadvantaged groups? Navigating in the world of big data can be difficult and, although the disadvantages can be many, the advantages are even greater. With data analysis numerous societal problems can be solved, improving the life quality of many in aspects like health care, criminal justice, fraud detection, and promoting innovation. Where the road of big data will take us is left to the future to show - but as a first step, the right software and tools are needed. This is where Cloudera and the CDH come in.

References

[1] C. Ji, Y. Li, W. Qiu, Y. Jin, Y. Xu, U. Awada, K. Li, and W. Qu, "Big data processing: Big challenges and opportunities," *Journal of Interconnection Networks*, vol. 13, pp. 177–180, 2012.

[2] X. Wu, X. Zhu, G. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, 2014.

[3] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile networks and applications*, vol. 19, no. 2, pp. 171–209, 2014.

[4] M. Al-Mekhlal and A. A. Khwaja, "A synthesis of big data definition and characteristics," in *2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*. IEEE, 2019, pp. 314–322.

[5] S. Sagiroglu and D. Sinanc, "Big data: A review," in *2013 international conference on collaboration technologies and systems (CTS)*. IEEE, 2013, pp. 42–47.

[6] G. s. Bhathal and A. S. Dhiman, "Big data solution: Improvised distributions framework of hadoop," in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2018, pp. 35–38.

[7] A. Gawer, M. A. Cusumano *et al.*, *Platform leadership: How Intel, Microsoft, and Cisco drive industry innovation*. Harvard Business School Press Boston, MA, 2002, vol. 5.

[8] C. Inc, "Cloudera introduction," Accessed: Dec. 11, 2020. [Online]. Available: https://docs.cloudera.com/content/www/en-us/documentation/enterprise/5-5-x/PDF/cloudera-introduction.pdf.

[9] A. S. Foundation, "How to contribute to apache hadoop," Accessed: Dec. 7, 2020. [Online]. Available: https://cwiki.apache.org/confluence/display/HADOOP/How+To+Contribute.

[10] M. C. Cohen, "Big data and service operations." *Production Operations Management*, vol. 27, no. 9, pp. 1709–1723, 2018.

[11] T. M. Le and S. Liaw, "Effects of pros and cons of applying big data analytics to consumers' responses in an e-commerce context," *Sustainability*, vol. 9, p. 798, 2017.

[12] A. Rashid, J. Weckert, and R. Lucas, "Software engineering ethics in a digital world," *Computer*, vol. 42, no. 6, pp. 34–41, 2009.

[13] DataGuidance, "Comparing privacy laws: Gdpr v. ccpa," Accessed: Dec. 27, 2020. [Online]. Available: https://fpf.org/wp-content/uploads/2018/11/GDPR_CCPA_Comparison-Guide.pdf.

[14] A. Sinclair, "License profile: Apache license, version 2.0," *IFOSS L. Rev.*, vol. 2, p. 107, 2010.

[15] C. Inc., "Legal: Cloudera patents," Accessed: Dec. 9, 2020. [Online]. Available: https://www.cloudera.com/legal/patents.html.

[16] M. Kornacker, J. Erickson, N. Li, L. Kuff, H. N. Robinson, A. Choi, and A. Behm, "Background format optimization for enhanced sql-like queries in hadoop," Oct. 25 2016, uS Patent 9,477,731.

[17] A. Fitzpatrick, M. Mcgrath, and R. G. Lennon, "Legal issues surrounding data storage on the cloud," in *2012 5th Romania Tier 2 Federation Grid, Cloud High Performance Computing Science (RQLCG)*, 2012, pp. 53–56.

[18] D. Song, E. Shi, I. Fischer, and U. Shankar, "Cloud data protection for the masses," *Computer*, vol. 45, no. 1, pp. 39–45, 2012.

[19] C. Inc., "Cloudera's privacy and data policies," Accessed: Dec. 11, 2020. [Online]. Available: https://www.cloudera.com/legal/policies.htmlhosting.

APPENDIX: CONTRIBUTION STATEMENT

The intention is and was from the beginning to distribute the workload evenly. All the planning, discussions on content and reviewing were done equally together. The individual contributions are presented below.

Matilda has been responsible for the ecosystem section; including everything from the literature search and construction of the figure, to the analysis and the writing of the content. She was also responsible for structuring and writing the introductory part, including literature search. Furthermore, she did the literature search, content structuring, analysis and writing of the legal section. And, lastly, she wrote the summery section together with Gloria.

Gloria has been responsible for the business section; including everything from the literature search to the analysis and writing of the content. In addition, she did the literature search, content structuring, analysis and writing of the ethics section. Together with Matilda she wrote the summary section. Lastly she was responsible of the abstract.