JOURNAL of COMPUTER SCIENCES INSTITUTE

JCSI 13 (2019) 283-287
Wysłane: 2019-06-18
Przyjęte: 2019-06-27

Overview of Big Data platforms

Gabriel Wróbel [a,b,*], Maciej Daniel Wikira [a]
[a]Institute of Computer Science, Lublin University of Technology, Nadbystrzycka 36B, 20-618 Lublin, Poland
[b] University of Oulu, Oulu, Finland

Abstract. The primary purpose of this paper is to present and provide main advantages and disadvantages of most popular big data platforms as well as their comparison in terms of ease of installation, work, performance and price, in order to find the most suitable solution to work with big sets of data. Nowadays, the data is largely analyzed by scientists not related to IT, so the ease of use and presentation of data is extremely important. The purpose of the assessment was to indicate the best IT tool for analyzing data from the point of view of a young analyst or scientist graduating and entering the labor market.

Keywords: big data; data analysis; platform; tool comparison

*Autor do korespondencji.

E-mail address: g.wrobel@pollub.edu.pl

Przegląd platform Big Data

Gabriel Wróbel [a,b,*], Maciej Daniel Wikira[a]
[a]Politechnika Lubelska,Instytut Informatyki, Nadbystrzycka 36B, 20-618 Lublin, Polska
[b]Uniwersytet Oulu, Oulu, Finlandia

Streszczenie. Głównym celem niniejszej pracy jest prezentacja głównych zalet oraz wad najbardziej popularnych platform big data, jak również porównanie ich pod względami łatwości instalacji, funkcjonalności, wydajności oraz ceny co pozwoli na wskazanie rozwiązania najlepiej dostosowanego do pracy z dużymi zbiorami danych. W dzisiejszych czasach dane są przetwarzane przez analityków niezwiązanych z branżą IT, w związku z czym bardzo istotne są kwestie łatwości użytkowania i prezentacji danych. Celem oceny jest wyznaczenie najlepszego narzędzia z branży IT dla analizy danych z perspektywy młodego analityka lub naukowca kończącego edukację i wchodzącego na rynek pracy.

Słowa kluczowe:big data; analiza danych; platforma; porównanie narzędzi

*Autor do korespondencji.

Adres e-mail: G.wrobel@pollub.edu.pl

1. Introduction

For decades, data was being collected from large number of websites, devices and sensors that can be used to carry out analyses in various fields of science and life. Banking, telecommunications, tourism, insurance, e-business, and energy –these are just some of the industries in which Big Data analysis is presentnowadays. In case of banks, modern tools allow, for example, toexamine the age of customers who use credit cards most often.

Big Data makes it possible even to examine monthly averagebills of customers who have given up their services. For the analysis of big datasets,big set of data analysis software tools is currently being used.These tools themselves function very well but can be problematic for users who are not familiar with the technology. To facilitate the work with large data sets, big data platform solutions have been developed that are designed to facilitate the use of analysis tools and increase the efficiency of working with big data. Platform data isa collection of tools that allows complex data analysis, machine learning, data storage and visualization. There is a possibilityto use solutions from various companies that can operate locally, such as Cloudera, Hortonworks, MapR Platform, as well as cloud solutions such as Amazon AWS, GoogleCloud, Microsoft Azure. Platforms Cloudera, Hortonworks, MapRand HDInsight (Microsoft Azure).Main advantages and disadvantages of these solutions have beenpresented in this article as well as their comparison based on difficulty of installationand price[6].

2. Big Data Tools

To clearly understand why platforms are so important in working with big data,it is important to consider what features platform can include. Therefore, Hadoop is introduced as the most typical set of tools.

The simple definition of the Hadoop operating principle is saving files and processing data. It is easy to imagine afile larger than the disk capacity of a standard PC. By thetraditional way, it is not possible to save such a file. Hadoop allows one to save files larger than the common disk capacity, so they can be stored on a given server or matrix, by distributing data to multiple clusters (matrices, servers). The second element of Hadoop is the ability to process this data. If a huge amount of data should be made available, traditionally it must be sent to a local device, which usually cannot cope with this task. Hadoop reverses this situation and, thanks to the MapReduce component, moves the processing tools towards the data[5].

Hadoop is developed by the Apache Software Foundation. An important element of Hadoop is HDFS. HDFS (Hadoop Distributed File System) is a technology that provides effective scaling of the storage layer. HDFS is a Java-based file system that has been adapted to work even on hardware with little capacity. Another important element of Hadoop is

the Data Processing Framework, which in the form of MapReduce is used to work with data. MapReduce runs a series of processes, each of which is a separated Java application that searches data. It is worth to highlight that queries are not being used in this case as they are in a relational database. In Hadoop, there are tools such as Hive (initially developed by Facebook) that allowsone to convert the query language into MapReduce tasks[5].

Nowadays, in data platforms solutions like Spark or YARN (Yet Another Resource Negotiator) can be found, which are in fact new generations of Hadoop and MapReduce solutions.This solves some limitations, but the main principles remain the same [5].

3. Platform Presentation

A. MapR

The MapR is an enterprise-class distribution for the Hadoop and Spark platforms. The MapR platform has been designed to provide the highest ease of use, performance, and reliability for users who need tools to analyze large data sets. The solution includes a full set of tools necessary for work. It is possible to work using a file system designed specifically for MapR (MapR-FS).In addition, it isalso possible to manage the documentary database MapR-DBand work using the streams of MapR. All the tools available in the Hadoop environment can be used, so there is alsoan opportunity to use HDFS file system or mapReduce. In case of MapR there is an alternative to mapReduce, as the platform supports the YARN architecture that manages files and tasks between clusters[1].

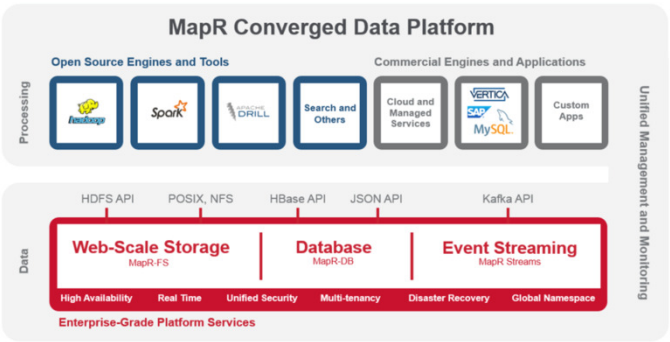

Fig.1. Visualization of MapR Data Platform architekture

Figure 1 shows the architectural layout of the tools contained in the MapR data platform as well as components and links between them[1].

MapR platform introduces snapshots that are remembering the situation of data at the specifiedmoment. This solution allowsto use the memory effectivelywhile keeping only the changes from the last snapshot. The abovementioned solution minimizes the possibility of data loss in the event of a failure.

MapR has a strong security based on the user's authorization, which occurs each time the user attempts to access the file. The platform checks whether a given user has access rights to the requested document[1].

MapR does not use name nodes to store file location information. Instead it provides high availability for MapReduce JobTracker and Direct Access NFS via server CLDB[1].

B. Cloudera

Cloudera provides a fully integrated and scalable platform designed to work with large and constantly growing, diverse data sets. The products and solutions introduced by Cloudera enable working with the Apache Hadoop tool and related tools for manipulating, analyzing, protecting and securing data[2]. The service architecture is shown on figure 2.

Cloudera provides the following products and tools:

- CDH – complete distribution of Apache Hadoop as well as other open-source projects such as Apache Impala or Cloudera Search. CDH is a solution, related to data security, providing software integration and hardware integration[2].
- Apache Impala – solution which enables parallel work with data using the SQL engine which in combination with Apache Hadoop makes available a wide range of BI capabilities. This is possible thanks to highly optimized architecture. The tool allows to direct SQL queries to files in the HDFS file system, divided between clusters using MapReduce or loaded into Hive tables. Impala has a YARN resource management component that makes it possible to construct SQL queries on a working cluster. Impala management is possible from the level of Cloudera Manager and secure the sentry framework[2].
- Cloudera Search –solution that allows searching the data placed in Hadoop or HBase systems in nearly real time. The search provides batch indexing, parallel text search, but does not require skills related to programming or skills related to the construction of SQL queries. Cloudera search is a fully scalable, flexible solution, and is included in the CDH. With this tool, we do not need to transfer data to perform business tasks[2].
- Cloudera Manager –the application used to monitor, manage, and diagnose problems occurring in the CDH platform. The solution works using the administrator or web client, which makes managing the platform much easier and more intuitive. Cloudera Manager contains an API with the help of which we can configure the Cloudera manager and also get information on the condition of clusters or matrices, and use this data in our own monitoring applications[2].
- Cloudera Navigator – complete tool for data management and data protection in the CDH platform. The solution enables administrators and analysts to explore data placed in Hadoop. Cloudera navigator makes it easy for businesses to store data according to the law through audits, data management, life cycle management and data encryption management [2].

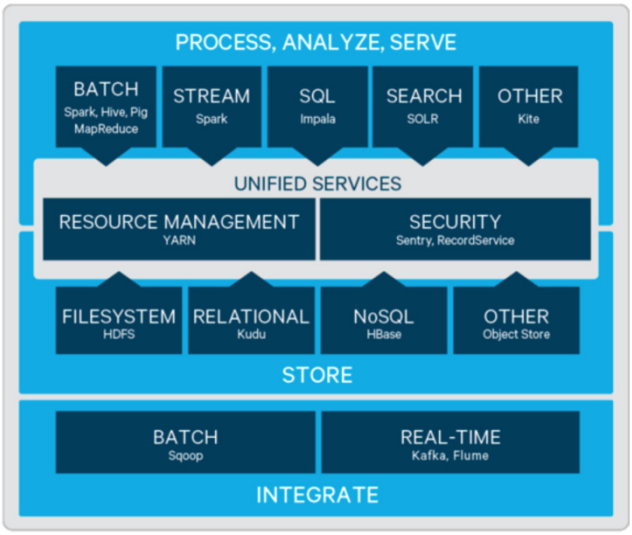

Fig. 2. Visualization of CDH Data Platform architekture.

C. *Hortonworks*

Hortonworks Data Platform (HDP) is an open support platform for Apache Hadoop, which provides a stable foundation for developing big data solutions in the Apache Hadoop ecosystem. HDP strongly focuses on container architecture which makes the solution more flexible and easy to scale. HDP includes tools such as TechPreview, TensorFLow, Apache Zeppelin or Apache Spark, which enablesapossibility to use machine learning and deep learning.These solutions can be very helpful in long-term data analysis. Moreover, the platform provides the possibility of deep learning through GPUs, which significantly streamlines the process[3].

Hortonworks introduces hybrid architecture so that the platform can be used in the cloud solution as well as on-premises[3]. The visualization of such architecture is presented on figure 3.

HDP provides higher availability of data with multiple name nodes, at significantly lower TCO with Erasure Coding.Erasure Coding enables certain features for data protection which until now have mostly been found in object data stores [3].

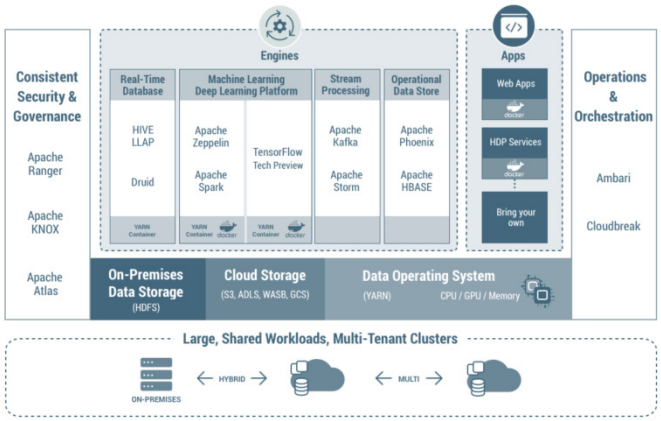

Fig.3. Visualization of HDP Data Platform architekture

In the last quarter of 2018, the Hortonworks and Cloudera corporations announced a merger, which meant that the two most competing solutions will now create one new solution [3].

D. *HDInsight (Microsoft Azure)*

Azure HDInsight is a cloud distribution of Hadoop service components from the Hortonworks Data Platform (HDP) platform. Azure HDInsight facilitates and speeds up the processing of huge amounts of data. It can be used with the most popular open source platforms such as Hadoop, Spark, Hive, LLAP, Kafka, Storm, R and more. With these platforms, implementations of various scenarios related to extracting, transforming and loading data, data storage, machine learning and the Internet of Things (IoT) are possible [4].

Using HDInsight, it is possible to perform interactive petabyte queries against structured or non-structured data in any format,and also to create models that combine them with business analysis tools.

Using the HDInsight service, streamed data can be processed, and received in real time from various devices [4].

There are following cluster types in the HDInsight service:

- Apache Hadoop –a platform using HDFS, YARN resource management, and a simple MapReduce programming model for parallel processing and analysis of batch data [4].
- Apache HBase – NoSQL database based on the Hadoop platform that provides random access and high consistency for large amounts of unstructured and partially structured data – potentially billions of rows multiplied by millions of columns[4].
- ML Service – server designed for hosting and managing parallel, distributed R processes. It enables data analysts, statisticians, and R language programmers to access scalable, distributed analysis methods in the HDInsight service[4] on-demand.
- Apache Storm – a distributed computing system that works in real time for a fast processing of large data streams. Storm is offered as a managed cluster in the HDInsight service[4].
- Apache Interactive Query preview (AKA: Live Long and Process) – cache memory in the memory for interactive and faster execution of Hive queries[4].
- Apache Kafka – open source platform that is used to create pipelines of streamed data, which also provides applications to handle this data. Additionally the Kafka platform includes a message queue function that allows users to publish and subscribe to data streams [4].

HDInsight is a comprehensive solution that does not require a technical person on the client's side because the support of all services remains within the service provider (Microsoft Azure). Due to the fact that HDInsight is a solution maintained in the cloud, all data is transferred to the servers of the service provider which relievesthe client of the need to have hardware architecture[4].

4. Platform Comparison

Undoubtedly, the big data platform solution makes it much easier to work with large data sets, but it is worth considering which platform should be used for that purpose. The following

paragraph will present several aspects that can facilitate the decision.

A. Knowledge of computer science

Knowledge in the field of computer science can be crucialin this case. If the user needs to perform more complex analytical activities and does not have system administration skills, he may need the help of a technical person. Such situationsgenerate higher investment costs. Knowledge about information systems refers mainly to the LINUX system in this context.The reason for that is the fact, that most platforms were designed to work on UNIX systems[5, 6].

In this case, it may be meaningful to compare the way and the complexity of preparing the work environment in which each platform operates. This comparison is shown in table 1.

Table 1. Table of installation complexity

Complexity of installation	
Cloudera	Installation and configuration of the CDH platform with the manager is an intermediate administrative task in which the installer will require knowledge of the UNIX system commands and basic knowledge in the field of operating system administration. Cloudera also provides images of virtual machines for a quick start of work as well as a file dockerfile through which a container can be run with an already configured platform. However, it is not a solution for enterprise and only for educational purposes[2 ,3].
MapRon-premise	Installation and configuration of the MapR platform is a complex task in terms of operating system administration skills. The producer provides a package with software in libraries available from the LINUX terminal level. The installation requires familiarity with the UNIX family of command systems and basic knowledge in the field of operating system administration[1].
MapR cloud	Customer is provided with an interactive form to choose the size of the solution he needs, depending on the price. Simple and basic service that does not require knowledge in the field of operating system administration[1].
HDInsight (Microsoft Azure)	The customer is provided with an interactive form to choose the size of a solution he needs, depending on the price.Simple operation providing a fairly extensive range of configuration options. The entire configuration does not require knowledge of the administration of operating systems and provides numerous configuration options[4].

In the comparisonabove, in terms of skill requirements and knowledge demanded from the client, HDInsight is on the lead, but right behind it, MapR is an equally interesting solution because it provides the opportunity to work in the cloud where the responsibility for administrative issues lies within the service provider. The CDH platform from Cloudera with Hortonworks comes up the worst in the given list since it is expects knowledge from the user and does not provide the tool.

B. Law Aspect

The usefulness of the solution is always contingent on the task to be performed and, in this case, some solutions may have advantages in one of the cases and be completely useless in another one. An example of such a phenomenon can be sensitive data that cannot be stored outside the state due to law provisions in a given country. In this case, the cloud solution may be unhelpful, because it may happen that a cloud service provider does not have a servers in a given country and the client is forced to invest in his own solution.Here, MapR has solution for working on AWS but also on-premises,what makes this platform very flexible. Cloudera solutions with Hortonworks can be more useful in such situation than solution introduced by Microsoft Azure (HDInsight) but it is not so flexible as MapR[1, 4].

C. Pricing

All producers analysedin this paper have plans for additional financing. The simplest and the cheapest offer packages were taken into account for the balancing of the statement.

In this case, Cloudera with Hortonworks offers its platform for commercial software solution for $ 2,000 per year, and also provides this platform for free for educational purposes for 60 days. MapR on-premises is available for free on trial license for 30 days and there is alsoan option of downloading the sand-box on one'sPC. The enterprise version is paid, and for a specific price,a contact with the producer must be made. For a cloud solution from MapR in the community version, the producer gives a price of $ 2.40 for the cluster's hour of work while working in the AWS cloud. HDInsight from Microsoft Azure does not provide a trial version. It is possible to ordera paid service from the cluster's working hours, which is about 10 € per hour and there is an option to enable and disable the service depending on your needs[1, 2, 4].

Regardless of the fact that the components of the platforms are mostly available under an open source license, the most reasonable solution seems to be Cloudera, which makes the software available in an annual subscription, regardless of use. Other producers do not quote prices on the commercial market or count the price per hour of work,which is normal in a cloud environment. However, in the cloud environment the leader turns out to be a MapR.

5. Summary

Nowadays, while working with large data sets,it is almost indispensable to see correlations in all areas of research.

With the help of open platforms, efficient work is possible, even for people who are not connected with computer science.

Platforms nowadays are very close to each other and very often overlap and cooperate with each other. It is very difficult to choose the right and, at the same time, the best one to work with Big Data because each of them has its advantages and disadvantages in various respects.

Bibliography

[1] MapR Producent website https://mapr.com/docs/51/MapROverview/c_overview_intro.html*(website)* [05/2019]

[2] Cloudera Producent website https://www.cloudera.com/documentation/enterprise/5-13-x/topics/introduction.html *(website)* [05/2019]

[3] Hortonworks Producent website https://hortonworks.com/products/data-platforms/hdp/ *(website)* [05/2019]

[4] Microsoft Azure Producent website https://docs.microsoft.com/pl-pl/azure/hdinsight/hadoop/apache-hadoop-introduction *(website)* [05/2019]

M. Siudziński, Hadoop article https://itwiz.pl/hadoop-czyli-przetwarzanie-rozproszone-open-source/*(website)*[05/2019]

[5] R.Wasiluk, P.Muryjas: The assessment of usefulness modern IT tools of data analysis Big Data, Institute of Computer Science, Lublin University of Technology, 2017