

A Comparative Study between Big Data Solutions HortonWorks, Cloudera and Microsoft Azure HD Insight

Dange M. S., Sanusi Sulaiman

*Federal University Gusau
Teachers Service Board*

Date of Submission: 21-11-2022

Date of Acceptance: 30-11-2022

ABSTRACT—The rapid growth of digital data over the past decade due to the proliferation of applications and smart devices has led to the introduction of the concept of big data. Several publishers, including HortonWorks, Cloudera, MapR, IBM Infosphere, BigInsights, Pivotal HD, and Microsoft HDInsight, had launched commercialised Hadoop distributions to manage vast quantities of data in order to extract value from this mass of data. A comparative study between HortonWorks, Cloudera, and Microsoft HDInsight is being conducted to observe the strengths and weaknesses of each distribution, allowing users to select the distribution that best meets their needs.

KEYWORDS—big data solutions, HortonWorks, Cloudera, Microsoft Azure HDInsight, Hadoop

I. INTRODUCTION

Being In this age of information explosion, conventional storage architecture and technology are incapable of storing or processing massive amounts of data. Doug Laney introduced the 3Vs to define big data, which are volume, variability, and velocity[1]. Because the characteristics of big data makes it difficult to collect, store, manage, process, or analyse, the Apache Hadoop framework was created to address the scaling issue.[2]. Companies, such as HortonWorks, Cloudera, MapR, IBM Infosphere, BigInsights, Pivotal HD, and Microsoft HDInsight, have launched their own Apache Hadoop-based data solutions platforms with application features, utilities, and capabilities to capitalise on this rapidly expanding market. The features, performance, capabilities, integration, and service and support offered by these Hadoop distributions vary. Variations in ready-to-use distributions enable users to select solutions based on their specific needs. In this paper, we conducted a comparison of HortonWorks, Cloudera, and

Microsoft Azure HDInsight, three popular distributions..

II. HISTORY AND EVOLUTION OF THE DISTRIBUTIONS/SERVICES

A. HORTONWORKS

HortonWorks was established in 2011 by Yahoo and Benchmark capital. The HortonWorks distribution comprises open-source components licensed by Apache; it is a simplified adoption of the Apache Hadoop platform [3]. This platform includes the Hortonworks Data Platform (HDP) used for storing, processing, and analysing big data. Horton work's license is not commercialized, their support and training are rather commercialized. In October 2018, Cloudera announced the acquisition of HortonWorks, and the merger was completed the next year January [4].

B. CLOUDERA

The coalition of engineers from Google, Yahoo, Facebook, and Sleepycat Software became the genesis of Cloudera Inc. in the year 2008. Over the years Cloudera partnered with several stakeholders in the computing and Information Technology world, such partnerships include Oracle Corporation, Dell, SAS institute etc. and acquired several others [5]. In 2019, Cloudera took a bold swing in big data management and processing by a merger with HortonWorks and making all its software open source [4].

C. MICROSOFT AZURE HDINSIGHT

Microsoft introduced Windows Azure (former name of Microsoft Azure) in October 2008 which is the first model that delivers software through the web which known as Software as a Service (SaaS). Windows Azure was launched commercially in 2010 by providing Platform as a Services (PaaS) services. After the rise of open-source software (OSS) and virtual machine

solutions, Azure had expanded the business by building a new model based on Infrastructure as a Services (IaaS) and rename it as Microsoft Azure in April 2014 because the cloud computing services had gone beyond Windows from continued improvement and expansion [6]. Microsoft HDInsight came in 2013 as cloud distribution of Hadoop service under Azure [7].

III. HIGHLIGHTS OF THE DISTRIBUTIONS/SERVICES AND THEIR COMPONENTS

Spike in the volume of data generated results in urge needs of big data storage and management solution. The existence of various big data solutions makes the industry relatively competitive, to attract and fulfil needs from different end-users, each solution possesses unique services and components.

A. HORTONWORKS

HortonWorks Data Platform which is also known as HDP is an open-source Hadoop based data platform that owns zero proprietary applications, purely relied on the deployment of the Apache Hadoop project. Components of Apache Hadoop in HDP can be divided into 3 layers which are core Hadoop, essential Hadoop and supporting components. Core Hadoop is the set of basic components of Apache which includes Hadoop Distributed File System (HDFS), YARN and MapReduce. Essential Hadoop is a set of components designed to ease the process of working with core Hadoop and supporting components is a set of components that allow users to monitor Hadoop installation and connect Hadoop to a larger computing environment [8].

Unlike other distributions, HortonWorks deployed Apache Ambari as an administration console to manage Hadoop clusters, Apache Stinger for query handling, Apache Solr for data searches, HCatalog that enable connection of Hadoop to other enterprise applications. Furthermore, HDP deployed Apache Druid components which is a data store designed for online analytical processing queries on event data. It allows user to run business intelligence application with minimal latency and enable quick reflection of streaming data [9].

HDP is a good choice for users that is new to Hadoop or small enterprise where its open-source nature lowers the cost where HortonWorks only charge on support and training. For future enhancement, HDP allows migration to Cloudera

Data Platform or integration with cloud services providers such as Microsoft and Amazon.



Figure 1 is showing the framework for HortonWorks distribution.

B. CLUDERA

Like other distributions, Cloudera provides Private Cloud and Public Cloud solutions for users to choose from, on top of that, Cloudera had also introduced Hybrid Cloud which is a hybrid data cloud platform that allows the management of data anywhere without concern about performance, scalability, or security. Hybrid Cloud services possess strength from both CDP Private Cloud and CDP Public Cloud which is more powerful than traditional cloud services [10].

Cloudera provides different types of services for different needs from users were including Data Engineering, Data Flow, Data Hub, Data Warehouse, Machine Learning and Operational Database. Different hourly rate is applied to services respectively where they range from USD 0.04 to USD 0.30 every hour for one concurrent user. Cloudera also provides additional tools such as data science workbench, enterprise data hub, GPU acceleration HDP enterprise plus for high-level complex needs from users [11].

With Cloudera, users can run real-time analytics with the deployment of the Apache Kudu project and support from running analytical queries in SQL with the Apache Impala project. Despite Apache Hadoop project deployment, Cloudera launched a proprietary administration console, Cloudera Manager, a centralized interface that ease cluster deployment and management with zero downtime for maintenance allows users to manage aggregates logs across all services and hosts. Built-in health checks, backup and disaster recovery features decrease user workloads and lower the risk of running critical work [12].

As one of the most deployed implementations of Hadoop in use distributions, Cloudera is more suitable for an enterprise that wants to set up their Data Hub with intention of applying analysis action on it.



Figure 2 is showing the framework for Cloudera distribution.

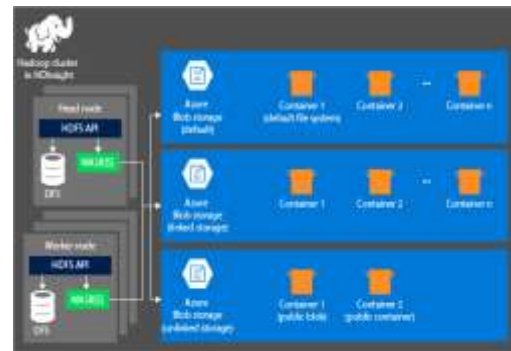


Figure 3 is showing abstract view of the HDInsight architecture of Azure Storage.

C. MICROSOFT AZURE HDINSIGHT

Microsoft Azure provides a wide range of services that can be divided into fourteen categories which included Compute, Networking, Storage, Web + Mobile, Containers, Databases, Data + Analytics, AI + Cognitive Services, Internet of Things, Enterprise Integration, Security + Identity, Developer Tools, Monitoring + Management and Microsoft Azure Stack. As part of the Azure family, integration or cooperation between services and Azure HDInsight had massively increased the ways of handling data [13].

Besides that, for better deployment and management of clusters, Microsoft provides Azure Marketplace with over 30 widely used Hadoop and Spark-based applications to fulfil various user needs. Applications include monitoring tools for HDInsight Spark, HDInsight Kafka, HDInsight Storm, HDInsight Hadoop and so on, tools to build web services on HDInsight HBase, distributed machine learning platform with linear scalability, performance accelerator, data quality tool, backup, recovery, migration tool. With a few clicks and a few minutes, an application can be easily deployed and run in HDInsight.

Through collaboration with R Server, Microsoft claimed that more accurate models can be trained in a shorter time by using Microsoft R Server. R Server for HDInsight allows user to input thousands of more data for machine learning model training and the speed of training is 50 times faster than open-source R with utilizing transparent parallelization and multi-threaded math libraries [14].

As a distribution that launched commercially 11 years ago, Microsoft HDInsight is a mature big data solution that provides a variety of functions, but it is not friendly to non-existing Microsoft Services subscribers due to its pricing and lack of integration with other services such as Amazon AWS, Google Cloud Storage or Cloudera.

VI. COMPARISONS AMONG THE DISTRIBUTIONS/SERVICES AND THEIR COMPONENTS (ANALYSIS/RESULTS)

A. FEATURE

For observing differences between distributions, a comparison between two mainstream popular Hadoop based big data platforms HortonWorks and Cloudera with cloud distribution of Hadoop based big data distribution introduced by Microsoft, Azure HDInsight. We had compared them from 6 sectors which are basic, security, data access, data science, data movement and cloud storage. All the information stated in the table below is based on [15],[16] and [17].

Basic:

- **Latest Edition:** Latest version of distributions introduced by publishers.
- **Base:** Basis of distributions
- **Administration Console:** Components set up administration console
- **Programming Language Supported:** Default programming language supported by distribution
- **Others Language Supported:** Programming language supported with additional components
- **Hadoop Specific Language Supported:** Hadoop specific programming language supported
- **Hadoop UI:** Components set up Hadoop UI
- **Scripting Platform:** Platform for scripting

Security:

- **Perimeter Security:** Perimeter used to secure data
- **Authentication:** Access Management Control
- **Authorization:** Authorization Management Control

- **Encryption for Data at Rest:** Way of encryption applied to data while at rest
- **Encryption for Data in Transit:** Way of encryption for data during transit.

Data Access

- **Metadata Management:** Components used to manage metadata
- **Metadata Storage:** Storage used to store metadata
- **Integration with Non-relational Database:** Components used to integrate with non-relational database
- **Integration with RDBMS:** Components used to integrate with relational database
- **Dealing with Streaming Log:** Components used to deal with streaming log
- **Parallel Query Execution:** Components used to execute parallel query
- **Real-Time Data Ingestion:** Components used to ingest with real-time data

Data Science:

- **Data Analysis:** Components used for analysing data

- **Data Visualization:** Components used to create visuals
- **Machine Learning:** Components used to run Machine Learning algorithm
- **Machine Learning Library: Hadoop based Machine Learning Library used**

Data Movement:

- **Workflow Manager UI:** Interface used for workflow management
- **Workflow Engine:** Components used to run workflows
- **Data Movement Management In:** Components used to manage data movement into distributions
- **Data Movement Management Out:** Components used to manage data movement out of distributions
- Cloud Storage
- **Cloud Storage Services Integration:** Cloud storage services available to integrate with distributions
- **Access Cloud Data:** Components used to access cloud data from distributions
- **Copy Cloud Data:** Components used to copy cloud data

Basic			
	HortonWorks	Cloudera	Microsoft HDInsight
Latest Edition	HortonWorks HDP 3.0.0	Cloudera Enterprise 6.0, Cloudera Express 6.3.2	HDInsight 4.0
Base	HDFS, YARN, MapReduce (MR2)	HDFS, YARN, MapReduce	HDFS, YARN, MapReduce
Administration Console	Ambari	Cloudera Manager	Azure Monitor logs, Ambari
Programming Language supported	Python, Java	C, C++, Go, Java, Python, PHP	Java, Python, .NET, Go
Other Languages Supported	Scala	Jython, Scala	Clojure, Jython (Python for Java), Scala
Hadoop specific language supported	Pig, Hive, Spark	Apache Groovy	Pig Latin for Pig jobs, HiveQL for Hive jobs and Spark SQL
Hadoop UI	Hue	Hue	Hue
Scripting Platform	Pig	Pig	Azure Portal, Azure PowerShell, Azure Classic CLI, HDInsight .NET SDK, Azure Resource Manager Template

Security			
	HortonWorks	Cloudera	Microsoft HDInsight
Perimeter Security	Apache Knox	Apache Knox	VNET, NSG
Authentication	PAM authentication, Hadoop Group Mapping Service Provider, Kerberos	Lightweight Directory Access Protocol (LDAP), MIT Kerberos, Red Hat Identity Management, Microsoft Active Directory	Azure Active Directory Domain Services
Authorization	Apache Ranger, Access Control Lists (ACL) on HDFS	Apache Sentry	Apache Ranger
Encryption for Data at Rest	Transparent Data Encryption (TDE), Key Management Store (KMS) by Ranger	HDFS Transparent Encryption, Navigator Encrypt, Navigator Key Trustee	Transport Layer Security (TLS), Internet Protocol Security (IPSec)
Encryption for Data in Transit	RPC encryption, Data Transfer Protocol, HTTPS encryption, SASL (JDBC)	HDFS Transparent Encryption, SASL (RPC, JDBC), TLS (Avro RPC, JDBC, ODBC)	Transport Layer Security (TLS), Internet Protocol Security (IPSec)

Data Access			
	HortonWorks	Cloudera	Microsoft HDInsight
Metadata Management	HCatalog, WebHCat	Impala	HCatalog, Azure Data Catalog
Metadata Storage	Hive	Hive	Hive, Ambari, Oozie
Non-Relational Database	HBase, Accumulo, Phoenix	HBase	HBase
Integration with RDBMS	Sqoop	Sqoop, Kudu	Sqoop
Dealing with streaming data	Flume	Spark	Kafka, Spark, Storm, HBase
Parallel Query Execution	Hive	Impala	Hive
Real-Time Data Ingestion	-	NiFi	Kafka

Data Science			
	HortonWorks	Cloudera	Microsoft HDInsight
Data Analysis	Spark, Zeppelin, Data Analytics Studio	Cloudera Data Science Workbench (CDSW)	HiveQL, Microsoft Power BI
Data Visualization	Zeppelin	Cloudera Data Platform (CDP) Data Visualization	Microsoft Power BI, Zeppelin
Machine Learning	Spark	Cloudera Machine Learning (CML) workspaces	HDInsight Spark*,
Machine Learning Library	Mahout, Spark MLlib	Spark MLlib	MMLSpark**

* an Azure-hosted offering of Apache Spark

** Microsoft Machine Learning Library for Apache Spark

Data Movement			
	HortonWorks	Cloudera	Microsoft HDInsight
Workflow Manager UI	Ambari	Cloudera Manager	Azure web console
Workflow Engine	Oozie	Oozie	Oozie, Blaze, Spark, Hive
Data Movement into Hadoop	Data Movement Integration Suite (DMI Suite): Falcon, Oozie, Sqoop, Flume	CDP Data Hub, Apache NiFi	Power Query
Data Movement out of Hadoop	Data Movement Integration Suite (DMI Suite): Falcon, Oozie, Sqoop, Flume	CDP Data Hub, Apache NiFi	Power Query

Cloud			
	HortonWorks	Cloudera	Microsoft HDInsight
Cloud Storage Services Integration	Amazon S3, Azure Data Lake Store (ADLS), Windows Azure Storage (WASB)	Amazon S3, Azure Data Lake Store (ADLS), Google Cloud Storage (GCS)	-
Access Cloud Data	Hive, Spark	Azure Blob File System Driver (ABFS) connector, google-access-key.json,	-
Copy Cloud Data	DistCp, FS Shell commands	DistCp	-

A. Strength and Weakness

Based on rating and reviews on [18] and [19], Cloudera has the highest rating out of these three big solutions which is 8.4 out of 10, followed by Microsoft HDInsight with a rating of 8.3 and HortonWorks with a rating of 7.0.

In the reviews from end-users, we can observe that the reason why Cloudera stands out from these three big data solutions is because of its ability to process a large batch of data, scalability and user-friendly administration console, Cloudera Manager that allows users to manage, deploy, and monitor aggregates logs easily.

Even though users are satisfied with Cloudera, some suggestion is given on log visibility and costs. Other than that, after-acquired HortonWorks, Cloudera had deployed some of the features from HDP into the latest version of Cloudera Data Platform where some of the users found that the components are poorly integrated with existing components in CDP and making the management process more complex than it used to be.

According to the feedbacks, the 30 types of Hadoop and Spark applications in Azure Marketplace that ease of deployment process and easy transformation of data are aspects that attract user attention. As part of the Microsoft family, Microsoft Azure HDInsight possess an advantage for existing Microsoft services users where users will be familiar with graphic interfaces design that is similar to other Microsoft products.

The major weakness of HDInsight is its poor performance with large data set, users had faced corruption issues while loading large volumes of data. Moreover, we observed that some users are satisfied with the pricing of Azure HDInsight while some are not, this is due to some Azure services being free for existing Windows users. In terms of fully releasing the potential of Azure HDInsight, subscription to other services or products in Azure and Microsoft will be essential which will be costly especially for those non-Microsoft users.

Even though HortonWorks Data Platform has the lowest rating, consumers are satisfied with the overall performance of HDP from the deployment of Apache Ranger and Kerberos for

security to easy integration with other tools and components. Furthermore, as a contributor to Apache Software Foundation, HortonWorks followed tightly to the pace of Apache Hadoop versions and projects.

The major factors that hold back users in HDP are the complex installation process and high resources requirements which had increased the difficulty of setting up the data platform. Besides that, poor graphic interface design decreases users' satisfaction, simple interface design of Apache Ambari web UI does not provide many features for users compared to other distribution, more resources are needed to observe and monitor the health of clusters.

To summarize, Cloudera has the highest rating followed tightly by Microsoft and HortonWorks. Each data distribution possesses its strengths and weakness allows users to choose according to their specifications and requirements. Since Cloudera had acquired HortonWorks, the future performance and enhancement are making the industry look forward, so as continuous enhancement and expansion of Microsoft Azure HD Insight.

V. DISCUSSION AND RECOMMENDATION (APPLICATION DOMAIN)

A. DISCUSSION

With the involvement of applications and smart devices in every sector of human daily life, the volume of digital is increasing drastically every second so is the need for big data distribution. In the below section, we will be discussed about real-life use cases of HortonWorks, Cloudera and Microsoft Azure HDInsight in healthcare, education, and public sector.

HEALTHCARE

Cloudera collaborates with IBM to fill the urgent demands in the healthcare sector where they provide ways to connect data with predicting and driving agility to adopt new solutions disease tracking areas which have been helpful to detect propensities and patterns for diseases and diagnosis, especially during a pandemic. Through analysis, the data provides the ability to keep on track of the spreading of disease, identification of clusters, healthcare resources monitoring and treatment results. Besides this, a free access platform contained peer-reviewed papers and licensed databases available to assist qualified scientists and academics to keep on track with the latest information. The platform also provides AI-

based virtual medical decision support to support clinicians on drug and disease information [20].

In partnership with Microsoft, a team from Virginia Tech had developed an on-demand, cloud computing model on Microsoft Azure HDInsight that allows faster and cheaper DNA sequencing analysis that has the potential to lead to a wide range of breakthroughs in medical and pharmaceutical fields. This model provides significant cost savings compared to the regular standard way of genomics analysis and allows easy sharing, access virtually anywhere with any devices which ease the large-scale collaborative research [21].

EDUCATION

Carnegie Mellon University built a centralized, integrated structure with dynamic, custom reporting tools and access to real-time data and events with Azure SQL Database, Azure HDInsight, and Power BI. The intention of this model is to look for ways to improve energy and operational efficiency in buildings. Through this model, a researcher from Carnegie Mellon University had stated that will deployment of the model, plug load energy consumption was reduced by 30%. The research team is looking forward to applying this energy savings model on a citywide level [22].

As the largest university-affiliated research centre in United State, Johns Hopkins University applies its research to areas such as air and missile defence, space science and medicine. Through these research works, highly sensitive data is collected and a secure, robust big data solution platform is needed. Out of all the Hadoop based distribution, Johns Hopkins University had chosen HortonWorks to collect and dispense information in a secure manner. HortonWorks Data Flow (HDF) works along with Apache NiFi to ingest information from different sources and compile it into a single place for future analysis. HortonWorks had assisted on the speed of reporting where the time consumed reduced from 2 days to 5 minutes [23].

PUBLIC SECTOR

As part of local administration organization of Istanbul, Istanbul Metropolitan Municipality (IMM) responsible for entire provincial territory of 5343 km² and over 16 million residents. A great amount of data is generated and collected everyday by the organization moreover the organization had faced problems in sharing and accessing of data among different department or subsidiaries. Cloudera

along with Kafka, Spark, CKAN data management platform, Tableau and Vertica Analytics help to solve the issues. After implementation of Cloudera, IMM managed to modernize traffic management and planning, run pandemic and crisis management more efficiently [24].

City of Barcelona work with Microsoft and created a Big Data BI solution that runs on hybrid cloud based on Azure, HDInsight, and SQL Server in year 2013. Through this platform, user able to view real time business intelligence based on public data sources. The intention of this project is to increase quality of life for citizens and assist to discover investment opportunities in neighbourhoods [25].

B. RECOMMENDATION

Based on the knowledge and understanding that we made through this research process a few recommendations can be given on these big data solutions. We hoping to see an improvement in Microsoft Azure pricing where lower the price allows new users to join the Microsoft ecosystem and more details or user-friendly documentation regarding integration between Microsoft HDInsight with other Microsoft products.

Improvement in cluster management user interface will be a strength for HortonWorks, where Apache Ambari is too simple and not able to support users' needs. Since HortonWorks merged with Cloudera, a new distribution that possesses strengths from both distributions will be a piece of exciting news to the industry.

VI. CONCLUSION

To conclude, HortonWorks is acquired by Cloudera, but it still maintains its open-source nature. Microsoft Azure HDInsight provides cloud-based services with working side by side with other Microsoft products while Cloudera provides hybrid clouds that gained merit from both private and public cloud. All in all, users should understand their own needs before deploying a big data solution since every distribution has its own strengths and constraints, there is no one fit distribution available.

VII. REFERENCES

- [1]. Patgiri, R., & Ahmed, A. (2016). Big data: The V's of the game changer paradigm. 2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS). <https://doi.org/10.1109/hpcc-smartcity-dss.2016.0014>
- [2]. FaridhRonianto, M., Asror, I., Prabowo, S., & Nugraha, F. A. (2019). Performance evaluation SQL-on-hadoop: A case study of hortonworks and Cloudera. *Journal of Physics: Conference Series*, 1192, 012016. <https://doi.org/10.1088/1742-6596/1192/1/012016>
- [3]. Hortonworkssuccessstory. Sucess Story. (n.d.). Retrieved December 16, 2021, from <https://successtory.com/companies/hortonworks>
- [4]. Novet, J. (2019, January 3). Cloudera Is Gearing up to Take on Amazon, but Investors Are Skeptical. Retrieved December 16, 2021, from <https://www.cnbc.com/2018/12/28/as-cloudera-closes-hortonworks-merger-it-takes-aim-at-amazon.html>.
- [5]. About | cloudera. (n.d.). Retrieved December 16, 2021, from <https://www.cloudera.com/about.html>
- [6]. (n.d.). A Look Back At Ten Years Of Microsoft Azure. Retrieved from <https://www.forbes.com/sites/janakirammsv/2020/02/03/a-look-back-at-ten-years-of-microsoft-azure/?sh=6b33615f4929>.
- [7]. Meisner, J. (2013, October 28). Announcing Windows azure hdinsight: Where big data meets the cloud. *The Official Microsoft Blog*. Retrieved December 16, 2021, from <https://blogs.microsoft.com/blog/2013/10/28/announcing-windows-azure-hdinsight-where-big-data-meets-the-cloud/>
- [8]. 1.HDP components. 1. HDP Components - Hortonworks Data Platform. (n.d.). Retrieved December 14, 2021, from https://docs.cloudera.com/HDPDocuments/HDP2/HDP-2.3.0-Win/bk_HDP_Install_Win/content/iug_hdp_components.html
- [9]. Chapter 1. introduction - hortonworks data platform. (n.d.). Retrieved December 14, 2021, from https://docs.cloudera.com/HDPDocuments/HDPS/HDPS-3.0.0/bk_solr-search-installation/content/ch_hdp-search.html
- [10]. CDP Hybrid Cloud | Cloudera. (n.d.). Retrieved December 14, 2021, from <https://www.cloudera.com/products/cloudera-data-platform/cdp-hybrid-cloud.html>

- [11]. Cloudera pricing | Cloudera. (n.d.). Retrieved December 14, 2021, from <https://www.cloudera.com/products/pricing.html>
- [12]. Cloudera manager: Hadoop Administration Tool. (n.d.). Retrieved December 14, 2021, from <https://www.cloudera.com/products/product-components/cloudera-manager.html>
- [13]. What is Azure-Microsoft Cloud Services: Microsoft Azure. What is Azure-Microsoft Cloud Services | Microsoft Azure. (n.d.). Retrieved December 14, 2021, from <https://azure.microsoft.com/en-us/overview/what-is-azure/>
- [14]. Alicia Li Principal Program Manager. (n.d.). Top 8 reasons to choose Azure hdinsight: Blog Y Actualizaciones de Azure: Microsoft Azure. Blog y actualizaciones de Azure | Microsoft Azure. Retrieved December 14, 2021, from <https://azure.microsoft.com/es-es/blog/top-8-reasons-to-choose-azure-hdinsight/>
- [15]. HDP documentation. HDP. (n.d.). Retrieved December 14, 2021, from <https://docs.cloudera.com/HDPDocuments/HDP3/HDP-3.1.5/index.html>
- [16]. Cloudera Documentation. Cloudera Product Documentation. (n.d.). Retrieved December 14, 2021, from <https://docs.cloudera.com/>
- [17]. JasonWHowell. (n.d.). Azure hdinsight Documentation. Azure HDInsight | Microsoft Docs. Retrieved December 14, 2021, from <https://docs.microsoft.com/en-us/azure/hdinsight/>
- [18]. Trustradius: Software reviews, software comparisons and more. (n.d.). Retrieved December 14, 2021, from <https://www.trustradius.com/>
- [19]. Gartner, I. (n.d.). Enterprise software & service reviews: Gartner peer insights. Gartner. Retrieved December 14, 2021, from <https://www.gartner.com/reviews/home>
- [20]. IBM and Cloudera Partnership. IBM. (n.d.). Retrieved December 14, 2021, from <https://www.ibm.com/analytics/partners/cloudera>
- [21]. University Transforms Life Sciences ... - virginia tech. (n.d.). Retrieved December 14, 2021, from https://seec.cs.vt.edu/files/VirginiaTech_DataInsights_Final_1.pdf
- [22]. University improves operational efficiency, cuts energy consumption by 30 percent with Bi Solution. Microsoft Customers Stories. (n.d.). Retrieved December 14, 2021, from <https://customers.microsoft.com/en-us/story/731538-university-improves-operational-efficiency-cuts-energy>
- [23]. Johns Hopkins University | Customer Success | Cloudera. (n.d.). Retrieved December 14, 2021, from <https://www.cloudera.com/about/customers/johns-hopkins-university.html>
- [24]. Istanbul Metropolitan Municipality | Customer Success ... (n.d.). Retrieved December 16, 2021, from <https://www.cloudera.com/about/customers/istanbul-metropolitan-municipality.html>
- [25]. Team, S. Q. L. S. (2013, October 30). Revolutionizing City Planning in the 21st century with windows azure, HDInsight, and SQL server. Microsoft SQL Server Blog. Retrieved December 16, 2021, from <https://cloudblogs.microsoft.com/sqlserver/2013/10/30/revolutionizing-city-planning-in-the-21st-century-with-windows-azure-hdinsight-and-sql-server/>