Comparative study of Big Data Hadoop solutions: Cloudera, Azure, and AWS

Siang Lee Khaw
School of Computer Sciences
University Science Malaysia
Penang, Malaysia
sianglee.khaw@student.usm.my

***Abstract*— In recent years, there are a tremendous amount of data generated by new technologies like social media, research articles, daily online shopping transactions, etc. These data reflect the concept of Big Data due to its velocity, volume, and variety [1]. New technical architectures and analytics tools were needed to reveal the business value from these data. One of the Big Data technology that supports the processing and storage of extremely large data sets in a distributed computing environment is Hadoop. In this paper, I will perform a comparative study on the Hadoop distributer that manages Big Data namely Cloudera, Windows Azure HDInsight, and Amazon Elastic MapReduce.**

Keywords—Hadoop distribution, Cloudera, Windows Azure HDInsight, Amazon Elastic MapReduce

I. INTRODUCTION

New technologies like social media, daily transactions, scientific data logs have produced massive amounts of data that need to be captured, collect, store and analyze to find useful insight. Several vendors offer ready-to-use distributions to deal with Big Data [3], namely Cloudera, Windows Azure HDInsight, Amazon Elastic MapReduce, etc. Each distribution has its approach for a Big Data system and the choice of selecting the distribution will be based on various parameters depending on several requirements. I will focus on an evaluation by Forrester Wave [7] on the same Hadoop distributions on which they used 35 evaluation criteria grouped into three high-level buckets: Current Offering, Strategy, and Market presence [2].

II. HISTORY AND EVOLUTION

A. Cloudera Distribution

Cloudera was founded by Hadoop experts from Facebook, Google, Oracle, and Yahoo in 2008. This distribution is largely based on the components of Apache Hadoop and it is complemented by essentially house components for cluster management. The aim of Cloudera's business model is not only to sell Licenses but to sell support and training as well. Cloudera offers a fully open-source version of its platform (Apache 2.0 license). [3]

On 3 January 2019, Cloudera announced the completion of its merger with Hortonworks, Inc. Cloudera will deliver the first enterprise data cloud, running in any cloud from the Edge to AI, on a 100% open-source data platform. An enterprise data cloud supports both hybrid and multi-cloud deployments, providing enterprises with the flexibility to perform machine learning and analytics with their data, with no lock-in. [10]

B. Windows Azure HDInsight

In 2014, Microsoft partnership with Hadoop software developer and distributor Hortonworks that deploys Hortonworks Hadoop on Windows Azure. Azure HDInsight is a managed, full-spectrum, open-source analytics service in the cloud for enterprises. Users can use open-source frameworks such as Hadoop, Apache Spark, Apache Hive, LLAP, Apache Kafka, Apache Storm, R, and more. [11]

C. Amazon Elastic MapReduce

Amazon EMR is an Amazon Web Services (AWS) tool for big data processing and analysis. Amazon EMR is a managed cluster platform that simplifies running Hadoop frameworks, on AWS to process and analyze vast amounts of data. Users can use Amazon EMR to transform and move large amounts of data into and out of other AWS data stores and databases, such as Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB. [12]

III. HIGHLIGHTS OF DISTRIBUTIONS

A. Cloudera Distribution

Cloudera CDH provides the following products and tools.

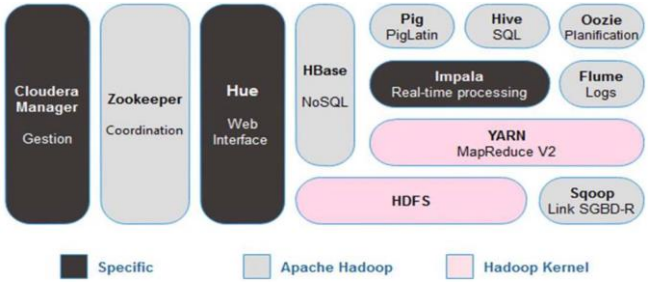

Figure 1. Cloudera Distribution [3]

- CDH — The Cloudera distribution of Apache Hadoop provides security and integration with various hardware and software solutions.
- Apache Impala — It is a massively parallel processing SQL engine for interactive analytics and business intelligence. It queries Hadoop data files from different sources like MapReduce jobs or Hive tables.
- Cloudera Search — It provides real-time access to data stored in Hadoop and HBase.
- Cloudera Manager — It is an application used to deploy, manage, monitor, and diagnose issues with CDH deployments. It includes the Cloudera Manager API, which is used to obtain cluster health information and metrics, as well as to configure Cloudera Manager.
- Cloudera Navigator — It is a data management and security tool for the CDH platform. It enables administrators, data managers, and analysts to explore

the data in Hadoop and simplifies the storage and management of encryption keys.

- Hue - It is a web-based interactive query editor in the Hadoop stack that helps in visualizing and sharing data. [4]

In 2019, Cloudera start to offer Cloudera Data Platform (CDP). Initially delivered as a public cloud service.

Figure 2. Cloudera Data Platform [13]

- Data Warehouse and Machine Learning services as well as a Data Hub service for building custom business applications powered by their new Cloudera Runtime open source distribution.
- A unified control plane to manage infrastructure, data, and analytic workloads across hybrid and multi-cloud environments
- Consistent data security, governance, and control that safeguards data privacy, regulatory compliance, and prevents cybersecurity threats across environments
- 100 percent open source, supporting your objectives to avoid vendor lock-in and accelerate enterprise innovation
- A clear path for extending your existing CDH and HDP investment to the cloud. [13]

B. Windows Azure HDInsight

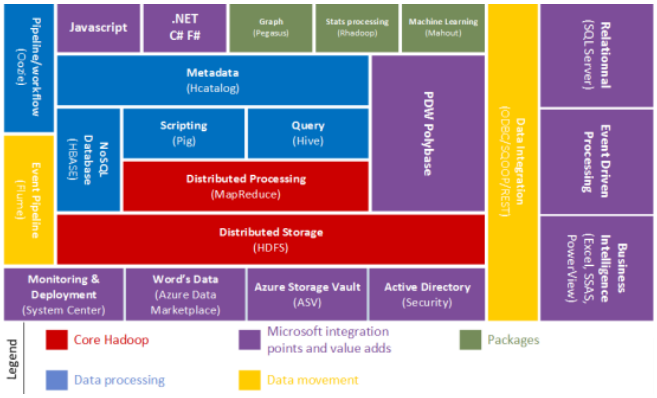

Figure 3. HDInsight Ecosystem [14]

Microsoft provides the highest availability guarantee in the industry with a 99.9% service level agreement, ensuring continuity and protection against catastrophic events. Azure also offers 24/7 enterprise support and cluster monitoring. [8]

HDInsight clusters are configured to store data directly in Azure Blob storage, which provides low latency and increased elasticity in performance and cost choices.

Below are capability lists of Azure HDInsight:

- Cloud-native - Azure HDInsight enables the user to create optimized clusters for Hadoop, Spark, Interactive Query (LLAP), Kafka, Storm, HBase, and ML Services on Azure.
- Low-cost and scalable - HDInsight enables the user to scale workloads up or down. Users can reduce costs by creating clusters on demand and paying only for what they use.
- Secure and compliant - HDInsight also meets the most popular industry and government compliance standards.
- Monitoring - Azure Monitor logs to provide a single interface with which you can monitor all your clusters.
- Global availability – Available in more than 50 regions around the world.
- Productivity – User can use their preferred development environments. HDInsight supports Visual Studio, VSCode, Eclipse, and IntelliJ for Scala, Python, R, Java, and .NET support. Data scientists can also collaborate using popular notebooks such as Jupyter and Zeppelin.
- Extensibility - HDInsight enables seamless integration with the most popular big data solutions with one-click deployment. [11]

C. Amazon Elastic MapReduce

Amazon EMR provides a managed Hadoop framework as a web service. The data cross dynamically with scalable Amazon EC2 instances. Amazon EMR supports Amazon S3 (EMRFS), the Hadoop Distributed File System (HDFS), and Amazon DynamoDB as the data stores. Amazon EMR can run popular frameworks such as Apache Spark, HBase, Presto, HUE, Flink, and more. [15]

Amazon EMR is available across 12 regions worldwide. Amazon EMR automatically configures Amazon EC2 firewall settings that control network access to instances, and customers can launch clusters in a virtual private cloud. Amazon EMR provides a development environment based on Jupyter Notebook, that helps analysts, developers, and data scientists prepare and visualize data. Developers can build applications, collaborate with peers and do interactive analysis using EMR clusters.

Below are the advantages of using Amazon EMR: [17]

- Increased speed and agility – Users can dynamically create, add or remove the existing cluster. Organizations save cost and time it takes to allocate resources for experimentation and development.
- Reduced administrative complexity – Hadoop is deploying as a service in AWS, all the configuration and maintenance is handled by the AWS team.
- Integration with other AWS services - Hadoop environment with other services such as Amazon S3, Amazon Kinesis, Amazon Redshift, and Amazon DynamoDB to enable data movement, workflows, and analytics across the many diverse services on the AWS platform.
- Pay for clusters as needed – Many Hadoop jobs may only occur a few times per year. Amazon EMR can spin up the

working cluster easily, save the result then shut down to save infrastructure cost.

- Availability and disaster recovery – Amazon EMR can be launched in any number of available regions. Users can easily launch a new cluster in other regions when a problem happens to a specific region.
- Flexible capacity – Amazon EMR support Auto scaling that dynamically scales out and scales in nodes. [17]

IV. COMPARISONS AMONG THE DISTRIBUTORS

A. *Features comparison*

Cloudera CDP is relatively new, so the below comparison is based on Cloudera CDH.

	Cloudera	**HDInsight**	**Amazon EMR**
Deployment Model	On-Premises Hosted	Cloud	Cloud
Autonomous	No	Yes	Yes
Out-of-the-box Data Processing Engines	Installation required	MapReduce, Hive, Pig, Spark, HBase, Storm	MapReduce, Hive, Pig, HBase, Cascading, Impala, Spark, Presto
Data Store	On-Premises	Azure	AWS
Setup	Manual	Automatic	Automatic
Management	Support and 3rd Party Consulting	No Big Data-specific Support	No Big Data-specific Support
Economic Structure	Software License and Support, Infrastructure Purchase and Personnel	Elastic compute pricing	Elastic compute pricing
Scalability	Fixed Cluster	Manual scaling, elastic, on-demand, no graceful downscaling	Manual scaling, elastic, on-demand
Costing	High	Low	Low

Figure 4. Big data solutions comparisons

B. *Forrester Wave report*

Forrester Wave [7] evaluate Hadoop distributions on which they used 35 evaluation criteria grouped into three high-level buckets: Current Offering, Strategy, and Market presence. Forrester's evaluation of Hadoop's distributions of Big Data revealed that in 2014, the clear leader is Amazon EMR followed by Cloudera CDH and Azure HDInsight follow as strong performers Referring to Figure 5 [7].

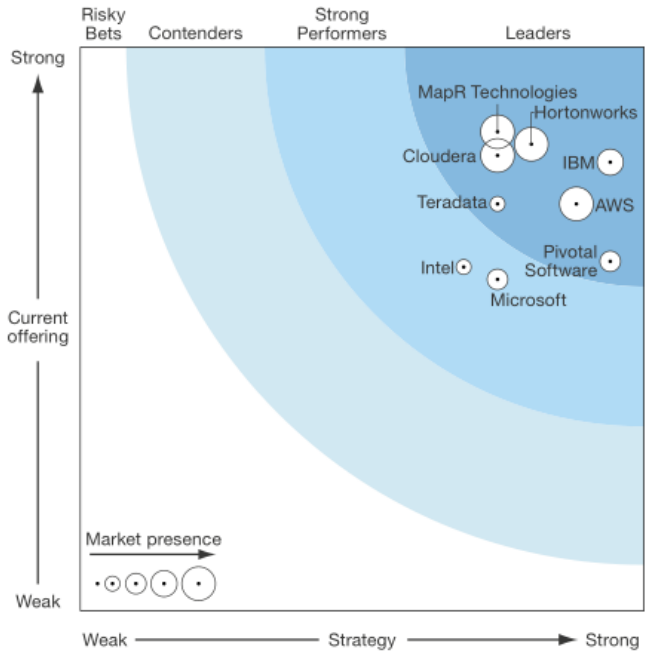

Figure 5. Forrester Wave Big Data solution Q1 '14 [7]

Below is the Forrester Wave report for Q1 2019, it clearly shows that Amazon EMR is the top leader with all three criteria on the strong side. Followed by Windows Azure HDInsight. Cloudera is also placed at the edge of the leader chart but the market presence is still low because Cloudera CDP is still relatively new and first launching in September 2019.

Figure 6. Forrester Wave Cloud Hadoop Q1 2019 [16]

V. DISCUSSION

Cloudera offers the best technical expertise which helps customers to harness the most from the data. Telecommunication service providers will benefit most from Cloudera Data scientists. The service provider can set up an

on-premise cluster that can perform Churn analytics on demand. Service providers can gain a real-time view of the network to allocate resources more efficiently. Use of Cloudera Machine Learning to detect fraud in real-time and unlock new revenue by exploring IoT & connected ecosystems. Cloudera DataFlow is a scalable, real-time streaming analytics platform that can easily ingest, process, analyze, monetize IoT uses case. [18]

I would recommend Windows Azure HDInsight for the public sector and Government sector. Microsoft Azure Government delivers a cloud platform built upon the foundational principles of security, privacy and control, compliance, and transparency. Azure Government delivers a dedicated cloud enabling government agencies and their partners to transform mission-critical workloads to the cloud [19].

I would recommend Amazon EMR for the education sector. Amazon AWS enables education institutions to create a customized infrastructure that is more efficient and flexible. Reduce costs with the AWS pay-as-you-go model. The ability to scale up and down as traffic changes make it possible to ramp up their capacity temporarily, at a fraction of the cost. Institute can also leverage this model to temporarily create a new cluster for testing purposes without the huge cost of Hadoop cluster ownership in the premise [20].

VI. CONCLUSION

With the rapid rising of structure and unstructured data, it is time for an organization to re-evaluate its current IT infrastructure setup, data storage management, and analytics. On-premise legacy systems will remain for low volume, highly sensitive, and valuable data. High performance, cost-effectiveness, scalability, and streamlined architecture of Hadoop and cloud will benefit the organization in the long run.

Although all three vendors have the same Hadoop platform. They have slights different in their product integration and product offering.

Future work may include optimizing the platform with a user-friendly interface, improving real-time data processes with multi-stream processing, and returning high accuracy results.

REFERENCES

[1] Amrit Pal, Dr. Sanjay Agrawal, A Study of Data Management Technology for Handling Big Data, International Journal of Computer Science and Mobile Computing, Vol.3 Issue.9, September- 2014

[2] Allae Erraissi, Abdessamad Belangour, Abderrahim Tragha, A Big Data Hadoop building blocks comparative study, International Journal of Computer Trends and Technology (IJCTT) – Volume 48 Number 1 June 2017.

[3] Allae Erraissi, Abdessamad Belangour, Abderrahim Tragha, A Comparative Study of Hadoop-based Big Data Architectures, International Journal of Web Applications Volume 9 Number 4 December 2017.

[4] Vanika, Aman Kumar Sharma, A Comparative Study of Hadoop-Based Big Data Architectures, IJSART - Volume 4 Issue 8 – AUGUST 2018.

[5] Ionuț ȚĂRANU, Big Data Analytics Platforms analyze from startups to traditional database players, Database Systems Journal vol. VI, no. 1/2015.

[6] Bakshi Rohit Prasad, Sonali Agarwal, Comparative Study of Big Data Computing and Storage Tools: A Review, International Journal of Database Theory and Application Vol.9, No.1 (2016), pp.45-66.

[7] Mike Gualtieri, Noel Yuhanna, The Forrester Wave™: Big Data Hadoop Solutions, Q1 2014, Forrester Research, Inc. February 27, 2014.

[8] Noel Yuhanna, Mike Gualtieri, The Forrester Wave™: Big Data Hadoop Cloud Solutions, Q2 2016, Forrester Research, Inc. April 22, 2016.

[9] Microsoft Azure HDInsight, https://2xbbhjxc6wk3v21p62t8n4d4-wpengine.netdna-ssl.com/wp-content/uploads/2016/10/1555_MSFT-HW_AllUp_datasheet_r2t3_zv.pdf

[10] Cloudera and Hortonworks Complete Planned Merger, https://www.cloudera.com/about/news-and-blogs/press-releases/2019-01-03-cloudera-and-hortonworks-complete-planned-merger.html

[11] What are the Apache Hadoop and Apache Spark technology stack?, https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-overview

[12] What Is Amazon EMR? - Amazon EMR, https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-what-is-emr.html

[13] Enterprise data cloud | Cloudera, https://www.cloudera.com/about/enterprise-data-cloud.html

[14] HDInsight : Le Big Data selon Microsoft | OCTO Talks, https://blog.octo.com/hdinsight-le-big-data-selon-microsoft/

[15] Which is Right Hadoop Solution for You? - IT Cheer Up, http://www.itcheerup.net/2018/08/hadoop-solution/

[16] Forrester Wave HARK Cloud Hadoop and Spark, https://www.cloudera.com/campaign/forrester-wave-hark-cloud-hadoop-spark.html

[17] Apache Hadoop on Amazon EMR - Amazon Web Services, https://aws.amazon.com/emr/features/hadoop/

[18] Telecommunications Data Analytics, Big Data in Telecom, https://www.cloudera.com/solutions/telecommunications.html?tab=3

[19] Azure Government Documentation -Quickstarts, Tutorials, https://docs.microsoft.com/en-in/azure/azure-government/

[20] Cloud Computing for Education, https://aws.amazon.com/education/