Survey on Open Source Frameworks for Big Data Analytics

Chao-Hsu Chen and Chien-Lung Hsu
Department of Information Management, Chang Gung University,
No. 259, Wenhua 1st Rd., Guishan Dist., Taoyuan City 33302, Taiwan
E-mail: M0345208@stmail.cgu.edu.tw, clhsu@mail.cgu.edu.tw
Kuo-Yu Tsai
Department of Applied Mathematics, Chinese Culture University,
No. 55, Hwa-Kang Road, Yang-Ming-Shan, Taipei 11114, Taiwan
E-mail: cgy13@ulive.pccu.edu.tw

ABSTRACT

The main contribution of this study is to provide a feasible open source big data framework and comparison of characteristics of modules upon different layers so that whenever big data processing technologies is essential for industries and business applications to deploy with. In accord with comparison to successful models, open source big data framework and modules with different characteristic can be quickly referred to and adopted. Based upon this study, we can develop specific open source big data framework and modules for different industries in future.

KEYWORDS

Open Source, Big Data, Framework

1 INTRODUCTION

To the huge amount of growing data, Google has developed Google File System [1] and MapReduce [2] and also developed a variety of open-source software system. This leads to the development of large data handling software like Apache Hadoop [3] and Hadoop File System [4]. Big data services and solutions are experiencing a sustained growth, which also reflect the growing number of major vendors such as IBM [5], Oracle [6], HPE [7], Microsoft [8], SAS [9] and Cloudera [10].

Thus, comparative analysis and benchmarking big data platform has become increasingly important. There are many well-known large companies hunting for big data talented people and resources to construct large data applications [11, 12]. They do can take advantage of their huge corporate data to make business strategies faster, more efficient, reducing the resources and time-consuming, and thus enhance their competitiveness. In the construction of a complete big data applications, usually in data collection, storage, management, processing, rendering visualization, privacy control, business model and other sectors, it needs experts and expertise to resolve. Therefore, data, domain expert and communication specialists also appears to play an important role in big data era.

Big data has evolved through volume, variety, velocity, 3V [13] to volume, variety, velocity, value, veracity, 5V [14] in big data analysis capabilities. Many companies have faced one-day amount of data at a rate of tens, hundreds, from TB (terabyte) increases to PB (petabyte) level, so that the traditional database becomes difficult to handle the amount of data whereas volume is an important factor in these days.

Velocity is important since data nowadays is increasing faster and faster, such as mobile computing. Along with the popularity of social networks, data increases faster than traditional enterprise applications. With the flow of faster information, data processing and analysis have to speed up to keep everything up. Variety refers to the diversity of information, the internet is now not only the information, apart from information we post pictures, videos, and backup data. Value

means the ability to manage these increasing big data ecosystem. Veracity is relatively important since the correctness and accuracy of the information are very important these days. In this new era, big data is the heart of all the information. Many vendors rely on the analysis of their new generation systems to achieve and deal with this huge data.

In this paper, we will have a clearer understanding of the advantages and disadvantages of each open source platform usage of big data platform in the industry. We can save more time in referencing analytical framework. Finally, by the trend of big data platforms by international development to make useful recommendations of large data analysis applications.

2 BIG DATA INFRASTRUCTURE

2.1 Hadoop Stack Analytics

Hadoop platform as shown in the following two most important members as shown in Figure Figure 1 Hadoop Stack with different components. Hadoop Distributed File System (HDFS) [15] is to store data across a cluster of machines providing high availability and fault tolerance of distributed file system. Hadoop YARN [16] handles resource management and scheduling job across the cluster.

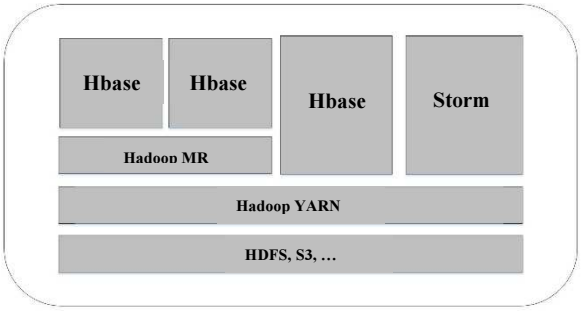

Figure 1 Hadoop Stack with different components

- ***MapReduce***

MapReduce [17] programming model used in Hadoop was proposed by Google's Dean and Ghemawat. The basic scheme in Hadoop MapReduce processing can be divided into two parts, known as mappers and reducers. At a high level, mappers read out data from HDFS to be processed, and produce results to reducers. Reducers are used in the polymerization intermediate results to produce the final output which is written to HDFS again. A typical Hadoop job involves running on different nodes in the multiple cluster of mappers and reducers. A good use of MapReduce can be found out in parallel data processing [18].

- ***MapReduce wrappers***

A specific packaging (Wrapper) for MapReduce is currently being developed. These packages provide better control of the source code to MapReduce to assist with.

- **Apache Pig:** Yahoo [19] developed a SQL-like environment being used in many organizations such as Yahoo , Twitter , Facebook, AOL, LinkedIn, etc.
- **Hive:** Facebook [20] developed another MapReduce packaging. Both packages provide a better environment in which the program can develop easier where the app developers do not have to deal with complex MapReduce program code.

- ***Limitations of MapReduce***

One of the major disadvantages of the MapReduce algorithm is its efficiency reduced when it is running iterative formula. Mapper [12] repeatedly reading the same data from the hard drive. The results of must be able to write to your hard drive each time before they are passed to the next. This bottleneck reduces MapReduce performance.

2.2 Berkeley data analytics stack (BDAS)

Spark developers also proposed what is known as a complete Data Processing Stack called Berkeley Data Analytics Stack (BDAS) [12, 21,] as in Figure **錯誤! 所指定的樣式的文字不存在文件中。**2. At the bottom of this stack , there is Tachyon [22] which is of HDFS component. The main advantage of Tachyon than Hadoop HDFS is that it can have more

aggressive memory usage. Another feature of Tachyon is its compatibility with Hadoop MapReduce. MapReduce can be run on Tachyon without any modification.

In BDAS, Tachyon upper layer is Apache Mesos. As a resource isolation and sharing to distributed applications in cluster manager, it has support to Hadoop, Spark, Aurora [23], and other applications. Mesos on scalability can be increased up to tens of thousands of nodes. In BDAS architecture, the third component running on the Mesos, Spark is part of playing the role of Hadoop MapReduce.

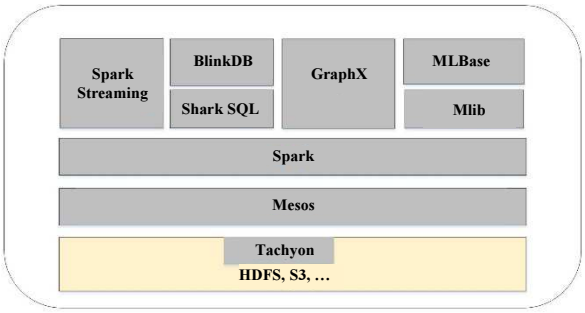

Figure 錯誤! 所指定的樣式的文字不存在文件中。 Berkeley Data Analysis Stack (BDAS) components

3 OPEN SOURCE BIG DATA PLATFORMS PRELIMINARY APPROACH

In this paper, a variety of open-source platform will massively discussed regarding scalability, data processing IO performance, data size, iterative support, real time processing and other advantages and disadvantages [11, 12]. Upon data collections and consulting domain experts, Hadoop Big Data Stack [15] and Berkeley data analytics stack (BDAS) [12, 21] different components and different data layer as in below table. There is fewer discussion regarding visualization and analysis at below table in BDAS.

Table 1 Big Data Infrastructure

	Open Source Hadoop Stack	BDAS
B.I.	Pentaho, SAP Cognos	Pentaho, SAP Cognos
Visualization	EDW, SAS, SAP, Qlikview, Tableau, D3.js, Kibana, Datawatch	
Analysis	EDW Connector, EDW, Hive, Impala, Stinger, Drill, Mahout, Mlib	Mlib
Data Mgmt	Oozie, Chukwa, Flume, Flumetd, Zookeeper, Spark SQL, Scribe, Logstash, Elasticsearch	Spark Streaming, Spark SQL, Blink DB, Graph X, ML Base
Data Access	Sqoop, Hive, Pig, Hbase, Apache Storm, Avro	Hadoop Map Reduced, Spark
Data Processing	Hadoop Map Reduced, Hadoop Yarn	Mesos
Data Storage	HDFS, S3, Hbase	HDFS, S3, Hbase, Tachyon

3.1 Characteristics and Discussion

With the explosive growth of information, more and more enterprises are deploying a private cloud system or hire a public cloud system to handle large data. Doug Cutting, creator of both Lucene and Hadoop, and Mike Cafarella originated Nutch and Hadoop history. Following is history of Hadoop creation [10].

Currently ways of analyzing big data platform can be classified into the following four [24, 25]：

- Transaction type RDBM Systems
  - **Enterprise Hubs:** IBM [26], Oracle [6], and Sybase [27] has traditional enterprise application analysis and processing transactions.
  - **Departmental Marts:** Microsoft SQL [8] and MySQL are suitable for small or medium business (SMB).

- Analytic Platforms
  - **MPP Database:** Massively parallel processing (MPP) data base providing vendors are Teradata [28], Microsoft[8] and EMC Greenplum [4].

- **Analytical Appliance:** IBM Netezza [26], EMC Greenplum [4], and Oracle Exadata [6] has analysis application.
- **In-Memory Systems:** SAP HANA [29] can provide in memory computing which is performance oriented data processing method.
- **Columnar:** SAP's Sybase IQ Hewlett Packard's Vertica, Paraccel, Infobright, Exasol, Calpont, and Sand [24] have the data stored by column rather than row manner that access to information easier.

- Hadoop Distributions: Hadoop Distributed File System (HDFS) [15] store data across commodity machine cluster, while providing high availability and fault tolerance using distributed file system.

- NoSQL Databases: These database mean not only SQL servers.
  - **Key Value Pair Databases:** Cassandra, Hbase, and Basho Riak [24] are this type of databases.
  - **Document Stores:** JSON, MongoDB and Couchbase [24] belong to this type of databases.
  - **SQL MapReduce:** Teradata's Aster Data [28] and EMC Greenplum [4] have data processing features using MapReduce.
  - **Graph Systems:** These systems contact people through multimedia companies.
  - **Unified Information Access:** Attivio, MarkLogic, and Splunk can handle both structured data and unstructured data.
  - **Other:** There are also many NoSQL databases resulting from the different uses of application and information [24].

3.2 Infrastructure and Analytics

Infrastructure and analytics can be distinguished in following layers.

- Data Storage Layer: Data storage layer [30] can be described as follows.
  - **Hbase:** Hbase [4] is an open source distributed database that Google run large data model written in Java [31, 32].
  - **S3:** Amazon S3 (Simple Storage Service) is provided by online file storage Web Services such as REST，SOAP and Bit Torrent [33].
  - **HDFS:** Hadoop Distributed File System is inspired by the Google File System which can store very large files in the machines across a large cluster [34].
  - **Tachyon:**Tachyon [21] is a memory centric distributed storage system to achieve reliable data stored in the cache shared across a cluster framework [22]. Tachyon avoids frequent disk read load data sets. And is also compatible with Hadoop. File system content is stored in memory of the body of all cluster nodes. Thus, the system achieves higher storage speed than that of traditional disk -based system like HDFS [32].

- Data Processing Layer: Characteristics for data processing layer can be found as follows.
  - **Yarn:** MapReduce after Hadoop 0.23 has a comprehensive overhaul of MapReduce which become MRv2 or Yarn. The basic concept is to give MRv2 JobTracker, resource management and job scheduling / monitoring two functions into separate programs [4, 16]. It has ResourceManager (RM) and ApplicationMaster (AM). The two main components are the Scheduler and Applications Manager [35].
  - **Mesos:** Apache Mesos is an open source cluster manager developed at the University of California Berkeley. It provides an effective isolation and sharing of resources across a distributed application or

framework [4, 36]. Currently, there are at least 50 organizations use Mesos including some large companies Twitter, Airbnb and Apple [37].

- **Map Reduce:** MapReduce is a programming model which is for processing and generating large data sets related to the implementation of parallel, distributed algorithm across the cluster [2, 4, 38]. MapReduce programs executed by a Map ( ) method of filtering and sorting and Reduce ( ) method, the executive summary of the operation. Map Reduce is a part of Apache Hadoop [38].
- Yarn and Mesos pros and cons [39] can be seen as in below table.

Table 2 Yarn and Mesos

		Yarn	Mesos
1	MapReduce [4, 39]	java	C++
2	Data processing [39]	In memory	In memory + CPU
3	Operating system [39]	Unix	Linux
4	Setups [40]	Specific instructions	Flexible
5	Coding [40]	3 times	1 time
6	Cluster [4, 40]	Hadoop like	Specific infrastructure
7	Security [4, 40]	Kerberos basic inheritance Hadoop security	Lack of implementation
8	Deployment [40]	Direct	Cannot directly deployed
9	Documentation and specs [40]	Less	Mature
10	Usage on Hadoop [40]	Transparent	Less transparent
11	Usability [40]	Yahoo, Hortonworks	Self-applications which have compatibility concerns.

- Data Access Layer: Characteristics for data access layer can be found as follows.
  - **Sqoop:** Sqoop acts in between relational databases and Hadoop command -line interface for transmitting data applications [41]. It supports a single table or SQL query incremental load and can run into the database which is updated several times to save jobs. Import can be used to fill Hive or table of Hbase [4, 41]. Export to Hadoop data can be converted into a relational database. Microsoft uses Sqoop based connectors to help transport from Microsoft SQL Server databases to Hadoop. Couchbase has provided Couchbase Hadoop server connector with Sqoop device.
  - **Hive:** Apache Hive [4, 42] is establishing a database on top of Hadoop infrastructure to provide data collection, query and analysis [32]. Apache Hive which has been first developed by Hive has now been used by other companies like Netflix and other development [43]. Amazon is also using Apache Hive software on Amazon Web Services for Amazon Elastic Map Reduce.
  - **Pig:** Apache Pig [44] is the data analysis program which consists of a high-level language for analyzing data, infrastructure and the platform used to evaluate those programs with large data sets [45]. Their structure is suitable for a large number of parallel processing, thus enabling them to handle very large data sets [4, 32]. Main features include easy to program and optimize the opportunities and scalability.
  - **Hbase:** Hbase [4] is an open source distributed database that Google run large data model written in Java [31, 32].
  - **Storm:** Apache Storm [46, 47] a free open source distributed real-time computing systems. Storm has a lot of use cases such as realtime analytics, online machine learning,

continuous computing, distributed RPC, ETL, and more[4, 32].

- **Avro:** Apache Avro [48, 4] is a data serialization system. Avro provides:
  1.Rich data structures.
  2.Small and fast binary data format.
  3.Container type files to store persistent data.
  4.Remote procedure call（RPC）.
  5.Simple dynamic language integration.
- **Map Reduce:** MapReduce is a programming model for processing and generating large data sets related to the implementation of parallel , distributed across the cluster algorithm [2, 4, 38]. MapReduce is executed by a Map( ) method of filtering and sorting and Reduce( ) method, the executive summary of the operation. Map Reduce is part of the Apache Hadoop [38].
- **Spark:** Apache Spark [4, 49, 50] a fast and large-scale data processing engine. It was originally developed at the University of California Berkeley campus AMPLab and is an open source cluster computing framework. It was donated to the Apache Software Foundation. Comparing to two-stage disk of Hadoop MapReduce, the in-memory operation of Spark for certain applications provide performance up to 100 times faster.

- Data Management Layer: Characteristics of data management layer can be found as follows.
  - **Oozie:** Oozie [4, 27, 51] is Direct Acyclical Graphs(DAG) of MR workflow scheduling system . Its coordinator can offer job by time (frequency) and data availability.
  - **Chukwa:** Chukwa [27, 52, 53] is a large-scale log aggregation and analysis system. It is also an open source data collection system for monitoring large-scale distributed systems. Chukwa is built on Hadoop Distributed File System (HDFS) and Map Reduce framework and thus also inherits the scalability and robustness of Hadoop. It includes a flexible and powerful toolkit for displaying, monitoring and analyzing of the results, so that the data collected can be fully utilized.
  - **Flume:** Flume [4, 27, 54] a distributed, reliable service which can be used effectively to collect and aggregate a large number of mobile service log data.
  - **Zookeeper:** Zookeeper [4, 27, 55] is used to maintain configuration information, naming, providing distributed synchronization, and services for centralized services. All of these types of services use distributed programs or being used by another application.
  - **Spark SQL:** Spark SQL [4, 56] replaced Shark SQL. Spark SQL query can plan structured data, or use the familiar SQL data frames API available in Java, Scala, Python and R.
  - **Mlib:** MLlib [57] is suitable for Spark API using Python, NumPy and starts at Spark 0.9 interoperability .Hadoop can use any data source (such as HDFS, HBase or local files). It is easy to insert into Hadoop workflows. And its algorithm is a hundred times faster than that of Map Reduce.
  - **Logstash:** Logstash [4, 58] is an event management and logging tool. It is generally used in a larger system log collection, processing, storage and search activities.
  - **Scribe:** Scribe [59, 60] is a server used to gather real-time streaming data from a large number of server log data. It can be modified without the client's scalable, and robust fault or any particular machine on the network.

- Data Analysis Layer: Characteristics of data analysis layer can be found as follows.
  - **EDW:** Enterprise data warehouse (EDW) [61] is a system for reporting and data analysis. It is the central repository from one or more different sources of comprehensive data. They store current and historical data, and is used across the enterprise to create a knowledge-based analysis. Examples may include reports from comparisons of trends from quarterly and annually data to detailed daily sales analysis.
  - **Mahout:** Apache Mahout's [27, 62] goal is to build high-performance machines to quickly create scalable applications learning environment. The three main components are scalable algorithms, new Scala + Spark and H2O (Apache Flink ongoing) algorithm and Mahout which is a mature Hadoop's MapReduce algorithm.
  - **Hive:** Apache Hive [4, 42] is to establish a database on top of Hadoop infrastructure to provide data collection , query and analysis [32]. Apache Hive which is initially developed by Facebook has now been used by other companies like Netflix and other development [43]. Amazon is also using Apache Hive software on Amazon Web Services for Amazon Elastic Map Reduce.
  - **Impala:** Impala [4, 63] is Cloudera's open source on Apache Hadoop cluster of computers which run massively parallel data processing (MPP) of the SQL query engine .
  - **Stinger:** Stinger [4, 64] is a continuation of speed, scale and a familiar SQL in the Hive. The three delivery schedules can be achieved in an open community of Apache Hive. These three goals are speed, size and SQL query.
  - **Drill:** Apache Drill [4, 27, 65] supports data-intensive distributed applications interactive analysis of large data sets of open-source software framework. It is an open source version of Google Dremel system and it can be used as Google Big Query of infrastructure services.

- Data Visualization Layer: Data visualization is an important layer of big data. Comparison [66] on this can be found as follows.
  - **SAS:** Statistical Analysis System (SAS) [67, 68] is a high level analysis system developed by the SAS Institute to use in software development, multivariate analysis, business intelligence, data management and predictive analysis software packages.
  - **SAP:** Systems, Applications & Products in Data Processing (SAP) [29, 69] is an application server that includes an in memory, column-oriented, relational database management system. It is previous known as "SAP High-Performance Analytic Appliance".
  - **Qlikview:** Qlikview [70] is an in-memory, business discovery tool which also is a Business Intelligence application that helps organizations big and small in data discovery. QlikView can deliver data visually which provides context of the data in rich but simple format. It consists of QlikView desktop, server and publisher.
  - **Tableau:** The visualization capabilities of Tableau [71, 72] are diverse and highly insightful. Tableau features such as "word clouds" or "bubble maps" are great to enhance comprehension. Its tree maps also provide the facility to add context for graphics. The tree

maps are mainly used to display relative proportions of multiple categories of a variety of information. There also is a capability for laying out the dashboard via “overlaps” is also a big powerful feature. It enables efficient use of screen space.

- **D3.js:** D3.js(Data-Driven Documents) [73, 74] can produce dynamic, interactive data visualization using JavaScript library on Web browser. It uses a widely implemented SVG, HTML5 and CSS standards. Even from its earlier Protovis framework. D3.js can control the final visual effect.
- **Elk:** Combining popular Elasticsearch, Logstash and Kibana, Elasticsearch company has established an end - to-end ELK stack [75, 76]. It provide immediate actionable insights for any type of structured and unstructured data sources.
- **Datawatch:** Datawatch [77] has the technique which relies on in-memory OLAP (Online Analytical Processing) perspective, which includes a tree display through a series of visualization. This allows the user to load data, select variables and hierarchy, and navigate through the resulting visualization, filtering, amplification and drilling (also known as slicing and cutting), to identify outliers, correlations and trends.

- Business intelligence: Business intelligence platforms [78, 79] are also very important in big data and benchmarking can be found as follows.
  - **Pentaho:** Pentaho [80, 81] provides business analysis. It is an open source business intelligence (BI) product that provides data integration, OLAP services, reporting, monitoring, data mining and ETL functionality. Hitachi acquired Pentaho in 2015.
  - **SAP:** Systems, Applications & Products in Data Processing (SAP) [29, 69] is an application server that includes an in memory, column-oriented, relational database management system. It is previous known as "SAP High-Performance Analytic Appliance". SAP attracts larger organizations with complicated needs.
  - **Cognos:** Cognos [82, 83] can resolve to help understand , monitor and manage business performance including business reporting and analysis, profitability measurement, budgeting, forecasting and optimization of cost management. It is a fast and effective technology to provide multi-dimensional business intelligence data.

4 CASE STUDY- FACEBOOK

Facebook [11, 84] is collecting data from two sources. MySQL tier contains user data whereas web servers generate event based log data. Web server data is collected to Scribe servers and then executed in Hadoop clusters. The aggregated log data from Scribe server is written to HDFS. The data in HDFS is compressed periodically and transferred to Hive Hadoop for further processing. Data analysis queries in Facebook are specified with graphical user interface called HiPal or Hive command line interface.

Table 3 Facebook Infrastructure

	Architecture
B.I. [11, 85]	Pentaho, SAP Cognos
Visualization [11, 85]	EDW, SAS, SAP, Qlikview, Tableau, D3.js, Kibana, Datawatch, Project Palantir, Friend Wheel, Touch Graph Browser, Mutual Friend network, Nexus, HiPal
Analysis [11, 84, 85]	EDW Connector, EDW, Hive, Impala, Stinger, Drill Mahout, Mlib,
Data Mgmt. [11, 85]	Oozie, Chukwa, Flume, Flumetd, Zookeeper, Spark SQL, Scribe, Logstash, Elasticsearch, HiPal, Databee

Data Access[11, 86, 87]	Sqoop, Hive, Pig, Hbase, Apache Storm, Avro, Presto, Giraph
Data Processing[11, 40]	Hadoop Map Reduced, Hadoop Yarn, Cassandra
Data Storage[11, 87]	HDFS, S3, Hbase, Scuba

Table 4 Facebook Infrastructure Case Study

B.I.	Pentaho[1]	SAP Business[1]	Cognos[1]							
Visualization	EDW[1]	SAS[1]	SAP[1]	Qlikview[1]	Tableau[1]	D3.js[1]	Kibana[1]	Datawatch[1]		Project Palantir, Friend Wheel, Touch Graph Browser, Mutual Friend network, Nexus, HiPal
Analysis	EDW Connector[1]	EDW[1]	Hive[1]	Impala[1]	Stinger[1]	Drill[1]	Mlib[1]	Mahout[1]		
Data Mgmt	Oozie[1]	Chukwa[1]	Flume[1]	Flumetd[1]	Zookeeper[1]	Spark SQL[1]	Elastic search[1]	Logstash[1]	Scribe[1]	HiPal, Databee
Data Access	Sqoop[1]	Hive[1]	Pig[1]	Hbase[1]	Apache Storm[1]	Avro[1]	Hadoop Map Reduced[2]	Spark[2]		Presto, Giraph
Data Processing	Hadoop Map Reduced[1]	Hadoop Yarn[1]	Mesos[2]							Cassandra
Data Storage	HDFS[1, 2]	S3[1, 2]	Hbase[1, 2]	Tachyon[2]						Scuba

Notes: 1: Open Source Hadoop Stack 2. BDAS (Berkeley data analytics stack) Right most column is the proprietary components.

5 CONCLUSION

From the case study above, it is found that most of the real world application can be based on our general big data framework such as HDFS in storage layer, HIVE and PIG in data access layer, Tableau and Qlikview in visualization layer but other components maybe customized into proprietary software or by suggestions of domain experts.

The overall framework hierarchy and characteristics discussed with experts remain the same. Only difference may be specific needs and reference architecture of different proprietary domain. For example, even the same type of social media big data companies such as Facebook and Twitter has very similar general framework but they have separate proprietary customized software.

ACKNOWLEDGE

The authors gratefully acknowledge the support from Taiwan Information Security Center (TWISC) and Ministry of Science and Technology (MOST), under the grants MOST , 105-2923-E-182 -001 -MY3, 105-2221-E-182-053, 105-2632-H-182-001, and 105-2221-E-034 -020 -.

REFERENCES

[1] S. Ghemawat, H. Gobioff, and S.-T. Leung, “The Google File System,” in Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles. 2003: New York, NY, USA.

[2] J. Dean and S. Ghemawat, “MapReduce: A Flexible Data Processing Tool,” Communications of the ACM, vol. 53, no. 1, pp. 72-77, 2010.

[3] Apache Software Foundation, Hadoop 1.2.1 Documentation, Available from: http://hadoop.apache.org/docs/r1.2.1/index.html.

[4] T. White, Hadoop: The Definitive Guide. the 4th Edition. 2015: O' Reilly Media.

[5] IBM website, IBM Stream Computing, Available from: https://www.ibm.com/analytics/us/en/technology/stream-computing/.

[6] Oracle website, Oracle and Big Data: Big Data for the Enterprise, Available from: https://www.oracle.com/ big-data/index.html

[7] Hewlett Packard Enterprise website, Big Data Solutions, Available from: https://www.hpe.com/us/en/solutions/big-data.html.

[8] Miscrosoft website, Big Data, Available from: http://www.microsoft.com/enterprise/it-trends/big-data/default.aspx#fbid=m1BCtxTTPjr.

[9] A. Tattersall and M.J. Grant, "Big Data - What Is It and Why It Matters," Health Information & Libraries Journal, vol 33, no. 2, pp. 89-91, 2016.

[10] Cloudera website, Cloudera Enterprise: The World'S Most Popular Apache Hadoop Solution, Available from: http://www.cloudera.com/content/www/en-us/products.html.

[11] P. Pääkkönen and D. Pakkala, "Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems," Big Data Research, vol. 2, no. 2, pp. 166-186, 2015.

[12] D. Singh and C.K. Reddy, "A Survey on Platforms for Big Data Analytics," Journal of Big Data, vol. 1, no. 8, 2014.

[13] Gartner, Inc., Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data, Available from: http://www.gartner.com/newsroom/ id/1731916.

[14] Y. Demchenko, C. de Laat, and P. Membrey, "Defining architecture Components of the Big Data Ecosystem.," in Proceedings of 2014 International Conference on Collaboration Technologies and Systems (CTS), 2014, pp. 104-112.

[15] Apache Software Foundation, HDFS architecture guide, Available from https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html#Introduction.

[16] V.K. Vavilapalli, A.C. Murthy, C. Douglas, S. Agarwal, M. Konar, R. Evans, T. Graves, J. Lowe, H. Shah, S. Seth, B. Saha, C. Curino, O. O'Malley, B. Reed, and E. Baldeschwieler, "Apache Hadoop YARN: Yet Another Resource Negotiator," in Proceedings of the 4th annual Symposium on Cloud Computing, 2013, Article No. 5.

[17] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Communications of the ACM, vol. 51, no. 1, pp. 107-113, 2008.

[18] K.H. Lee, Y.J. Lee, H. Choi, Y.D. Chung, and B. Moon, "Parallel Data Processing with MapReduce: A Survey," ACM SIGMOD Record, vol. 40, no. 4, pp. 11-20, 2011.

[19] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins, "Pig Latin: a Not-so-foreign Language for Data Processing," in Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, 2008, pp 1099-1110.

[20] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, and R. Murthy, "Hive: a Warehousing Solution over a Map-reduce Framework," in Proceedings of the VLDB Endowment, vol. 2, no. 2, 2009, pp. 1626-1629.

[21] AMPlab UC Berkeley, Berkeley Data Analysis Stack, Available from: https://amplab.cs.berkeley.edu/software/.

[22] Alluxio website, Open Source Memory Speed Virtual Distributed Storage, Available from: http://http://www.alluxio.org/.

[23] Apache Software Foundation, Aurora Project Incubation Status, Available from https://incubator.apache.org/projects/aurora.html.

[24] W.Eckerson, Categorizing Big Data Processing Systems, Beye Network, Avaiable from: http://www.-eye-network.com/blogs/eckerson/archives/2012/02 categorizing_bi.php

[25] M.B. Nirmala, "A Survey of Big Data Analytics Systems: Appliances, Platforms, and Frameworks," Handbook of Research for Cloud Infrastructures to Big Data Analytics, IGI Global, 2014, p.p. 393-419.

[26] IBM website, Big Data- Big Data Solutions That Work for You, Available from: http://www-03.ibm.com/software/products/en/category/bigdata.

[27] M. Maier, Towards a Big Data Reference Architecture, Master's thesis, Eindhoven University of Technology, 2013.

[28] W. O'Connell, I.T. Ieong, D. Schrater, C. Watson, G. Au, A. Biliris, S. Choo, P. Colin, G. Linderman, E. Panagos, J. Wang, and T.Walter, "A Teradata Content-based Multimedia Object Manager for Massively Parallel Architectures," in Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, 1996, pp. 68-78.

[29] SAP website, SAP HANA, Available from: http://go.sap.com/index.html.

[30] S. Neumann, Storing Apache Hadoop Data on the Cloud - HDFS vs. S3, Available from: https://www.xplenty.com/blog/2014/03/storing-apache-hadoop-data-cloud-hdfs-vs-s3/.

[31] Apache Software Foundation, Apache HBase, Available from: http://hbase.apache.org/.

[32] J. Roman, The Hadoop Ecosystem Table, Available from https://hadoopecosystemtable.github.io/.

[33] Amazon, Amazon S3, Available from: http://aws.amazon.com/tw/s3/.

[34] Hadoop Wiki, Hadoop Distributed File System, Available from: https://wiki.apache.org/hadoop/HDFS.

[35] Apache Software Foundation, Apache Hadoop YARN, Available from: https://hadoop. apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html.

[36] Wikipedia, Apache Mesos, Available from: https://en.wikipedia.org/wiki/Apache_Mesos.

[37] Apache Mesos, Organizations Using Mesos, Available from http://mesos.apache.org/documentation/latest/powered-by-mesos/

[38] Wikipedia, Map Reduce, Available from: https://en.wikipedia.org/wiki/MapReduce.

[39] Quora, How Does YARN Compare to Mesos?, Available from: https://www.quora.com/How-does-YARN-compare-to-Mesos.

[40] Cloudtimes, Facebook's Big Data: New Concept in Data Management, Available from: http://cloudtimes.org/2012/09/03/facebooks-big-data-new-concept-in-data-management/.

[41] Apache Software Foundation, Apache Sqoop, Available from: http://sqoop.apache.org/.

[42] Apache Software Foundation, Apache Hive, Available from: https://hive.apache.org/.

[43] Wikipedia, Hive, Available from: https://en.wikipedia.org/wiki/Apache_Hive.

[44] Apache Software Foundation, Apache Pig, Available from: https://pig.apache.org/.

[45] Wikipedia, Pig (Programming Tool), Available from: https://en.wikipedia.org/wiki/Pig_(programming_tool).

[46] Apache Software Foundation, Apache Storm, Available from: http://storm.apache.org/.

[47] Wikipedia, Storm (event processor), Available from: https://en.wikipedia.org/wiki/Storm_(event_processor).

[48] Apache Software Foundation, Apache Avro, Available from: https://avro.apache.org/docs/current/.

[49] Apache Software Foundation, Apache Spark, Available from: http://spark.apache.org/.

[50] Wikipedia, Spark, Available from: https://en.wikipedia.org/wiki/Apache_Spark.

[51] Apache Software Foundation, Apache Oozie Workflow Scheduler for Hadoop, Available from: http://oozie.apache.org/.

[52] Apache Software Foundation, Apache Chukwa, Available from: https://chukwa.apache.org/.

[53] J. Boulon, A. Konwinski, R. Qi, A. Rabkin, E. Yang, and M. Yan, "Chukwa: a large-scale monitoring system," in Proceedings of Cloud Computing and its Applications (CCA '08), 2008.

[54] Apache Software Foundation, Apache Flume, Available from: https://flume.apache.org/.

[55] Apache Software Foundation, Apache Zookeeper, Available from: https://zookeeper.apache.org/.

[56] Apache Software Foundation, Spark SQL, Available from: https://spark.apache.org/sql/.

[57] Apache Software Foundation, Mlib, Available from: http://spark.apache.org/mllib/.

[58] Elastic, Logstash, Collect, Enrich & Transport , Available from: https://www.elastic.co/products/logstash.

[59] Quora, What's the Difference between Thrift and Scribe?, Available from: https://www.quora.com/Whats-the-difference-between-Thrift- and-Scribe.

[60] Wikipedia, Scribe (log server), Available from: https://en.wikipedia.org/wiki/ Scribe_(log_server).

[61] Wikipedia, Data Warehouse, Available from: https://en.wikipedia.org/wiki/Data_warehouse.

[62] Apache Software Foundation, Apache Mahout, Available from: http://mahout.apache.org/.

[63] Cloudera. Apache Impala (Incubating), Available from: http://www.cloudera.com/content/www/en-us/products/apache-hadoop/impala.html.

[64] Hortonworks. Enterprise Big Data Solutions, Available from: http://hortonworks.com/innovation/stinger/.

[65] Apache Software Foundation, Apache Drill, Available from: https://drill.apache.org/.

[66] E.M. Forster, G. Wallas, and A. Gide, Data Visualization- Discover, Analyze, Explore, Pivot, Drilldown, Visualize Your Data…"How do I know what I think until I see what I say?", Available from: https://apandre.wordpress.com/tools/comparison/.

[67] SAS website, Analytics Software & Solutions, Available from: http://www.sas.com/zh_tw/home.html.

[68] Wikipedia, SAS(Software), Available from: https://en.wikipedia.org/wiki/SAS_(software).

[69] Wikipedia, SAP SE, Available from: https://en.wikipedia.org/wiki/SAP_SE.

[70] Qlik websiet, Qli, Available from: http://www.qlik.com/.

[71] Tableau website, Tableau, Available from: http://www.tableau.com/.

[72] Wikipedia, Tableau Software, Available from: https://en.wikipedia.org/ wiki/Tableau_Software.

[73] Wikipedia, D3.js, Available from: https://en.wikipedia.org/wiki/D3.js.

[74] Mike Bostock, Data-Driven Documents, Available from: http://d3js.org/.

[75] Christopher'blog, Visualizing Data With Elasticsearch, Logstash and Kibana, Available from: https://blog.webkid.io/visualize-datasets-with-elk/.

[76] Scaleway, How to Collect and Visualize Your Log with ELK Stack., Available from: https://www.scaleway.com/docs/how-to-use-the-elk-stack-instant-apps/.

[77] DataWatch website, DataWatch, Available from: http://www. datawatch.com/.

[78] Business Application Research Center, The BI Survey 10, p.p. 4-19, 2011.

[79] Butler Analytics, Enterprise BI Platforms Compared, Available from: http://www.butleranalytics.com/enterprise-bi-platforms-compared/.

[80] Pentaho, A Comprehensive Data Integration and Business Analytics Platform, Available from: http://www.pentaho.com/.

[81] Wikipedia, Pentaho Suite, Available from: https://en.wikipedia.org/wiki/Pentaho.

[82] GoliInfo, Cognos Business Intelligence and Enterprise Performance Management, Available from: http://www.cognos-bi.info/.

[83] IBM, Cognos, Available from: http://www-01.ibm.com/software/analytics/cognos/.

[84] A. Thusoo Z. Shao, S. Anthony, D. Borthakur, N. Jain, J.S. Sarma, R. Murthy, and H. Liu, "Data Warehousing and Analytics Infrastructure at Facebook," in Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, 2010, p.p. 1013-1020 .

[85] S. Schroeder, 6 Gorgeous Facebook Visualizations, Available from: http://mashable.com/2009/08/21/gorgeous-facebook-visualizations/#Rg9f0Wo7FOqG.

[86] L. Chan, Presto: Interacting with Petabytes of Data at Facebook, Available from: https://www.facebook.com/notes/facebook-engineering/presto-interacting-with-petabytes-of-data-at-facebook/10151786197628920/.

[87] J. Wiener and N. Bronson, Facebook's Top Open Data Problems, Available from: https://research.facebook.com/blog/facebook-s-top-open-data-problems/.