A Survey of Apache Hadoop Framework in Support of Big Data

Leonardo B. A. de Ana, Paul HK Kim
School of Informatics and Computing
Indiana University
Bloomington, IN 47408, USA
{leonardo.deana, hyeunkyu.kim}@ge.com

***Abstract*–Advancements on the internet adoption over the last years have drastically increased the volume and complexity of the different Web services we see nowadays and that led to the generation of huge amounts of data. To process these data, both performance and availability are critical factors and that is where Apache Hadoop, a framework for storage and processing data at large-scale, is helping large corporations to address these issues. As we are experiencing a hype of Hadoop, a lot of companies are leveraging the framework to create their data lakes, traditional and advanced analytics, transactional data and finally using Hadoop as an application platform itself. Hadoop has been considered an effective tool and is currently used by large companies such as Facebook, GE, IBM, Yahoo! among others and companies which are late in the industrial internet business are making Hadoop a priority for 2015. This paper will introduce the main concepts of the Apache Hadoop framework and how is supporting Big Data in large companies, explaining its deployment models, advantages and disadvantages, applications, components, and the predictions of new trends on this tool.**

***Keywords*–Apache Hadoop Framework; Map Reduce; HDFS; YARN;**

1. INTRODUCTION

The Big Data subject itself merges with Hadoop[1] framework history. This open source software platform managed by the Apache Software Foundation[2] has proven to be very helpful in storing and managing vast amounts of data in a low cost and efficiently manner. Big data platforms can increase
capacity and performance by adding nodes at a linear cost increase. This shift will open a new era to businesses organizations, allowing them to mine their data for greater insights by combining different sets of data to have better understanding of consumer behaviors and interests as well as operational excellence such as supply chain optimization.

One of the key technologies that have been at the core of the big data initiative landscape is Apache Hadoop. Hadoop provides an ecosystem with multiple tools, Hadoop Distributed file System (HDFS)[3] and Hadoop MapReduce[4] – that store and process large datasets in a scalable and cost effective way.

Starting from analyzing how Big Data and Hadoop merge together, in this paper we focus on a deep understanding of Hadoop core components , Map Reduce, HDFS and the new YARN[5] , passing through its mainly deployment models, exploring its advantages and disadvantages, and how the framework is supporting Big Data for large companies.

Also we will touch Hadoop subprojects, our predictions for its new trends in the following years and finalizing with an overview of a data lake and how Hadoop is helping to become a reality.

[1] https://hadoop.apache.org/

[2] http://www.apache.org/

[3] https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

[4] http://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

[5] http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html

2. BIG DATA AND HADOOP

One of the great computational challenges of our time is to store, manipulate and analyze effectively the vast amount of data we produce today. Corporate systems, services, Web systems, social media, among others, produce together a huge amount of data, reaching the size of petabytes daily [1]. There is a huge potential in information from these data and many companies do not know how to get value from them. Most of these data is stored in an unstructured manner and in systems, languages and very different formats, which in many cases are incompatible.

These giant data sets are becoming a valuable source of information. This is because the ones who hold such data are now evaluated not only by the innovation of its applications, but also for the data they maintain and especially for the potential they may bring with it. Google Company does not have a high added value only by its powerful Web search algorithm and its many available services, but also for maintaining a large volume of data from its members. Data itself that passing through analysis tends to become valuable, enabling the creation of smart solutions for its users.

Another example where it can be seen a huge amount of data is in Facebook. Initially, it used in its infrastructure a relational database for data storaging. However, in a rapid expansion, the company went from an amount of 15 terabytes in 2007 to 700 terabytes in 2010, making this infrastructure improper [2]. In 2012, with about 900 million active users, and maintaining analytical information about these users, this number has reached the petabytes scale. Today, Facebook is considered the world`s largest social network and also one of the most valuable companies.

Based on these and other applications that have a huge amount of Data, it comes the concept called Big Data [3]. This term mentions not only the volume but also to its variety and the necessary speed for processing it. As there are many resources for generating these data, it brings an extensive variety of structured and unstructured formats. For generating structured data, we have as examples enterprise systems and Web applications. For generating unstructured data, we have these systems' logs, Web pages, social media, smartphones, images, videos, sensors, microphones. Finally, "Big Data" is also related to speed as many of the new applications need responses in a short period of time and, in many cases, in real time.

Big Data applications make a computing engine to create solutions that analyze large databases, process their heavy calculations, identify certain behaviors and make available specialized services in their domains, however, it could need the use of supercomputers. These calculations can consume hours or days of processing in conventional architectures. While in constantly evolution, conventional computing resources are insufficient to meet the growing complexity of new applications.

Driven by the great computing demand and physical restrictions of conventional architectures, parallel and distributed computing came as an alternative to address some of these great computational challenges. Currently, this computing model has a crucial role in processing and extracting information regarding Big Data applications. This computation is typically performed in clusters and computing grids that with a set of ordinary computers, can add high processing power at a relatively low cost associated.

Although the parallel and distributed computing is a promising mechanism to support the handling of Big Data applications, some of its characteristics inhibit their use by new users. Dividing a task in subtasks and then running them in parallel in several processing units is not trivial. In Addition, if the size and the division of the sub-tasks are not proper dimensioned, that can totally compromise application`s performance. Besides that, the programmer must extract the dependency between the application data, determine a load balancing and algorithm and a scheduling algorithm for the tasks to ensure the efficient of using computing resources and the application`s recovery or non-stop execution in case of a machine failure.

In that context it was developed the Apache Hadoop, a framework for processing large amounts of data in clusters and computational grids. The idea of promoting solutions to the distributed systems challenges in a single

framework is the central point of the Hadoop project. In that framework, problems such as: data integrity, nodes availability, application scalability and failover occurs transparently to user. In addition, its programming model and storage system promote fast data processing, superior to other similar technologies. Currently, besides being consolidated in the business world, the Apache Hadoop framework has also achieved growing support of the academic community, thereby providing scientific and practical studies.

The Apache projects that perform bellow set of functions are detailed in Figure 1. This set of projects and technologies represent the core of Enterprise Hadoop. Key technology powerhouses such as Microsoft, SAP, Teradata, Yahoo!, Facebook, Twitter, LinkedIn and many others are continually contributing to enhance the capabilities of the open source platform, each bringing their unique capabilities and use cases. As a result, the innovation of Enterprise Hadoop has continued to outpace all proprietary efforts[14].

2.1 HADOOP COMPONENTS

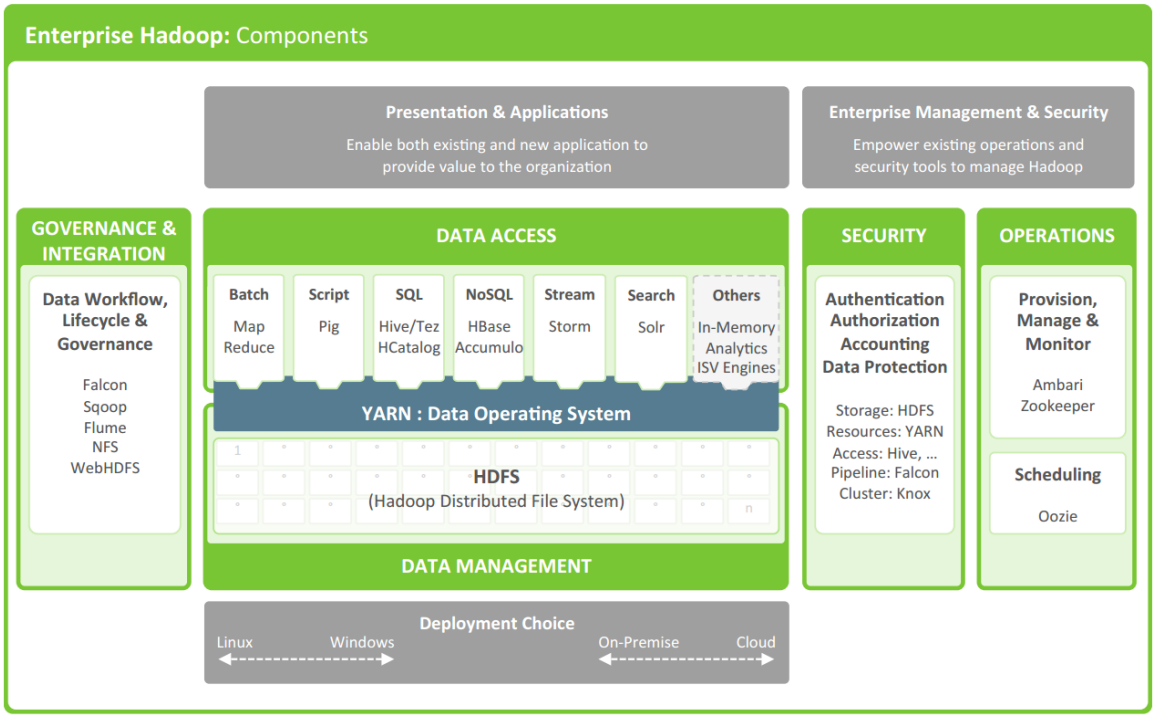

Figure 1. Hadoop Components

- **Data Management**: Hadoop Distributed File System (HDFS) is the core technology for the efficient scale out storage layer, and is designed to run across low-cost commodity hardware. Apache Hadoop YARN is the pre-requisite for Enterprise Hadoop as it provides the resource management and pluggable architecture for enabling a wide variety of data access methods to operate on data stored in Hadoop with predictable performance and service levels.

- **Data Access**: Apache Hive [6] is the most widely adopted data access technology, though there are many specialized engines. For instance, Apache Pig[7] provides scripting

[6] https://hive.apache.org/
[7] https://pig.apache.org/

capabilities, Apache Storm[8] offers real-time processing, Apache HBase[9] offers columnar NoSQL[10] storage and Apache Accumulo[11] offers cell-level access control. All of these engines can work across one set of data and resources thanks to YARN. YARN also provides flexibility for new and emerging data access methods, for instance Search and programming frameworks such as Cascading.

- **Data Governance & Integration**: Apache Falcon[12] provides policy-based workflows for governance, while Apache Flume[13] and Sqoop[14] enable easy data ingestion, as do the NFS and WebHDFS interfaces to HDFS.

- **Security**: Security is provided at every layer of the Hadoop stack from HDFS and YARN to Hive and the other Data Access components on up through the entire perimeter of the cluster via Apache Knox[15].

- **Operations**: Apache Ambari[16] offers the necessary interface and APIs to provision, manage and monitor Hadoop clusters and integrate with other management console software

2.2 HADOOP ADVANTAGES

Apache Hadoop is currently considered one of the best tools to processing huge amounts of data. Among the benefits of using it we can highlight these top 5 major advantages as per Michele Nemschoff[4]:

- **Scalable**: Hadoop is a highly scalable storage platform, because it can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel. Unlike traditional relational database systems (RDBMS) that can't scale to process large amounts of data, Hadoop enables businesses to run applications on thousands of nodes involving thousands of terabytes of data.

- **Cost effective**: Hadoop also offers a cost effective storage solution for businesses' exploding data sets. The problem with traditional relational database management systems is that it is extremely cost prohibitive to scale to such a degree in order to process such massive volumes of data. In an effort to reduce costs, many companies in the past would have had to down-sample data and classify it based on certain assumptions as to which data was the most valuable. The raw data would be deleted, as it would be too cost-prohibitive to keep. While this approach may have worked in the short term, this meant that when business priorities changed, the complete raw data set was not available, as it was too expensive to store. Hadoop, on the other hand, is designed as a scale-out architecture that can affordably store all of a company's data for later use. The cost savings are staggering: instead of costing thousands to tens of thousands of pounds per terabyte, Hadoop offers computing and storage capabilities for hundreds of pounds per terabyte.

- **Flexible:** Hadoop enables businesses to easily access new data sources and tap into different types of data (both structured and unstructured) to generate value from that data. This means businesses can use Hadoop to derive valuable business insights from data sources such as social media, email conversations or clickstream data. In addition, Hadoop can be used for a wide variety of purposes, such as log processing, recommendation systems, data warehousing, market campaign analysis and fraud detection.

- **Rapid**: Hadoop's unique storage method is based on a distributed file system that basically 'maps' data wherever it is located on a cluster. The tools for data processing

[8] https://storm.apache.org/
[9] http://hbase.apache.org/
[10] http://nosql-database.org/
[11] https://accumulo.apache.org/
[12] http://falcon.apache.org/
[13] https://flume.apache.org/
[14] http://sqoop.apache.org/
[15] https://knox.apache.org/
[16] https://ambari.apache.org/

are often on the same servers where the data is located, resulting in much faster data processing. If you're dealing with large volumes of unstructured data, Hadoop is able to efficiently process terabytes of data in just minutes, and petabytes in hours.

- **Resilient to failure**: A key advantage of using Hadoop is its fault tolerance. When data is sent to an individual node, that data is also replicated to other nodes in the cluster, which means that in the event of failure, there is another copy available for use.

When it comes to handling large data sets in a safe and cost-effective manner, Hadoop has the advantage over relational database management systems, and its value for any size business will continue to increase as unstructured data continues to grow.

2.3 HADOOP DESADVANTAGES

As Apache Hadoop is a new framework and evolving, some its features are not mature enough to support all situations that are being requested. In addition, as a disadvantage, are added other difficulties inherent by the parallel programming model. It is emphasized that some of these drawbacks listed below, as far as possible, constantly being mitigated with each new version of the framework. We can highlight the main disadvantages of using Hadoop for big data as per bigdatacompanies.com[5]:

- **Security Concerns**: Just managing a complex application such as Hadoop can be challenging. A classic example can be seen in the Hadoop security model, which is disabled by default due to sheer complexity. If whoever's managing the platform lacks the knowhow to enable it, your data could be at huge risk. Hadoop is also missing encryption at the storage and network levels, which is a major selling point for government agencies and others that prefer to keep their data under wraps.

- **Vulnerability**: Speaking of security, the very makeup of Hadoop makes running it a risky proposition. The framework is written almost entirely in Java, one of the most widely used yet controversial programming languages in existence. Java has been heavily exploited by cybercriminals and as a result, implicated in numerous security breaches. For this reason, several experts have suggested dumping it in favor of safer, more efficient alternatives.

- **Not Fit for Small Data:** While big data isn't exclusively made for big businesses, not all big data platforms are suited for small data needs. Unfortunately, Hadoop happens to be one of them. Due to its high capacity design, the Hadoop Distributed File System or HDFS lacks the ability to efficiently support the random reading of small files. As a result, it is not recommended for organizations with small quantities of data.

- **Potential Stability Issues**: Hadoop is an open source platform. That essentially means it is created by the contributions of the many developers who continue to work on the project. While improvements are constantly being made, like all open source software, Hadoop has had its fair share of stability issues. To avoid these issues, organizations are strongly recommended to make sure they are running the latest stable version, or run it under a third-party vendor equipped to handle such problems.

2.4 HADOOP IN LARGE COMPANIES

The effectiveness obtained by Hadoop can be found by checking the amount of large companies from different branches which are using Hadoop for either educational or business purposes. One of the biggest Hadoop developers and project contributors have been the company Yahoo!. Next we will present a list of some of the major institutions using Hadoop, however, a more comprehensive list can be found on the project website[6]

- **Adobe**(www.adobe.com)
    - Where used: have about 30 nodes running HDFS, Hadoop and HBase in clusters ranging from 5 to 14

nodes on both production and development. Have planned a deployment on an 80 nodes cluster.

- **Ebay**(www.ebay.com)
  - Where used: 532 nodes cluster (8 * 532 cores, 5.3PB).
  - Heavy usage of Java MapReduce, Apache Pig, Apache Hive, Apache HBase
  - Using it for Search optimization and Research.

- **Facebook**(www.facebook.com)
  - Where used: Use Apache Hadoop to store copies of internal log and dimension data sources and use it as a source for reporting/analytics and machine learning.
  - Currently they have 2 major clusters:
    - A 1100-machine cluster with 8800 cores and about 12 PB raw storage.
    - A 300-machine cluster with 2400 cores and about 3 PB raw storage.
    - Each (commodity) node has 8 cores and 12 TB of storage.
  - Heavy users of both streaming as well as the Java APIs. They have built a higher level data warehousing framework using these features called Hive and also developed a FUSE implementation over HDFS.

- **General Electric(**www.ge.com):
  - Where used: GE has invested in a data analysis platform called Predix for its industrial data lake which leverages the massively parallel processing of Apache Hadoop Framework. This combination will have numerous applications across many industries and types of hardware, from jet engines and locomotives to medical scanners.

- **Google / IBM** (www.google.com / www.ibm.com)
  - Where used: University Initiative to Address Internet-Scale Computing Challenges[7]

- **Linkedin**(www.linkedin.com)
  - Where used: have multiple grids for analysis and search of similarity between user profiles.
  - Hardware:
    - ~800 Westmere-based HP SL 170x, with 2x4 cores, 24GB RAM, 6x2TB SATA
    - ~1900 Westmere-based SuperMicro X8DTT-H, with 2x6 cores, 24GB RAM, 6x2TB SATA
    - ~1400 Sandy Bridge-based SuperMicro with 2x6 cores, 32GB RAM, 6x2TB SATA

- **Twitter**(www.twitter.com)
  - Where used: Use Apache Hadoop to store and process tweets, log files, and many other types of data generated across Twitter. They store all data as compressed LZO files.
- **Yahoo!** (www.yahoo.com)
  - Where used: Used to support research for Ad Systems and Web Search. Also used to do scaling tests to support development of Apache Hadoop on larger clusters
  - More than 100,000 CPUs in >40,000 computers running Hadoop
    - Biggest cluster: 4500 nodes (2*4cpu boxes w 4*1TB disk & 16GB RAM)

3.1 HADOOP FRAMEWORK WITH MAPREDUCE

As Apache Hadoop has become successful in its role in enterprise data architectures, the capabilities of the platform have expanded significantly in response to enterprise

requirements. The key elements of a Hadoop platform in its early days are enabled storage - HDFS and computing engine –MapReduce. As shown in Figure 2[13], MapReduce and HDFS played key roles among components, while other are built around the core.

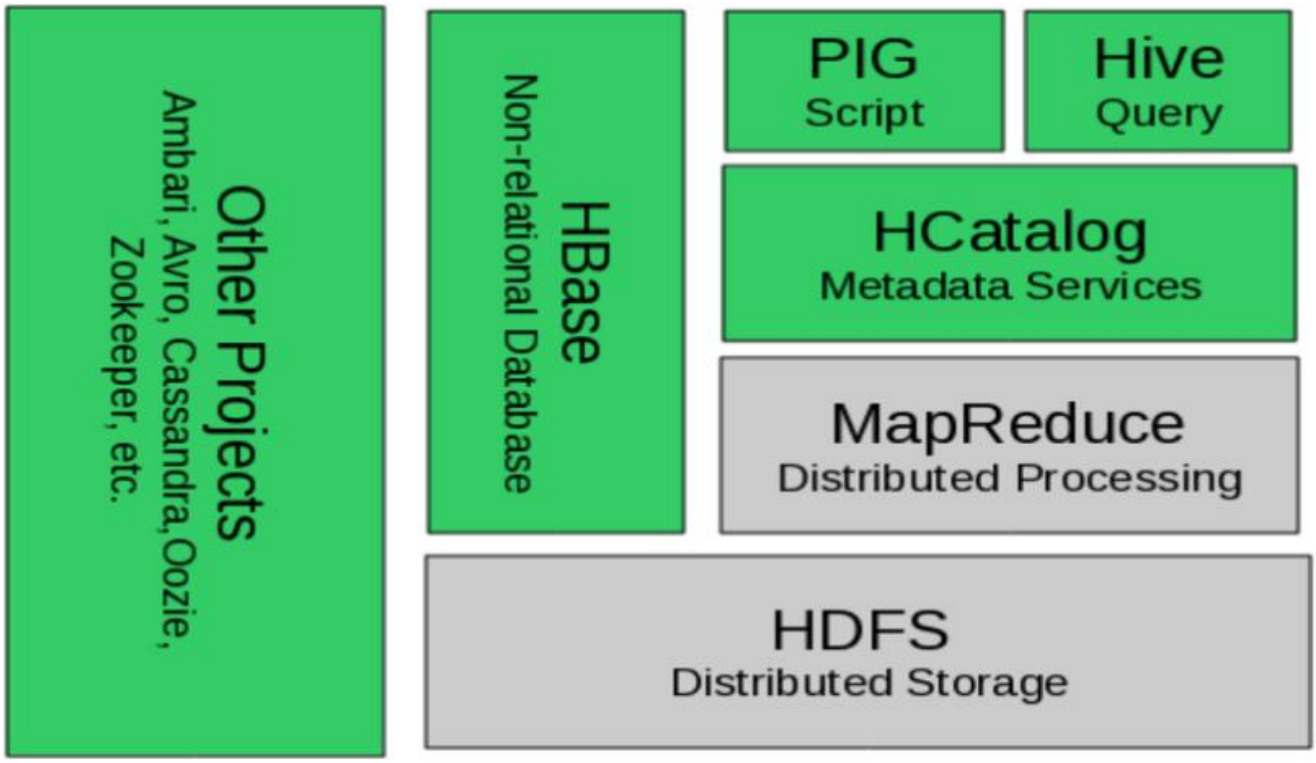

Figure 2. The Hadoop 1.0 ecosystem with HDFS and MapReduce

- **HDFS:** The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. Files in HDFS are divided into large blocks, typically 64MB, and each block is stored as a separate file in the local filesystem.

  HDFS is implemented by two services: the NameNode and DataNode. The NameNode is responsible for maintaining the HDFS directory tree, and is a centralized service in the cluster operating on a single node. Clients contact the NameNode in order to perform common file system operations, such as open, close, rename, and delete. The NameNode does not store HDFS data itself, but rather maintains a mapping between HDFS file name, a list of blocks in the file, and the DataNode(s) on which those blocks are stored. In addition to a centralized NameNode, all remaining cluster nodes provide the DataNode service. Each DataNode stores HDFS blocks on behalf of local or remote clients. Each block is saved as a separate file in the node's local filesystem. Because the DataNode abstracts away details of the local storage

arrangement, all nodes do not have to use the same local filesystem. Blocks are created or destroyed on DataNodes at the request of the NameNode, which validates and processes requests from clients. Although the NameNode manages the namespace, clients communicate directly with DataNodes in order to read or write data at the HDFS block level.

- **MapReduce**: Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

  The MapReduce processing model consists of two separate steps. The first step is an embarrassingly parallel map phase where input data is split into discreet chunks that can be processed independently. The second and final step is a reduce phase where the output of the map phase is aggregated to produce the desired result. The simple, and fairly restricted, nature of the programming model lends itself to very efficient and extremely large-scale implementations across thousands of low cost commodity servers (or nodes).

  One of the keys to Hadoop performance is the lack of data motion where compute tasks are moved to the servers on which the data reside and not the other way around (i.e., large data movement to compute servers is minimized or eliminated). Specifically, the MapReduce tasks can be scheduled on the same physical nodes on which data are resident in HDFS, which exposes the underlying storage layout across the cluster. This design significantly reduces the network I/O patterns and keeps most of the I/O on the local disk or on a neighboring server within the same server rack.

3.2 HADOOP FRAMEWORK WITH YARN

The YARN (Yet Another Resource Negotiator) project was started by the core development team to give Hadoop the ability to run non-MapReduce jobs within the Hadoop framework. YARN provides both full compatibility with existing MapReduce applications and new support for virtually any distributed application. Figure 3[13] illustrates how YARN fits into the new Hadoop ecosystem.

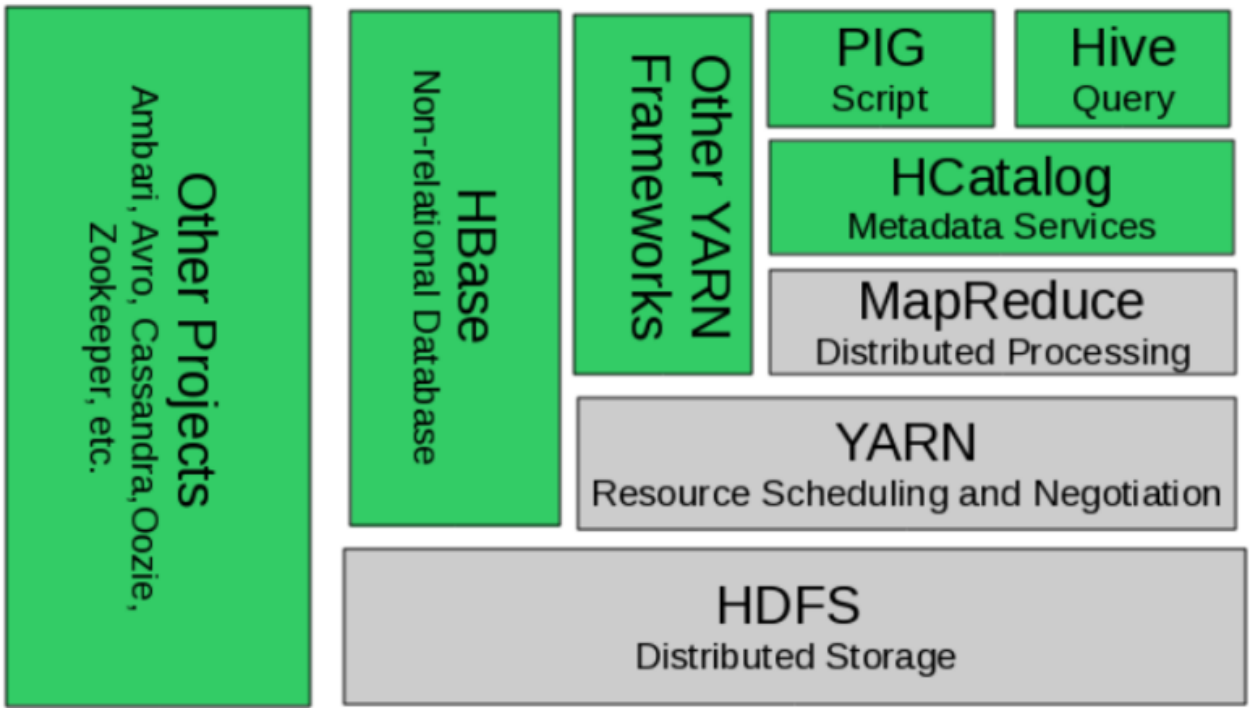

Figure 3. New Hadoop framework with YARN

YARN[13]: The fundamental idea of YARN is to split up the two major responsibilities of the JobTracker, in other words resource management and job scheduling/monitoring, into separate daemons: a global ResourceManager and per-application ApplicationMaster (AM). The ResourceManager and per-node slave, the NodeManager (NM), form the new, and generic, operating system for managing applications in a distributed manner.

The ResourceManager is the ultimate authority that arbitrates resources among all the applications in the system. The per-application ApplicationMaster is, in effect, a framework specific entity and is tasked with negotiating resources from the ResourceManager and working with the NodeManager(s) to execute and monitor the component tasks. The ResourceManager has a scheduler component, which is responsible for allocating resources to the various running applications subject to familiar constraints of capacities, queues etc.

The Scheduler is a pure scheduler in the sense that it performs no monitoring or tracking of status for the application, offering no guarantees on restarting failed tasks either due to application failure or hardware failures. The scheduler performs its scheduling function based on the resource requirements of an application by using the abstract notion of a resource container, which incorporates resource dimensions such as memory, CPU, disk, network etc.

The NodeManager is the per-machine slave, which is responsible for launching the applications' containers, monitoring their resource usage (CPU, memory, disk, network), and reporting the same to the ResourceManager. One of the crucial implementation details for MapReduce within the new YARN system is the reuse of the existing MapReduce framework without any major surgery. This step was very important to ensure compatibility for existing MapReduce applications and users.

3.3 HADOOP DEPLOYMENT MODELS

There are mainly 4 different deployment models of adopting Hadoop in business

organization. Before making a decision out of 4 following options, enterprises have to consider at least 4 factors carefully which will drive successful implementation of Hadoop technology.

The first area is price over performance ratio which is significantly important compared to other factors. The Hadoop-as-a-service model is basically cloud-base and uses virtualization technology to automate operation process. In comparison, the other models typically use physical machines directly. The second area is data privacy, which is a common pain point when storing data outside of company owned infrastructure. The next area is the productivity of developers and data scientists which can determine the organic capacity of an organization. The fourth area is data enrichment, which involves leveraging multiple datasets to uncover new insights. Common challenge is that the storage of these multiple datasets increases the volume of data, resulting in slow connectivity. Therefore, considering both current and future needs is critical when you design your Hadoop adoption strategy. [10]

- **Full customization with one-premise**: With this model, businesses purchase commodity hardware, then they install software and operate it by themselves. This option gives businesses full control of the Hadoop cluster in order to freely customize Hadoop to meet their specific needs and requirements.

- **Hadoop appliance**: This preconfigured Hadoop cluster allows businesses to skip detailed technical configuration decisions and to focus on data analytics based on their collective data set in a scalable way.

- **Hadoop hosting**: Much as with a traditional ISP model, organizations rely on a service provider to deploy and operate Hadoop clusters on their behalf.

- **Hadoop as a Service**: Hadoop-as-a-Service refers to a cloud service that allows users to deploy Hadoop clusters on demand, run MapReduce jobs, and tear down the clusters when the jobs are completed. This option gives businesses instant access to Hadoop clusters with a pay-per-use consumption model, providing greater business agility.

4.1 HADOOP NEW TRENDS

We are experiencing a hype of Apache Hadoop framework. A lot of companies who are late in the industrial internet business are leveraging Hadoop to create their data lakes, traditional and advanced analytics, transactional data and making the platform a priority in 2015.

Several companies involved in analyzing digital information including General Electric, Hortonworks, IBM, Pivotal and Verizon would develop their products and services on a common core of Hadoop's key components[8].Common standards often follow early development of software and hardware. If more companies use the same components, it usually helps with learning and certification, application development, and new products.

Following Forrester[9] predictions for Hadoop in 2015 and the following years we describe this list of trends:

- **Hadoop talents increase:** is expected that what businesses and service providers have sometimes suffered from a shortage of Hadoop experts will quickly disappear as enterprises turn to their existing application development teams to implement projects such as developing MapReduce jobs using Java. CIOs tend to increasingly leverage in-house Hadoop talent rather than high-priced consultants.

- **Hadoop cluster in the cloud**: we expect to see more Hadoop clusters getting to the cloud in the following years as enterprises could spin up a thousand nodes to perform a particular job for a few hours and then knock back the compute and network resources to next to nothing when not needed. Off course is not suggest for all cases, but for really big experiments we see Hadoop getting benefit from the cloud instead of using on-premise systems.

- **Application platform**: Hadoop is changing its core analytical purpose and becoming an application platform with better resource management features provided by YARN, database options such as HBase and in-memory overlay Apache Spark. It's far beyond the analytics and sooner or later more database options and middleware will run directly on Hadoop.

- **Hadoop distributions:** we expect that other large enterprise vendors will create their own Hadoop distribution as IBM and Pivotal already have. This will increase competition and customer choice.

- **Hadoop becomes part of then operating system:** operating system vendors can potentially include Hadoop and make it a configurable option within their operating systems and virtualization platforms. If a technology management professional wanted to add a node to join a Hadoop cluster for example they could simply "turn it on" through configuration.

4.2 OTHER SUBPROJECTS

- **Avro [17] :** Data ranking system based on schemas. Its composition is given by a repository of persistent data, a compact format of binary data and supports for remote procedure calls (RPC).

- **Chukwa[18]:** System specialized in collection and analysis of large scale systems logs. Uses HDFS to store the files and MapReduce to reports generation. It has the advantage of an auxiliary tool kit, very powerful and flexible, which promises to improve visualization, monitoring and analysis of the collected data.

- **Hbase:** database created by the company Power Set in 2007, later becoming an Apache Software Foundation project. Considered an open source version of *BigTable* database, created by Google, it is a distributed, scalable database that supports structured and optimized storage for large tables.

- **Hive:** framework developed by the Facebook's staff, and have become an open source project on August 2008. Its main functionality is to provide an infrastructure that allows the use of *Hive QL*, a query language similar to SQL and other relational data concepts such tables, columns and rows, to facilitate the complex analysis made to non-relational data of a Hadoop application.

- **Pig:** a high-level data flow oriented programming language and an execution framework for parallel computing. Their use does not change the configuration of Hadoop cluster as it is used in client-side mode, providing a language called *Pig Latin* and a compiler capable of transforming Pig type programs in MapReduce programming model sequences.

- **ZooKeeper[19]:** framework created by Yahoo! in 2007 with the objective of providing a management service for high performance distributed applications, which provides means to facilitate the following tasks: node configuration, distributed process and service groups' synchronization.

4.3 VISION OF A DATA LAKE

With the continued growth in scope and scale of analytics applications using Hadoop and other data sources, then the vision of an enterprise data lake can become a reality. In a practical sense, a data lake is characterized by three key attributes[14]:

- **Collect everything:** A data lake contains all data, both raw sources over extended periods of time as well as any processed data.

[17] http://avro.apache.org/
[18] http://chukwa.apache.org/
[19] https://zookeeper.apache.org/

- **Dive in anywhere:** A data lake enables users across multiple business units to refine, explore and enrich data on their terms.

- **Flexible access:** A data lake enables multiple data access patterns across a shared infrastructure: batch, interactive, online, search, in-memory and other processing engines.

As a result, a data lake delivers maximum scale and insight with the lowest possible friction and cost. As data continues to grow exponentially, then Enterprise Hadoop and EDW investments can provide a strategy for both efficiency in a modern data architecture, and opportunity in an enterprise data lake.

4. CONLUSION

The release of Apache Hadoop YARN provides many new capabilities to the existing Hadoop big data ecosystem. While the scalable MapReduce paradigm has enabled previously intractable problems to be efficiently managed on large clustered systems, YARN provides a framework for managing both MapReduce and non-MapReduce tasks of greater size and complexity. We conducted primary research on features and mechanism of Apache Hadoop, a rapid, scalable, cost effective and resilient to failure framework, also covering the way it is being applied in business world.

There are several different ways of adopting Hadoop solutions based on service level and we also researched on how it is being used in large IT firms such as Yahoo, Facebook, IBM and Google as well as GE.

To realize the value in your investment in big data, use the blueprint for Enterprise Hadoop to integrate with 'Enterprise Data Warehouse' and related data systems. Building a modern data architecture enables your organization to store and analyze the data most important to your business at massive scale, extract critical business insights from all types of data from any source, and ultimately improve your competitive position in the market and maximize customer loyalty and revenues.

As we are experiencing a hype of Apache Hadoop framework we see a lot of demand and talent increase inside business and new jobs emerging to address this new area. Hadoop is becoming far beyond analytics and turning itself to an application platform having a lot of big players in the business, such as GE, IBM and Yahoo! to name a few, making its on distributions of the framework and increasing costumers choice and market competition. We also expect and would not be surprised to see Hadoop becoming part of the operating systems making it a configurable option within the systems.

Hadoop is enabling the generations of the big data lakes, increasing the possibilities of Industrial Internet and Big Data and that is why many companies is making investments on it a priority for this year.

REFERENCES

[1] White, T. (2010). Hadoop: The Definitive Guide. O'Reilly Media.

[2] Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Zhang, N., et al. (2010). Hive –A Petabyte Scale Data Warehouse Using Hadoop. *Architecture*, 996-1005.

[3] IBM. (2011). Bringing big data to the Enterprise. *IBMcom*.

[4] "Big data: 5 major advantages of Hadoop" December 2013. [Online] Available: http://www.itproportal.com/2013/12/20/big-data-5-major-advantages-of-hadoop/

[5] "5 big disadvantages of Hadoop for big data" 2015.[Online] Available:http://www.bigdatacompanies.com/5-big-disadvantages-of-hadoop-for-big-data/

[6] "Hadoop Powered by" 2015 [Online]Available: http://wiki.apache.org/hadoop/PoweredBy

[7] "Google and IBM Announce University Initiative to Address Internet-Scale Computing Challenges"2007 [Online] Available:

http://googlepress.blogspot.com.br/2007/10/google-and-ibm-announce-university_08.html

[8] “IBM, G.E. and Others Create Big Data Alliance” February 2015. [Online] Available:http://bits.blogs.nytimes.com/2015/02/17/ibm-g-e-and-others-create-big-data-alliance/

[9] “Forrester’s Hadoop Predictions 2015 ” November-2014.[Online] Available: http://blogs.forrester.com/mike_gualtieri/14-11-04-forresters_hadoop_predictions_2015

[10] Accenture Technology Labs (2013) “Hadoop Deployment Comparison Study”

[11] HDFS Java API [Online] Available: http://hadoop.apache.org/core/docs/current/api/

[12] HDFS source code [Online] Available: http://hadoop.apache.org/hdfs/version_control.html

[13] Arun Murthy (2014)“Apache Hadoop Yarn-Moving beyond MapReduce and Batch Processing with Apache Hadoop 2”

[14] Hortonworks (2014) “A Modern Data Architecture with Apache™ Hadoop -The Journey to a Data Lake”