**Monroe College**

**CS 675: Big Data Management and Analytics**

**Final Project Documentation**

**Group 1**

Gita Gurung

Madeline (Zhiyang) Lin

Priyanka Chaudhari

Sukriya Gurung

Professor Syed Farid

December 2, 2022

# **0. Introduction**

Analytics is a broad term that encompasses the processes, technologies, frameworks and algorithms to extract meaningful insights from data. Raw data in itself does not have a meaning until it is contextualized and processed into useful information Data Analytics is this process of applying qualitative and quantitative techniques when examining data, with the goal of providing valuable insights.

Big data is defined as collections of datasets whose volume, velocity or variety is so large that it is difficult to store, manage, process and analyze the data using traditional databases and data processing tools. In the recent years, there has been an exponential growth in data generated by information technology, industrial, healthcare, Internet of Things, and other systems.

In this final documentation, we would like to discuss how to analyze big data using Hadoop framework. First, we will cover the overview of Hadoop framework, including the MapReduce, Yet Another Resource Negotiator (“Yarn”) and Hadoop Distributed File System (“HDFS”). Next, we will elaborate on the topics of real time analysis, text analysis and blockchain. Then, we will cover the data governance. Last, we will discuss how we apply Hadoop framework to machine learning and data science fields.

# **2. Hadoop Framework**

## **2.1 Hadoop Overview**

We live in the age of big data, where the data volumes we need to work with on a day-to-day basis have outgrown the storage and processing capabilities of a single host. Big data brings with it two fundamental challenges: how to store and work with voluminous data sizes, and how to understand data and turn it into a competitive advantage.

Hadoop fills a gap in the market by effectively storing and providing computational capabilities for substantial amounts of data. Hadoop is a platform that provides both distributed storage and computational capabilities. It offers a way to parallelize and execute programs on a cluster of machines.

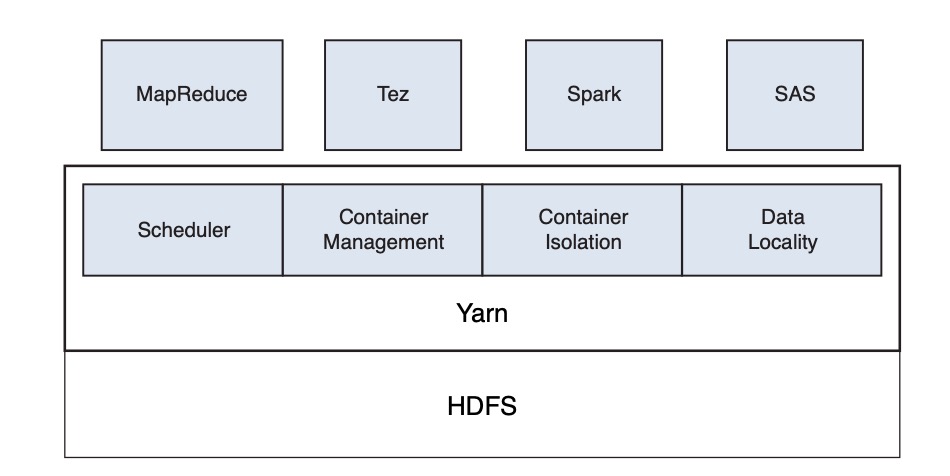


Figure 1: Hadoop Architecture

## **2.2 Hadoop Components**

## **2.2.1 MapReduce**

MapReduce is a java-based programming paradigm of Hadoop framework that provides scalability across various Hadoop clusters. It distributes the workload into various tasks that can run in parallel.

MapReduce performs **the Map Job** and **the Reduce Job**. The Map Job breaks down the data sets into key-value pairs or tuples. The Reduce Job then takes the output of the map job and combines the data tuples to into smaller set of tuples. The reduce job is always performed after the map job is executed.

Hadoop’s MapReduce implementation provides us with a fault-tolerant mechanism for large scale data analysis of heterogeneous structured and unstructured data sources.

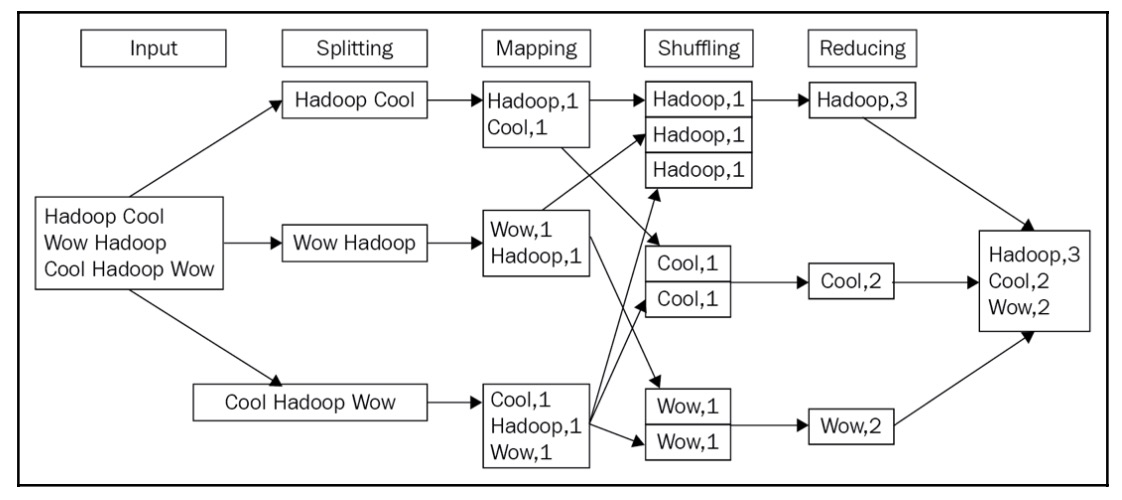


Figure 2: MapReduce Stages Overview

## **2.2.2 Yarn**

### **2.2.2.1 Yarn Overview**

MapReduce was introduced with the idea of dealing with processing massive data in parallel without compromising its efficiency. Soon it became a leading technology for big data due to its scalability and reliability. The open-source concept made this technology widely popular among academia and industry for big data processing. To maintain its popularity, the next step it needed was to improve its performance especially when many of the clusters are processing multiple jobs. The common issue was executing MapReduce jobs all at once when resources are limited. It was crucial to schedule their executions which require resource allocation to jobs and performance shouldn’t be compromised. However, the lack of proper management tools was not possible to utilize available resources efficiently leading to the development of Hadoop 2.0.

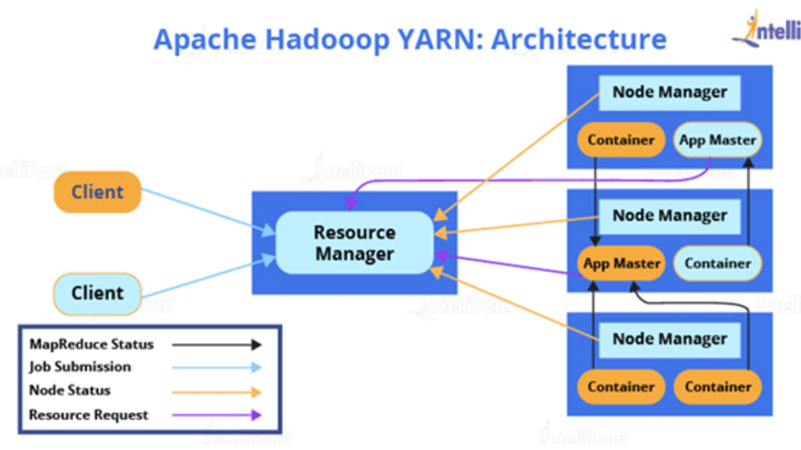
YARN is a core component of Hadoop 2.0. It is a resource management computing platform that is designed completely differently than traditional MapReduce. It is a layer that separates the resource management layer and processing components layer. The goal is to improve resource utilization and reduce the makespan of a given set of jobs. In YARN they removed “slot” which used to be a building block in the old version of MapReduce that prevented distinguishing “maps and reduce” tasks while allocating resources. Alternate to that each task must make a resource request of 2G memory and 1 CPU core and then it will be allocated to a node with sufficient capacity.

Figure 3: Hadoop 2.0

FIFO scheduler is an example of a widely adopted scheduling in YARN, however, it doesn’t consider the best solution for cluster resources. It is because FIFO requires both CUP and memory-intensive jobs sequentially which leads to unnecessary resource idleness.

### **2.2.2.2 YARN Architecture**

YARN consists of three important elements which are ResourceManager (“RM”), ApplicationMaster (“AM”), and NodeManager (“NM”).

Figure 4: Apache Hadoop YARN Architecture

The RM is the master server that can run multiple services, and most importantly the resource scheduler that decides how to assign the resources. RM knows the location of the data node and how many resources are in use.

The AM negotiates resources for a single application, the application runs in the first container allotted to it. Each AM requests resources from the RM and then works with containers provided by NM. There could be many NM in a single cluster.

The NM is the slave, when it starts it sends information to RM and offers resources to the cluster, the resource capacity depends on the amount of memory and virtual core. At runtime resources scheduler decides how to use this capacity. The application manager is an interface that maintains a list of applications that has been submitted.

## **2.3.1 HDFS**

HDFS is a java-based file system for scalable and reliable storage of large datasets. The main concept behind HDFS is that it divides a file into blocks instead of dealing with a file as a whole. This allows many features such as distribution, replication, failure recovery, and more importantly distributed processing of the blocks using multiple machines.

HDFS has two main components: the **NameNode** and the **DataNode**. The NameNode contains all the metadata of all content of the filesystem: filenames, file permissions, and the location of each block of each file, and hence it is the most important machine in HDFS. The DataNodes connect to the NameNodes and store the blocks within HDFS. They rely on the NameNodes for all metadata information regarding the content in the filesystem. If the NameNode does not have any information, the DataNode will not be able to serve information to any client who wants to read/write to the HDFS.

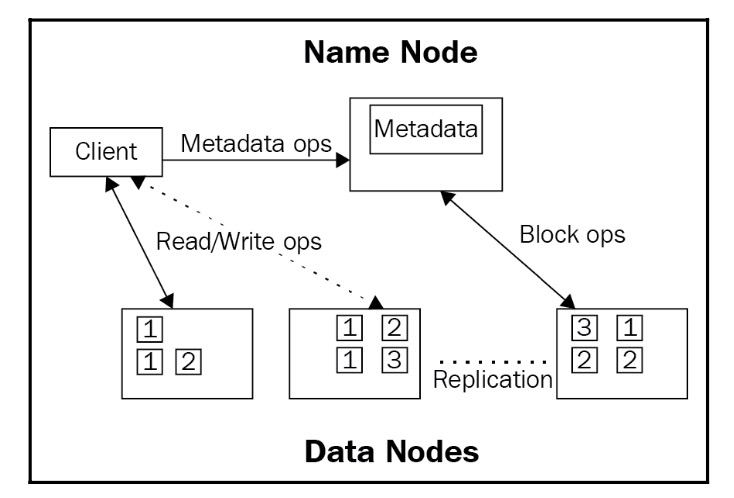


Figure 5: Map of files to blocks in the Name Node and replicas within the Data Node

# **3. Data Science**

## **3.1 Data Science Overview**

Data Science is the exploration of data via the scientific method to discover meaning or insight and the construction of software systems that utilize such meaning and insight in a business context.

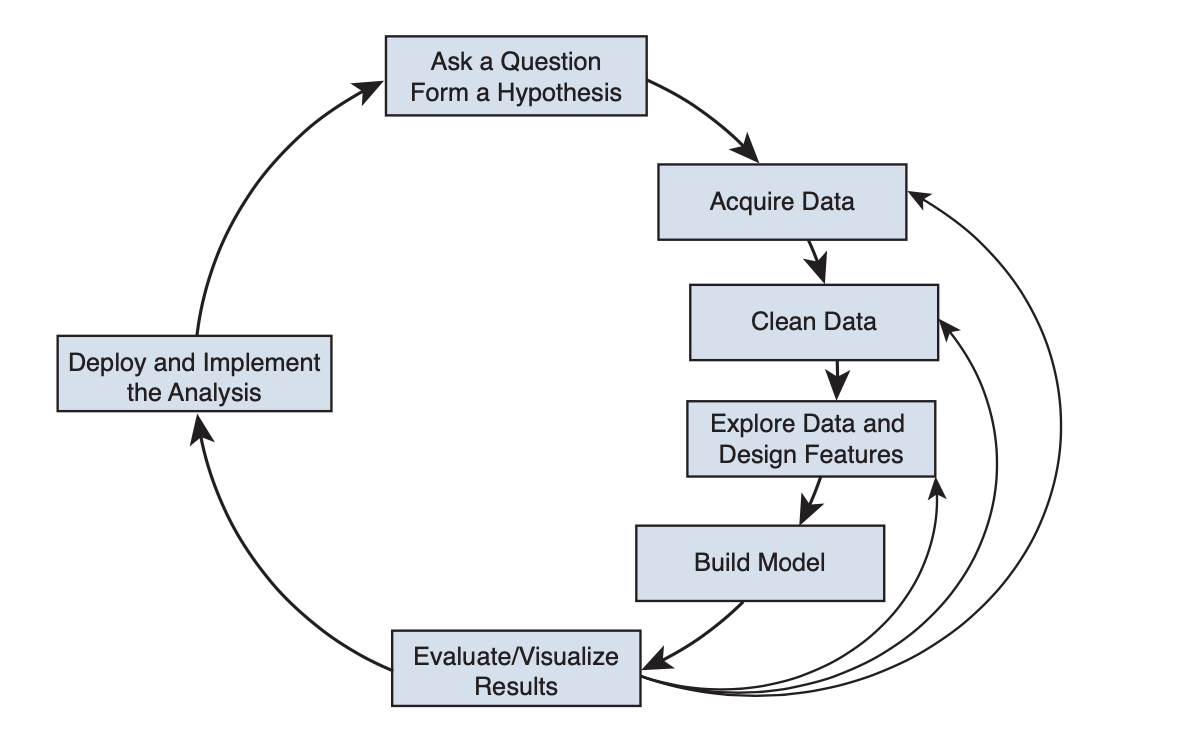


Figure 6: Data Science Flow Chart

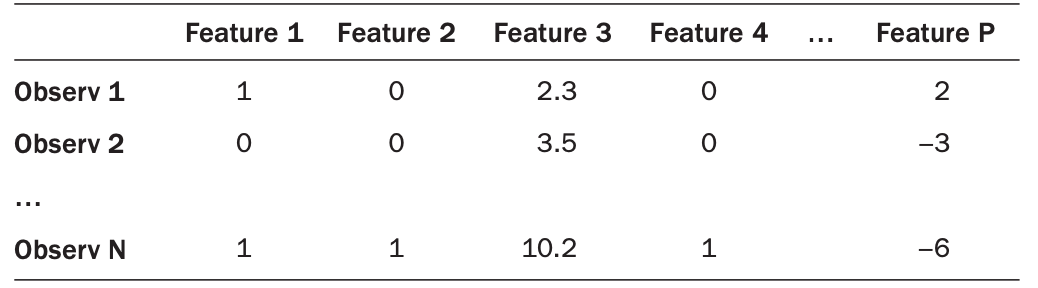
Data scientist has a set of tools to perform their job. Key data processing engines available on Hadoop are HDFS, YARN, Hive, Pig, Spark, Sqoop and Fume and popular modeling tools are R and Python. Hadoop facilitates data scientists to perform data acquisition, data quality analysis and cleaning, statistical computing, and visualization with minimal resistance due to its attributes such as cost-effective fault tolerant storage, multi-language tooling, scheme on read, robust scheduling and resource management, multiple levels of distributed systems abstractions, scalable creation of models and scalable execution of models.

## **3.2 Machine Learning**

Machine learning is a combination of statistics, computer science, and applied mathematics. Deep learning is based on neural networks. Deep learning algorithms consist of multiple layers of neural networks.

In machine learning, there are some terminologies including observation, feature, and target.

* Observation: Each observation is a representation, as data, of some object or entity such as an email message, a customer, a piece of equipment, etc.
* Feature: Each observation is represented as a vector of features (also called variables or attributes) of the observation.
* Target: Target represents our desire to predict the feature value.

Figure 7: Machine Learning Terminology

Task types in machine learning include supervised learning and unsupervised learning. Supervised learning is to learn an association between the inputs (features) and output (target variable) by using the examples provided. Unsupervised learning refers to a feature mix of observations without a target variable, and thus is often used for exploratory analysis or classification to gain insight or as a step before supervised learning. Unsupervised learning includes clustering, anomaly detection, recommender systems, and market basket analysis.

Data sets are growing faster than Moore's law. Scalable tools like Hadoop are one way to mitigate this issue that data comes from "everywhere". Big data with new and accurate features provides for better training of machine learning techniques. However, more instances do not always make machine-learning techniques perform better. The performance of a machine learning technique depends on the quality of the training set and how well the model generalizes to unseen data points.

# **4. Applications**

## **4.1 Blockchain**

The blockchain is a decentralized ledger, or list, of all transactions across a peer-to-peer network. Blockchain in Hadoop can impose security while making transactions, enables the possible inter-cluster communication, race between miners for validating and authorizing transactions, integrating the global economy with electronic currency for any kind of trading goods, materials, money, etc.

The BlockHDFS is the proposed solution which enhances the security of HDFS in a user

transparent way by storing metadata of files from HDFS in a blockchain. BlockHDFS consists of

three components: an HDFS cluster including the NameNodes and DataNodes, a permissioned

blockchain network such as Hyperledger Fabric, and a NodeJS Client which acts as a bridge

connecting the HDFS cluster and the blockchain network. In BlockHDFS, the blockchain is

responsible for storing the metadata of the files.

Integrating blockchain with distributed file systems such as HDFS can potentially improve security and traceability. In the context of this paper, the original design of Hadoop is more optimized for file processing instead of security-by-design. Hence, in this paper, we proposed a new approach to introduce blockchain (and more specifically, Hyperledger) to enhance the security of the HDFS ecosystem. In the current implementation, we have only added minimal metadata to the blockchain, but with BlockHDFS, one can easily add more features suited to their application needs. For future work, BlockHDFS can be extended to work in real-time with the file system and track all data between NameNode and DataNodes of the HDFS in the secure ledger with multiple nodes.

## **4.2 Text Analysis**

Text analysis is a machine learning technique used to automatically extract valuable insights from unstructured text data. Companies use text analysis tools to quickly digest online data and documents, and transform them into actionable insights. Text analysis delivers qualitative results while text analytics delivers quantitative results. If a machine performs text analysis, it identifies important information within the text itself, but if it performs text analytics, it reveals patterns across thousands of texts, resulting in graphs, reports, tables, etc. Text analysis tools allow businesses to structure vast quantities of information, like emails, chats, social media, support tickets, documents, and so on, in seconds rather than days, so we can redirect extra resources to more important business tasks.

### **4.2.1 Text Analysis Methods & Techniques**

**Text Classification:** It is the process of assigning predefined tags or categories to unstructured text. Sentiment analysis, topic modeling, language detection, and intent detection. Sentiment analysis uses powerful machine learning algorithms to automatically read and classify opinion polarity (positive, negative, neutral) and beyond, into the feelings and emotions of the writer, even context and sarcasm. The topic analysis automatically organizes text by subject or theme. Intent detection is often used to automatically understand the reason behind customer feedback.

**Text Extraction:** It extracts pieces of data (keywords, prices, company names, and product specifications) that already exist within any given text like news, reports, product reviews, and more. Keywords are the most used and most relevant terms within a text, words, and phrases that summarize the contents of the text. A named entity recognition extractor finds entities, which can be people, companies, or locations, and exist within text data.

**Word Frequency:** It measures the most frequently occurring words or concepts in a given text using the numerical statistics TF-IDF (term frequency-inverse document frequency).

**Collocation:** It helps to identify words that commonly co-occur. It identifies hidden semantic structures and improves the granularity of the insights by counting bigrams and trigrams as one word.

**Concordance:** It helps identify the context and instances of words or a set of words. It can also be used to decode the ambiguity of the human language to a certain extent, by looking at how words are used in different contexts, as well as being able to analyze more complex phrases.

**Clustering:** Text clusters are able to understand and group vast quantities of unstructured data. Although less accurate than classification algorithms, clustering algorithms are faster to implement, because you don’t need to tag examples to train models.

### **4.2.2 How to Analyze Text Data**

**Data Gathering:** Data is gathered about the brand, product, or service from both internal and external sources. Internal data is generated every day, from emails and chats to surveys, customer queries, and customer support tickets. Customer Service Software, CRM, Chat, Email, Surveys, Databases, and Product Analytics are some examples of internal data. You can use web scraping tools, APIs, and open datasets to collect external data from social media, news reports, online reviews, forums, and more, and analyze it with machine learning models.

**Data Preparation:** In order to automatically analyze text with machine learning, we will need to organize data. However, it’s important to understand that automatic text analysis makes use of a number of natural language processing techniques like Tokenization, part-of-speech Tagging, and Parsing. Tokenization is the process of breaking up a string of characters into semantically meaningful parts that can be analyzed while discarding meaningless chunks. Part-of-speech tagging refers to the process of assigning a grammatical category, such as noun, verb, etc. to the tokens that have been detected. Parsing refers to the process of determining the syntactic structure of a text. Dependency parsing is the process of using dependency grammar to determine the syntactic structure of a sentence. Constituency parsing refers to the process of using constituency grammar to determine the syntactic structure of a sentence.

**Analyze your Text Data**

**Machine Learning-based System:** It can make a prediction based on what they learn from past observations. These systems need to be fed multiple examples of texts and the expected predictions for each which is called training data. The more consistent and accurate your training data, the better the ultimate prediction will be.

**Machine Learning Algorithms:** The Naïve Bayes family of algorithms is based on Baye’s Theorem and the conditional probabilities of occurrence of the words of a sample text within the words of a set of texts that belong to a given tag. Support Vector Machines is an algorithm that can divide a vector space of tagged texts into two subspaces: one space that contains most of the vectors that belong to a given tag and another subspace that contains most of the vectors that do not belong to that one tag. Deep learning is a set of algorithms and techniques that use artificial neural networks to process data much as the human brain does.

**Visualize your Text Data**

**MonkeyLearn Studio:** It is an all-in-one data gathering, analysis, and visualization tool. Deep learning machine learning techniques allow you to choose the text analyses you need and chain them together to work simultaneously.

**Google Data Studio:** Google’s free visualization tool allows you to create interactive reports using a wide variety of data. Once you’ve imported your data you can use different tools to design your report and turn your data into an impressive visual story.

**Looker:** Looker is a business data analytics platform designed to direct meaningful data to anyone within a company. The idea is to allow teams to have a bigger picture of what’s happening in their company.

**Tableau:** Tableau is a business intelligence and data visualization tool with an intuitive, user-friendly approach. It allows organizations to work with almost any existing data source and provides powerful visualization options with more advanced tools for developers.

### **4.2.3 DataServices Text Analysis and Hadoop**

DataServices first loads the source text in its own memory and then runs the text analysis on its own server/engine. The text sources are usually unstructured text or binary files such as Word, Excel, PDF files etc. If these files reside on a Hadoop cluster as HDFS files, DataServices can also push down the TDP transform as a MapReduce job to the Hadoop cluster.

### **4.2.4 Why run DataServices Text Analysis within Hadoop?**

### Running the text analysis with Hadoop can be an appealing approach if the total volume of source files is big and at the time the Hadoop cluster has enough resources. Then, the text analysis might run much quicker inside Hadoop than within DataServices. The text analysis extracts entities from unstructured text and as such it will transform unstructured data into structured data. This is a fundamental prerequisite in order to be able to run any kind of analysis with the data in the text sources. You only need the text sources as input for the text analysis.

### **4.2.5 Apache Hadoop-based effective sentiment analysis on demonetization and COVID-19 tweets**

### Sentiment analysis is the process of user’s sentiments as positive or negative estimations. For analyzing a large amount of data, Hadoop is suitable for large file procession and large storage of wider data units. Using the Apache Flume, the demonetization and covid 19 tweets were collected from the web server and aggregated into HDFS for Apache Pig Analysis. Apache Flume is a distributed open-source tool for aggregating log activities from the web server into HDFS. The three layers of Hadoop are used to process the client request, allocate jobs to their task trackers, and store the data as blocks into HDFS. The processed data can be uploaded and stored in HDFS. The study obtained a number of positive, and negative sentiment, and their association rule mining algorithm.

The study collected data on social media from the early phase of COVID-19 spread. Tokenization is the process of breaking a stream of sentences into a collection of words. After preprocessing the data, the resultant dataset was uploaded into HDFS. Data are categorized into positive, and negative. The unrelated tweets are also declared neutral in the demonetization and Covid-19 dataset. After uploading the dataset into HDFS, the dataset is analyzed by Apache Pig. A dataset with positive, negative, and neutral labels was created as a training dataset. The data is pre-processed to remove some of the unnecessary attributes from the dataset. An assortment of pre-processing steps is done and the data is cleaned for analysis. Thus, the data is fitting for all the learning algorithms. After the process of tokenization, compute the rating of tweet id, tweet for categorization of Tweets. In the Hadoop ecosystem, apache pig and apache flume are used. After categorizing the dataset, three machine learning techniques such as support vector machine, Naïve Bayes and Recurrent neural networks were chosen for performing the classification task.

## **4.3 Big Data Integration – Real Time Analysis**

Big data analytics is the process of examining large data sets to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. Big Data refers to an extremely large volume of data and data sets that include structured and unstructured data coming in from multiple sources. These datasets are so voluminous that traditional data processing software is unable to capture, manage, or process them. As a result, big complex data can be used to address business problems that were previously inaccessible. Big Data is a combination of *variety, volume, veracity, velocity, value* and *variability*.

Earlier Big Data was often recognized by **3Vs** but these have expanded and now can be characterized by **6Vs.**

1. The data collected from different sources which are relevant to the organization can be characterized as a **volume.**

2. Information comes in a **variety** of formats. They range from organized, conventional databases, also unstructured content, email, video, sound, etc.

3. The third most important characteristic is the speed or **velocity** at which the data is collected.

4. One of the additional characteristics is **veracity**, this means the level of authenticity of the collected data.

5. Once the data is collected and structured, the next important characteristic to measure is its **value** i.e the value that it holds to the organization

6. **Variability** helps us understand the various ways the collected data can be used.

**Real-Time** data is a boon. As we know that decision-making is an integral part of any business, data has become a crucial part of this process. From planning and forecasting to strategizing and testing, it’s the data that the data analyst needs. For modern data-driven businesses, meaningful data helps create intuitive decisions. When we talk about data analytics, a data analyst BI (“Business Intelligence”) supports data-driven business decision-making. This helps in keeping data operational. There is an increasing need for real-time processing of these enormous volumes, such as the 200 million emails, 300,000 tweets and 100 hours of YouTube videos that are passing by every minute of the day. Real-time processing reduces storage requirements while providing more responsive, accurate and profitable responses.

Theplatforms that can handle real time streaming big data are Apache Storm, Apache Kafka, IBM Infosphere Streams, etc.

**5. Data Governance**

## **5.1 Big Data Governance Best Practices**

Hadoop was initially developed without security or privacy considerations. Hence, there has been a huge gap in the data management tools, structure, and operations of Hadoop in the past. Without proper data governance tools and measures, the Hadoop data lake can become a data swamp. Data protection involves three categories of considerations: sensitive data protection, data sharing considerations, and adherence to governance policies.

Sensitive data protection is concerned with policies that ensure your operations are meeting regulatory and compliance requirements for sensitive data. This is predicated on classifying your data into critical, high-priority, medium-priority, and low-priority. All critical and high priorities are regarded as sensitive. Some organizations refer to such sensitive information as private information (“PI”).

In terms of data sharing considerations, it is a critical governance consideration – in particular in large organizations, multi-division, and global companies whose data generation and consumption span not only functional lines but countries, divisions, and affiliates such as corporate partners and franchises.

The flip side of enforcing governance policies is adherence. When it comes to data protection at a high level, we need to apply tokenization prior to data ingestion into the lake. Data tokenization is an important strategy to protect data. When tokenizing data, it is advisable to store the token keys on the token servers on your premise in a different subnet, in particular, if the data resides in the cloud. Tokenization replaces the digits with a randomly generated alphanumeric text ID (known as a “token”) which cannot be identified without de-tokenization using the initial key that was generated in the token server.

## **5.2 Big Data Governance Framework**

Data governance refers to the rules, structures, processes, and practices that allocate the roles, responsibilities, and rights of participants in big data analytics. This governance policy is intended to meet all relevant compliance with applicable laws and objectives of operating the big data platform in a safe and sound manner.

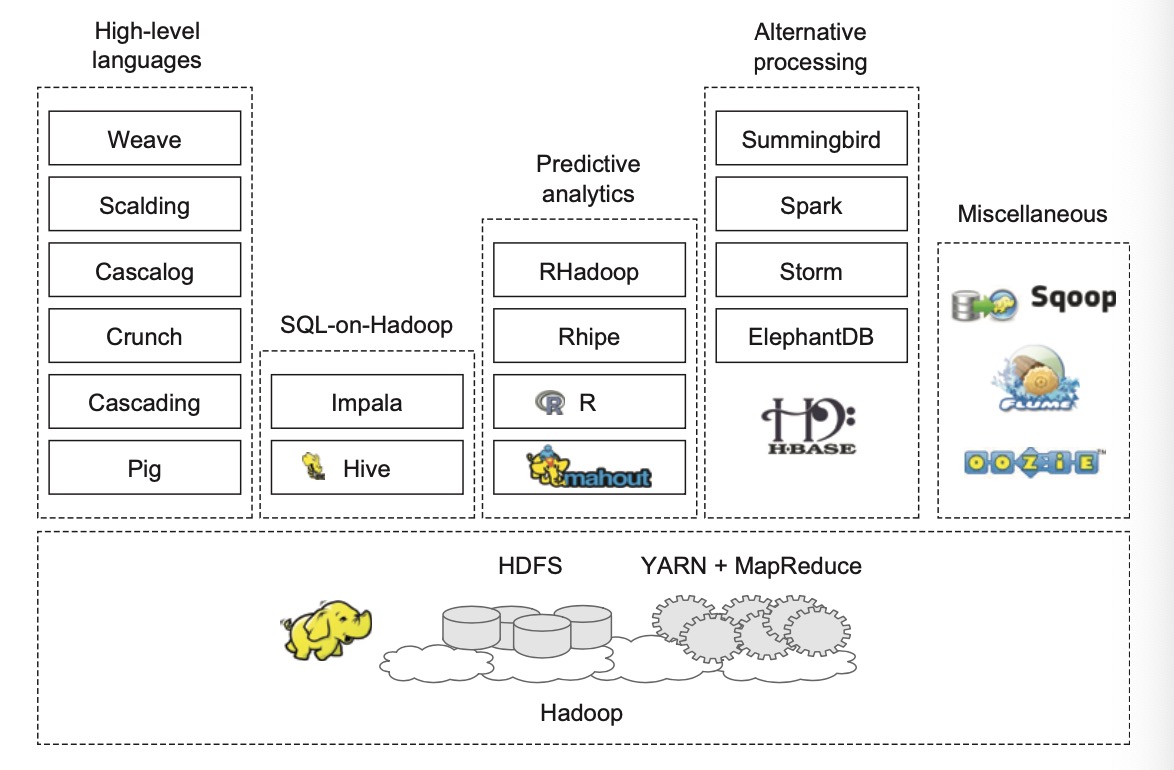
The framework consists of eight capabilities including organization, metadata management standards, data classification, security, privacy, and compliance standards, data usage agreement (“DUA”), security operations and policies, information lifecycle management, and data quality standards.

Big data governance is federated and organized by the data council. The data council meets regularly, typically monthly, to address changes in policy, manage data issues, audit results and issues related to compliance, security, and privacy. There are four data governance forums that are managed by the data council: 1) Metadata management, 2) Quality management, 3) Data governance, 4) Security/privacy and compliance. When it comes to data classification, it is an important step toward data and application management. As discussed in the Ch9 lecture notes, there are four data types: raw, keyed, validated, and refined.

As for data security, privacy and compliance standards, the security policies and processes outlined in this governance framework are aligned with the enterprise data security, privacy, and compliance rules. Four pillars of big data security are perimeter, access, visibility and protection.

Information lifecycle management governs the standards and policies of data from creation and acquisition to deletion. As for data quality standards, it defines requirements to classify, document, and manage data to ensure that the big data platform’s critical and high-priority data meets established quality requirements. The objective of this standard is to ensure that the platform’s data is managed and of sufficient quality. All data whether internally or externally obtained is to be classified into four categories including critical data, high-priority data, medium-priority data, and low-priority data. There are five metrics that are used when defining data quality rules for critical and high-priority data elements. They are accuracy, validity, completeness, timeliness, and consistency.

# **6. Summary**

Figure 8: Hadoop Overview

Many companies such as Amazon Web Services, Meta, Yahoo, Twitter, eBay, J.P. Morgan use Hadoop to compute Big Data. It has become the requirement for each and every company which is responsible for data management to manage various types of structured and unstructured data. Hadoop was initially designed to collect information from various sites. The above is a summary of the Hadoop ecosystem. In this documentation, we introduced about the three major components of Hadoop, which are MapReduce, Yarn and HDFS. Also, we discussed briefly about data science and machine learning and their applications. We talked about how we apply Hadoop framework in blockchain, text analysis, real-time analysis, etc. Moreover, we mentioned about the data governance, which is an indispensable element in the big data century. Hadoop comes in handy when dealing with huge amounts of data containing a variety of different data.

**Reference**

Lecture Note: Overview of big data analytics

Lecture Note: Introduction to Big Data

Lecture Note: Hadoop in Heart Beat

Lecture Note: Practical Data Science with Hadoop and Spark

Lecture Note: Hadoop data modeling

Lecture Note: Hadoop machine learning

Lecture Note: Introduction to Yarn

Lecture Note: Big Data Governance

Lecture Note: Blockchain-integrated Hadoop Distributed File System For Secure Provenance Traceability

Lecture Note: Big-Crypto Big Data, Blockchain and Cryptocurrency

Lecture Note: LoT Big Data provenance scheme using blockchain on Hadoop ecosystem

Lecture Note: Twitter sentiment analysis using Hadoop

Lecture Note: DataServices text analysis and Hadoop

Lecture Note: Data munging with Hadoop

Lecture Note: Sentiment analysis of tweets using Hadoop

Lecture Note: Real time analysis

Lecture Note: Data acquisition

Yao, Y., Wang, J., Sheng, B., Lin, J., & Mi, N. (2014). Haste: Hadoop Yarn Scheduling based on

task-dependency and resource-demand. *2014 IEEE 7th International Conference on Cloud Computing.* https://doi.org/10.1109/cloud.2014.34