Modern Applied Science; Vol. 13, No. 7; 2019
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

A Comparative Study on Big Data Analytics Frameworks, Data Resources and Challenges

Flasteen Abuqabita[1], Razan Al-Omoush[1] & Jaber Alwidian[1]

[1] Department of Data science, Princess Sumaya University for Technology, Jordan, Amman

Correspondence: Flasteen Abuqabita, Department of Data science, Princess Sumaya University for Technology, Amman, Jordan. E-mail: fal20188050@std.psut.edu.jo

Received: May 1, 2019 Accepted: May 30, 2019 Online Published: June 20, 2019

doi:10.5539/mas.v13n7p1 URL: https://doi.org/10.5539/mas.v13n7p1

Abstract

Recently, huge amount of data has been generated in all over the world; these data are very huge, extremely fast and varies in its type. In order to extract the value from this data and make sense of it, a lot of frameworks and tools are needed to be developed for analyzing it. Until now a lot of tools and frameworks were generated to capture, store, analyze and visualize it. In this study we categorized the existing frameworks which is used for processing the big data into three groups, namely as, Batch processing, Stream analytics and Interactive analytics, we discussed each of them in detailed and made comparison on each of them.

Keywords: Directed Acyclic Graph, Big Data, Interactive processing, IOT

1. Introduction

In era of the data and information technologies the growth of data are increasing day by day, in which the size of data that generated every day by the internet are exceeded two Exabyte (Inoubli,Aridhi, Mezni, Maddouri, Inoubli, Aridhi, 2018), by 2020 the growth of data in all over the world is expected to reach 44 zettabyte (Charles, 2019), tens of terabyte exchanging every day online.(Padgavankar & Gupta, 2014) Big data marketing was cost $16.1 billion in 2014. Another report in 2014 predicted that marketing would cost $32.4 billion by 2017. (Tsai, Lai, Chao, & Vasilakos, 2015).

In health sector- as an example of big data application (K. U & M. David, 2015), (Li, 2016) data are shared around the world in order to make a research to save our live,a although this area suffer from the privacy issues(Benjelloun, & Lahcen,2018) and there is lows grantee the patient privacy,to save data is also important because security breach of Big Data cost a lot, company may lose about $1.1 billion (Garg & Sharma, 2014) but the data from this sector is very important and it needs to be studied in suitable platforms to make this health systems and health tools much stronger (Zandi, Reis, Vayena & Goodman, 2019), also data mining and extracting data may give us good health knowledge and avoid the ambiguity in this area. (Li, 2016), this is the main aim of big data management. (Mishra, Dhote, S. Prajapati, & Shukla, J. P. 2015).

The three v's as we will see later refers to volume, velocity, variety is the concepts of big data, the analysis framework should be able to possess, store and manage the big data that is high volume and rapidly change in different types of data. (P., D & Ahmed, 2016) The analysis method basically based on the vary of Map/Reduce and machine learning. (Jo, Basanta-Val, Steed, Song, & Lv, 2017).

The big data platforms are needs to process or analysis big businesses (Khan,Yaqoob, Hashem, Inayat, Mahmoud Ali, Alam, Gani, 2014)., it is difficult to take care into every single detail from gathering data into analysis, but platforms can do with raw data. (Mishra, 2015) most of the computers today cannot handle all of the dataset (Tsai, 2015) in each platform has its individual focus, batch processing, real-time analytic or interactive one (P.,2016), our survey aims to divide the platforms into three types of platform.

In this study we will talk about big data development in section 1, then will present some of big data important resources in section 2, in section 3 there is a summary for knowledge discovery in the data set, the types of analytics frameworks with some examples in section 4, section 5 will be about machine learning frameworks examples and finally in section 6 there are challenges in big data.

2. Big Data Development

What is big data and how each papers defined it? Most of the paper consider at least the 3V'S- Volume, Varity Velocity (Srinuvasu, M, Koushik, & Santhosh, 2018), to called the data as big data, some consider much more V's but we will choose these ones.

The following is some of big data definitions, big data is huge amount of structured and unstructured data (Tsai et la..,2015), (Srinuvasu et al..,2018), and it may carry some fake information, data comes from different domains and different resources (Jo et al.., 2017), create big data is most than find useful information from it. (Tsai et la.., 2015), (Martin, Swennen, Depaire, Jans, Caris, & Vanhoof, 2015), Big data is not only a data, but is a concept which actually explains about the gathering huge data, organizing and analyzing the data to getting information from raw data. (Mishra et al.., 2015) it is about the way of transforming „big data" to „real value (Zhu, Li, & Tang, 2019).

The main challenge not collecting big data but How mange it? the traditional data was managing by different techniques (Jo et al.., 2017).

What is big data and how the research papers defined the big data concepts, many papers describe the data by the following three concepts:

Volume: huge size of data that created and gathered. (Bhadani, & Jothimani,2017) it is measured in Petabytes of data (Mishra et al.., 2015) and also it is raw, semi-structured or unstructured format that's difficult to realize (Li, 2016), unless using paralyze distributed systems and avoiding the bottleneck processing in one node (Tsai et la.., 2015), as we will see the new systems sends the structure to the node to solve this problem. there is a paper defined it as a volume of streaming, live streaming data which is collected from sensors (Wang, 2011).

Variety: different types of data, structured, unstructured (P., 2016) semi-structured (Alam, Sajid,alib, & Niaz, 2014), that are being generated and collected. (Bhadani, 2017) which mean different data format (Benjelloun et al..,2018), this is one of the biggest big data challenges because dealing with these type being more difficult when changing rapidly. (Srinuvasu et al..,2018), the data is growing exponentially, (Alam et al.., 2014)maps, images, text, raster and vector data are different types of data in a complex structure (Li, 2016) comes from different fields as internet (Rong, Gong, & Gao, 2019). big data platform should deal with. structured data is the data that has predefined model to store in, this data not more than 5% of all data (Bhadani, 2017).

Velocity: Is the rate of generation data (Bhadani, 2017), (P., 2016) or it is the speed of creation data from many online resources (Lee, Cho, & Kim, (2019), data streams all the times. (Zhu et al.., 2019)there is paper define it as capacity of application to process data stream generated or handle it continuously and constantly which need to be analyzed in real time or almost in real-time (Tariq & nasser,2015).

3. Big Data Important Resources

3.1 IOT

Many papers consider IOT as a major driver for big data,(Tudoran, Nicolae & Brasche, 2016) It is become a new source of big data with other social media resources(Bhadani,2017), (Gil,Johnsson, Mora, & Szymański, 2019)from every phase several aspects of the Big Data life cycle, that is sensors observe streaming, continues amount of data (Li, 2016), (P., 2016) which doubled every two years as Moore's low.

IOT also is tends as linked by technology and ecosystems (Williams, Hardy, & Nitschke, P. 2019) the article below introduces a new platform that make the process on IOT which generate from smart homes sensors. The authors propose the use of fog nodes and cloud system for data-driven services and high light the challenges of difficulties and resource needs for online and offline data from processing, storing, and organization analysis. (Yassine, Singh, Hossain, & Muhammad, 2019)

The main challenge that big data facing is knowledge gaining from IOT. The need of infrastructure for IOT device generates continuous streams of data and develop tools to extract good information from these data, analyzing them by machine learning and computational intelligence strategy's is the only solution to work with big data from IOT prospective. (P., 2016), (Chen, & Jin, 2012)

3.2 Online Social Data and Mobiles

Social media being one of the biggest resource of big data (Zhu et al.., 2019),(Li, 2016),(Jo, 2017), thoughts and records streams continuously (Garg ,2014), Facebook for example has more than one billion users over 618 million of them are regular users producing more than 500 terabytes of new data day. (Higashino, L'Heureux, Capretz, Hayes, Allison, & Grolinger, 2014) it is also an example of big data velocity (Alam et al.., 2014).

Social media is connected with mobiles by mobiles application, day after day these applications works better and people connected strongly with it, not only in social media applications but also smart cities applications. (Gil, 2019) the cheap of internet and the availability of mobile in each place make it more easy than before.

Internet of Things (IOT) and social media, has, now, become an important source of big data. The data can be captured from many areas as farming, industry, medical care, etc. (Bhadani, 2017) these are some kinds of data comes from social media calls, tweet, comment, text, net surf, browse websites, every day we do about hundreds of them and also we exchange messages in many format by the social media applications. (Wang, 2011).

3.3 Smart City

Large amount of data that comes from household appliances, smart home equipment would sense your activities at home, digital glasses would transmit what you see, and gene data are also massively gathered now. (KULITZY,1957).Devices require high cost and efficient resources, big data analytics and also the physical system. (Yassine,2019)

We can't separate the smart city from IOT and mobile technologies they are all connected with each other and all also connected with cloud computing as we will see below data receive from smart cities as a stream system in high volumes, high velocity and from various sources(Yassine,2019)

,(K.U,2015), smart city where many applications have been developed for its ecosystem an important source of complexity within the IOT.(Gil,2019)

The importance of the infrastructure comes from It supporting the capability with machine learning algorithms and improving new algorithms. The models which build by the machine learning may measure efficiently with the data in big data environment. (Jo,2017)

Smart city may be considered as an application of big data, this application allows people to have better services, customer experiences, and also to prevent and detect illness much easier than before. (K.U,2015).

4. Knowledge Discovery in Datasets

Process data to get knowledge is known as KDD (knowledge discovery in datasets), the figure 1 is show the phases.(Tsai et al.., 2015), (Gyamfi, & Williams,2019)

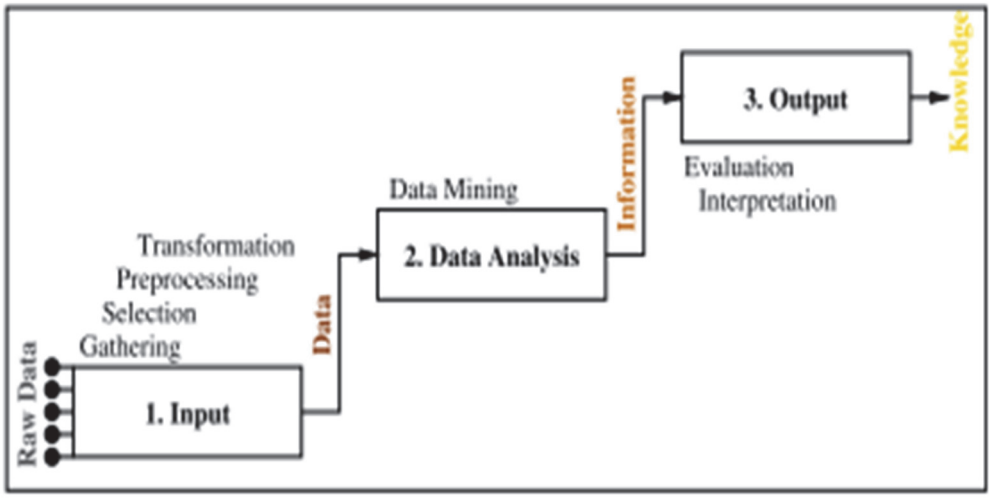

Figure 1. KDD

The most phase we care about in our survey is the analysis one, why? To help stockholder take decisions (P., 2016), traditional data process help but it cost much for example accompany make a compare between systems analysis on one terabyte of data and it found that it costs $37k using traditional database system, $5K for a database one and only $2K for Hadoop cluster (Garg, 2014) although we will care on the analysis part also which is done after the output phase when the data mining algorithms applied, some of the field specific algorithms are developed. An example is the prior algorithm which is suitable algorithm designed for the association or metaheuristic algorithms

for data mining, genetic and k-mean algorithm for clustering (Tsai et al..,2015), but this image works also if we consider the output data as a row one other area.

There are a few efficient methods work on analyzing a huge amount of data which can be dividing to many based of, like Density-based approaches, Divide and conquer approaches, Incremental learning approaches, distributed computing approaches, they can analyze data in response time(Tsai et al..,2015), also we need storage and computing area suitable for dealing with big data, cloud computing is use in the big data platforms and framework this.(Mishra, 2015) we should not forget that big data applications which use cloud computing, needs to support tools that helps scientist to analysis data (P., 2016)

Reduction methods are very important; these techniques make the data processing less expensive(Tsai,2015), because the data volume will reduce (Garg, 2014) the following is an examples:

1) PCA (PRICIPLES COMPOUNNET ANALYSIS)

It is an example of dimensional reduction method applied to reduce the input data size quickens the procedure of data analytics (Tsai, 2015), how is it work? The dimensions of point have been reduced and there is loss in data but we hope not loss the most important once (Dimensionality.R, 2010).

2) Sampling

It is also an example of reduction method that uses to increase the computing time when analysis data (Tsai, 2015), this strategy has four types they are simple random sample, stratified sampling, sampling without replacement, and sampling with replacement.

The sample size and sample type very important to fulfill the goal of finding useful data, for example the size of sample will not give an idea about the data if it very small.

The reason of sampling also is come from the data size, sometimes it is very difficult to apply the sampling on the big data, it used more for single-machine clustering (Tsai, 2015).

3) *Transformation*

This technique is transformed a feature from it format to another one Smoothing, Aggregation, Generalization, Normalization, Attribute Construction (Data.T, 2019) there is also other technologies based on machine learning and arithmetical approaches to find useful information from raw data (Tsai, 2015)

4) Clustering and Clustering

Clustering is the popular data mining techniques, the mean idea is to divide the data into k unlabelled group by tree-based, naive Bayesian and support vector machine, while classification is separated the group into k labeled group, while classification is focused on relationship, clustering care on association. (Tsai, 2015)

5. Analytics Frameworks

In this section we will discuss briefly some of the most powerful frameworks which is used for handle the huge amount and rapidly generated data, this section organized as follow, 1. Batch processing, 2. Stream analytics, 3. Interactive analytics, and for each one of them we will discuss the used frameworks.

5.1 Batch Processing

Batch computing is execution of large blocks of data which have already been stored in a database (Tudoran et al.., 2016).all at once –as a batches -at the same time (Martin et al..,2015), these blocks are store up during working time and then executed when the system become idle (Martin, 2015)in another word batch computing is execute of static job which is not in the interactive state, for instants in the smart city we could collect a huge blocks of new location data, then update all people locations all at one time. Briefly batch computing deals with jobs that start and finish (Chardonnens, Cudre-Mauroux, Grund & Perroud,2013).this type of analysis applied by Apache Hadoop, apache Dryad, Apache Mahout Frameworks. The figure bellow presents the data which have been already stored in disk and queries are submitted to the data “Query-Driven”. (Figure 2)

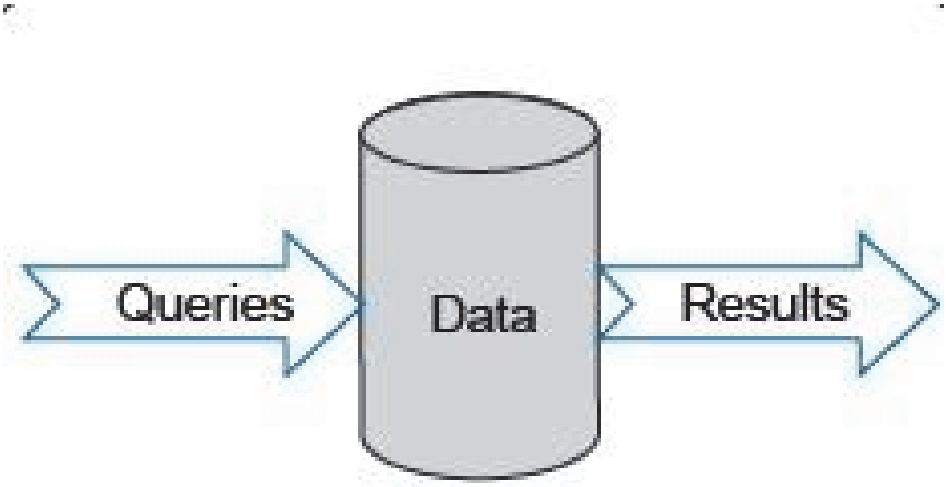

Figure 2. Batch Processing

5.1.1 Hadoop/MapRreduce

Hadoop is an open source framework/platform (Wang,2011), (Tudoran,2016). which is developed by Apache Software Foundation.(Mishra, 2015),written in java (Khan,2014), and started in 2006 (Khan,2014) it is considered to be the most widely tool for dealing with big data, where it is dedicated for processing and storing unstructured data, it inspired by Google file system that used Map/Reduce which split an application to small blocks and run it on multi nodes across a cluster (P.,2016). Hadoop consist of two core component:

1) *Hadoop distributed file system (HDFS) (*Jambunathan, Varsha, S Venkatesan, 2016), (Khan,2014), (Tudoran,2016) it's the storage layer for Hadoop it stores the data in form of small memory blocks and distribute them on the nodes across the cluster (Wang,2011),it is designed specifically to handle the huge amount of data and it is designed to be fault tolerance.
2) *Map/Reduce* It is a programming paradigm, used to processing huge amount of data on distributed systems (Higashino et al..,2014) it executes the task in parallel by distribute them across the node in the cluster (Map phase), and tack the result from the nodes and use this result as input in reduce phase. Map/Reduce consist of two functions map function and reduce function,map function used for filtering and sorting where reduce function used for aggregation and grouping, to be more clear let give you an example,suppose there is a big file contains a huge amount of numbers and the task is to find the maximum number of all numbers in the file, map function split the file into small blocks and produce a key value pair for them,and apply the task on each blokes in parallel then store the result in an intermediate file (ex. local Disk drive),reduce function used these result as input for it and aggregate it then apply the task on it and so on until fined the final result (maximum number).(Higashino, L'Heureux, Capretz, Hayes, Allison, & Grolinger, 2014) Hadoop is a high latency platform if we compare it with Drill and Storm because the use of map/reduce framework. (Tsai, 2015)

5.1.2 Apache Mahout

It is a project introduced by apache software foundation at 2009 as machine learning library which is run on top of Apache Hadoop via Map/Reduce, it is a distributed computing apache *(*Jambunathan,2016),Mahout written in java and scale, it is aimed to support machine learning technique and algorithms for scalable and intelligent big data application analysis, there are a numbers of algorithm in Mahout which are based on classification, clustering and regression techniques, however the number of available algorithms are in increase, but various of it are still missing. (Biem, Bouillet, Feng, Ranganathan, Riabov, Verscheure & Moran, 2010)

5.1.3 Apache Dryad

It is a programming framework for distributing and parallel computing, which introduced by Microsoft it is a powerful module that can increase the ability of processing and scaling from small cluster to a bigger one (Philip Chen, & Zhang,2014) ,this framework permits users to utilize the resources of a cluster for processing data programs in parallel style the cluster contains thousands of machines each one of them have many processors, they

running programs in parallel without knowing anything about each other's. The processing of programs is structured as an Directed Acyclic graph where the vertices are the computations and edges are the channels /data flows,each one of vertices have diverse input and output channels [56], The job in drayed framework is considered as a "Graph Generator" which can create any "directed acyclic graph "[56] Drayed is very powerful framework that contained complete functions including creation, monitoring and management of different jobs, resource managing, visualization, and finally fault tolerance. Recently at 2011 Microsoft has discontinued dryad since it been contributing to Hadoop. (Philip Chen, & Zhang,2014) (Table 1) present a comparison on batch processing frameworks.

Table 1. Batch computing Framework Comparison

Criteria	Framework		
	Map/reduce	Dryad	Mahout
Main Functionality	Distrubuted programming system	Distributed programming system	Processing machine learning algorithm
Job processing	Process Map/Reduce Job (many input few output)	Processing are arbitrary DAG(arbitrary input / output)	Process Map/Reduce Job
complexity	Simple	Complex (process more complicated computations)	----------
Advantages	support fault tolerant, data consistency, concurrency and,very high processing performance (Faster than Dryad)	support fault tolerant, data consistency, concurrency and,high processing performance	Provides wide collection of various machine learning algorithm,

5.2 Streams Analytics

It is a big data technology and it is a function of analytical layer of big data processing (Zhu et al..,2019). which processing a sequence of data objects that permanently generated at high rate, this sequences of data are probably unlimited and have to be processed within small period of time likely near real time response with low latency (García-Gil, Ramírez-Gallego, García, & Herrera,2017),(P., 2016),.(Padgavankar, 2014),(Garg, 2014), (Tudoran, Nicolae, & Brasche, 2016) this is also should happen in input and output process of data (Li,2016)for instance, using stream processing you can receive a warning if the temperature reached freezing status based on data stream coming from temperature sensors. However most of stream data that need this type of processing is generate from IOT (Yassine,2019), (Charles, 2019), sensors, loges, in big data environment we need to process these kind of data that generated in high rate from social media for example as records or thoughts (Tariq, 2015), we do not need to store all these data just the important of it unlike batch processing which store all the data then make processing, updating, and store the valuable part of it. In order to execute stream processing, a lot of frameworks have been developed (Lee et al.., 2019) The figure shows that in the stream computing the data are processed in motion mode before it is stored to the disk, substantially bring the data to the analysis. "Data-driven" (Figure 3)

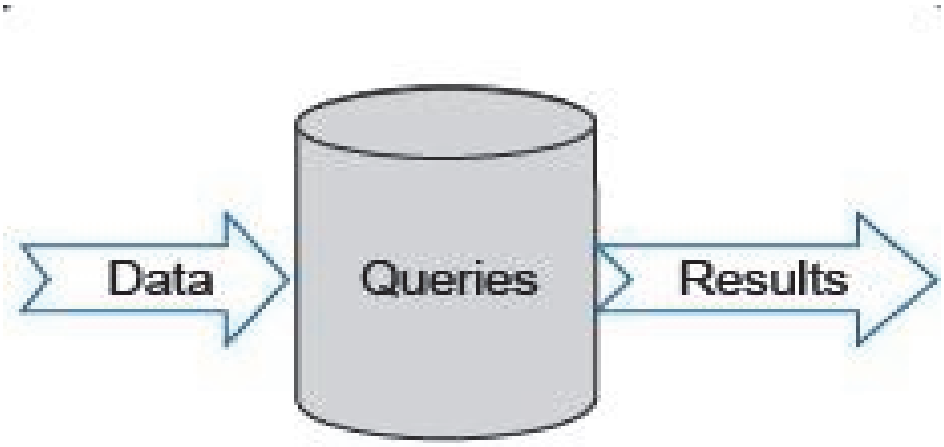

Figure 3. Stream Analytics

5.2.1 Spark

It is an open source processing framework,developed by apache software foundation and written in Scala, spark is primary used for batch processing, actually it was produce to over comes the problem of high latency of Map/reduce (P., 2016), (Yassine, 2019)

,in which it acting at the same way of Map/Reduce but the different lies in where the data is processing and when to write it on the disk,However spark is very powerful tool it can processing large scale of datasets,structured,unstructured and semi-structured in different techniques it can process data as a batches or as streams (Tudoran, 2016).spark is a fast engine analytic (Bhadani, 2017), (P., 2016), (Zhu et al.., 2019), (Tudoran, 2016),it can analyze data using sql and Machine learning libraries which is built on top of spark core, it is powerful in data integration (Wang,2011). The prominent feature of spark is having the concept of RDD (Resilient distributed datasets) (Bhadani, 2017), (Tudoran, 2016) which allow storing data in any type and processing it in memory with the concept of lazy execution this type of execution happened when an action occurred. Spark support very powerful libraries which build on top of spark core these four libraries are Sql library, MLib, Spark stream, and graph-x.

5.2.2 Apache Storm

Is a distributed stream processing framework (Padgavankar, 2014), (Tudoran, 2016) used to process a enormous amount of structured or unstructured data with high velocity rate in near -real time response (Lee et al.., (2019) Storm cluster run master /slave architecture,the master node run a daemon called “Nimbus”, where slave nodes run a ”supervisor” daemon, “Nimbus ” manage and assigned the tasks -which named as topologies – between the worker nodes(Martin, 2015). Storm is designed to be a directed acyclic graph (Tudoran, 2016) which consist of nodes and edges,the edges perform the data transfer between the nodes, where the nodes are categorized for two type,”spouts” and “blots”,” spouts”are perform the source of the data/stream where “blots” are the transformation/operation that performs on the stream of data (Inoubli, 2018) it is best choice for big stream data (P.,2016) (Bhadani,2017), Storm have the fault tolerant feature in which if one of the nodes fails master node will reassign the topology for another one (Inoubli, 2018).

5.2.3 ApacheFlink

Is an open source distributed processing frameworkused specifically to process streams but it can also deals with batch processing (Tudoran, 2016) written in java and Scala, it is dedicated to process large scale of data with low latency and high throughput (García-Gil, 2017). Flink has a lot of features the most shinning one is the ability to process data in real time mode, also it offers a fault tolerance technique (Tudoran, 2016) in order to deal with system frailer (Apache Flink, 2019). It deals with iterative processing on data coming from different sources such as Kafka and flume (García-Gil, 2017). Flink has various libraries such as Gelly that used for graph processing, FlinkMl it is a machine learning library, Flinksql and Table API is supporting Sql like language for processing queries and FlinkCEP which is library used for complex event processing and pattern detection, and continuous event stream analysis, those libraries are not fully self-contained on flink core rather it embed on API,flink use “Dataset API” for processing batches while “DataStream API “for stream (Apache Flink, 2019).

5.2.4. Apache Sazma

Is an open source distributed stream processing framework (Tudoran,2016), written in java and Scala, it was developed by LinkedIn then detonate to Apache software foundation sazma dedicated to process huge amount of massages as they are received one by one with low latency and fault tolerant strategies (García-Gil, 2017) The main goal of developing sazam is to serve the use cases that requires processing with very high throughput and no single frailer –the data cannot be loss - (García-Gil, 2017) The type of streams that processed by sazam are massage streams,it processes the stream form the consumers in which sazam relies on Apache Kafka for massaging and it build on top of Hadoop YARN for parallel distribute processing and managing the recourses [18].

5.2.5 IBM Infosphere Streams

It is a framework introduced by IBM to deal with unbounded streams of data at high performance,it can be scaled up to a huge number of nodes and deal with both type structured and unstructured streams of data [(Biem, 2010)·IBM streams can process complex data streams at high rate with very low latency.It is offer a stream process language (SPL) which allow users to developed stream application using high level programming language.The type of stream data is tuples which is a sequence of immutable key-value pairs that can hold structured or un structured data,these tuples are transfer through data flow graph (DAG) in which the edges present the transfer line or the stream line which carry the tuples and nodes present the operations that will process the stream data (sort, join,filter) (Philip Chen,2014), (Table 2) present a comparison between common Stream processing Frameworks.

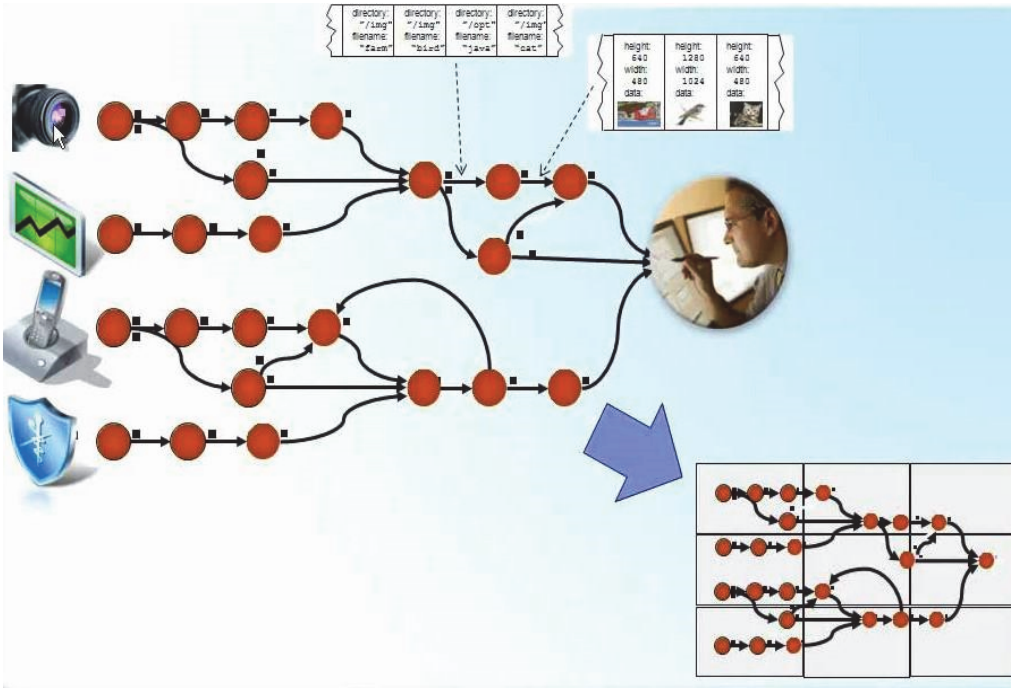

Figure 4. IBM Infosphere Stream

Table 2. Stream Analytics Frameworks Comparison

Criterea	Framework						
	Apache Spark /spark streaming	**Apache Storm**	**Apache Sazma**	**Apache Flink**		**IBM streams**	**Infphosfer**
Stream Type	DStream(Discretized	Tuple	Messages	Data	Steams	Tuple	

	streams)			/event	
Streaming strategy	Process streams as micro batches	Process streams item by item	Process streams item by item	Process streams item by item	Process event streams items by item
Latency	A few seconds	Sub-second	Very low	Very low(real time)	Very low
Processing Guarantee	Exactly-once	At least-once	At least-once	Exactly-once	At least-once
Programming languages support	Java-Scala-Python,R	JVM languages (java,scala),python	Only JVM(Java,scala)	Java,SQL	Java,c++,c,scala,python
Throughput	Very high	High	Very high	Very High	Very High
Source of streams	HDFS, Kafka,any DBMS	Spouts	Consumer Messages from Kafka	Kafka, HDFS, any DBMS	DBMS, GHDFS, XML file
storing state across steams	Stateful	Stateless	Stateful	Stateful	Stateful
Batch processing support	Yes	NO	Yes	Yes	No
schdular	Mesos,Yarn Stand alone	ZooKeper	YARN Stand alone	ZooKeper	YARN
Advantages	• Support both batch and stream mode processing • Guarantee arriving streams in order	• provides near real time processing • very low latency at high stream rate	• Guarantee arriving messages in order	• Guarantee arriving messages in order • can handel event time	• provides high scalability and performance • Guarantee arriving messages in order
Disadvantages	• High memory usage • Not suitable for sensitive data that require very low latency stream processing	• Does not guarantee arriving stream in order • process at least one stream this imply data duplicate	• Not fixable in supporting programming languages,only JVM • data might be duplicated	• Fink still small not very famous project	• process at least one stream this imply data duplicate.

5.3 Interactive Analytic

It is an interactive analysis processes on data, which enable querying the streams of Big Data to meet the scaling Varity types of data with size of terabytes interactively in response time (Garg, 2014), The user is directly linked to the computer and can relate with the system, the data can be matched, revised, and analyzed in graphic or tabular schema or both at the same time, the most tools used this type are Apache drill. The main challenge in interactive processing come from the small tasks, these task is divided in Map/Reduce to a smaller one, which is not efficient to deal with. (Higashino et al.., 2014)

5.3.1 Apache Drill

It is an open source distributed query engine framework used for handle interactive analysis of huge datasets (Hausenblas, & Nadeau, 2013), Drill is inspired by Google's Dermel and developed by Apache software foundation (P., 2016),the main aim of Apache Drill is to process the ad-hoc queries in a very low latency mode (Apache. D, 2019), it is intended to handle up to petabytes of data extended across several thousands of servers, in very fast speed which required by business intelligent analytic environment (Hausenblas, 2013). Drill is specifically focus on non-relational data base but it can also work with relational one it support a various NOSQL data base like Hbase,MongoDB.Amazon s3,HDFS and others.It combatable with BI/tools like Tableau and Microstage,it can run on Hadoop or on any distributed cluster (Apache. D, 2019).

5.3.2 Impala

In 2015 Impala is evaluated, impala is an open source SQL engine,which runs on hundreds of machines as distributed architecture,it is provide an interactive process.Impala support massively parallel processing (MPP)engine,when compared with Hive and Spark SQL it is much faster than both.Impala is providing suitable performance using scans, aggregations, and joins to give queries, it consider as low latency apache and fault tolerance, the stored data in impala is in Parquet files, this apache doesn't use Hadoop but it install set of daemons on each Data Node for local processing,this strategy is to avoid bottleneck problem. (Rodrigues, Santos,& Bernardino, 2019), The run time of impala is very short run time it is supported HiveQL, although it is compatible with BI it is not a visualization apache data. In comparing with apache Drill, both of them shares a lot of similarities, both are MPP SQL engines and inspired by Google's Dermel, both are powerful and support low latency processing and fault tolerant, the most prominent difference is, Drill is schemaless, it does not need pr-defined schema, while Impala required schema to be pre-defined (Charles, 2019).

6. Frameworks for Machine Learning

Many papers have studied machine learning, because it important for building intelligent decision(Higashino et al..,2014) ,it is important find systematic ways to working with big data like dealing with industrial-scale problems, there is a general-purpose proposed framework which systematically addresses data- and model-parallel challenges in large-scale ML by many ML programs that is fundamentally optimization-centric and fault-tolerance, dynamic scheduling.etc (Jo, 2017). Machine learning used with statistical strategy in order to analysis data, like using data mining (Tsai, 2015)

The design of Map/Reduce now is compatible with batch processing, the future needs a new version of its can be integrate with machine learning. (Table 3) present the Usageof common Machine learning analytics frameworks.

Table 3. The Usage of common Machine learning analytics frameworks.

Criteria	**Frameworks and machine learning**			
	Hadoop	***Spark***	***Flink***	***Storm***
ML library support	Mahout	Spark-MLib	FlinkML	Trident-ml
Used Algorthims	• **Clustering** (K-Means,/fuzzy k-Means, Canopy**,**Dirichlet,Mean-Shift.) • **Classification /LogisticRegression** (NaiveBayes, SVM,Logistic and linear regression)	•**Clustering** (K-means , streaming k-means, Gaussian matrix) • **Classification /LogisticRegression** (NaiveBayes, SVM,Logistic and linear regression,Rando	• **Clustering (**K-Nearest neighbors join) • **Classification /Logistic Regression (** SVM**)** , some linear regression algorithms) • Collaborative filtering	•**Clustering** (K-means) •**Classification /Logistic Regression** Perceptron,, winnow, Passive-Aggressive,AROW)

	• Callaborative filtering algorthim (ALS) (Ingersoll,2009)	m Forests,isotonic) •Collaborative filtering algorithm like (ALS) (Spark.Mlib,2019)	algorithm like (ALS) **(**Flink.Ml,2019)	(Trident-ml, 2015)
Coverage of Algorithms	High	Very High	Low(still growing)	low
Integrattion	Build on top of Hadoop	Build on top of spark core & compatible with others spark library	Build on Flink	Trident-ml build on Trident which is abstraction build on storm

7. Challenge of Big Data

7.1 Cloud Storage

Cloud computing is a huge computing concepts which access a massive amount of data so that many clients can use it, it is a paradigm for sharing on-demand services and computing resources via a Varity broad network access(Gamal, Rizk, Mahdi,& Elhady,2018),(Garg,2014).The development of virtualization software and technologies make the systems to be a platform (Alkhashai, & Omara, 2016), it is like a true physical systems in which the virtual computer has infrastructures including memory-huge touchy information is stored (Trident-ml, 2015), (Garg, 2014) -processors, disk and operating systems, these virtual computers are mean of cloud computing which is the most important technique used in big data,clouds also links the physical machine with virtual ones (Garg,2014)

In order to apply big data technology on cloud computing we have to integrate distributed map/reduce framework and cloud computing to provides powerful Petabyte scale computing.

The advantages of cloud computing on big data is providing scalability and extensibility on storing huge amount of data but there is a critical challenge that face big data in cloud, which is the data that lies in cloud is shared by others, also the cost and time required to load large amount of data on the cloud, the difficulty of handling the distribution processing over the cloud, this make the big data a good thing that push the cloud environment for high level development.

7.2 Security and privacy

Security is a big data challenge (Khan, 2014), (Jo, 2017), (Padmavalli, 2016), because of many reasons, first the big data platform contains many types of data structured, unstructured and semi-structured each has different need to be secure, security can't be sure. (Jo, 2017) Second the parallel in processing data is a new challenge that we should grantee the secure of the data. Third challenge Is come from real time analytic that how can the platform save the privacy and make a real time analysis, and also big data is stored in a cloud as a distributed file that make the security difficult more and more. (Benjelloun, 2018)

The secure platform should be flexible with big data so it can integrate with the new technologies at the same time it should take care with the above security problems, and it should be comparable with big data the traditional techniques (encoding for example) slow the performance. (Benjelloun, 2018)

In MapReduce responsibility is to save mange data in mapper and reducer,the solution has been suggested to createan Accountable MapReduceso This solution develops a set of auditors to perform accountability tests on the mappers and reducers in real-time by monitoring the results of these tests, any unauthenticated assess on mappers or reducers can be detected (Higashino et al.., 2014) .

Data replication in big data platforms have displayed some security faults, the generation of multiple copies, put the data in risk, the policies define the data that are stored, process, analyzed, but still there is now grantee to save this data (Khan, 2014).Those policies should be able to efficiently determine what is allowed or not, and keep track in each layer of data processing from data to service in the big data systems. (Jo, 2017)

8. Conclusion

Finally, the magic word is not the big data but how to find knowledge from it, There is no rubbish data as we think before, every small piece of data is important, outliers may be the most important thing to study, analysis using machine learning needs to improve its techniques and sub sets which use to learn the machine that is the key word, and of cores the real time processing is important we actually hope to find the best frameworks to introduce for stockholders so they can find best decision at the specific time without drop any piece of data. In this survey we discuss what does big data mean and we present the most affected sources which responsible for generate data volume, the KDD phases, were also presented, we discuss briefly the most Known analytics framework and categorized it into three classes, Batch analytic, stream analytic and interactive analytics, finally some of dig data challenges have also been presented.

References

Alkhashai, H. M., & Omara, F. A. (2016). An enhanced task scheduling algorithm on cloud computing environment. *International Journal of Grid and Distributed Computing*. https://doi.org/10.14257/ijgdc.2016.9.7.10

Apache, D. (2019). Retrieved from https://drill.apache.org/

Apache, F. (2019). Retrieved from https://flink.apache.org/

Apache Spark-Mlib. (2019). Retrieved from https://spark.apache.org/docs/1.4.1/mllib-guide.html

Benjelloun, F. Z., & Lahcen, A. A. (2018). Big Data Security. *Web Services*, 25–38. https://doi.org/10.4018/978-1-5225-7501-6.ch003

Bhadani, A., & Jothimani, D. (2017). Big Data: Challenges, Opportunities and Realities.https://arxiv.org/pdf/1705.04928

Biem, A., Bouillet, E., Feng, H., Ranganathan, A., Riabov, A., Verscheure, O., … Moran, C. (2010). *IBM infosphere streams for scalable, real-time, intelligent transportation services*. 1093. https://doi.org/10.1145/1807167.1807291

Chardonnens, T., Cudre-Mauroux, P., Grund, M., & Perroud, B. (*2013*). Big data analytics on high Velocity streams: A case study, *IEEE International Conference on Big Data*, Silicon Valley, 784-787. https://doi.org/10.1109/BigData.2013.6691653

Charles. (2019). Retrieved from https://hackernoon.com/a-few-facts-to-take-into-account-about-big-data-market-growth-eaf7c993f0fd

Chen, X. Y., & Jin, Z. G. (2012). Research on Key Technology and Applications for Internet of Things. *Physics Procedia*, *33*, 561–566. http://dx.doi.org/10.1016/j.phpro.2012.05.104

Christophe, Drayed-Microsoft Research. (2004). Retrieved from https://www.microsoft.com/en-us/research/project/dryad/

Data Transformation. (2019). Retrieved from http://www.lastnightstudy.com/Show?id=42/Data-Transformation-In-Data-Mining

Dimensionality Reduction, Wikipedia. (2010). Retrieved from https://en.wikipedia.org/wiki/Dimensionality_reduction

Flink, M. L. (2019). Retrieved from https://ci.apache.org/projects/flink/flink-docs-stable/dev/libs/ml/

Gamal, M., Rizk, R., Mahdi, H., & Elhady, B. (2018). Bio-inspired load balancing algorithm in cloud computing. *Advances in Intelligent Systems and Computing*. https://doi.org/10.1007/978-3-319-64861-3_54

García-Gil, D., Ramírez-Gallego, S., García, S., & Herrera, F. (2017). A comparison on scalability for batch big data processing on Apache Spark and Apache Flink. *Big Data Analytics*, *2*(1). https://doi.org/10.1186/s41044-016-0020-2

Garg, P., & Sharma, V. (2014). An efficient and secure data storage in Mobile Cloud Computing through RSA and Hash function. *Proceedings of the 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques, ICICT 2014*. https://doi.org/10.1109/ICICICT.2014.6781303.

Gil, D., Johnsson, M., Mora, H., & Szymański, J. (2019). Review of the Complexity of Managing Big Data of the Internet of Things. *Complexity*, *2019*, 1–12. https://doi.org/10.1155/2019/4592902

Gyamfi, A., & Williams, I. (2019). Big Data and Knowledge Sharing in Virtual Organizations. i(January). https://doi.org/10.4018/978-1-5225-7519-1.ch004

Hausenblas, M., & Nadeau, J. (2013). Apache Drill: Interactive Ad-Hoc Analysis at Scale. *Big Data*, *1*(2), 100–104. https://doi.org/10.1089/big.2013.0011

Higashino, W. A., L'Heureux, A., Capretz, M. A. M., Hayes, M., Allison, D. S., & Grolinger, K. (2014). *Challenges for MapReduce in Big Data*. https://doi.org/10.1109/services.2014.41

Higashino, W. A., L'Heureux, A., Capretz, M. A. M., Hayes, M., Allison, D. S., & Grolinger, K. (2014). Challenges for MapReduce in Big Data. https://doi.org/10.1109/services.2014.4

Ingersoll, G. (2009). Introducing Apache Mahout intelligent applications, 1–18.

Inoubli, W., Aridhi, S., Mezni, H., Maddouri, M., Inoubli, W., Aridhi, S., … Compara-, E. N. A. (2018). *A Comparative Study on Streaming Frameworks for Big Data To cite this version : HAL Id : hal-01835437.HAL Id : hal-01835437*.

International Journal of Computer Science and Mobile Computing A Review on the Role of Big Data in Business. In *International Journal of Computer Science and Mobile Computing* (Vol. 3). Retrieved from www.ijcsmc.com.

Jambunathan, V., & Venkatesan, S. (2016). A Review on Big Data Challenges and Opportunities. *International Journal of Latest Technology in Engineering Management & Applied Science* V(Xi): 2278–2540. Retrieved from www.ijltemas.in

Jo, M., Basanta-Val, P., Steed, A., Song, H., & Lv, Z. (2017). Next-Generation Big Data Analytics: State of the Art, Challenges, and Future Research Topics. *IEEE Transactions on Industrial Informatics*. https://doi.org/10.1109/tii.2017.2650204

K. U, J., & David, M. J. (2015). Issues, Challenges and Solutions: Big Data Mining. https://doi.org/10.5121/csit.2014.41311

Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Mahmoud Ali, W. K., Alam, M., … Gani, A. (2014). Big data: Survey, technologies, opportunities, and challenges. *Scientific World Journal*. https://doi.org/10.1155/2014/712826

Lee, Y., Cho, H., & Kim, S. (2019). Advances in Computer Communication and Computational Sciences. In *Advances in Computer Communication and Computational Sciences* (Vol. 760). https://doi.org/10.1007/978-981-13-0344-9

Li, Songnian et al. (2016). Geospatial Big Data Handling Theory and Methods: A Review and Research Challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*. https://doi.org/10.1016/j.isprsjprs.2015.10.012

Martin, N., Swennen, M., Depaire, B., Jans, M., Caris, A., & Vanhoof, K. (2015). Batch procebing: Definition and event log identification. *CEUR Workshop Proceedings*, *1527*(December), 137–140.

Mishra, S., Dhote, V., S. Prajapati, G., & Shukla, J. P. (2015). Challenges in Big Data Application: A Review. *International Journal of Computer Applications*. https://doi.org/10.5120/21651-4962

P., D., & Ahmed, K. (2016). A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools. *International Journal of Advanced Computer Science and Applications*. https://doi.org/10.14569/ijacsa.2016.070267

Padgavankar, M. H., & Gupta, S. R. (2014). *Big Data Storage and Challenges*. Retrieved from www.ijcsit.com

Padmavalli, M. (2016). Big Data: Emerging Challenges of Big Data and Techniques for Handling, *18*(6), 13–18. https://doi.org/10.9790/0661-1806041318

Philip Chen, C. L., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, *275*, 314–347. https://doi.org/10.1016/j.ins.2014.01.015

Rodrigues, M., Santos, M. Y., & Bernardino, J. (2019). Big data processing tools: An experimental performance

evaluation. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(2), 1–24. https://doi.org/10.1002/widm.1297.

Rong, M., Gong, D., & Gao, X. Z. (2019). Feature Selection and Its Use In Big Data: Challenges, Methods, and Trends. *IEEE Access*, *7*, 1–1. https://doi.org/10.1109/ACCESS.2019.2894366

Societal, Economic, Ethical and Legal Challenges of the Digital Revolution: From Big Data to Deep Learning, Artificial Intelligence, and Manipulative Technologies1. *Magyar n??OrvosokLapja*, *20*(4–5), 216–223. https://doi.org/10.2139/ssrn.2594352

Srinuvasu, M. A., Koushik, A., & Santhosh, E. B. (2018). Big Data Challenges and Solutions. *International Journal of Computer Sciences and Engineering*, *5*(10), 250–255. https://doi.org/10.26438/ijcse/v5i10.250255.

Tariq, R. S., & Nasser, T. (2015). Big Data Challenges." *Computer Engineering & Information Technology*. https://doi.org/10.4172/2324-9307.1000133.

Trident-ml. (2015). Retrived from https://github.com/pmerienne/trident-ml

Tsai, C. W., Lai, C. F., Chao, H. C., & Vasilakos, A. V. (2015). Big data analytics: a survey. *Journal of Big Data*, *2*(1). https://doi.org/10.1186/s40537-015-0030-3

Tudoran, R., Nicolae, B., & Brasche, G. (2016). *Data Multiverse: The Uncertainty Challenge of Future Big Data Analytics*. https://doi.org/10.1007/978-3-319-53640-8

Tudoran, R., Bogdan, N., & GötzBrasche. (2016). Data Multiverse: The Uncertainty Challenge of Future Big Data Analytics. https://doi.org/10.1007/978-3-319-53640-8_2

Wang, A. (2011). Retrieved from https://umbrant.gitlab.io/blog/2011/dryad_mapreduce_review.html

Williams, S. P., Hardy, C. A., & Nitschke, P. (2019). Configuring the Internet of Things (IoT): A Review and Implications for Big Data Analytics Configuring. *Proceedings of the 52nd Hawaii International Conference on System Sciences*, (January), 5848–5857. Retrieved from http://hdl.handle.net/10125/60020

Yassine, A., Singh, S., Hossain, M. S., & Muhammad, G. (2019). IoT big data analytics for smart homes with fog and cloud computing. *Future Generation Computer Systems*, *91*(September), 563–573. https://doi.org/10.1016/j.future.2018.08.040

Zandi, D., Reis, A., Vayena, E., & Goodman, K. (2019). New ethical challenges of digital technologies, machine learning and artificial intelligence in public health: A call for papers. *Bulletin of the World Health Organization*, *97*(1), 2. https://doi.org/10.2471/BLT.18.227686

Zhu, J. Y., Li, V. O. K., & Tang, B. (2019). A five-layer architecture for big data processing and analytics. *International Journal of Big Data Intelligence*, *6*(1), 38. https://doi.org/10.1504/IJBDI.2019.10018535