ALGORITHMS AND TOOLS OF BIG DATA: A BIBLIOGRAPHIC REVIEW

Carlos Andrés Cortés Núñez

UNIVERSIDAD CATÓLICA DE COLOMBIA

FACULTAD DE INGENIERÍA

PROGRAMA DE INGENIERÍA DE SISTEMAS

BOGOTÁ D.C.

2015

ALGORITHMS AND TOOLS OF BIG DATA: A BIBLIOGRAPHIC REVIEW

Carlos Andrés Cortés Núñez

Trabajo de Grado

Director

Mario Martínez Rojas

M.Sc.

UNIVERSIDAD CATÓLICA DE COLOMBIA

FACULTAD DE INGENIERÍA

PROGRAMA DE INGENIERÍA DE SISTEMAS

BOGOTÁ D.C.

2015

Nota de Aceptación

Aprobado por el comité de grado en cumplimiento de los requisitos exigidos por la Facultad de Ingeniería y la Universidad Católica de Colombia para optar al título de ingeniero de Sistemas.

_____

Ingeniero Mario Martínez
Director

_____

Ingeniero Manuel Báez
Revisor metodológico

Bogotá, 18, Noviembre, 2015.

Table of contents Pag

List of tables Pag

List of Images Pag

List of Annexes

Annex 1. 2007-Combinatorial Algorithm for Compressed Sensing

Annex 2. 2007-Near-Optimal Algorithm for estimating the Entropy of a Stream

Annex 3. 2009-Data Stream Algorithms200-Barbados Workshop on Computational Complexity

Annex 4. 2010-Data-Parallel Algorithms and Techniques

Annex 5. 2011-Introduction to IO Efficient Algorithms-MIT

Annex 6. 2011-Introduction to Parallel Algorithm Analysis

Annex 7. 2011-Introduction to Streaming Algorithms

Annex 8. 2011-IO-Efcient Merge Sort

Annex 9. 2012-Clustering-Stanford University

Annex 10. 2012-Graph Streams algorithm

Annex 11. 2012-IO-Algorithms

Annex 12. 2012-IO-Algorithms2

Annex 13. 2012-MapReduce Algorithms.

Annex 14. 2013-Algorithm_and_approaches_to_handle_large_Data-_A_Survey

Annex 15. 2013-Algorithmic Techniques for Big Data

Annex 16. 2014-Genetic-Algorithm-and-its-application-to-Big-Data-Analysis

Annex 17. 2014-Parallel Algorithms for Geometric Graph Problems

Annex 18. Date-Clustering-Stanford University

Annex 19. External-Memory Graph Algorithms

Annex 20. Fast Parallel GPU-Sorting Using a Hybrid Algorithm

Annex 21. Fundamental Parallel Algorithms for Private-Cache Chip Multiprocessors

Annex 22. IO complexity of graph algorithms

Annex 23. K-Center-Algorithm

Annex 24. K-means Advantages of Careful Seeding

Annex 25. Lower Bounds in streaming (algorithm)

Annex 26. Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High

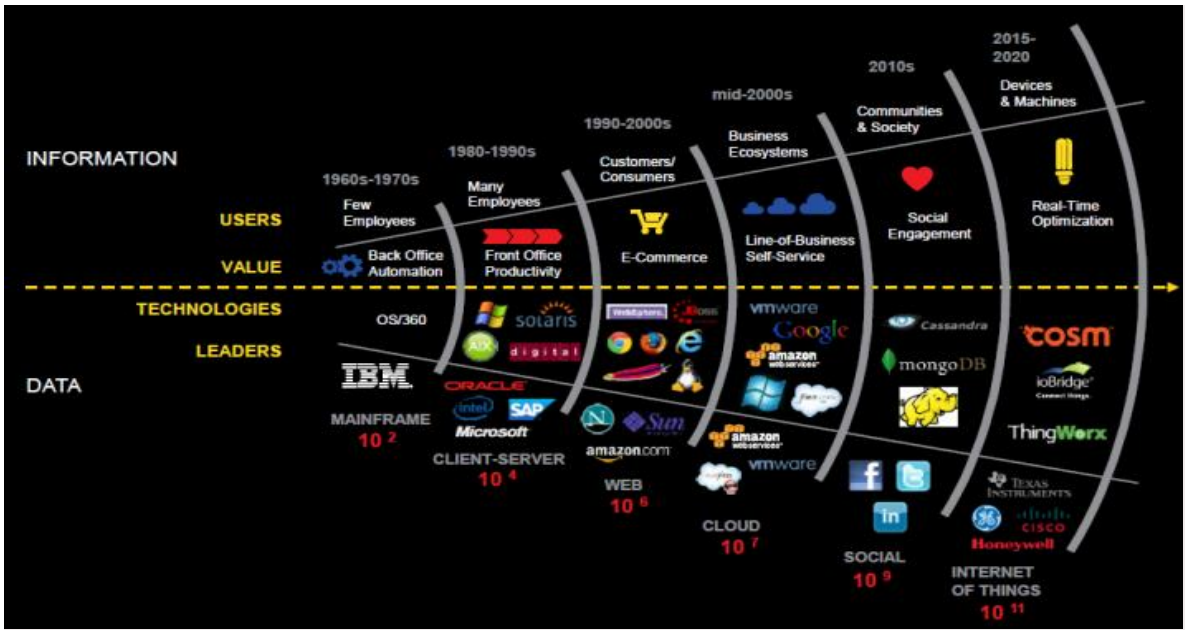
Annex 27. Scalable KMeans++

Annex 28. Streaming Graph Algorithms-MIT

INTRODUCTION

Big data is among us, it's present on all parts of our life and like time it's never stops growing. Having in mind that all data will tend to become big data is necessary to change the customs ways of dealing for two simple reasons, the increase in the cost and the decrease of effectives. A quick glance in the evolution of data and information shows the leaders and the technologies that have been developing a way to work with big data, all this having as an essential factor the value given to the user.

Image 1. Evolution of information and data



Ravi Kalakota ph. D. Big Data [on line]. [August 1 2015]. Available on Internet: < URL: http://practicalanalytics.co/2013/06/11/nsa-prism-the-mother-of-all-big-data-projects/>

Big data is a well-known term present on different fields with huge appliances on many levels in organizations and marketing fields. With the years, the internet of things and "business operations that have been gathering and storing information

from their customers, suppliers and employees"[1], have contribute for an exponential growth of data through time. "With the use of specialized algorithms big data can be analyzed computationally to reveal patterns, trends and associations"[2] to get value from data, while algorithms were evolving, even more data was collected and stored onto a point that without the evolution of the algorithms parallel to the grow of information, analyzing the data would be less efficient to impossible, giving birth to a new scientific paradigm know as data intensive scientific discovery (DISD).

Different sources like lectures, courses, scientific documents and articles, where gathered and cited for creating a state of the art meta-analysis for algorithm in big data. With the use of a framework for analyze and deliver the results of combining different studies having a predefined set of rules that defines methods of data extraction.

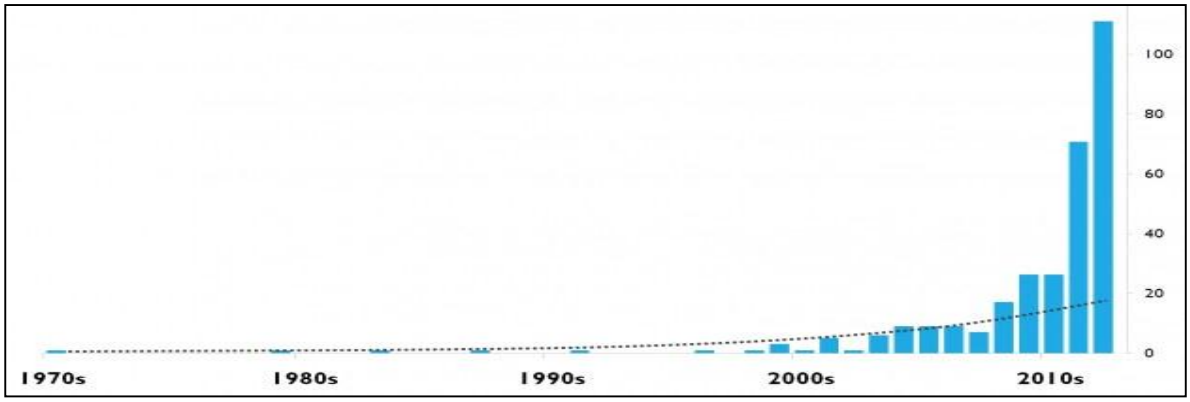
According to Needle in a Data stack Report from McAfee most of current organizations in the world are not prepared to face the security breaches and risk that comes when handing big data, most of them may have the latest antivirus software and tracing log functionally for intern transactions, but when it comes to big data a new set of rules must be follow to unsure security on big data.

In scientific literature and studies the number of papers per year has considerably grown since the middle 2000s, exploring new possibilities with hardware components advancements and digital models for storing and handling data.

[1] McKinsey & Company. Big Data [on line]. [August 1 2015]. Available on Internet : < URL: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation>

[2] OXFORD DICTIONARIES. Big Data [on line]. [August 1 2015]. Available on Internet : < URL: http://www.oxforddictionaries.com/us/definition/american_english/big-data >

Image 2. Number of “Big data” papers per year.



Scopus. Big Data [on line]. [August 1 2015]. Available on Internet: < URL: http://www.researchtrends.com/issue-30-september-2012/the-evolution-of-big-data-as-a-research-and-scientific-topic-overview-of-the-literature/>

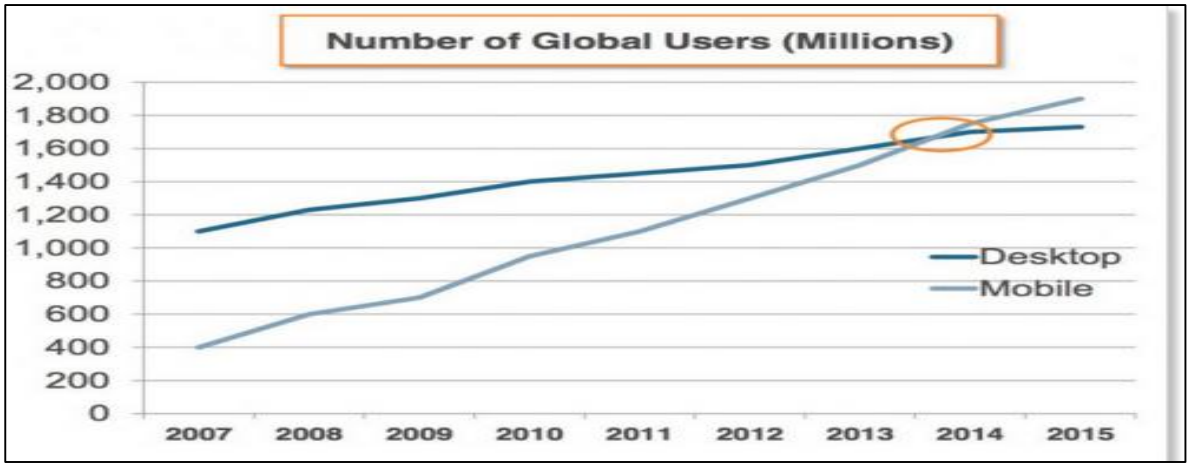
1. PROBLEM STATEMENT

A new scientific paradigm is born as data intensive scientific discovery, also known as Big Data problems. “A large number of fields and sectors, ranging from economic and business activities to public administration, from national security to scientific researches in many areas, involve with Big Data problems.”[3] Already the market is showing great business opportunities for capable companies in terms of data analysis and data visualization as a juncture to be a key differentiator in decision making and the strategies for competitive markets.

With the internet of things the explosion of information is considered a problem for traditional ways for saving, organization and analyzing it, leave behind opportunities and therefore not fully taken into consideration for a real business scenario. Considering the latest statistics on mobile users alone (that have already overcome the use of desktop computer due to the easy acceptance and easy acquisition) we can get a hold of the upcoming opportunities for big data management, implying the involvement of algorithms and tools that have evolved with the need to catch the rapid increase of data.

[3] Chen, Philip. Zhang, Chun-Yang. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Macau, China: University of Macau. 2013. p. 1.

Image 3. Number of Desktop and Mobile Global Users



ComScore. Number of Global users [on line]. [August 1 2015]. Available on Internet: < URL: http://www.smartinsights.com/mobile-marketing/mobile-marketing-analytics/mobile-marketing-statistics/>

2. OBJECTIVES

2.1 GENERAL OBJECTIVE

Characterize algorithms and tools for Big Data from meta-analysis of publications according to their impact factor.

2.2 SPECIFICS OBJECTIVES

- Specify the dependent and independent variables associated with the characterization for Big Data algorithms.
- Identify the variability and representativeness of the characteristics of algorithms for Big Data according to a systematic record of publications.
- Identify areas of open research.

3. BIG DATA

3.1 THEORETICAL

"Big Data is defined as extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions"[4], where it consists from text files to videos and any other information that thought analysis can be a key differentiator for businesses seeking to gain competitive advantage. Alone big data means a large amount of data that with traditional searching algorithms does not result in useful information for a specific purpose, with the coming of new ways to engage the analysis various algorithms stand up by their agility and optimal performance on large amount of data.

When talking about big data is pretty common to hear about the four V, which reefer to volume, variety, velocity and veracity from IBM Big Data & Analytics Hub. The words are pretty self-explanatory where volume is measured in bytes and means the scale of data, variety is the different form of data coming from text files to videos, the velocity that refers to the analysis of stream data and veracity deals with the uncertainty of the data. These four V's are a perfect characterization of what big data can be look at it, dividing into four different aspects to relate to a specific measurement where all data can be identify and organized.

3.1.1 Volume. According to Fortune magazine, up to 2003, the human race had generated just 5 Exabytes (5 billion Gigabytes) of digital data. That amount of data

[4] OXFORD DICTIONARIES. Big Data [on line]. [August 1 2015]. Available on Internet : < URL: http://www.oxforddictionaries.com/us/definition/american_english/big-data >

was produced in just two days in 2011, and by 2013 we were generating more than that every 10 minutes. With the total volume of data stored on the world's computers doubling every 12 to 18 months, we truly live in the age of big data.[5]

3.1.2 Variety. One of defining characteristics of big data is the variety of sources and content involved, which opens up a whole range of new opportunities to create value from this torrent of bits and bytes. Some sources are internal to the enterprise, like the list of customer purchases generated by a transaction processing system. Other sources, like tweets, geo-location data and public records, are external. And the data comes in different formats, some of it structured like conventional database entries, some of it semi-structured like images with metadata, and the rest completely unstructured like text, graphics, raw imagery (e.g. satellite imagery), audio files or streaming video.[6]

3.1.3 Velocity. To get the most out of big data, all this content must be processed dynamically to generate immediate results. In other words, velocity is of the essence. With modern advances in analytical algorithms (Big analytics) and data transmission infrastructures, it is now becoming possible to feed data into business processes on the fly. Certain kinds of data are only useful if the content can be analyzed as soon as the data has been generated. Online fraud, for example, needs to be detected straight away, and streaming video from traffic monitoring cameras needs to be analyzed constantly to determine road traffic patterns in real time.[7]

3.1.4 Veracity. Big data also changes the value of data, both in a monetary sense and in terms of its usefulness. Data quality in a given situation — in other words

[5] Thales Group.Big Data [on line]. [August 18 2015]. Available on Internet : < URL: https://www.thalesgroup.com/en/worldwide/big-data/four-vs-big-data>
[6] Ibid., p. 1.
[7] Ibid., p. 1.

the integrity and veracity of the information — depends on two factors. First, the data may be incomplete or incorrect, or structured in a way that makes it hard to analyze, in which case the credibility of the source and the quality of the content need to be verified. Second, organizations clearly store vast quantities of data, but what is much less clear are the types of data that are worth analyzing. Preliminary investigations may be needed to root out the weak signals from the noise and clutter, and identify the types of data with the potential to become "business drivers". Defining the objectives as early as possible is the best way to avoid expending resources on analyzing data with little operational value. [8]

3.2 UNSTRUCTURED DATA

"One of the many problems to resolve in big data is to take advantage of all the unstructured data that organizations have, dealing with more than 80% of their information unstructured."[9]

"Text mining and neuro-linguistic programming (NLP) helps companies tap unstructured data to analyze customer feedback, such as emails, blog posts, customer comments and queries, so they can extract meaningful insights. For instance, these capabilities have the power to:"[10]

- Manage real-time data feeds.

[8] Ibid., p. 1.
[9] BLUE OCEAN. UTILIZING CUTTING-EDGE UNSTRUCTURED DATA ANALYTICS TO SOLVE TODAY'S BUSINESS CHALLENGES [on line]. [August 19 2015]. Available on Internet : < URL: http://www.blueoceanmi.com/blueblog/utilizing-cutting-edge-unstructured-data-analytics-to-solve-todays-business-challenges/>
[10] Ibid., p. 1.

- Convert unstructured text into structured data.
- Reduce manual efforts and turnaround time significantly.
- Scale solutions to continually manage voluminous data.

3.2.1 Artificial Intelligence. "AI has been used with research methodologies in medicine, robot control, defense and the remote sensing industries. One key distinction is that AI diagnostic methods emphasize the algorithm—as opposed to standard predictive analytics techniques—and are able to predict business scenarios and trends by leveraging historic data"[11]

3.2.2 Machine Learning. "Machine learning enables rapid processing of large amounts of customer-centric data, including customer conversations in the form of calls, email and chats. It has the ability to handle huge volumes of data, enabling companies to incorporate multiple data streams without compromising the accuracy or relevance of the results. For instance, companies are using machine learning to build a single view of customers so they can develop the next best offer or predict churn."[12]

3.2.3 Sentiment Analytics. "Sentiment Analytics allows analysts to tap into the world of social media and combine it with cutting-edge algorithms such as NLP in order to decode complex human interactions and evolving languages to understand what customers are really saying. For instance, movie makers can use algorithms and computational models to learn about their audiences so they can script a movie that will have maximum impact."[13]

[11] Ibid., p. 1.
[12] Ibid., p. 1.
[13] Ibid., p. 1.

3.2.4 High-Performance Computing. High-Performance Computing has the ability to handle huge volumes of data on an almost real-time basis, thereby transforming the way data used to be analyzed. As unstructured data consists of nearly 80 percent of a company's total data, the boundaries between unstructured data analytics and high-speed computing powers continue to blur. The large volume of unstructured data combined with its ever-increasing velocity demands that companies move away from traditional computing systems. Further, as algorithms become more complex in nature, executing them requires higher computing power to be effective. [14]

3.3 METAHEURISTIC

"Metaheuristic optimization deals with optimization problems using metaheuristic algorithms. Optimization is essentially everywhere, from engineering design to economics and from holiday planning to Internet routing. As money, resources and time are always limited, the optimal utility of these available resources is crucially important.

Most real-world optimizations are highly nonlinear and multimodal, under various complex constraints. Different objectives are often conflicting. Even for a single objective, sometimes, optimal solutions may not exist at all. In general, finding an optimal solution or even sub-optimal solutions is not an easy task."[15]

[14] Ibid., p. 1.

[15] Scholarpedia.Metaheuristic Optimization[on line]. [August 1 2015]. Available on Internet : < URL: http://www.scholarpedia.org/article/Metaheuristic_Optimization>

“The field of Information Theory refers big data as datasets whose rate of increase is exponentially high and in small span of time; it becomes very painful to analyze them using typical data mining tools. Such data sets results from daily capture of stock exchange, insurance cross line capture, health care services etc. In real time these data sets go on increasing and with passage of time create complex scenarios. Thus the typical data mining tools needs to be empowered by computationally efficient and adaptive technique to increase degree of efficiency by using adaptive techniques.

The statistics that have heavily contributed are the ANOVA, ANCOVA, Poisson’s Distribution, and Random Indicator Variables. The biggest drawback of any statistical tactics lies in its tuning. With exponential explosion of data, this tuning goes on taking more time and inversely affects the through put. Also due to their static nature, often complex hidden patterns are left out.” [16]

[16] Munawar, Hasan. Data Genetic Algorithm and its application to Big Data Analysis. New Delhi, India: University of Dehradun.2014. p. 1.

4. META ANALYSIS

Different sources like lectures, courses, scientific documents and articles, where gathered and cited for creating a state of the art meta-analysis for algorithm in big data. With the use of a framework for analyze and deliver the results of combining different studies having a predefined set of rules that defines methods of data extraction.

For the documents, lectures and articles standardization the named convention is the defined the following way:

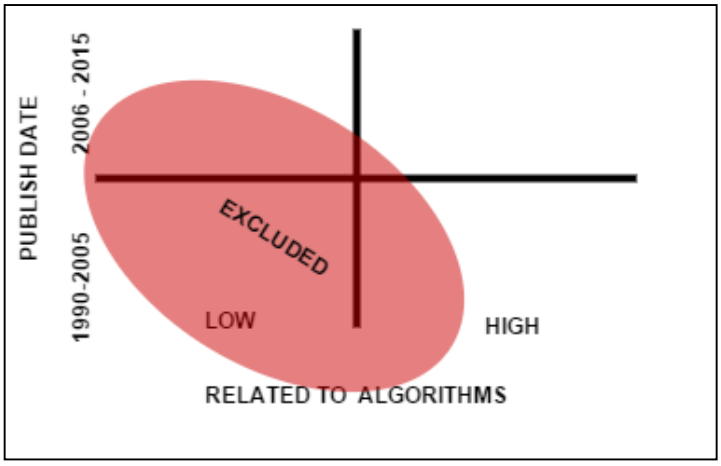
Date of publish – Name of the Articles – Author – Institution.

4.1 REASON FOR FULL-TEXT ARTICLES EXCLUDED

According to the main objectives of this document, there have been identified several dependent and independent variables to sort lectures, document and articles by importance given for the variables context.

For the first selection, the independent variable taken into consideration is the publishing data, and the dependent variable is the relatedness of the documents, lectures, articles to algorithms.

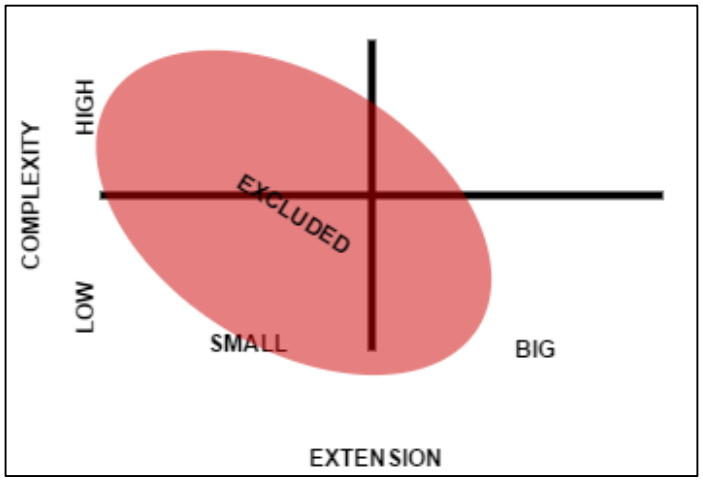
Image 4. Publish Date vs. Relation to Algorithms



Source: The Author.

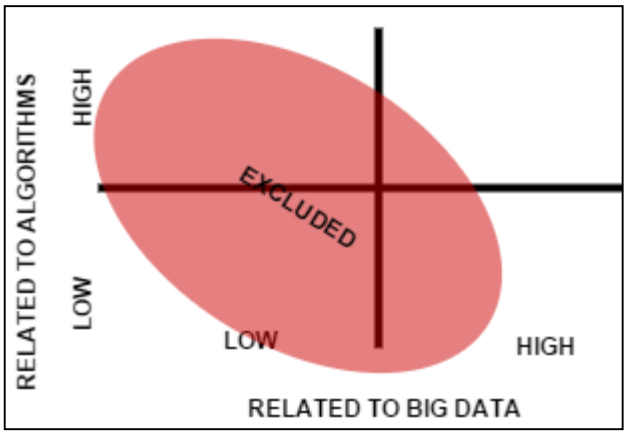
For the second pass, and to follow up the goal of the document, the extension and complexity have been considered for eligibility, taking from the result the document, lectures and article that bets fit the relation on extensibility and complexity.

Image 5. Extension vs. Complexity



Source: The Author.

Image 6. Topic Relation



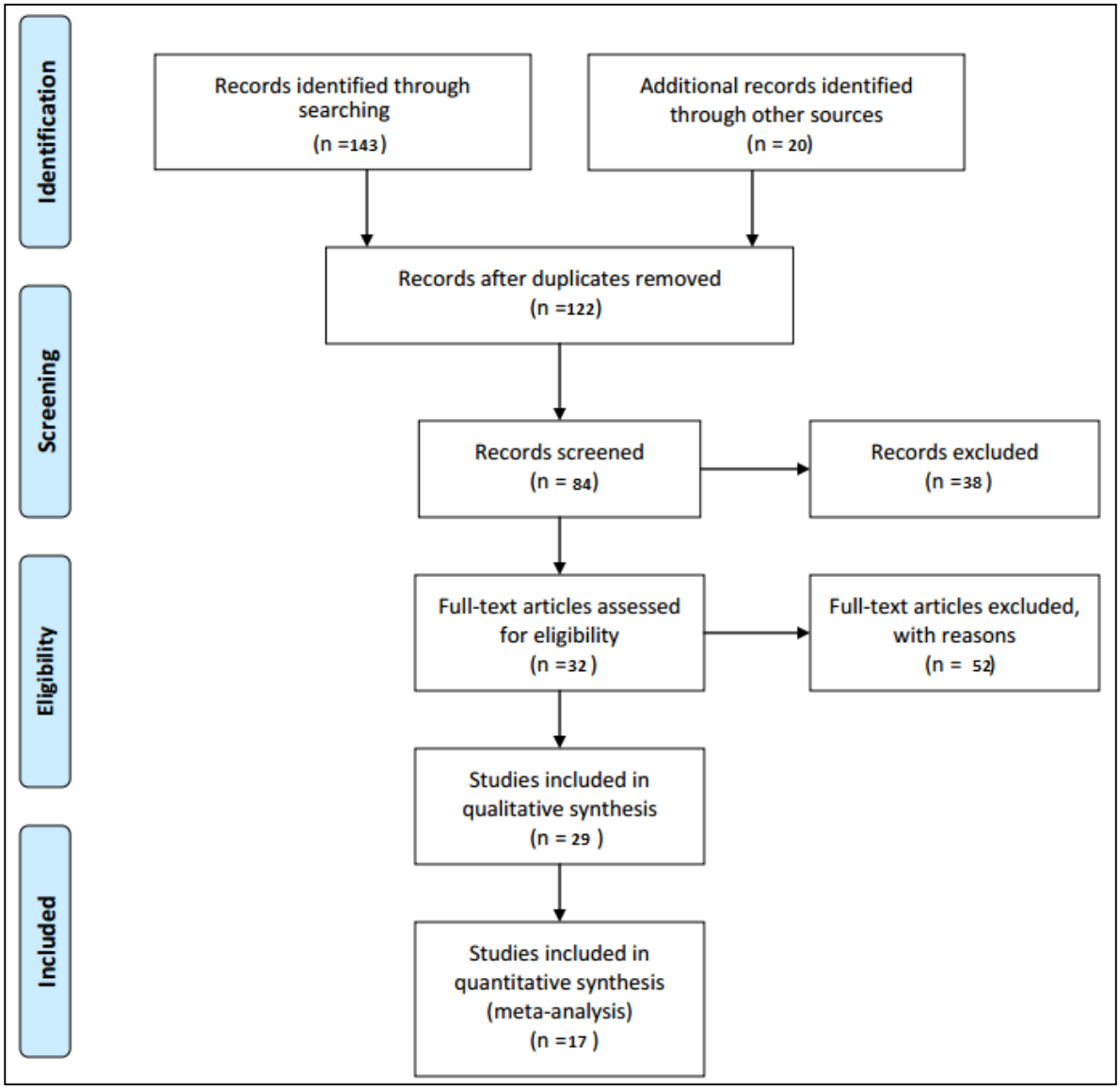
Source: The Author.

4.2 INFORMATION SOURCES

Articles, lecture and articles come from across multiple courses from different colleges on the United States and around the world, they consists of 143 documents from different studies and scientific publications, information gathered between July 2015 and October 2015, selecting documents from 2002 to 2015.

4.3 STUDY SELECTION

Table 1. PRISMA FRAMEWORK.



PRISMA Group [on line]. [August 1 2015]. Available on Internet: <http://www.prisma-statement.org/documents/PRISMA%202009%20flow%20diagram.pdf>

The Data collection process will consist of the objective of finding advancements in the field of big data algorithms, taking in consideration lectures and online courses from different colleges around the world.

Variables:

Independent Dependent

- Date of Publish.
- Related to algorithms.
- Related to Big Data.
- Wide Background.
- Complexity.
- Extension.
- Source.

4.4 SYNTHESIS OF RESULTS

Due to the variability and representativeness of the characteristics of algorithms for Big Data, the method of analyzing and taking into consideration different dimensions called variables differentiated by its type, dependent and independent. Through a systematic record of publications where lectures, document and articles where identified, classified and reviewed to find the most representative to the objectives of this document.

4.5 RESULTS

4.5.1 Study Characteristics. For each study, selected in the meta-analysis, algorithmic techniques were extracted to fulfill the objectives of the document, giving a list of the topics included in each document.

After the identification of the documents, a classification according to topic was implemented to fit the results. The following are the categories that documents fall into.

Clustering.

Data Stream Systems.

Data Structure.

Dimension reduction.

Distributed algorithms.

Genetic Algorithms.

Graph Sketches.

Graph structure.

Hashing Algorithms.

IO Algorithms.

K-means.

Large scale machine.

MapReduce.

Distributed Data.

Mining.

Multi-Core computing.

Parallel Algorithms.

2007-Combinatorial Algorithm for Compressed Sensing

- Combinatorial methods
- Combinatorial Algorithm for Compressed Sensing
- Hamming Trick
- Generalizing to k-sparse signals

2007-Near-Optimal Algorithm for estimating the Entropy of a Stream

- Computing the entropy of a stream
- Higher-order entropy
- Entropy of a random walk

2009-Data Stream Algorithms200-Barbados Workshop on Computational Complexity

- Data Streams
- Streaming Algorithms via Sampling
- Some Applications of CM-Sketch
- Graph and Geometry Problems in the Stream Model
- Counting Triangles

- Maximum Weight Matching
- K-Center Clustering
- Distance Estimation in a Graph
- Functional Approximation
- Arbitrary Dictionary
- An orthogonal basis: Haar wavelet
- Massive Unordered Distributed Data and Functional Monitoring
- Massive Unordered Distributed Data

2010-Data-Parallel Algorithms and Techniques

- The PRAM Model
- Example of a PRAM algorithm
- Work-Depth presentation of algorithms
- Goals for Designers of Parallel Algorithms
- Some final Comments
- Default assumption regarding shared assumption access resolution
- NC: A Related, Yet Different, Efficiency Goal
- On selection of material for these notes
- Level of material
- Techniques: Balanced Binary Trees; Problem: Prefix-Sums
- Application - the Compaction Problem
- Recursive Presentation of the Prefix-Sums Algorithm
- The Simple Merging-Sorting Cluster
- Technique: Partitioning; Problem: Merging
- Technique: Divide and Conquer; Problem: Sorting-by-merging

2011-Introduction to IO Efficient Algorithms-MIT

- Von Neumann Architecture
- Memory as Disk

- Cache
- Memory Hierarchy
- Block Transfer
- Scalability
- External Memory Model
- Fundamental Bounds
- Difference between N and N=B

2011-Introduction to Parallel Algorithm Analysis

- Petri Nets
- Critical Regions Problem
- Amdahl's and Gustafson's Laws
- Logical Clocks
- DAG Model
- PRAM Model
- Message Passing Model
- Programming in MPI (Open Source High Performance Computing).
- Bulk Synchronous Parallel
- Bridging Model for Multi-Core Computing
- MapReduce
- General Purpose GPU

2011-Introduction to Streaming Algorithms

- Network Router
- Telephone Switch
- Ad Auction
- Flight Logs on Tape
- Streaming Model
- Decreasing Probability of Failure

- Markov Inequality
- Chebyshev's Inequality
- Chertoff Inequality

2012-Clustering-Stanford University

- Introduction to Clustering Techniques
- Points, Spaces, and Distances
- Clustering Strategies
- The Curse of Dimensionality
- Angles between Vectors
- Hierarchical Clustering
- Hierarchical Clustering in a Euclidean Space
- Efficiency of Hierarchical Clustering
- Alternative Rules for Controlling Hierarchical Clustering
- Hierarchical Clustering in Non-Euclidean Spaces
- K-means Algorithms
- K-Means Basics
- Initializing Clusters for K-Means
- Picking the Right Value of k
- The Algorithm of Bradley, Fayyad, and Reina
- Processing Data in the BFR Algorithm

- The CURE (Clustering Using Representatives) Algorithm
    - Initialization in CURE
    - Completion of the CURE Algorithm

- Clustering in Non-Euclidean Spaces
    - Representing Clusters in the GRGPF Algorithm
    - Initializing the Cluster Tree
    - Adding Points in the GRGPF Algorithm

- Splitting and Merging Clusters

- Clustering for Streams and Parallelism
- The Stream-Computing Model
- A Stream-Clustering Algorithm
- Initializing Buckets
- Merging Buckets
- Answering Queries
- Clustering in a Parallel Environment

2012-Graph Streams algorithm

- Graph Streams
- K-Edge Connectivity
- Proof of Lemma
- Spanners
- Sparsifier
- Basic Properties of Sparsifiers
- Spectral Sparsification

2012-IO-Algorithms

- Massive Data
- Random Access Machine Model
- Hierarchical Memory
- Slow I/O
- Scalability Problems
- External Memory Model
- Fundamental Bounds

- Scalability Problems: Block Access Matters

2012-IO-Algorithms2

- I/O-Model
- External Search Trees
- B-trees
- Secondary Structures
- Weight-balanced B-tree
- Weight-balanced B-tree Insert
- Weight-balanced B-tree Delete
- Persistent B-tree
- Buffer-tree Technique

2012-MapReduce Algorithms

- MapReduce makes parallel programming easy
- Unordered Data
- Matrix Transpose
- Hadoop(Open source version of MapReduce)
    - What is MapReduce?
    - Modeling MapReduce
    - Dealing with Data Skew

2013-Algorithm_and_approaches_to_handle_large_Data-_A_Survey

- Data mining
- Architecture
- Algorithm
- Potential Application
- Issues with Big Data

2013-Algorithmic Techniques for Big Data

- Introduction to Data Streams
- Finding Frequent Items Deterministically
- Lower Bound for Deterministic Computation of Distinct elements
- Basic Concentration Inequalities

2014-Genetic-Algorithm-and-its-application-to-Big-Data-Analysis

- Genetic Algorithm
- Why and Where GA?
- The Chromosome:
- Representation of the Chromosome
- The Idea of Random Population
- Paradox of Exponential Degradation or Combinatorial Explosion
- Big Data Analysis using the Genetic Algorithm

2014-Parallel Algorithms for Geometric Graph Problems

- The Model
- Minimum Spanning Tree
- The Unit Step Algorithm

External-Memory Graph Algorithms

- PRAM Simulation
- Time-Forward Processing
- An Algorithmic Framework for List Ranking.

Fast Parallel GPU-Sorting Using a Hybrid Algorithm

- Vector-Mergesort
- Partitioning of input stream into sublists

Fundamental Parallel Algorithms for Private-Cache Chip Multiprocessors

- THE MODEL
- Multiway Partitioning
- SORTING
- Distribution Sort
- Merge sort
- BOUNDS FOR SORTINGALGORITHMS IN THE PEM MODEL

2013 - CSCI8980 Algorithmic Techniques for Big Data

- Linear Sketch as Dimensionality Reduction Technique
- Johnson-Linden Strauss Lemma
- Nearest Neighbor Problem
- Locality Sensitive Hashing

2012 – Clustering – Stanford University.

- Hierarchical Clustering in a Euclidean Space.
- K-means Algorithms.
- The Algorithm of Bradley, Fayyad, and Reina.
- Processing Data in the BFR Algorithm.
- The CURE Algorithm.

2011 - Introduction to Streaming Algorithms Jeff M. Phillips

- Network Router
- Streaming Model

2009 - Data Stream Algorithms - S. Muthu Muthukrishnan

- Count-Min Sketch.
- Streaming Algorithms via Sampling.
- Graph and Geometry Problems in the Stream Model.
- The Matador's Fate and the Mad Sofa Taster: Random Order and Multiple Passes.

Lower Bounds in Streaming - PiotrIndyk - MIT

- Communication Complexity
- Space complexity of L2 norm estimation

2007 - A Near-Optimal Algorithm for Estimating the Entropy of a Stream - AMIT CHAKRABARTI

- Computing the entropy of a stream
- Efficient Implementation

2013 - Algorithm and approaches to handle large Algorithm and approaches to handle large

- Architecture
- Algorithm
- Issues with Big Data

2014 - Genetic Algorithm and its application to Big Data Analysis - Munawar Hasan

- The Chromosome.
- Representation of the Chromosome.
- Paradox of Exponential Degradation or Combinatorial Explosion.

- Big Data Analysis using the Genetic Algorithm.

2007 - Sketching, Streaming and Sub-linear Space Algorithms - Piotr Indyk

- Combinatorial Algorithm for Compressed Sensing
- Hamming Trick
- Generalizing to k-sparse signals

Parallel Algorithms for Geometric Graph Problems

- Solve-And-Sketch Framework
- Minimum Spanning Tree
- Hierarchical Partitions
- The Unit Step Algorithm

4.6 DISCUSSION

4.6.1 Limitations. For limiting and narrow results several independent and dependent variables have been identify to grade each of the document collected. The limitations will be the variables itself that dictates the selection or rejection of documents. The investigation will be defined in lectures, document and articles about big data and algorithms for large data sets, from deferent sources like tutorials, courses and lectures.

4.6.2 Conclusions. A vast view of the selected document included in the repository for the Meta-analysis, were all main topics where extract for easier identification when searching for different topics related to big data.

Thanks to the use of the PRISMA framework and the guiles lines defined to construct this document and the definition of a state-of-the-art document, the meta-analysis have come out with a up to date research on big data algorithms, identifying and selecting from different sources documents, lectures and articles the investigation have produce a well-organized repository for future investigation on the subject.

5. OPEN FIELDS OF RESEARCH

5.1 BIG DATA SECURITY

According to Needle in a Data stack Report from McAfee most of current organizations in the world are not prepared to face the security breaches and risk that comes when handing big data, most of them may have the latest antivirus software and tracing log functionally for intern transactions, but when it comes to big data a new set of rules must be follow to unsure security on big data.

"With organizations gathering and storing customers, providers and employee information in different types of storage, which may be shared across multiple segments in the organization, can become a huge threats to privacy and malicious intents for this information, threats like these can only be prevented through careful analysis of behaviors viewed against what is considered normal within an organization. Careful monitoring of processes may have prevented situations like this." [17]

The following image show how quickly organizations are able to respond in case of an attack, having in mind that 66% of the attack comes from foreign sources, while 44% comes from intern source, such as employees and clients.

[17] THALESGROUP. Big Data [on line]. [August 1 2015]. Available on Internet : < URL: https://www.thalesgroup.com/en/worldwide/big-data/four-vs-big-data>

A lesson learn from the McAfee report is that security information from all points of vulnerability must be gathered and analyzed in real time in order to identify correlations and patterns that indicate attempts to breach defenses.

5.1.1 Common Techniques for Securing Big Data. "Collect All Security Information, to achieve risk-based security intelligence, address APTs, and improve security monitoring, businesses need to store and analyze the right information. This goes way beyond log management. Without an automated approach and high-performance systems, this can be a real challenge. Deploying technologies that provide intelligent detection and automated collection will give organizations greater external threat and internal user context."[18]

"Synthesize Actionable Insights in Real Time. The volume, velocity, and variety of information have pushed legacy SIEM systems to their limit. Now, with the pressing need to clearly identify complex attacks, organizations need advanced analytics that go beyond pattern matching to true risk-based analysis and modeling backed by a data management system that can keep up with complex real-time analytics."[19]

"Store and Investigate Long-Term Trends. While real-time analysis of data is essential to derive security value from SIEM, organizations also need to be able to research long term trends and patterns. Beyond just finding a needle in a data stack, APT detection needs to be even more granular to find the right needle in a stack of needles. Organizations store approximately 11 to 15 terabytes of data a week, but 58% of firms store this data for just three months. There is no one-size-fits-all best practice, but organizations should be aware that advanced threats can occur over months or years by going under the radar of many blocking

[18] McAfee. Big Data [on line]. [August 18 2015]. Available on Internet : < URL: http://www.mcafee.com/uk/resources/reports/rp-needle-in-a-datastack.pdf >
[19] Ibid., p. 1.

technologies. Not retaining the data impedes an organization's ability to find, understand, and eliminate these insidious threats."[20]

"Increase Threat Visibility. To be effective, SIEM analysis has to go beyond an IP address and should provide security professionals with an understanding of the nature of the external system. Many SIEMs support threat feeds, but even more important is the breadth of the threat feed and the way it is used. Effective threat feed implementations use this data to perform a real-time reputation check, immediately alerting upon interaction with a known threat and pulling the reputation of the external source into the risk score."[21]

"Customize Security Information. Organizations that have an advanced, easy-to-use SIEM also must be able to customize their SIEM deployment based on risk which results in a stronger opportunity to detect APTs, insider abuse, and other hard-to-uncover attacks. At a minimum, that process requires having an understanding of which data is sensitive, which services are most critical, and who trusted users are that access these systems and services. A strong SIEM solution will have a risk-based engine where these parameters can easily be added to make risk prioritization meaningful."[22]

"Monitor and Block. Many organizations are frequently confused about monitoring versus blocking. Successful businesses understand what they can block and what they cannot and put a monitoring program in place to detect threats that can leverage available services, data, and resources. This is the mantra of prevent what you can, monitor what you can't. At the heart of any strong security program is the protection of the confidentiality, availability, and integrity of assets. An effective SIEM will orchestrate this monitoring through collection of all security-

[20] Ibid., p. 1.
[21] Ibid., p. 1.
[22] Ibid., p. 1.

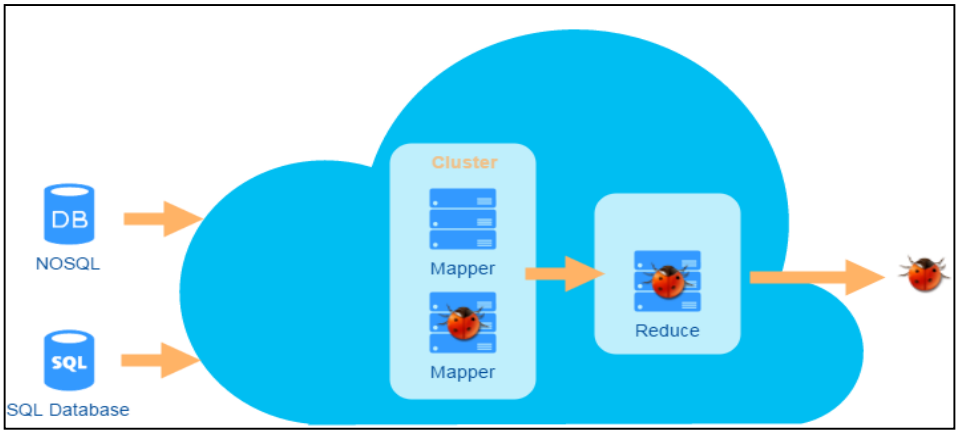
relevant events, align it to context, and perform analytics to detect suspicious or malicious activity."[23]

"Create Synergy between It and Security. There needs to be greater understanding and cooperation between security and IT. Security and IT convergence is not at the stage it should be in most organizations, and IT departments often believe assets are protected when, in fact, they are not."[24]

5.1.2 Threats for Big Data. Secure Computations in Distributed Programming Frameworks. "Distributed programming frameworks utilize parallelism in computational and storage to process massive amounts of data. The MapReduce framework splits an input file into multiple chunks. In the first phase of MapReduce, a Mapper for each chunk reads the data, performs some computation, and outputs a list of key/value pairs. In the next phase, a reducer combines the values belonging to each distinct key and outputs the result. There are two major attack prevention measures: securing the mappers and securing the data in the presence of an untrusted mapper. Untrusted mappers could return wrong results, which will in turn generate incorrect aggregate results"[25]

[23] Ibid., p. 1.
[24] Ibid., p. 1.
[25] ISACA. Big Data [on line]. [August 1 2015]. Available on Internet : < URL: http://www.isaca.org/groups/professional-english/big-data/groupdocuments/big_data_top_ten_v1.pdf>

Image 7. MapReduce diagram with untrusted mappers



Source: The Author

“Originally developed by Google, the MapReduce website describes it as "a programming model and software framework for writing applications that rapidly process vast amounts of data in parallel on large clusters of compute nodes." It's used by Hadoop, as well as many other data processing applications”. [26]

Security Best Practices for Non-Relational Data Stores. “The truth is that NoSQL has not been designed with security as a priority, so developers or security teams must add a security layer to their organizations NoSQL applications.”[27]

“Unlike traditional RDBMS, NoSQL databases do not have a schema. Hence, segregating permissions on a schema, table, column or row makes no sense”[28]

[26] BDISYS. Big Data [on line]. [August 5 2015]. Available on Internet : < URL: http://www.bdisys.com/27/1/17/BIG%20DATA/HADOOP>
[27] Computer Weekly. Securing NoSQL applications [on line]. [August 15 2015]. Available on Internet : < URL: http://www.computerweekly.com/tip/Securing-NoSQL-applications-Best-practises-for-big-data-security>

“Non-relation data stores popularized by NoSQL databases are still evolving with respect to security infrastructure. For instance, robust solutions to NoSQL injection are still not mature. Each NoSQL DBs were built to tackle different challenges posed by the analytics world and hence security was never part of the model at any point of its design stage. Developer using NoSQL databases usually embed security in the middleware. NoSQL databases do not provide ant support for enforcing it explicitly in the database. However, clustering aspect of NoSQL poses addition challenges to the robustness of such security practices.”[29]

“Companies dealing with big unstructured data sets may benefit by migration from a traditional relational database to a NoSQL database in terms of accommodating/processing huge volume of data. In general, the security philosophy of NoSQL databases relies in external enforcing mechanisms. To reduce security incidents, the company must review security polices for the middleware assign items to its engine and at the same time toughen NoSQL database itself to match its counterpart RDBs without compromising on its operational features.”[30]

“NoSQL data stores are basically vulnerable to the same security risks as traditional RDBMS data stores, so the usual best practices for storing sensitive data should be applied when developing a NoSQL-based application. These include:

- Encrypting sensitive database fields;

[28] Team Shatter. The Year of NoSQL Data Breaches [on line]. [August 09 2015]. Available on Internet : < URL: http://www.teamshatter.com/uncategorized/2011-%E2%80%93-the-year-of-nosql-data-breaches/>
[29] ISACA. Big Data [on line]. [August 18 2015]. Available on Internet : < URL: http://www.isaca.org/groups/professional-english/big-data/groupdocuments/big_data_top_ten_v1.pdf >
[30] Ibid., p. 1.

- Keeping unencrypted values in a sandboxed environment;
- Using sufficient input validation;
- Applying strong user authentication policies."[31]

Kerberos Authentication Mechanism. The user tries to log on to the client by providing user credentials. The Kerberos service on the client computer sends a Kerberos authentication service request to the Key Distribution Center (KDC). The request contains the user name, the service information for which the ticket-granting ticket (TGT) is requested, and a time stamp that is encrypted using the user's long-term key, or password.

Authentication service sends encrypted TGT and session key. The KDC gets the long-term key, or password, for the user from Active Directory, and then decrypts the time stamp that was passed along with the request. If the time stamp is valid, the user is genuine. The KDC authentication service creates a logon session key and encrypts the copy with the user's long-term key. The authentication service then creates a TGT, which includes user information and a logon session key. Finally, the authentication service encrypts the TGT with its own key and passes both the encrypted session key and the encrypted TGT to client.

Client requests server access from TGT. The client decrypts the logon session key using its long-term key, or password, and caches it locally. Additionally, the client stores the encrypted TGT in its cache. When accessing a network service, the client sends a request to the KDC ticket-granting service (TGS) with information, including the user's name, an authenticator message encrypted using the user's logon session key, the TGT, and the name of the service (and server) that the user wants to access.

TGS sends encrypted session key and ticket. The TGS on the KDC decrypts the TGT using its own key and extracts the logon session key. Using the logon session key, it decrypts the authenticator message (which is usually a time stamp). If the authenticator message is successfully decrypted, the TGS

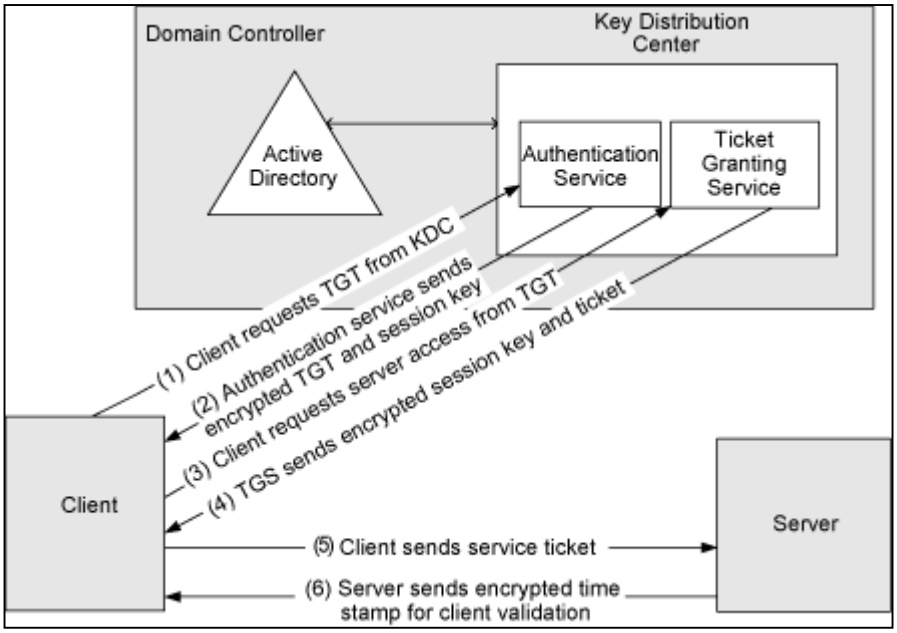
[31] Ibid., p. 1.

extracts user information from the TGT, and using the user information creates a service session key for accessing the service. It encrypts one copy of the service session key with the user's logon session key, creates a service ticket with the service session key and user information, and then encrypts the service ticket with the server's long term key (password). The TGS then sends the encrypted service session key and service ticket to the client.

Client sends service ticket. When a client accesses the service, it sends a request to the server. The request contains the authenticator message (time stamp), which is encrypted using the service session-key and the service ticket.

Server sends encrypted time stamp for client validation. The server decrypts the service ticket and extracts the service session-key. Using the service session-key, the server decrypts the authenticator message (time stamp) and evaluates it. If the authenticator passes the test, the server encrypts the authenticator (time stamp) using the service session-key and then passes the authenticator back to the client. The client decrypts the time stamp, and if it is the same as the original, the service is genuine and the client proceeds with the connection.

Image 8. Kerberos authentication protocol.



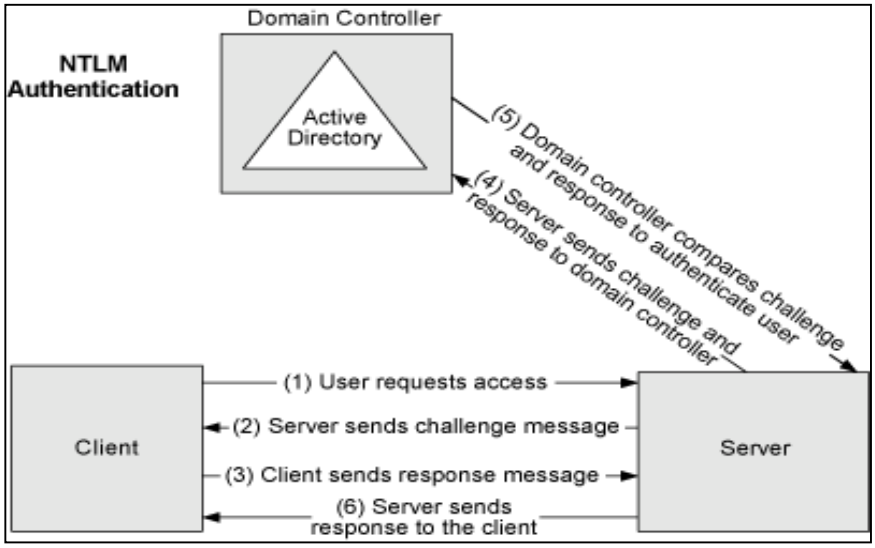
Source Windows Authentication in ASP.NET 2.0. Microsoft [on line]. [August 1 2015]. Available on Internet: <https://msdn.microsoft.com/en-us/library/ff647076.aspx>

NTLM Authentication Mechanism. NT LAN Manager is the authentication protocol used in Windows NT and in Windows 2000 Server work-group environments. It is also employed in mixed Windows 2000 Active Directory domain environments that must authenticate Windows NT systems. When Windows 2000 Server is converted to native mode where no down-level Windows NT domain controllers exist, NTLM is disabled. Kerberos v5 then becomes the default authentication protocol for the enterprise.

User requests access. The user tries to log on to the client by providing user credentials. Before logon, the client computer caches the password hash and discards the password. The client sends a request to the server, which includes the user name along with the request in plain text. Server sends challenge message. The server generates a 16-byte random number called challenge, or nonce, and sends it to the client. Client sends response message. The client uses a password hash generated from the user's

password to encrypt the challenge sent by the server. It sends this encrypted challenge in the form of a response back to the server. Server sends challenge and response to domain controller. The server sends the user name, the original challenge, and the response from the client computer to the domain controller. Domain controller compares challenge and response to authenticate user. The domain controller obtains the password hash for the user, and then uses this hash to encrypt the original challenge. Next, the domain controller compares the encrypted challenge with the response from the client computer. If they match, the domain controller sends the server confirmation that the user is authenticated. Server sends response to the client. Assuming valid credentials, the server grants the client access to the requested service or resource.[32]

Image 9. NTLM protocol.



Source Windows Authentication in ASP.NET 2.0. Microsoft [on line]. [August 1 2015]. Available on Internet: <https://msdn.microsoft.com/en-us/library/ff647076.aspx>

[32] MICROSOFT. Big Data [on line]. [August 18 2015]. Available on Internet : < URL: https://msdn.microsoft.com/en-us/library/ff647076.aspx>

Secure Data Storage and Transaction Logs. “Data and transaction logs are stored in multi-tiered storage media. Manually moving data between tiers gives the IT manager direct control over exactly what data is moved and when. However, as the size of the data set has been, and continues to be, growing exponentially, scalability and availability have necessitated auto-tiering for big data storage. New mechanism are imperative to thwart unauthorized access and maintain the 24/7 availability.”[33]

“A manufacture wants to integrate data from different divisions. Some of this data is rarely retrieved, while some divisions constantly utilize the same data pools. An auto-tier storage system will save the manufacturer money by pulling the rarely utilized data to a lower (and cheaper) tier. However, this data may consist in R&D results, not popular but containing critical information. As lower-tier often provides decreased security, the company should study carefully tiering strategies.”[34]

“Tiered storage strategies that assign data based on value as well as access and retention needs, can cut enterprise data storage costs and address storage capacity issues. An effective tiered storage strategy has to balance storage costs, data lifecycle management practices, storage technology priorities and data access speeds”.[35]

For making data storage processes more effective there are 3 main aspects that worth understand before implementing a solution:

[33] ISACA. Top ten big data security and privacy challenges [on line]. [August 18 2015]. Available on Internet : <URL: http://www.isaca.org/groups/professional-english/big-data/groupdocuments/big_data_top_ten_v1.pdf>

[34] Ibid., p. 1.

[35] COMPUTERWEEKLY. Big Data Best Practices [on line]. [August 18 2015]. Available on Internet : < URL: http://www.computerweekly.com/report/Tiered-storage-strategies-and-best-practices>

- Data Classification
- Data Backup
- Data storage.

5.2 BIG DATA INFRASTRUCTURE

"Big data is simply any data set that has grown too big to be efficiently worked on in real-time with traditional database tools"[36] like any other infrastructure architecture in big data we need to attend the main consideration when developing system ranging from scalability to performance, passing thought flexibility and straightforward operational.

In order to give a better understanding of the challenges faced when putting up a big data infrastructure we can compare the components for traditional data and big data.

Table 2. Traditional Data vs Big Data

Components	Traditional Data	Big Data
Architecture	Centralized	Distributed
Data volume	Terabytes	Petabytes to exabytes
Data type	Structured or transactional	Unstructured or semi-structured
Data relationships	Known relationship	Complex/unknown relationships
Data model	Fixed schema	Schema-less

Source: Juniper Networks. Big Data [on line]. [August 1 2015]. Available on Internet: <http://www.juniper.net/us/en/local/pdf/whitepapers/2000488-en.pdf>

[36] COMPUTERWEEKLY. Big Data Security [on line]. [August 18 2015]. Available on Internet : < URL: http://www.computerweekly.com/tip/Securing-NoSQL-applications-Best-practises-for-big-data-security>

“Before you even start your big data project, you need to make sure you have the capability in place to manage it effectively. Whether we are talking about transactional application servers, specialist appliances for applications such as business intelligence, or the supercomputers used for digital simulation you need to have significant amounts of processing power at your disposal, simply to deal with the vast volumes of data typically processed in big data implementations.

For example, running a NoSQL database like MongoDB that requires high I/O operations on flash or SSD storage ensures the best performance where it really matters – speed.”[37]

5.2.1 Solid State Drives. “SSDs are based on NAND Flash memory are well-suited for big data applications because they provide ultra-fast storage performance, quickly delivering an impressive return on investment. SSDs can be deployed as host cache, network cache, all-SSD storage arrays, or hybrid storage arrays with an SSD tier.”[38]

“Depending on the big data application, either enterprise class or personal storage SSDs may be used. Enterprise SSDs are robust and durable, offering superior performance for mixed read/write workloads, while personal storage SSDs typically cost less and are suitable for read-centric workloads. It is important to understand the workload performance and endurance requirements before making a decision.”[39]

[37] INFORMATION-AGE. Big Data [on line]. [August 18 2015]. Available on Internet : < URL: http://www.information-age.com/technology/information-management/123457900/how-build-big-data-infrastructure#sthash.yi36Ozan.dpuf>
[38] MICRON. Big Data [on line]. [August 18 2015]. Available on Internet : < URL: https://www.micron.com/~/media/documents/products/technical-marketing-brief/brief_ssds_big_data.pdf>
[39] Ibid., p. 1.

5.2.2 SSD Benefits. "Deliver good sequential I/O and outstanding random I/O performance. For many systems, storage I/O acts as a bottleneck, while powerful, multicore CPUs sit idle waiting for data to process. SSDs remove the bottleneck and unleash application performance, enabling true processing throughput and user productivity.

- Nonvolatile Retain data when power is removed; no destaging required, like DRAM.
- Low Power Consume less power per system than equivalent spinning disks, reducing data center power and cooling expenses.
- Flexible Deployment Available in a unique variety of form factors and interfaces compared to other storage solutions"[40]

5.3. BIG DATA FOR BUSINESS

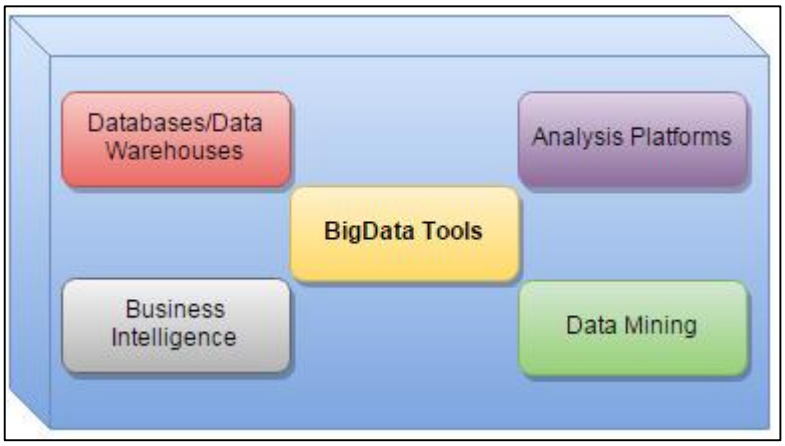
Data have shown exponential grow rate, especially in the last few years, where different fields are in need of new ways to store, process and analyze data to retrieve new profit from all the information.

This has become high priority, giving birth to tools that emphasize in being more cost-effective. There can be identified four main fields where tools have made important advances for treating big data.

[40] MICRON. Big Data [on line]. [August 18 2015]. Available on Internet : < URL: https://www.micron.com/~/media/documents/products/technical-marketing-rief/brief_ssds_big_data.pdf>

Four big trending can be point it out from the big data tools, in which the tools listed will be categorized into the described segment that suit his purpose.

Image 10. Big Data Tools.



Source: The Author.

5.3.1 Hadoop Meets All. "Hadoop is essentially an open-source framework for processing, storing and analyzing data. The fundamental principle behind Hadoop is rather than tackling one monolithic block of data all in one go, it's more efficient to break up & distribute data into many parts, allowing processing and analyzing of different parts concurrently"[41].

"The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on

[41] DATACONOMY. Big Data [on line]. [August 18 2015]. Available on Internet : < URL: http://dataconomy.com/understanding-big-data-infrastructure/>

hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures"[42]

"Hadoop is a whole ecosystem of different products, largely presided over by the Apache Software Foundation. Some key components include:

- HDFS The default storage layer.
- MapReduce Executes a wide range of analytic functions by analyzing datasets in parallel before 'reducing' the results. The "Map" job distributes a query to different nodes, and the "Reduce" gathers the results and resolves them into a single value.
- YARN Responsible for cluster management and scheduling user applications.
- Spark Used on top of HDFS and promises speeds up to100 times faster than the two-step MapReduce function in certain applications. Allows data to loaded in-memory and queried repeatedly, making it particularly apt for machine learning algorithms."[43]

5.3.2 Hadoop Security. "Mutual Authentication with Kerberos RPC (SASL/GSSAPI) on RPC connections SASL/GSSAPI was used to implement Kerberos and mutually authenticate users, their processes, and Hadoop services on RPC connections"[44]

"Pluggable Authentication for HTTP Web Consoles meaning that implementers of web applications and web consoles could implement their own authentication

[42] HADOOP. Big Data [on line]. [August 18 2015]. Available on Internet : < URL: https://hadoop.apache.org/>
[43] DATACONOMY. Big Data [on line]. [August 23 2015]. Available on Internet : < URL: http://dataconomy.com/understanding-big-data-infrastructure/>
[44] INFOQ.Big Data [on line]. [August 23 2015]. Available on Internet : < URL: http://www.infoq.com/articles/HadoopSecurityModel/>

mechanism for HTTP connections. This could include (but was not limited to) HTTP SPNEGO authentication. Enforcement of HDFS files permissions Access control to files in HDFS could be enforced by the Name Node based on file permissions - Access Control Lists (ACLs) of users and groups. Delegation Tokens for Subsequent Authentication checks these were used between the various clients and services after their initial authentication in order to reduce the performance overhead and load on the Kerberos KDC after the initial user authentication. Specifically, delegation tokens are used in communication with the Name Node for subsequent authenticated access without using the Kerberos Servers."[45]

"Block Access Tokens for Access Control to Data Block When access to data blocks were needed, the Name Node would make an access control decision based on HDFS file permissions and would issue Block access tokens (using HMAC-SHA1) that could be sent to the Data Node for block access requests. Because Data Nodes have no concept of files or permissions, this was necessary to make the connection between the HDFS permissions and access to the blocks of data. Job Tokens to Enforce Task Authorization Job tokens are created by the Job Tracker and passed onto Task Trackers, ensuring that Tasks could only do work on the jobs that they are assigned. Tasks could also be configured to run as the user submitting the job, making access control checks simpler"[46]

"Pluggable Authentication to HTTP SPNEGO Authentication Although the 2009 security design of Hadoop focused on pluggable authentication, the Hadoop developer community decided that it would be better to use Kerberos consistently, since Kerberos authentication was already being used for RPC connections (users, applications, and Hadoop services). Now, Hadoop web consoles are configured to use HTTP SPNEGO Authentication, an implementation of Kerberos for web consoles. This provided some much-needed consistency. Network Encryption Connections utilizing SASL can be configured to use a Quality of Protection (QoP) of confidential, enforcing encryption at the network level – this includes

[45] Ibid., p. 1.
[46] Ibid., p. 1.

connections using Kerberos RPC and subsequent authentication using delegation tokens. Web consoles and MapReduce shuffle operations can be encrypted by configuring them to use SSL. Finally, HDFS File Transfer can also be configured for encryption"[47]

5.3.3 Hadoop Cluster. "The Hadoop Cluster Hadoop, which includes a distributed file system known as Hadoop Distributed File System (HDFS) and MapReduce, is a critical big data technology that provides a scalable file system infrastructure and allows for the horizontal scale of data for quick query, access, and data management. At its most basic level, a Hadoop implementation creates four unique node types for cataloging, tracking, and managing data throughout the infrastructure: data node, client node, name node, and job tracker."[48]

"The capabilities of these four types are generally as follows:

- Data node the data nodes are the repositories for the data, and consist of multiple smaller database infrastructures that are horizontally scaled across compute and storage resources through the infrastructure. Larger big data repositories will have numerous data nodes. The critical architectural concern is that unlike traditional database infrastructure, these data nodes have no necessary requirement for locality to clients, analytics, or other business intelligence.
- Client The client represents the user interface to the big data implementation and query engine. The client could be a server or PC with a traditional user interface.
- Name node the name node is the equivalent of the address router for the big data implementation. This node maintains the index and location of every data node.

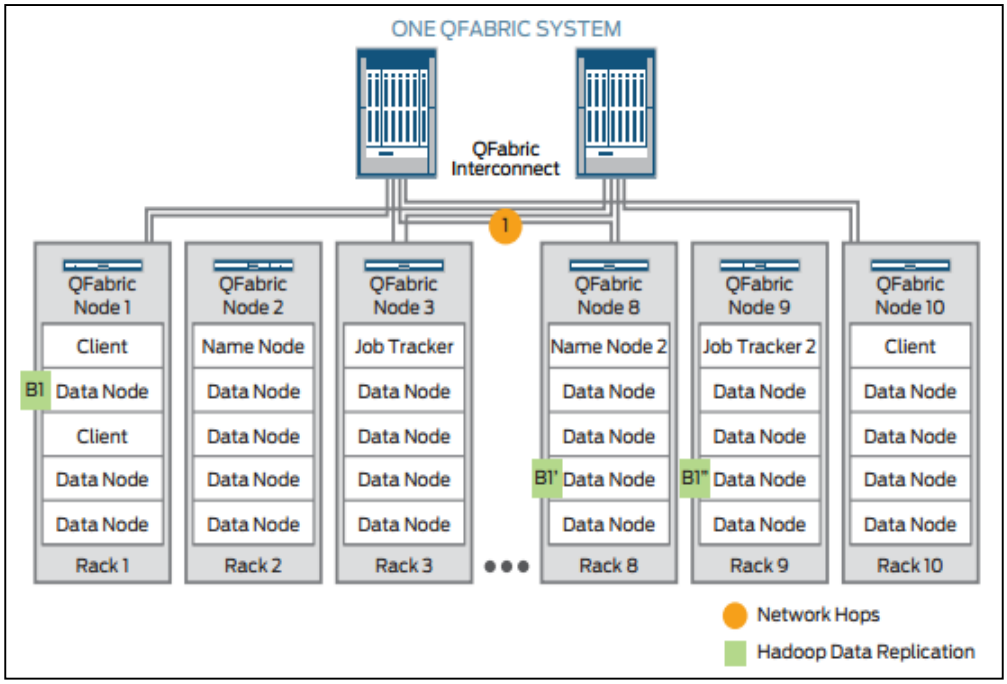
[47] Ibid., p. 1.

[48] JUNIPER.Big Data [on line]. [August 23 2015]. Available on Internet : < URL: http://www.juniper.net/us/en/local/pdf/whitepapers/2000488-en.pdf/>

- Job tracker the job tracker represents the software job tracking mechanism to distribute and aggregate search queries across multiple nodes for ultimate client analysis. Within each data node, there may exists several tens of server or data storage elements and its own switching tier connecting each storage element with the overall Hadoop cluster. Big data infrastructure purposely breaks the data into horizontally scaled nodes, which naturally adds latencies across nodes. This is important as locality and network hops represent potential latencies in the architecture. For example, Figure # shows a granular representation of a more sophisticated Hadoop big data implementation that illustrates a typical architecture for the individual data nodes in a cluster. Not only is it particularly more resource and potentially management intensive than a simple diagram may indicate, but from a pure performance perspective, the additional hops between client, job tracker, and individual data nodes are more significant. The job tracker is required to keep track of the five hops associated with top-of-rack switch 3 in order to access data with top-of-rack switch 8. These individual hops also represent latencies and potential performance bottlenecks."[49]

[49] Ibid., p. 1.

Image 11. Hadoop



Source: Juniper Networks. Big Data [on line]. [August 1 2015]. Available on Internet: <http://www.juniper.net/us/en/local/pdf/whitepapers/2000488-en.pdf>

5.3.4 Databases and Data Warehouses

"The Apache Cassandra database is the right choice when you need scalability and high availability without compromising performance. Linear scalability and proven fault-tolerance on commodity hardware or cloud infrastructure make it the perfect platform for mission-critical data. Cassandra's support for replicating across multiple datacenters is best-in-class, providing lower latency for your users and the peace of mind of knowing that you can survive regional outages."[50]

[50] CASSANDRA.Big Data [on line]. [August 18 2015]. Available on Internet: < URL: http://cassandra.apache.org/>

“Cassandra's data model offers the convenience of column indexes with the performance of log-structured updates, strong support for de-normalization and materialized views, and powerful built-in caching”[51]

“MongoDB By offering the best of traditional databases as well as the flexibility, scale, and performance required by today’s applications, MongoDB lets innovators deploy apps as big as they can possibly dream. From startups to enterprises, for the modern and the mission-critical, MongoDB is the database for giant ideas”[52]

“MongoDB, Inc. is the company behind the database for giant ideas. We build MongoDB and the drivers, and sell software and services to make your life easier. We also run MongoDB University and sponsor the community, hosting MongoDB World and MongoDB Days all over the globe. With offices across North America, Europe, and Asia, we are close to where you do business”[53]

“MongoDB was founded in 2007 by the people behind DoubleClick, ShopWiki and Gilt Group. At DoubleClick, the site could never be down and there were daily challenges with processing, storing, and scaling data, forcing them to write their own software to handle specific problems. It was in these trenches that the team had the insight for MongoDB. They asked themselves, What do we wish we had while at DoubleClick?”[54]

[51] Ibid., p. 1.
[52] Mongodb.Big Data [on line]. [August 18 2015]. Available on Internet: < URL: https://www.mongodb.com/company?jmp=dotorg-form&_ga=1.11303793.459559017.1446154053/>
[53] Ibid., p. 1.
[54] Ibid., p. 1.

The company was founded to harness the power of the cloud for more efficiency, to scale horizontally, and to make operations easier for scale at development. Today, MongoDB boasts more than 10 million downloads, thousands of customers, and more than 1,000 partners.

5.3.5 Business Intelligence

"The Jaspersoft package is one of the open source leaders for producing reports from database columns. The software is well-polished and already installed in many businesses turning SQL tables into PDFs that everyone can scrutinize at meetings"[55]

"The JasperReports Server now offers software to suck up data from many of the major storage platforms, including MongoDB, Cassandra, Redis, Riak, CouchDB, and Neo4j. Hadoop is also well-represented, with JasperReports providing a Hive connector to reach inside of HBase"[56]

"Pentaho is another software platform that began as a report generating engine; it is, like JasperSoft, branching into big data by making it easier to absorb information from the new sources. You can hook up Pentaho's tool to many of the most popular NoSQL databases such as MongoDB and Cassandra. Once the databases

[55] Info World.Big Data Tools [on line]. [August 18 2015]. Available on Internet: < URL: http://www.infoworld.com/article/2616959/big-data/7-top-tools-for-taming-big-data.html />
[56] Ibid., p. 1.

are connected, you can drag and drop the columns into views and reports as if the information came from SQL databases”[57]

[57] Ibid., p. 1.

6. CONCLUSIONS

From the study made on big data, it can now be deduce that big data is present everywhere, from your house and your personal devices, to corporation and government. Is now clear that big data is everything to satellite communication (information that cannot be saved due to his size) to social networking (joins and graphs) where traditional ways of handling and saving this information are becoming less efficient every time big data gets bigger. This is the main reason for constructing a meta-analysis for showing the state-of-the-art on the development of new way to handle big data.

Huge opportunities for organizations have been brought up by different studies, where it already shows the profit made for handling correcting big data, it's important to stand out the problems that many organization will have in the near future, where organization ready to handle will have more open field and paths for business while organization that are not prepared will have difficulties to caching up the ones that does.

Thanks to the use of the PRISMA framework and the guiles lines defined to construct this document and the definition of a state-of-the-art document, the meta-analysis have come out with a up to date research on big data algorithms, identifying and selecting from different sources documents, lectures and articles the investigation have produce a well-organized repository for future investigation on the subject.

REFERENCES

BaseLineMag.Utilizing Cutting-Edge Unstructured Data Analytics [on line]. [August 1 2015]. Available on Internet: < URL: http://www.baselinemag.com/analytics-big-data/utilizing-cutting-edge-unstructured-data-analytics.html>

BDISYS. Big Data [on line]. [August 5 2015]. Available on Internet :< URL: http://www.bdisys.com/27/1/17/BIG%20DATA/HADOOP>

Chen, Philip. Zhang, Chun-Yang. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Macau, China: University of Macau. 2013. p. 1.

COMPUTERWEEKLY. Big Data [on line]. [August 18 2015]. Available on Internet :< URL: http://www.computerweekly.com/tip/Securing-NoSQL-applications-Best-practises-for-big-data-security>

COMPUTERWEEKLY. Big Data [on line]. [August 18 2015]. Available on Internet :< URL: http://www.computerweekly.com/report/Tiered-storage-strategies-and-best-practices>

DATACONOMY. Big Data [on line]. [August 18 2015]. Available on Internet :< URL: http://dataconomy.com/understanding-big-data-infrastructure/>

HADOOP. Big Data [on line]. [August 18 2015]. Available on Internet :< URL: https://hadoop.apache.org/>

INFOQ. Big Data [on line]. [August 23 2015]. Available on Internet :< URL: http://www.infoq.com/articles/HadoopSecurityModel />

INFORMATION-AGE. Big Data [on line]. [August 18 2015]. Available on Internet :< URL: http://www.information-age.com/technology/information-management/123457900/how-build-big-data-infrastructure#sthash.yi36Ozan.dpuf>

ISACA. Big Data [on line]. [August 18 2015]. Available on Internet :< URL: http://www.isaca.org/groups/professional-english/big-data/groupdocuments/big_data_top_ten_v1.pdf >

JUNIPER. Big Data [on line]. [August 23 2015]. Available on Internet :< URL: http://www.juniper.net/us/en/local/pdf/whitepapers/2000488-en.pdf/>

McAfee. Big Data [on line]. [August 18 2015]. Available on Internet: < URL: http://www.mcafee.com/uk/resources/reports/rp-needle-in-a-datastack.pdf >

McKinsey & Company. Big Data [on line]. [August 1 2015]. Available on Internet: < URL: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation>

MICRON. Big Data [on line]. [August 18 2015]. Available on Internet :< URL: https://www.micron.com/~/media/documents/products/technical-marketing-brief/brief_ssds_big_data.pdf >

MICROSOFT. Big Data [on line]. [August 18 2015]. Available on Internet :< URL: https://msdn.microsoft.com/en-us/library/ff647076.aspx>

Munawar, Hasan. Data Genetic Algorithm and its application to Big Data Analysis. New Delhi, India: University of Dehradun.2014. p. 1.

OXFORD DICTIONARIES. Big Data [on line]. [August 1 2015]. Available on Internet: < URL: http://www.oxforddictionaries.com/us/definition/american_english/big-data >

Scholarpedia. Metaheuristic Optimization [on line]. [August 1 2015]. Available on Internet: < URL: http://www.scholarpedia.org/article/Metaheuristic_Optimization>

THALES GROUP. Big Data [on line]. [August 1 2015]. Available on Internet: < URL: https://www.thalesgroup.com/en/worldwide/big-data/four-vs-big-data>

ThalesGroup.Big Data [on line]. [August 18 2015]. Available on Internet: < URL: https://www.thalesgroup.com/en/worldwide/big-data/four-vs-big-data>