Big Data In computer Cyber Security Systems

Amani Mobarak AlMadahkah

The Publics Authority for Applied Education and Training- Secretarial and Office Administration Institute -Girls-Computer Department- Kuwait

Summary
while Big data regularly become a important issue of research and has been used everywhere in many industries, Big data security has been increasingly concerned .Nevertheless, there is noticeable challenge between Big data security and the common use of Bigdata.In this paper we firstly reviewed the definition of Big data ,the rise of big data. Then ,we present the needs for bigdata and data analytics and show the advance in bigdata analytics. After that we address some security challenges in bigdata . We highlight on the cyber security and how can we secure the bigdata ,Also we will number some of the top security tools in fight against cybercrime .finally mention some of applications of bigdata techniques in learning .

Key words:

API	*Application programming interface*
GB	*Gaia bytes*
HPC	*High performance computing*
HDFS	*Hadoop distributed file system*
MPI	*Message –passing interface*
ACL	*Access Control List*
EDW	*enterprise data warehouse*
OLAP	*On Line Analytical Processing*

1. Introduction

Big data has become an important topics in science, health, engineering, medicine, finance, business and finally society itself. More than 2.5 quintillion bytes of data are created every day [1]. Almost 90%of data in the world today had been created in last two years alone. Examples of big data include the quantity of data shared every day in internet, mobile phone location data, YouTube videos viewed and twitter feeds. In the latest years, data produced by learning environments started to have enough raising the necessitate for big data technologies and tools to handle them. In another words we can say Big data is shorthand for advance learning in technology that can open the door to a new approach to understanding the world and making the

The revelation that started in computing processes more than two decades had led to huge amount of digital data being amassed by corporations. Advances in digital sensors; proliferation of communication systems, especially mobiles and devices; massive scale logging of system events; and rapid movement toward paperless organizations also had led to massive collection of data sources within organizations. Also the businesses dependence increasing in technology ensures that the data would keep growing even faster.

The computer industry started looking at other options, namely parallel processing to provide more economical solution. [1]

2. What is big data?

The expression “big data" refers to "massive amount of digital information's which has been stored or transmitted in computer systems" . Another definition of big data “any amount of data that is too huge or too multifaceted that predictable applications are not sufficient to process them". The term also refers to the tools and technologies used to hold "Big data".

Definition and characteristics of big data ;3V's" big data is high- volume, high velocity and high -variety information assets that demand cost -effective ,innovative forms of information processing for enhanced insight and decision making .

There is a lot of data all the time growing at 50 % a year or more than doubling every two years, estimates IDC, It's not just more steams of data, but entirely new ones. Internet improved access to information is also increasing the big data inclination. For example; In 2009 Washington opened the door for more data by starting Data.gov, a Web site that makes any one can access easily to all kind of government data. Data now is available and understandable to computers. Most of the big data is uncontrollable stuff like words, video on the Web, images and steams of sensor data, its traditional database.

Since 2011 the interest in a area known as big data has increased, and the enormous majority of computer science research, big data has received significant public and media interest. Headlines "Big data :the greater, good or attack of privacy?" and "Big data is opening doors , but maybe too many "speak volumes as to the common perception of big data .From the begging it obvious that the big data intertwined with substantial technical and socio-technical issues but an exact defection is unclear.

Big data is mostly associated with two ideas: data storage and data analysis. Despite the sudden interest in big data ,these concepts are far from new ,there for "big" implies significant , complexity and challenge ,Unfortunately the term "Big" also invites

Manuscript received April 5, 2016
Manuscript revised April 20, 2016

quantification and therein lies the difficulty in furnishing a definition.

- Oracle defines Big data as; the derivation of value from traditional relational database driven business decision making , augmented with new sources (such as blogs, social media, sensor networks, images est.)of un structured data which vary in size ,structure , format and other factors. It focused upon infrastructure
- Intel describes big data through quantifying the expenses of its business partners. It also suggest that the organizations which were surveyed deal extensively with unstructured data and place an emphasis on performing analytics over their data.
- Micro soft provides a remarkable succinct definition:"Big data is the tern increasingly used to describe the process of applying serious computing power- the latest in machine learning and artificial intelligence to seriously massive end of ten highly complex sets of information" ,It states that big data requires the application of significant computer power .

Notably
All definitions make at least one of the following assertions:

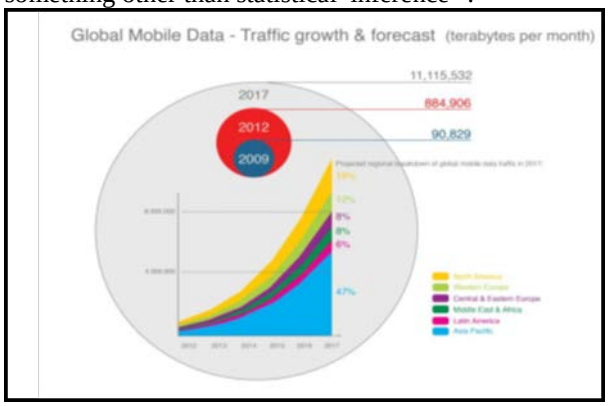
- Size: the volume of datasets in critical factor.
- Complexity: the structure, behavior and permutations of datasets in critical factor.
- Technologies: tools and techniques which are used to process a sizable or complex datasets is a critical factor.

The definitions surveyed here all encompass at least one of these factors, most encompass two ,therefore the following detention suggested "Big data is a term describing the storage and analysis of large or complex data using a serious of techniques including , but not limited to :NoSQL, Map Reduce and machine learning". [1-2]

figer 1

3. The rise of big data

The first point and central feature of Big data as a phenomenon is the un -presented growth in the volume and variety of high-frequency digital structured and the un structured data which badly emitted by and picked up about humans populations behaviors and beliefs, each year since 2012, over 1.2 zeta byte of data have been produced -1021 bytes; enough to fill 80 billon 16 GB I phones that would the earth more than 100 times.
The volume of these data is increasing just like a human population of digital data .And just like a human population with a sudden outburst of fertility gets larger and younger, The proportion of digital data produced recently (i.e. new baby data) is growing .It has been said many times that 90% of world's data was created over just two years, Although the assertion is almost impossible to source or corroborate. The second development is what has primarily been called 'the industrial revolution of data 'big data' Mike Hoeeigan at the US BLS defined big data as "non sampled data ,characterized by the creation of database from electronic sources whose primary purpose is something other than statistical inference ".



figer2

4. The need for big data and big data analytics

The needs of new security challenges because of the growing of security intelligence requires big data and big data analytics.
Firstly, an organization wants to hold its outdated security data for longer time to do analysis on the data. The historical analysis has the possibility of finding longer running attack methods and identifies declines in security over time.
Secondly, data sources not usually hired for security would be helpful for the organization to qualify what assets and entities need to be protected or detected. Like; identifying users that usually work with sensitive data, and systems

that are serious to essential business processes. Data sources such as email, social media content, corporate documents, and web content maybe helpful to add additional context to traditional security data but largely unstructured data .

Next, a variety of analytics can be performed to expose security visions from these huge data sets and need more processing time. Also this operation need to be done asynchronously to the real-time analysis that traditional security intelligence specializes in. Once the analysis is complete, the vision have to be feedback to the real time components to make the overall solution more effective over time.

Finally, a renewed emphasis needs to be placed on investigative analysis that could be branded as hoc before it is codified. That can offer the specificity of an organization and it business environment this will be very important for security intelligence solution gain interrelated consciousness for frustrating targeted attacks.

5. Advance in Big data Analytics

Why must we rush to advanced analytics? first :change is uncontrolled in business ,as seen in the multiple "economies" in recent years .Analytics Led us to discover the changes and how we react .Second, as we move slowly out the decline and into the recovery ,there are more business chances should be held ,So advanced analytics is the best way to know new customers segments ,identify the best suppliers, associate products of similarity ,understand sales seasonality, and so on. The need to advanced analytics means that a lot of organizations realize the advanced analytics for the first time and there for they are confused about how to go about it. Even if we have related experience in data warehousing, reporting, and online analytic processing (OLAP) ,we will discover that the business and technical requirements are different for advanced forms of analytics.

The user organizations are implementing specific forms of analytics, particularly what is sometimes called advanced analytics .This is a collection of related techniques and tool types, usually including predictive analytics, data mining ,statistical analysis and compelled SQL .We could also add to the list data visualization, artificial intelligence, natural language processing ,and database capabilities that supports analytics(like Map Reduce, in-database analytics, in-memory databases, columnar data stores).

Better we call the advanced analytics as "discovery analytics "why? Because what users are trying to achieve. In other word some people call it "exploratory analytics" With big data analytics, the user is typically a business analyst who is trying to discover new facts that never known before by anyone in enterprise. To do that, the analyst needs large volumes of data with huge detail. This is often data that the enterprise has not tapped for analytics. A business analysts would hold several terabytes of detailed data leftover from operational applications to get a view of recent customer behaviors .The analyst might mix that data with historic data from data warehouse.

Discovery analytics against big data can be enabled by different types of analytic tools ,including those based on SQL queries ,data mining ,statistical analytics ,fact clustering ,data visualization . TDWI research big data analytics natural language processing, text analytics, artificial intelligence and so on. It's an store of tool types, and know-how users get to know their analytic requirement before deciding which tool type is appropriate to their needs. Many of these techniques appeared in 1990s and they have been around for years .The difference today is that far more user organizations are actually using them. Because most of these techniques adapt well to large, multi-terabyte data sets with minimal data preparation.

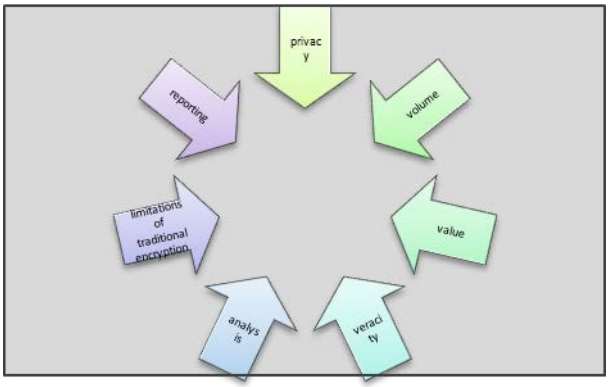
Today, where is bigdata for advanced analytics managed and operated on? where would you prefer that bigdata for advanced analytics be managed and operated on?

- The EDW :enterprise data warehouse ;is much used and preferred platform for analytics ,At most two-thirds of users surveyed report using an EDW today, and two -third say's a performance .Some EDW were originally designed by users for reporting, performance management , and OLAP(on line analytical s processing), Also It can handle advanced analytics - in terms of scalability and query performance and some cannot.
- New types of analytic platforms are coming .A few users report using cloud-based analytic platforms today, and many users would perform them.
- TDWI expects various types of clouds to become common platforms for analytics within a few years.
- Hadoop: is so hyped we would talk about it later.

figer3

6. Big Data security challenges and risks

The great opportunity that big data presents for the enterprises by tapping into varieties and volumes of data ,Scientists ,product managers ,marketers , executives , and others can take benefit from informing plans and decisions , discover new chances for optimization , and deliver breakthrough innovations. Without the right security and encryption solutions the big data could be really big problem.



figer4

In spite of the applications of big data analytics to security problems has significant promise, we have to mention some challenges:

6.1 . First challenge is the Privacy:

of avoiding data responses (using data only for the purposes that it was collected).Recently, privacy trusted largely on www.computer.org/security 75 technological limitations on the ability to extract, analyze, and correlate potentially sensitive datasets .However advance in big data analytics brought us tools extract and correlate this data ,That would make data violation much easier .That make developing the big data applications a must without forgetting the needs of privacy principles and recommendations . Al the activities produced in communications commission works like (telecommunications companies, Health Accountability data, and any Federal trade commission's) have been broad in system coverage and mostly could cause interpretation. The large scale collection and storage of data would be attractive to many people especially (whom using this data for advertising and marketing), Also government (finding this data necessary for the national security or for low prosecution), and for law breakers (they would like to steal the identities).That why we need from the big data designers creating a suitable safeguards to prevent abuse of these big data stores.[6-1]

6.2. Second challenge, the veracity: which means (authenticity - reality) (the data provenance problem) .why? Because it's difficult to be sure that each data meets the trustworthiness that our analysis algorithms require to produce the accurate results. Therefore, we need reconsider the authenticity and integrity of used data in out tools .we can take advantages from adversarial machine learning and from strong statistics to identify and moderate the effects of unkindly inserted data.[6-2]

6.3. Third challenge, the volume: volume which means (storage) ,The amount of data created every day through internet is in the order of Exabyte ,That's make the capacity of hard disks nowadays in the range of terabytes, Its large enough and it will get larger in future. The traditional RDBMS tools will be unable to store or process such as big data .to solve this challenge ,the database that don't use traditional SQL based queries are used. Compression technology might be a good choice to compress the data at rest and in memory.

6.4. Forth challenge, Analysis: Analyzing the huge size of data and the different in structure because the generated data to several types of online sites ,analysis the data may consume a lot of time and resources .defeating this, scaled out architectures could be used for processing the data in disseminated methods. Splitting data to small pieces and processing it in huge number of computers available during the network and the processed data is aggregated.

6.5. Fifth challenge, limitations of traditional encryption approaches: However there are many of encryption offerings around, most of them engage in one specific aspect. For example we can use transparent data encryption capabilities from our data base vendor, but what happens when that data gets exported to big data environments? Also, what about all other data sources and systems in play? we also have to know if the vendor store the keys with the data or no ?. some vendors offer big data encryption capabilities , It secure only specific big data nodes ,not the original data sources that are fed into big data environments or the analytics that come out of the environments, Further ,the encryption in big data offerings not secure for the configurations information and also for the log files.

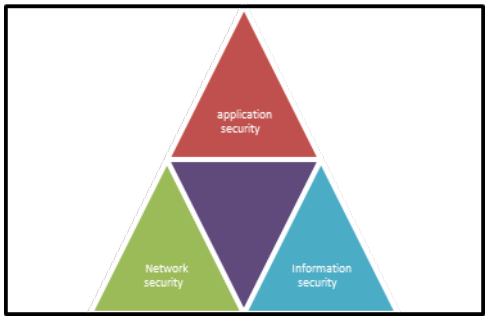
6.6. Sixth challenge, Reporting: When huge amount of data are involved because the Traditional reports display of statistical data in the form of numbers, It would be hard to interpret by human beings. To get over this matter we need to represent the reports in a form that can be easily recognized by looking into them.

7. What is cyber security?

Cyber security, we can also call it information technology security, focuses on protecting computers ,networks, programs and data from unintended access, change or destruction. Another definition; Cyber security is the body of technologies ,processes and practices designed to protect networks ,computers ,programs from

attack ,damage or unauthorized access .In computing context ,The term security implies cyber security .According to a December 2010 analysis of U.S spending plans ,the federal government has allotted more than 13$ billion to cyber security over the next five years.[7-1]

Why is cyber security Important? Governments, corporations, financial institutions ,hospitals and other business collect ,process and store confidential information on computers and transmit that data across networks or other computers .Because of difficulty and the growing volume of cyber attacks , constant attention is required to protect sensitive personal information and business as well as safeguard national security .On March 2013 The nation's top intelligence officials warned that cyber attack and digital spying are the top threats to national security ,eclipsing terrorism.[7] Ensuring cyber security requires corresponding efforts thought an information system .We can mention these elements of cyber security : Application security, Information security, Network security, Disaster recovery business continuity planning and End user education.

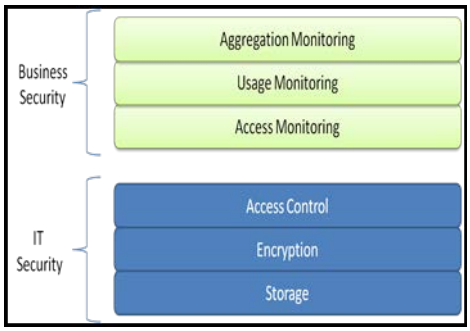


figer5 - Elements of cyber Security

One of most challenging element of cyber security is the quickly and constantly evolving nature of security risks. The traditional approach was focusing on resources on the most important system competent and protect against the biggest known threats, which causes leaving some less important system components undefended and also some dangerous risks not protected against. Such as approach in insufficient in current environment. Adam Vincent, CTO-public sector (a security services provider to federal agencies including Defense Department organizations) described the problem "The threat is advancing quicker than we can keep up with it. The threat changes faster than our idea of risk .It's no longer possible to write a large white paper about the risk to particular system. You would be rewriting the white paper constantly."[7-2]

8. Securing big data

Cyber criminals are never going to stop attacking, and with such a big target to protect, it is cautious for any enterprise utilizing big data technologies to be unbeaten as possible in securing its data. [8]. The big data and its technology get deeply involve into the enterprise many questions around compliance and security raise up , Protecting the information has been massively important ,The hygiene factors in securing the information still can be accessible to people with right permissions , Because of that its still at risk from cyber or social engineering attacks. The means that security of big data is about layering also organizing it in two types: business security and IT security.



figer 6- TYPES OF SECURTY

The IT solutions: means technical solutions which can be helpful accessing control, encryption but they do not really help to prevent information being stolen.[8] .There is two different type security:

- IT Traditional.
- ACL -Access Control List.

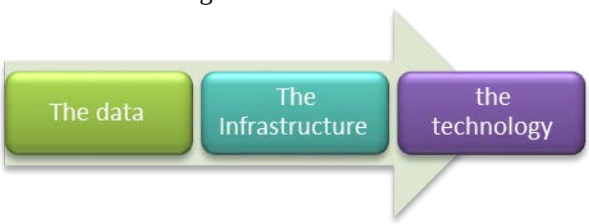
Both are done to match traditional solutions with RDBMS but unfortunately stop in term of security which means; they do not address the real threats with cyber-attacks .An ACL Its good if people do what they are expected to do Approaches part of the problem not the solution.

NO human defined and managed system can handle this huge amount of information, but Machine Learning algorithms have just munch up this sort of data and create the models .It is for sure more massive and complex than the sorts of ACL's ,encryption criteria and basic security policies that IT is used to. These Algorithms needs more monitoring. The primary point is that encryption and ACLs provide only a basic hygiene factor when it comes to securing big data. The risk and value of information is increasing and by creating Big data solutions businesses are creating more valuable data, and there for more at risk information solutions. This means that Information security needs to become a fundamental part of information Governance and that new ways of securing that information are required.

This is where big data comes to its own rescue through the use of large data sets which enable new generations of algorithms to identify and alert based on the risk and the right way to handle it .That needs to think about Information Security as a core part of the Metadata that is captured and governed around information.

Big data security is a new generation of challenges and new generation of risks, these requiring new generation of corporation and new generation of solutions so the information security is not left to few people in the IT department.

Securing big data comes with it special and exclusive challenges further than a high-value target. Big data security challenges arise because of incremental differences, not fundamental ones. The different between big data environments and traditional data environment include: The data collected, grouped, and analyzed for big data analysis. The infrastructure used to store and house big data. The technologies applied to analyze structured and unstructured big data



figer7- The different between Big data environments and traditional data environments

First ,The data : each data source will likely have its own access restrictions and policies ,making it hard to balance appropriate security for all data sources with the need to collect and extract meaning from the data .For example ,a big data environment may include a dataset with proprietary research information, or dataset requiring regulatory compliance, and separate dataset with personally identifiable information (PII).Protecting big data requires balancing analysis with security requirements on case-by-case basis.

Also, many of repositories collect data at high volumes and velocity from a number of different data sources; also they could have their own data transfer workflows. Multiple repositories with these connections can increase the attempt to attack surface for adversary. A big data system receiving feeds from 20 different data sources may present an attacker with 20 viable vectors to attempt to win an access to a cluster.

Second, The Infrastructure is another big data challenge, It is the distributed nature of big data environments. Related with a single high-end database server, distributed environment are difficult and defenseless to any possible attack. The big data distribute geography ,physical security controls need to be standardized across all accessible locations .Scientists in any organization might need accessing to information, perimeter protection would be very necessary also complicated to ensure access to users while protecting the system from possible attacks. The possibility that the configuration of servers may at risk and not be stable would increase.

Third, the technology, especially big data programming tools, including Hadoop and NoSQL databases, were not originally designed with security in mind. Let us take hadoop for example, originally did not authenticate services or users, and didn't encrypt data that's transmitted between nodes in the environment. This would cause weakness for authentication and network security. However NoSQL, database lack some of security features provided by traditional databases, such as role-based access control. The advantage of NoSQL is it allows for the flexibility to include new data types on the fly, but defending security policies for this new data is not straightforward with these technologies.

How can we use Big data to stop cybercrime? Lately threat monitoring .Automated information collection and intelligence sharing. Smarter links between many security systems and layers (including physical protections). A continuous and real-time loop of monitoring data and behavioral signals. Many cyber security experts think that it is necessary and useful weapons to fight the plague of data breaches. So –called 'security analytics' would allow systems to adjust automatically their risk profile (i.e., go on high alert) once any system in the “threat intelligence network" detects a threat –be it malware a reprobate bordering or distrustful log activity.

- Beyond big data to data-driven insights for cyber security: Big data and analytics are the most effective defenses against cyber interruptions. Better, faster, actionable security information reduce the critical time from detection to remediation, enabling cyber warfare specialists to proactively defend and protect the network. Teradata delivers a single, automotive ecosystem integrating information security, cyber security, network operations data, analytics and reporting.

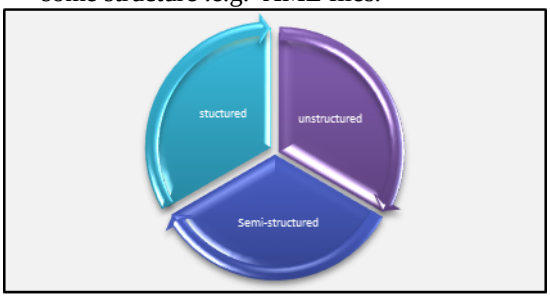
- Is more Cyber Defense always better? Challengers, or hackers continue developing a powerful attacking moments to ruin what considered to be high effective cyber defenses .Today with available hacker’s resources, they can easily move around a defense –in-depth strategy to break data systems. A common response to evolving attacks is to either add more security tools or increase the sensitivity of the security tools already in place – or both.
- Unfortunately, as cyber-attack worsen and businesses respond with greater force, existing staff resources are taxed, yielding less effective security results.
- Cyber strategy: Evolving Cyber Threats Demand a New Generation of Cyber strategy; Hackers today have evolved from hobbyists to professionals. They are well

trained and well funded, and also they can run the range from social activities and state-sponsored operators to criminal syndicate members .As a result, cyber crime is on the rise, costing $118 billion in business losses annually and climbing .Security professionals used to be confident they could lock down and secure their networks to prevent cyber attack. Now the attitude is that cyber attacks are expected .And the load is on the IT teams to find out how to detect and remediate an attack before data is compromised. [7-3]

9-The Top security tools In fight against cybercrime.

Before we talk about the tools and technologies let us recognize the different types of data we are processing it:
we have three different types of data

- ✓ structured Data: This is the data that we used in traditional RDMS and is divided into well-defined structures. It has schema that will be checked during write & read operation .e.g. Data in RDMS like Oracle, My SQL server est.
- ✓ unstructured: This is important to understand, 80% of world data is unstructured or semi structured .These are the data which are on its raw form and cannot be processed using RDMS. Example: Face book, twitter data. Data does not have any structure and it can be any form –Web server logs ,E-mail, Image sect.[9]
- ✓ Semi-structured: Data is not strictly structured but have some structure .e.g. XML files.



figer 8-Kinds of data

the challenges faced in processing Big data technologies are overcome by using various techniques. The most popular techniques used like:

- Regression: is used in predicting values of dependant variable by estimating the relationship among variables using statistical analysis
- Nearest Neighbor: in this technique the value are predicted based on the predicted values of the records that are nearest to the record that needs to be predicted.
- Clustering: it involves group of records that are similar by identifying the distance between them in an n-dimensional space where n is the number of variables.
- Classifications: is the identification of category/ class to which a value belongs to ,on the basic of previously categorizes values.[9-1].Several Open source tools exist which help taming Big data some of top tools are mongo DB-(across platform document oriented data base management system)Hadoop- Map Reduce- (will be explained.)-Orange (python based tools for processing and mining big data) - weka (java bases tool for processing large amount of data, algorithms used in mining data). - SAP HANA (is a priority in-memory RDBMS capable of handling large amount of data). [9-2].let us take closer look to some of these tools like:

A. **Hadoop:** Hadoop Is a software framework that can be installed on commodity Linux cluster to license large scale distributed data analysis .No hardware modification is needed other than possible changes to meet minimum recommended RAM ,disk space ,etc. Requirements per node. The initial version of Hadoop was created in 2004 by Dough Cutting (and named after his son's elephant).Hadoop became a top-level Apache Software Foundation project in January 2008. It provides the robust, fault-tolerant Hadoop Distributed File System (HDFS), inspired by Google's file system, as well as Java –based API that allows parallel processing across the nodes of the cluster using the Map Reduce paradigm .Use of code written in other languages, such as Python and C is possible through Hadoop Streaming, a utility which allows users to create and run jobs with any executable as mapped and /or the reducer. Furthermore Hadoop comes with Job and Trackers that keep track of the programs, execution across the nodes of cluster. So Hadoop is a frame work that allows for distributed processing of large data sets across clusters of computers using simple programming modules .There are four modules in Hadoop:

1. Hadoop Common: the common utilities that support the other Hadoop modules.
2. Hadoop Distributed File System (HDFS): distrusted file system that provides high-throughput access to application data.
3. Hadoop YARN: a frame work for job scheduling and cluster resource management.
4. Hadoop Map Reduce: AYARN – based system for parallel processing of large data sets.

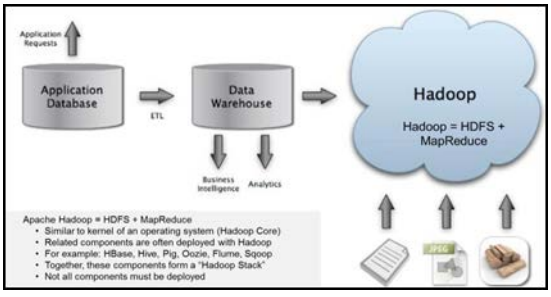
HDFS file system –There are some drawbacks to HDFS use. HDFS handles continuous updates (write many) less well than traditional relational database management system. Encompasses HDFS (Hadoop Distributed File System) and Map Reduce programming framework.

- ❖ Why Hadoop Because ; Hadoop is very useful but it's not the only way of interacting with the data. This

reason Hadoop is the industry standard for handling big data because it's very scalable. As more you throw of computing resources at it as more you would receive a better performance and higher data processing capabilities working with hadoop primarily might feel a little discouraging and the barriers to use appear higher especially because of the Java API. It's unimportant to increase more resources; scaling five nodes cluster to thousand node cluster does not terribly increase the load to the administrate. You do not have to worry about the spiteful issues associated with big data spanning across multiple disks and multiple machines .It takes care the resilience, fault tolerance, and scalability issues. Hadoop is also fault tolerant. It a disk or node, or even a whole rack of nodes goes down your data is replicated across the cluster in such way that prevents losing any data. Also the running jobs that processing data also are fault tolerant, restarting tasks when necessary to ensure that all the data is correctly processed. In addition to Hadoop itself, There are multiple open source projects built on top of Hadoop .Major project are described such below

table 1

Hive	▪ Created for people with SQL background, Data ware house developed at Face book used for hoc querying with SQL type query language and for more complex analysis. Users define tables and columns. Data loaded into retrieved reports, analysis .It designed for batch offer real-time queries. The queries written are similar to SQL. ▪ It process completely structure data not for unstructured data.
PIG	▪ A high –level data flow language (Pig-Latin) and execution framework whose compiler produces sequences of Map reduce programs for execution within Hadoop, its batch for processing data. ▪ It can be used for both structured and unstructured data.
https://developer.yahoo.com/blogs/hadoop/comparing-pig-latin-sql-constructing-data-processing-pipelines-444.html	



figer9

B. Map Reduce: we can also use Map Reduce (Map Reduce has their own vision of Hbase), in - memory databases and analysis frameworks like Spark and Shark, as well as graph databases like unlimited Graph and Titan. This is by no resources a complete list, the point is different use cases needs different tools and many different ways of interacting with big data. Hadoop can be used for everything, is the hit for which everything is a secure but it's not necessarily the perfect tool for every use case. Map Reduce highlights parallelism in data retrieval .Jobs are parceled out or 'mapped' to a number of subsidiary nodes, with result handed backup (reduced) in the ultimate of the original tasks.

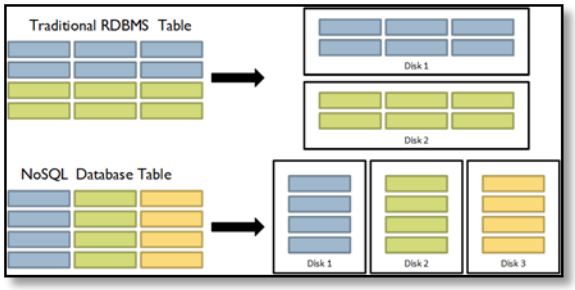
C. **NoSQL:** It's a platform such as Cassandra , Mongo DB, and others .It's techniques that directly address some of limitations of traditional .relational data store when analysis of a body of data is the priority .They can be highly distributed systems ,developed in many cases to deliver better performance in data management and retrieval at Internet scale .They are well suited for the analysis of an entire body of recorded data to discover patterns, trends, and anomalies which makes them compelling candidates for handling large and diverse bodies of security-relevant data. These techniques incline to hold high distributed, fault-tolerant architecture on commodity hardware. Historically businesses have been storing the data in relational databases with some normalization and IT groups are contended maintaining and querying in this model. With new tidal wave of data that organizations are looking to store and take advantages from NoSQL solutions are becoming much more essential due to the resilience and scalability factors. All of the NoSQL solutions fundamentally offer an system just like other technologies can be at the top of the mass in hierarchy.

There are two important factors when analyzing large data sets and Hadoop does both well:

- the ability to run processes across multiple nodes in the data center and file system that can store results in single view. A lot of companies have tried their own techniques and technologies to achieve this but, with using Hadoop they really became satisfied these two requirements in a dependable, open -source way.
- The barriers to use Hadoop is understanding the API and architecture, setting it up, and using the eco-system have become considerably lower. For example using Map Reduce abstracts away a lot of the issues that initially made Hadoop difficult to use. It's becoming easier to quickly start using Hadoop primarily due to additional companies creating tools and APLs with their own optimizations and stacks that let us focusing on business problems and less on the infrastructure.

Another advantages of NoSQL is It's great tolerance for flexibility in embracing a wider variety of data and data structure compared to relational systems, which often

requite data to conform to a defined schema on (or before) ingestion. When structure is needed to define a dataset for analysis, NoSQL techniques allow more precise definitions later in the process.
This further suggests how NoSQL techniques may enable security organizations to take in a wider variety of data, such as unstructured data from both internal and external sources, and binary content such as images or video (which may be managed via the application of structured or semi-structured metadata).



figer 11-comparison between Traditional RDBMS table and NoSQL database table

10-applications of big data techniques in learning

Firstly the learning started in the classroom was based on three models:

table 2

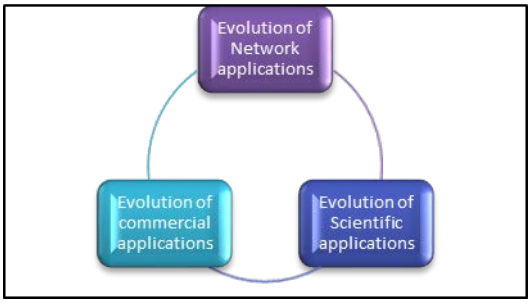
Behavioral	Cognitive	Constructivist
It relay on observe changes in student behavior to assess the learning out come	It based on the active involvement of teacher in learning	The student have learn on their own from knowledge available to them.(learning in network)

Recent learning methods depend greatly on online activities (on line learning management systems) to improve the learning experience .students has started using smart phones to access learning content. As the learning environment have become accessible anywhere through the internet. The data available from student activities and also the data of educational institutions which use applications to mange courses, classes and students. The data is so huge so the educational institutions start using "big data" technologies to process the educational data.

- **Application Evolution**: The propose of big data and big data analysis is describing datasets and analytical technologies in large-scale complex programs, which need to be analyzed in advanced analytical methods.

The data driven applications have appeared in the past decades for example ,as early as 1990's business intelligence has become a prevailing technology for business application and network search engines based on massive data mining processing emerged in the early twenty-first century .Some potential and influential applications from different field and their data and analysis characteristic as follows:

- ***Evolution of commercial applications:*** Evolution of commercial applications: The past business data was mostly structured data, which collected by companies from old systems and stored in RDBMSs. Analytical technologies used in such systems were prevailing in 1990's and it was very simple .e.g., reports, instruments panels, special queries, search-based business intelligence, online transactions processing, interactive visualization, score cards, predictive modeling, data mining.
- ***Evolution of Network applications***: Text analysis ,data mining, and webpage analysis technologies have been applied to the mining of email contents and building search engines ,nowadays ,most applications are web-based ,also their application field and design goals. Network data accounts for major percentage of global data volume . has become a common platform for interconnected pages, all kinds of data, like (text, images, videos, pictures, and interactive contents. etc.).So many advanced technologies used for semi-structured or structured data emerged at the right moment. For example, the image analysis technology may extract useful information from pictures e.g. face recognition. Multimedia analysis technologies can be applied to the automated videos systems for business, low enforcement, and military applications. Different users might use daily and celebrity news publish their social and political opinions, and provide different applications with timely feedback.
- ***Evolution of Scientific applications***: Scientific research in many fields causing massive data with high-throughput sensors and instruments, such as genomics, ocean logy, astrophysics, and environmental research. The U.S National Science Foundation (NSF) has announced the big data research Interactive to promote research efforts to extract knowledge and insights from large and complex collections of digital data.



Figer 12- Application Evolution

Conclusion:

Big data technologies are changing the whole world, everything from internet of things to gathering both more qualitative and more quantitative data will lead to better decision-making and insight. By leveraging big data technologies effectively ,organizations can be more efficient and more competitive .Privacy advocates and data organizers criticize the history of big data as they watch the growing ubiquity of data collection and increasingly tough uses of data enabled by powerful processors and unlimited storage .Researchers ,business ,and entrepreneurs strongly point to concrete or anticipated innovations that may be dependent on the default collection of large data sets. Also the quick growth of the internet has bought with it an exponential increase in the type and frequency of cyber attacks. Many well-known cyber security solutions are in place to counteract these attacks.

The huge argument today is how should privacy risks be weighed against big data rewards?

especially the recent controversy over leaked documents revealing the massive scope of data collection, analysis .Big data creates tremendous chance for the world economy not only in field of security ,but also in marketing and credit risk analysis to medical research and built-up planning. At the same time the unexpected benefits of big data are tempered by concerned that advances of data ecosystem will turn over the power relationships between government, business and individuals, and lead to racial or other profiling. Discrimination over criminalization, and other restricted freedoms.

Finally: It is really very important to understand the security and privacy implications resulting from big data implementations supporting non information security functions. Specifically, security required executives should be aware of who Big data increases attack surface of hackers and understand how to protect against link ability threats.

Insert acknowledgment, if any.

References:

[1] Motivation for Big Data-Pro Apache Hadoop-Date: 08 September 2014-Sameer Walker Affiliated with, Madhu Siddalingaiah. http://link.springer.com/chapter/10.1007/978-1-4302-4864-4_1#page-2

[2] Undefined By Data: A Survey of Big Data Definitions-Jonathan Stuart Ward and Adam Barker-School of Computer Science-University of St Andrews, UK{jonthan.stuart.ward, adam.barker}@st-andrews.ac.uk-http://arxiv.org/pdf/1309.5821v1.pdf

[3] BIGDATA ANALYTICS - 5th QUARTER BY PHILIP RUSSOM -TDWI research. http://www.tableau.com/sites/default/files/whitepapers/tdwi_bpreport_q411_big_data_analytics_tableau.pdf.

[4] Big Data Analytics for Security- Posted by Alvaro A. Cárdenas, Pratyusa K. Manadhata, Sreeranga P. Rajan-http://www.infoq.com/articles/bigdata-analytics-for-security.

[5] http://www.umuc.edu/cybersecurity/about/cybersecurity-asics.cfmumuc[7-1]http://whatis.techtarget.com/definition/cybersecurity

[6] http://whatis.techtarget.com/glossary/Security-Threats-and-untermeasures

[7] teradata-http://www.teradata.com/Cyber-Security-http://bigdata.teradata.com/US/Success-Stories/Innovations-and-Insights/

[8] Securing Big Data - Part 1-Posted by Steve Jones at Tuesday, January 06, 2015

[9] http://service-architecture.blogspot.com/2015/01/securing-big-data-part-2-understanding.html.

[10] unstructured data in big data environment-http://www.dummies.com/how-to/content/unstructured-data-in-a-big-data-environment.html

[11] http://ictactjournals.in/paper/IJSC_Paper_6_pp_1035_1049.pdf

[12] ICTACT JOURNAL ON SOFT COMPUTING: SPECIAL ISSUE ON SOFT COMPUTING MODELS FOR BIG DATA, JULY 2015, VOLUME: 05, ISSUE: 04 1035

[13] APPLICATION OF BIG DATA IN EDUCATION DATA MINING AND LEARNING ANALYTICS – A LITERATURE REVIEW -Katrina Sin1 and Loganathan Muthu2-1Faculty of Education and Languages, Open University Malaysia, Malaysia-http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-12-1

[14] -https://developer.yahoo.com/blogs/hadoop/comparing-pig-latin-sql-constructing-data-processing-pipelines-444.html

[15] M.Chen et al., Big data: Related Technologies, Challenges and future Prospects,:"-Springer Brief in computer science-http://link.springer.com/chapter/10.1007/978-3-319-06245-7_6#page-2.