CLOUD COMPUTING PAPERS FA18

Gregor von Laszewski
Geoffrey C. Fox

laszewski@gmail.com

Cloud Computing Papers

Gregor von Laszewski

Cloud Computing Papers

1 Distributed TensorFlow FA18-523-68

Selahattin Akkas
sakkas@iu.edu
Indiana University
hid: fa18-523-68
github:

- this is a draft, review has not been started due to this
- second review. No forther review possibly before we grade.

Keywords: Distributed TensorFlow, TensorFlow

1.1 Abstract

It is non-practical to do computation on a single machine for Big Data applications. Likewise, it is also non-practical to train machine learning algorithms using large datasets on a single machine. One of the widely used Deep Learning framework TensorFlow supports distributed learning. In this paper, Distributed TensorFlow's architecture will be explained.

1.2 Introduction

1.3 Parameter Server

1.4 TensorFlow Cluster

1.5 Parameter Server

1.6 Shared Variables

1.7 Synchronus Data Parallelism

1.8 Asynchronus Data Parallelism

1.9 In-graph replication

1.10 Between-graph replication

1.11 Asynchronus training

1.12 Synchronus training

2 Big Data Analytics in E-commerce

Bo Li
bl15@iu.edu
Indiana University Bloomington
hid: fa18-523-85
github:

Keywords: E-commerce, Consumer Behaviors, Python, TensorFlow, Big Data, Deep Learning

2.1 Abstract

In recent years, online shopping has become a more popular way of consuming. Traditional retailers are eager to find a way to maintain revenue, online retailers are also finding ways to extend the market. The rise of ‘big data’ had impacted on marketing research and practice a lot. As technology developed, we have huge data on consumer behaviors which could be very detailed and accurate, but how to mine the data is a problem. In this article, we talk about the application of big data in the consumer behaviors data, subsequently discuss how the TensorFlow can translate the data into valuable conclusions in consumers’ behaviors research.

2.2 Introduction

The Big Data has a common definition, Big Data always comes with 3V: volume, velocity, and veracity. The Internet allows us to do almost all work online and keep records of our actions. If you listen to a song in the playlist, maybe iTunes will record it as part of your individual activity log, which could be the dataset that explores your interest. If you often use Uber to commute between your home and your company, maybe they could picture your daily life including the places that you have spent time in. If you use your device to safari on the internet, your action of clicking on several links could also be recorded and researched since the action contains you using habits and preferences [1].

The online shopping data includes the consumer's all kinds of information: age, job, education, catalog preferences, price sensitivity, etc. But some of them are not presented to us directly, mining the consumer behaviors is an appropriate way to get access to the hidden information. How could we mine something meaningful to explore consumer behaviors and provide valuable insight? The answer lies in several using cases and the understanding of market research and also human psychology research.

In order to extract the knowledge behind the commercial data generated by hundreds of thousands of consumers for the use of leading managers to make the decision, it is necessary to conduct a deep analysis to the commercial data, instead of generating simple reports [2]. The deep analysis could hardly be done by SQL since the process relies on complex models. Without those models, it is impossible to get a profound understand of the commercial data [3]. People will not only need to find out what is happening now, but also need to use data to make some predictions in order to make preparations for the future events. For example, if the manager is able to predict the loss of the customer in the future, they can use discount to attract the users again [4].

In the context of big commercial data, the traditional OLAP operations are not enough anymore to meet the requirements, we also need path analysis, time series analysis, graph analysis, what-if analysis and some complex statistical models. Time series analysis, a useful method in the commercial data analysis since we have got lots of the trading historical data. The managers want to get some patterns in the data in order to fine some chances to improve the revenue. By the trend analysis, they can even predict some chances in advance. In the financial area, analysts are able to develop some software to conduct the time series analysis of the trading data, and find some profitable trading patterns. After further verification, they can use those profitable trading patterns to conduct the real trade and make profits.

2.3 Technology Background

TensorFlow is an open source software library for numerical computation using data flow graphs [5]. TensorFlow could help developers to transform from code to graph, which could benefit developer in understanding their work, and the term tensor, is generated in the process, as the tensor will go from the beginning

to the end of the graph, so the technology is called TensorFlow. The process of computation could be done in CPU or GPU, as we know, in blockchain, GPU works better than CPU since the fundamental design of GPU fits better in the computation of mining coins. TensorFlow also has a data visualization module called TensorBoard, which contains the common drawing tools as well as some useful templates for the developers to visualize their data. There is no doubt that the graph will be more clear than codes especially when the structure of data is very complex. And the graph could give the readers a direct presentation of the data, which is worthwhile since it could reduce the communication cost between different developers.

There is a team in Google called Brain team, which is the initial developer of TensorFlow, there original purpose of the development of this module is to improve the efficiency of machine learning. For example, if the deep learning task is to predict a result based on a training dataset, the more layers you have, the more accuracy you will have. But more layers will cost much more time, so TensorFlow is created to solve the problem. One more important thing is that in the previous version of TensorFlow, it began to support distributed computing which means more resources could be deployed in the process so the efficiency will be boosted.

To support more developers, Python API and C APIs are also available in TensorFlow. One thing that must be pointed out is that, although TensorFlow supports many kinds of languages, the Python API is the most efficient one since Python does better in the feedback process which focuses on improving the model. And there are also more examples in Python since most of the machine learning work is done by Python.

2.4 Big Data Applications in E-commerce

Since the design of the website is getting more complex than before, the users may conduct different operations in different pages of the website, but most of them are very import to provide an essential information for us to find the customer’s preferences.

> ***“There is a way to achieve that which is called four rights. Talk to the right audience, through the right channel, with the***

> ***right message, at the right time" [6].***

> ***"Customer acquisition: Marketing will target high-value customer segments identified by behavior analytics and study behavior patterns to determine the best potential offers. Customer engagement: Behavior patterns will be used to generate personalized next-best, cross-sell and up-sell offers, while behavioral customer segmentation will be used for more general customer marketing offers. Customer retention: Behavior patterns will be used to detect possible customer churn and generate next-best retention offers" [6].***

The strategic meaning of big data is that deploying professional analysis on those meaningful datasets generated from the E-commerce trading. Some meaningful things are hidden behind the big data, mining them is the main task of the application. The improvement in the value of data is the most important part of the benefit of using big data, especially compared with the previous situation, that the data useless since the huge amount [7]. Such information has been stored in the log document, but it is too massive and fragmented to analysis it with limited technologies and techniques.

Big data could help the enterprise to have a more profound understanding by analyzing users' behavior data, which allows the enterprise to establish strategies with more specific aims. This could make the enterprise to be competitive in the market and win more consumers' hearts. For example, the user may want to buy a guitar for himself, and he has browsed several kinds of guitars and could hardly to make his decision. In the E-commerce domain, users have generated huge amount data about their every action: browsing products, clicking on the details of products, adding the products into the wishing list, adding the products into the cart, delete the products from the wishing list, clicking dislike product button and querying more information to the seller of the products [8]. The big data technologies could base on his operation history to find his acceptable price level and the specific version of guitar (such as with or without pick up system), then we can put such requirements into our database to find the appropriate guitars and push those potential choices to the user's interface, which could realize an improvement in sales transformation rate.

TensorFlow could be a kind tool to analyze consumers' behaviors since the

model of recommendation is doable in TensorFlow. Due to the feature of distributed computation, the efficiency of the model is good enough for a small scale recommendation system. To have a more profound understanding of user behavior, a whole lifecycle of the user is needed to be established for the analysis of user behavior.

2.5 Features of User Behavior in the E-commerce Platform

The online platform is the main platform for e-commerce which lists the product in different ways and provides the whole chain of finishing the trade. Different from traditional commerce, online e-commerce has some special features which could be used in the big data.

There are fewer limitations for consumers in the B2C pattern since the online platform can run for 24 hours if it is well maintained. Consumers are able to conduct any operation (browsing, selecting, finishing the trade) at any time in anywhere.

The trading cost is much less than the traditional commerce pattern. For the consumer, time cost, transportation cost and delivery cost are lower than the traditional commerce pattern. Their trading action is much simplified by the online shopping systems, the trade can be done by several clicks on the mobile device.

The online product can offer a more attractive price due to the advantages of the internet. Comparing with the traditional commerce, online sellers have less item to pay for maintaining the shop. There is a lot of extra costs for the real store.

Customized service. The recommendation system is able to recommend the most wanted goods for each customer based on their user behavior and the big data technology, which is the traditional commerce cannot achieve due to the cost. The customized service can benefit a lot in the transformation between browsing and buying.

More kinds of product and no space limitations. Since the information of the product is much smaller than the product itself, and the online store can exhibit

them all at the same time, so there are more choices presented for the consumers.

The information is easy to get. Every item in the online platform has been labeled by the system, so the search of the item is very convenient for consumers, the cost is significantly lower than the traditional commerce.

2.6 APPLICATIONS

Due to the hotness of machine learning and deep learning, there are a lot of applications in every domain. Search ranking and recommendation are the most common two applications.

> ***"Recommendation systems in particular benefit from specialized features describing past user behavior with items" [9].***

Just like search ranking, recommendation systems also have a problem of the balance between memorization and generalization. Memorization can be seen as the representing of the relationship between the products and users, which can be extracted as vectors. Generalization is to generate rare feature combinations in order to serve for the recommendation systems [10].

TensorFlow, with so many advantages in machine learning, is very appropriate for the recommendation system. Since the features of products could be learned by multi-labels classification, and the user's features could be learned in his historical actions in the online platform, which has the record of every consumer's trading history. When we have both features of products and users, we can establish a recommendation system by matching the two objects. Besides, the model can be judged by the dot product between the two vectors.

To establish such a recommendation system, we need to fit the TensorFlow since it is tensor that flows in the whole model. The transformation from dataset to tensor is a necessary step to conduct the model. And the users', as well as the products' character tensors, need to be transformed into the presentations of users and products by the embedding function. The next step is to generate the recommendation by the pair the presentations of users and products. Such pair of presentations contains the most match user and product calculated by the model,

the vectors in the model contain all the information of the user and the product. The last step is to compare the generated score and the actual comment from users to define the result's quality, which is called loss function [11].

TensorRec scores recommendations by consuming user and item features (ids, tags, or other metadata) and building two low-dimensional vectors, a "user representation" and an "item representation". The dot product of these two vectors is the score for the relationship between that user and that item,the highest scores are predicted to be the best recommendations.

The representation function in TensorRec can be set up by developer's preferences, it could extract the features of users as well as products. It can be very convenient for developers to set the parameters independently since the scenario varies in different cases [11].

2.7 CONCLUSION

Information is booming in recent years, data and internet techniques are spreading in everywhere, with the significant effect on consumers' deciding pattern and purchasing pattern. The digital economic on the internet has become the focus of all the domain. For the biggest group in the digital economics in the internet, online consumers are the focus in the specific domain. How to draw the picture of the users and get the key feature of their behaviors have become a hot topic. Besides, the social network analysis is also important in the commercial data analysis. Different buyers may be divided in different groups, the members in the same group could have the same interests. Social network is the study of social entities (people in an organization, called actors), and their interactions and relationships [12]. The interactions and relationships can be represented with a network or graph, where each vertex (or node) represents an actor and each link represents a relationship. From the network we can study the properties of its structure, and the role, position and prestige of each social actor. We can also find various kinds of sub-graphs, e.g., communities formed by groups of actors [13]. Social network analysis is useful for the e-commerce because the group of buyers is essentially a virtual society, and thus a virtual social network, where each page can be regarded as a social actor and each hyperlink as a relationship. Many of the results from social networks can be adapted and extended for use in the Web context [14]. The ideas from social network analysis are indeed

instrumental to the success of Web search engines.

Benefitted by the internet, we have all the records of most of the online activities of users, but the relationship between those actions and the user's features is ambiguous. Basing on the TensorFlow, we are able to use a low-dimensional vector to represent the user's features as well as the products' features. The algorithm allows us to extract the key point of users as well as products, which provide a base for the recommendation of the product.

The big data technology can help us to mine the black box of the relationship between the actions and the features. Several factors which measure the user's preferences can represent the user, and those factors are also the key parameters in the model. Once we get a clear picture of the user, we are able to customize the recommendation, which can not only improve the user's experience and also improve the revenue of the online retailer.

3 SAP FA18-523-86

Jeff Liu
liujeff@iu.edu
Indiana University
hid: fa18-523-86
github:

3.1 Introduction

ERP (Enterprise resource planning) [15], a systemized management theory based on information technology, has become into an important and popular modern enterprise management tool for providing a management platform for decision-making operations for enterprises. A good ERP system, it is just a set of software but a management idea. It can not only fully adapt to the management and business processes of the enterprise, but also achieve rapid deployment and challenges in technology. SAP) [16] is a leading ERP software, most of the world's top 500 are in use, aalthough the SAP license, maintenance updates and related training will cost a lot of money, but it come up with the improvement of operational efficiency and information processing cost saving, so SAP becomes a first choice for the business operation of large enterprises. SAP (Systems Applications and Products in Data Processing) is ERP software, from the back end to the company management level, from the factory warehouse to the storefront, from the computer desktop to the mobile terminal, SAP provides ERP solutions, and can provide comprehensive services for enterprises of various industries and different levels.

3.2 Implementation

SAP Modules and Functions: There are 2 Types of SAP ERP Modules. Number one is Functional Modules and second one is Technical Modules. All SAP Modules integrated with each other with functionality and provide us best solution for Business[17]. Most important SAP Modules that Bunnies implement for their business are

1. SAP FICO module
2. Human Resource Management (SAP HRM), also known as Human Resource (HR)
3. Production Planning (SAP PP)
4. Material Management (SAP MM)
5. Financial Supply Chain Management (SAP FSCM)
6. Sales and Distribution (SAP SD)
7. Project System (SAP PS)
8. Financial Accounting and Controlling (SAP FICO)
9. Plant Maintenance (SAP PM)
10. Quality Management (SAP QM) security module [18].

SAP Functions: 1. SAP Business Objects provides comprehensive business intelligence capabilities that give users the ability to make effective and informed decisions based on solid data and analysis results. All users from high-level analysts to ordinary business users have access to the information they need, with less IT support [19]. 2. SAP CRM can help you reduce costs and improve decision making while helping enterprises differentiate to gain a long-term competitive advantage. It helps to increase the competitive advantage and bring higher profits [20]. 3. SAP Business Objects Information Management provides comprehensive information management capabilities to help deliver consolidated enterprise data in a timely and accurate manner, both structured and unstructured. it helps users provide data for key action plans such as business transaction processing, business intelligence, data warehousing, data migration, and master data management 4. SAP Business Objects helps you leverage the value of your company's data and make your business more agile and competitive by increasing your organization's collaboration, insight and confidence [19]. 5. SAP ERP is one of the top five suites of SAP Business Suite and the most powerful core suite of SAP in the market. The SAP ERP application software supports the basic functions of the business process and operational efficiency of the enterprise and is customized to their specific needs [19]. 6. SAP HR supports the entire process of recruiting, deploying, developing, motivating, and ultimately leaving valuable employees, improving these processes from the beginning to the end [21]. 7. SAP PLM, one of the core suites in the SAP Business Suite, provides collaborative engineering, custom development, project management, financial management, quality management and more throughout the product and asset lifecycle. 8. SAP Supply Chain

Management is a member of the SAP Business Suite. The suite uses modular software that works with other SAP and non-SAP software to enable organizations to perform basic business upgrades. 9. SAP Supplier Relationship Management SRM is a sub function of the SAP Business Suite business application. This integrated suite expands the value of SAP Business Suite by automating the process of commodities and services from purchase to payment [20].

3.3 CONCLUSIONS

In summary, SAP is an ideal EPR tool for big companies. SAP system is very expensive, but the system is also very powerful, and can be adjusted differently for each customer's different needs. First, SAP's did good at customer management and carefully examine the customer's relevant information, asset status and so on. Secondly, SAP's most powerful and outstanding function is sales management, from order opening, process determination, cost analysis, performance tracking, delivery arrangements, accident handling, payment tracking, a series of powerful and powerful Features.

4 Title: Big Data Analytics in Computer Vision fa18-423-08

Yuli Zhao Indiana University Bloomington, Indiana yulizhao@iu.edu

format incorrect

github:

4.1 Abstract

Computer vision is designed to imitate human vision. In order to imitate human vision, enabling computer to see is not enough, computer needs to be able to analyze what it sees. What computer sees is essentially frames of images which means nothing to it, but computer is able to analyze the images and extract useful information out following specific instructions from the programmer. Due to the nature of computer, it is able to perform analyzes on a large scale with specific instructions. This paper will discuss how computer vision is used in determining water turbidity through recognizing the secchi disk and extracting measurements as useful information, and how computer vision is applied to a big data set.

4.2 Introductions

Secchi disk is an eight inch circular disk has alternating black and white color on the surface. It is used to determine the turbidity or the transparency of the water. The way it works is that secchi disk is lowered into the water slowly using a tape that has measurements, and the researcher would record the measurement when the secchi disk is no longer visible. But one single measurement does not offer many information to be useful, useful information can only be extracted by researchers when there are many more measurements. But it would take researcher enormous amount of time to keep in track with each and every measurement. And computer vision is needed here to replace all that workload. Instead of having researcher record every measurement, a camera can be placed by the measurement that has visual on both the tape measurement and the secchi

disk. The process of lowering secchi disk can be repeated in different locations, and each process will return result in the format of a video. And computer will be able to extract the necessary information from the video and mitigate the workload for the researcher.

5 Big Data Security and Privacy hid-sp18-709, hid-sp18-710

Andres Castro, Uma M Kugan
andrescastro@iu.edu, umakugan@iu.edu
Indiana University
hid: sp18-709, sp18-710
github:

Keywords: big data, security, privacy

5.1 Introduction

Each organization has unique needs when it comes to Big Data. These needs cannot be described with one defined structure alone, and likewise, the information that they use does not come with defined data types. Because of this, there is the need for the Big Data Platform. Big Data is gaining more popularity because of its ability to connect to a number of devices in the so-called ***Internet of Things*** (IoT), producing a huge dump of data that needs to be transformed into information assets. It is also very popular to buy additional on-demand computing power and storage from public and private clouds to perform intensive data-parallel processing. These things not only create the way for Big Data expansion but also boosts security and privacy issues. Big Data security is the process of securing data and their processes both within and outside the organization. Big Data deployments are valuable targets for intruders and, because of this, security becomes a never ending concern for any organization. A single unauthorized user gaining access to an organization's big data could in and of itself acquire all the valuable information that the company possesses which could result not only in monetary loss but also be detrimental to its business and to its brand name. In current trends, security teams work towards continuously monitoring networks, hosts and application behavior across their organization's data. Traditional methods of securing firewalls are no longer enough to secure a company's data assets and Big Data platforms need to be secured with a mix of both traditional and newly developed security tools, as

well as big data analytics for monitoring security throughout the life of the platform [22].

5.2 What is Big Data

Big Data, by definition of its name, is an extensive variety and heavy volume of data that can be entered or transferred at high velocity, and include data sets coming from dynamic sources of data and applies technologies to analyze these data sets. It is a term usually used to define huge and complex data sets that do not fit into any traditional system. Most recently, the term ***Big Data*** tends to refer to the use of predictive, user behavior analytics, or certain other advanced data analytics that extract value from data sets. These analytics provide more insights about the data which indeed help businesses understand their trends which will eventually, in good theory, help their growth [23].

For example, a company that works with waste management, can collect data on the waste production and human activities from very diverse sources, then interpret the findings of Big Data to make optimal decisions [24].

5.3 Big Data Needs Big Security

The amount of data collected by organizations and individuals around the world is growing on a daily basis, and the volume of the data being collected is expected to continue to grow exponentially. It is believed that the 90% of the data we have currently have in the world has been collected in the past few years. Velocity, volume and variety of Big Data comes results in privacy, security and compliance issues as well. Some of the data stored in Big Data platforms is very sensitive and regulations need be put in place, strictly controlling specific aspects of the data and who has access to the data. Proper measures have to be taken to control any weaknesses to cyber threats.

There are requirements for security measures already in place. Big Data platforms are subject to compliance mandates by government and industry regulations, including GDPR, PCI, Sarbanes-Oxley (SOX), and HIPAA [25]. These measures place regulations on company practices and implementations that ensure proper data security and monitoring. These regulations are mandatory, and failing to comply could result in severe penalties, from heavy

fines to legal actions.

While these requirements are important, traditional security mechanisms that have been in place for securing structured static data are no longer sufficient. With technological advances also comes a need to continually assess weaknesses in the new systems, to protect itself from new cyber threats and hacking strategies, and to create user friendly platforms for client that do not compromise the data being collected or stored. These developments are often far ahead of regulation, and individual entities need to be continually monitoring and enhancing their platforms to ensure protection of its data and systems. Big Data needs bigger security to protect its data, applications and infrastructure. Securing data not only protects the brand, reduces costs and avoids any legal issues, it also helps in retaining the brand name and increases revenue and growth [26].

5.4 Big Data Security Challenges

Recent adoption of cloud storage has increased the amount of data collected by organizations and hence it has become of vital importance to secure these data platforms as well. Data security issues are generally caused by the lack of proper tools and measures provided by traditional anti-virus software. Routine security checks to detect patches are no longer enough to handle real time influxes of data. Streaming real time data demands a great amount of attention focused on security and privacy solutions. Databases are no longer static. Big Data security's motto is to restrict unauthorized users and intruders from getting into a platform and also to block the encryption of data both in-transit and at-rest. The adoption of cloud storage creates a need to pay particular attention to the in-transit, or the continually expanding and modifying databases. Big Data security tools must be in place at all stages of data i.e. on incoming data, data stored in the platform and also on the data that goes out to other applications or outside party [27].

5.4.1 Access Control

Access control, in the context of Big Data, is controlling who can access data by using security settings. The different platforms that use Big Data need to be able to identify critical data, data origination and also who has access to the data. In this capacity, data access is not only protecting from external access, but also

protecting data from those who have internal access as well [28].

User access should be controlled via a policy-based approach that automates access based on user and role-based settings. This manages different level of approvals in order to regulate who has access to the critical data and to protect the big data platform against inside attacks [29].

5.4.2 Audit Control

Big Data analytics can be used to analyze different types of logs in order to identify malicious activity. It also can regularly audit all the working directories inside the organization in order to check for any unauthorized access to any sensitive or privacy data. In reality, not all attacks are identified in the exact moment when the attack occurs. In order to perform a root cause analysis of the incident, data security professionals need to have access to audit logs which allow them to trace attacks back to the point of entry, exact time, modifications or weaknesses. In case of data breach, some firms are required to turn over their audit logs to stakeholders and possibly affected companies and heavy fines are imposed for failure to comply [26].

5.4.3 Real Time Compliance Control

Real time security monitoring is always very challenging due to the number of false positive alerts generated by security programs. Because of the frequency of false positive alerts, they are usually ignored. Big Data analytics may help provide more meaningful insights that could result in real time detection .

5.4.4 Non Relational Databases Privacy

Non Relational Databases are still not fully matured. This poses a severe threat to securing the data and it is often difficult for security and governance team to keep up with the demand. NoSQL databases primarily focus on how to handle high volume of data without paying much attention to their security needs.

5.4.5 End-Point Input Validation

Many organizations collect their data from End-Point devices. It is very

important to ensure that data coming from these devices is not infected. Proper steps must be taken to make sure data is coming from an authentic source and it is legitimate. Incoming data from End-Point devices such as smart phones is growing tremendously and filtering or validating data from these sources is a very big challenge [26].

5.4.6 Securing Transaction Logs and Data

Data in any organization many be stored at various levels (tiers) of the storage structure depending on the need and usage of the data. Increase in the transfer of data within the organization enforce for the need of auto-tiering for Big Data storage whereas auto-tiering does not maintain the log of where the data is stored and hence security is a big concern.

5.4.7 Securing Distributed Framework

Distributed framework enforces parallelism. This means that data is distributed across multiple nodes to achieve faster processing of large volumes of data. This increases the security concern of the framework and the data that exists there. Most companies use a distributed framework like MapReduce in which mappers read and compute and reducers combine the output from each mapper. If mappers are not secured, there is the chance of data being compromised [29].

5.4.8 Data Provenance

It is very important to know the original data that is coming to the platform so that we can better classify them. Data Origin should be consistently monitored but in reality due to the high volume it is becoming a big concern for data security. Provenance metadata is growing significantly as well and protecting metadata is very crucial for any organization [29].

5.5 Big Data Security Stakeholders

In the digital era, the traditional way of securing the data, changing passwords frequently, firewall protection is just not enough to keep up with the growth of data produced by Internet of Things(IoT), Smart Devices, Bring Your Own Devices (BYOD) and several customer friendly apps that is coming out

everyday. “Even though end user has the biggest responsibility with securing his own data, unfortunately, end users are not fully aware of the cyber security issues and they do not have the appropriate knowledge to discover the world wide web in complete safety” [30].

Big Data deployment is not possible to handle by any single business unit or with single tech team. It involves several business units, infrastructure, information technology, security, compliance, programmers, testers and product owners are all involved in big data deployment. They are all responsible for Big Data Security. Information Technology and Security team is responsible for drawing the policies and procedures. Compliance officers together with security team will protect compliance, such as automatically encrypting personally identifiable information before it is easily accessible. Administrators will automate these process to protect their environment. Even though every organizations have their policies and control laid in place to protect their biggest asset, phishing attacks can come in any form as a simple email. Frequent internal audit within the company can help us periodically check if all privacy, security and compliance are all in place.If not, proper measures can be taken right away to avoid any legal issues.

“The average annualized cost of cyber crime based upon a representative sample of 237 organizations in six countries by Ponemon Institute in their 2016 Cost of Cyber Crime Study and the Risk of Business Innovation sponsored by Hewlett Packard Enterprise is 9.5million U.S. dollars” [31]. In any organization, loss of information is the most expensive consequence of a cyber crime. The cyber attack may results in business disruptions, data or information loss, loss of revenue, damage to equipment and last but not the least it damages the brand. So it is big time to protect and secure the big data and the environment from all angles.

5.6 Best Practices for securing Big Data

There are three fundamental principles used in defining security goals: confidentiality, integrity, and availability. Confidentiality is the ability to keep sensitive data safe from third parties and unauthorized access. Integrity in this context means to avoid unauthorized modification of the data. Finally, availability means always being able to access the data and resources. These

three concepts are known as the CIA triad, and is used as base principles when discussing and designing security practices [32].

To meet these goals, there are four main branches of security that apply to Big Data: Authentication, Encryption, Data Masking and Access Control [33].

5.6.1 Authentication

Because of its nature (large sizes of data, linking different sources, sharing access with third parties, etc), some of Big Data's features are highly susceptible to different privacy, security and welfare risks [34].

Privacy can be defined as the condition of confidentiality, protecting information from third parties. To support privacy, there have been different Authentication methods that both verify and validate entities who attempt to access the information. This ensures that only authorized entities are able to access the data or resources.

With Big Data, it is important to choose a proper authentication method, with the least computation complexity as possible, to allow dynamic security solutions within large Data Centers and also to avoid incrementing the traffic unnecessarily. Choosing an overbearing authentication method can cause both delay and storage issues. Because of this, it is important to tailor the security to the needs of the specific network ???.

5.6.2 Cryptography

There are multiple understandings of how data moves through stages, also known as Data Life cycles. Cryptography- define in terms of security. CITA.

From the perspective of cryptography, there are three phases in the Data Life Cycle: Data in Transit, Data in Storage, and Data in Use. Different cryptography techniques will be implemented depending on which stage of the life cycle the data is in [32].

There are different cryptographic tools that not only keep data secure at each point in its life-cycle, but also enable richer use of the data. The main tool is Encryption. Encryption takes pieces of data in plain text and use a cryptographic

key to produce a version of the data that can only be read using the cryptographic key. Without the key, the information is illegible. There are two types of encryption: secret key encryption and public key encryption. Secret key encryption is when the same key is used for both encrypting and decrypting data. There are scenarios when one of the keys can be made public. For instance, if the locking key is kept private but the unlocking one is made public, this security can be used to prove authenticity [32].

There are different standards for encryption. The most well known and commonly used is Advanced Encryption Standard (AES). This standard sets guiding principles to ensure that data is encrypted in a manner that meets security needs and allows the recovery of original data [32].

5.6.3 Data Masking

By definition, Big Data works with large volumes of heterogeneous data sets using software to manage the data and to provide predictive analysis. Data masking works by replacing sensitive data with non-sensitive values, yet preserves the data integrity. For instance, replacing names with code names, or social security numbers with a key number. By doing this, different parties can access information without putting sensitive data at risk [35].

Five laws for data masking have been developed by Securosis Research. The first law is that data masking should not be reversible. This means that the data should not be unmasked easily using reverse engineering. The second law is that data that has been masked has to represent the original data set. For example, it has to belong in the same context. The third law states that data masking should maintain application and database integrity. This means that the process of data masking should not modify or affect the data in the databases in a negative way. The fourth law emphasizes that non-sensitive data can be masked, but it should not be masked if it can not be used to make sensitive data vulnerable. For instance, when masking information about a person, it is correct to mask the person’s name, email address and social security number, but other information like gender, or favorite colour, would be useless to mask. Finally, data masking must be a repeatable process, using a standard to reproduce the steps taken to mask the data, allowing to troubleshoot possible problems in the process [35].

5.6.4 Access Control

As it was explained in the Challenges section, Access control, allows some entities to access the data or resources, while denying its use to other entities. Through security settings.

Some authors add that the inferences drawn from data should also be a cause for concern, because they can identify traits and patterns that could expose vulnerabilities. They propose that organizations who use the protected data should disclose their decisions criteria in order to apply access control in a broader spectrum. By doing so, it would be sufficient to diminish privacy concerns by de-identifying the data, or denying access to certain parts of the data that coul be used to make entities or data vulnerable. Some of these authors say that by doing this, it would not only reduce the privacy risk, it would also salvage large amounts of data for alternative use. This de-identification can also be achieved through data masking, pseudonymization, aggregation, among other methods [36].

5.6.5 Physical Security

It is always better to build and deploy Big Data platforms in their own data center. If deployed in a cloud, the organization must diligent to ensure that the cloud provider’s data center is physically well secured. Access should be restricted to strangers and staff who have no official responsibilities in the designated areas or interacting with the data sources. Data centers should be properly monitored at all times and video surveillance and security logs are important tools to achieve this.

5.7 Future of Big Data Security

To think about Future of Big Data Security, it is necessary to engage the conversation of what the trends are in Big Data and what technologies are expanding and changing the horizon. There are many new technologies and solutions that are shaping the future of the Big Data, but because of the length and focus of this document, there are three main areas that will be covered: Virtualization and Cloud Computing, IOT Security, External Password Vaults and Penetration Tests.

5.7.1 Virtualization and Cloud Computing

Virtualization is a way of deploying resources at multiple levels, such as hardware, network infrastructure, application and desktop centralized managing and using dynamically the physical resources. This makes the system flexible and less costly than traditional environments and giving management new tools to optimize the use of resources [37].

Since virtualization can be developed in so many levels, including cloud computing and by multiple service providers, it is natural that the system requirements of users and organizations move towards a variety of solutions that may include Infrastructure as a Service (IaaS) frameworks from public clouds such as the ones offered by Amazon, Microsoft, Google, Rackspace, HP, among others, or even Private clouds, maintained and many times even set up by internal IT departments [38].

These cloud computing technologies are being used to solve data-intensive problems on large-scale infrastructure. Thus, integrating big data technologies and cloud computing for data mining, knowledge discovery, and decision-making [39].

5.7.2 IOT Security

The Internet of Things (IoT) is the name given to the large network of physical devices that does not match the typical concept of computer networks, this includes all kinds of objects. The large and growing amount of devices and diverse uses given to them, makes IoT generates very important Big Data streams. Making it necessary to develop new systems and data mining techniques for this new paradigm [40].

In this IoT paradigm, each new opportunity opens doors to new threats as well. This makes it necessary to develop techniques to ensure trust, security and privacy. Different Authors write about the possible ways to face these challenges, and some, they consider three main axes to articulate the solutions: Effective security - used in very small embedded networks, context-aware privacy and user-centric privacy, and the third one is the systemic and cognitive approach for IoT security - where the interaction between people and the IoT can

be envisioned as a set of nodes and tensions [41].

All this to say that in order to approach privacy and security in this new paradigm, many new theories and techniques have been developed since old security products and techniques may not suffice the needs of the different IoT users and communications.

5.7.3 External Password Vaults

Password vaults are applications that store multiple passwords and encrypt them storing them in a database [42].

There are small Password Vaults that can be stored locally on a system, or larger options that can be integrated into larger systems, providing additional security options, like generating real time temporary passwords for effective password rotation (I.E. Cyberark External Password Vault) [43].

These techniques are key to articulate authentication and a proper data access while using multiple services such as Cloud infrastructure and IoT.

5.7.4 Penetration Tests

After applying all the security techniques and strategies, and after putting in place all necessary security and privacy policies, the most important step is validating the strength of the security of the system. For some time, companies have started to perform tests that consist on simulating an attack from the perspective of an attacker, this method is known as Penetration test and it allows to actively evaluate and assess the security of a system [44].

The tester identifies the threats faced by an organization from hackers and suggest changes to improve the security and minimize the vulnerabilities and close the possible loop holes in the network [44].

5.8 CONCLUSIONS

Big Data as a constantly evolving and ever changing branch of information technologies resembles an ecosystem that since it covers gathering data from so

many sources, processing it and generating new information, there will be many entities and interests involved that will need to be protected. The features of Big Data such as Volume, Variety and Velocity bring new challenges to security and privacy protection. To protect the integrity and availability, security providers and local IT departments, will have to diversify their security and privacy strategies and policies, in order to keep pace with the growth and evolution of this new ecosystem.

The authors would like to thank Dr. Gregor von Laszewski for his support and suggestions in writing this paper.

5.9 Work Breakdown

- Uma Kugan Research for Section Big Data Needs Big Security, Big Data Security and Challenges.
- Andres Castro Benavides Research for Section Best practices and Future
- Editing:: Andres Castro Benavides and Uma Kugan, Gregor von Laszewski

REFERNCES

[1] S. C. Matz and O. Netzer, “Using big data as a window into consumers’ psychology,” ***Current Opinion in Behavioral Sciences***, vol. 18, pp. 7–12, 2017.

[2] S. Gavaris, “Use of a multiplicative model to estimate catch rate and effort from commercial data,” ***Canadian Journal of Fisheries and Aquatic Sciences***, vol. 37, no. 12, pp. 2272–2275, 1980.

[3] T. W. Sidle, “Weaknesses of commercial data base management systems in engineering applications,” in ***Design automation, 1980. 17th conference on***, 1980, pp. 57–61.

[4] C. J. Hoofnagle, “Big brother’s little helpers: How choicepoint and other commercial data brokers collect and package your data for law enforcement,” ***NCJ Int’l L. & Com. Reg.***, vol. 29, p. 595, 2003.

[5] M. Abadi ***et al.***, “Tensorflow: A system for large-scale machine learning.” in ***OSDI***, 2016, vol. 16, pp. 265–283.

[6] Datameer, “Six ways to create better customer behavior analytics.” online, Feb-2018 [Online]. Available: https://www.datameer.com/blog/six-ways-create-better-customer-behavior-analytics/

[7] L. M. Powell ***et al.***, “Field validation of secondary commercial data sources on the retail food outlet environment in the us,” ***Health & place***, vol. 17, no. 5, pp. 1122–1131, 2011.

[8] J. X. Dempsey and L. M. Flint, “Commercial data and national security,” ***Geo. Wash. L. Rev.***, vol. 72, p. 1459, 2003.

[9] P. Covington, J. Adams, and E. Sargin, “Deep neural networks for youtube recommendations,” in ***Proceedings of the 10th acm conference on recommender systems***, 2016, pp. 191–198.

[10] H.-T. Cheng ***et al.***, “Wide & deep learning for recommender systems,” in ***Proceedings of the 1st workshop on deep learning for recommender systems***,

2016, pp. 7–10.

[11] xingangzhongzhinao, “A recommendation system based on tensorflow.” online, Jan-2018 [Online]. Available: https://blog.csdn.net/shebao3333/article/details/78966926

[12] C. M. Hoehner and M. Schootman, “Concordance of commercial data sources for neighborhood-effects studies,” ***Journal of Urban Health***, vol. 87, no. 4, pp. 713–725, 2010.

[13] M. Zanker, M. Jessenitschnig, D. Jannach, and S. Gordea, “Comparing recommendation strategies in a commercial context,” ***IEEE Intelligent Systems***, vol. 22, no. 3, 2007.

[14] R. J. Birk, T. Stanley, G. I. Snyder, T. A. Hennig, M. M. Fladeland, and F. Policelli, “Government programs for research and operational uses of commercial remote sensing data,” ***Remote Sensing of Environment***, vol. 88, nos. 1-2, pp. 3–16, 2003.

[15] wikipedia.org, “ERP.” Web page, Nov-2018 [Online]. Available: https://en.wikipedia.org/wiki/Enterprise_resource_planning

[16] SAP, Inc., “SAP: Software solutions | business applications and technology.” Web page, Nov-2018 [Online]. Available: https://www.sap.com/index.html/

[17] Saudi ERP & Website Solution Blog, “Complete list of sap erp modules.” Web page, Jun-2015 [Online]. Available: https://solutiondots.com/blog/sap-erp-modules/

[18] Eshna Verma, “Overview of sap modules.” Web page, Nov-2018 [Online]. Available: https://www.simplilearn.com/sap-modules-sap-fi-sap-co-sap-sd-sap-hcm-and-more-rar111-article

[19] SAP, Inc., “Business intelligence software.” Web page, Nov-2018 [Online]. Available: https://www.sap.com/products/bi-platform.html

[20] SAP, Inc., “CRM and customer experience.” Web page, Nov-2018 [Online]. Available: https://www.sap.com/products/crm-commerce.html

[21] SAP, Inc., “Core hr and payroll.” Web page, Nov-2018 [Online]. Available: https://www.sap.com/products/human-resources-hcm/core-hr-payroll.html

[22] J. Moura and C. Serrao, “Security and privacy issues of big data,” ***CoRR***, vol. abs/1601.06206, pp. 1–2, 2016 [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1601/1601.06206

[23] A. Hey, S. Tansley, and K. Tolle, ***The fourth paradigm: Data-intensive scientific discovery***. REDMOND, WASHINGTON: Microsoft Research, 2009 [Online]. Available: https://books.google.com.my/books?id=oGs_AQAAIAAJ

[24] V. Yenkar and M. Bartere, “Review on ‘data mining with big data’,” ***International Journal of Computer Science and Mobile Computing***, vol. 3, no. 4, pp. 97–102, 2014.

[25] C. O’Neill, “Big data needs big security. Here’s why.” Imperva, Redwood Shores,California, 2017 [Online]. Available: https://www.imperva.com/blog/2017/02/big-data-needs-big-security-heres/

[26] S. Rajan, W. van Ginkel, and N. Sundaresan, “Top ten big data security and privacy challenges,” Cloud Security Alliance, 2012 [Online]. Available: https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Top_Ten

[27] C. Taylor, “Big data security,” Datamation, Foster City, California, 2017 [Online]. Available: https://www.datamation.com/big-data/big-data-security.html

[28] A. RAHMANI, A. AMINE, and M. R. HAMOU, “A mathematical model of access control in big data using confidence interval and digital signature,” ***Computer Science & Information Technology***, vol. 5, pp. 183–198, 2015.

[29] P. Buttler, “Big data needs big security. Here’s why.” Dataconomy, Berlin, Germany, 2017 [Online]. Available: http://dataconomy.com/2017/07/10-challenges-big-data-security-privacy/

[30] Realdolmen, “Cyber security.” Belgium, 2017 [Online]. Available: http://www.realdolmen.com/en/blog/who-responsible-for-data-security-your-company

[31] P. R. Department, “Ponemon institute© research report.” Traverse City, Michigan, 2016 [Online]. Available: https://www.ponemon.org/local/upload/file/2016%20HPE%20CCC%20GLOBA

[32] A. Hamlin, N. Schear, E. Shen, M. Varia, S. Yakoubov, and A. Yerukhimovich, ***Cryptography for big data security***. Boca Raton, Florida: Taylor & Francis CRC Press, 2016, pp. 241–288.

[33] K. Abouelmehdi, A. Beni-Hssane, H. Khaloufi, and M. Saadi, “"Big data security and privacy in healthcare: A review",” ***"Procedia Computer Science"***, vol. 113, no. Supplement C, pp. 73–80, 2017 [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050917317015

[34] N. Kshetri, “Big data′ s impact on privacy, security and consumer welfare,” ***Telecommunications Policy***, vol. 38, no. 11, pp. 1134–1145, 2014.

[35] R. Archana, R. S. Hegadi, and T. Manjunath, “A big data security using data masking methods,” ***Indonesian Journal of Electrical Engineering and Computer Science***, vol. 7, no. 2, pp. 449–456, 2017.

[36] O. Tene and J. Polonetsky, “Big data for all: Privacy and user control in the age of analytics,” ***Nw. J. Tech. & Intell. Prop.***, vol. 11, p. xxvii, 2012.

[37] K. Padmini, “Securing data management based on key technologies in cloud computing,” ***International Journal of Advance Research in Computer Science and Management Studies***, vol. 3, no. 2, pp. 165–172, 2015.

[38] G. von Laszewski, F. Wang, H. Lee, H. Chen, and G. C. Fox, “Accessing multiple clouds with cloudmesh,” in ***Proceedings of the 2014 acm international workshop on software-defined ecosystems***, 2014, pp. 21–28 [Online]. Available: http://doi.acm.org/10.1145/2609441.2609638

[39] R. Kune, P. K. Konugurthi, A. Agarwal, R. R. Chillarige, and R. Buyya, “The anatomy of big data computing,” ***Software: Practice and Experience***, vol. 46, no. 1, pp. 79–105, 2016.

[40] A. Bifet, “Mining internet of things (iot) big data streams.” in ***SIMBig***, 2016, pp. 15–16.

[41] A. Riahi, E. Natalizio, Y. Challal, N. Mitton, and A. Iera, “A systemic and cognitive approach for iot security,” in ***Computing, networking and communications (icnc), 2014 international conference on***, 2014, pp. 183–188.

[42] R. Chatterjee, J. Bonneau, A. Juels, and T. Ristenpart, “Cracking-resistant password vaults using natural language encoders,” in ***Security and privacy (sp), 2015 ieee symposium on***, 2015, pp. 481–498.

[43] G. S. Nelson, “Practical implications of sharing data: A primer on data privacy, anonymization, and de-identification,” in ***SAS global forum proceedings***, 2015, p. XXIII.

[44] C. N. Shivayogimath, “AN overview of network penetration testing,” ***International Journal of Research Engineering and Technology***, vol. 3, no. 7, pp. 408–413, 2014.