BIG DATA

Supervised by:

Dr/Nisreen Al-Abyad

By:

1-Yousef khaled ALWaer

2-Mahmoud mohamed Seif

3-Yousef Saeed Ebrahim

4-Mohamed Ebrahim moussa

5-Abd Elrahman Mohamed Salem

2020

Table of Contents

Executive Summary

As soon as you hear the word Big Data for the first time, large-sized data comes into your mind. This is true at first glance, but since 2003 the situation has changed a lot. The concept of big data began to introduce more dilemmas which can be summarized in the word 4V', illustrating the development that has occurred in the data in the last years. The letters stand for: Volume, velocity, variety and veracity.

But as usual, nothing is perfect. Despite the many advantages and uses of big data, it has a number of disadvantages including the risk of information leakage. One example of such data is social data.s

Social data appeared with the appearance of social platforms .Since that time collecting members' data and analyzing it become a must for social media platforms companies. Social data analytics allow the platforms to have more members every day and increase profits. But this process has many defects. It violates our privacy and may use collected data to track us. All countries should apply laws and restrictions to avoid harms of social data analytics on their people.

Big Data has many applications in real life, as Big Data helps the police in combating crime. By analyzing information, the police can predict the occurrence of some crimes in the future, making them able to prevent some crimes from happening. Big data is used in the field of online shopping, as a company like Amazon uses Big Data to suggest to the user what products suit him by analyzing his information on social media. Music streaming is one of the areas in which big data is used. A company like Spotify uses Big Data to know what kind of music the user prefers, and accordingly suggests songs to him. Big Data is also used in TV streaming, as a company like Netflix uses it to know the opinions of viewers and make decisions based on it.

To talk about something concrete and massively important, it would be about the use of big data in the health sector. No longer are we in the era of healthcare that is solely dependent on human error, but rather, we are moving towards a future in which precision in diagnosis and early

intervention and detection of diseases exist. From clinical trials to EHRS to the various studies done on society, contribution to the gathering of data relevant to our cause here keeps magnifying, helping in both: improvement of the healthcare system and reduction of the unnecessary expense. But things are not always that ready for us to devour and take advantage of them, many difficulties appear to be standing in our way of advancement.

To be honest, there are so many difficulties that face IT professionals. Some of them matter such as dirty data, company policies or financial support, and lack of data science talent. These are the main difficulties obtained from a survey in 2017 by asking 16000 IT professionals, and there are other minor difficulties that used to be a big deal such as processing, scalability, poor data quality, etc. But for us humans, big data is worth money and effort for its great benefits in a variety of fields.

Introduction

When talking specifically about Big Data, it gets confusing because until this time no one can fully define it. But for the sake of simplicity, Big Data can be interpreted as a huge amount of data that has a few additions. 1In most big data circles, these are called the 4V's: volume, velocity, variety, and veracity. Relating to each one of these characteristics, problems appeared that needed new technical solutions.

For Volume: the problem of size. In 2003, Google and Yahoo produced an integrated eco-system known as Hadoop, which divides the large files required to be stored into a larger number of small chunks that are distributed among several supercomputers to carry out these storage operations. Thus was the problem of volume and lack of tools solved.

And due to the emergence of new types of data, dealing with different types of data, each requiring a different high speed, got to be the problem associated with the second characteristics: velocity. A technical solution invented to solve this was through the Apache Software Foundation establishing a framework known as Spark. The special thing about Spark is that it can deal with different speeds of processing so that there are no differences in the basic operations of data processing.

Another problem due to the different types of data was that new ways to store all the different data types was lacking after the impossibility of storing all kinds of data. We needed a framework like MongoDB, which is characterized by its ability to store and arrange data regardless of its type and in the absence of the need to place it in tables.

Of course, all these technologies are not without purpose. However, the goal of gathering and analysing such gathered information is to convert it into knowledge, through which it is possible to know the appropriate ways to exploit that data. For example, knowing that half the sales of a store is done through its website, helps the store improving its sales returns. One way of doing this is if the store made more offers available on that particular website, thus appealing to a large number of customers and potentially increasing its sales.

Such data may also be required through any of the various social platforms. The following are the most important methods of gathering data through these platforms.

Social data

With the great evolution of computer devices and networks in late 1990s , people utilizes that evolution to satisfy their need to communicate with each other by establishing personal connections , at that period the terminology “SOCIAL MEDIA” appeared .In fact , social platforms appeared long time ago, in 1844 ,Samuel Morris managed to send the first telegraph message from Baltimore to DC Washington and the data transferred was dots and dashes.

The beginning of 21th century was marked by the appearance of social media platforms we know today. *Facebook* as an example was launched in 2004 by Harvard student Mark Zuckerberg to reach after a few months to 1M MAUs(Monthly active users) to reach in 2020 to 2.23B MAUs, to be the market leader platform.

LinkedIn was launched in 2002 with only 81K MAUs to reach in 2020 380M MAUs. It was and still the best platform for many job seekers to get their best job opportunity and also for human resources mangers to have the best employees.

Statistics shows that there are about 3.5 B users who use some kind of social media which equates 45% of earth’s population.

The data that being collected from members like name, age , gender, marital status ,address ,……..vice versa to join social platform is called social data. This data is being classified and filtered using machine learning algorithm .This process is called Social data analytics. This process is double-edged weapon.

On one hand, it allows marketers for services and products in social platform to target their customer based on detailed specifications that go beyond the data segregated. This process is called “MICROTARGETING”.

In this way, marketers make more profits than advertising without Microtargeting .This is also better for customer as it absolves the need for boredom and frustration of having to go through advertisements that have little to do with him/her. In this way social platform win more money and attract more users .

Social platform also enables a brand owner whether it's a company or a person to develop their brand. It uses the data analyzed to get information about people who like or disliked their content. In this way it tells the brand owner whether or not their brand message achieves its goals or not. If not, it helps them to reach their target based on the data analyzed.

But social data analytics has some demerits. The process of collecting data about us and classifying it is against privacy law. This data also might be in somewhere that is vulnerable to be hacked. So , the hacker can use this data against us which is called data breaches.

Microtargeting may also be used for discrimination and social harm. It's possible to target category of people with offensive advertisements and abusive content .

Some platforms can provide fake news, bots and filter bubbles. They target public figures and institutions and even countries .They can lead to social and political harm as the information that informs people is manipulated, potentially leading to misinformation and undermining democratic and political processes as well as social well-being.

But, we can avoid most of social data analytics demerits. Companies operating these platforms have to follow anti violation policy with users and to be transparent about how they use the user data.

They also have to evaluate user behavior and advertisements to prevent damage, social harm, discriminative and abusive contents.

As we live today in a world that has no borders and becomes only controlled by cyber geography, all countries should do their best keep their cyber security and privacy of their people. They can restrict social platforms with legislations that can reach to banning.

Applications of Big Data in Real Life

Crime fighting

By collecting and analyzing data, the police can classify the areas in terms of severity, and this percentage is due to the number of crimes that occur in this area in a certain period of time.

The police appoint employees who collect and analyze a large number of data, and by analyzing that data, they can predict the occurrence of some crimes in the future in a certain place and time.

The police take that data about the location and time of the expected crime, and increase the number of police officers in this place, and thus sometimes they can prevent some crimes from happening.

Thus, Big Data has a role in decreasing the number of crimes at the country level, and increasing the percentage of safety among citizens.

Online shopping

Everyone knows that Amazon is one of the largest online shopping companies, and part of the reason for its success is due to its use of Big Data, where that the company collects data by taking a look at the user's search history and browsing history, and compares that data with the data of customers who have already bought products from them.

If this happens and there is similarity between the data, the company will suggest to the user the products that the customer who with similar information has bought, because there is a high probability that the user will buy a product from these products, because his thinking is similar to that of the customer who bought the product before him.

If the user has previously bought a product from the company, then the company sees what products are related to this product, and suggests these products to the user, because there is a high probability that the user will buy another product from these products, which leads to an increase in the company's sales and Increase the company's profits.

Music streaming

If we take Spotify as an example, Spotify has a large collection of songs on its platform, it divides these songs into more than one category according to the type of song and singer, so when a new user comes to register on the platform, the site displays many types of songs and singers' pages, And asked him to choose the type of songs he preferred, and according to the user's choice, the site suggests to him an album consisting of the user's favorite song type.

And the site comes every week and sees what is the most type the user has listened to during this week, and proposes to him a group of songs of the same type, because the opportunity for the user to listen to these songs is great, which brings profit to the company, and also the site sees any The songs were the most listened to this week in the world, and he collects these songs in one album, and suggests it to all users.

TV streaming

Netflix is one of the largest companies that offer online entertainment content worldwide, and part of its success is due to its use of Big Data, as the company, by collecting information and analyzing people's opinions on social media platforms, knew that the user does not like to wait a week between Episodes of the series, and also know who are the favorite actors of most people and the types of stories that most people prefer.

Through this information, the company is making a series with a small number of episodes, an exciting story, and a good cast, and it determines the date of the series' launch and launches the series with its full episodes on this date, in order to give the user complete freedom to choose when to watch the series, which makes the user Happy, and this led to an increase in the number of annual subscriptions on the Netflix platform, and an increase in the company's profits.

Big Data and Health

To talk about Big Data in the health sector, and its uses and ways of generation, let's first state the main objective we are after.

The general objective of using big data in health is improving healthcare. Reducing expenses can also be considered part of that same objective.

5.1 Ways of generation

Ways of generation of relevant data greatly differ, but fall within the scope of the following:

1. Electronic Health Records (*EHRS*)

   The main source to consider is, obviously, patient records. But instead of having these records as paper records, they are digitized. What is now known as EHRS is simply the digitized, systemized patient records and histories. With most healthcare providers moving toward digitization, we get to have a large valuable database. This database may be at the scale of a city, a nation, or even a union of nations. Only the size of such database is a hint of what valuable insights it will give us with the right data analysis techniques.

2. Clinical trials

   Clinical trials are a great source of health-related information. They are typically performed for every new vaccine or treatment. What they aim at is studying the effects of the new treatment. These effects may differ according to the health state of each individual being tested. Prior diseases, like genetic diseases, may be part of the calculation of side effects. These trials when performed on a large scale, produce massive and valuable data. They may consist of multiple experimental levels. These levels are a must for each treatment before releasing it to the masses. The list written with each medicine, of side effects and reasons for use, is provided as a result of these trials.

The last example for this that we have witnessed, is the release of the last vaccine for COVID-19.

3. Telemedicine

Telemedicine is simply defined as the remote delivery of healthcare services. It is typically used when the direct clinician-to-patient contact is not available for any reason whatsoever. However, it is projected that it will be gradually adopted at a large scale as technology advances. Different types of telemedicine exists, but the most common are:

- Interactive Medicine: Real-time communication between physicians and patients.
- Remote Patient Monitoring: Monitoring patients residing at home using mobile health devices.
- Store and Forward: Sharing patient information, what we might call it now EHRS, between physicians in different locations.

4. Genomic data

What a *genome* signifies is the whole genetic side of any human being. Collection of genome sequences on a large scale results in a large genetic database. Integrating this kind of data with other kinds of available data, previously mentioned, a greater perception of the genetic nature of diseases is gained.

All of Us is an example of a research done to collect such data. This research program is done by the NIH in the US. It is an effort toward collecting the genomic data of one million people or more of the US citizens. Its main goal is to accelerate research and improve healthcare.

5. Behavior and Socio-economic status (SES)

As with everything, the social and economic side of things determines a lot.

And so it does here, mainly determining three things:

- Health care

  As social and economic classes differs, so does the level of access to health care, and the quality of such care.
- Environmental exposure

  Each class is surrounded and affected by certain environmental factors. These factors may help us greatly predict which class is more prone to a certain disease. Environmental factors constitute what is known as risk factors of diseases which we will talk about later.
- Health behavior

  Behaviors generally differ among classes. These behaviors may be a valuable source in our analysis.

5.2 Uses

So the insistent question that keeps manifesting now is: what will be done with all of this data? A lot of things, all of which satisfy our general objective here: improving healthcare and reducing expense. The following are the most prominent ones:

- Early Diagnosis

  Having studied diseases well and their different manifestation on patients, we now know a great deal about the signs and symptoms of each disease. Detecting such symptoms early helps in catching diseases early. Some diseases are degenerative by nature, like Parkinson disease and Hepatitis C, and so early diagnosis helps catching these before reaching a degenerative state.

- Disease prevention

  Disease prevention is not early diagnosis. It is, as its name tells, preventing patients from catching diseases in the first place. To know how that may be accomplished we first need to familiarize ourselves with what risk factors are.

Each disease has what is called risk factors. These may either be genetic factors, or environmental ones. The more one is exposed to such factors, the more one is vulnerable to catching their diseases. With enough patient records and data on the nature of diseases, vulnerable individuals are detected. They are helped and advised to reduce their exposure on such factors, if they are environmental.

- Managed Care

With healthcare providers having EHRS for each patient, it gets easier to give each patient the specialized care they need. That leads to two main things:

1. Reduction of unnecessary expenses, both for the patient and for the healthcare provider, previously spent due to wrong diagnoses.

2. Better and more accurate treatment for each patient.

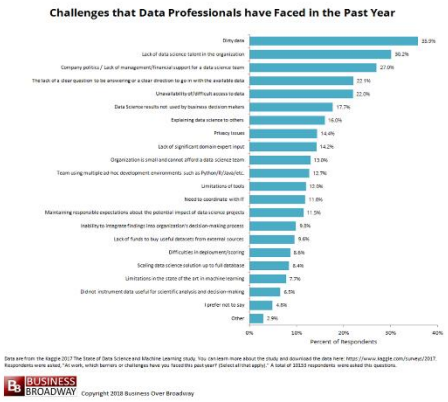
- Prediction of Outcomes

Prediction of the outcomes of the various treatments developed and of the development of symptoms gets to be easier.

Difficulties facing IT professionals

After we talked about the pros of Big Data, let's talk about the downsides. Like anything in life, big data is a double-edged sword. When we talk about big data, we should put a question mark about users' personal data. But here we will talk especially about difficulties facing IT (data) professionals. We will divide the problems into two categories: major difficulties and minor difficulties. To define what major difficulties are, they are simply difficulties that are still present and that still matter. On the other hand, minor difficulties are difficulties that used to cause a lot of trouble but it's no longer that important.

6.1 Major difficulties



For major difficulties, there are many, but we will only talk about the top three. These results are arranged according to a research done in 2017. This research was done by Kaggle (a company acquired by google that is specialized in machine learning and data science). They asked 16000 IT professional to arrange the difficulties they face. The results came as following: 36% voted for dirty data, 30 % for lack of data science talent and 27% for lack of management support. Those are the three difficulties we will talk about.

Definitions

1. Dirty data: It's called also rogue data, as it's in accurate or inconsistent or inaccurate. It can contain mistakes like spelling or punctuation error , incorrect data associated with a field or outdated , or a data that is duplicated as result of receiving the same data for the same person from different resources , but actually they can be cleaned using data cleaning . also for collecting huge amount of data , if it contains impurities it may result in bad decisions and bad data environment . unclear definitions is one of its reasons .

2. Lack of data science talent: the amount of data on the internet increased by 50 times between 2010 and 2020 . while the number of IT professionals increased by 1.5 times for the sane period and this created a big gap between the amount of data available and number of people who can handle it . so untrained data scientist can lead to insufficient knowledge of infrastructure needed or benefits of data they are collecting . so this can lead to waste of time , effort and money on collecting data that will never be used . so IT department must organize numerous trainings for data scientist . also for untrained data scientists , they can get lost in variety of big data technologies in the market .

3. Company policies/Financial support: let's talk first about company politics , as we know our data is so valuable for many companies , but at the same time , some companies like Facebook force restriction for anyone to reach our data , as these data are personal and it can be used illegally for advertisement reasons , so a big social media website like Facebook which have 2.7 billion user , is like a treasure for advertisement companies .

but to show the most relevant product to the most possible customer , they need our data to be used for this purpose , so Facebook put some restriction on our data so that it's not available for anyone to use . we I say our data , I mean like the search history , interests and so on . but sometimes data get leaked for security reasons as in 2019 when the phone numbers of 419 million users got leaked . this is because the interest in protecting data isn't as much as creating data and collecting them . as Only about half the data on the internet need to be secured actually isn't secured .

For financial support : as we know , big data costs a lot of money , as the amount of data on internet duplicates with time . so we need more modern hardware components , new hires , so much electricity as the process of collecting data and analyzing it , is done on computers and servers that consumes so much power . also we need to develop so many complicated algorithms ,

so the process of collecting data becomes more efficient . so the development department consumes so much money . to understand the importance of developing algorithms , it can reduce power consumption by 5 to 100 times .

6.2 Minor difficulties

Now after we talked about major difficulties, let's talk about minor one. We will take about things that used to be a big deal but it's no longer that importance. Also some of the things we will talk about isn't in the survey that made by Kaggle. We will only talk about 5 topics and their definitions.

Definitions

1. **Heterogeneous data:** most organizations and companies collect data from various resources and locations. So this creates data with high variability of data types and formats. This will result in a difficulty in integrating the data with each other. So it becomes untruthful, so IT scientists must analyse data formats in the early stage of project. For example: some studies show that sugar is linked to obesity and other studies say the opposite. These studies was to shoe the effect of sugar to obesity (regardless of being a source of calories).

So why this big difference? This is due to many variables as every person has its own and distinct medical history, demographics and diagnostic test.

2. **Scalability:** as we said before the amount of data duplicates radically every year . as the we said that the amount of data on internet increased by 50 times between 2010 and 2020 . so this needs more time , effort , money and IT developers . but to be honest the big data Revenue from big data and business analytics worldwide in 2019 is 189.1billion dollar . for example YouTube was sold to google in 2006 for 1.65 billion dollar for many reasons , one of them is storage , YouTube now worth like 300 billion dollar . as when YouTube became popular between youth . they started uploading many videos on it . this decreased the storage in YouTube servers radically . but at the same time google offered to buy it for a good offer , so it was sold to google . also long time ago , storage was big deal . if you look at the dimensions and price of a normal HDD nowadays and 10 years ago . you will find a big difference. This made storage less in dimensions and for lower prices.
3. **Complexity of managing data quality:** as data from different sources will run into problem of data integration , as the data doesn't have a specific form . it may be a text as search history or videos or photos or even voice recordings . so the problem here is to integrate all these different types of data together to extract a useful information . also to be honest , no one can say that this data is 100% accurate . as we said before that the data can contain impurities .

4. **Tricky process of converting big data into valuable insights :** well actually I can't explain this difficulty without giving example . imagine that a company specialized in collecting data . this company collect the research history of some users in a certain region . this company found that these people in these days are looking for a certain outfit . another company that make clothes used this information . so the clothing company decided to produce this outfit in big amounts . after it did that , a famous football player changed his outfit and start to appear with a new outfit . then the people will want to dress like him . so they will buy the outfit that this player is wearing . so the company that produced the first outfit that is based on big data lost a lot of money . this can be so tricky for any big data company or organization .
5. **Processing :** the processing power of the CPU versus GPU , as in 1999 when Nvidia made the first graphics card for gaming and graphics , it was then used for processing as GPU contains hundreds and thousands of Cuda cores but at lower frequencies than CPU

which helps then to do many easy tasks in parallel . there is a video to show the difference between speed of processing between GPU and CPU .

This video was made by Nvidia , it was to show the difference between CPU and GPU in painting a photo . I know that you are wondering what is the relation between painting a photo and data analysis.

Well actually , the screen you are watching through it this report or even your mobile screen . it contains thousands of pixels , each pixel has its own color which changes simultaneously . so that the photo or any color on your screen can be shown . and guess what. GPU is the hardware part that is responsible for arranging this process for thousands of pixels . which means that GPU can process so many simple operation at the same time. The following is the link to the video that I am talking about: https://www.youtube.com/watch?v=-P28LKWTzrI .

Conclusion

Unfortunately, there is no perfect or near perfect thing in our world, as it was mentioned in the previous parts of this report, despite all those great benefits that result from our employment of big data, such as calculating the user experience in some applications,[6] and using them in health to avoid the outbreak of epidemics and early detection of diseases despite all The difficulties that Big Data faces, for example, the lack of purity of data and other risks, but the worst risk of big data is the risk of information leakage, which at that moment turns your smart phone into a tracking device that enables those who pay the higher price to obtain your private information and a company got involved in that Facebook in 2016.

Sources

[1] https://www.youtube.com/watch?v=FEvMyjKDAQs&t=779s

[2] https://en.wikipedia.org/wiki/Apache_Hadoop

[3] https://databricks.com/session/solving-real-problems-with-apache-spark-archiving-e-discovery-and-supervision

[4] https://www.mongodb.com/nosql-explained

[5] https://www.rd-alliance.org/group/big-data-ig-data-development-ig/wiki/big-data-definition-importance-examples-tools

[6] https://www.linkedin.com/pulse/5-biggest-risks-big-data-bernard-marr

[7] https://www.bbc.com/news/business-49099364

[8]https://locowise.com/blog/what-is-big-data-analytics-on-social-media

[9]https://www.business.com/articles/big-data-social-media-strategies/

[10]https://www.bbc.com/news/technology-47135058

[11]https://freedomhouse.org/report/freedom-on-the-net/2019/the-crisis-of-social-media/social-media-surveillance

[12] https://www.smartdatacollective.com/police-are-using-big-data-to-predict-future-crime-rates/

[13] https://www.mygreatlearning.com/blog/understanding-customers-with-big-data-the-amazon-way/

[14] https://datafloq.com/read/amp/5-ways-big-data-affects-your-personal-life/5735

[15] Benefits and challenges of Big Data in healthcare: an overview of the European initiatives (nih.gov)

[16] National Institutes of Health (NIH) | National Institutes of Health (NIH) — All of Us

[17] https://businessoverbroadway.com/2018/03/18/top-10-challenges-to-practicing-data-science-at-work/

[18] https://www.statista.com/statistics/551501/worldwide-big-data-business-analytics-revenue/

[19] https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-phone-numbers-data-breach-privacy-a9092641.html

[20] https://en.wikipedia.org/wiki/Kaggle

[21] https://www.techrepublic.com/article/challenges-facing-data-science-in-2020-and-four-ways-to-address-them/

[22] https://acuvate.com/blog/challenges-faced-by-data-scientists/