An investigation on the feasibility of Shoe Size Recognition Model

Wanga Mbabe
3566769@myuwc.ac.za
Robert Sobukwe Rd, Bellville Cape Town, 7535

Due Date: May,31, 2020

1 Introduction

1.1 Overview of the Project

Fashion retail is a highly competitive market, where stock management is an significant factor in companies' competitiveness. Precise sales forecasts are therefore important for this environment to succeed. If sales forecasting is unreliable, stock or summary circumstances may arise that directly and instantly affect the competitiveness of the business. If the sales forecast is not correct, situations of stock or overstock can occur that will have a real and tangible impact on the profitability of the business [6]. The effect is not limited to performance in profitability, since an inefficient predictive system can also affect the Quality of Customer Service. For instance , if a customer faces a stock-out situation, he can choose to shop with another distributor[2]. Furthermore, it is also known that a long range of suppliers (such as suppliers of raw materials, manufacturers , distributors and retailers) in the fashion industry are involved in order to place orders before the level of demand for the products is properly understood[3].
Sales predictions are a complex topic given their importance because the quality of a product depends heavily on the personal savor of the customer, which varies widely[7]. Furthermore, the lifecycle is usually very short, replacing new products every new season without any historical sales figures. Fashion collections also consist of a huge number of different products in many different sizes, which are consistent with many different storage units[5]. Precise sales forecasts are therefore essential for planning and business improvement. This paper is focused on the improvement of the stock management by analysing the demands of the customer. To avoid the overstocking and under stocking the company needs to know its customers needs on a deeper level. This work is focused on identifying the shoe size of the customers that visit Takkie Town Store in order to understand what shoe sizes needed to be stock for different types of shoes.

1.2 Objectives of the Project

The objectives of this project is to:

- Develop a Machine Learning Model with trained with still side images of different shoe sizes and shapes.
- The model will convert the length of the shoe in the image, and convert it to standard UK shoe size which customers in South Africa use to identify their personal shoe size.
- The ML model will use the video feed to predict the shoe sizes that entered the store over a given(n) period of time. The data collected, shoe size,timestamp and frequency will be uploaded into a Hadoop File System (HDFS), along with data provided from the retailer such as actual sales and purchase order data
- The data in the HDFS will be used for Business Intelligence (BI) and Analytic. This data can

assist the retailer to get an actual customer profile in terms of the size shoes the customers wear that is visiting the specific store. This will aid in demand forecasting when the retailer replenish their shoes stock levels. Visualisation of the data will be through Reports and Dashboards for user friendly interpretation

1.3 The Need for the Project

With increasing competition, each retailer needs to correctly cope up with the impending demand. The retail firms need to take into account various factors including the lead time and the seasonality of goods. The project is focused on the improvement of the stock management by analysing the demands of the customer. To avoid the overstock-ing and under stocking the company needs to know its customers needs on a deeper level. This work is focused on identifying the shoe size of the customers that visit the Store in order to understand what shoe sizes needed to be stock for different types of shoes.

1.4 Overview of Existing Systems and Technologies

In this section, we review some of the recent video-based recognition methods. In video face recognition, given a test video of a moving face, the first step is to track a set of facial features across all the frames of the video. From the tracked features, one can extract a few key frames that can be used for matching with exemplars in the gallery. Significant work has been done on face tracking using 2D appearance-based models [8]. The 2D approaches; however, do not provide the 3D configuration of the head, and are not robust to large changes in pose or viewpoint. To deal with this problem, several methods have been developed for 3D face tracking. Cascia [4] proposed a cylindrical face model for face tracking. An extension of this work was proposed by Aggarwal [1] that uses a particle filter for state estimation. 3d face recognition system consists some of the use cases that can be implemented in the Shoe Size Recognition Model. But Shoe Size Recognition Model is mainly concerned with Shoe size features, unlike the face recognition system which contains face features.
Main Technologies associated with Shoe Size Recognition Model:

- Keras
- Scikit-learn.
- Tensorflow
- Diagram and design tools (Visio, Nclass, Draw.IO, Microsoft project)
- Video/Camera

1.5 Scope of the Project

Main actor of this system:

- Store manager / Administration staff

Main use cases associated:

- View statistical details

1.6 Deliverance

A Shoe Size Recognition Model. This consists of different classifications and functionalities for various stage. For example sizes of different types(formal,sandal, boots etc.), Since many number of things that are involved, different statistical details will be provided for different shoe type size.

2 Feasibility Study

2.1 Feasibility Financial

The Shoe Size Recognition Model consist of multimedia data transfer, bandwidth required for the operation of this application is very low.
The model will follow the freeware software standards. No cost will be charged from the potential customers. Bug fixes and maintaining tasks will have an associated cost. At the initial stage the potential market space will be the local stores.
Beside the associated cost, there will be many benefits for the customers. Especially the extra effort that is associated stock management will be significantly

reduced while the effort to create descriptive statistical reports will be eliminated, since reports generation is fully automated. From these it's clear that the project is financially feasible.

2.2 Technical Feasibility

The main technologies and tools that are associated with Shoe Size Recognition Model are:

- python
- Google colab
- Airtable
- Keras
- Scikit-learn
- Tensorflow
- Diagram and design tools
  1. Draw.IO,
  2. Visio,

Each of the technologies are freely available and the technical skills required are manageable. Time limitations of the product development and the ease of implementing using these technologies are synchronized.
From these it's clear that the project is technically feasible.

2.3 Resource and Time Feasibility

Resource feasibility Resources that are required for the project includes

- Programming device (Laptop)
- Programming tools (freely available)
- Programming individuals

So it's clear that the project has the required resource feasibility.

2.4 Risk Feasibility

Risk feasibility can be discussed under several contexts.
Risk associated with size:

- Amount of reused software:
  Though the main logics are implemented throughout the project, The model will use some JSP libraries to incorporate additional functionalities such as to support file uploads.
- Number of projected changes to the requirements for the product? Before delivery? After delivery:
  The requirements are clearly identified before the implementation phase. Being a general product (not specific to a single user) the requirements will be changed only if new functionalities are added to the system.

Business impact risks:

- Reasonableness of delivery deadlines:
  Being a 14 weeks project, the project will have several deadlines and deliverables that are scheduled successively. Depending on the coding and designing cost and effort, the deadlines are quite reasonable.
- Amount and quality of product documentation that must be produced and delivered to the customer:
  Customer will be provided with a complete online user manual. As the software is implemented as a freeware and open source system, the code will be available for free.

Customer related risks:
Shoe Size Recognition Model is a general type of product (not designed just for a single store). Before implementing the system in a store, there will be some basic modifications required.
Development environment risks
Is a software project management tool available? Airtable will be used as the main project management tool.

Are tools for analysis and design available? The model will require designing software and Draw.IO (database design) will be used for that.

Process issue risks

Shoe Size Recognition Model will follow the Agile software development process. This provides the flexibility to accommodate changing software requirements of Shoe Size Recognition Model .

Technology risks

Is the technology to be built new? All the technologies are very well established and old enough (but not obsolete).

2.5 Social/Legal Feasibility

Shoe Size Recognition Model uses freely available development tools, and provide the system as an open source system. Only the maintenance cost will be charged from potential customers. JSP Software libraries that are used in this system are free open source libraries. Since this new system eliminates the effort to make statistical distributions, it will have a great impact in a fashion industry.

References

[1] Gaurav Aggarwal, Ashok Veeraraghavan, and Rama Chellappa. 3d facial pose tracking in uncalibrated videos. In *International conference on pattern recognition and machine intelligence*, pages 515–520. Springer, 2005.

[2] Daniel Corsten and Thomas W Gruen. Stock-outs cause walkouts. *Harvard Business Review*, 82(5):26–28, 2004.

[3] He Huang and Qiurui Liu. Intelligent retail forecasting system for new clothing products considering stock-out. *Fibres & Textiles in Eastern Europe*, 2017.

[4] Marco La Cascia, Stan Sclaroff, and Vassilis Athitsos. Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models. *IEEE Transactions on pattern analysis and machine intelligence*, 22(4):322–336, 2000.

[5] Na Liu, Shuyun Ren, Tsan-Ming Choi, Chi-Leung Hui, and Sau-Fun Ng. Sales forecasting for fashion retailing service industry: a review. *Mathematical Problems in Engineering*, 2013, 2013.

[6] Min Xia and Wai Keung Wong. A seasonal discrete grey forecasting model for fashion retailing. *Knowledge-Based Systems*, 57:119–126, 2014.

[7] Min Xia and Wai Keung Wong. A seasonal discrete grey forecasting model for fashion retailing. *Knowledge-Based Systems*, 57:119–126, 2014.

[8] Shaohua Kevin Zhou, Rama Chellappa, and Baback Moghaddam. Visual tracking and recognition using appearance-adaptive models in particle filters. *IEEE Transactions on Image Processing*, 13(11):1491–1506, 2004.

Machine Learning assisted Shoe Size Recognition Model:Security and ethical issues

Heinrich Davids
Department of Economic and Management Sciences
University of the Western Cape
Cape Town, South Africa
2330838@myuwc.ac.za

***Abstract*—Technology has been the cause of hopes, fears and ethical consideration for many years now. The ubiquitous nature of technology makes it almost impossible to engage with it in some form or another. With the arrival of Big Data, the positive and negative connotations of technology have increased. Big Data can be identified by the four V's, Volume, Variety, Veracity and Velocity. Many researchers claim there are plenty more V's or characteristics, but these four are the most prevalent in characterizing Big Data. Organisations are increasingly benefiting from Big Data innovations, whilst equally battling to prevent data breaches and using data in an ethical manner. This research aims to identify the most common security and ethical issues organistions face, and how these issues are applicable to installing a Shoe Size Recognition camera in a fashion retailer store. The purpose if the camera is too identify shoe sizes of customers walking into the store, and use the data collected to improve on demand forecasting.**

Keywords—Big Data, Technology, Organisation, Environmental, Security, Ethics, Privacy, Choice, Trust, Awareness, Machine Learning, Analytics, Hadoop Distributed File System

I. Introduction

The world today is overwhelmed by data and it is increasing day by day [1]. Big Data (BD) is believed to be the new salvation for businesses to reinvigorate their competitive position in the market [2]. BD drives economic growth through technological innovation in businesses [2]. BD refers to data in large volumes, which are continuously generated from various sources (sensors, human beings, machines, equipment), and from multiple environments [3]. Big Data can be identified by the four V's, Volume, Variety, Veracity and Velocity. Many researchers claim there are plenty more V's or characteristics, but these four are the most prevalent in characterizing Big Data. Volume refers to the large size (Gigabytes to Petabytes) of BD; Variety refers to the multiple sources where the data is coming from, for example Social Media platforms, Video, Audio, sensors and ERP systems; Veracity refers to the unpredictable nature of BD, sometimes the data will be accurate and relevant and other times it can be inaccurate and irrelevant; Velocity refers to the speed at which the data is generated and collected.

It is the ubiquitous nature of BD that is giving it a bitter sweet connotation among society's members and business professionals. Every click on the internet or movement with a smart device is in the hands of internet business companies, like Amazon, Google and Facebook [2]. Business professionals are monitored more closely to track their performance and measure their KPIs, for example, truck drivers are monitored via Intelligent Transportation Systems. To make it worse, many of the above mentioned companies has had some form of data breaches in recent years, with many customers' information been leaked to malicious entities [2].

The fashion retail industry has traditionally relied on creativity, intuition and historic Point of Sale (POS) data for designing, buying and merchandising [4]. BD analytics can be used for a number of purposes inter alia, trend analysis, market identification, measuring influencers' impact and understanding the customer[4]. In a time of global uncertainty, with COVID-19 causing unpredictable markets and irregular spending patterns, BD can provide a much needed competitive advantage and improved profitability. Trend forecasting in the fashion industry offer great insight to planners, incorporating Google Trends technology. Google Trends can predict present and near future fashion trends with BD analytics [5]. Demand forecasting is as crucial in the fashion retail industry as it is in any. Knowing your customer's needs not only means the company sells it's products, but also eliminate overproduction which leads to production wastage and also reduce excess inventory which in turn leads to products being marked down and sold for much lower prices [4].

Historical POS data are deemed insufficient for accurate demand forecasting in the fashion retail industry [4]. Companies can only gather data on clients who actually made a purchase, non-purchasing customers are non-existent to retailers, which means they lose valuable information on their actual clientele. Tracking all feet that enters a shoe retailer can give insight to "all" customers walking in the store. Gender classification and shoe size identification by means of a Machine Learning (ML) algorithm, using the Convolution Neural Network technique to predict the gender and shoe sizes of customers walking into the store. This information can be invaluable to shoe retailers. The question is, is it too invasive to the customers, is it legal and what customer data will be saved by the company?

II. Literature Review

A. *Security*

Companies can split their security awareness and policies into three categories, Technological, Organisational and Environmental [6]. For this reason the Technology, Organisation and Environmental (TOE) framework will be applied to discuss Security considerations applicable to implementing a Shoe Size

Recognition Model (SSRM) in the fashion retail industry.

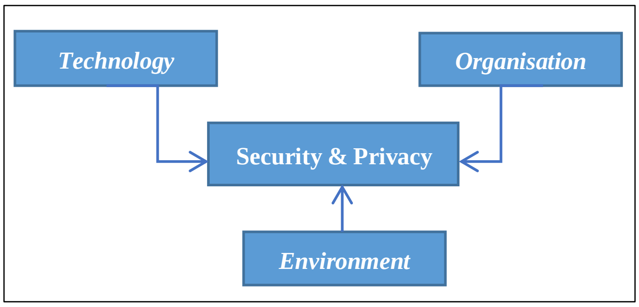


Figure.1 Adapted TOE framework

1) Technology

This section entails the organisation's internal equipment and processes. For BD a strong security solution are needed to ensure data integrity, confidentiality and availability [6]. Traditional security protocols used for traditional Relational Database Systems (RDBMS) for example are not sufficient in a BD environment [4]. BD is generated and analysed at unprecedented speeds, either in batch, real time or near-real time, and this makes it difficult to maintain data protection [6]. Another distinction from traditional data is the data security life cycle [9]. Traditional security technologies does not take into account the whole process of the data security life cycle, for example, data can be shared over multiple devices and cloud storing platforms, BD security need to consider the protection of the data across all platforms/ devices.

Another unique characteristic of BD that needs to be considered in terms of data security is the various formats and sources of the data being collected [6]. This variety characteristic of BD usually means the data is unstructured more than structured. Most security technology is not capable of protecting unstructured data [6]. With rapid transfer of data streams from local servers to BD platforms, for example cloud based storage; a cloud based firewall is needed for data protection and to prevent data loss [6]. Cloud based firewalls provide, availability, scalability and extensibility [7], which means backups in case of site failure, scalability to handle high bandwidth capabilities and a secure communication path. Cloud storage may pose security problems and lead to privacy issues when the data are hosted in a server that is publicly accessible, so companies needs to ensure sufficient vetting of cloud service providers.

Live video stream will be uploaded from an Internet Protocol (IP) camera to a cloud based platform directly from the camera for the video feed storage. The upload will happen via the store WiFi network. The store is part of the corporate brand Pepkor, which means there are already extensive data security protocols in place to ensure a secure upload connection, hence no extra security mechanisms will be needed in terms of the upload. Attention should be given when selecting the cloud storage service to ensure end to end protection and privacy of video footage of Shoe City customers. Cloud computing environments store data from different devices, so it can be a major security risk [11]. The Proof of Concept (POC) of the solution will be based on community based cloud service, and a thorough vetting process will have to be done before selecting a cloud service to ensure adequate protection is in place. Important data protection techniques to look out for are data confidentiality, integrity and data access controllability [11].

Video footage should be periodically removed from the cloud server as a precaution, once analysis has been done and the feedback from the SSRM has been uploaded to HDFS the video footage can be deleted. The cloud based storage service should also offer enough storage capacity and routine backups to prevent massive data loss.

2) Organisation

The Organisational aspect of the framework deals with company culture, strategies, structure and policies [7]. Even with the most advanced security protocols in place, if the staff does not adopt the culture of protecting company assets, the company is not protected. Organisational culture and awareness on security and privacy, is crucial in eliminating human-related security breaches [6]. In 2009 approximately 221 million personal identifying information records became public knowledge because of external information breaches [8]. A good example of human-related security breaches are through phishing [8]. Phishing is when an email user gets an email and the email user reveals confidential information such as usernames, passwords and bank account details by clicking on website link or follows certain instructions. This information can be used to infiltrate the company's networks and use the data for illicit purposes.

To avoid a catastrophic event of a data breach it is crucial companies have sufficient protection mechanisms in place [6], but for a successful security culture in a company technology is not enough, top management needs to promote the security culture and provide the necessary support to the staff [9]. A lack of top management involvement and support may deter all efforts made by IT professionals. Another important organisational security aspect when implementing BD is the level of employee competencies or understanding of BD protection mechanisms [3]. Adequate training should be provided to staff to ensure a good understanding of the company's policies and culture. Employees should acquire data privacy and security training through SETA programs as research suggests training is an effective method of expanding employees' knowledge [10]. A study done by [10], suggests prevented phishing attack incidents dropped a further 10% after employees were trained on the awareness of phishing attacks.

The in-store camera does not pose many data security risks in terms of organisational culture and policies, as all data will be uploaded to a remote cloud server. Adequate training should be provided to staff make sure they do not divulge the WiFi password to any unauthorised people and also to treat the camera company property and not tamper with it. If employees put a memory card in the camera, they can get access to the data and this might pose an ethical problem when company data fall into unauthorised hands. The store manager will have access to dashboards that presents the results of the SSRM calculations, for better planning. This information is propriety to the fashion retailer and might not be divulged to any unauthorised people. Thus from an Organisational security perspective, sufficient training must be given and adequate awareness of data confidentiality and security should be created.

3) *Environment*

The environmental context refers to the environment where the company conducts its business in, which entails the industry, government and the industry [12]. The external environment always has an impact on the internal environment, and is harder to control. In the BD environment, the collection of data from numerous sources is typically involved and many times this data comes from other companies [13]. Data mining companies has become very popular over the last five years. These companies collect data on behalf of other companies, for example, companies that mine social media websites and sell customer interests to corporates. Once sensitive data are being shared cross organization, security, privacy and confidentiality issues arise [6]. Both companies at either side of the data transfer should have adequate security mechanisms in place, and a clear understanding about the intellectual property of the data should be established beforehand. Companies can use data in unethical or even illegal ways, so the data owners should be establish to prevent legal complications and also reputations being tarnished by the acts of other entities.

Integration between companies, like a retailer and their supplier, also poses a security risk as the companies exchange information or data between each other. Companies conceal their security profiles as they want to hide their vulnerabilities to protect their assets [14], so the companies won't necessarily know the security level implemented in the other company. The relationship between system integration and level of security countermeasures is dependent on the firm size, industry and other external factors [15]. In 2018 a study found, the greater the integration between companies, the larger investment companies make in security controls and countermeasures [14].

Another environmental aspect companies need to address when considering their BD security profile is the use of third party tools [6]. Usually in a BD environment companies use third party tools/ applications to store, analyse, access and share data [6]. For example, HDFS is prominently used for BD storage and processing within a BD environment [7]. Dependence on BD third party applications from companies using their services are prevalent because companies does not have the internal security mechanisms in place, thus they rely on third party's security infrastructure [6].

In terms of the SSRM, the data will not be integrated to external companies so no evident considerations need to be taken on that. The data will only be used for the company where the camera is installed in the store. The data will be uploaded to a HDFS, so security and privacy considerations are applicable in terms of the SSRM. Using the HDFS could be beneficial in terms of security mechanisms that do not currently exist within the client company. The data generated inside the retail store, which make the company the data owners of all the data and are responsible for ethical usage of the data.

B. *Ethics*

BD analytics makes use of algorithms to analyse huge datasets to detect patterns, similarities, and other economic and social value [16]. Many researches claims BD analytics has come under criticism recently for having unethical consequences to various stakeholders [17] [18]. Breaching of privacy, discrimination against customers and individual profiling are amongst the issues that has been raised [16]. Companies argue they collect data on consumers to deliver a more personalised and better quality service to the consumers, but consumers feel the incentives are more for the companies than for their needs [16]. The main BD ethical issues from an individual's perspective can be divided into four groups namely, Privacy, Trust, Awareness and Choice [16].

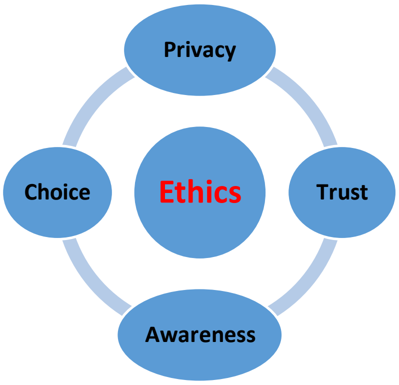


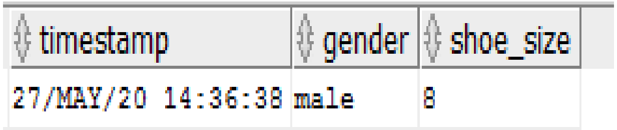
Figure.2 Privacy, Trust, Awareness and Choice

1) *Privacy*

Privacy in the context of BD ethics can be defines as the individuals' capacity to restrict and control how companies use their personal information [19]. This entails how companies access, modify and use personal data. Individuals feel even if they give consent to companies to collect and share their data, they want to decide what data are being collected. For example, if I give consent for companies to collect my Google search history, I still want to restrict my online purchasing data. The information given to companies can be filtered down to exclude personal information on individuals, and only show online searches or purchases and not demographic data, but the aggregation process might re-identify the individual data without the individual's knowledge [20].

Individuals also feel they should be able to modify the data about themselves, to avoid misrepresentation of themselves [16]. Individuals might make purchases on behalf of someone else then get inundated by online suggestions and unsolicited advertisements about those products which are irrelevant to them. Another concern individuals have in terms of their privacy being violated is, sensitive data about them might get shared and create discomfort or make them victims of discrimination [16]. It is a known fact that certain government agencies "spy" on citizen's online activity, and might misread online activity of individuals.

Installing a video camera in the store to only collect data of their feet/ shoes is not invasive to the customer, but some customers might feel their consent is necessary to collect the data. For this a disclaimer needs to be displayed where the customer can see it. The disclaimer will count as consent given by the customer that the company may collect and use the data collected on the customers. No personal data on the customer will be collected, only their shoe size and their gender that will be added on to the SSRM at a later stage. Below an example of the data in the HDFS:



timestamp	gender	shoe_size
27/MAY/20 14:36:38	male	8

Figure.3 Example of data uploaded in HDFS

As illustrated above, no personal customer information will be stored by the SSRM. The information collected should be stipulated in the disclaimer for the customer's knowledge.

2) Trust

Individuals should be able to trust the company's collecting their data, use the data in a responsible, predictable manner and not behave in an opportunistic manner, which leads to inappropriate ways their data are being used [16]. Mistrust is created when companies partake in unauthorised monitoring, unsolicited intrusions and security of personal data [16]. A good example of unauthorised monitoring is when a website does not disclose it uses cookies to store information of the user's browsing activity. Individuals need to feel confident companies collect their data only with their consent and use it for clearly stipulated purposes [20]. Otherwise data collection would lead to individuals feeling they are being monitored while going about everyday life activities, which is too invasive.

Another untrustworthy act from companies is sending unsolicited advertisements, emails and promotional offers, which can be difficult to opt out for individuals [21]. Their personal data could have been sold and widely distributed, so unsubscribing from one company does not usually stop the individual from being inundated by spam. Spam is usually promotional communication sent to individuals via emails or advertisements, many times without individuals consenting. Hence individuals needs to feel confident companies will ensure their data's security [16], throughout the while data security life cycle [9].

For the SSRM to be a feasible solution for the client, no subscriptions or personal contact information are needed. This means the company cannot partake in any untrustworthy acts with the customer's data. The data collected are only gender and shoe size, no email address, contact number or social media information. Also the data collected for the SSRM cannot be linked to the existing POS or customer data the company has, as no personal information will be saved in the HDFS.

3) Awareness

BD analytics is a fairly new term to a lot of consumers. Hence their knowledge of the topic is limited. Awareness of BD is the concept of the individual's understanding of what BD is, their rights regarding BD and understanding who holds the data [16]. Ethical issues arise when there is lack awareness from the individual. To eliminate this, individuals need to educate themselves in general BD policies, regulations and laws that exist to protect them from unknowingly consenting to something they do not understand. For example, in South Africa we have the Protection Of Personal Information Act, 2013 (POPI act) to "*promote the protection of personal information processed by public and private bodies*" [22]. The POPI act is quite extensive to protect the personal information of the individual and hold data owners accountable for misconduct, for example, the data owner are responsible for loss of personal information, unlawful access and processing of personal information [22]. If individuals are aware of the protection the law provides, their trust levels may rise in terms of consenting to their personal information being collected.

Individuals also need to be aware of what data are being collected about them and what third parties have access to their data [23]. Terms and conditions need to be clear for a layman to understand it and not contain mostly IT professional jargon to ease the opting out process if needed.

Customers walking into the shoe retailer will not be aware they are consenting to their data being collected via the SSRM. They will also not be aware what the retailer is going to do with the data collected. For them to be aware of it, a well informative disclaimer needs to be put up. The noninvasive nature of the SSRM/ camera and the nature of the data being collected do not heed for written consent from the customer. The disclaimer should advise the customer to ask further questions should him/ her need more information. The customer is accustomed to security cameras inside the store which they understand are there for their personal protection. If the disclaimer explains the extra camera's (to measure their shoe size) purpose is to ensure they get their shoe size in store upon their visits might make them more trustworthy and won't make them feel their rights has been violated.

4) Choice

Our own choices are one of the most fundamental things about being human. Especially a customer, a customer wants to make their own choices and these choices may change a lot. BD analytics has the ability to restrict individual's choices [24]. BD analytics can profile individuals by age, behaviors, location and gender. This leads to companies categorizing individuals according to the data collected and, only includes them in communication for services and products based on their profile [16]. As a result of this, customers may face a less-than-free market and lose out on choice [25]. Many customers feel the offers on "Takealot Daily Deals" are tailored to their historic purchasing patterns, and all clients do not get the same discounts, although this is just a suspicion there are companies that use similar marketing strategies. Companies also target certain profiles and concentrate product/ service specific marketing on them, also offering rewards and discounts until customers start buying their products [17]. This is perceived as unethical as companies use personal data to manipulate customers in buying their products/ services.

As previously mentioned the SSRM will not be collecting any personal information on customers, so it makes it impossible target specific profiles for marketing of products. The data collected is for demand forecasting for all existing and potential customers.

III. Conclusion

BD can offer many useful benefits to companies, but it also presents unprecedented challenges. The ubiquitous nature of BD makes it invaluable to companies. This same nature makes it complex to manage. Privacy and Security preservation of the data is one of the biggest challenges companies face. Traditional security mechanisms are insufficient to protect BD, mainly because of the volume, veracity, velocity and variety characteristics. BD protection strategies should not only consider what the organization is doing internally, but from a Technological, Organisational and Environmental aspect. All these areas are where BD is vulnerable to security breaches. Thus companies should analyse and invest accordingly.

Ethical considerations are just as important as security when it comes to BD. If the source of data feels their rights or privacy has been violated, or if the company has irresponsible or malicious intent, then they will not be consensual in giving up the data. Companies need to make sure their data sources are treated respectfully and their need and rights are taken into consideration. The four main ethical aspects companies need to take into account are Privacy, Trust, Awareness and Choice. These considerations ensure mutual respect and a good relationship for future data collection.

IV. References

[1] A. Seetharaman, Indu Niranjan, Varun Tandon, and A. S. Saravanan. Impact of big data on the retail industry. Corporate Ownership and Control, 14(1):506–518, 2016.

[2] Wei Fang, Xue Zhi Wen, Yu Zheng, and Ming Zhou. A Survey of Big Data Security and Privacy Preserving. IETE Technical Review (Institution of Electronics and Telecommunication Engineers, India), 34(5):544–560, 2017.

[3] S. Vijayakumar Bharathi. Prioritizing and Ranking the Big Data Information Security Risk Spectrum. Global Journal of Flexible Systems Management, 18(3):183–201, 2017

[4] Big Data in fashion: transforming the retail sector. Journal of Business Strategy, 2019.

[5] Emmanuel Sirimal Silva, Hossein Hassani, Dag Øivind Madsen, and Liz Gee. Googling fashion: forecasting fashion consumer behaviour using google trends. Social Sciences, 8(4):111, 2019.

[6] Technological, Organizational and Environmental Security and Privacy Issues of Big Data: A Literature Review. Procedia Computer Science, 100:19– 28, 2016.

[7] Supriya Haribhau Pawar. A study on big data security and data storageinfrastructure.International Journal, 6(7), 2016.

[8] Dhirendra Sharma, Management Program, Michael Cusumano, and Thesis Supervisor. Enterprise Information Security Management Framework [EISMF] Enterprise Information Security Management Framework [EISMF]. 2011.

[9] Agata McCormac, Dragana Calic, Marcus Butavicius, Kathryn Parsons,Tara Zwaans, and Malcolm Pattinson. A reliable measure of InformationSecurity Awareness and the identification of bias in responses.Australasian Journal of Information Systems, 21:1–12, 2017.Vaibhav Kumar Sarkania and Vinod Kumar Bhalla. International Journal of Advanced Research in. Android Internals, 3(6):143–147, 2013

[10] Tianjian Zhang. Knowledge Expiration in Security Awareness Training.Annual ADFSL Conference on Digital Forensics, Security and Law, (c):197–212, 2018.

[11] Vaibhav Kumar Sarkania and Vinod Kumar Bhalla. International Journal of Advanced Research in.Android Internals, 3(6):143–147, 2013.

[12] Shiwei Sun, Casey G. Cegielski, Lin Jia, and Dianne J. Hall. Understandingthe Factors Affecting the Organizational Adoption of

[13] Big Data.Journal ofComputer Information Systems, 58(3):193–203, 2018.

[14] K. Hayashi. Social issues of big data and cloud: Privacy, confidentiality, andpublic utility. In2013 International Conference on Availability, Reliabilityand Security, pages 506–511, 2013.

[15] Richard Baskerville, Frantz Rowe, and Fran ̧cois Charles Wolff. Integrationof information systems and cybersecurity countermeasures: An exposure torisk perspective.Data Base for Advances in Information Systems, 49(1):33–52, 2018.

[16] Michael R Galbreth and Mikhael Shor. The impact of malicious agents onthe enterprise software industry.Mis Quarterly, pages 595–612, 2010.

[17] Ida Someh, Michael Davern, Christoph F. Breidbach, and Graeme Shanks.Ethical issues in big data analytics: A stakeholder perspective.Communi-cations of the Association for Information Systems, 44(1):718–747, 2019.

[18] Shoshana Zuboff. Big other: surveillance capitalism and the prospects of aninformation civilization.Journal of Information Technology, 30(1):75–89,2015.

[19] Marcus R Wigan and Roger Clarke. Big data's big unintended consequences.Computer, 46(6):46–53, 2013.

[20] France Belanger and Robert E Crossler. Privacy in the digital age: a re-view of information privacy research in information systems.MIS quarterly,35(4):1017–1042, 2011.

[21] Solon Barocas and Helen Nissenbaum. Big data's end run around proceduralprivacy protections.Communications of the ACM, 57(11):31–33, 2014.

[22] Alexander Halavais.Bigger sociological imaginations: Framing big so-cial data theory and methods.Information, Communication & Society,18(5):583–594, 2015.

[23] https://www.justice.gov.za/inforeg/docs/InfoRegSA-POPIA-act2013-004.pdf

[24] Kate Crawford and Jason Schultz. Big data and due process: Toward aframework to redress predictive privacy harms.BCL Rev., 55:93, 2014.

[25] Shoshana Zuboff. Big other: surveillance capitalism and the prospects of aninformation civilization.Journal of Information Technology, 30(1):75–89,2015.

[26] Mike Ananny. Toward an ethics of algorithms: Convening, observation,probability, and timeliness.Science, Technology, & Human Values, 41(1):93–117, 2016.

Machine Learning assisted Shoe Size Recognition Model

1st Jinwei Liu
Department of Computer Science
University of the Western Cape
Cape Town, South Africa
3758155@myuwc.ac.za

I. INTRODUCTION

A. Related Work

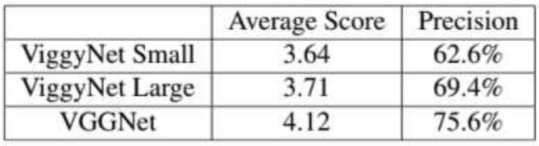
In this section, similar projects that have been conducted shoe recognition or Big Data application on supply chain will be discussed. The main theme of this paper is relevantly new which is using advanced machine learning to detect the presence of shoes and record it, the record will be used to calculate the shoe size and convert into valuable information for decision-making. The papers that related to the study can be partially related, such as object detection or Hadoop framework.

In the past decade, computer vision and data science application in Supply Chain Management (SCM) have gained more attention and the development of these industries are rapid. South Africa is still currently using the traditional way to manage the supply chain, the absence of intelligence in the supply chain makes it very costly.

A paper written by Khosla and Venkataraman [6] proposed that using convolutional neural networks to solve shoe classification and retrieval problems. During the course, Khosla and Venkataraman experimented with different network architecture and fine-tuning for the dataset. In the classification section, the authors prepared three approaches, the first approach follows the general approach of building a Convolutional Neural Network (CNN) which is a select moderately large dataset and train a network from scratch, by tuning the parameters to obtain the best accuracy. This is straightforward to solve the classification problem. The first back approach utilises the transfer learning technique to use an established network and look for the fine-tune for the parameters to obtain the best accuracy. The above two approaches require large computational resources. And in case of failure of both approaches, problem reduction will be applied to reduce the number of classifications, this will allow the model to learn easily.

The second objective of the paper which is retrieval, ti recommend shoes that with similar looking. The performance evaluation for the retrieval problem is much more difficult as it has no absolute answer like the classification problem. The authors planned to calculate significant Confidence Interval (CI) or the result will be evaluated by real humans to determine the success.

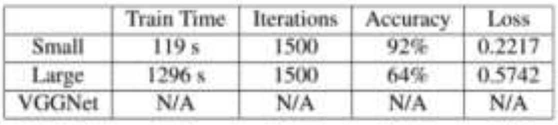
After solving limited computational resource and messy dataset issues, 32,000 men's and women's shoe images were collected and each image has its own short description text file which included name, category, brand, label and et al. Three networks were used and compared in both classification and retrieval problem.



	Average Score	Precision
ViggyNet Small	3.64	62.6%
ViggyNet Large	3.71	69.4%
VGGNet	4.12	75.6%

Figure 1: Result of three networks on classification.

Source: Adapted from [6].



	Train Time	Iterations	Accuracy	Loss
Small	119 s	1500	92%	0.2217
Large	1296 s	1500	64%	0.5742
VGGNet	N/A	N/A	N/A	N/A

Figure 2: Result of three networks on retrieval problem.

Source: Adapted from [6].

This project established a good starting for shoe classification problem and demonstrated the feasibility of using Convolutional Neural Network (CNN) to solve shoe classification. The authors advised the future developers to have access to high-quality image dataset to improve the accuracy of the network, the labelling of the dataset also plays an important role in improving the model accuracy. Furthermore, the authors advised that a stronger computational resource can also improve the model.

Unlike the paper of Khosla and Venkataraman, Mottaghi et al. (2014) [7] used a different approach towards object detection. The authors conducted further analysis of the effect of context in detection and segmentation. PASCAL VOC 2010 is chosen, PASCAL VOC 2010 is test dataset which consists of 20 object classes, such as a person, bird cat, cow, dog, bicycle, car, chair, sofa, etc. In Figure 3, as you can observe that PASCAL VOC 2010 is not like the normal training images dataset, every pixel of PASCAL VOC 2010 is labelled as the figure below. The object segmentation in the image is different.

The authors discovered that it is difficult for existing models to deal with PASCAL imagery. To overcome the problem, the authors proposed a novel deformable part-based model, this model can not only exploits the local context around the object candidate but also the global context at the level of the whole picture. This model has significantly improved in object detection, it is able to detect an object in any scale and most effective on the tiny objects.

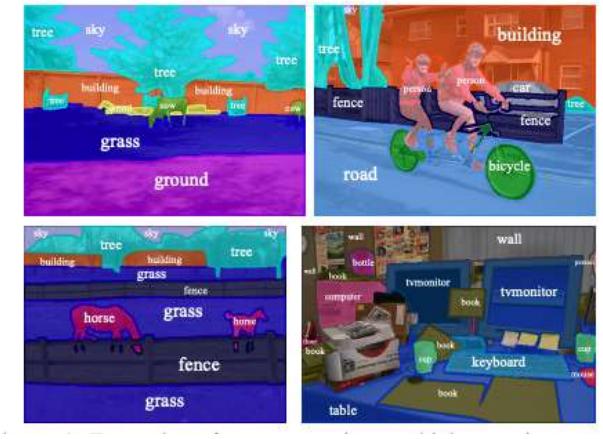


Figure 3: Examples from PASCAL VOC 2010.

Source: Adapted from [7].

Girshick [4] introduced Fast R-CNN which is developed to overcome the problem of R-CNN and SPPnet. The training of R-CNN is a multi-stage pipeline, because of that, the training becomes time-consuming and expensive in space, last and most importantly that the object detection of R-CNN is slow. The new Fast R-CNN is not only fast on object detection but also a higher detection quality, the training has become single-stage, using a multi-task loss, training can update all network layers and no disk storage is required.

The architecture of the Fast R-CNN consists of several convolutional and max pooling layers, these layers are used to produce conv feature map. Then pooling layer extracts a feature vector from the feature map. Each feature vectorial then go through fully connected layers that branch into two sibling output layers.

To compare how effective the Fast R-CNN is, Fast R-CNN, R-CNN and SPPnet gone through three testing dataset VOC 2007, VOC 2010 and VOC 2012. The results suggested that Fast R-CNN is clean and fast network, the improvements that were made are very significant.

Ivanov, Dolgui, Sokolov, Werner and Ivanova(2016) [8] proposed using a dynamic model and an algorithm to solve the short-term supply chain scheduling. The study aimed to optimise the short-term supply chain process which is a multi-stage, multi-objective, flexible flow-shop problem. The team created a dynamic scheduling environment which takes factors like unavailability of the machine, fluctuations of processing time and technological constraints. The result of this study provided a theoretical solution for a real-world scheduling problem and this will help the future development for a sophisticated system for smart factory 4.0.

Yan et al. (2014) [10] proposed using Could of Things as a base to create an integrated management system for a supply chain. The objective of this study is to tackle the current supply chain problem which is lacking in real-time information and agility, improving the supply chain process efficiency. The authors also emphasised the importance of information sharing and collaborations between the links in the supply chain, if information sharing and collaboration is poorly done, it lowers the visibility of the supply chain and leads to information delay and distortion, which can result in the expense of the supply chain increases.

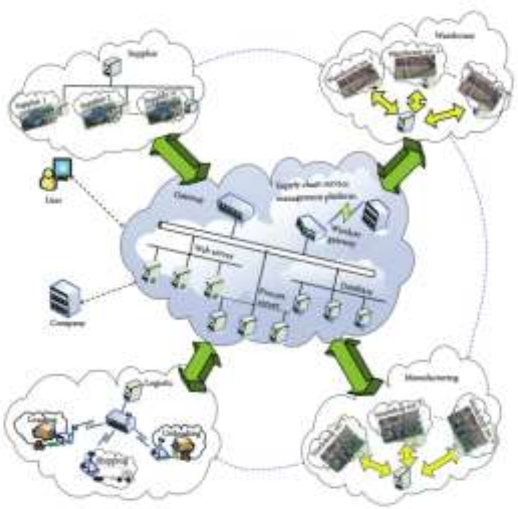


Figure 4: The CoT model for supply chain integration and management.

Source: Adapted from [10].

As shown in Fig4, the study divided the whole supply chain into four sections which includes supplier management, warehouse management, manufacturing management and logistics management. All four sections are connected with each other through CoT, the full connected architecture enable timely communication between each section and allow the tracking of the parcels become much easier.

Each step from manufacturing, warehouse, supplier to client, information about the stocks, communication between different parties are all recorded. These processes created large amount of data, and through CoT to process these data, each party of the integration and management is able to extract relevant information.

However, the traditional non relational database is unable to handle such large heterogeneous data. In order to avoid this situation, Hadoop framework is used to improve data processing. The deployment of Hadoop framework changes the raw data into valuable information which improve the efficiency of the supply chain. Hadoop framework also able to streaming all the data to incorporate with CoT. Hadoop can store large amount of data compare to the traditional database which data storage can be a big problem.

In Figure 5, the whole supply chain is divided into four sections, and using SOA architecture. Each modules are connected with the others. The development of service management platform are based on CoT and the Hadoop which keep producing valuable information and communicating with four main sections.

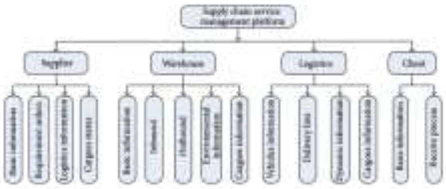


Figure 5: Functional modules of the supply chain service management platform.

Source: Adapted from [10].

Xing et al. (2015) [3] proposed a new platform for Machine Learning on Big Data. This paper is close related to the main theme of our paper which involved both topics Machine Learning and Big Data. Xing et al. suggested in order to fix the space and time bottleneck problem by using Petuum- a system for writing data and model-parallel ML programme.

Jaggi and Kadam (2016) [9] published research paper about Apache spark. In the paper, it indicated that Big data analytics helps the business or an organisation to extract valuable information in vast amount of unstructured data. Big data analytics can help supply chain management become much smoother, and the supply chain will be more intelligent than before. The data generation pace of the supply chain today is super fast, consider the supply chain in a global scale.

In the paper, authors also indicated that the reasons to choose Apache Spark over Hadoop. The three main reasons are computing speed, ease of use and environment. Apache Spark is much faster than Hadoop, in the case of the supply chain, time can be a critical factor. Apache Spark is much easier to operate than Hadoop. Spark can run on different environments even on cloud, also accessible.

Figure 6: Apache Spark logo.

Source: Adapted from [9].

A fast Big Data analytics that capable of processing terabytes of data is very beneficial for an organisation. Today’s supply chain management does demand a tool like Apache Spark. Tests have proven that Spark is 100 times faster than Hadoop in memory. This fast data processing engine can help the organisation stay competitive.

In the comparison of Apache Spark, Singh and Kaur (2014) [2] published a paper nohow Hadoop handling the Big Data challenges. The three major challenges of Big Data also as known as the 3V, volume, velocity, variety. MapReduce framework which is one of the main components of Hadoop, it splits the data into small parts, this process is called data segmentation. Data segmentation is able to handle the large volume of data. The Hadoop filesystem is capable of handling any forms of unstructured data.

In conclusion, Hadoop is capable of handling Big Data challenges. It has features such as MapReduce, filesystem and fault tolerant to make it reliable.

Blum, Springenberg, Wülfing and Riedmiller [5] did another work on object recognition in RGB-D data. Another new approach for object recognition by using convolutional k-means descriptor. Unlike the traditional neural network, the descriptor “studies” the interest point and combine all the information. As figure 7 demonstrated below, the descriptor pinpointing both the colour and the depth.

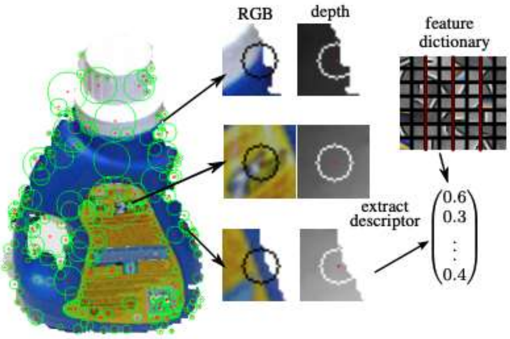
To efficiently process large RGB-D images, the authors made the following modifications in the original algorithm. Since the dimensionality of all images is high, the model will only learn feature responses of the interest point nearby. A fourth channel was added to the response to the RGB-D images. Lastly, to improve the unsupervised algorithm, the pre-processing was modified and bootstrap was introduced. The two modifications were added to help the model to cope with high dimensionality images.

The result of the new descriptor shows that this approach has successfully learnt meaningful features from RGB-D images. The authors suggest that learned feature descriptor can be a valuable tool shortly to improve real-world object recognition application’s accuracy.

Identify applicable funding agency here. If none, delete this text

Figure 7: Example of general descriptor extraction procedure.

Source: Adapted from [5].



Suthaharan (2013) [1] proposed that combine both Big Data and Machine Learning to tackle network intrusion traffic problem. Suthaharan first defined that how is network intrusion traffic problem possesses the characterises of Big Data. Then Suthaharan addressed the network topology, communication challenges and security challenges. On the top of the traditional machine learning approach, Suthaharan proposed the application of Representation Learning and Machine Lifelong Learning to tackle the problem.

The result of the study revealed that the feasibility of using Representation Learning and Machine Lifelong Learning on the network intrusion traffic problem.

References

1. Langlois J A, Maggi S and Crepaldi G 2000 Workshop on hip fracture registries in Europe *Aging Clin. Exp. Res.* **12** 398–401
2. Singh K and Kaur R 2014 Hadoop: Addressing challenges of Big Data *Souvenir 2014 IEEE Int. Adv. Comput. Conf. IACC 2014* 686–9
3. Xing E P, Ho Q, Dai W, Kim J K, Wei J, Lee S, Zheng X, Xie P, Kumar A and Yu Y 2015 Petuum: A new platform for distributed machine learning on big data *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* **2015-Augus** 1335–44
4. Girshick R 2015 Fast R-CNN *Proc. IEEE Int. Conf. Comput. Vis.* **2015 Inter** 1440–8
5. Blum M, Springenberg J T, Wülfing J and Riedmiller M 2012 A learned feature descriptor for object recognition in RGB-D data *Proc. - IEEE Int. Conf. Robot. Autom.* 1298–303
6. Khosla N and Venkataraman V Building Image-Based Shoe Search Using Convolutional Neural Networks
7. Mottaghi R, Chen X, Liu X, Cho N-G, Lee S-W, Fidler S, Urtasun R and Yuille A L 2014 The Role of Context for Object Detection and Semantic Segmentation in the Wild. BT - 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014 *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 891–8
8. Ivanov D, Dolgui A, Sokolov B, Werner F and Ivanova M 2016 A dynamic model and an algorithm for short-term supply chain scheduling in the smart factory industry 4.0 *Int. J. Prod. Res.* **54** 386–4
9. Jaggi H S and Kadam S S 2016 Integration of Spark framework in Supply Chain Management *Procedia Comput. Sci.* **79** 1013–20
10. Awad, Hussain & Nassar, Mohammad. (2010). Supply Chain Integration: Definition and Challenges. Lecture Notes in Engineering and Computer Science. 2180.

Big Data Engineering Technologies

Lwando Maciti
Computer Science Department
University of the Western Cape
Bellville, Cape Town, South Africa
3756147@myuwc.ac.za

***Abstract*- Access to the internet using technological devices has become easier than it was before. As a result, large amounts of data are being produced daily from different technological platforms and devices such as smartphones, laptops, smartwatches, and other IoT devices. These large amounts of data are often unstructured and meaningless. However, it becomes of high value once it has been processed and analyzed. The relevant organizations may find value from the trends and patterns identified from the data – what they do with theses analysis is what gives the data real meaning and hence the term "Big Data". This paper aims to discuss the different ways of producing large amounts of data and its challenges and how big data technologies can be used to process and analyze large amounts of data to obtain value from it.**

Keywords—Big Data, Big Data Analytics, Data Engineering

I. Introduction

Everyone owning or using a technological device produces data even without realizing it. For example, we produce data from social media posts, electronic transactions, sensor devices such as temperature scanners, and even from the pictures we take on our phones. This data can either be very useful or of no use at all, depending on what problem the data is going to be used to solve.

There are various fields where data can be used, such as Machine Learning, where data is used to develop classification or predictive models and in Business Intelligence which involves the descriptive analysis on how the business is performing and how it has previously performed [1]. Big data is not just big data because it consists of large amounts of data, which gives it its essence is the value it provides from the analysis made on it. Data can be a very important asset in an organization as it can help them identify areas of improvement which consequently improves business productivity. Big data analysis generally improves decision making within organizations or an individual. Data seems to be becoming an asset in businesses, so much that they are investing in improving their data infrastructures [2].

Big data engineering can be formally described as the systematic process of ensuring that data is made available to its stakeholders when required. It is a newly formed field with its focus being "the design, implementation, and maintenance of distributed information systems" [2]. Big data is often produced in such large amounts of quantities that it cannot be captured, stored, and analyzed using normal databases [3]. Hence, advanced systems need to be engineered to facilitate the capturing, storage, and analysis of this data. To do so, the nature of the data needs to be understood, as well as the purpose it will be serving should be clearly defined. As a result, technologies such as Hadoop and many others have been developed to facilitate the storage and processing of big data.

II. Background

The main objective of engineering big data is to ensure that all data needs required by the information systems have been addressed. Big data engineering is a guided process consisting of the following important components, data handling a representation, data architecture (expert systems), data construction, handling issues concerned with application and management, tools for specifying and developing data and implementation, and design [2].

Data analysis is regarded as an important tool in data engineering, it makes use of statistical models and techniques to evaluate data [2]. Data analysis also helps to establish important relationships found in the data, it allows for comparisons between data variables to be made and allows us to predict possible future events or forecasting [4]. Figure 1 below outlines the steps to follow when performing data analysis.

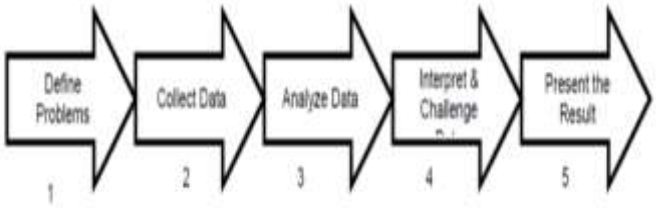


Figure 1 Data analysis procedure [2].

A. Define the problem

This is where you clearly define the nature and scope of the problem you aim to solve using big data engineering, this will also help in identifying ways of collecting the required data.

B. Collect data

Identify the data collection source suitable for the problem you want to solve, for example, social media posts, temperature sensors et cetera and start the data collection process.

C. Analyze the data

Perform data cleansing and processing and develop statistical models to identify any trends and important patterns in the data.

D. Interpret the information

Add meaning to the data by explaining and drawing conclusions about the identified trends and patterns mean.

E. Present the results

This consists of the written reports containing the visualization of the results, also making note of all the limitations of the experiment and the conclusions you can and cannot make [2][4].

III. THREE FORMS OF BIG DATA

A. *Structured Data*

The structure of this type of data is well defined. The data types and structure are predefined in a schema, and all the records are split into corresponding rows and columns. This type of data is easy to read and manage – it is normally presented in a relational database. The main sources of this data are usually humans and machines, data can be manually added into the database system by its users. However, this type of data "accounts for only 20% of the total available data" [5][6]. Figure 2 below is a typical example of a database schema and a table in a database system.

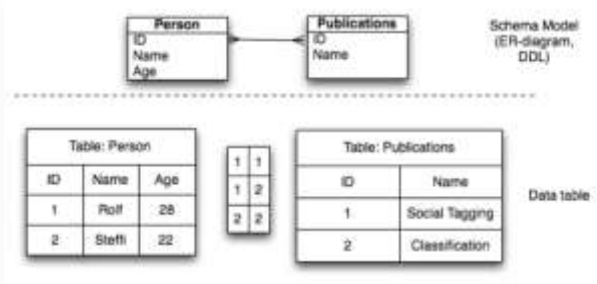


Table: Person		
ID	Name	Age
1	Rolf	28
2	Steffi	22

1	1
1	2
2	2

Table: Publications	
ID	Name
1	Social Tagging
2	Classification

Figure 2 Database table and schema example [6].

B. *Semi-structured Data*

This type of data is not predefined in a schema; however, it is possible to define it in the form of a schema if required. It is not so different from structured data; the main distinguishing factor is that semi-structured data cannot be stored in a relational database. This type of data also allows for variations in its structure, for example, it can contain duplicate and there can also exist minor changes to the structure. CSV files and no SQL are popular examples of semi-structured data [5][6]. Figure 3 below shows a typical example of an RDF graph which represents unstructured data.

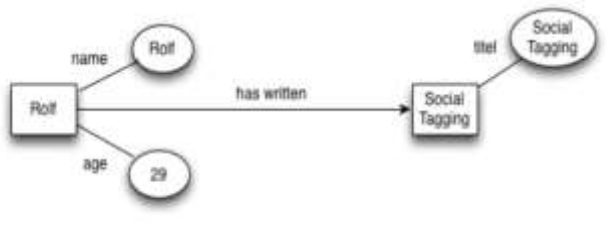


Figure 3 RDF Graph example [6].

C. *Unstructured Data*

Unstructured data is data that has no defined structured or format of storage. A traditional database system can not be used as a means of storing this kind of data. Any kind of data can be stored in unstructured data since it has no restrictions on the format or structure of the data. However, this limits the constraints that can be applied to the data, this also adds complexity to the storage and management of the data [5][6]. A typical example of this type of data could be data retrieved from twitter posts. Different forms of data can be retrieved from a twitter post, for example, images, text, video, and the number of likes and retweets. Table 1 below represents the growth rate of unstructured data produced by different companies according to [7].

TABLE 1 Growth rate of unstructured data:

Company	Generated Data
Digital Study	1,227 Exabyte in 2010 Predicted 1.8 Zettabyte data creation annually in 2011 2.7 Zettabytes in 2013
YouTube	48 hours of new video uploaded every minute
Facebook	34,722 Likes received every minute 100 terabytes uploaded daily 30+ Petabytes (stores, accesses, and analyzes) 30 Billion Pieces of content shared monthly
Domain Name	571 new websites are created every minute
Web store	More than 2.5 petabytes hourly
Twitter	Roughly 175 million tweets every day More than 465 million accounts
Boeing 787	40 terabytes (TB) per hour of flight
Oil drilling	Up to 2.4 TB per minute
Automated manufacturing facility	Approximately 1 TB per hour
Large retail store	Approximately 10 gigabytes (GB) per hour
Global data center IP traffic	8.6 Zettabytes annually
Universe Data generated by IoE	400 ZB
The world	Creates 2.5 quintillion bytes of data per day

Figure 4 Growth of unstructured data [7].

IV. MANAGING BIG DATA USING BIG DATA TOOLS

As we can see from table 1 that data is rapidly growing from the extensive production of data from the different platforms. This data is too large to store and manage using traditional database systems [3]. This may be because the current computer hardware may not be able to handle a large amount of data processing and manipulation. This leads to the need of developing advanced systems that can handle computations of large amounts of data. The Apache Hadoop software was developed as a possible solution to this problem.

A. *Hadoop*

Hadoop is an open-source software developed by Apache to solve the problem of processing large amounts of data by providing a "reliable shared storage and analysis system". [8]. Hadoop consists of two main components, Hadoop File Systems (HDFS) which handles data storage and MapReduce which provides the analysis of the data [8][9].

The downside about Hadoop is that it is more suitable for batch jobs, which is the processing of data that has already been captured, therefore, it is not suitable for live data streaming or interactive data analysis [10].

B. *HDFS*

Hadoop File Systems is a distributed file system used by Hadoop for reliable storage and access to large volume data. “It consists of blocks, name node (master) and data node (slave)” [11]. These blocks indicate the minimum amount of data that it can perform a read or write operation [11].

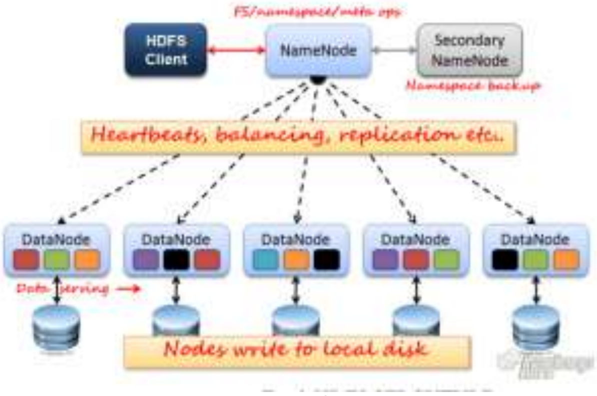


Figure 5 Structure of an HDFS [11].

Some of the pros of HDFS are that it can contain high bandwidth to support map-reduce operations. It is inexpensive and data can be written once and read several times. The cons of HDFS are that it is quite complex to manage the different clusters and joining multiple databases takes a lot of computational time [11].

C. *MapReduce*

Map-reduce is a tool used by Hadoop for the parallel processing of large amounts of data in many clusters. It consists of two nodes, the Job tracker node, and the Task tracker node. The job tracker node allocates tasks and resources to the task node, the task node then executes the allocated tasks. “It contains two important functions, the MAP() and the REDUCE() function” [11].

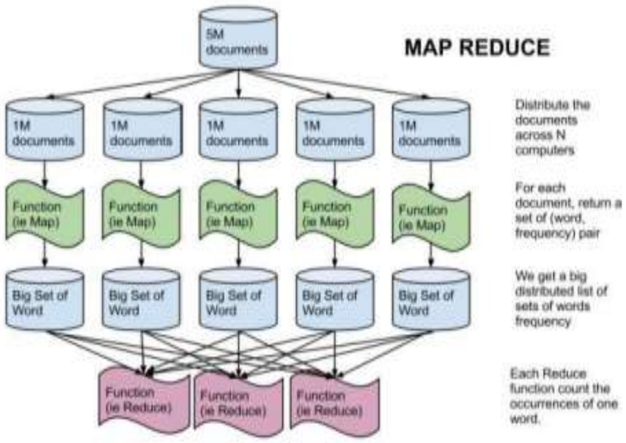


Figure 6 Map Reduce structure [11].

The following are the two map-reduce functions:

- *MAPPER ()*
  This function takes grouped data and divides it into multiple splits and individual elements are broken into key/value tuples [11].
- *REDUCER ()*
  It retrieves the input from the map () function and joins it the data tuples into smaller tuples. This function initiates after the map () function to produce the final output [11].

Map-reduce supports various programming languages including java. It cannot be used for intensive interactive data analysis and it only works for batch processing [11].

D. *SQOOP*

The main function of the Sqoop framework is to coordinate the transfer of data between Hadoop and a database system in a parallel manner. It “refers to SQL to Hadoop” and vice versa. The ETL processes, “Extraction, Transformation and Loading can be performed using Sqoop”. Sqoop provides the movement of a mixture of data with easy integration. It can import a single table in a database up to an entire database into HDFS [11].

V. Real-time Data Processing (Stream-processing)

The Hadoop framework is more suitable for batch jobs or batch processing since its map-reduce function does not support stream processing, this means that it cannot provide real-time data analytics. This indicates a need for real-time (non-batch processing) data processing techniques [10]. Therefore, we will be looking at stream processing as a means of obtaining real-time data processing or interactive data analysis.

Data streams refer to data that is continuously generated at any rate. It is characterized by the “continuous arrival of data objects, unordered arrival of data objects, and an unbounded size of a stream [12]. Typical examples could be twitter likes, clickstreams, and message streams [10].

Map-reduce does not support streaming, however, by treating the batches of data as small chunks allow for map-reduce to handle streaming through a process called micro-batching. Map-reduce implementations such as Spark support this method of streaming through a process “called discretized stream or DStream” [10]. Data streams are applicable in several scenarios for example, in continuous queries, where queries are made on data streams to monitor changes in data for the continuous response to queries. Datastream mining can also obtain valuable information from stream data [12].

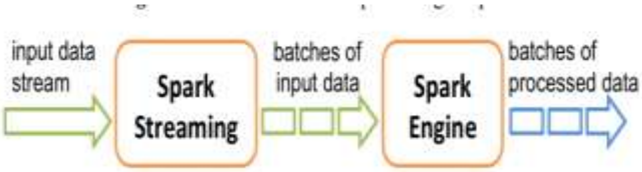


Figure 7 Stream processing schema with Spark [10].

VI. CHALLENGES WITH BIG DATA

Datastores are becoming large and data engineering is facing the difficulty of simplifying the access and management of the data [2]. The extensive production of data gives rise to the high variability in the data. Data is now produced in many different types and formats as a result of the increase in usage of smart devices and the Internet of Things (IoT). The high speed of data production exceeds the capacity of current software and, consequently, these software systems are not able to handle the processing of these large amounts of data. Since the data is generated from various sources, there could be issues of misspelling or misclassification which could make the data inaccurate and consequently lowers the overall quality of the data [13]. Other issues to be concerned with are, privacy issues where the use of personal data is applicable. Data security is another challenge; data needs to be protected and secured to prevent the damage that may be caused when it lands on the wrong hands.

VII. CONCLUSION

The different forms of big data have been outlined and the examples thereof. From that we can deduct that, big data is mostly generated from unstructured data from different smart or IoT devices. We also revealed that stream processing is a more useful way of processing data as it provides the ability to interactively analyze and query data. Big data can be applied in many fields, including machine learning, Business Intelligence, and data mining. In as much as the idea of big data engineering is exciting, the extensive production of data leads to several challenges such as data quality issues, data processing issues, and data volume challenges. Lastly, an important thing to note about big data is that big data is not just big data on its own, but what makes it big data is the processing and analysis you do with it.

REFERENCES

[1] M. Niño and A. Illarramendi, “Entendiendo El Big Data: Antecedentes, Origen Y Desarrollo Posterior,” *Dyna New Technol.*, vol. 2, no. 3, p. [8 p.]-[8 p.], 2015, doi: 10.6036/nt7835.

[2] K. Eze, “The Essence of Data Engineering,” no. November, pp. 2–4, 2018.

[3] L. Wang and C. A. Alexander, “Big data in design and manufacturing engineering,” *Am. J. Eng. Appl. Sci.*, vol. 8, no. 2, pp. 223–232, 2015, DOI: 10.3844/ajeassp.2015.223.232.

[4] “Data analysis, interpretation, and presentation,” *Use 137Cs Soil Eros. Assess.*, pp. 34–39, 2019, DOI: 10.18356/a2ceb52f-en.

[5] P. Tiwari, “Comparative Analysis of Big Data,” *Int. J. Comput. Appl.*, vol. 140, no. 7, pp. 24–29, 2016, DOI: 10.5120/ijca2016909400.

[6] R. Sint, S. Schaffert, S. Stroka, and R. Ferstl, “Combining unstructured, fully structured and semi-structured information in semantic wikis,” *CEUR Workshop Proc.*, vol. 464, no. May, pp. 73–87, 2009.

[7] A. C. Eberendu, “Unstructured Data: an overview of the data of Big Data,” *Int. J. Comput. Trends Technol.*, vol. 38, no. 1, pp. 46–50, 2016, DOI: 10.14445/22312803/ijctt-v38p109.

[8] S. Blazhievsky, “Introduction to Hadoop, MapReduce, and HDFS for Big Data Applications,” p. 67, 2013.

[9] P. J. Charles, S. T. Bharathi, and V. Susmitha, “Big Data – Concepts, Analytics, Architectures – Overview,” *Int. Res. J. Eng. Technol.*, vol. 5, no. 2, pp. 125–129, 2018.

[10] S. Shahrivari, “Beyond batch processing: Towards real-time and streaming big data,” *Computers*, vol. 3, no. 4, pp. 117–129, 2014, DOI: 10.3390/computers3040117.

[11] S. Suguna, “Improvement of HADOOP Ecosystem and Their Pros and Cons in Big Data,” *Int. J. Eng. Comput. Sci.*, no. May 2016, 2016, DOI: 10.18535/ijecs/v5i5.57.

[12] D. Namiot, “On Big Data Stream Processing,” *Int. J. Open Inf. Technol.*, vol. 3, no. 8, pp. 48–51, 2015, [Online]. Available: http://injoit.org/index.php/j1/article/view/225.

[13] Q. Maqbool and A. Habib, “5Big Data challenges,” *Control Eng.*, vol. 66, no. 3, p. 33, 2019, DOI: 10.4172/2324-9307.1000133.