# HOW CAN ORGANISATIONS INTEGRATE THEIR DATA-ORIENTED PROCESSES WITH THE USE OF BIG DATA, AND EFFECTIVELY UTILISE THESE TECHNIQUES TO IMPROVE THEIR BUSINESS PROCCESSES?

## IS5102 ASS3 Group E

Artisom Hramyka (140022099), Chun Paul HO (180029517), Siqi Chen (180010145), Yani Hu (180003766) & Stepan Mincev (180028634)

## Contents

Introduction ....3
Big Data Dimensions ....4
Data Processing ....5
Big Data creation ....5
Big Data collection ....5
Big Data transmission ....5
Big Data pre-processing ....6
Discretisation ....6
Normalisation ....6
Feature extraction ....6
Feature indexing and encoding ....6
Incomplete data problems ....7
Big Data storage ....7
NoSQL Databases ....8
MongoDB ....8
Cassandra ....8
Apache Hadoop ....9
Big Data analysis ....9
Big Data Coding Languages ....10
R ....10
Python ....10
Other Big Data Software Tools ....10
Apache Spark ....10
Apache Hive ....11
Big Data in Industry: ....11
Retail: ....11
Marketing: ....11
Product Development: ....12
ASOS: ....12
Voluminous (quantity of data) ....12
Variety (numerous sources) ....13
Velocity (frequency of data) ....13
Veracity (uncertainty of data) ....13
Big Data gathered from social Media ....13
Types of Big Data used by e-commerce retailers (ASOS) ....14
Transaction/business activity data ....14
Click-Stream Data ....14
Video data ....15

Voice data ....15
Drive Innovation ....15
Big Data Problems and Limitation ....16
Data Privacy ....16
Data Security ....16
Data Storage and Transportation ....17
Image 1- Data Estimation Transfer Time Google (2017) ....18
Data Discrimination ....18
Data Quality ....19
Data Volume ....20
Data Change ....20
Lack of consensual data quality standards ....20
Data Quality versus Quantity ....20
Scalability ....21
Data Ownership ....21
Future research directions ....22
Effective data cleaning and life cycle of data ....23
Storage and dynamic data placement ....24
Develop algorithms and Scalability software platforms ....24
Ethical considerations ....25
Conclusion ....25
Finished references ....26

# Introduction

In 2001 Big Data has been defined by Gartner as Data and Analytics of high volume, velocity and varied “information assets which require cost-effective innovations for information processing to ensure enhanced insight, decision making and process automation” (Gartner,2018). Big Data is not a new discovery and was conceptualised in the 1960s during the origins of data processing centres. In 2005 the world realised the sheer amounts of data generated through emerging social networks. With the recent emergence of Big Data, companies and organisations today are realising the importance to specialise in this area and have even gone the extra step to hire expert data scientists, whilst academic and software developers rush to devise programs relevant for data science. Big Data is often defined as the data sets which cannot undergo traditional processing and require innovation. Big Data innovations can also be implemented in Data Mining which is exponentially increasing in popularity.

One problem amongst the public today though is defining data science itself. This confusion can potentially lead to a misinterpretation of what Big Data is, but also have negative consequences on their context applications. Companies today are fighting for a competitive advantage, by finding ways to manipulate and exploit Big Data, however as more organisations look to do this it is imperative that they understand the problems and limitations which could arise. Manyika et al.,(2011) estimated that a retail company could boost their operating margin by over 60%. Within retail there is prediction that those unable to utilise big data will be left behind by laggards (Brynjolfsson et al., 2011). Companies now have been looking to develop their Big Data competencies. Our report will investigate what exactly Big Data is, the business applications of Big Data in retail, Big Data software tools available, and arising problems. The contextual application of this report is primarily focusing on e-commerce retail companies, as opposed to the traditional social media perspectives to Big Data.

Focusing on the software tools which are at organisations disposal today, Hadoop (bespoke Open Source Framework), Hbase and Couch DB have been under the scope of media. Along with the different software applications specialising in Big Data Analytics requires specific coding languages. Open-source frameworks including Hadoop and Spark were critical towards the emergence and continuous growth of Big Data as they permit user friendly storage solutions. Since the development of open-source frameworks the sheer volume of big data grown at a rate in which companies are unable to keep up in terms of processing power. Companies are constantly trying to improve their Big Data algorithm processing plateaus to reach a “knee”, however after a certain point it is no longer proportionate to required resources (Kailser et al. 2013). The issues of Big Data for organisations are later explored in this report.

## Big Data Dimensions

As mentioned Big Data holds a definition, however many academics recognise that Big Data cannot solely be defined with one statement, rather it holds dimensions. Initially, big data was characterised by three dimensions, which has progressively been developed into a now recognised 6V model. This section will identify each dimension.

Volume:

- Singh and Zikopoulos et al. categorise the magnitude of data which is being both generated and collected is continuously increasing, recently increasing at a faster rate of petabytes rather than terabytes. This increase has been recognised to be particularly due to the increase in video data which require different data management technologies (Gandomi, 2015)

Velocity:

- Academics agree that velocity refers to the rate of generation of data. Gandomi and Haider have stressed the critical role of time, as the increasing rate of data generation should be processed and analysed in real-time.

Variety:

- Similarly, Singh and Zikopoulos et al. categorise variety as the different types of data that is being generated and captured. This definition is extended to the variety of data stemming from structured to semi-structured and lastly unstructured. Cukier has stated that structured data is known as data which can be organised using pre-defined data models and constitutes merely 5% of current data and is decreasing annually. Contrarily, unstructured data cannot be organised using pre-defined data models, rather requires data management file systems. Lastly, semi-structure data refers to data that falls between the two categories such as Extensible Mark-up Language.

The initial 3V model has only been developed further with the following three dimensions.

Veracity:

- Implemented by IBM, veracity denotes the unreliability of data. Unreliable also referred to as ‘irrelevant’ data is largely seen from social media as sentiment analysis through social media is uncertain.

Variability:

- Added by SAS, variability refers to the inconsistencies in variation of the flow rate of data, caused due to geographical locations having unique semantics.

Low-Value Density:

- The most recent dimension of Big Data, recognised by Sun and Heller in 2012, refers to data which in its original form is unusable. Examples include website logs which cannot be to obtain business value, rather it must be analysed to potentially predict customer behaviour.

The 6V dimensions are continuously developing as the concept of big data is continuously being developed.

## Data Processing

In this section we will discuss the lifecycle of Big Data from its creation and decision making. Along the way we will analyse state-of-the-art tools and techniques used for Big Data processing.

## Big Data creation

The first part of data lifecycle is data creation. The way the data is created affects what kind of data it is. For example, Facebook has friendships and Twitter has followings. You can collect data about who is friends with who on Facebook and who follows who on Twitter; although these concepts are somewhat similar they have a major difference: friendship is a two-sided relationship while following is one sided. Thus, the data created and consecutively collected is different. If some entrepreneurs were to, say, create a social network “Faceter” with intention to collect some specific form of information from network usage, they might have incorporated their vision on what data they would like to collect into the design of the product.

## Big Data collection

After data is created the next step is data collection. There are various ways to collect data. There are two things to keep in mind:

1. type of data collected, for eg. log files, text, voice, images, video.
2. tools for collecting data, for eg. mobile devices, cameras, sensors, smart watches, accelerometers

## Big Data transmission

After large amounts of data have been collected there is another task on the way: transfer this data. Data needs to be transferred to a data storage, so it can be passed to processing infrastructure for processing and analysis. There are two phases of this process:

1. Transferring from collecting tool to data centre
2. Transferring within a data centre

## Big Data pre-processing

Data collected might not have the best quality, this is where pre-processing techniques are used. Analytical Tools are being used to increase the quality of data, organisations work with, resulting in better quality results. Garcia et al. state various Big Data pre-processing techniques, available today for companies to use these are discussed below (Garcia et al. 2016).

## Discretisation

Discretisation is transformation of continuous variables into discrete intervals. This might come in handy for various Machine Learning algorithms.

## Normalisation

Normalisation is a technique for producing a set of suitable relations that support the data requirements of an enterprise.

## Feature extraction

Feature extraction, extracts new set of features by combining original features. Techniques include:

1. Polynomial expansion - expands the set of features into a polynomial space.
2. Vector Assembler
3. Single Value Decomposition
4. Principal component analysis

### *Feature Selection*

Selecting relevant features without losing too much information, finding a threshold in some sense. Techniques include:

1. VectorSlicer
2. RFormula
3. Chi-Squared selector

## Feature indexing and encoding

Indexing or encoding features from one type to another. Techniques include:

1. StringIndexer: converts a column of string into numerical values, where values represent frequency of string occurrences
2. OneHotEncoder: converts a column of strings into binary columns where each column represents a unique string from the original column and row has the value of one if this row corresponded to that unique column string in the original column
3. Vector indexer: automatically decides which features are categorical and transform them to category indices

## Incomplete data problems

There is a high possibility that the data will have missing values caused either by human error, machine-error or natural input error causing gaps. This issue must be faced early in the process so that it does not affect future steps. There are various techniques for imputing missing values.

## Big Data storage

Companies need to think about how to store their big data. As the nature of Big Data is commonly unstructured (although a specific volume size is usually not defined, big data storage usually refers to volumes that grow to terabyte or petabyte scale (Oracle.com, 2016; SearchDataManagement, 2018), we cannot use traditional relational databases. Therefore, industries and academia use different type of storage such as Big Data Storage (Techopedia.com, 2018) - storage infrastructure which is specifically used for Big Data storage, retrieval and managing.

There are various types of storage systems available for industries to use. The traditional method for managing structured data is known as Relational Database Management System (RDBMS). RDBMS utilises a personal database and schema to store and retrieve data. Traditionally for grander datasets, data warehouses have been used, however due to atomicity, consistency, isolation and durability (ACID), RDBMS does not support the scaling of data, and is incapable of handling semi-structured and unstructured data, leading to the creation of NoSQL management systems. Below are some of the more important software systems which are being used to store big data, including NoSQL(Strohbach et al. 2016).

- Distributed File Systems - file systems that provide a way to store unstructured data in a reliable way. One of the most widespread such systems is Hadoop Distributed File System (Shvachko et al. 2010). HDFS is one of the most used file systems is partly because it is integral part of Apache Hadoop framework (White 2012). HDFS distributes data across some number of servers which both host data and execute tasks (Shvachko et al. 2010).
- NoSQL (Not only SQL) Databases - sometimes called “Not Relational”(Rick Cattell 2010) are databases that do not have to be relational and do not

necessarily follow ACID(atomicity, consistency, isolation and durability) qualities.

- NewSQL Databases - a newly developed form of relational databases that provide NoSQL-like level of scalability following ACID guarantees.
- Big Data Querying Platforms - platforms that provide SQL-like query functionality on top of traditional big data storage systems like DFS or NoSQL.

Each of them has their own pros and cons and it is up for a company to decide on which benefits they are willing to have and which sacrifices they are willing to make.

## NoSQL Databases

We have decided to focus largely on NoSQL databases which are generating an increase in popularity. With the rise of the Internet web2.0 website, the traditional relational database has become unable to cope with the ultra-large-scale pure dynamic website, exposing many insurmountable problems. NoSQL have evolved very rapidly due to their own characteristics. NoSQL includes many products such as the most classic MongoDB, Redis, Hadoop, and Cassandra produced by many major websites.

## MongoDB

- MongoDB is a Document store of NoSQL database. The data is stored in MongoDB in the form of a document (corresponding to a relational database record, sometimes mixed). The document is a JSON string that matches the user’s reading habit. JSON is well supported in mainstream computer languages such as Java and Python. Once the data is read from MongoDB, it can be used without conversion. Facebook as an example, likes, comments, tags, avatars, such attributes can be nested directly in the document in the JSON subdocuments, one query directly to get all the content, no need to do multi-join join. MongoDB is published under a combination of the Server-Side Public License and the Apache License. MongoDB is known to be used by the City of Chicago, Codecademy, Foursquare, IBM, Orange S.A., The Gap, Inc., Uber, and Urban Outfitters.(Esayas, 2015)

## Cassandra

- Cassandra is a hybrid, non-relational database like Google's BigTable.Cassandra was originally developed by Facebook and later turned into an open source project. It is an ideal database for social networking cloud computing. The main feature of Cassandra is that it is not a database, but a distributed network service composed of a bunch of database nodes. An operation of Cassandra will be copied to other nodes, which means that the efficiency can be greatly improved. The main customers are twitter and Facebook.

## Apache Hadoop

- Apache Hadoop is a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation (Apache, 2018). In other word, Hadoop's biggest feature is to solve the problem of reliable storage and processing of big data (which is big enough to be stored on one computer and cannot be processed in the required time).

- Hadoop is used in a wide range of applications. Yahoo! Inc. launched what they claimed was the world's largest Hadoop production application in 2008. The Yahoo! Search Webmap is a Hadoop application that runs on a Linux cluster with more than 10,000 cores and produced data that was used in every Yahoo! web search query(Yahoo,2008). In 2010, Facebook claimed that they had the largest Hadoop cluster in the world with 21 PB of storage (Hadoop blog, 2010). In June 2012, they announced the data had grown to 100 PB (Facebook, 2012) and later that year they announced that the data was growing by roughly half a PB per day. As of 2013, Hadoop adoption had become widespread: more than half of the Fortune 50 used Hadoop (Eatontown, 2012).

## Big Data analysis

Analysis is integral and one of the most important part of Big Data. Various tools and techniques have been proposed for performing analysis. Here, we will first discuss different kinds of analytics. Then, we will follow it by the tools used for their implementation.

Sivarajah et al. 2017 proposed to divide analytics into 4 types:

- Descriptive analytics - once the business data has arrived the first thing to do is to analyse what is currently happening in the business. Describes current state of situation.
- Inquisitive analytics - testing business hypothesis by probing data
- Predictive analytics - prediction of what kind of events businesses might expect in the future
- Prescriptive analytics - based on the results from predictive analytics, choose an action for enhancing business performance
- Pre-emptive analytics - identifying precautionary steps for the future that will help to avoid undesirable outcomes

Various tools are available for Big Data Analysis. As with Big Data Storage one of the most commonly used is Hadoop MapReduce(Devi, 2018; Import.io, 2018).

## Big Data Coding Languages

There are numerous tools which are used by data scientists to analyse and process data. Such tools include a variety of modern languages, frameworks and data storage technologies which support the handling of big data. Some of such tools are listed below.

### R

- An open-source statistical computing language which grants a diverse range of statistical and visual techniques thereby providing analytical insights from data. Through effective data handling, storage facility and due to supporting vector operations it allows for fast data processing. On top of key features that standard programming languages hold, additionally it holds packages through Comprehensive R Archive Network and supports data mining, data wrangling and machine learning algorithms making it a useful tool for big data handling.

### Python

- Much like R, Python is a popular open-source programming language, which hosts many third-party packages and community contributed modules. Packages such as NumPy, Scikit or Panda allow for scientific computing within Python, by supporting data mining, data pre-processing, computing and modelling. NumPy is recognised as the base package, as it grants the user the widest functionality with multi-dimensional arrays and matrices. Many developers utilise Python due to its simple user interface as it is it effective for quick analysis, and due to its effective integration of spark.

Other object-oriented languages which are growing in popularity due to their ability to handle big data problems effectively include Scala.

## Other Big Data Software Tools

### Apache Spark

- Apache Spark is a fast computational in memory cluster software, through Scala it uses Hadoop to store data. Used as an addition due to its built in API to Java, Scala, Python and R. Due to its 80 level operators it allows for interactive querying. Due to it being supported by Resilient Distributed Data (RDD) Framework it offers faster computation. It delivers faster speed due to its built-in memory processing. It can deliver nearly real time results, but requires significantly more memory (Datamation, 2018)

## Apache Hive

- An open source platform providing functions which enable querying and management of large datasets existing in distributed storage such as Hadoop Distributed File System. Apache Hive enables custom mappers and reducer codes if there is a lack of desired logic.

There are multiple other tools and frameworks that support big data. Some of which are listed below:

- **Apache Pig**
- **Amazon Elastic Compute Cloud**
- **BlinkDB**
- **Tachyon**
- **And others**

## Big Data in Industry:

The notion of big data has directly affected most industry sectors. Ranging from the integration of clinical, public health and behavioural data within the healthcare industry to the adoption of RTM Dx within the Law Enforcement industry. Growing necessity to analyse a vast range of data from the web has meant most industries and organisations have needed to adopt new software and methods to remain competitive. Retail has been an early adopter of big data analytics; the following will discuss the application of big data analytics in retail with primary focus on ASOS, a world leading online apparel retailer.

## Retail:

The development of e-commerce, online purchasing, social media communication and GPS smartphone interactions have led to the increase in data, which requires specific data-driven customisation to extract the quality of data (Brown, 2011). Retailers must maintain customer information; specific demographics have certain purchasing patterns. Customer purchasing patterns affect retailers and how they group items; Agrawal and Srikant proposed a specific data mining technique (Market Basket Analysis) which is used by majority of retailers to group their items. MBS is a modelling technique based on the idea that buyers of certain items are more inclined to purchase another group of items. In recent years retailers within e-commerce use MBS with recommender systems to section and target their audience, achieved by collecting click stream data, observing customer behaviour and real-time recommendations.

## Marketing:

Marketing analytics aid organisations by evaluating their marketing performance, analysing consumer behaviour and analysing customer purchasing patterns

(Bhadani, 2016). Thus, organisations can successfully format their marketing strategies according to marketing trends. Marketing strategies include programmatic advertising, placement of advertisement, dynamic pricing and personalised products (Soares, 2012).

## Product Development:

Product Development holds a risk as such organisations integrate external data sources, such as Social media (Facebook, Instagram, Twitter, etc.), and internal data sources, such as customer relationship management systems to comprehend their target audience when developing a new product. Development requires strict analysis and planning to mitigate potential risk, however if done correctly it can increase customer value or indorse brand engagement (Anastasia, 2015).

## ASOS:

E-commerce organisations are one of the fastest adopters of big data analytics, primarily due to vast competition and the need to play a competitive role within the market (Koirala, 2012). ASOS is one of the leading apparel suppliers, with 18.4 million active customers (ASOS, 2018), ASOS utilises the most up-to-date methods to secure a position against competitors such as H&M, Zara and others. ASOS utilise both structured and unstructured data. Structured data is demographic information about customers, such as age, gender, name etc, compared to unstructured data which is largely based on psychological information, such as hobbies, interests and other personal information. Managing both types of data contains challenges, particularly generating meaningful insights from vast amounts to data to increase conversions. Within e-commerce key big data characteristics are distinctive according to 4 of the 6V's voluminous, variety, velocity and veracity (Akter, 2016).

## Voluminous (quantity of data)

With emergence of web technologies, there has been a staggering increase in the quantity of big data, particularly seen within the e-commerce environment. The quantity of data within e-commerce has the potential to improve organisational decision-making processes. ASOS utilize large volumes of data to analyse customer choices and customer feedback from millions of reviews (Davenport, 2007). However due to the vastness, it is unlikely that such data is clean and free of errors, despite causing challenges it enables for real-time decision making. To effectively use the massive quantity of big data, ASOS utilize Amazon's sophisticated recommendation engine to ensure superior customer satisfaction and dynamic pricing systems, thus ASOS can adjust prices against competitors (Goff, 2012).

## Variety (numerous sources)

Variety is a critical attribute of big data; the severity of variety amongst data from various sources and formats, which contain multidimensional data fields, has meant retailers such as ASOS use analytical models. Analytical models utilize structure, semi-structured and unstructured data such as customer profiles, purchasing history or seasonal buying patterns to optimize the store concept and or promotional campaigns. Additionally, ASOS utilizes sentiment analysis which lets them assess real time responses and adjust accordingly. Sentiment analysis has become a strong role as consumers are increasingly relying on peer sentiment and recommendations (Bhadani, 2016).

## Velocity (frequency of data)

Retailers must understand the velocity of big-data and the importance of prioritizing and syncing it into business processes, decision making and performance improvement (Beulke, 2011). The velocity of data has meant ASOS perform real-time data management, including real-time basket analysis, to maintain a constant flow of new products without comprising delivery dates. Additionally, apparel retailers can track individual customer data, including individual click stream data, to leverage their behavioral analysis. ASOS is increasingly updating its customer data in real-time to maintain strict changes in customer behavior (Brown, 2011).

## Veracity (uncertainty of data)

Uncertainty of data types has become an issue due to the quantity of data which retailers generate from the web, as such it is key to screen out ‘bad’ or ‘irrelevant’ data. High data quality is an essential requirement of big data analytics, hence ASOS utilizes screening technology, particularly on data generated from social media (Facebook generates 30 billion pieces of information every month). Additionally, e-commerce retailers often use data fusion to combine multiple less reliable sources, thus creating efficient data, examples include social media comments attached to geospatial location information (Schroeck, 2012).

## Big Data gathered from social Media

Social media analytics are the synthesis of user behaviour. The availability of data on consumers’ through web browsing, online shopping behaviour, customers’ feedback and marketing research on social networks allow organisations to gain timely and extensive insights into consumers. Therefore, organisations can focus their market intelligence strategies based on different objectives such as advertising and product launches, publicity and brand management; promoting customer loyalty, providing personalised services to customers, keeping a tab on market trends and competitors. Specifically, most social media, such as Facebook, Twitter, Instagram, YouTube, analyse user behaviour from the following points.

Tracking cookies: Many social media websites track its users across the web by using tracking cookies. If a user is logged into account and simultaneously browses other websites, these social media can track the sites they are visiting.

Facial recognition: One of Facebook's latest investments has been in facial recognition and image processing capabilities (Avantika, 2018). Facebook can track its users across the internet and other Facebook profiles with image data provided through user sharing. This is also the future development trend of social media.

Tag suggestions: Most social media suggest who to tag in user photos through image processing and facial recognition.

Analysing the Likes: A recent study conducted showed that it is viable to predict data accurately on a range of personal attributes that are highly sensitive just by analysing a user's Likes. Work conducted by researchers at Cambridge University and Microsoft Research show how the patterns of Likes can very accurately predict your sexual orientation, satisfaction with life, intelligence, emotional stability, religion, alcohol use and drug use, relationship status, age, gender, race, and political views—among many others.

## Types of Big Data used by e-commerce retailers (ASOS)

E-commerce retailers capture various types of data, which have been broadly categorised as transaction/business activity data, click stream data, video data and voice data. The following points will highlight the importance of each data type and how ASOS utilises each data type to track consumer shopping behaviour.

## Transaction/business activity data

Transaction/business activity data appear because of exchanges between customer and companies and appear in the form of structured data. ASOS utilises a predictive modelling technique (collaborative filtering) based on user activity data to produce unique personalised recommendations. This technology is utilised by multiple other e-commerce organisations, many state that collaborative filtering generates up to 30% of their sales (Brown, 2011). Activity data has a direct impact on marketing, ASOS accumulates detailed customer profiles and uses them to personalise advertisement in aims of increasing customer loyalty.

## Click-Stream Data

Click-Stream data originates from user interactions with online advertising, social media or e-commerce businesses. Within recent years, social media and online advertisements hold key importance for management, particularly for strategic decision making. ASOS largely utilises click-stream data when undergoing sentiment analysis, as described above.

## Video data

Video data refers to data gathered from live images. As software develops, organisations like ASOS develop image software analysis which captures video data. Video data enables organisations to enhance their offerings, a famous case of successful use of video data is seen with Netflix. Netflix used their visualisation and demand analytics tool to understand preferences and developed their high grossing program “House of Cards” within the US.

## Voice data

Voice data is data typically generated from phone calls or online customer service. ASOS is using advanced capabilities to analyse customer transcripts from calls and have been using it to form a closer understanding of certain customer profiles.

ASOS is one of thousands of retailers who aim to constantly develop their software to continue to gather any form of competitive advantage over competitors. Big Data analytics is used vigorously, with the growth of e-commerce, organisations must maintain a strict relationship with their customers, by gathering individual customer profiles from various data types; aim to maximise customer value.
ASOS is merely one example of big data analysis put into practice, there are multiple uses big data holds within the real world (Oracle, 2018):

**Predictive Maintenance**

- Mechanical failures can often be hidden within structured or unstructured data. Through analysing these indications organisations have the potential to locate issues before problems arise, saving costs and maximise equipment up-time.

**Fraud and Compliance**

- Big Data lets organisations identify patterns which can be used to identify fraud.

**Machine Learning**

- Big data is one of the key reasons why developers can develop artificial intelligent software. The availability of big data facilitates such development

**Operational Efficiency**

- Big data has meant organisations can analyse and assess production, customer feedback and returns thereby limiting outages or predicting future customer demands. Additionally, big data can be used to improve decision making within the current market demand.

## Drive Innovation

- By studying interdependencies, big data allows organisations to efficiently innovate. By examining trends organisations can improve planning and financial decisions to deliver new products or services, as highlighted with Netflix's success of 'House of Cards'.

## Big Data Problems and Limitation

### Data Privacy

With advancements in Big Data applications and services, technology boycotts will not be sufficient in terms of consumer protection. Traditionally analytics would mine sensitive information, the information is then subject to scrutiny and data discrimination. Organisations in the past used various methods of de-identification (anonymisation/ encryption) to separate data from identities. However, Tene & Polonetsky (2012) found that anonymised data can be "re-identified and linked to the identity holders". To reduce the issue of re-identification, a solution would be to treat data as identifiable entities. However, this may cause firms to refrain from de-identification measures hence, raising security and privacy risks when accessing data.

An exponential growth in data collection, storage and processing will make accessing sensitive data much easier. Although new legislation will be introduced such as the EU GDPR 18, governing authorities must ask themselves who is to enforce the legislation onto companies. The GDPR 18 aims at protecting consumers by giving them back control of personnel data. Penalties for offending companies is a 4% of annual turnover for those who are non-compliant and use the sensitive information of EU citizens for malicious purposes. Businesses must be ethical today and show transparency regarding their intended use of sensitive data, if a company can be transparent then they should seize the given opportunity. Firms should inform consumers about what data they hold and why it is held.

### Data Security

Linking in to privacy is compliance, certain domains like social media and health information can spark fear amongst individuals as the data holding companies may questionably have too much information about them. One of the greatest threats to personal security would be the unregulated accumulation of data by social media companies (Kailser et al., 2013). A greater concern arises from the fact so many people willingly comply to surrendering their information.

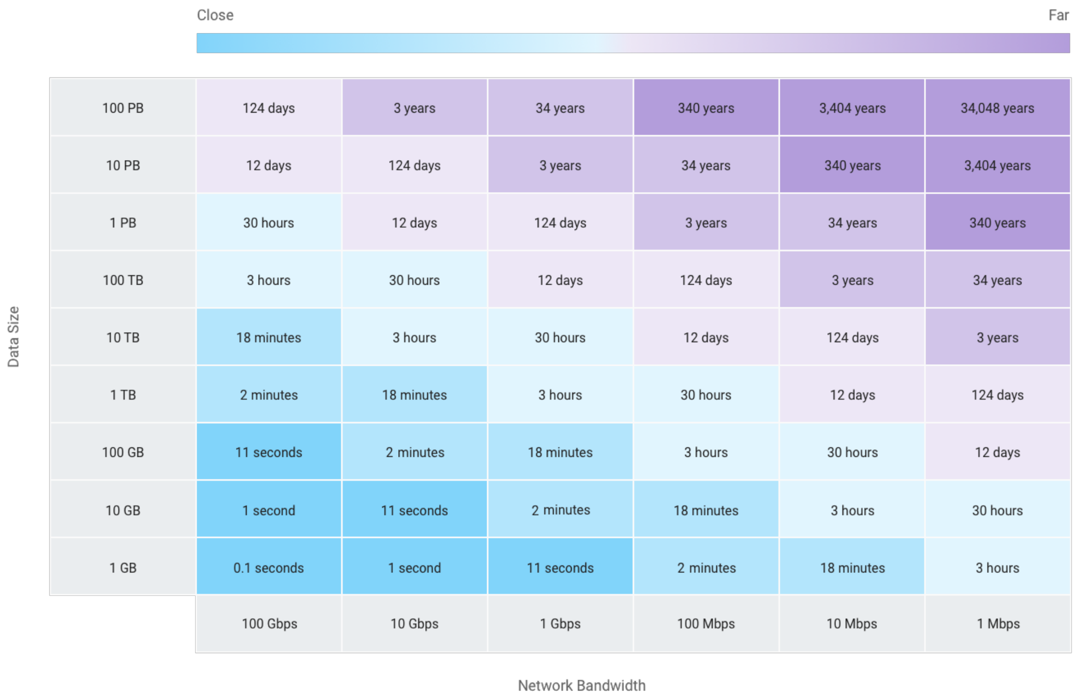
Regarding security, Gandhi et al.(2017) states that big data analyses with colossal data loads are "correlated, analysed and mined for meaningful patterns". Focusing on the issue of safeguarding, Gandhi et al. (2017) highlights that the preservation of sensitive information is a major issue and that information security is also becoming an issue related to big data analytics. The point raised about preserving sensitive information compliments a Data Ownership question adapted from Kailser et al. (2013) which questions the necessity of data storage. A solution to improving Data Security would be to implement various authentication and authorisation practices but also encrypt data (Gandhi et al, 2017). With Big Data covering a wide variety of applications including scale of networks;,differing devices, lack of intrusion systems and real-time security monitoring; there has been an arousal in attention towards information security. Thus, extra attention must be paid to developing rigorous multi-level security models and a system for prevention, for not only Big Data Security but also Information security.

## Data Storage and Transportation

As data volume has been more than doubling on a yearly basis, it is imperative that data handlers and specialists can safely store the sheer incoming data loads. Today new storage mediums are being invented or improved, however simultaneously once a medium has been invented and improved there is an instantaneous boom in data quantity, this is particularly the case as all citizens these days can create data not just professionals.

Current medium technologies are limited to approximately 4TB/ disk. An exabyte requires 25 disks, even if an exabyte could be processed, it could only directly attach the requisite number of disks. (Kailser et al.,2013). This would distort and take priority over communications networks. An example to help explain this would be network with a 1 gigabyte per second speed which only has 80% effective transfer rate and a bandwidth sustainable of 100mb. To transfer an exabyte it would take 2800 hours given that the transfer is consistent and sustainable. Transmitting the data would in fact take longer, a potential remedy to this issue would be to process the data in location in which it is created and transfer the results only. Google (2017) informs potential cloud customers of the variable effects in which data location and bandwidth available has on transfer time and speed. Interestingly Google use factors: data size, bandwidth and intended use to help define close; this representation of close data is depicted in Image 1 as low bandwidths and large data volumes result in a much longer estimated transfer time.

## Image 1- Data Estimation Transfer Time Google (2017)



Close

Far

Data Size

| | | | | | | |
|---|---|---|---|---|---|---|
| 100 PB | 124 days | 3 years | 34 years | 340 years | 3,404 years | 34,048 years |
| 10 PB | 12 days | 124 days | 3 years | 34 years | 340 years | 3,404 years |
| 1 PB | 30 hours | 12 days | 124 days | 3 years | 34 years | 340 years |
| 100 TB | 3 hours | 30 hours | 12 days | 124 days | 3 years | 34 years |
| 10 TB | 18 minutes | 3 hours | 30 hours | 12 days | 124 days | 3 years |
| 1 TB | 2 minutes | 18 minutes | 3 hours | 30 hours | 12 days | 124 days |
| 100 GB | 11 seconds | 2 minutes | 18 minutes | 3 hours | 30 hours | 12 days |
| 10 GB | 1 second | 11 seconds | 2 minutes | 18 minutes | 3 hours | 30 hours |
| 1 GB | 0.1 seconds | 1 second | 11 seconds | 2 minutes | 18 minutes | 3 hours |
| | 100 Gbps | 10 Gbps | 1 Gbps | 100 Mbps | 10 Mbps | 1 Mbps |

Network Bandwidth

Kailser et al. (2013) suggested that data triages are to be implemented and that data only critical to downstream analysis should be used for downstream analysis. A triage will prioritise on imperative data first and then transfer medium and low priority data afterwards. This potentially could speed up Big Data transfers although it would also increase the number of big data transfers.

## Data Discrimination

As data holders (companies and organisations) are holding more and more data regarding individuals, is it still unethical to discriminate based on the data held. In the past and still today credit scores are used to help decide whether an applicant is eligible for a financial loan, similarly the insurance industry is driven by secondary data. With Big Data Permitting businesses to become greater marketeers and service providers, consequently this can be a tool for enhanced data discrimination (Bernard, 2017).

A problem which complements the issue of data discrimination is that many consumers accept the fact that they are analysed and evaluated in search for a supposedly better experience. However, consumers should question whether constant analysis makes it easier for them to access the resources they need (Ramirez et al., 2016). There are legal practices in place such as the “Fair Credit Reporting Act” and Federal Trade Commission Act, although there needs to a strict enforcement so that companies adhere to this legislation. A solution to reduce Data Discrimination would be to ensure that algorithms provide representative samples of consumers, to complement this the companies holding the data should acknowledge and act upon any Data biases. Outcomes of Big Data should then be compared against traditional practices.

## Data Quality

Another issue related to Big Data storage is the quality of the data opposed to the quantity of the data. With the emergence of Big Data there has been a great level of attraction from industrial, academic and governmental organisations; an example is the US government investing $200 million to launch the “Big Data Research Initiative” (Li & Chen, 2012). The ever-increasing development and utilisation of big data has also grown in popularity across the medical, financial and retail fields; but also has become of great social value. This has led to an increase in the rapid acquisition and analysis of big data which has been recognised as a useful tool in predicting and understanding consumer demands, bettering service and forecasting, preventing risks. Although utilising and analysing big data should be an accurate verification and validation of high quality data. Therefore, it is imperative that data holders implement a quality framework to assess and process data.

Looking at the “Big Data 4V characteristics”, the extraction and processing of high-quality data from large varied complex data sets has become an urging issue which requires close attention. Adapting Cai & Lin (2015) main challenges of data quality the four challenges of data quality will be analysed:

1. Data Diversity
2. Data Volume
3. Data Change
4. Lack of consensual data quality standards

## Data Diversity

Looking at data diversity, organisation used to use only primary data which they had generated, however today the primary data analysed by enterprises has now overwhelmed this past trend. Sources of big data are extremely diverse ranging from internet (including mobile) and IoT device datasets (Li & Liu,2013); but also, experimental data (Demchenko, et al.,2013).

Three data types can be relevant to the quality of Big Data: structured, semi-structured data and unstructured data. Unstructured data accounts for over 80% of all existing data (video, audio and documents). Semi-structured data can take the form of software, spreadsheets or reports. With companies gathering big data varying in structure from an array of sources, a major challenge for all data handlers would be integrating the mixed data structures (McGilvray, 2008). Having to process data from a plethora of sources and structures can often lead to conflicting inconsistencies or even cause contradiction amongst data sources (Cai & Zhu, 2015). For small volumes of data, manual search and programming methodologies can be used including Extract; Transform and Load. In the case of

Big Data, manual search and programming efforts are futile especially when processing petabytes and exabytes of data.

## Data Volume

The sheer volume of Big data can pose as a challenge in terms of assessing the quality data within a feasible time. Throughout the 1970s information doubled annually, after 2011 there was 1.8 Zettabytes (ZB) of data. Obviously, the issue of gathering and integrating this data arises to produce high-quality data. As unstructured data accounts for most of data, this will dramatically increase the time needed to convert unstructured data into structured data. All in all, this poses as a perpetual challenge to current data processing methodologies.

## Data Change

Dramatic and instantaneous changes in big data can make the shelf life of certain data short. Companies which are unable to collect the necessary data in real time, or satisfy data requirements over an extended period run the risk of their data being obsolete or even invalid. Analytical and processing attempts on outdated data consequently could lead to inaccurate and misleading conclusions. Currently there are very few software solutions to increasing data shelf life, but only for the processing of data.

## Lack of consensual data quality standards

For physical consumer products, the ISO (International Organisation for Standardisation) published the ISO 9000 to ensure a high quality of products. This publication from ISO is recognised in over 100 countries creating a sense of agreement regarding domestic and international trade. Contrastingly data studies regarding standards started in the 1990s, although no standard was published by the ISO (8000) until 2011(Wang et al., 2010). Cai and Zhu (2015) state over 20 countries are part of the ISO 8000 standard, however not without disputes. With few standards regarding the quality of data and no consensus harmed by international disputes; one can only assume that in the big data era no one solution will be able to solve international data quality discrepancies.

## Data Quality versus Quantity

Another emerging Big Data Challenge is the trade-off between the quality and quantity of data. With users gaining greater access to data there is usually an addictive nature to want more to perhaps be able to explain a relevant phenomenon. In contrast some users of big data may be more interested in the quality of data as opposed to quantity. Although, having access to a large high-quality data arrays help draw a more valuable and accurate conclusion. Kailser et al. (2013,p.996) suggests that the value and richness of data records decreases as volume increases.

When examining the trade-off between data quality and quantity it is important for the user to ask themselves what degree of accuracy do they require? A good example would be when analysing trends, where a high precise database system may not be necessary; although a large-scale processing Big Data environment would be required. To help make a more educated trade-off between data quality and quantity, one can ask themselves:

- What makes data irrelevant and relevant?
- How can I make sure that all data is reliable and precise, or even accurate to an approximated degree?
- What quantity of data will suffice to estimate or predict probabilities and accuracies?
- What is the criteria for assessing valuable data and does quantity benefit me?

## Scalability

Dealing with great quantities of data and instantaneous changes in volume can be problematic for data analysts and holders. Scalability is split into three different aspects: Data volume, Hardware Size and Concurrency. As the volume of data exponentially increases, sufficient hardware is becoming larger also. This can pose a challenge as they must be able to house a feasible analytical system which can support future sets of data, as well as algorithms for expanding datasets (Purohit, 2017). Both a scalable and distributed architecture is required for big data storage, processing and analysis; along with a sophisticated amount of concurrency. Kailser et al.(2013) states that algorithms have a “knee” in which performance will no longer increase with the resources in a linear fashion. This presents a further problem of writing a new algorithm which can increase the relationship plateau of performance and hardware.

## Data Ownership

Another major problem regarding Big Data is Data ownership, this poses as an ongoing challenge especially in the realm of social media. Petabytes of data are existent on the servers of giants such as Facebook, MySpace and Twitter; although the social media giants do not own this data but house it. The data held by the social media giants is often contested due to the location of where the data is held, creating a sense of dichotomy which can only be settled in court. Kaisler et al.(2012.) recognised this in the context of computing.

When focusing on data ownership, ownership itself becomes a small part of responsibility to promote accuracy. At an individual level this may not be so significant, however at an organisational level this should of be great significance and a top priority. Although again there is the issue of enforcement (Kailser et al., 2013). User agreements will not be sufficient in terms of enforcing data laws and legislation as no social media purveyor can feasibly check all data items on their own servers.

Applying the issue of Big Data ownership to a real-world contextual application would be related to the implementation of IoT devices. Darrow (2016) states an examplar scenario where a city's governing authority were to hire tech providers to gather data from sensors used in public transport. If the city council has hired a tech company to gather data regarding sensors of transport and road statuses, it can become unclear who has rightful ownership of this data as the council may have to pay in order to access and retrieve the data from the tech provider regarding their own services. These are all sources of rich data types.

The launch of several popular social media websites has spurred a great trend in Big Data analytics regarding the integration of first hand verified data with third party and public external data. This external data however has not gone through rigorous validation and verification. Unverified data can potentially lead to a compromise a dataset's fidelity; but also introduce irrelevant entities and lead to incorrect linkages. A consequence of this would be a negative effect on data processing conclusion drawn from mixed data including that of a verified and unverified types.

When questioning the ownership of Big Data, it is important to consider these questions adapted from Kailser et al. (2013):

1. When is the expected expiry of publicly accessible data?
2. After data has expired regarding validity, should it be erased from public data sets or websites?
3. Which location and methodology should be used to archive expired data? Does the data even need to be archived?
4. Who is to be held accountable for the data accuracy and fidelity?

## Future research directions

Future directions of big data could be discussed from three layers and four aspects. According Anagnostopoulos et al, future further research should address challenges from different computer science areas and those challenges could be classified by using bottom-up layers approach according to "get", "save" and "analyse" which are basic data handling process (Anagnostopoulos et al., 2016).

The challenges of every layers are shown in the Fig. 1. To discuss future research from widely aspects which include not only technical aspects but also other aspects, research directions will come from four aspects during the whole data processing procedure.

1.When data is first acquired, the cleaning procedure is needed to keep useful part. In this layer there are four main points,

2. How to clean data effectively and the life cycle of data are one of the future research questions.

3. When it comes to the "save" layer, limited data storage and how to transfer and sharing information among different data warehouses are another future research direction. In the last layer - "analyse", we face the problem that how to analyse data more effectively with the increasing data information need to be analysed in the future. Therefore, algorithms and scalability software platforms are needed to be developed in regarding the purpose of timeless of data collection and analysis.

4.The fourth challenge is ethical issues including: intellectual properties and personal information privacy. These could appear in all layers. This challenge will become more and more important in the future big data environment. Detailed further research directions about four aspects mentioned above are mentioned below.

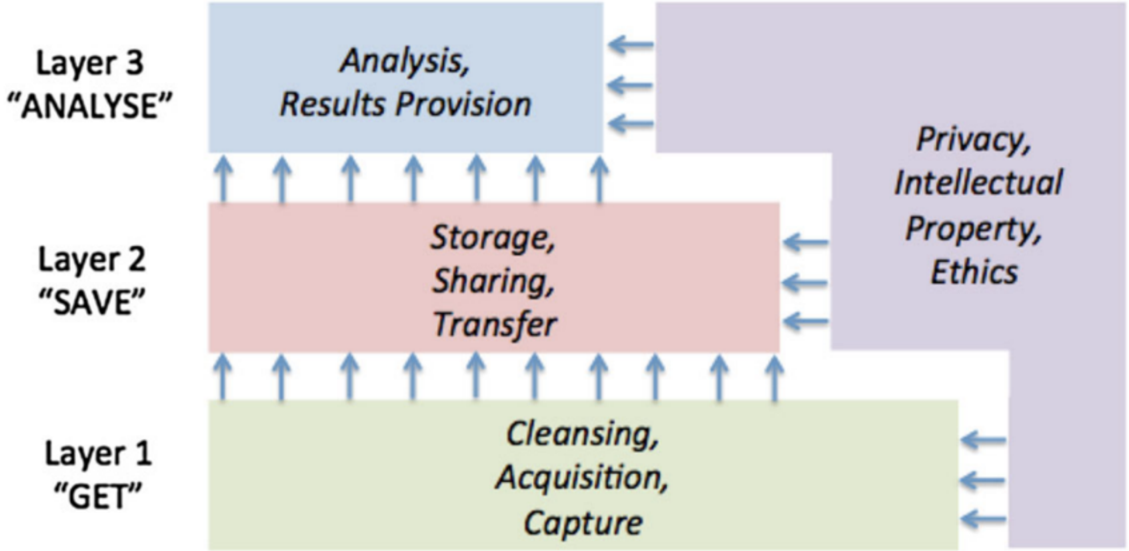


Fig.1(Anagnostopoulos et al., 2016)

## Effective data cleaning and life cycle of data

The first step of data processing is data capture, this time requires data to be cleaned as the data needs to fulfil some quality criteria so that can be used effectively. After cleaning, data could be reused, therefore filtering capable mechanisms need to be developed in terms of facilitating reproducibility (Anagnostopoulos et al., 2016) . At the same time one must consider how to define life cycle of data so that data can be used more effectively (Chen et al., 2014). However, most of the time data does not show clear procedures and may produce misleading conclusions. Problem solutions also need to be noticed during the “get” data process (Bhadani et al., 2016). In addition, latency issues appeared in the process of data capture, this also deserves further research including latency minimisation especially for time sensitive applications.

## Storage and dynamic data placement

The contradiction between limited data storage and increasing data information caused an increasing demand for storing data and information. Many efforts have been made towards this. For example, Google introduced MapReduce to deal with the massive amounts of data and cloud platforms are also developed for storing a large quantity of data (Dean et al., 2008). However, with the exponential amounts of increasing datasets and mobile devices developed, the demand of data storage will be continuously rising in the future. Purchasing and using available storage devices meet this demand (Khan et al., 2014). One of future research questions will be how to process highly diverse data capture with minimum latency; in addition the wide use of mobile devices like smartphones regarding storage and processing of unremarkable data of media resources demands that smartphones require deeply research in the future.

To effectively transfer across data warehouses could be a significant future research direction as inconsistent data from many different sources is a significant challenge in the future. Hence there is a need for further research about dynamic data placement. In this way, data could be placed dynamically in delivery networks, and big data can arrive at its destination without consuming any extra resources.

## Develop algorithms and Scalability software platforms

Future data analysis should be more effective and timeless. Therefore developing algorithms and more scalable software platforms are future research directions in terms of future big data. Combining the strengths of algorithms and analysing big data to design new scalable software platforms.

EC algorithms have to handle complex engineering and design issues, and machine learning algorithms meet basic demands during the processing of big data (Cheng et al., 2016). However, it cannot replace the domain requirements and algorithms will be needed to gain insights from the desired discipline.

There is an increasing demand for scalable software platforms for most companies. As data analysis are closely relevant to their profits, they pay most attention on how they analyse complex data from diverse resources. Scalable software platforms can help them analyse large amounts of complexed data among diverse data warehouses and datasets so data can be processed at a faster rate.

## Ethical considerations

Intellectual property and related issues could be the first ethical consideration for big data. Nowadays we have already developed the intellectual property license to deal with this problem. However, the role of such licenses  is to inform copyright holders and those looking to produce similar work. Property rights individual negotiation provides a framework for different licenses from multiple resources beyond big data environment, so that proof of ownership cannot be easily imitated.

More and more data will be transformed in the future as the development of cloud computing and internet of things (IOT), this provides a great opportunity to access data but causes an increasing comprise to personal information. The trade-off between convenience and security are the main problem most researchers have difficulty to struggle, and therefore should be a future ethical consideration regarding Big Data holders.

### Conclusion

Big Data is inevitably becoming increasingly important for businesses across the majority of market sectors. Huge companies such as Google, Facebook, LinkedIn are built around Big Data. There are studies which estimate that incorporating big data in a retail company could boost operating margin by 60%. Incorporating some sort of data handling mechanism will become a routine for businesses if they are willing to reap the benefits of insights that could be provided by data. For large businesses, it might be more sensible to create own Big Data infrastructure and hire data specialists. As for small businesses: data giants, such as Google, Amazon, provide their services for processing, storing, analysing Big Data.

In this report, we have set out to investigate what Big Data is, how is it processed, its business applications in retail and partly social media.

Various aspects of integrating Big Data into data-oriented processes have been outlined. Existing Big Data usage in industries has been overviewed. One of the largest retail companies, ASOS, has been analysed in terms of Big Data usage in order to provide the reader with an actual example of Big Data integration. We have given an overview of processes involved in Big Data analytics and various tools used for Big Data processing. Various options for Big Data storage were suggested. A synopsis of languages used to handle Big Data has been provided. Issues arising with an incorporation of Big Data into industrial processes were discussed.

## Finished references

- Akter, W. (2016). Big data analytics in E-commerce: a systematic review and agenda for future research. *Electronic Markets*, 173-194.

- Anagnostopoulos, I., Zeadally, S. and Exposito, E., 2016. Handling big data: research challenges and future directions. *The Journal of Supercomputing*, *72*(4), pp.1494-1516.

- Anastasia. (2015). *Big data and new product development*. *Entrepreneurial Insights*. Retrieved from http://www.entrepreneurial-insights.com/big-data-new-product-development/
- Andrew Ryan(2012) *Under the Hood: Hadoop Distributed Filesystem reliability with Namenode and Avatarnode*. [online] Available at: https://www.facebook.com/notes/facebook-engineering/under-the-hood-hadoop-distributed-filesystem-reliability-with-namenode-and-avata/10150888759153920 [Accessed: 15 Nov. 2018]

- Avantika(2018)How Facebook is Using Big Data - The Good, the Bad, and the Ugly. [online] Available at: https://www.simplilearn.com/how-facebook-is-using-big-data-article [Accessed: 15 Nov. 2018]

- Apache(2018)Apache Hadoop. Available at: http://hadoop.apache.org [Accessed: 15 Nov. 2018]
- ASOS. (2018). *About us*. Retrieved from https://www.asos.com
- Bhadani, A.K. and Jothimani, D., 2016. Big Data: Challenges, Opportunities, and Realities. In *Effective Big Data Management and Opportunities for Implementation* (pp. 1-24). IGI Global.

- Bello-Orgaz, G., Jung, J.J. and Camacho, D., 2016. Social big data: Recent achievements and new challenges. Information Fusion, 28, pp. 45-59.

- Bernard, M. (2017). *3 Massive Big Data Problems Everyone Should Know About*. Available: https://www.forbes.com/sites/bernardmarr/2017/06/15/3-massive-big-data-problems-everyone-should-know-about/#11a686506186 . Last accessed 15th Nov 2018.

- Beulke. (2011). *Big Data Impacts Data Management: The 5 Vs of big data. .* Retrieved from http://davebeulke.com/big-data-impacts-data-management-the-five-vs-of-big-data

- Bhadani, J. (2016). *Big Data: Challenges, Opportunities and Realities.* Pennsylvania: Management and Opportunities for Implementation.

- Brown, C. M. (2011, October). *Are you ready for the era of big.* Retrieved from Mckinsey: http://www.mckinsey.com/insights/strategy

- Cai, L. & Zhu, Y.. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal.* 14 (2), p1-10. DOI: http://dx.doi.org/10.5334/dsj-2015-002

- Chen, M., Mao, S., Liu, Y., 2014. Big data: A survey. Mobile Networks and Applications 19 (2), 171-209.

- Cheng, S., Liu, B., Shi, Y., Jin, Y. and Li, B., 2016, June. Evolutionary computation and big data: key challenges and future directions. In *International Conference on Data Mining and Big Data* (pp. 3-14). Springer, Cham.
-
- Darrow, B. (2016). *The Question of Who Owns the Data Is About to Get a Lot Trickier.* Available: http://fortune.com/2016/04/06/who-owns-the-data/ . Last accessed 15th Nov 2018.

- Davenport, H. (2007). The dark side of customer analytics. . *Harvard Business Review.*

- Dean, J. and Ghemawat, S., 2008. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), pp.107-113.
-
- Demchenko, Y., Grosso, P., de Laat, C., et al. (2013) Addressing Big Data Issues in Scientific Data Infrastructure. Procedures of the 2013 International Conference on Collaboration Technologies and Systems, California: ACM, pp 48-55.

- Dbms2.com.(2009)How 30+ enterprises are using Hadoop [online] Available at: http://www.dbms2.com/2009/10/10/enterprises-using-hadoop/ . [Accessed: 15 Nov. 2018]

- Esayas,Aloto (2015). Who's Using MongoDB and Why? [online] Available at: https://www.datavail.com/blog/whos-using-mongodb-and-why/ [Accessed: 15 Nov. 2018]

- Gandhi, R.V., Rathan Kumar, CH. & Vamshi Krishna,P.. (2017). BIG DATA: ISSUES AND CHALLENGES. *International Journal of Software & Hardware Research in Engineering*. 5 (7), p1-5.

- Goff, M. S. (2012). Need for speed: Algorithmic marketing and customer data overload. *McKinsey Quarterly*.

- Khan, N., Yaqoob, I., Hashem, I.A.T. et al., 2014. Big Data: Survey, Technologies, Opportunities, and Challenges. The Scientific World Journal, vol. 2014, Article ID 712826, 18 pages.

- Li, G. J., & Chen, X. Q. (2012) Research Status and Scientific Thinking of Big Data. Bulletin of Chinese Academy of Sciences 27(06), p 648–657.

- Li, J. Z., & Liu, X. M. (2013) An Important Aspect of Big Data: Data Usability. Journal of Computer Research and Development 50(6), pp 1147–1162.

- McGilvray, D. (2008) Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information, California: Morgan Kaufmann.

- Kaisler, S., W. Money, and S. J. Cohen. (2012). “A Decision Framework for Cloud Computing”, 45 th Hawaii International Conference on System Sciences, Grand Wailea, Maui, HI, Jan 4-7, 2012

- Kailser,S., Amour.F, Espinosa, J.A., & Money, W.(2013). Big Data: Issues and Challenges Moving Forward , IEEE Xplore. March 18 2013, Available: DOI https://doi.org/10.1109/HICSS.2013.645

- Koirala. (2012). *What is Big Data Analytics and its Application in E-Commerce?* . Retrieved from www.venturecity.com

- Oracle. (2018). *What is Big Data?* Retrieved from Oracle: https://www.oracle.com/uk/big-data/guide/what-is-big-data.html?fbclid=IwAR2fYsKGv218JOQAhDa35jsGfMXTL2wFhiYCr8OuY2ujZmYSW77lH8W7ttM

- Oracle. (2018). *Oracle Big Data*. Available: https://www.oracle.com/uk/big-data/guide/what-is-big-data.html. Last accessed 15th Nov 2018.

- Purohit, R. (2017). Issues and Challenges in Convergence of Big Data, Cloud and Data Science. *International Journal of Computer Applications (0975 – 8887)*. 160 (9), p7-12.

- Ramirez, E., Brill, J., Ohlhausen, M.K. & McSweeny, T.. (2016). *Big Data A Tool for Inclusion or Exclusion?*. Available: https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf. Last accessed 15th Nov 2018.

- Rowena(2014)Impact of Data and Analytics on Social Media in 2018. Available at: https://towardsdatascience.com/impact-of-data-and-analytics-on-social-media-in-2018-595a3bd4fb60 [Accessed: 15 Nov. 2018]

- Schroeck, S. S.-M. (2012). Analytics: The real-world use of big data. . *IBM Institute for Business Value*.

- Soares, S. (2012). *Big Data Governance: An Emerging Imperative*.

- Smith, M., Szongott, C., Henne, B. and Von Voigt, G., 2012, June. Big data privacy issues in public social media. In Digital Ecosystems Technologies (DEST), 2012 6th IEEE International Conference on (pp. 1-6). IEEE.

- Tene, O. & Polonetsky, J., (2012.) Privacy in the Age of big data

- Todd Clarke. hootsuite (2018)24+ Instagram Statistics That Matter to Marketers in 2019. [online] Available at: https://blog.hootsuite.com/instagram-statistics / [Accessed: 15 Nov. 2018]

- Tsou, M.H., 2015. Research challenges and opportunities in mapping social media and Big Data. Cartography and Geographic Information Science, 42(sup1), pp.70-74.

- Wang, J. L., Li, H., & Wang, Q. (2010) Research on ISO 8000 Series Standards for Data Quality. Standard Science 12, pp 44–46.

- Zikopoulos, P. and Eaton, C., 2011. Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media.

Individual Report

I have a lot of gains in this database group assignment. I will explain both the knowledge and the group experience. There are five people in our group. Part of what I write is the practical use of big data technology, with a focus on social media. Regarding this issue, I have divided the following structure.

First, this section explains the concept of social media. According to several typical examples, such as facebook, instagram, etc., then the main techniques used by several mainstream websites are explained. At present, the main information I searched for is mysql and some important NoSQL products. For example, Facebook and YouTube use Hadoop, Cassandra, MongoDB and other non-relational databases in addition to mysql. The use of these tools can greatly improve efficiency, because in a relational database, if users want to add a new attribute to the existing data, all the files from the data table to the application data layer need to be modified. Although some tools can automatically modify this content, this is still a complicated task, especially when updating the product online database. Some databases, such as MongoDB, do not have a Schema, so the user does not need to change the entire database, only need to make the necessary changes in the application. NoSQL databases contain many kinds. Key-value class: for example redis, riak. Document class include mongoDB, couchDB. Column-family class include HBase, Cassandra.The key-value class is also a key-value relationship, and the value is obtained according to a key; the document one is based on the classification of the document, puts the data of one document together, and queries according to the document ID; the column-family database is a famous three from Google. One of the papers derived from Google File System is a query by combining rows and columns into a key.In the data search phase, NoSQL database is found to be more and more popular in ultra-large-scale data storage, so there will be a more specific introduction at this point.

After analysing the big data collection techniques, the next step is to study the source of big data. In other words, what resources are collected by big data, and what behaviours of social media users are collected and analysed. In this part, according to scholars' research, it is mainly divided into four aspects.Tracking cookies,Facial recognition,Tag suggestions,Analysing the Likes. These data can be used to accurately predict a user's sexual orientation, intelligence, mood, religion, relationship status, age, gender, ethnicity, and other political perspectives.Finally, through the picture social network instagram to describe how big data is used to predict human behaviour.

When I was sorting out the network materials, I had a new understanding of the database. The background of NoSQL is probably because the computational cost of traditional relational databases has become very large in some analysis of data sets as the amount of data has expanded. NoSQL is strong with its "extensibility, big data, high availability, high performance, flexibility". Facebook's messaging applications, including Messages, Chats, Emails, and SMS systems, use HBase.The current popular databases are various. Many softwares are actually not competitive, but they make up for each other. It doesn't make sense to leave a specific business scenario to discuss good or bad. There is no best one, only the most appropriate one, and no one can guarantee which one will be completely replaced. Most of the databases are getting better and constantly improving themselves. For example, MySQL is constantly improving its JSON and geolocation processing capabilities, group replication development, etc.; MongoDB is enhancing the join function, providing more complex query capabilities; Redis also added geolocation processing capabilities. All of these software constitute a powerful database resource that allows computers to more fully collect, organise, retrieve, and analyse data.

Because we emphasise the theme of big data, we have reduced the content of social media and decided to expand the content of NoSQL. With Facebook as a case, we talked about the huge impact of data storage capacity enhancement and introduced the application of Canssandra. Deriving the problem areas that NoSQL will solve. At this point our team members discussed it.

Reasonable design is the key to a project. A reasonable division of labor is also a guarantee of efficiency. The structural design of our paper has been roughly completed in the first group meeting, and later only the details have been revised. Stephan played a very important leadership role in this step. Five members of the group, everyone should put forward a question, in addition, Artsiom also wrote the introduction, and summarised the whole article. Paul modified the format of paper. They have completed their work very satisfactorily.

From the perspective of member cooperation, this is the first time I have discussed problems with my classmates from different countries to learn and progress together. When I completed my part, I misunderstood a part of big data and social media, which added a lot of burdens and troubles to the team members. Another thing worth reflecting is that I can't manage time well, and it always takes a lot of time to complete my part of the task. In the next group work, I will seriously review the questions and communicate with the team members in time. Solve the problem with the utmost effort.