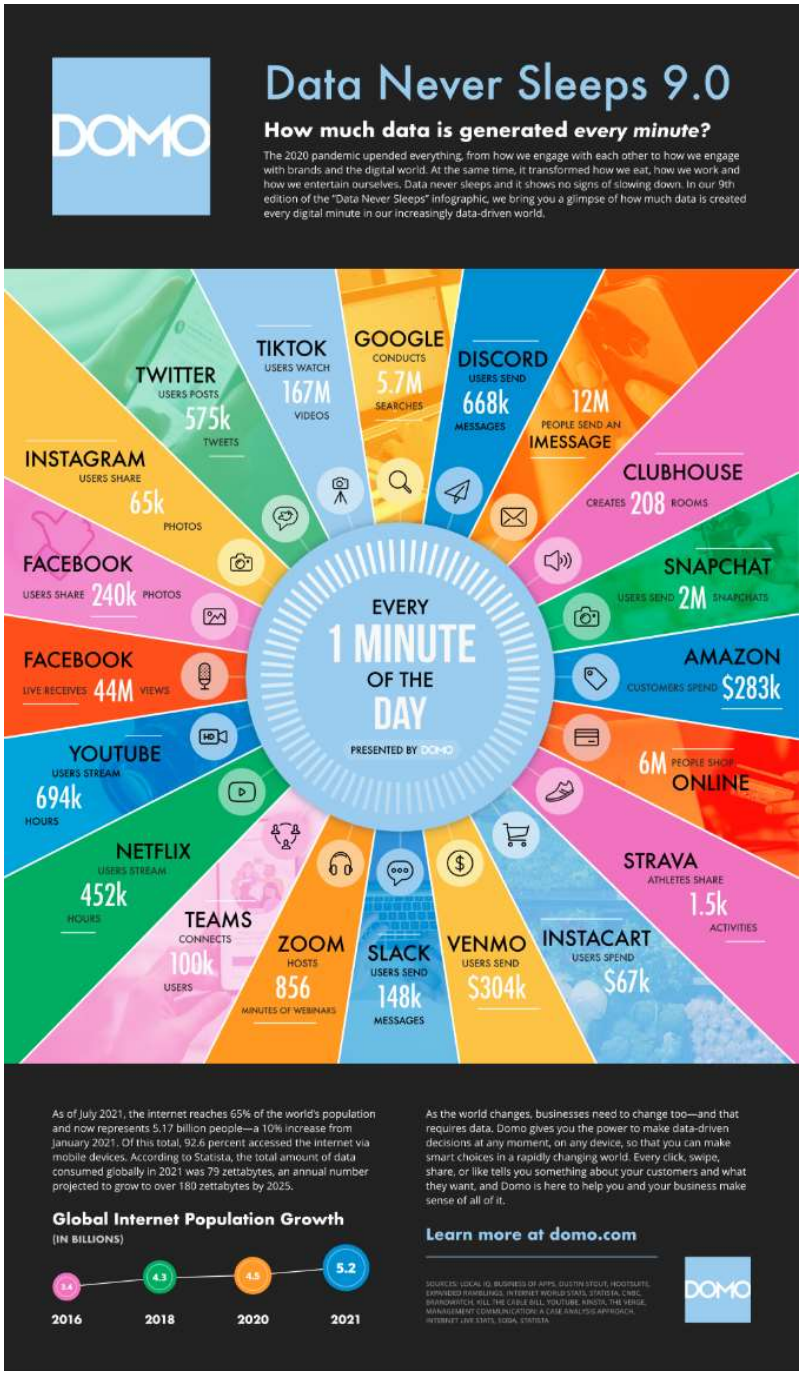
1.1 Introduction to the world of Analytics (and why Big Data is such a big thing)

Data analytics is not a new thing. People have been doing analytics since there is data, and methods that help people to describe, understand and use the data. Remember your first statistics classes some time ago? Remember the time you try to calculate the average and standard deviation of your records? Remember the analysis sport journalists compile to critic how well your favourite football team had performed? Likely we are already using or benefiting from analytics without consciously thinking about it.

Big Data Analytics brings the world of analytics to the next level. As the world becomes more and more connected, one can imagine the amount of data to be collected will continue to increase. Take a look at the figure below about how much data is generated every minute:



Source: Data never sleeps (James, J, 2021)

The figures mentioned in the image above should tell us the amount of data available today is way larger than one could imagine. In fact, the technology and approaches for the processing and analysis of such amount of data will have to be able to meet the requirements based on the current needs. Other than the amount, the technology also needs to manage the timeliness of the processing, and efficiency of the approaches.

It is in such situation that the past approaches begin to show their limitations. These approaches, defined as conventional analytics approaches, will require longer period of time to process and may also be incapable to deal with such large amount of data. Big data technology and solutions emerged in the last decade to answer these challenges and attempt to overcome the limitations.

1.1.1 What is Data Analytics?

Watch the following 9-minute video on the introduction to Data Analytics.

What is Data Analytics? - An Introduction (Full Guide)

https://youtu.be/yZvFH7B6gKI

Source: (CareerFoundry, 2021)

1.1.2 What is Big Data

Watch the following video of 1 minute and 50 seconds to learn the definition of Big Data.

What is Big Data?

https://youtu.be/eVSfJhssXUA

Source: (World Economic Forum, 2016)

1.1.3 Big Data - Definition and Overview

So what is your take on Data Analytics vs Big Data Analytics? Before you jump straight to Slide 1.1.5, perhaps you can take a short break to compare the two. Can you tell the differences between them now?

In summary, Big Data is not just a fancy word or a marketing jargon. It is in fact timely to promote and use Big Data now - Today, we will have more data than before, and the conventional data analytics technology and tools are no longer capable to process and analyse the large amount of data we may possess.

In general, we can define Big Data (or to be more precise, the characteristics of Big Data) using the 5Vs of Big Data: **Volume**, **Velocity**, **Variety**, **Veracity** and **Value**.

Volume

Volume refers to the sheer amount data that no longer fit on a single machine. This is also the reason why there is a need of specialized tools and frameworks to handle this large volume of data.

The reasons (culprits) behind this large amount of data are the lowering costs and increasing availability of sensors and data sources. Also, the lowering costs of data storage also contributes to the ever-increasing volume of data.

There is no fixed threshold how big should the volume be to qualify as big data. Typically, the term big data is used for massive scale data that is difficult to store, manage and process using traditional databases and data processing architectures.

Velocity

Velocity refers to how fast the data is generated Data can arrive at a data storage at very high velocities. Examples of such data includes social media data and sensor data. High velocity of data results in the volume of data accumulated to become very large in short span of time.

Furthermore, some applications have strict deadlines for data to be analysed. Hence, the data needs to be analysed in real-time. Specialised tools are required to ingest such high velocity data into the big data infrastructure and analyse the data in real-time.

Variety

Variety refers to the forms of the data. Big data (has the expectation that) comes in different forms such as structured, unstructured and semi-structured. Some examples of the variety of data include text data, audio, image, video, sensor data etc. Big Data Analytics systems need to be flexible enough to handle such variety of data.

Veracity

Veracity refers to how accurate is the data. This characteristic will be even more important when you have huge amount of data to be collected and processed. To extract value from the data, data needs to be cleaned to remove noise. Benefits of big data can be obtained from data-driven applications. Only when the data is meaningful and accurate, it can lead to useful and meaningful outcome. If there is any incorrect data, noise, or faulty data, the quality of the

outcome will be affected. Hence, the appropriate filtering and cleaning of data are absolutely important.

Value

Value of data refers to the usefulness of data for the intended purpose. Actually, the end goal of Big Data Analytics is about extracting value from the data. Not all data is useful, let alone if they can help bring value. Value of the data is also related to the veracity or accuracy of data. For some applications, value also depends on how fast one can process the data. Value is lost when insights are found too late too.

1.1.4 Problems that Big Data solves

So now you know what Big Data analytics is, and why we need it. Very good!

Can you think of what problems could Big Data be best used to solve? What situations or use cases will require one to seriously consider the deployment of Big Data technology? Please read the following resources and reflect how Big Data technology can be really useful and perhaps essential for modern businesses around us today.

Read the following articles to learn further:

Key readings

1. Google Cloud. (n.d.). What is big data? https://cloud.google.com/learn/what-is-big-data

2. Oracle. (n.d.). Top big data analytics use cases. https://www.oracle.com/my/a/ocom/docs/top-22-use-cases-for-big-data.pdf

3. ProjectPro. (2024). 6 big data use cases- How companies use big data? https://www.projectpro.io/article/5-big-data-use-cases-how-companies-use-big-data/155

4. Walmart Staff. (2017). 5 ways Walmart uses big data to help customers. Walmart. https://corporate.walmart.com/news/2017/08/07/5-ways-walmart-uses-big-data-to-help-custo\rs

1.1.5 Differences between Data Science, Data Analytics and Big Data

Watch the following 8-minute 41-second video to learn the differences between Data Science, Data Analytics and Big Data.

Data Science v/s Data Analytics v/s Big Data - What's the difference? | Whizlabs

https://youtu.be/Ceef30oLPvg

Source: (Whizlaps, 2022)

1.1.7 Big Data Technologies

In this slide, we will introduce you some of the more popular Big Data Technologies and its companies. Use the link provided to read up more about them. The technologies below are by no means exhaustive, but they should have covered the most popular or common ones used in the industry in the last decade.

Big Data Framework or Software Systems

Apache Hadoop

Apache Hadoop is a software library framework that enables processing of large data sets through distributed computing approach. It allows the use of simple programming models to perform the processing across clusters of computers. It provides high scalability using one or more (up to thousands) machines that will operate under the Hadoop system as a cluster, which as a result deliver a data analytics platform that delivers high-availability and reliable services. Key software in the Apache Hadoop system includes Hadoop Distributed File System (HDFS), MapReduce and YARN.

Watch the following 1-minute 53-second video to learn what is Hadoop.

What is Hadoop? An Introduction

https://youtu.be/gYQuu-gvmFU

Source: (Eye on Tech, 2019)

NoSQL

Traditional Relational Database Management Systems (RDBMS) are not designed to handle data with high volume and velocity. Hence, newer technology has been created to cater for this need. NoSQL (a.k.a. Not Only SQL) is one of the database technology developed in the late 2000s to handle Big Data problems. Different from RDBMS, NoSQL databases allows storage of huge amounts of unstructured data, such as data from IoT sensors, social media etc. Some of the benefits of using NoSQL over RDBMS including flexible schemes, its ability to scale using distributed computing, high performance for specific data models and access patterns as well as highly functional APIs and data types.

Watch the following 1-minute 6-second video to learn what is a NoSQL database.

What Is a NoSQL Database? | NoSQL Explained

https://youtu.be/f8t3Hh1RxVA

Source: (IBM Developer, 2020)

Apache Spark

Apache Spark was developed by researchers from University of California, Berkeley in 2010s and has gathered much interests and grew into a popular Big Data analytics engine today. It is an open-source analytics engine for large-scale data processing. It has gotten popular as many see Spark and its resilient distributed dataset (RDD) concept as an alternative to MapReduce, as the latter has faced limitation in big data processing. Spark can run on its own or on top of Apache HDFS, making it versatile and popular among Big Data practitioners.

Watch the following 2-minute 11-second video to learn what is Apache Spark.

What is Apache Spark?

https://youtu.be/p8FGC49N-zM

Source: What is Apache Spark, 2 minutes introduction (Databricks 2018)

Cloud-based Big Data Technologies

Amazon Redshift

For industrial level of big data storage and management, the supporting database service has to be robust and capable to handle the high volume and large scale processing. Amazon Redshift is a fully managed, scalable cloud data warehouse service offered by Amazon Web Services (AWS). Users will be able to implement real-time insights and predictive analytics on petabyte-scale data sets while not needing to own large number of on-premise hardware.

Watch the following 22-minute 38-second video to learn about Amazon Redshift.

Amazon Redshift Tutorial | Amazon Redshift Architecture | AWS Tutorial For Beginners | Simplilearn

https://youtu.be/7bfOllAyxlg

Source: (Simplilearn, 2020)

Google BigQuery

Google BigQuery is the large scale cloud data warehouse offered by Google. It also delivers capability to run analytics at scale and allow real-time insights and analytics. It supports querying using SQL language and this makes it easier for users who are already familiar with SQL to begin using BigQuery. One can even perform machine learning or visualisation of data via another product Data Studio.

Watch the following 4-minute 39-second video to learn what is BigQuery.

What is BigQuery?

https://youtu.be/d3MDxC_iuaw

Source: (Google Cloud Tech, 2020)

Snowflake

Snowflake is a cloud-based data storage and analytics service. It is capable of handling structured, semistructured and unstructured data in one single location. Being a cloud-based solution, users can enjoy managed data warehouse service focusing only on the storage and processing of data, and leave the hardware and its administration to the service provider.

Watch the following 8-minute 22-second video to learn what is Snowflake.

What is Snowflake? 8 Minute Demo

https://youtu.be/g9L5tM6d7vI

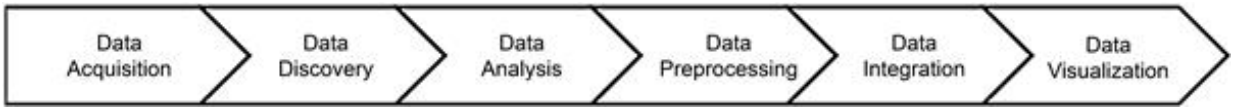
Source: (Snowflake Inc., 2021)

References

1. Apache Software Foundation. (n.d.). Apache Hadoop. https://hadoop.apache.org/
2. Apache Software Foundation. (n.d.). HDFS architecture. Apache Hadoop. https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html
3. Apache Software Foundation. (n.d.). MapReduce tutorial. Apache Hadoop. https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html#Overview
4. Apache Software Foundation. (n.d.). Apache Hadoop YARN. Apache Hadoop. https://hadoop.apache.org/docs/stable/hadoop-yarn/hadoop-yarn-site/YARN.html
5. MongoDB. (n.d.). What is NoSQL? https://www.mongodb.com/nosql-explained
6. AWS. (n.d.). What is NoSQL? https://aws.amazon.com/nosql/
7. Amazon Web Services. (n.d.). Databases on AWS: The right tool for the right job [Video]. https://youtu.be/WE8N5BU5MeI?si=TQkdyduuBIQghBJG
8. AWS. (n.d.). What is Apache Spark? https://aws.amazon.com/big-data/what-is-spark/
9. Apache Spark. (n.d.). Unified engine for large-scale data analytics. https://spark.apache.org/
10. AWS. (n.d.). Amazon Redshift. https://aws.amazon.com/redshift/
11. Google Cloud. (n.d.). BigQuery: From data warehouse to a unified, AI-ready data platform. https://cloud.google.com/bigquery
12. Snowflake. (n.d.). https://www.snowflake.com/en/

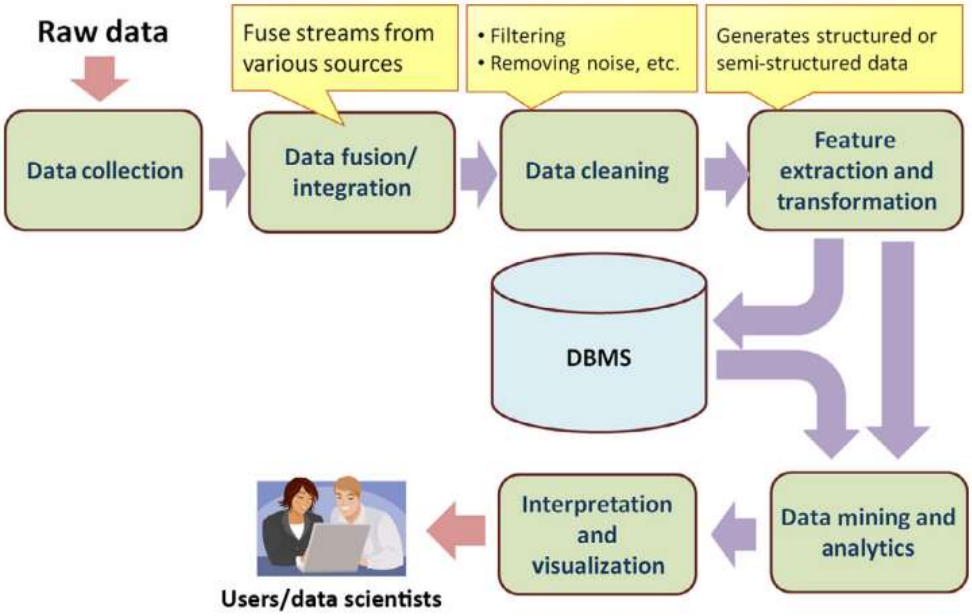
1.2 Big Data Processing

In this section, we will look into the technologies and methods developed and used for big data processing. The processing of big data is an important part in Big Data Analytics. Generally, Big Data processing involves multiple steps from the beginning of the analytics workflow until the final outcome. The following image depicts the common steps found in Big Data processing:



Source: The common steps in Big Data processing (Krishnan, 2013)

The above workflow showed how the journey of Big Data analytics may look like from acquisition until visualisation. In different literatures, you should see similar workflow like the above. For comparison, another resource depicts Big Data processing as follow:



Source: The flow of big data processing (Mehdipour, Noori & Javadi, 2016)

In general, we can conclude the following steps/phases in Big Data processing:

Data Acquisition

This is the first step in Big Data processing. It can be understood as the process of gathering and collecting of data. Acquisition may focus on only interested data, but in some situations, one may collect all available data even though only a certain selection of the data is of interests or needed. Normally, data collected in this phase are considered as raw data, but it is also possible to acquire data that has been processed.

For Big Data Analytics, data can be in both structured and unstructured formats.

Data Preparation

The data gathered from the data acquisition phase will very likely not be directly and immediately usable. One needs to perform one or more of the following processing to prepare the data for analysis and further purposes. Common data preparation techniques are like filtering, manipulation, conversion or even cleaning. Some data will have missing values, or noise. All these should be taken care of before one enters the next phase.

Data Analysis

This phase is what all have been waiting for - the analysis. Data that has been prepared through necessary processes can now be used for analysis. One may perform different kind of tasks such as data mining, machine learning, semantic analysis, data discovery and information extraction. One can expect this phase to produce results that lead the user to the "answers" he or she may ask before starting the Big Data Processing. However, answers should be arrived at in the next and final phase.

Data Interpretation and Visualisation

In final phase, one will want to make use of the outcome produced during the data analysis phase to understand and find the answers he or she seeks. The results could be a prediction based on the machine learning model, or information discovered through the data mining process. Have these outcome is long from completion of one's Big Data processing journey. A user will want to understand and interpret these results to find what he or she are interested in finding. Often, the interpretation will also be followed by the visualisation of the results. Sometimes, one may also use visualisation approaches to help interpret the outcome. A dashboard can also be developed to provide real time interaction between a user and the results of the analysis.

Within this workflow, data storage will also take place since it will not be possible to keep the large amount of data in computer memory throughout the different Big Data processing tasks.

1.3 Big Data Management

Once you have a better understanding on Big Data, the differences between Big Data Analytics and conventional data analytics, as well as Big Data Processing, you can easily understand more about Big Data Management. In principle, it has to do with the storage and management of Big Data for your organization.
Let's look at the 5-minute 30-second video below to get the first impression on this topic:

What is Big Data Management?

https://youtu.be/I7Dj-n-H3JM

Source: (Jotform, 2021)

You may find some of explanation are similar to the topics before, but the difference here is the focus on the organisation of the Big Data you will want to store, manage and process. Without a careful and proper plan for Big Data Management, the technology will not only help you to reach the insights you intend to discover, you may also not have the most efficient plan to carry out your Big Data project.
The appropriate management will also help better understanding and planning of your Big Data project. Also, it will also help ensure other factors like data quality, accessibility and appropriate strategies in your Big Data implementation and journey.
Read the article by SAS below to learn the five things about big data management.

Key reading

Loshin, D. (n.d.) Big data management: 5 things you need to know. SAS. https://www.sas.com/en_si/insights/articles/data-management/Big-data-management-5-things-you-need-to-know.html

1.3.1 Discussion: Identifying the key concepts of Big Data

Time: 30 minutes

Purpose: To identify the key concepts of Big Data

Task: Share your responses to the following question in the discussion forum.

Question:

Considering the content discussed in Topic 1.3 and the insights from the SAS article 'Big data management: 5 things you need to know', what are your key takeaways?

1.4 Challenges in Big Data processing and management

Big Data Processing and Management is not without its challenges. Let's look at some of the common challenges faced by organisations:

- Where are the relevant data?

  While we know data is abundant today, but insights are rare. The world produces more data than we can imagine day by day, but many still struggle to make sense out of the data, let alone finding useful insights.

  To many organisations, it is a challenging, if not daunting, task to identify which are the relevant data when they embarked on the data science journey. Without knowing which data may be more relevant, one may spend days or even months to find useful insights. This process is not only about tools, but one also needs to have.

- Accuracy and recency of the data

  Having data does not means they are all useful. Sometimes, we received more noise than usable data. Also, in some cases data can only be relevant if it is up-to-date. These two factors needs to be considered if you wish to embark on the data science journey.

- Unclear process to arrive at the intended outcome

There is no one size fit all solution when it comes to the processes in Big Data analytics. One has to first understand the problem, and analyse the available data to discover best ways to process the data in order to arrive at the intended insights. This is also why often we describe Big Data processing as a journey.

- Not understanding the needs for Big Data analytics

  When Big Data was first introduced, many jump on the bandwagon very quickly. However, they may not have understood where exactly will Big Data benefit their business or operation. Some have wrongly assumed that Big Data will automatically bring solutions to their problems even though they do not have the necessary data to begin with. The mismatch between expectations and real needs will be an important part in the change management for an organisation to adopt Big Data analytics.

- Costing

  The costing for Big Data may not be cheap, especially when one does not have an appropriate strategy or understanding on their organisations' needs and problems. Regardless if the implementation is on premise or in the cloud, by purchasing and deploying infrastructure without a proper plan, or just following the trend will cause the organisation to spend more money than necessary. The maintenance and also talent development costing are also important to be planned in the whole implementation costing.

- Talents are hard to find

  Finally, the shortage of Big Data talents is another challenge faced by organisation. The shortage of talent also lead to the challenge to retain talents in one organisation. Without talents with the skills required, organisations will not be able to benefit from the Big Data technology. This is also why we are glad that you took up this programme to up-skill yourself with the relevant data science skills!

Discussion: Challenges in Big Data processing and management

Time: 30 minutes

Purpose: To compare and contrast the challenges of Big Data processing and management

Task: Share your responses to the following question in the discussion forum.

Question: What challenges can you find in Big Data? Try to compare what you find in the Internet or other references. Then, share what will be your TOP THREE challenges and why you have selected them.

1.5 Conclusion and what's next?

Congratulations! You have made it to the end of Week 1! You have learnt the basics of big data analytics and its applications. By knowing the nature of the 5Vs of Big Data, you can evaluate which kind of situation will require a Big Data solution.