BIT34503 Data Science

CHAPTER 9: BIG DATA LANDSCAPE

BIG DATA LANDSCAPE

9.1 Small data versus big data

9.2 Big data analytics vs Data Science

9.3 Key elements in Big Data (3Vs)

9.4 Extracting values from big data

9.5 Challenges in Big data

9.1 Small data versus big data

- Small Data: It can be defined as small datasets that are capable of impacting decisions in the present. Anything that is currently ongoing and whose data can be accumulated in an Excel file. Small Data is also helpful in making decisions, but does not aim to impact the business to a great extent, rather for a short span of Small data can be described as small datasets that are capable of having an influence on current decisions. Almost everything currently in progress and the data of which can be acquired in an Excel file. Small data is also useful in decision-making but is not intended to have a large impact on business, rather for a short period of time.
- In nutshell, data that is simple enough to be used for human understanding in such a volume and structure that makes it accessible, concise, and workable is known as small data.

- Big Data: It can be represented as large chunks of structured and unstructured data. The amount of data stored is immense. It is therefore important for analysts to thoroughly dig the whole thing into making it relevant and useful to make proper business decisions.
- In short, datasets that are really huge and complex that conventional data processing techniques can not manage them are known as big data.

Feature	Small Data	Big Data
Technology	Traditional	Modern
Collection	Generally, it is obtained in an organized manner than is inserted into the database	The Big Data collection is done by using pipelines having queues like AWS Kinesis or Google Pub / Sub to balance high-speed data
Volume	Data in the range of tens or hundreds of Gigabytes	Size of Data is more than Terabytes
Analysis Areas	Data marts(Analysts)	Clusters(Data Scientists), Data marts(Analysts)
Quality	Contains less noise as data is less collected in a controlled manner	Usually, the quality of data is not guaranteed
Processing	It requires batch-oriented processing pipelines	It has both batch and stream processing pipelines

Database	SQL	NoSQL
Velocity	A regulated and constant flow of data, data aggregation is slow	Data arrives at extremely high speeds, large volumes of data aggregation in a short time
Structure	Structured data in tabular format with fixed schema(Relational)	Numerous variety of data set including tabular data, text, audio, images, video, logs, JSON etc.(Non Relational)
Scalability	They are usually vertically scaled	They are mostly based on horizontally scaling architectures, which gives more versatility at a lower cost
Query Language	only Sequel	Python, R, Java, Sequel
Hardware	A single server is sufficient	Requires more than one server

Value	Business Intelligence, analysis and reporting	Complex data mining techniques for pattern finding, recommendation, prediction etc.
Optimization	Data can be optimized manually(human powered)	Requires machine learning techniques for data optimization
Storage	Storage within enterprises, local servers etc.	Usually requires distributed storage systems on cloud or in external file systems
People	Data Analysts, Database Administrators and Data Engineers	Data Scientists, Data Analysts, Database Administrators and Data Engineers
Security	Security practices for Small Data include user privileges, data encryption, hashing, etc.	Securing Big Data systems are much more complicated. Best security practices include data encryption, cluster network isolation, strong access control protocols etc.
Nomenclature	Database, Data Warehouse, Data Mart	Data Lake
Infrastructure	Predictable resource allocation, mostly vertically scalable hardware.	More agile infrastructure with horizontally scalable hardware

9.2 Big data analytics vs Data Science

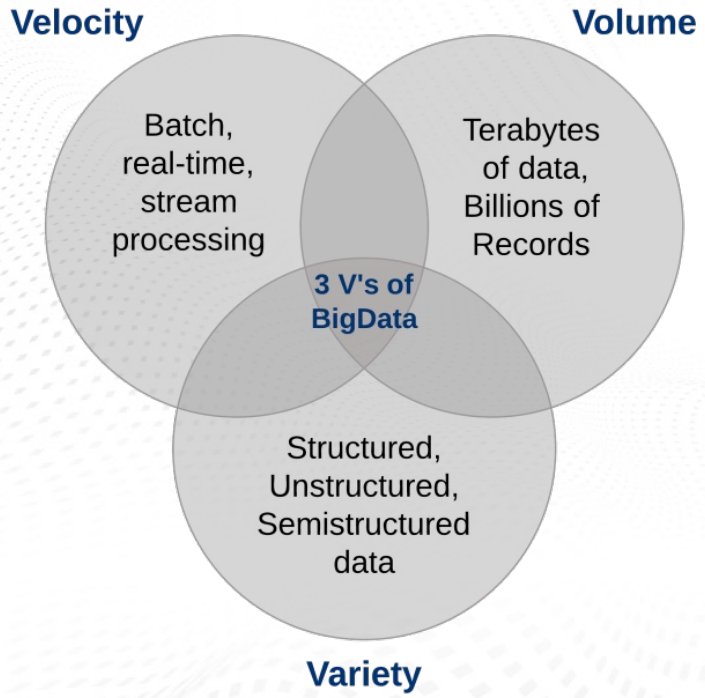
- **Big data** refers to any large and complex collection of data.
- **Data analytics** is the process of extracting meaningful information from data.
- **Data science** is a multidisciplinary field that aims to produce broader insights.

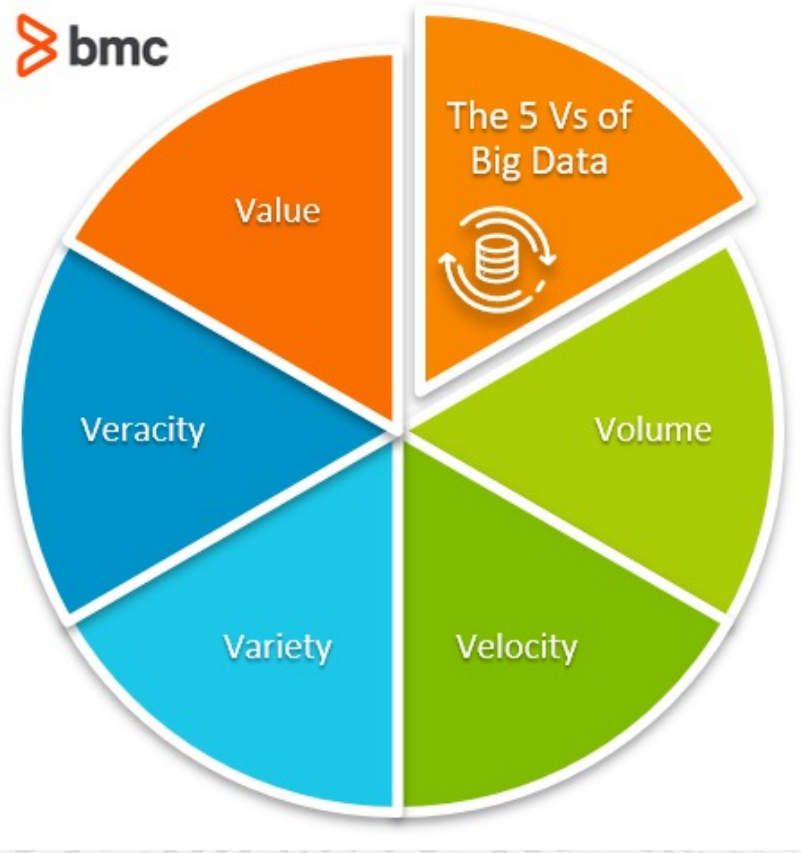
What is big data?

- As the name suggests, big data simply refers to extremely large data sets. This size, combined with the complexity and evolving nature of these data sets, has enabled them to surpass the capabilities of traditional data management tools. This way, data warehouses and data lakes have emerged as the go-to solutions to handle big data, far surpassing the power of traditional databases.
- Some data sets that we can consider truly big data include:
  - Stock market data
  - Social media
  - Sporting events and games
  - Scientific and research data

9.3 Key elements in Big Data (3Vs)

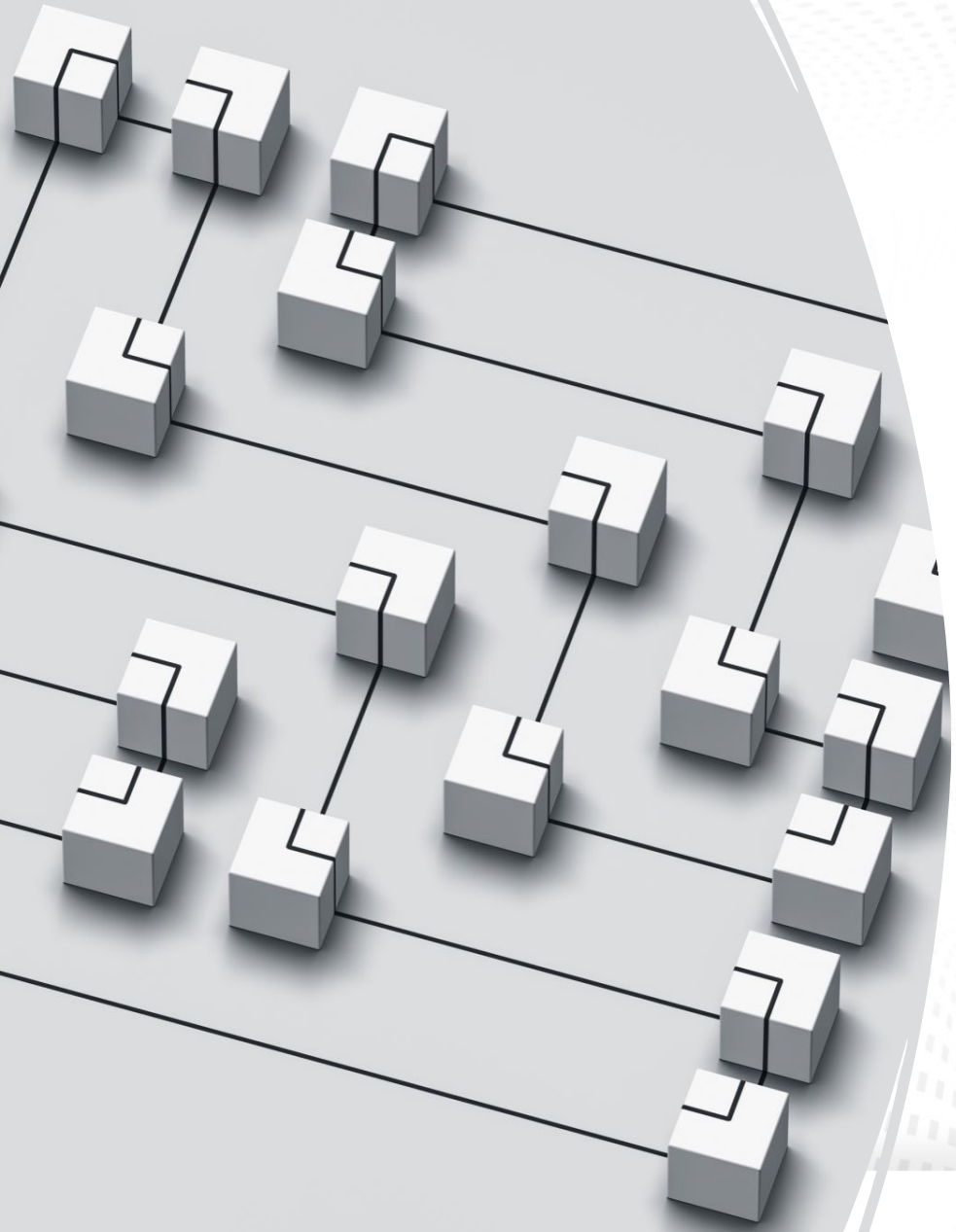
- Characteristics of big data





Characteristics of big data

- **Volume.** Big data is enormous, far surpassing the capabilities of normal data storage and processing methods. The volume of data determines if it can be categorized as big data.
- **Variety.** Large data sets are not limited to a single kind of data—instead, they consist of various kinds of data. Big data consists of different kinds of data, from tabular databases to images and audio data regardless of data structure.
- **Velocity.** The speed at which data is generated. In Big Data, new data is constantly generated and added to the data sets frequently. This is highly prevalent when dealing with continuously evolving data such as social media, IoT devices, and monitoring services.
- **Veracity or variability.** There will inevitably be some inconsistencies in the data sets due to the enormity and complexity of big data. Therefore, you must account for variability to properly manage and process big data.
- **Value.** The usefulness of Big Data assets. The worthiness of the output of big data analysis can be subjective and is evaluated based on unique business objectives.



Types of big data

- **Structured data.** Any data set that adheres to a specific structure can be called structured data. These structured data sets can be processed relatively easily compared to other data types as users can exactly identify the structure of the data. A good example for structured data will be a distributed RDBMS which contains data in organized table structures.
- **Semi-structured data.** This type of data does not adhere to a specific structure yet retains some kind of observable structure such as a grouping or an organized hierarchy. Some examples of semi-structured data will be markup languages (XML), web pages, emails, etc.
- **Unstructured data.** This type of data consists of data that does not adhere to a schema or a preset structure. It is the most common type of data when dealing with big data—things like text, pictures, video, and audio all come up under this type.

bmc

Structured data

- Difficult to collect
- Affordable to collect, process
- Limited insights
- Purpose-driven
- Requires active participation
- Transparency promotes privacy

Unstructured data

- Easy to collect
- Pricier to collect, process
- Nearly infinite insights
- Reusable
- Requires presence only
- Lack of transparency, privacy

Big data systems & tools

- When it comes to managing big data, many solutions are available to store and process the data sets. Cloud providers like AWS, Azure, and GCP offer their own data warehousing and data lake implementations, such as:
  - AWS Redshift
  - GCP BigQuery
  - Azure SQL Data Warehouse
  - Azure Synapse Analytics
  - Azure Data Lake
- Apart from that, there are specialized providers such as Snowflake, Databriks, and even open-source solutions like Apache Hadoop, Apache Storm, Openrefine, etc., that provide robust Big Data solutions on any kind of hardware, including commodity hardware.

What is data analytics?

- Data Analytics is the process of analyzing data in order to extract meaningful data from a given data set. These analytics techniques and methods are carried out on big data in most cases, though they certainly can be applied to any data set.

- The primary goal of data analytics is to help individuals or organizations to make informed decisions based on patterns, behaviors, trends, preferences, or any type of meaningful data extracted from a collection of data.
- For example, businesses can use analytics to identify their customer preferences, purchase habits, and market trends and then create strategies to address them and handle evolving market conditions. In a scientific sense, a medical research organization can collect data from medical trials and evaluate the effectiveness of drugs or treatments accurately by analyzing those research data.
- Combining these analytics with data visualization techniques will help you get a clearer picture of the underlying data and present them more flexibly and purposefully.

Types of analytics

- **Descriptive.** This refers to understanding what has happened in the data set. As the starting point in any analytics process, the descriptive analysis will help users understand what has happened in the past.
- **Diagnostic.** The next step of descriptive is diagnostic, which will consider the descriptive analysis and build on top of it to understand why something happened. It allows users to gain knowledge on the exact information of root causes of past events, patterns, etc.
- **Predictive.** As the name suggests, predictive analytics will predict what will happen in the future. This will combine data from descriptive and diagnostic analytics and use ML and AI techniques to predict future trends, patterns, problems, etc.
- **Prescriptive.** Prescriptive analytics takes predictions from predictive analytics and takes it a step further by exploring how the predictions will happen. This can be considered the most important type of analytics as it allows users to understand future events and tailor strategies to handle any predictions effectively.

Data analytics tools & technologies

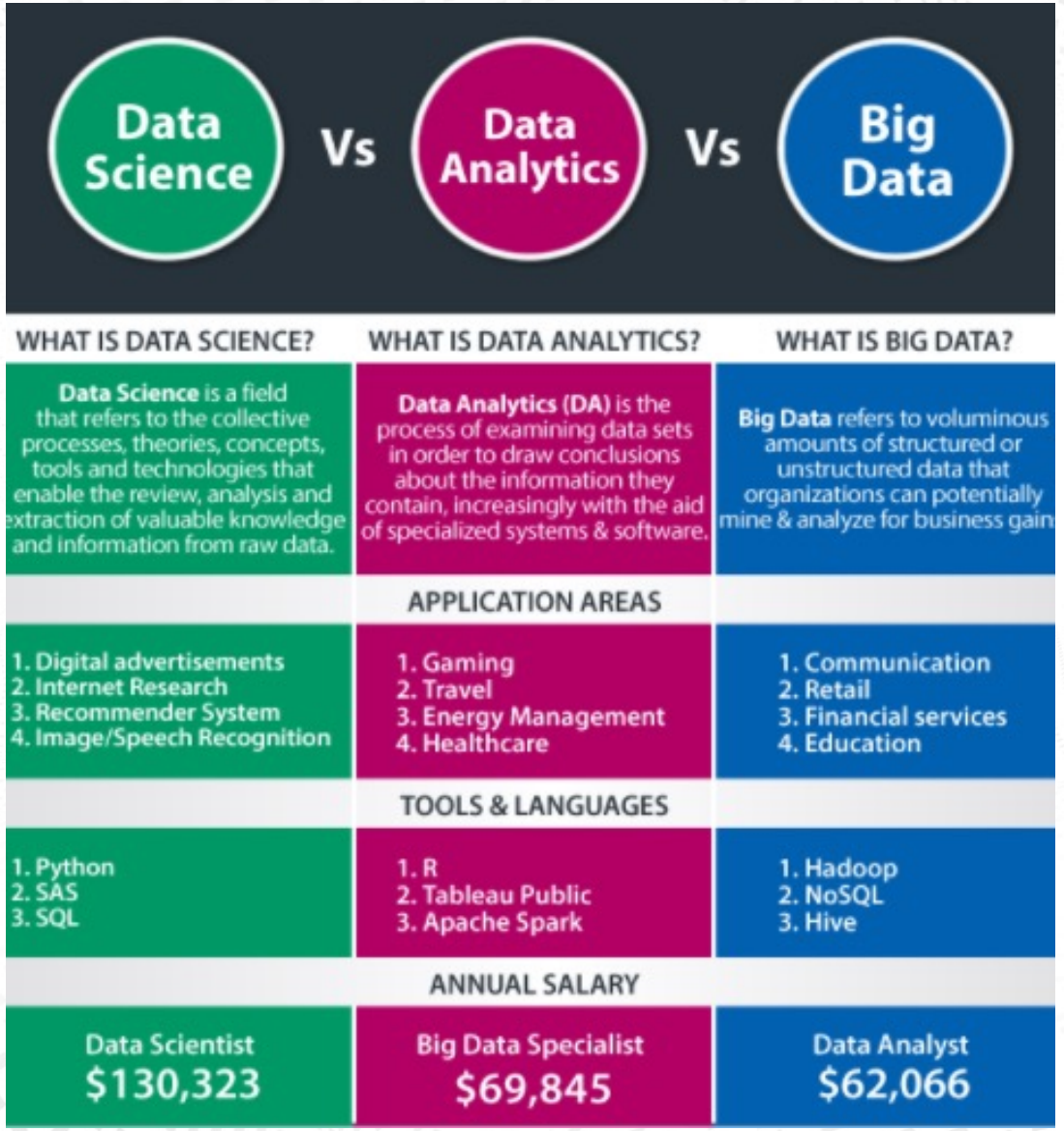
- There are both open source and commercial products for data analytics. They will range from simple analytics tools such as Microsoft Excel's Analysis ToolPak that comes with Microsoft Office to SAP BusinessObjects suite and open source tools such as Apache Spark.
- When considering cloud providers, Azure is known as the best platform for data analytics needs. It provides a complete toolset to cater to any need with its Azure Synapse Analytics suite, Apache Spark-based Databricks, HDInsights, Machine Learning, etc.
- AWS and GCP also provide tools such as Amazon QuickSight, Amazon Kinesis, GCP Stream Analytics to cater to analytics needs.
- Additionally, specialized BI tools provide powerful analytics functionality with relatively simple configurations. Examples here include Microsoft PowerBI, SAS Business Intelligence, and Periscope Data Even programming languages like Python or R can be used to create custom analytics scripts and visualizations for more targeted and advanced analytics needs.
- Finally, ML algorithms like TensorFlow and scikit-learn can be considered part of the data analytics toolbox—they are popular tools to use in the analytics process.

What is data science?

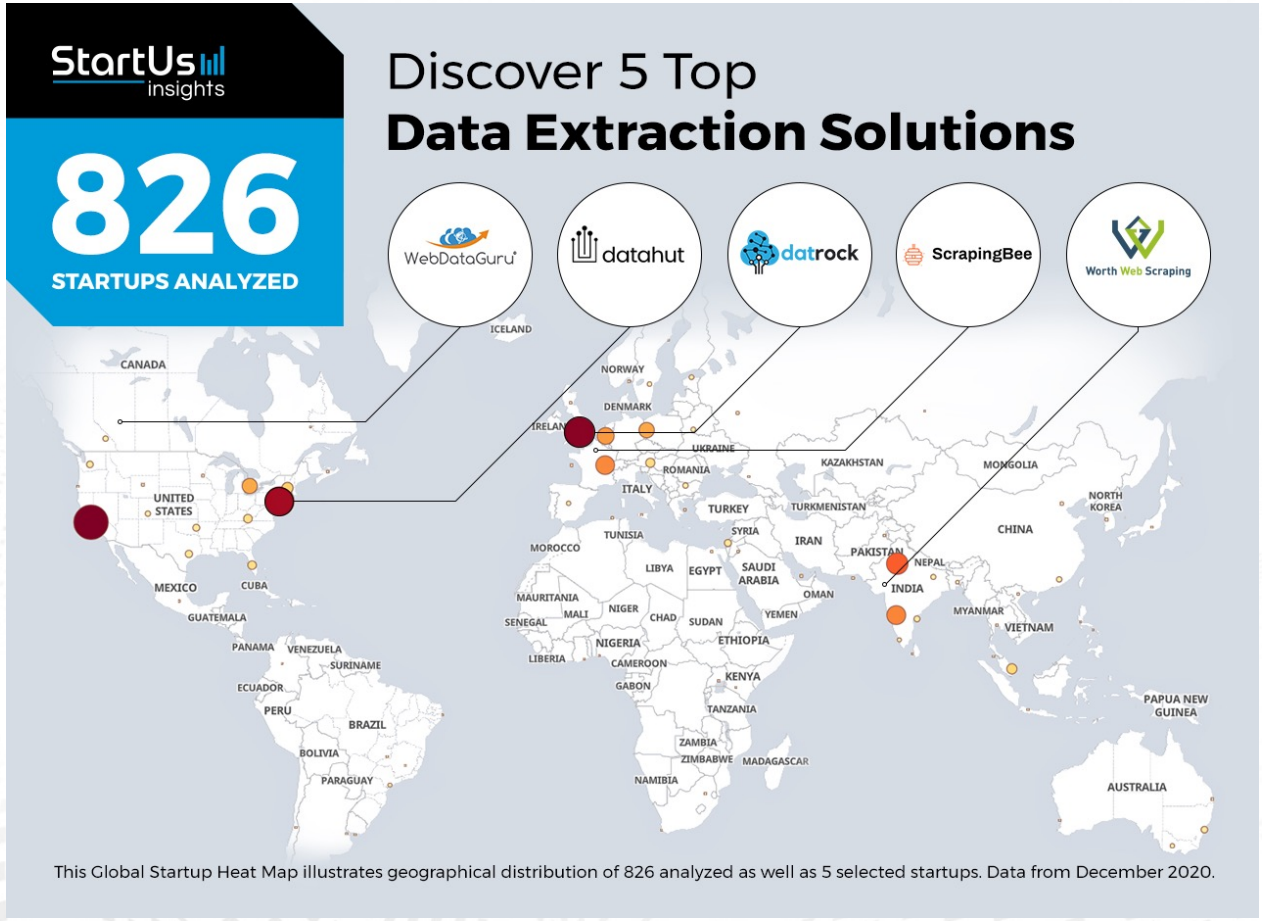
- Data science is a multidisciplinary approach that extracts information from data by combining:
  - Scientific methods
  - Maths and statistics
  - Programming
  - Advanced analytics
  - ML and AI
  - Deep learning
- In data analytics, the primary focus is to gain meaningful insights from the underlying data. The scope of Data Science far exceeds this purpose—data science will deal with everything, from analyzing complex data, creating new analytics algorithms and tools for data processing and purification, and even building powerful, useful visualizations.

Data science tools & technologies

- This includes programming languages like R, Python, Julia, which can be used to create new algorithms, ML models, AI processes for big data platforms like Apache Spark and Apache Hadoop.
- Data processing and purification tools such as Winpure, Data Ladder, and data visualization tools such as Microsoft Power Platform, Google Data Studio, Tableau to visualization frameworks like matplotlib and ploty can also be considered as data science tools.
- As data science covers everything related to data, any tool or technology that is used in Big Data and Data Analytics can somehow be utilized in the Data Science process.


Data Science
Vs
Data Analytics
Vs
Big Data
WHAT IS DATA SCIENCE?
WHAT IS DATA ANALYTICS?
WHAT IS BIG DATA?
Data Science is a field that refers to the collective processes, theories, concepts, tools and technologies that enable the review, analysis and extraction of valuable knowledge and information from raw data.
Data Analytics (DA) is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems & software.
Big Data refers to voluminous amounts of structured or unstructured data that organizations can potentially mine & analyze for business gain
APPLICATION AREAS
1. Digital advertisements
2. Internet Research
3. Recommender System
4. Image/Speech Recognition
1. Gaming
2. Travel
3. Energy Management
4. Healthcare
1. Communication
2. Retail
3. Financial services
4. Education
TOOLS & LANGUAGES
1. Python
2. SAS
3. SQL
1. R
2. Tableau Public
3. Apache Spark
1. Hadoop
2. NoSQL
3. Hive
ANNUAL SALARY
Data Scientist
$130,323
Big Data Specialist
$69,845
Data Analyst
$62,066

9.4 Extracting values from big data



This Global Startup Heat Map illustrates geographical distribution of 826 analyzed as well as 5 selected startups. Data from December 2020.

9.5 Challenges in Big data

Issues	Challenges
Ethical	often unclear when a business or a government department are gathering data about citizens
	People are often targeted without their knowledge, as massive amounts of social media posts, browser histories and web habits in general are often gathered by businesses to make marketing-related decisions.
	subsequently argued whether the collection and usage of these massive amounts of aggregated data to estimate personal behaviour might infringe personal privacy or not.

Data reliability	Big Data collection and evaluation is undoubtedly crucial not only to modern government and business strategies, but it could also play a role towards scientific purposes.
	whether Big Data will aid or mine the effectiveness of scientific and social-scientific methodology is a controversial issue: as the datasets available hold increasingly vast volume and variety about a very large number of areas, it is in fact argued that the modern advanced correlation techniques might substitute scientific hypothesis altogether.
	It could be argued that the size and variety of Big Data would not aid value of knowledge at all, as trying to discern valuable information amongst the huge volume of data can sometimes be an onerous task: an excessive amount of data could lead to imprecise information and, subsequently, erroneous knowledge.

Legal aspects regarding its security and its compliance to current regulations	**Big Data is made available from numerous resources, and as it may be problematic to determine whether certain sources are compromised, there exists a risk of compromising the security of the entire organisation's data by giving access to users with malicious intents. In fact, such a large volume of data does undoubtedly consist in a very valuable resource the misappropriation of which could come to mine various government agencies as well as the general public. It is subsequently crucial to employ efficient protection methods to avoid issues such as malicious data breaches, as well as effective persistency measures to prevent tragic loss of information due to physical causes.**
	businesses gathering new data types and methodologies are still expected to meet the legislative requirements placed on them by compliance laws: BigData can undoubtedly come to encompass "sensitive data", the processing of which isrestricted and prohibited in most cases

Related Readings

- https://www.geeksforgeeks.org/difference-between-small-data-and-big-data/
- https://www.bmc.com/blogs/big-data-vs-analytics
- https://www.devopsschool.com/blog/detailed-difference-among-big-data-data-science-and-data-analytics/
- https://www.analyticsvidhya.com/blog/2020/11/what-is-big-data-a-quick-introduction-for-analytics-and-data-engineering-beginners/


UTHM
Universiti Tun Hussein Onn Malaysia
1993-2023
AMANAH • PROFESIONAL • INOVATIF

Thank you