Big Data Management and Analytics

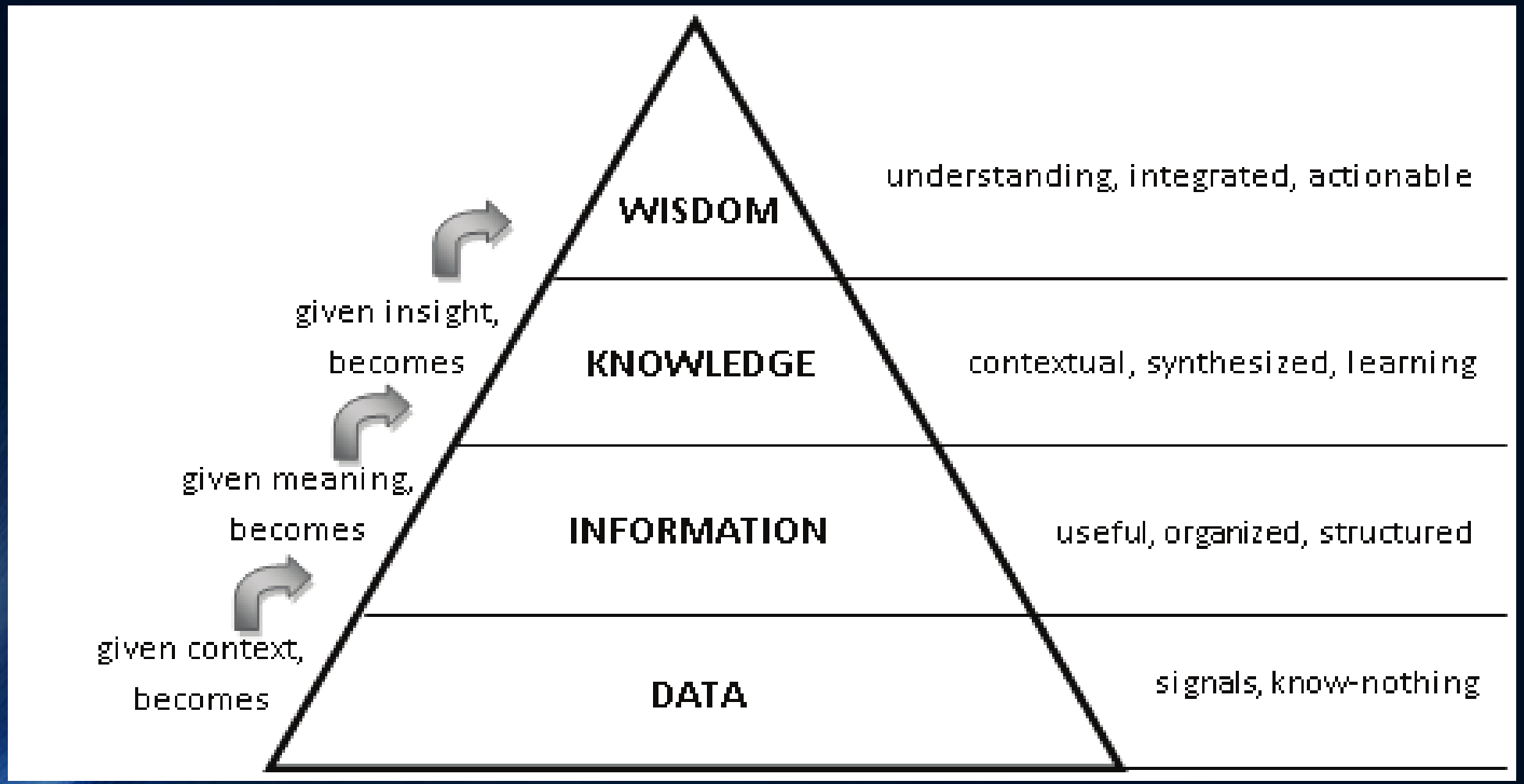
Khaled Mohammed Fouad

Director of MIS & DSS unit.

Agenda

- Introduction
- Big Data Management
- Big Data Analytics
- Big Data Analytics Tool, platforms and languages
- Big Data Analytics Use Cases

The role of data analytics in the organization



The Big Data Analytics Opportunities in the organization

- Operational excellence

It improves efficiency and productivity and results in higher profits.

- New products, services, and business models

It enable firms to create new products, services, and business models.

- Customer and supplier intimacy

Customers who are served well become repeat customers.

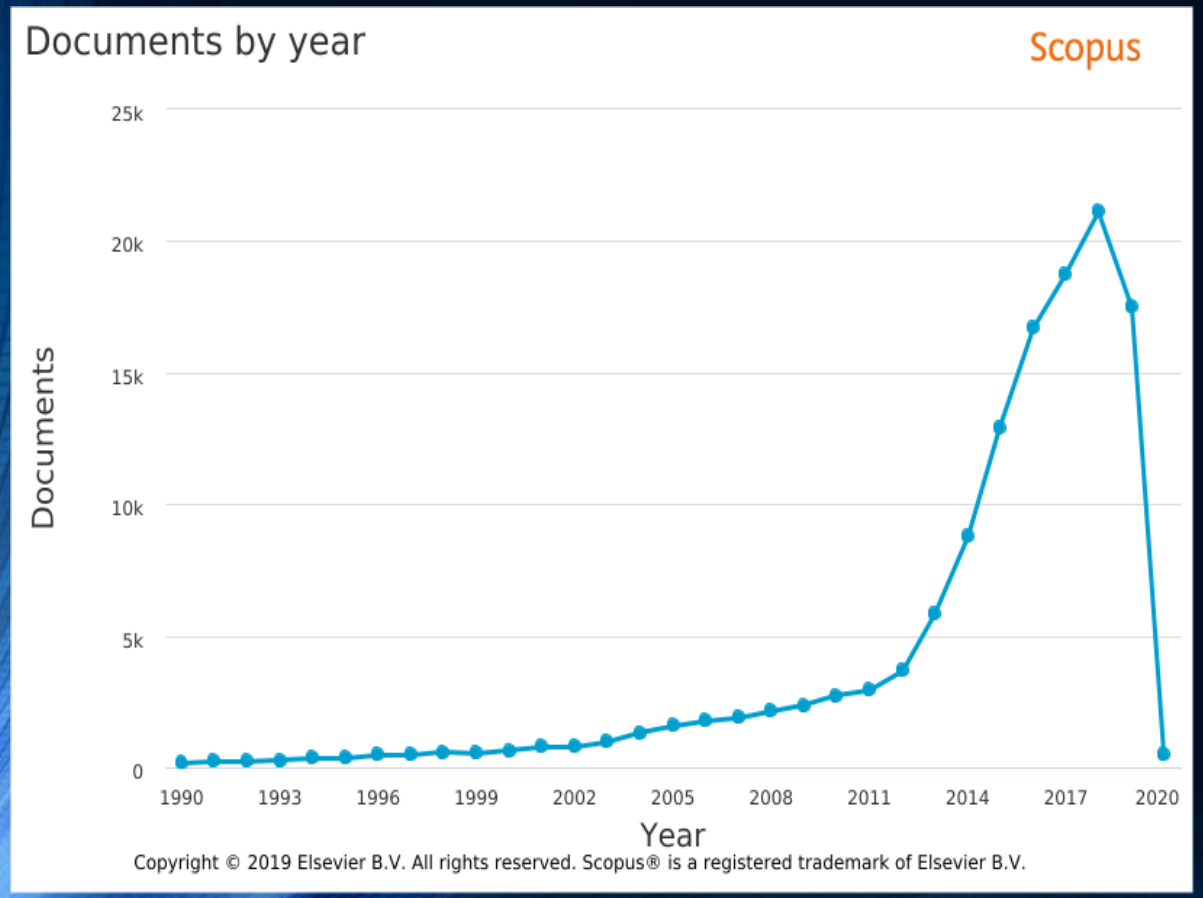
- Improved decision making

If managers rely on forecasts, best guesses, and luck,

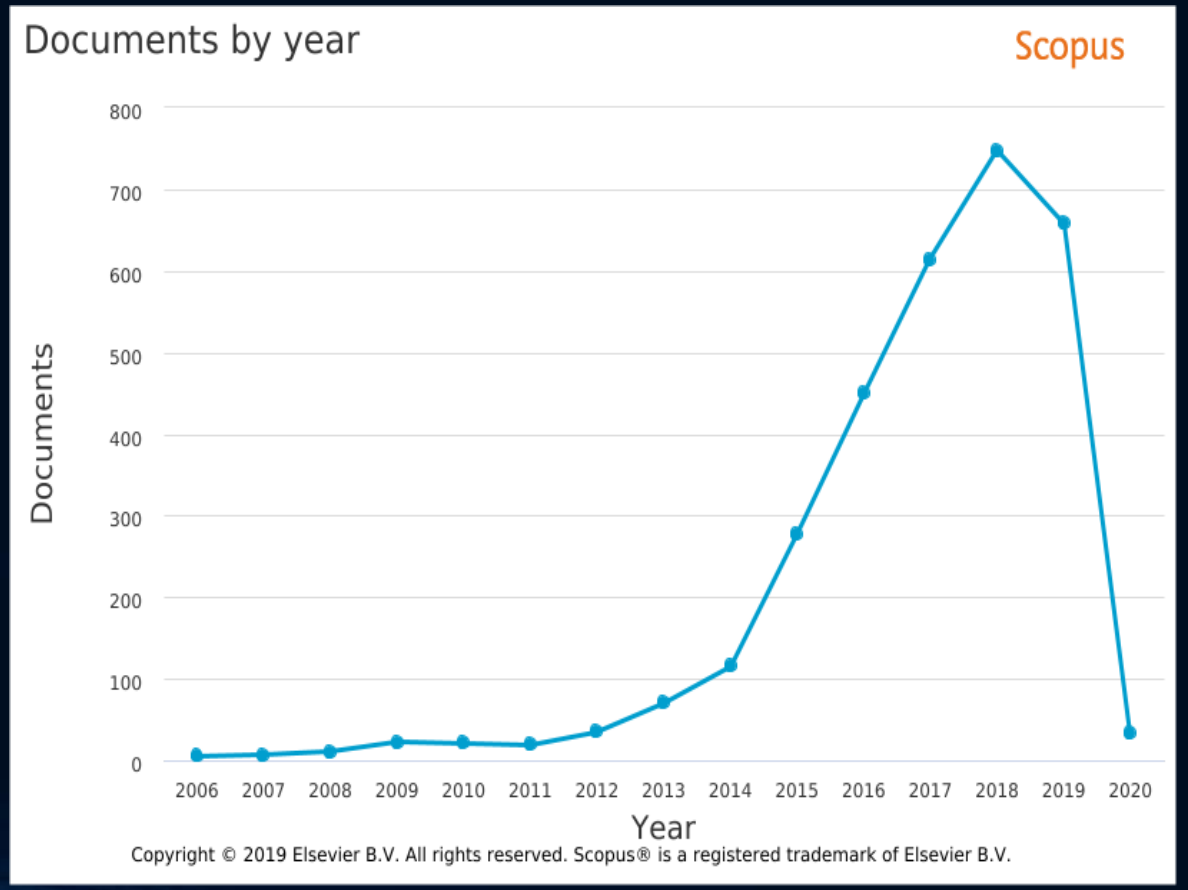
they will not take a right decision.

The Importance of Big Data is reflected on research Publication

No of papers in all world

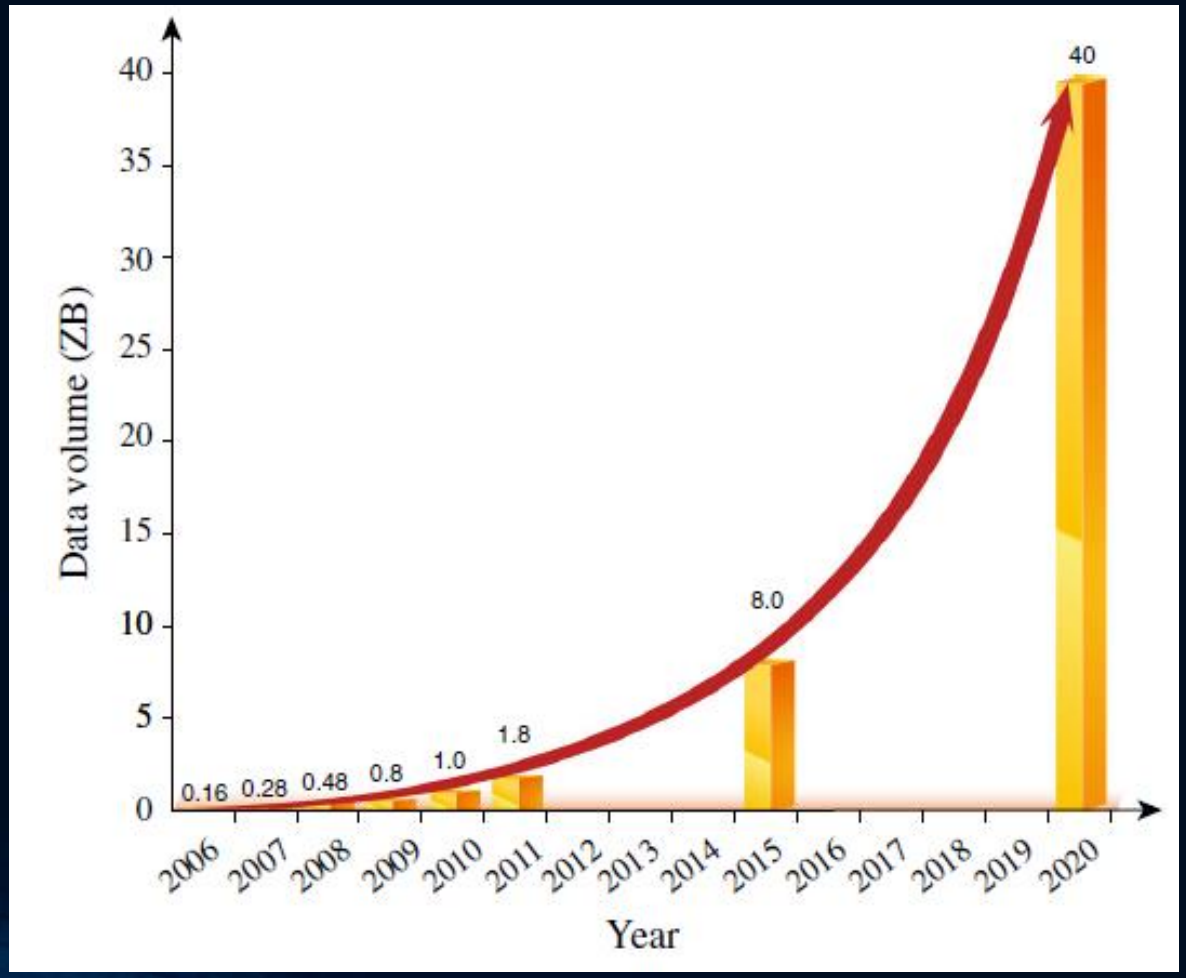


No of papers in the Arab world



Why has Big Data gain much attention

- Because of large amounts of data are generated from a variety of sources, the datasets increase at an exponential rate. [7][8]



Why has Big Data gain much attention

- The size of the generated data per day on the Internet has already exceeded two Exabyte (10^{18} bytes) [1, 7].

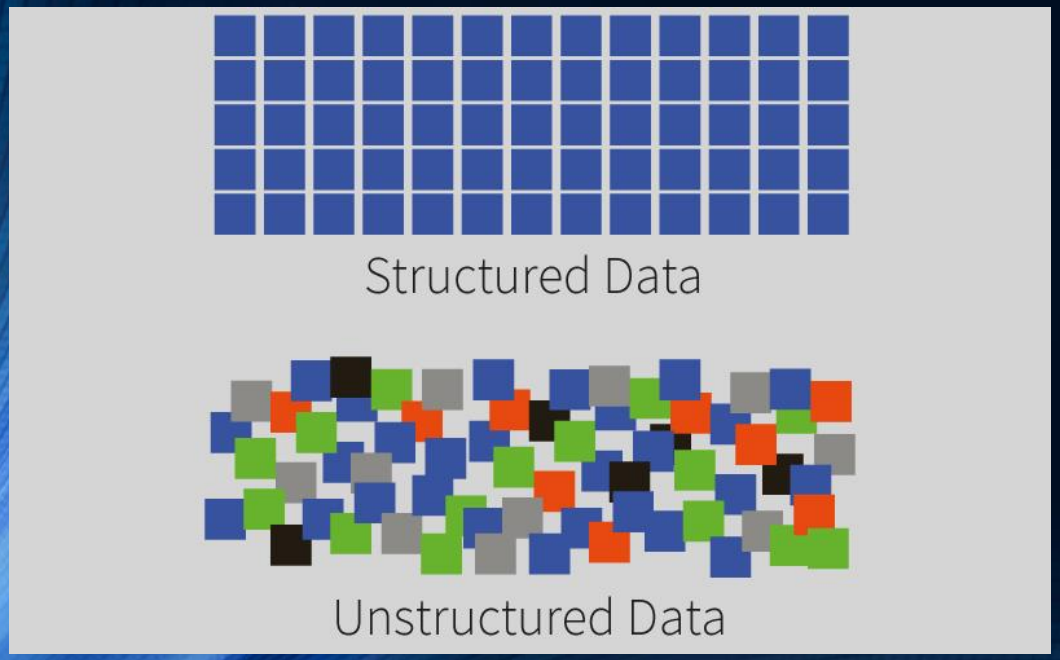
Notations of Data	Size	Comes Under
Bit	1/8 Byte	Data
Nibble	½ Byte	Data
Byte	1 Byte	Data
Megabyte	1,024 Kilobytes	Data
Gigabyte	1,024 Megabytes	Big Data
Terabyte	1,024 Gigabytes	Big Data
Petabyte	1,024 Terrabytes	Big Data
Exabytes	1,024petabytes	Big Data
Zettabyte	1,024 Exabytes	Bigger Than Big Data
Yottabyte	1,024 Zettabytes	Bigger Than Big Data
Googolbyte	10+1000's Bytes	Bigger Than Big Data

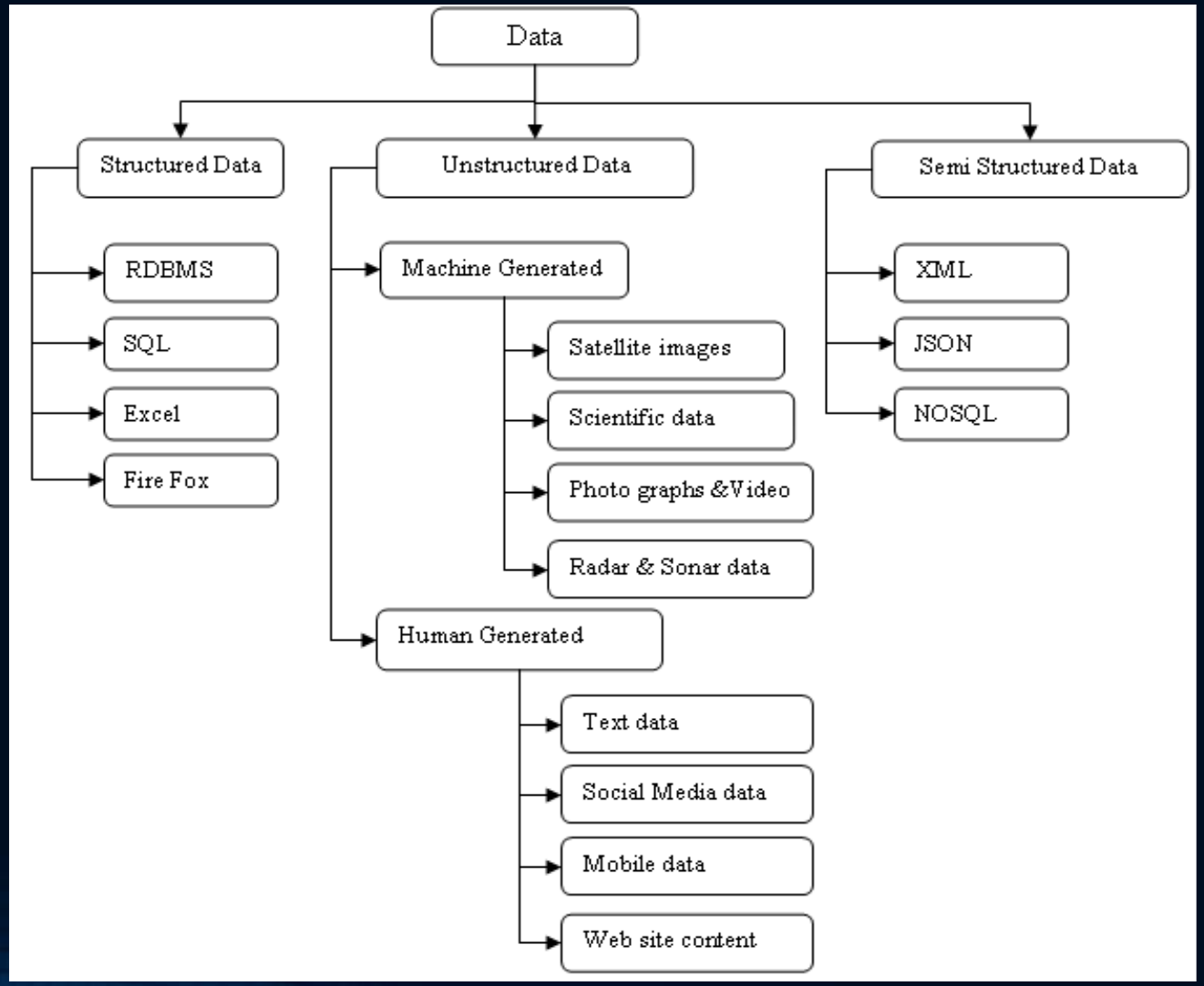
Why has Big Data gain much attention

- Within one minute [7],
  - 72 h of videos are uploaded to YouTube.
  - Around 30.000 new posts are created on the Tumblr blog platform.
  - More than 100.000 Tweets are shared on Twitter.
  - More than 200.000 pictures are posted on Facebook.

Big Data - Various types of data formats [1]

- Structured data
- Unstructured data
- Semi-structured data





Big Data - Big Data characteristics/dimensions [5, 6]

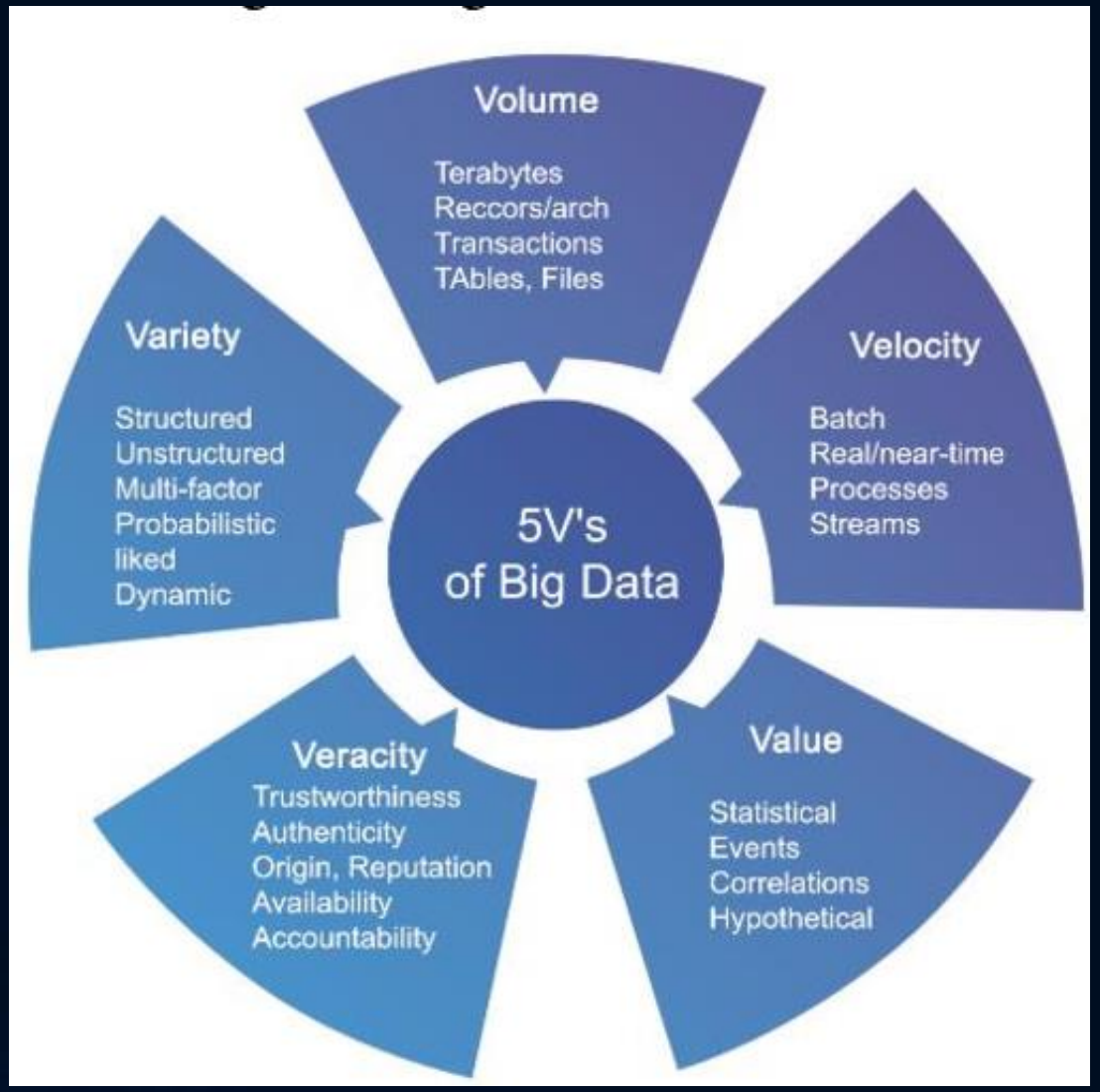
Volume: the massive amount of data that is required to be processed.

Velocity: the speed in which data must be stored and analyzed.

Value: the value that the data can add to the organization.

Veracity: the consistency of the data (assurance) and the reliability of the data.

Variety: the diversity of the data sources and formats.



Big Data - Big data Challenges [9]

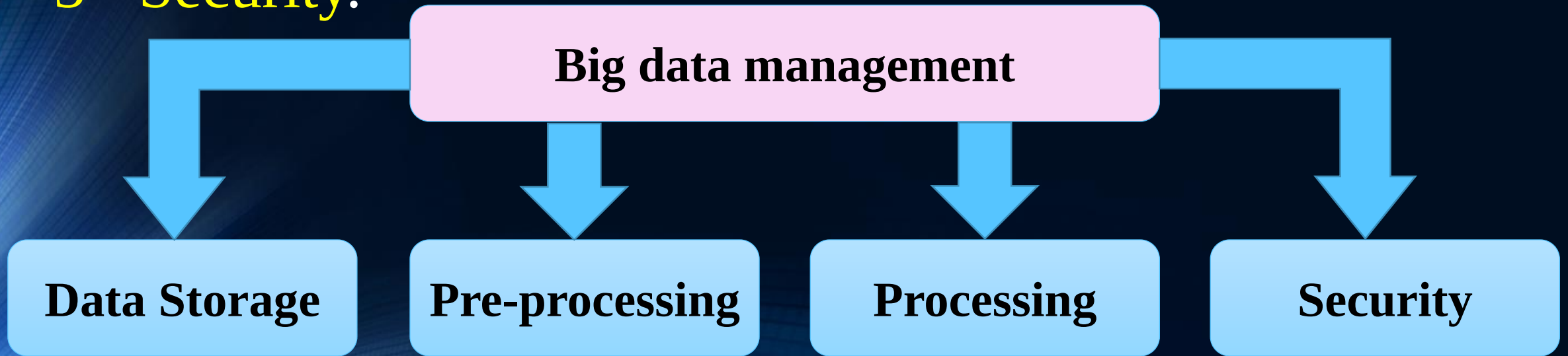
- Data volume
  The ability to capture, store, and process the huge volume of data in a timely manner.
- Data integration
  The ability to combine data quickly and at reasonable cost .
- Processing capabilities
  The ability to process the data quickly, as it is captured (i.e., stream analytics).
- Data governance (security, privacy, access).
- Skill availability (data scientist).
- Solution cost (ROI).

Agenda

- Introduction
- Big Data Management
- Big Data Analytics
- Big Data Analytics Tool, platforms and languages
- Big Data Analytics Use Cases

Big data Management - Taxonomy of big data management [10]

- The objective of big data management is ensuring the *effectiveness* of
  1- Big data storage,
  2- Processing applications and,
  3- Security.



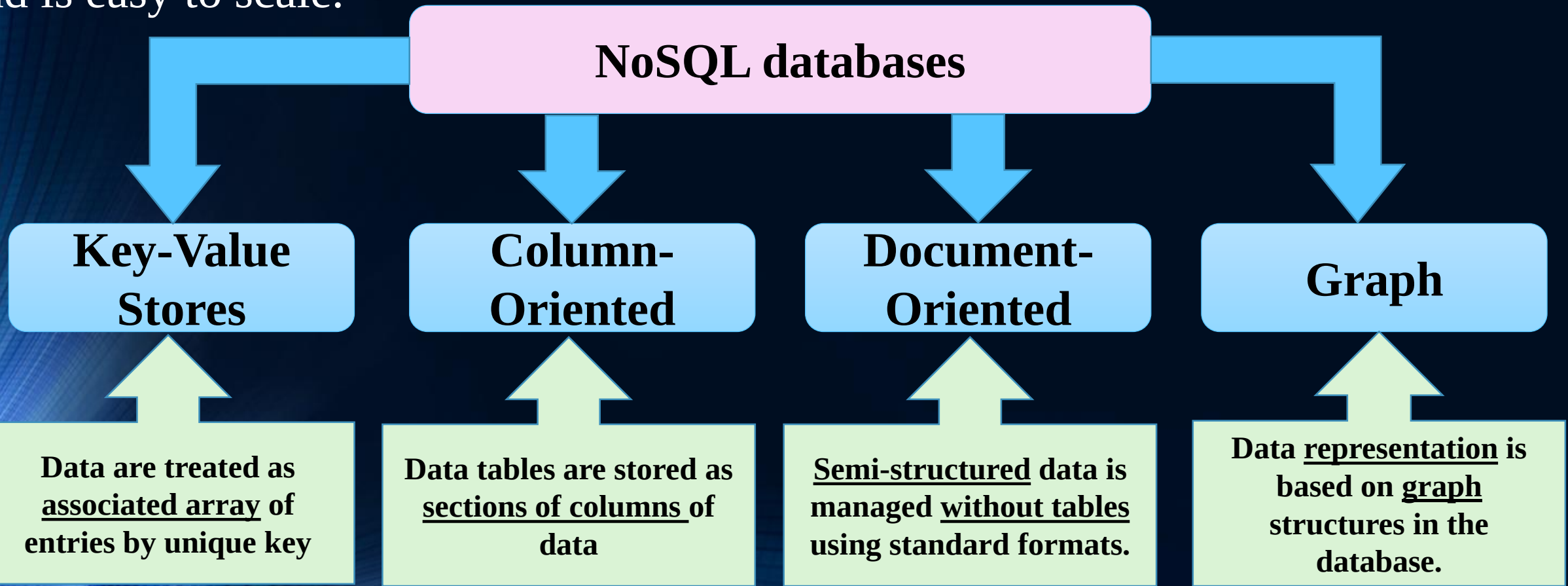
Big data Management - Data Storage [2]

- **Data Storage Technologies**
  - Relational databases require a schema before writing to the database.
  - It is too rigid to manipulate volumes of real-time data with diverse data structures.
  - The ACID properties are too strict.

Big data Management - Data Storage [2]

❑ NoSQL databases [11]

It is a non-relational DMS, that does not require a fixed schema, avoids joins, and is easy to scale.



Big data Management [2]

- **Pre-processing** [10]

1. Transmission is transferring raw data to a big data storage infrastructure.
2. Data cleansing is detecting and handling incomplete, inaccurate or irrational data.

Big data Management [2]

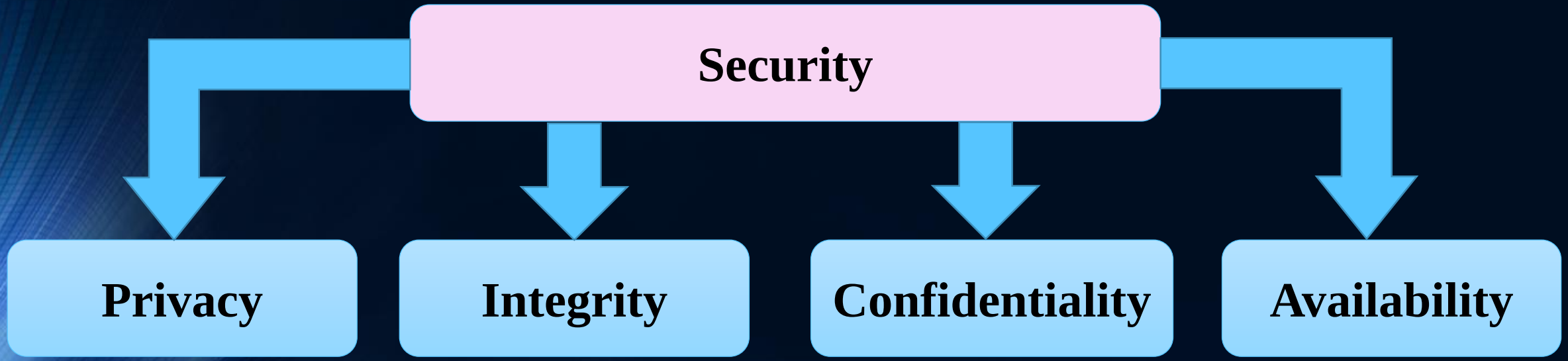
- **Processing** [10]

1. Classification classifies labelled objects into different groups or classes.
2. Prediction discovers a relationship between dependent and independent variables.

Big data Management [2]

- **Security** [10]

  - Data is generated from multiple sources, security has become a serious concern.



Big data Management - Security [2, 10]

1- Privacy

- It is required to find those violating privacy rules and to ensure that user data is not misused or leaked.

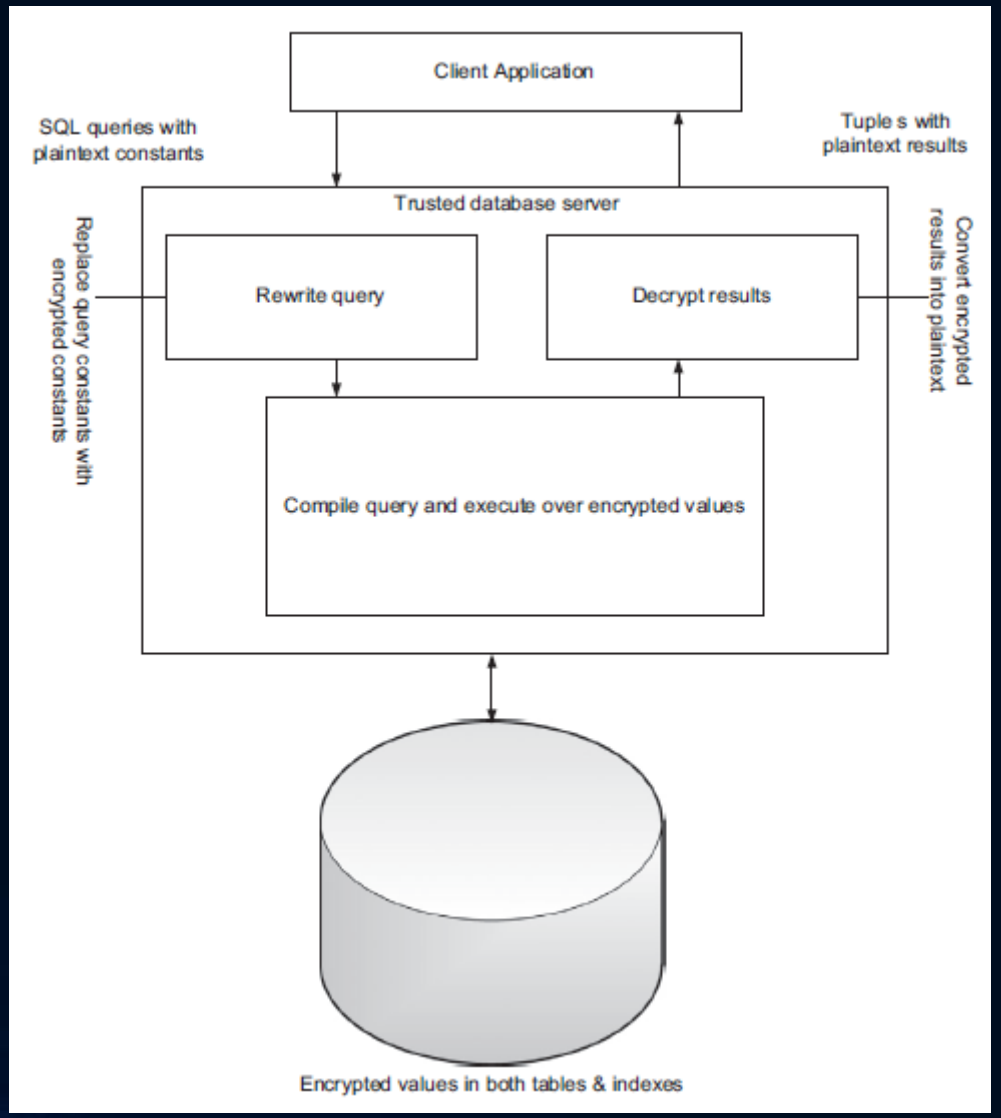
2- Integrity

- Data integrity is severe for large-scale collaborations and Integrity constraints aims to provide data consistency and accuracy.

Big data Management - Security [2, 10]

3- Confidentiality

- It can be achieved by protecting data from unauthorized and unintended users using encryption methods.
- Encryption levels are table encryption, disk encryption, and data encryption.



Big data Management - Security [2, 10]

4- Availability

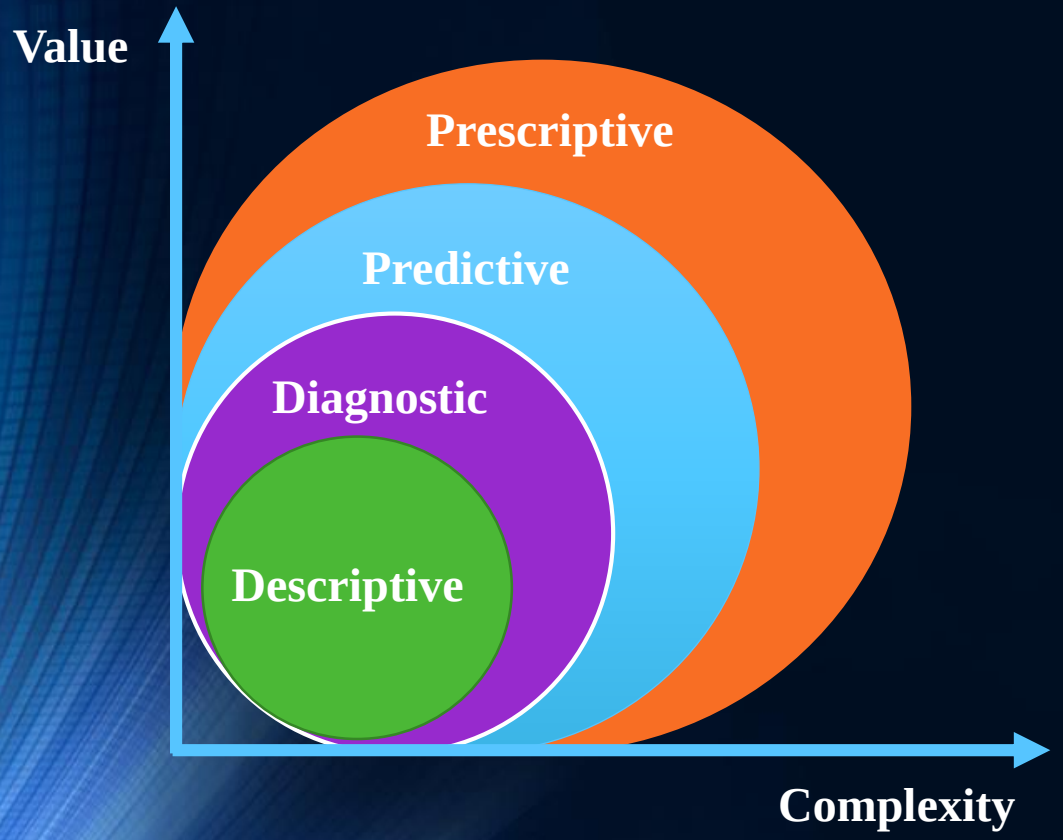
- In the cloud where large amounts of data reside, availability is crucial for data outsourcing.
- Different types of threats that cause data unavailability.
- To prevent users from working properly, request floods are submitted to a particular server that offers a particular service.

Agenda

- Introduction
- Big Data Management
- Big Data Analytics
- Big Data Analytics Tool, platforms and languages
- Big Data Analytics Use Cases

Big data Analytics

❑ The types of analytics approaches [13, 14]



Prescriptive: (What do I need to do? What should we do?) deploys the power of decision science, management science, and operations research methodologies to make the best use of allocated resources.

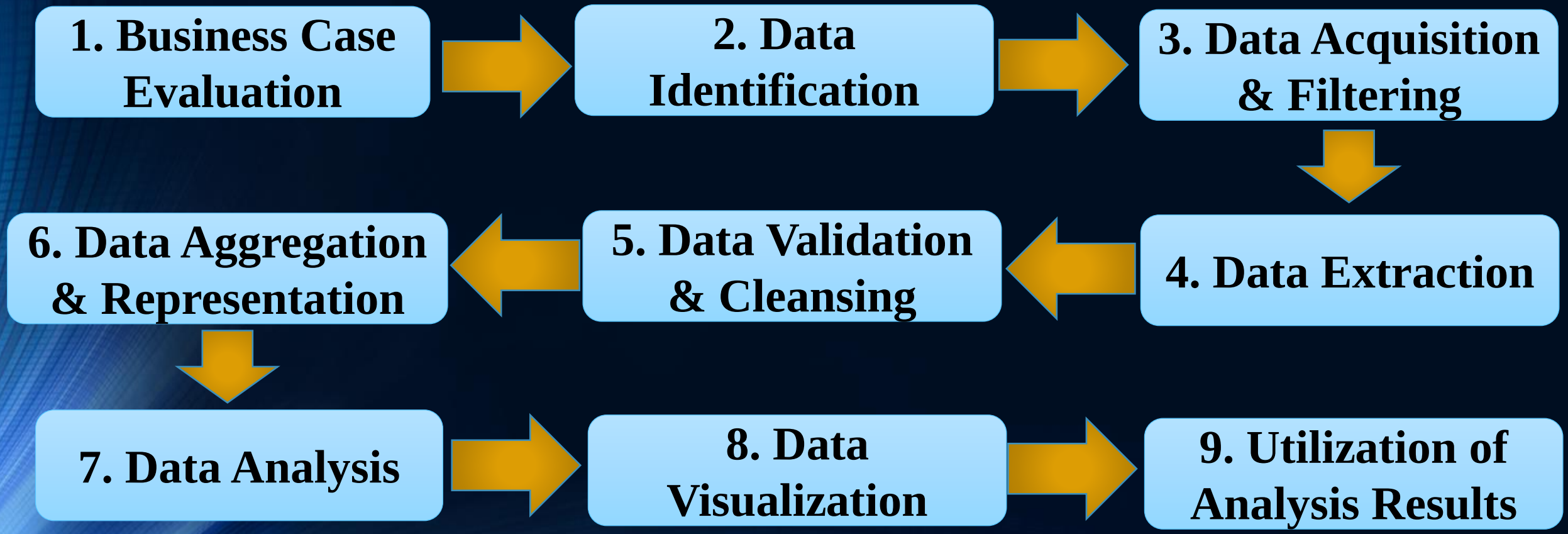
Predictive: (What's likely to happen? What could happen?) uses advanced statistical methods to identify predictive variables and build predictive models.

Diagnostic: (Why is it happening?) uses the analysis of past data to ascertain the cause of certain events.

Descriptive: (What is happening?) describes what is contained in a dataset. Descriptive analytics includes statistical measures.

Big Data Analytics [12]

- Big Data Analytics lifecycle



Big data Analytics - Big data Analytics lifecycle [12]

1. Business Case Evaluation

- It presents a clear understanding of the justification, motivation, and goals of carrying out the analysis.

2. Data Identification

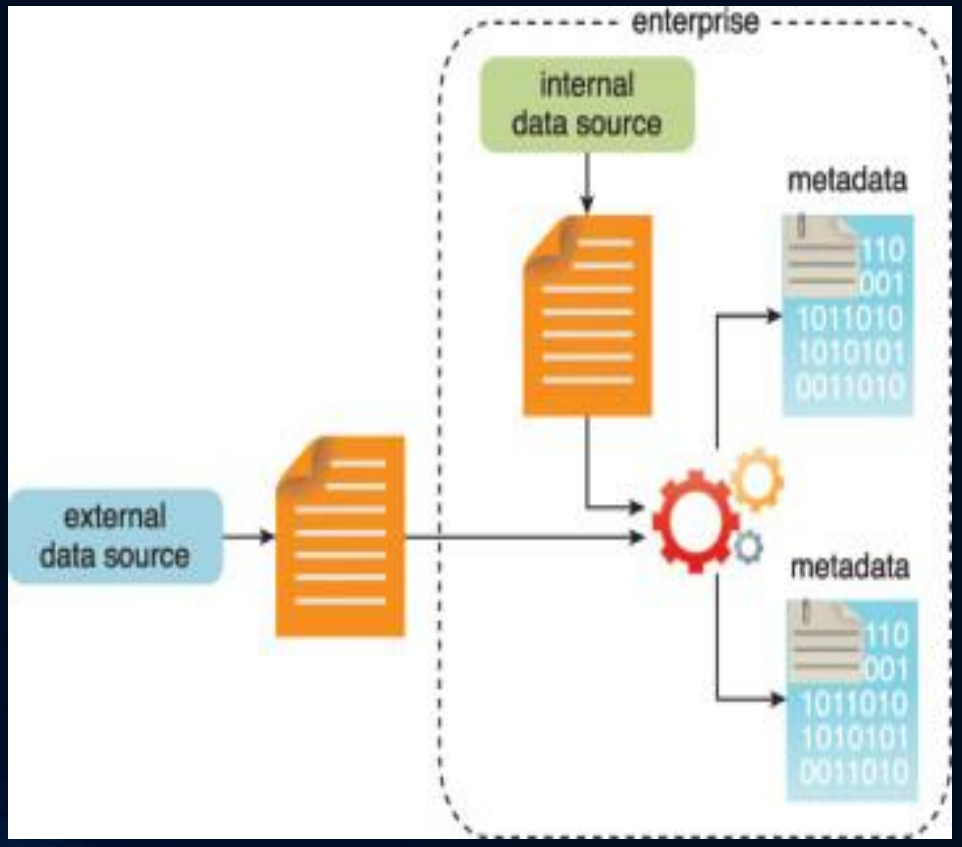
- It identifies the datasets required for the analysis project and its sources.

Big data Analytics - Big data Analytics lifecycle [12]

3. Data Acquisition & Filtering

- The data is gathered from all the data sources, then data is filtered for the removal of corrupt data.

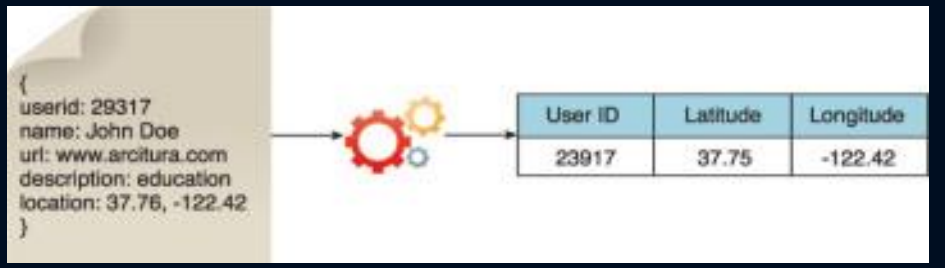
- Metadata is generated from both internal and external data sources.



Big data Analytics - Big data Analytics lifecycle [12]

4. Data Extraction

- It extracts various data,
- It transforms it into a format that the underlying big data solution can use for the purpose of the data analysis.



Big data Analytics - Big data Analytics lifecycle [12]

5. Data Validation & Cleansing

- It establishes validation rules and removing any known invalid data to enhance data quality.

Big data Analytics - Big data Analytics lifecycle [12]

6. Data Aggregation & Representation

- It integrates multiple datasets together to arrive at a unified view.

Big data Analytics - Big data Analytics lifecycle [12]

7. Data Analysis

- It performs the actual analysis task.
- It combines data mining and complex statistical analysis techniques to discover patterns and depict relationships between variables.

Big data Analytics - Big data Analytics lifecycle [12]

8. Data Visualization

- It uses data visualization techniques and tools to graphically provide the analysis results to users.

9. Utilization of Analysis Results

- It aims to determine how and where analyzed data can be further leverage.

Agenda

- Introduction
- Big Data Management
- Big Data Analytics
- Big Data Analytics Tool, platforms and languages
- Big Data Analytics Use Cases

Big Data Analysis Platforms and Tools [15, 16, 17]

- These platforms aims at examining large datasets, both structured and unstructured data to discover hidden patterns.
- These platforms support parallel and distributed computation.

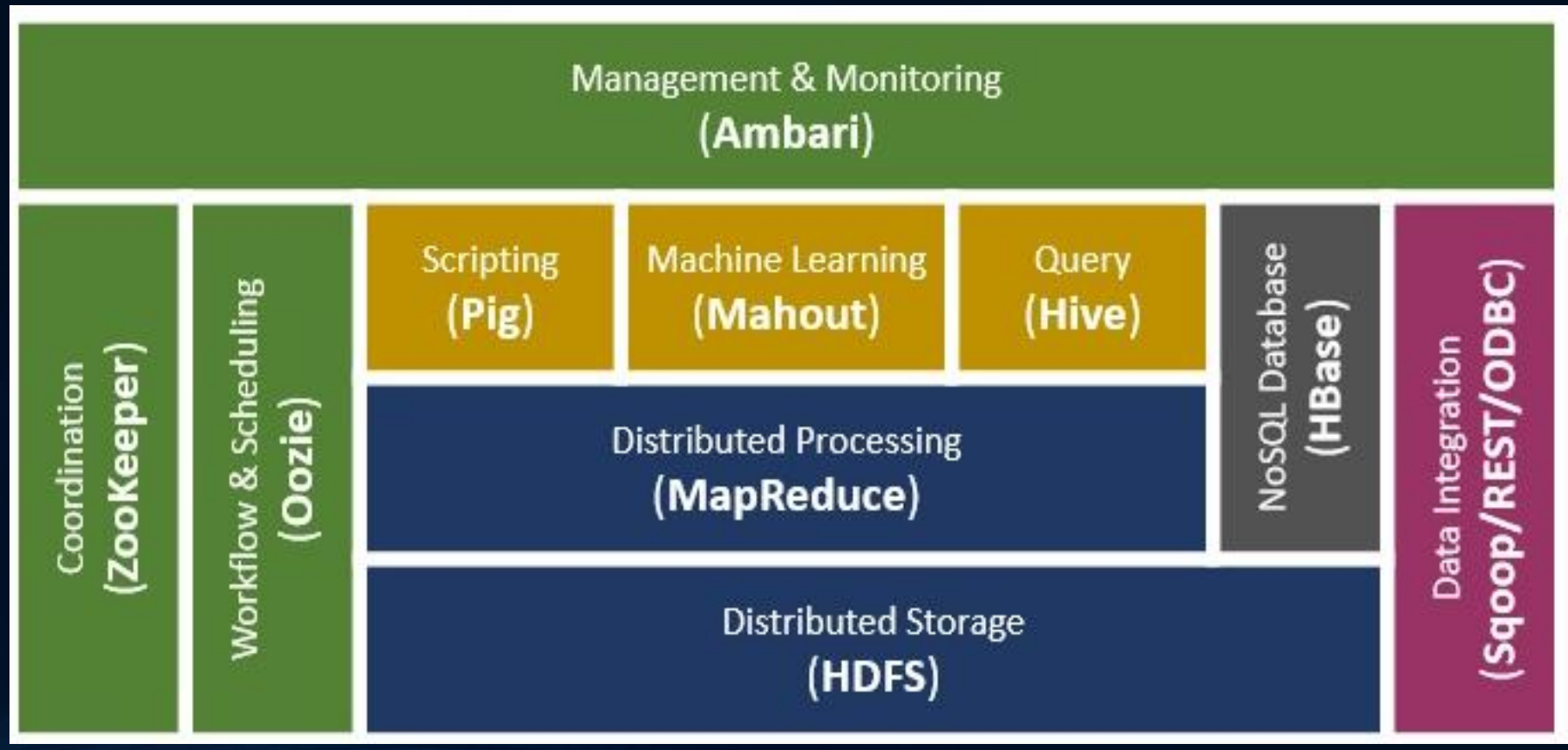
Big Data Analysis Platforms and Tools [15, 16, 17]

- **Hadoop**

  - It is an open-source apache framework.
  - Apache Hadoop is one of the most well-established software platforms that support data-intensive distributed applications.

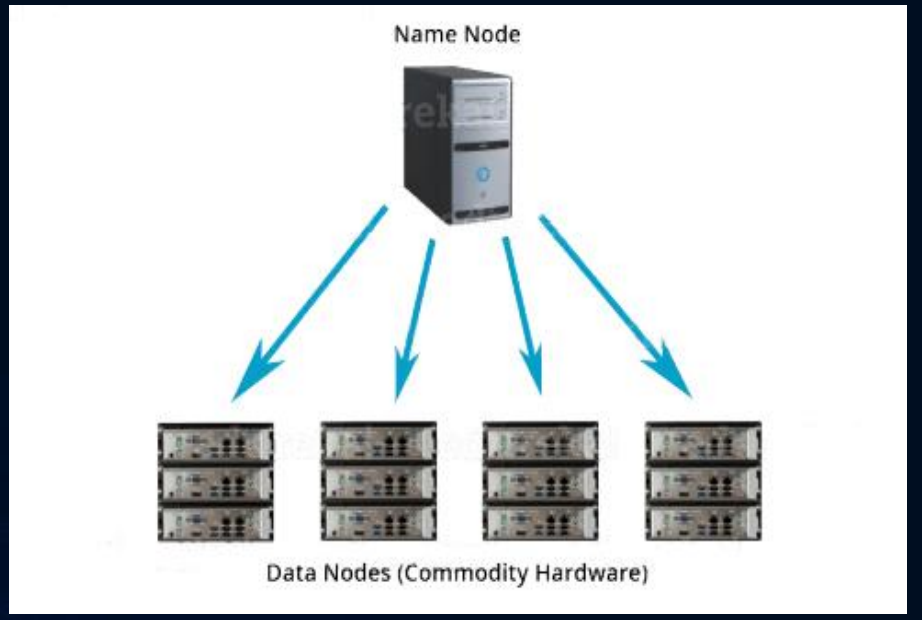
Big Data Analysis Platforms and Tools [15, 16, 17]

- **Hadoop (*Hadoop Ecosystem*)**



Big Data Analysis Platforms and Tools [15, 16, 17]

- **Hadoop Distributed File System (HDFS)**
  - It is the primary or significant component of the Hadoop ecosystem and is responsible for storing large datasets (structured or unstructured) across various nodes.



Big Data Analysis Platforms and Tools [15, 16, 17]

- **MapReduce- Distributed Data Processing Framework of Apache Hadoop**
  - MapReduce performs computational tasks in a batchwise manner, and it is a data-parallel architecture.
  - Instead of moving data to computation, MapReduce works on the concept that moves the computation to the data.
  - It performs two computational tasks; Map and Reduce.

Distributed machine learning tools [17]

- **Spark MLlib**
  - Apache Spark consists of a library of ML algorithms known as Mllib.
- **FlinkML**
  - FlinkML is Apache ML library. FlinkML aims to develop scalable ML algorithms.

Cloud-based machine learning tools [17]

- **Microsoft Azure ML**
  - It deploy predictive analytics tasks for the user's data in the Microsoft Azure cloud.

- **IBM Watson ML**
  - It is designed to provide an end-to-end ML platform for data scientists. It offers many optimizations that can ease installation, management, and can help accelerate performance.

Programming Languages [15]

- **Pig/Pig Latin**
  - Pig is a high-level language used for the analysis purposes .

- **Python**
  - Python is an open-source, easy-to-use, high-level and has large libraries.

- **Julia**
  - Julia was designed for parallel and distributed computation.

Programming Languages [15]

- **Go**

  - Go is an open-source statically typed programming language developed at Google.

- **R**

  - R is an open-source project for data statistical computing.

Agenda

- Introduction
- Big Data Management
- Big Data Analytics
- Big Data Analytics Tool, platforms and languages
- Big Data Analytics Use Cases

Use Case 1: Summary of Genomics

Application:

- Integrate data from multiple sequencing technologies and methods
- Develop highly confident characterization of whole human genomes as reference materials,

Current Approach:

- The storage of ~40TB NFS at NIST is full; there are also PBs of genomics data at NIH/NCBI.
- Use Open-source sequencing bioinformatics software from academic groups on a 72 core cluster at NIST supplemented by larger systems at collaborators.

Future:

- DNA sequencers can generate ~300GB compressed data/day which volume has increased much faster.
- Future data could include other 'omics' measurements, which will be even larger than DNA sequencing.
- So, the solution is using a Big Data platform.

Use Case 2: Twitter Data

Application: Information diffusion from Twitter Data

- Understand how communication spreads on social networks.
- Detect potentially harmful information spread at the early stage.

Current Approach:

- Acquisition and storage of a large volume (30 TB a year compressed) of continuous streaming data from Twitter (~100 million messages/day, ~500GB data/day increasing);
- Near real-time analysis of such data, for anomaly detection;

Future:

- Truthy plans to expand incorporating Google+ and Facebook.
- Need to move towards Hadoop/Indexed HBase & HDFS distributed storage.
- Use Redis as an in-memory database to be a buffer for real-time analysis.
- Need streaming clustering, anomaly detection and online learning.

Thank you

References

- [1] M. A. Srinuvasu, "Big Data : Challenges and Solutions International Journal of Computer Sciences and Engineering Open Access Big Data : Challenges and Solutions," no. October, 2017.
- [2] V. C. Storey and I. Song, "Data & Knowledge Engineering Big data technologies and Management : What conceptual modeling can do," Data & Knowledge Engineering, vol. 108, no. February, Elsevier B.V., pp. 50–67, 2017.
- [3] B. R. N. Singh and B. R. S. Reddy, "A Review on Big Data Mining in Cloud Computing," pp. 131–142, 2017.
- [4] E. Al Nuaimi, H. Al Neyadi, N. Mohamed, and J. Al-jaroodi, "Applications of big data to smart cities," J. Internet Serv. Appl., 2015.
- [5] B. Thuraisingham, "Big Data Security and Privacy," Proc. 5th ACM Conf. Data Appl. Secur. Priv. - CODASPY '15, pp. 279–280, 2015.
- [6] A. Fernando, C. Santos, and I. P. Teles, "Big Data: A Systematic Review," pp. 501–506, 2018.
- [7] W. Inoubli, S. Aridhi, H. Mezni, and M. Maddouri, "An experimental survey on big data frameworks," Future Generation Computer Systems, vol. 86, Elsevier B.V., pp. 546–564, 2018.
- [8] C. L. P. Chen and C. Zhang, "Data-intensive applications , challenges , techniques and technologies : A survey on Big Data," Information Sciences, vol. 275, Elsevier Inc., pp. 314–347, 2014.
- [9] A. Mathur and C. P. Gupta, "Big Data Challenges and Issues : A Review," Springer International Publishing, pp. 446–452, 2020.
- [10] A. Siddiqa et al., "Journal of Network and Computer Applications A survey of big data management : Taxonomy and state-of-the-art," vol. 71, pp. 151–166, 2016.

References

- [11] A. Siddiqa, A. Karim, and A. Gani, "Big data storage technologies : a survey," vol. 18, no. 8, pp. 1040–1070, 2017.
- [12] L. Rabhi, N. Falih, A. Afraites, and B. Bouikhalene, "ScienceDirect ScienceDirect Big Data Approach and its applications in Various Fields : Review Big Data Approach and its applications in Various Fields : Review," *Procedia Computer Science*, vol. 155, no. 2018, Elsevier B.V., pp. 599–605, 2019.
- [13] J. P. D. Comput, H. Tran, and J. Hu, "Privacy-preserving big data analytics - A comprehensive survey," *Journal of Parallel and Distributed Computing*, vol. 134, Elsevier Inc., pp. 207–218, 2019.
- [14] I. A. Ajah and H. F. Nweke, "Big Data and Business Analytics: Trends, Platforms, Success Factors and Applications," *Big Data Cogn. Comput.*, vol. 3, no. 2, p. 32, 2019.
- [15] P. Grover, "Big Data Analytics : A Review on Theoretical Contributions and Tools Used in Literature," *Global Journal of Flexible Systems Management*, vol. 18, no. 3, Springer India, pp. 203–229, 2017.
- [16] H. Kashyap, H. Afzal, A. Nazrul, and H. Swarup, "Big data analytics in bioinformatics : architectures , techniques , tools and issues," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 5, no. 1, Springer Vienna, pp. 1–28, 2016.
- [17] T. R. Rao, P. Mitra, and R. Bhatt, "The big data system , components , tools , and technologies : a survey," *Knowledge and Information Systems*, vol. 60, no. 3, Springer London, pp. 1165–1245, 2019.