SESSION 6

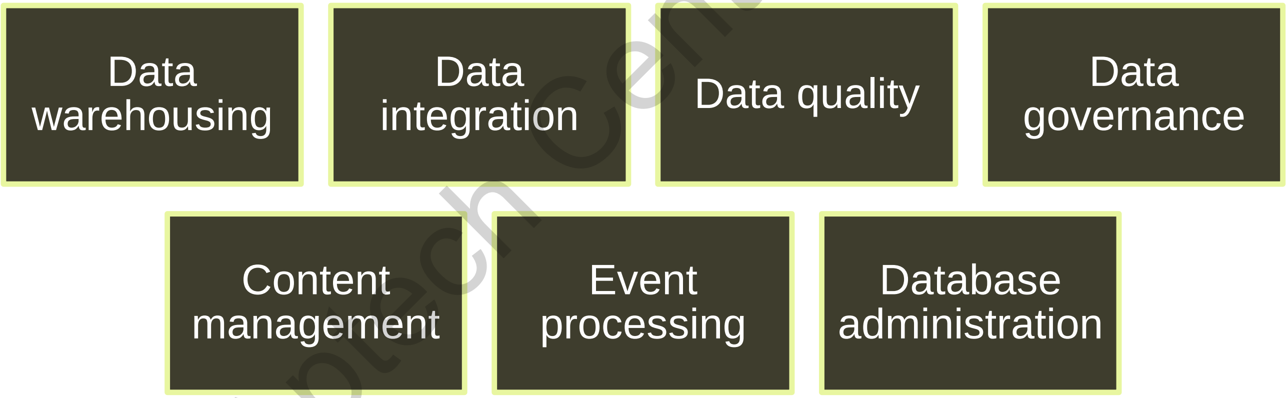
Data Security, Agility, and Clustering in Big Data

Objectives

- Describe data security in Big Data
- Explain data agility and clustering

Big Data Management (1-2)

- BDM is a huge challenge for organizations as the data is critical and requires high level of maintenance quality.
- It involves collection, storage, processing, and delivery of data.
- It encompasses the following data disciplines:



Data warehousing

Data integration

Data quality

Data governance

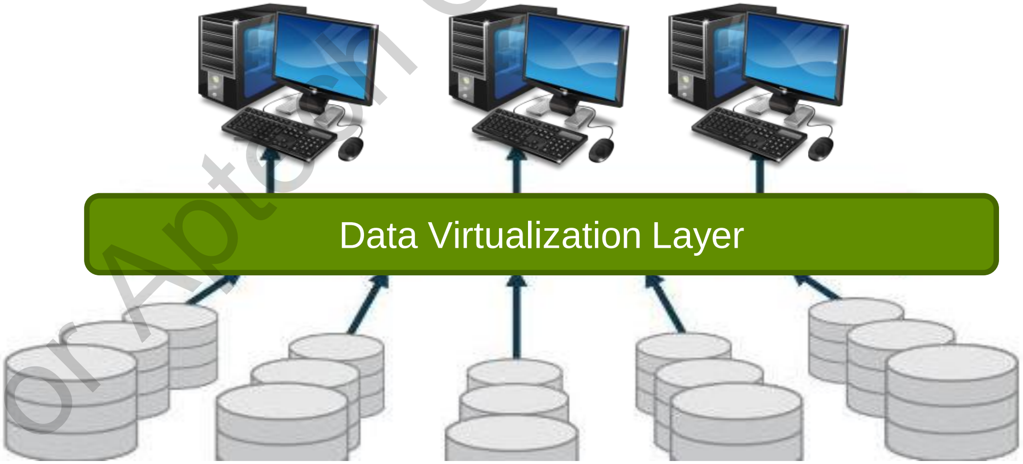
Content management

Event processing

Database administration

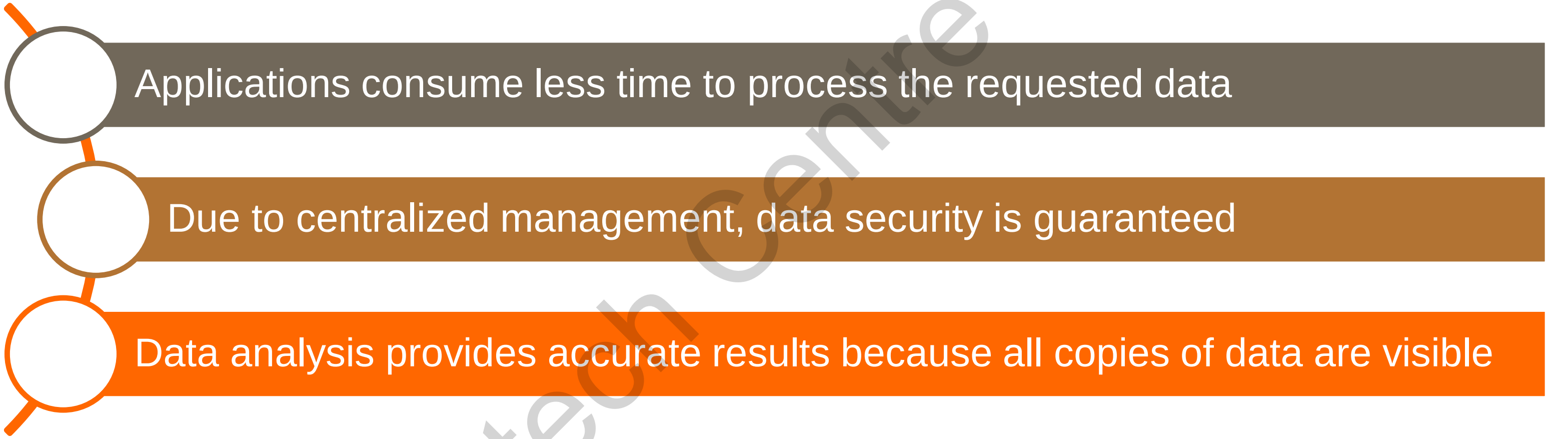
Big Data Management (2-2)

- Different BDM platforms and tools are available to manage Big Data.
- First step while managing Big Data is to reduce the amount of data and apply virtualization technology to virtualize the unique data set.
- After virtualization, multiple applications can reuse the same data footprint and smaller data footprint can be stored on any vendor-independent storage device.



Data Virtualization

- Reduces the database size by creating smaller footprints of data
- Due to smaller size, data management is improved in areas where:

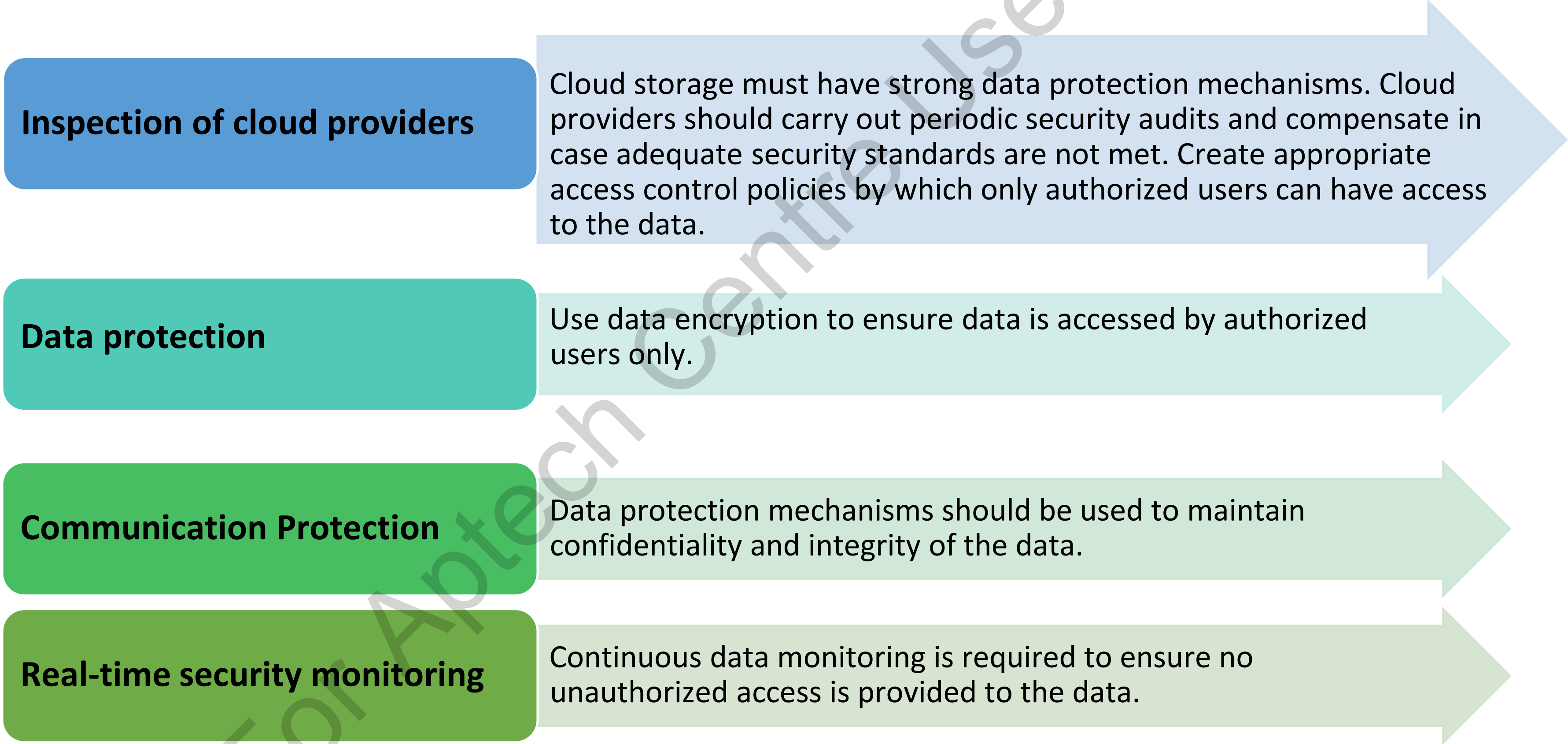


Applications consume less time to process the requested data

Due to centralized management, data security is guaranteed

Data analysis provides accurate results because all copies of data are visible

Security Challenges



Inspection of cloud providers

Cloud storage must have strong data protection mechanisms. Cloud providers should carry out periodic security audits and compensate in case adequate security standards are not met. Create appropriate access control policies by which only authorized users can have access to the data.

Data protection

Use data encryption to ensure data is accessed by authorized users only.

Communication Protection

Data protection mechanisms should be used to maintain confidentiality and integrity of the data.

Real-time security monitoring

Continuous data monitoring is required to ensure no unauthorized access is provided to the data.

Deploying Big Data Security

- Big Data provides replacements for log management systems of traditional SIEM and logging systems
- Big Data style analysis and more sophisticated pattern analysis enables to detect threat at an early stage
- To achieve data security, information can also be collected from physical security systems – CCTV
- Individual logs try to enhance Intrusion Detection Systems (IDS) and Intrusion Prevention Systems (IPS)

Data Agility

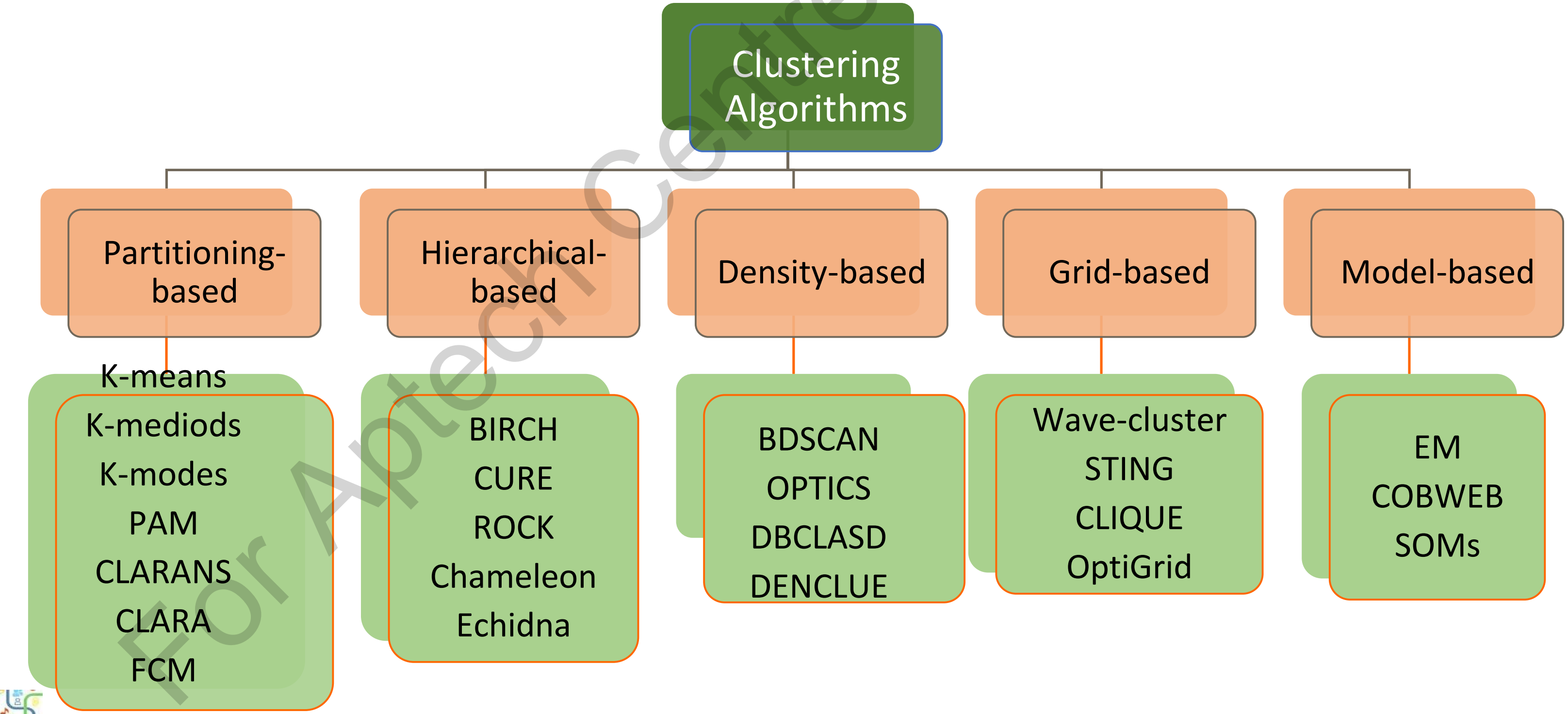
- Hadoop is the mainstream technology used to store and process Big Data at low cost
- Agility means how quickly data can be processed and output can be produced out of it
- Organizations have adapted new data exploration technologies such as Apache Drill:
  - □ It is an open-source framework with a low latency SQL query engine for Hadoop and SQL
  - □ It handles flat fixed schemas and built for semi-structured and nested data

Big Data Clustering

- Clustering methods are used to find similar data sets from Big Data
- Data clustering helps to divide data into meaningful groups
- There are two categories of Big Data clustering techniques:
  - Single machine clustering techniques
  - Multiple machine clustering techniques

Single Machine Clustering

- Data mining clustering algorithms enable users to use their tools to observe characteristics of each cluster
- Different clustering algorithms are as follows:



Multiple Machine Clustering

- In this, several multiple machine algorithms are used to perform clustering
- Two types of clustering methods are:

Parallel Clustering

- It divides the data partitions for distribution on different machines
- Speeds up the calculation and increases scalability

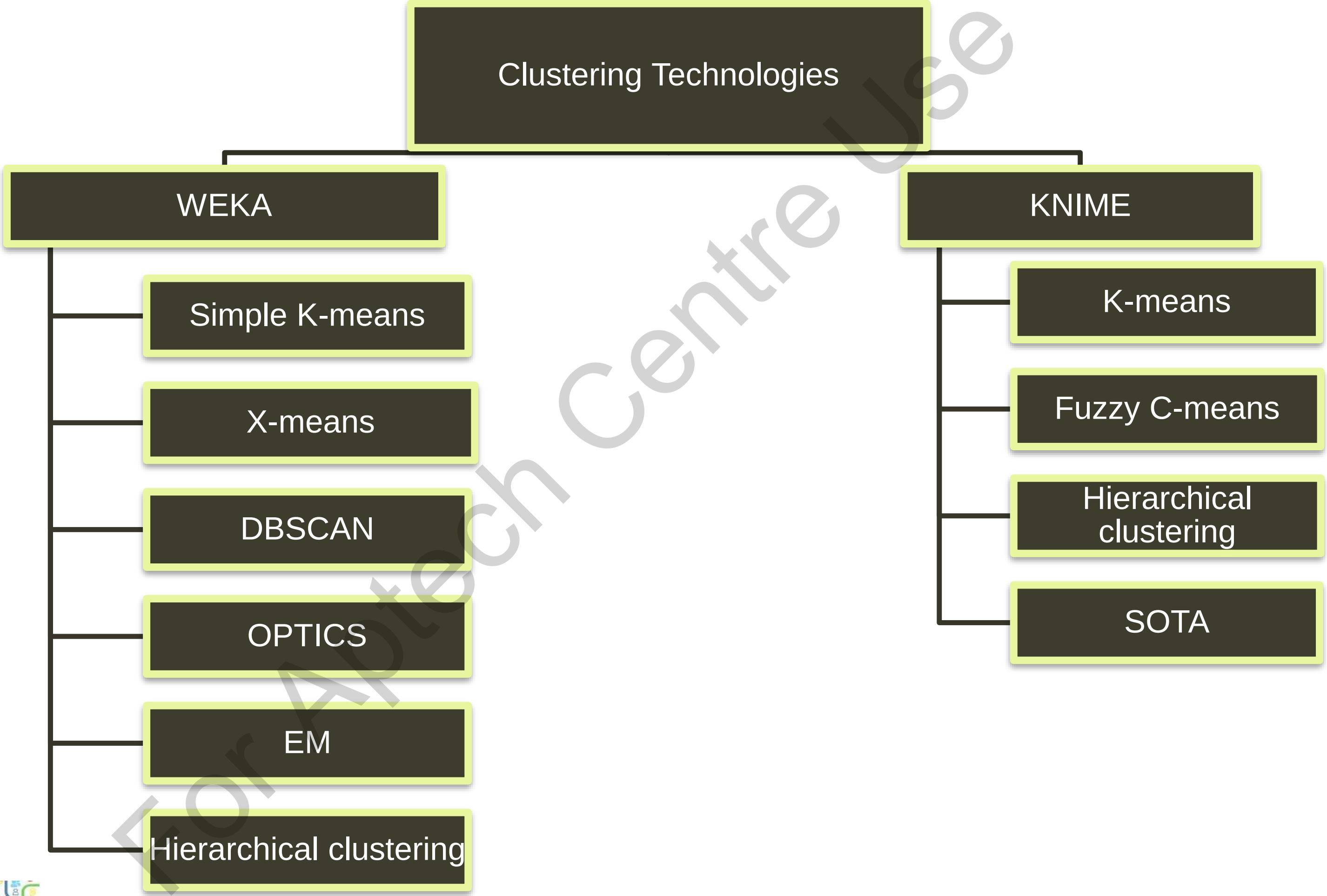
MapReduce-based Clustering

- It is a task partitioning mechanism for Big Data in a distributed environment
- The input data is first analyzed, cut into sub-tasks, and delegated to different servers
- The output from the servers is then collected and consolidated using the Reduce function

K-Means Method

- K-means is a popular clustering method, where K denotes the number of clusters
- The initial values of cluster centers are assigned and then each data item is assigned to the cluster and lastly, new centers for each cluster are calculated
- The K-means algorithm is easy to apply and cost effective
- Its disadvantage is that the data is too large to store in the main memory and accessed sequentially
- The X-means method is an extension of the K-means, which estimates the number of clusters

Clustering Technologies (1-2)



Clustering Technologies (2-2)

- Different applications to handle Big Data analytics:

Analytic databases

Data warehouse appliances

Columnar databases

NoSQL database

Distributed file systems

Summary (1-2)

- Big Data management is a big challenge for organizations because the data is critical and requires high level of maintenance quality while ensuring that data is easily accessible all the time.
- To ensure security and management of Big Data, organizations have to go beyond the relational databases and traditional data warehouse platforms to store non-transactional forms of data.
- Big Data management involves collection and storage, processing, and delivery of data.
- To ensure data security, organizations usually make sensitive user data anonymous and remove any unique identifiers for a user.

Summary (2-2)

- Big Data provides replacements for log management systems of traditional SIEM and logging systems.
- The Big Data style analysis and more sophisticated pattern analysis enables to detect threat at an early stage, thereby taking necessary preventive measures.
- Organizations have adapted new data exploration technologies, such as Apache Drill.
- Data clustering helps to divide data into meaningful groups. It can be done with the help of different clustering algorithms.