SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

Accredited by NAAC – UGC with 'A' Grade

Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

COURSE NAME : 19CS508 - BIG DATA ANALYTICS

III YEAR / V SEMESTER

Unit 2- **CLUSTERING AND CLASSIFICATION**

Topic 1 : Advanced Analytical Theory and Methods: Overview of Clustering

Clustering

- Cluster analysis, also known as clustering, is a method of data mining that groups similar data points together.
- The goal of cluster analysis is to divide a dataset into groups (or clusters) such that the data points within each group are more similar to each other than to data points in other groups.
- This process is often used for exploratory data analysis and can help identify patterns or relationships within the data that may not be immediately obvious.
- There are many different algorithms used for cluster analysis, such as k-means, hierarchical clustering, and density-based clustering.
- The choice of algorithm will depend on the specific requirements of the analysis and the nature of the data being analyzed.

Cont..

Properties of Clustering :

•**1. Clustering Scalability:** Nowadays there is a vast amount of data and should be dealing with huge databases. In order to handle extensive databases, the clustering algorithm should be scalable. Data should be scalable, if it is not scalable, then we can't get the appropriate result which would lead to wrong results.

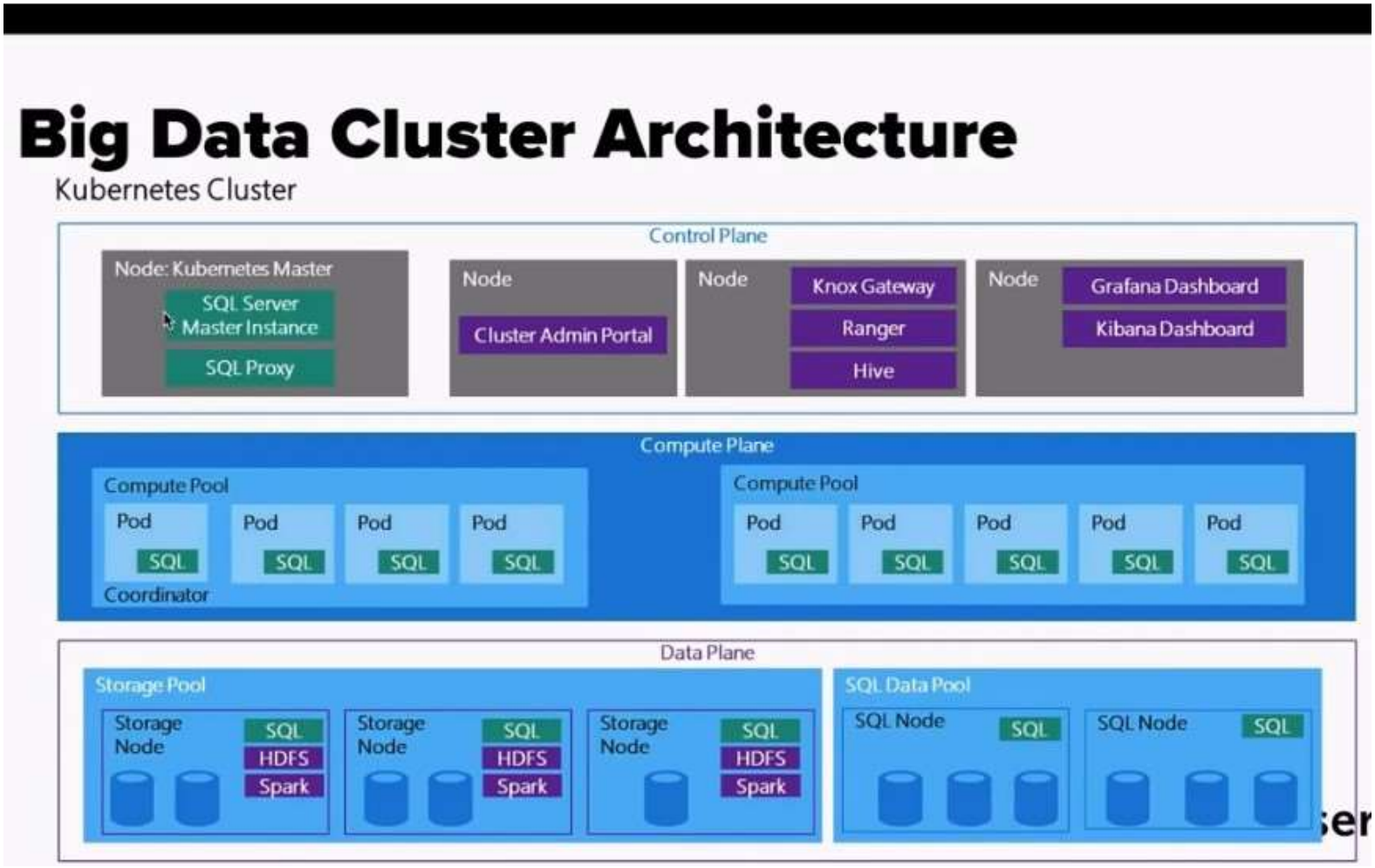
•**2. High Dimensionality:** The algorithm should be able to handle high dimensional space along with the data of small size.

•**3. Algorithm Usability with multiple data kinds:** Different kinds of data can be used with algorithms of clustering. It should be capable of dealing with different types of data like discrete, categorical and interval-based data, binary data etc.

•**4. Dealing with unstructured data:** There would be some databases that contain missing values, and noisy or erroneous data. If the algorithms are sensitive to such data then it may lead to poor quality clusters. So it should be able to handle unstructured data and give some structure to the data by organising it into groups of similar data objects. This makes the job of the data expert easier in order to process the data and discover new patterns.

•**5. Interpretability:** The clustering outcomes should be interpretable, comprehensible, and usable. The interpretability reflects how easily the data is understood.

Clustering Architecture



Clustering Methods

•**Clustering Methods:**

The clustering methods can be classified into the following categories:

1. Partitioning Method
2. Hierarchical Method
3. Density-based Method
4. Grid-Based Method
5. Model-Based Method
6. Constraint-based Method

CONT..

- **Applications Of Cluster Analysis:**
- It is widely used in image processing, data analysis, and pattern recognition.
- It helps marketers to find the distinct groups in their customer base and they can characterize their customer groups by using purchasing patterns.
- It can be used in the field of biology, by deriving animal and plant taxonomies and identifying genes with the same capabilities.
- It also helps in information discovery by classifying documents on the web.
- .

Activity

Advantages of Cluster Analysis:

1. It can help identify patterns and relationships within a dataset that may not be immediately obvious.

2. It can be used for exploratory data analysis and can help with feature selection.

3. It can be used to reduce the dimensionality of the data.

4. It can be used for anomaly detection and outlier identification.

5. It can be used for market segmentation and customer profiling.

Disadvantages of Cluster Analysis:

1. It can be sensitive to the choice of initial conditions and the number of clusters.

2. It can be sensitive to the presence of noise or outliers in the data.

3. It can be difficult to interpret the results of the analysis if the clusters are not well-defined.

4. It can be computationally expensive for large datasets.

5. The results of the analysis can be affected by the choice of clustering algorithm used.

6. It is important to note that the success of cluster analysis depends on the data, the goals of the analysis, and the ability of the analyst to interpret the results.

Assessment 1

1. List out the advantages of Cluster analysis
   a)_____
   b)_____
   c)_____
   d)_____

2. Identify the disadvantages of Cluster analysis
   a)_____
   b)_____
   c)_____
   d)_____

REFERENCES

1. Anand Rajaraman and Jeffrey David Ullman, "Mining of Massive Datasets", Cambridge University Press, 2012.
2. David Loshin, "Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph", Morgan Kaufmann/Elsevier Publishers, 2013
3. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data", Wiley publishers, 2015.

THANK YOU