BigData Analytics

Unit 3:
Understanding Data Processing Framework

Understanding Data Processing Framework

3.1. Principle features of Framework

3.2. Working of Framework

3.3. Techniques to optimize Framework Jobs

3.4. Uses of Data Processing Framework

3.5. Managing data in Ecosystem with ETL

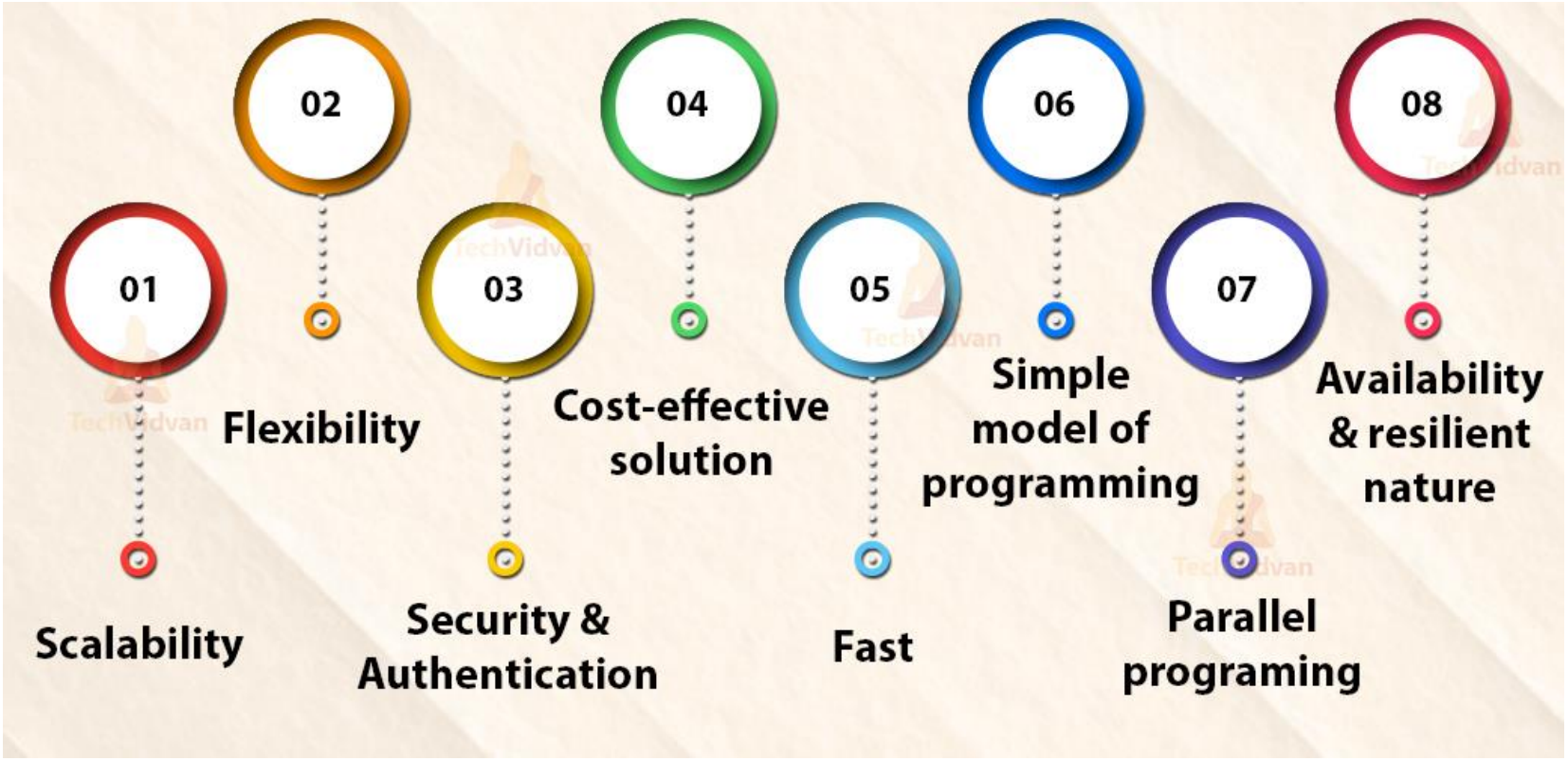
CE: MapReduce

Rewind…

- MapReduce keeps all the processing operations separate for parallel execution.
- Problems that are extremely large in size are divided into subtasks, which are chunks of data separated in manageable blocks.
- The subtasks are executed independently from each other and then, the results from all independent executions are combined to provide the complete output.

CE: 3.1.
Principle features of Framework

CE: 3.1. Features of MapReduce



CE: 3.1. Features of MapReduce (Conti…)

1. Scalability:

- Apache Hadoop is a highly scalable framework.
  - Because of *its ability to store and distribute huge data across plenty of servers*.

2. Flexibility:

- MapReduce programming enables companies to access new sources of data.
  - Because, *it enables companies to operate on different types of data*.

3. Security and Authentication:

- The MapReduce programming model *uses HBase and HDFS security* platform that allows access only to the authenticated users to operate on the data.
- Thus, *it protects unauthorized access to system data and enhances system security*.

CE: 3.1. Features of MapReduce (Conti…)

4. Cost-effective solution:

- Hadoop's scalable architecture with the MapReduce programming framework *allows the storage and processing of large data sets in a very affordable manner*.

5. Speed:

- MapReduce *can process huge unstructured data in a short time*.

6. Simple model of programming:

- This allows programmers *to develop the MapReduce programs which can handle tasks easily and efficiently*.

CE: 3.1. Features of MapReduce (Conti…)

7. Parallel Programming:

- It divides the tasks in a manner that allows their execution in parallel.
- The parallel processing *allows multiple processors* to execute these divided tasks. So the *entire program runs in less time*.

8. Availability and robust nature:

- Whenever the data is sent to an individual node, the same set of data is forwarded to some other nodes in a cluster.
- So, if any particular node suffers from a failure, then there are always other copies present on other nodes that can still be accessed whenever needed. This assures high availability of data.

9. Fault tolerance:

- One of the major features offered by *Apache Hadoop is its fault tolerance*.
- The Hadoop MapReduce framework has *the ability to quickly recognizing faults that occur*.

CE: 3.3.
Techniques to optimize Framework Jobs

CE: 3.5.
Managing data in Ecosystem with ETL

https://www.cleo.com/blog/knowledge-base-etl-integration#:~:text=The%205%20steps%20of%20the,the%20most%20important%20process%20steps.&text=Clean%3A%20Cleans%20data%20extracted%20from,the%20data%20prior%20to%20transformation.


Thank You