Big Data Processing Concepts

- Partitioning a large dataset into a smaller one can speed up processing.
- Big Data is often processed in parallel in a distributed fashion at the location in which it is stored.
- Important principle is the Speed, Consistency, and Volume (SCV) principle.

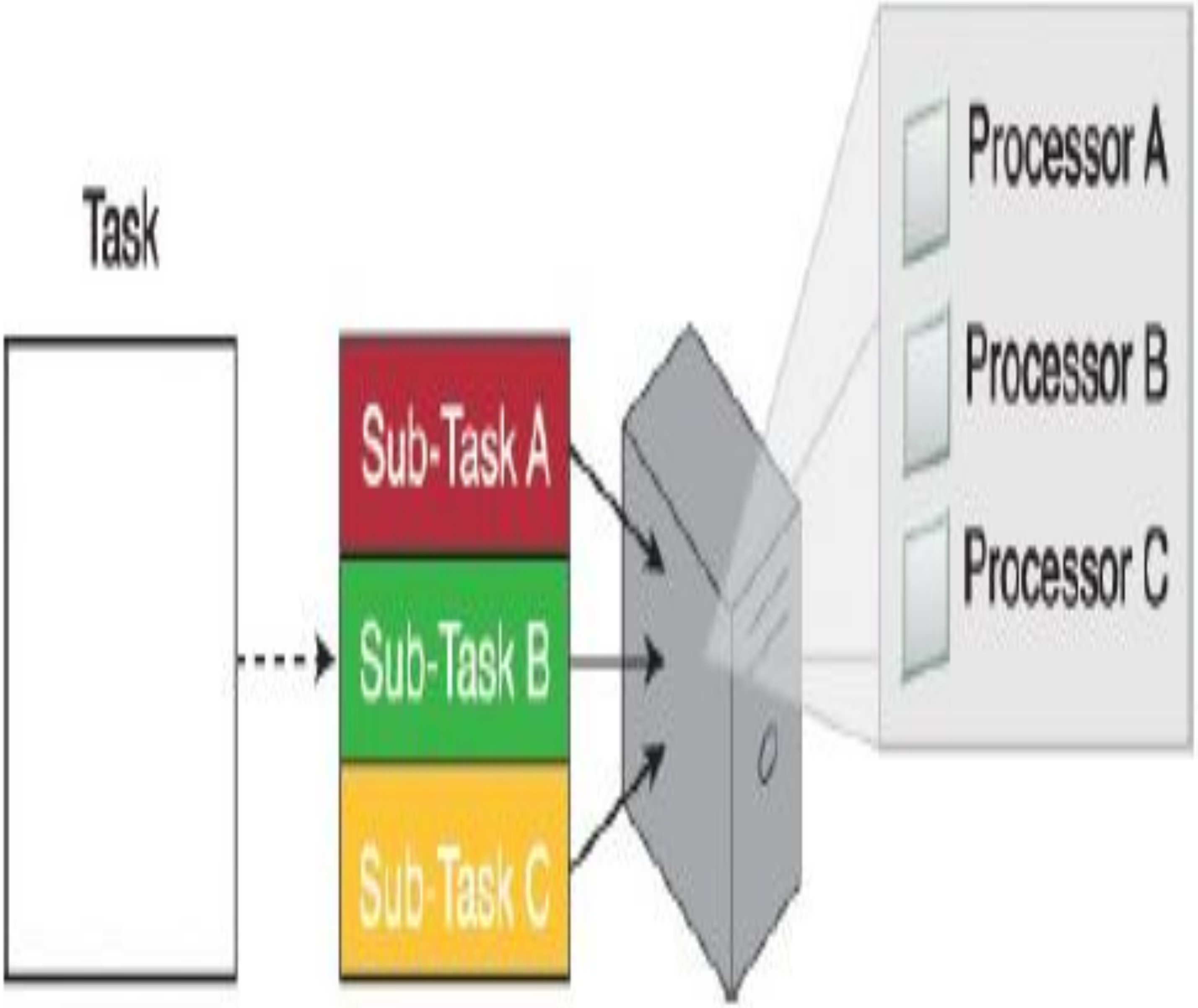
Important Concepts

- Parallel data processing
- Distributed data processing
- Hadoop
- Processing workloads
- Cluster

i. Parallel Data Processing

- It involve simultaneous execution of multiple sub-tasks that collectively comprise a larger task.
- The goal is to reduce the execution time by dividing a single larger task into multiple smaller tasks that run concurrently.

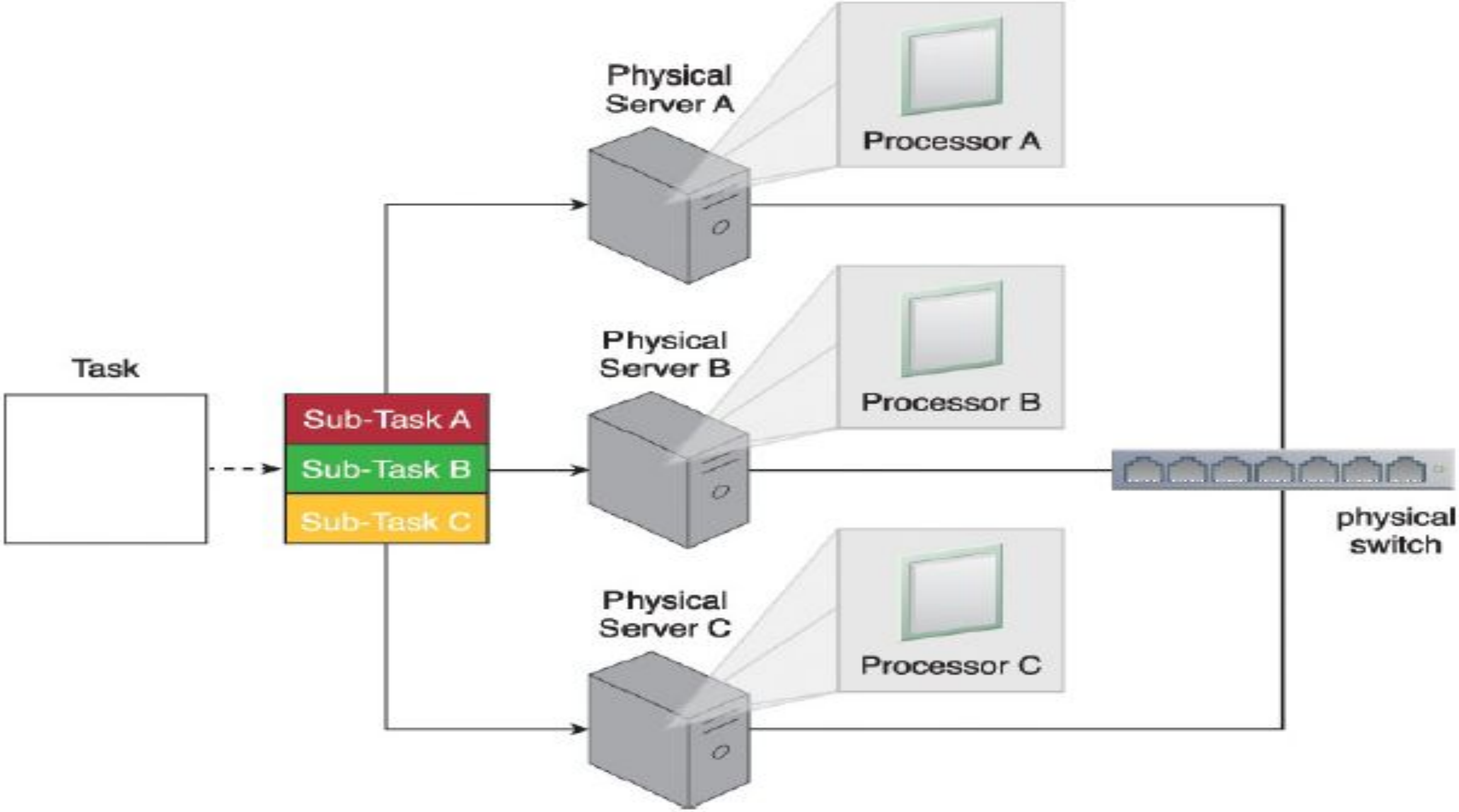
- It can be achieved through multiple networked machines.
- It is more typically achieved within the confines of a single machine with multiple processors or cores


Task
Sub-Task A
Sub-Task B
Sub-Task C
Processor A
Processor B
Processor C

ii. **Distributed Data Processing**

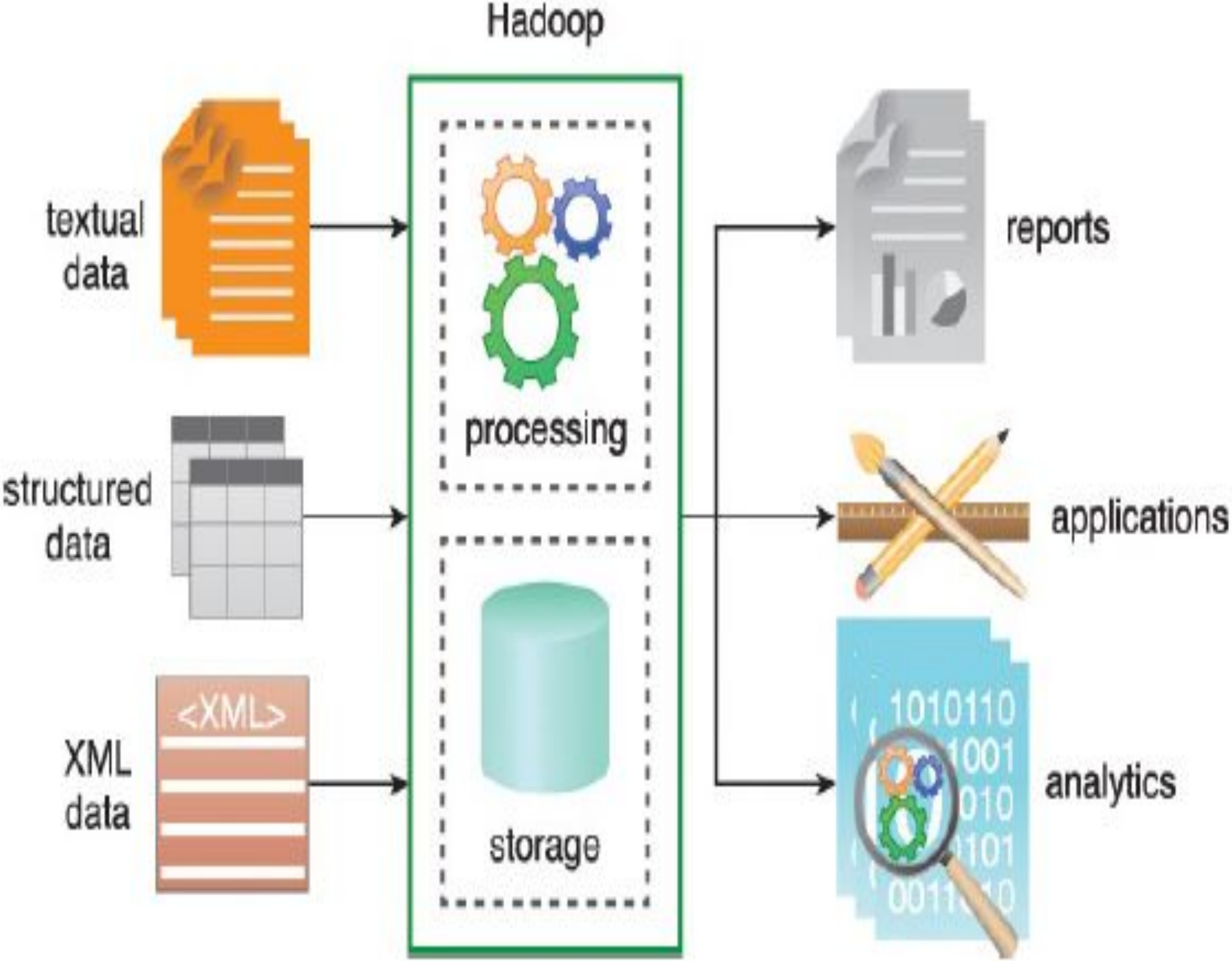
- It is related to parallel data processing.
- The principle of “divide-and-conquer” is applied.
- It is achieved through physically separate machines that are networked together as a cluster.

A task is divided into three sub-tasks that are then executed on three different machines sharing one physical switch.



iii. Hadoop

- It is an open-source framework for large-scale data storage and data processing.
- It has established itself as a de facto industry platform for contemporary Big Data solutions.
- It can be used as an ETL engine or as an analytics engine for processing large amounts of structured, semi structured and unstructured data.
- Hadoop implements the MapReduce processing framework


Hadoop
textual data
structured data
XML data
<XML>
processing
storage
reports
applications
analytics
1010110
1001
010
0101
0011 10

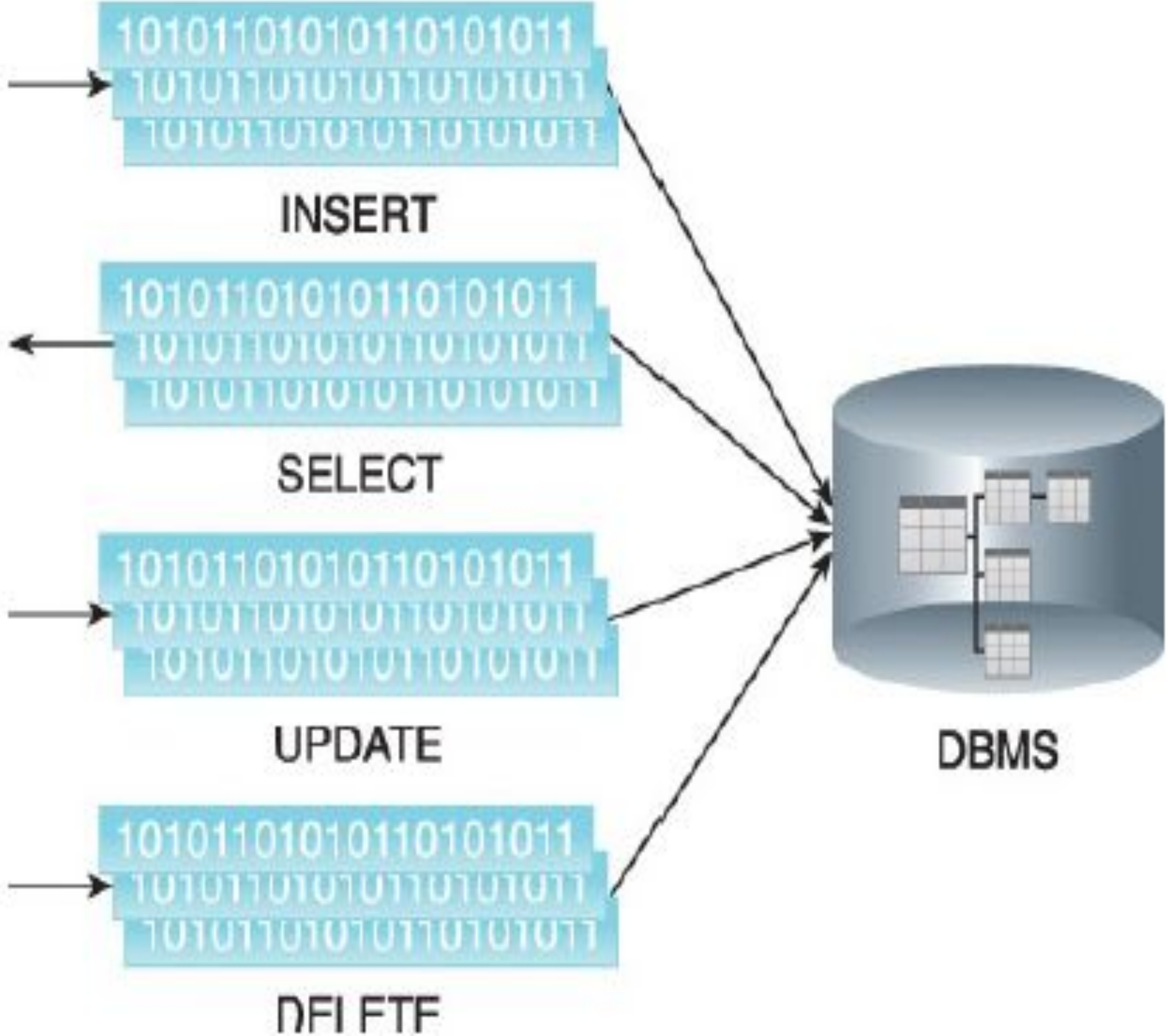
iv. Processing Workloads

- It is defined as the amount and nature of data that is processed within a certain amount of time.
- Workloads are usually divided into two types:
    - batch
    - transactional

i. Batch Processing

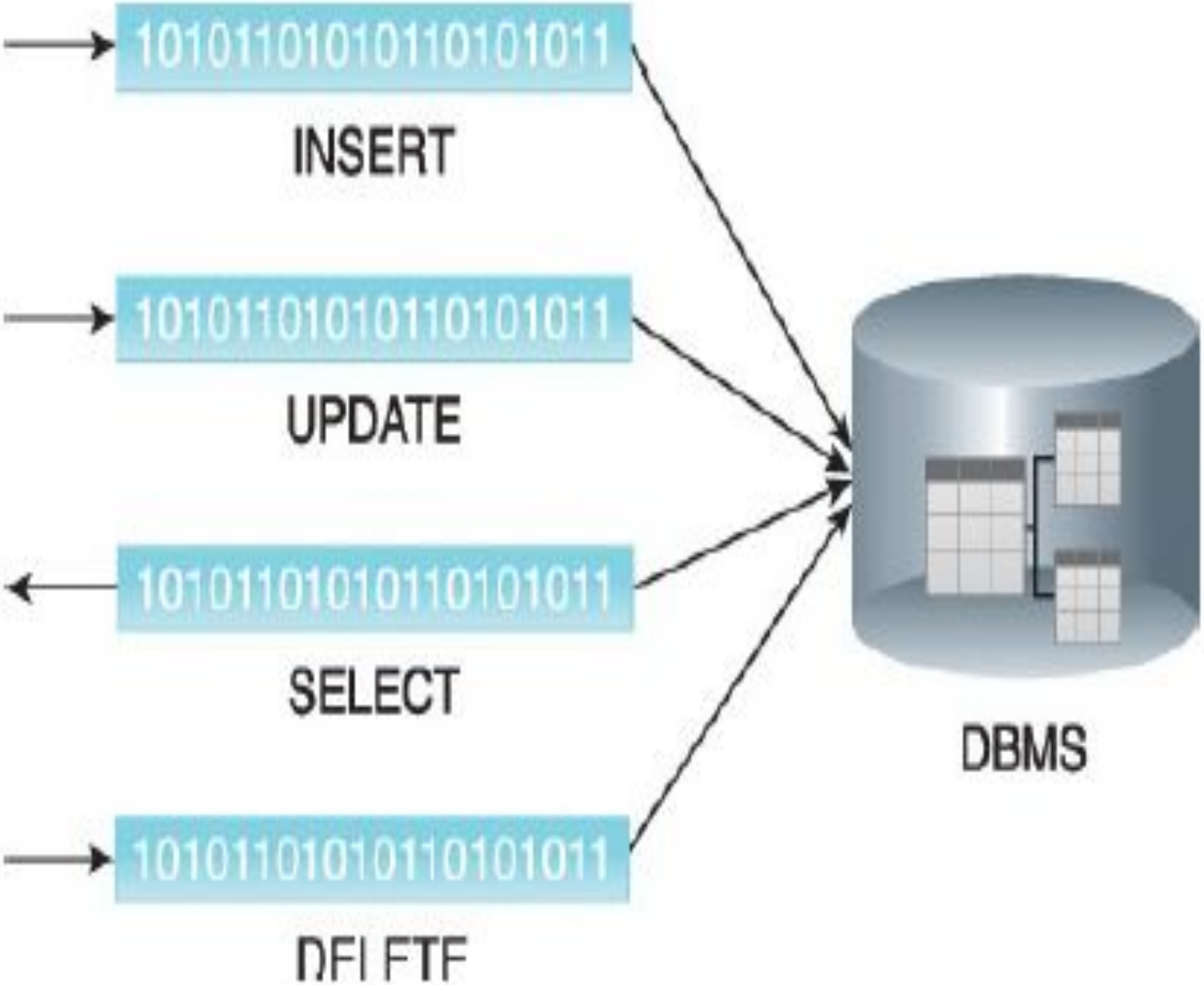
- Its also known as offline processing
- It involves processing data in batches
- Imposes delays
- Results in high-latency responses.
- Batch workloads involve large quantities of data with sequential read/writes and comprise of groups of read or write queries.

- Queries can be complex and involve multiple joins.
- OLAP systems commonly process workloads in batches.
- Strategic BI and analytics are batch-oriented as they are highly read-intensive tasks involving large volumes of data


INSERT
SELECT
UPDATE
DELETE
DBMS

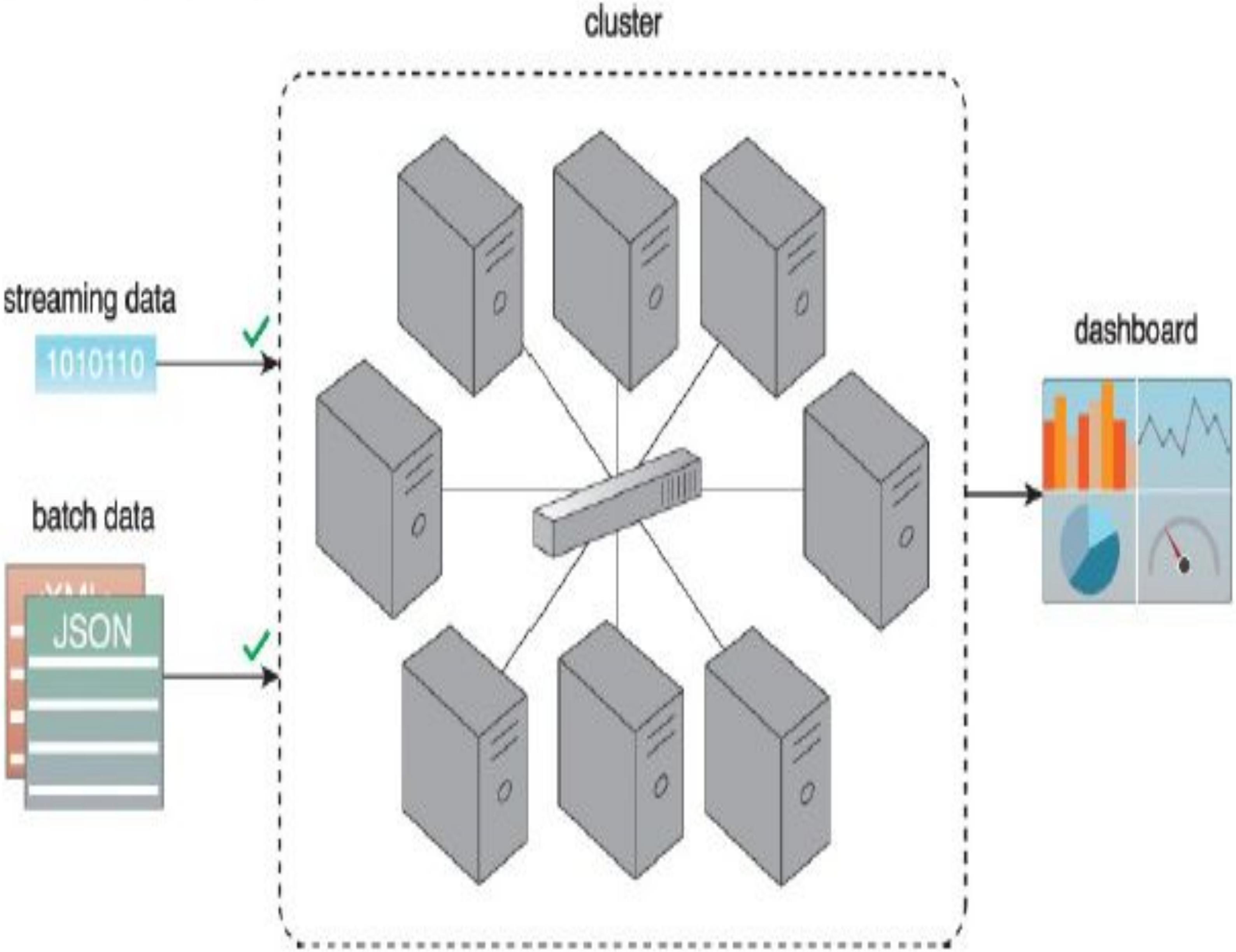
ii. Transactional processing

- Its also known as online processing.
- Data is processed interactively without delay
- Results in low-latency responses.
- Transaction workloads involve small amounts of data with random reads and writes.
- OLTP and operational systems, which are generally write-intensive, fall within this category.
- They are generally more write-intensive than read-intensive.


1010110101010110101011
INSERT
1010110101010110101011
UPDATE
1010110101010110101011
SELECT
1010110101010110101011
DELETE
DBMS

v. Cluster

- It provides mechanism to enable distributed data processing with linear scalability.
- They provide an ideal environment for Big Data processing and then processed in parallel in a distributed manner.
- When leveraging a cluster, Big Data datasets can either be processed in batch mode or real time mode


cluster
streaming data
1010110
batch data
XML
JSON
dashboard