Introduction to Big Data

ECAP456

Dr. Rajni Bhalla

Associate Professor

Learning Outcomes

After this lecture, you will be able to

- understand Programming Models For Big Data.

Popular Models

Popular Models for Big Data are:

- MapReduce,
- Directed Acyclic Graph,
- Message Passing,
- Bulk Synchronous Parallel,
- Workflow.

MapReduce

Process Huge
Amount of Data

MapReduce

Process Huge Amount of Data

In parallel

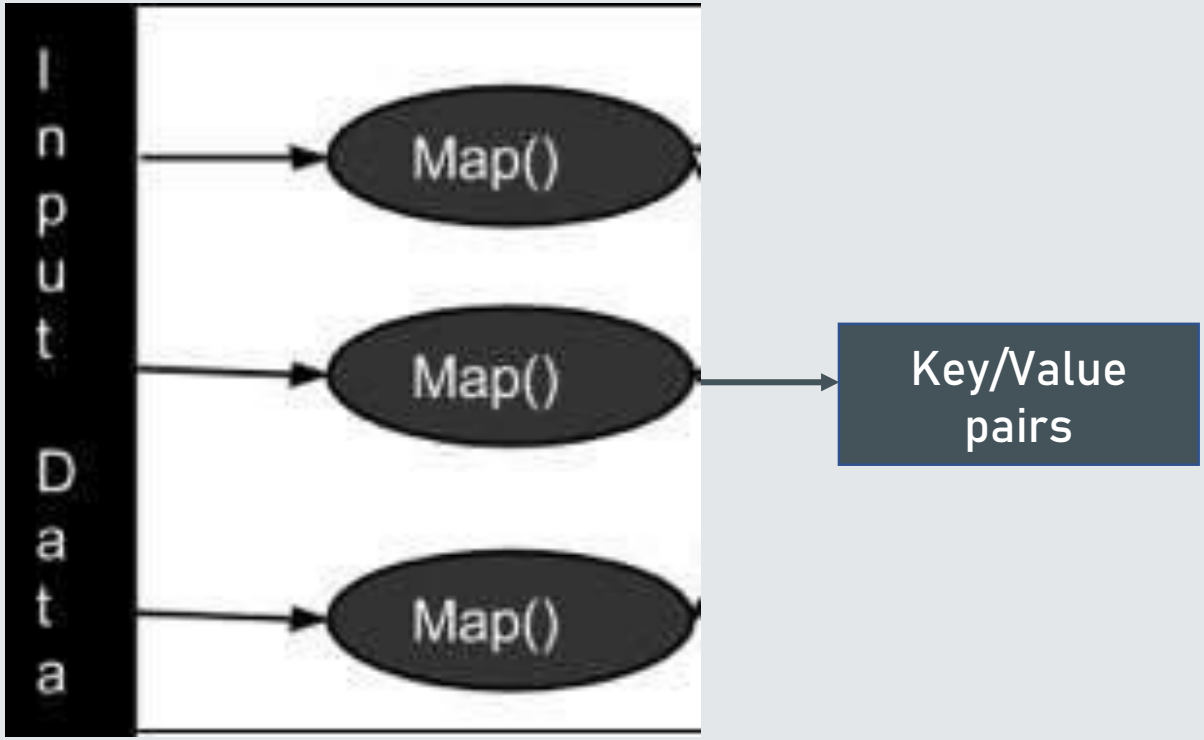
MapReduce

Process Huge Amount of Data

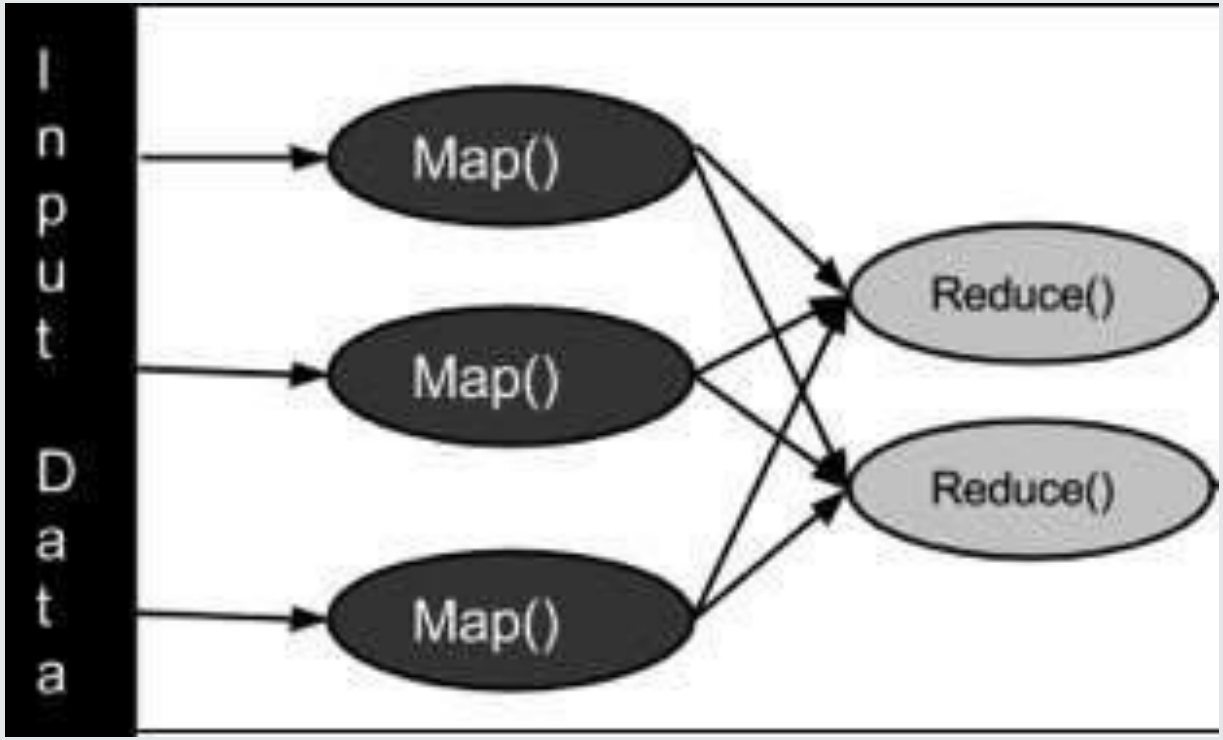
In parallel

On large clusters of commodity hardware

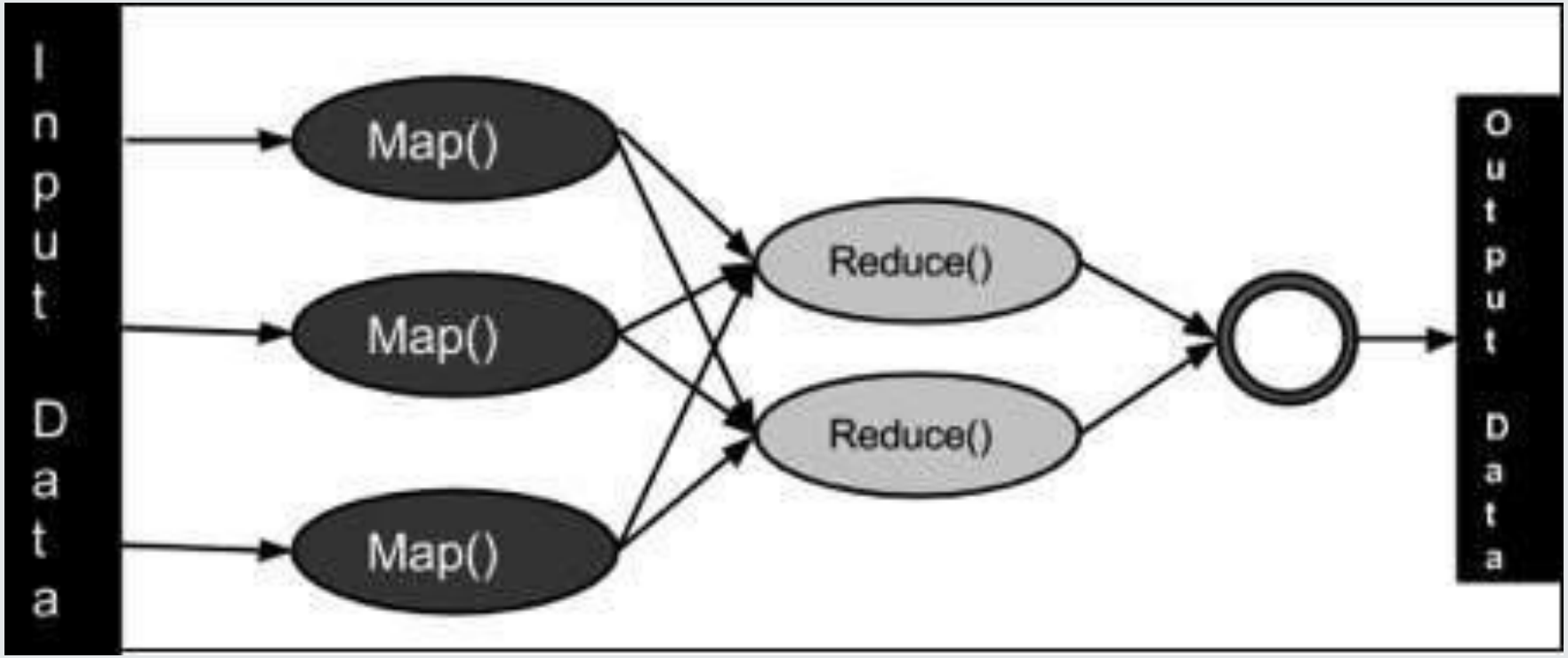
MapReduce



MapReduce



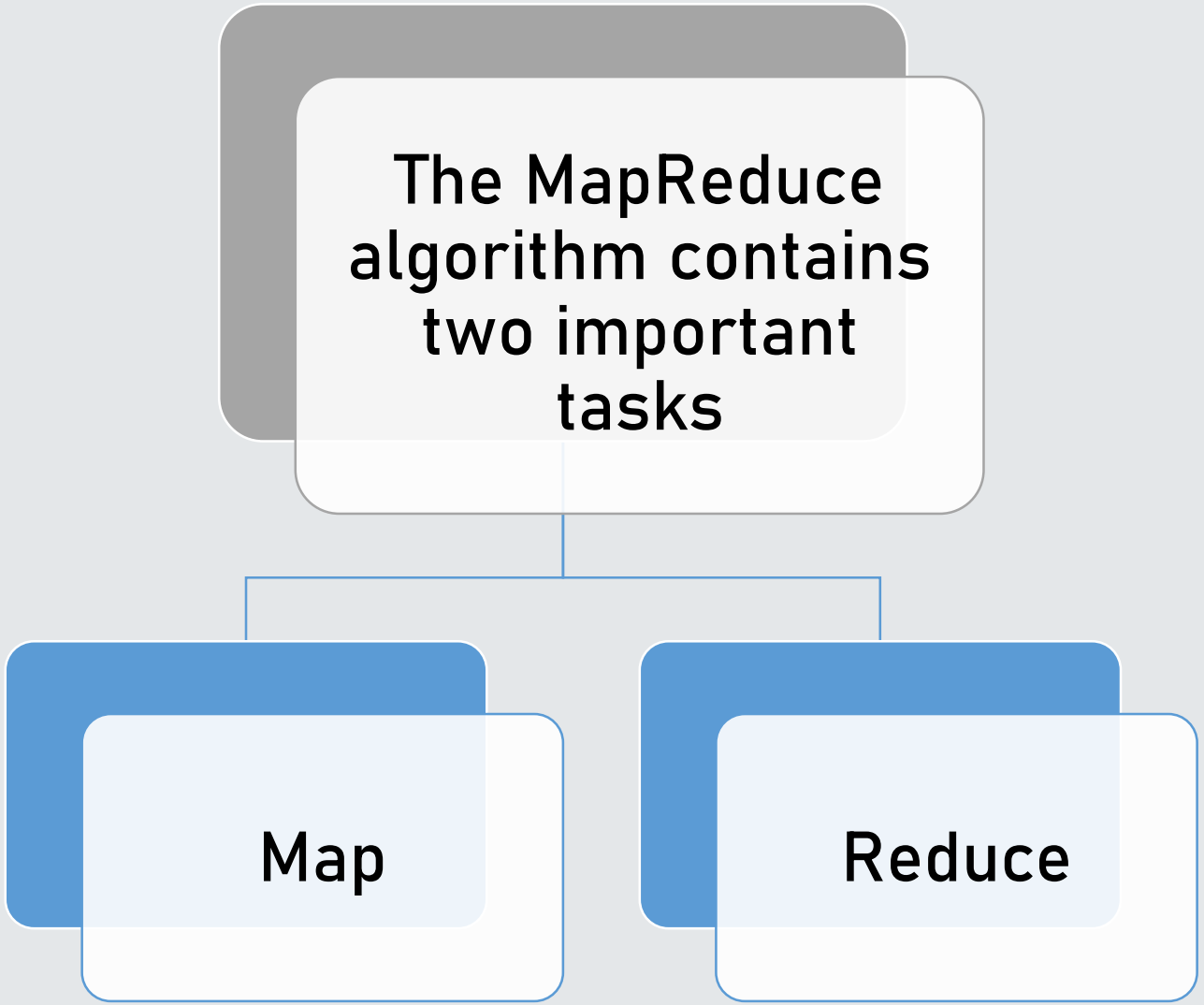
MapReduce



The Algorithm

- Generally MapReduce paradigm is based on sending the computer to where the data resides!
- MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

MapReduce


The MapReduce algorithm contains two important tasks
Map
Reduce

Advantage of MapReduce

- It is easy to scale data processing over multiple computing nodes.
- Under the MapReduce model, the data processing primitives are called mappers and reducers.
- Decomposing a data processing application into mappers and reducers is sometimes nontrivial.

Advantage of MapReduce

- But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change.
- This simple scalability is what has attracted many programmers to use the MapReduce model.

Directed Acyclic Graph,

Directed Acyclic Graph

Definition

In computer science and mathematics, a directed acyclic graph (DAG) refers to a directed graph which has no directed cycles.

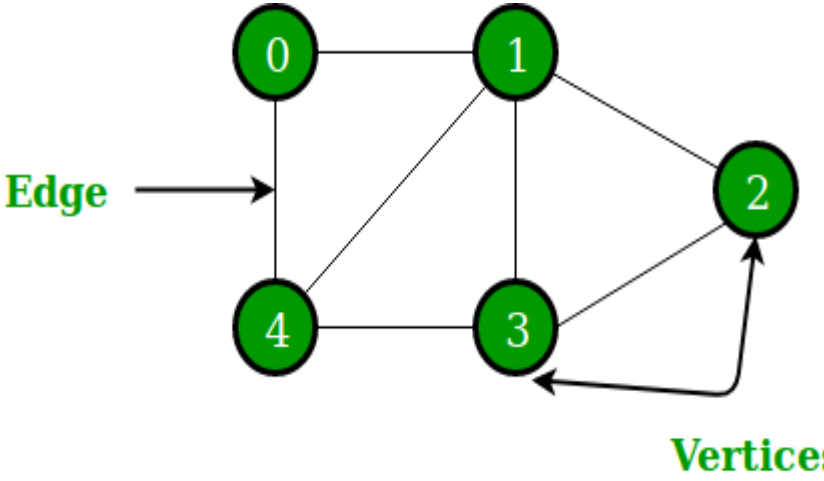
Explanation

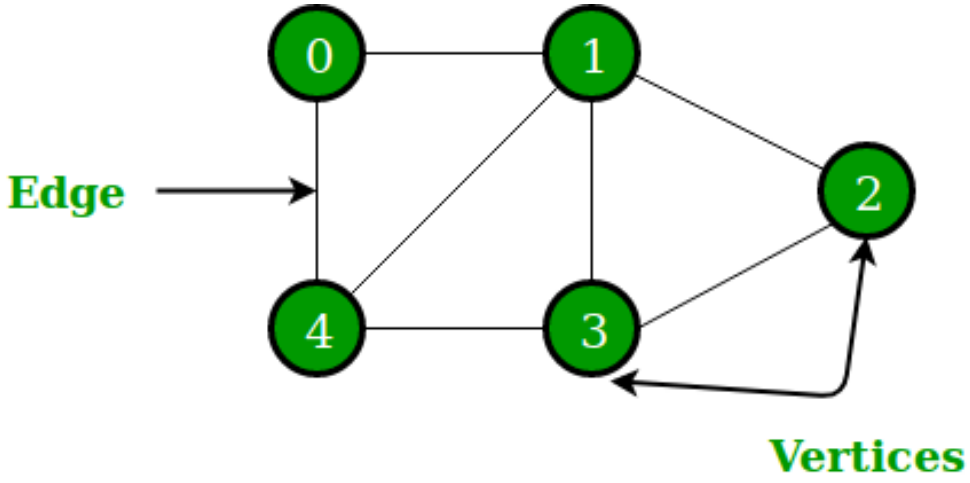
- In graph theory, a graph refers to a set of vertices which are connected by lines called edges.
- In a directed graph or a digraph, each edge is associated with a direction from a start vertex to an end vertex.

Explanation

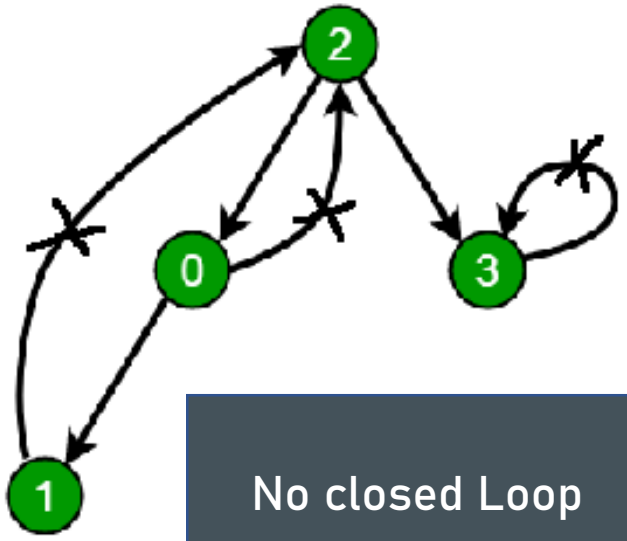
- If we traverse along the direction of the edges and we find that no closed loops are formed along any path, we say that there are no directed cycles. The graph formed is a directed acyclic graph.
- A DAG is always topologically ordered, i.e. for each edge in the graph, the start vertex of the edge occurs earlier in the sequence than the ending vertex of the edge.

Explanation



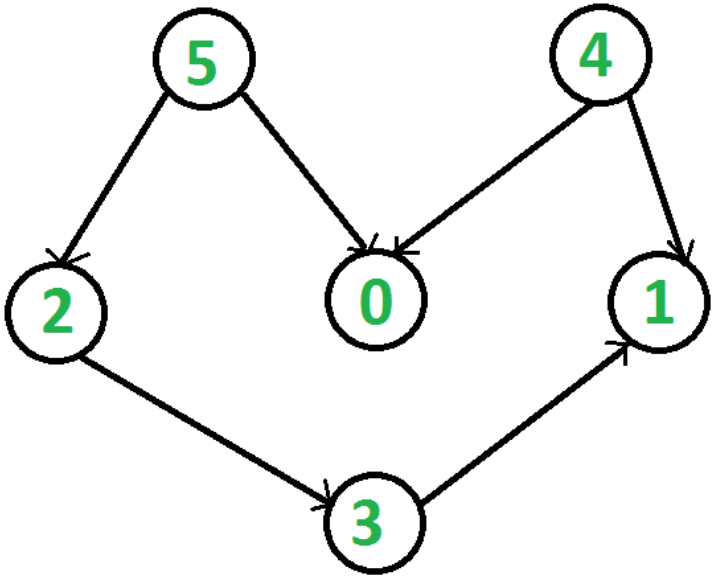


Directed Graph



No closed Loop

Directed Graph



Topological Order

Application Areas

Some of the main application areas of DAG are –

- Routing in computer networks.
- Job scheduling.
- Data processing.
- Genealogy.
- Citation graphs.

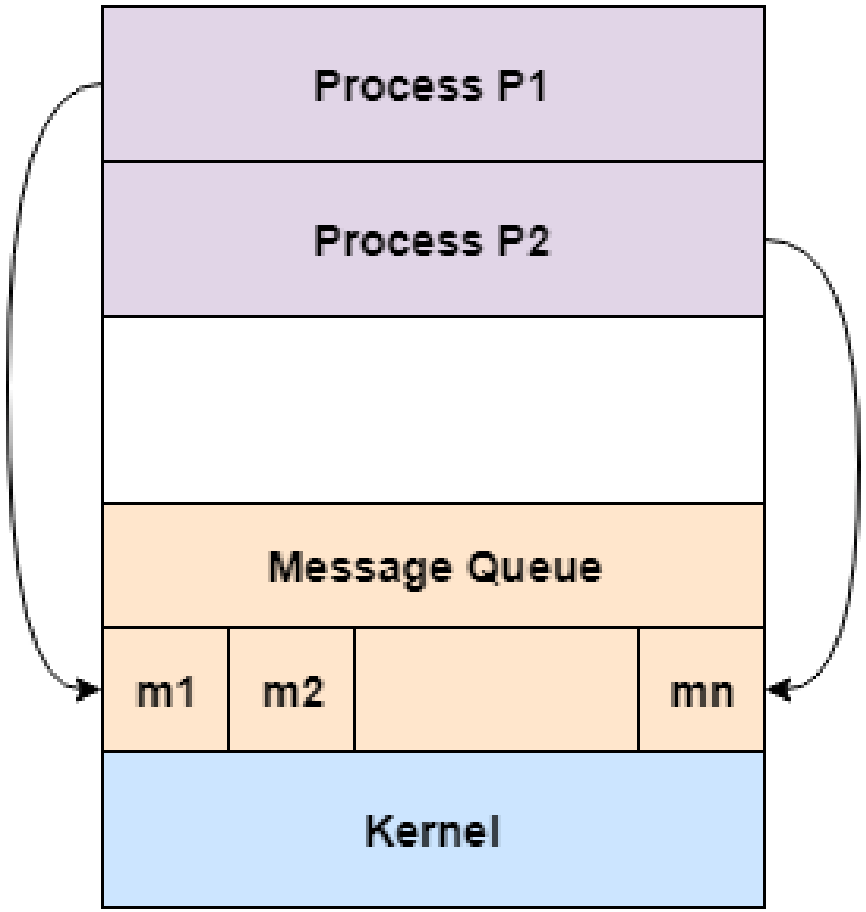
Message Passing

- Process communication is the mechanism provided by the operating system that allows processes to communicate with each other.
- This communication could involve a process letting another process know that some event has occurred or transferring of data from one process to another.
- One of the models of process communication is the message passing model.

Message Passing

- Message passing model allows multiple processes to read and write data to the message queue without being connected to each other.
- Messages are stored on the queue until their recipient retrieves them.
- Message queues are quite useful for inter-process communication and are used by most operating systems.

Message Passing



Message Passing Model

Advantages of Message Passing Mode

- The message passing model is much easier to implement than the shared memory model.
- It is easier to build parallel hardware using message passing model as it is quite tolerant of higher communication latencies.

Disadvantage of Message Passing Model

The message passing model has slower communication than the shared memory model because the connection setup takes time.

Bulk Synchronous Parallel

Bulk Synchronous Parallel

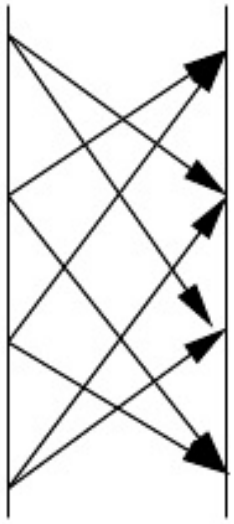
Another problem is that for graph algorithms such as DFS, BFS or Pert, MapReduce model is not satisfactory. For these scenarios, there is the BSP.

Bulk Synchronous Parallel

Superstep 0

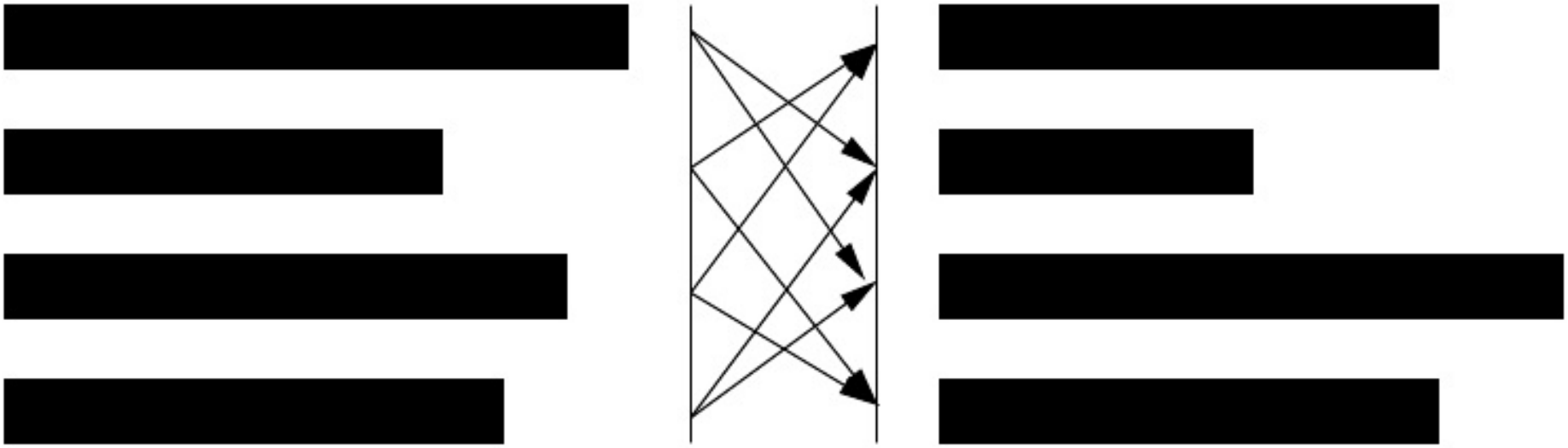
Bulk Synchronous Parallel

Superstep 0



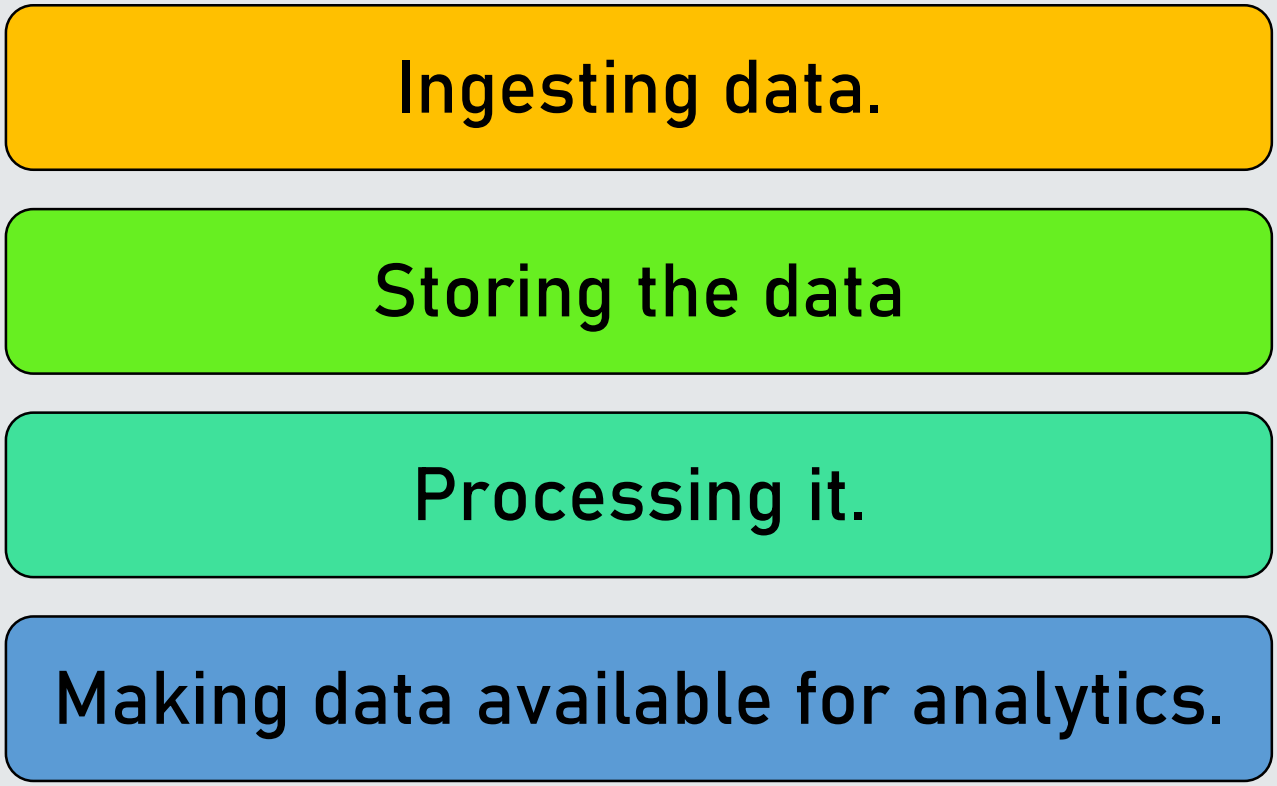
Sync

Bulk Synchronous Parallel



Five Reasons You Need a Step-by-Step Approach to Workflow Orchestration for Big Data

Workflow


Ingesting data.
Storing the data
Processing it.
Making data available for analytics.

Workflow

- The approach also requires having a reliable application workflow orchestration tool that simplifies the complexity of Big Data workflows, avoids automation silos, connects processes, and manages workflows from a single point.
- This allows you end to end automation, integration and orchestration of your Big Data processes, ensuring that everything is running successfully, meeting all SLAs, and delivering insights to business users on time.

Workflow

- Cobbling together disparate automation and orchestration tools that don't scale, may cause delays and put the entire project at risk.
- Here are some of the benefits of beginning your Big Data project with application workflow orchestration in mind and using a tool that supports these steps:

Workflow

- Improve quality, speed, and time to market.
- Reduce complexity in all environments – on premises, hybrid, and multi-cloud
- Ensure scalability and reduce risk
- Achieve better Integration
- Improve reliability
- Looking ahead

Workflow

- Improve quality, speed, and time to market
- Reduce complexity in all environments – on premises, hybrid, and multi-cloud.
- Ensure scalability and reduce risk
- Achieve better Integration
- Improve reliability
- Looking ahead

Workflow

- Improve quality, speed, and time to market
- Reduce complexity in all environments – on premises, hybrid, and multi-cloud
- Ensure scalability and reduce risk.
- Achieve better Integration
- Improve reliability
- Looking ahead

Workflow

- Improve quality, speed, and time to market
- Reduce complexity in all environments – on premises, hybrid, and multi-cloud
- Ensure scalability and reduce risk
- Achieve better Integration.
- Improve reliability
- Looking ahead

Workflow

- Improve quality, speed, and time to market
- Reduce complexity in all environments – on premises, hybrid, and multi-cloud
- Ensure scalability and reduce risk
- Achieve better Integration
- Improve reliability.
- Looking ahead

Workflow

- Improve quality, speed, and time to market
- Reduce complexity in all environments – on premises, hybrid, and multi-cloud
- Ensure scalability and reduce risk
- Achieve better Integration
- Improve reliability
- Looking ahead.


That's all for now...