1. Preparation

- Cluster Requirements:
    - Running Hadoop Services: Ensure Yarn resource manager, Namenode, Datanodes, and other crucial services are operational within the cluster.
    - Version Compatibility: Confirm your MapReduce application's version compatibility with the cluster's Hadoop version.
- Application Development:
    - Map and Reduce Functions: Write your Map and Reduce functions in Java, Python, or another supported language. Implement the logic for processing data in the Map phase and aggregating/combining data in the Reduce phase.
    - Job Configuration: Specify the input and output paths for the data, the number of reducers (optional), compression codecs (optional), and other relevant settings.

2. Job Submission

- Command-Line Submission: Use the hadoop jar command, mentioning the JAR file containing your compiled application, the main class to run, and the configuration arguments you defined.
- Alternative Methods: Consider tools like YARN web UI or client APIs for a graphical or programmatic interface to submit and manage jobs.

3. Job Execution Flow

- Job Scheduling: The JobTracker/Resource Manager in Yarn takes charge, dividing the input data into splits and assigning them to individual Map tasks on available nodes.
- Map Phase: Each Map task processes its assigned split, invoking your Map function on

each record within the split. This function generates key-value pairs as output, representing intermediate results.

- Shuffle and Sort: The key-value pairs are shuffled across the cluster based on their keys (hashing function used) and sorted within each reducer's input for efficient grouping.
- Reduce Phase: Reduce tasks receive groups of key-value pairs with the same key. Your Reduce function is invoked on each group of values, aggregating or combining them to produce final output.

4. Monitoring and Analysis

- Job Progress Tracking: Utilize cluster web UIs, command-line tools like mapred job -jobid <job_id> -status, or APIs to monitor the progress of Map and Reduce tasks, resource utilization, and overall job completion.
- Output Analysis: Access the output files stored in HDFS (typically) and analyze the results based on your desired insights.