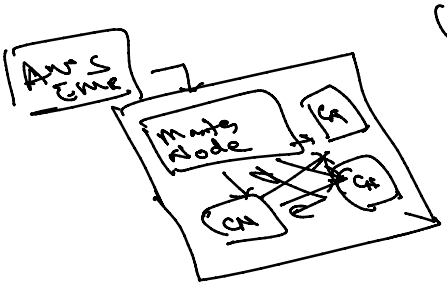
Day - 16 (11th June)
11 June 2024 10:02

① Amazon EMR (Elastic Map Reduce) :-

① Managed cluster platform that simplifies running of Hadoop Big data framework.



① Master/Primary Node :- → coordinating b/w core node. Hive PIG Sparks Hadoop Distributed File system
↓
Allot core of task or storage

② Core node :- store data (HDFS) & run the task.

③ Task node :- Run the task. It is not store any data.

YARN (Yet Another Resource Negotiator)

① Help the AWS EMR to Allocate the Resources
② Coordinate & schedule task among the node of the cluster.

Cluster Resource manager :- Mange the cluster & scheduling.

Hadoop :- Distributed processing. Process Huge Amount of Data.
HDFS

Sparks :- Distributed Processing Run of RAM & Cache memory. It bit costlier with Hadoop. Faster as compared to Hadoop.

Cost → Spot Instance.

Adv. of EMR :- (1) Low Cost → Spot Instance. (50 – 80%)

(2) Scalability → Scale in & scale out

(3) Integration with other AWS Services
(1) S3

(4) Security :- Amazon EMR integrate with AWS IAM for access control

Spark :- (1) Distributed Processing
(2) Run on the RAM/Cache memory

Architecture :-
(1) Driver :- Central Component of Spark App. Run main & high level fun
(2) Executor :- Execute task
(3) Cluster manager :- Manage resource of the cluster
(4) Distributed Storage :- Amazon S3, Apache Cassandra, HDFS