**Analytics**

1. Kinesis
2. Elastic Map Reduce (EMR)

**Kinesis**

1. Kinesis is a service for processing big data. Here data comes in rapidly and need quick processing
2. Provides ability for providers to send data in parallel from multiple sources. Kinesis consumes the data and streams it for the users sequentially. Each blog have its ID and using that it streams data sequentially to users. It reads the data blob by blob
3. Data will be there in kinesis for 24 hours and if needed we can extend it to max of 7 days. After 7 days we can move it to S3 if needed. Data have server side encryption
4. Used for data processing

**EMR**

1. EMR is a cluster which has a connector to connect with Kinesis. Hadoop application uses data from kinesis. Hadoop is a framework for processing big data
2. Take data from S3, load it to cluster, process it and generate output which can be stored back to a data store (eg. Redshift)

**Persistent Cluster and transitive cluster**

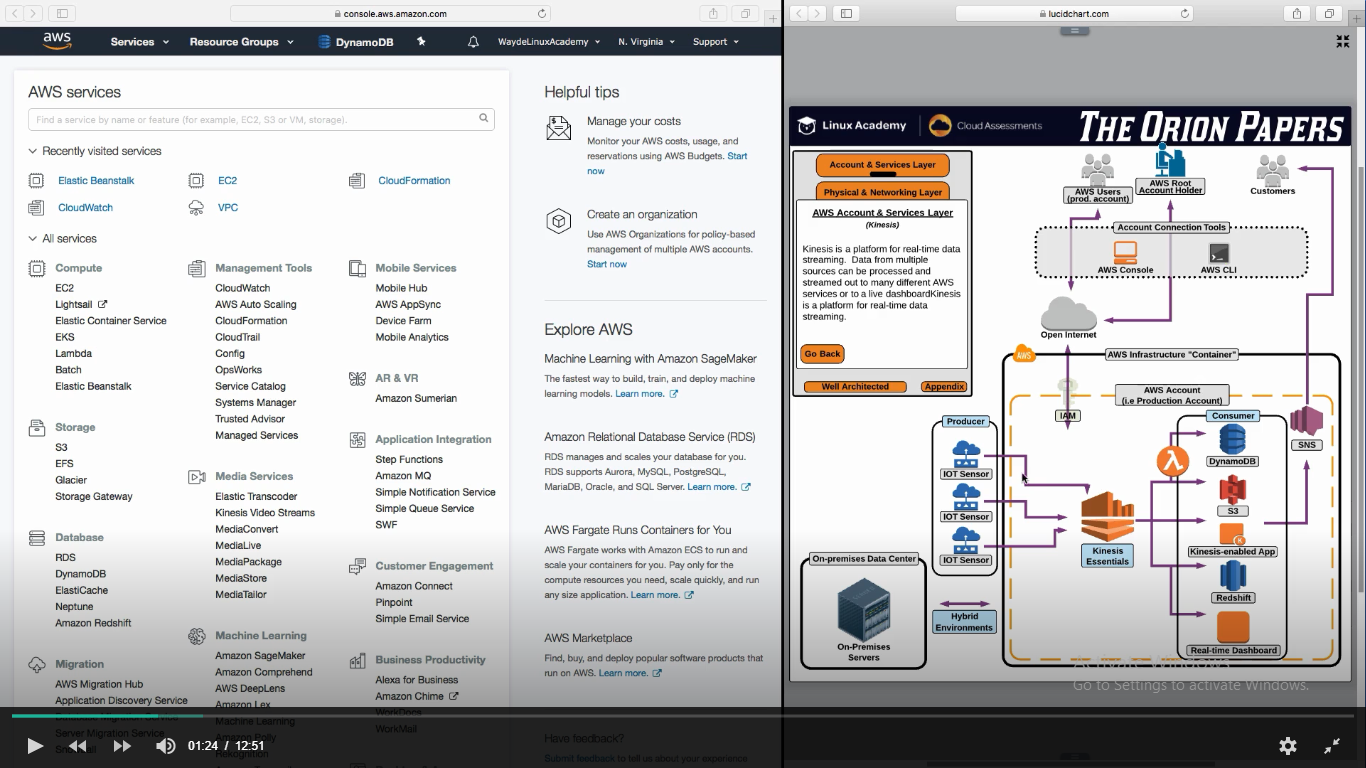
Clusters are scalable on the fly

**Persistent Cluster:**

1. It runs all the time

**Transient cluster:**

1. When we design the cluster we can give a set of task to be performed on the cluster. Once it finish all the step, it terminates the cluster

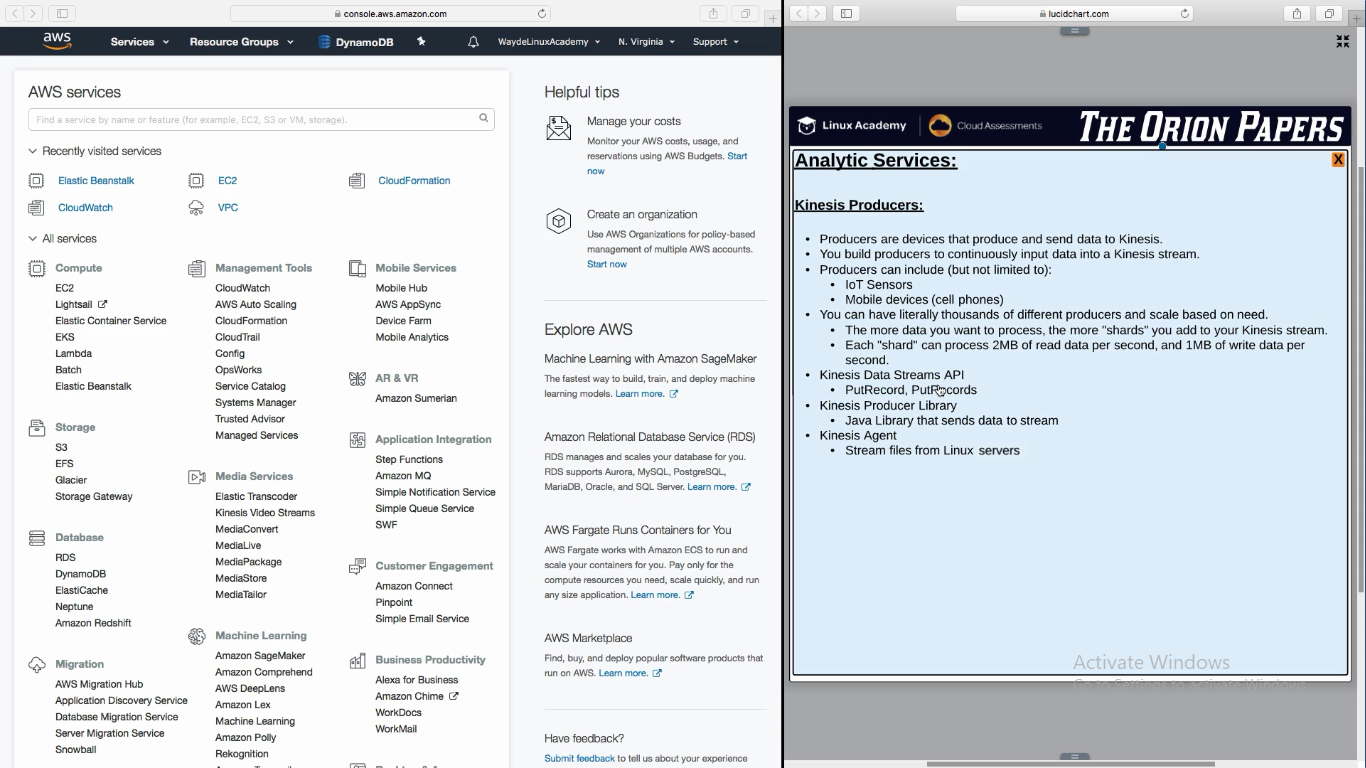


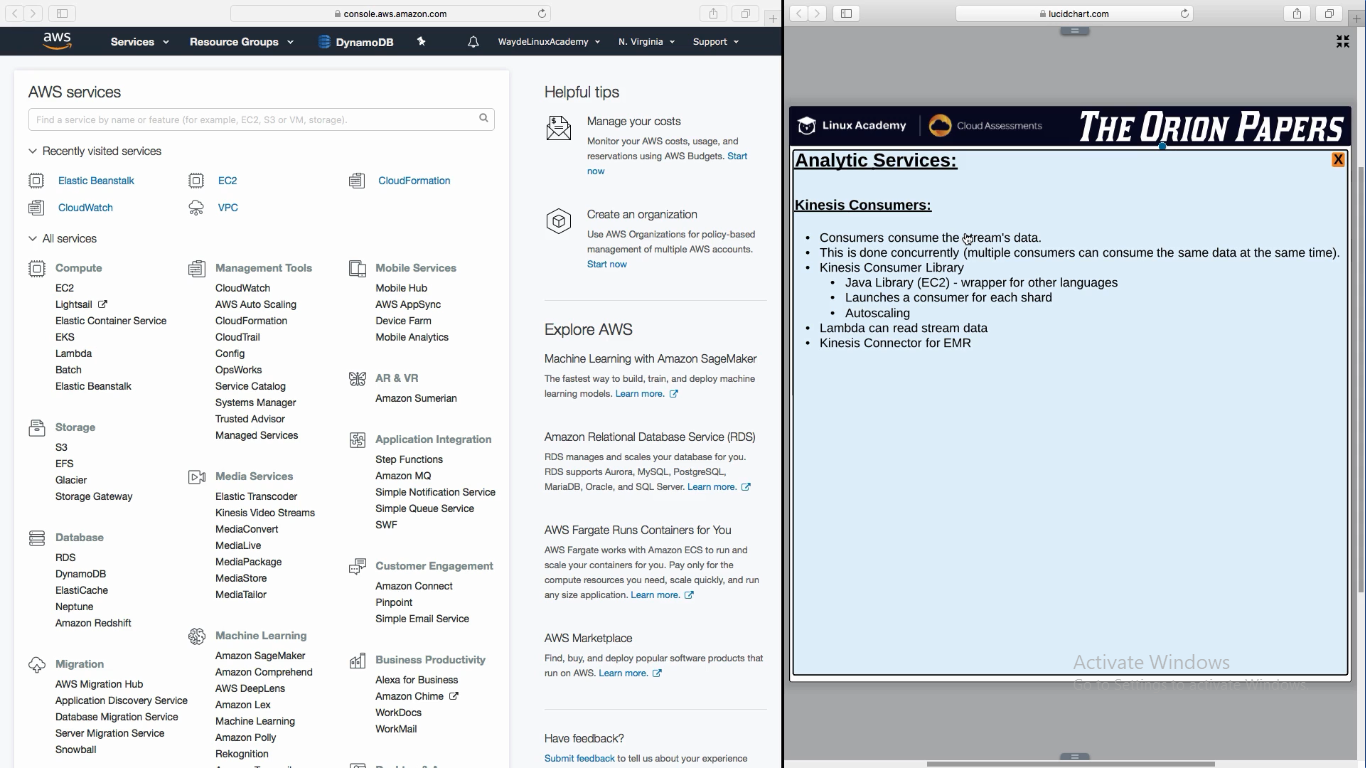


Distribute Data to shards using a partition key

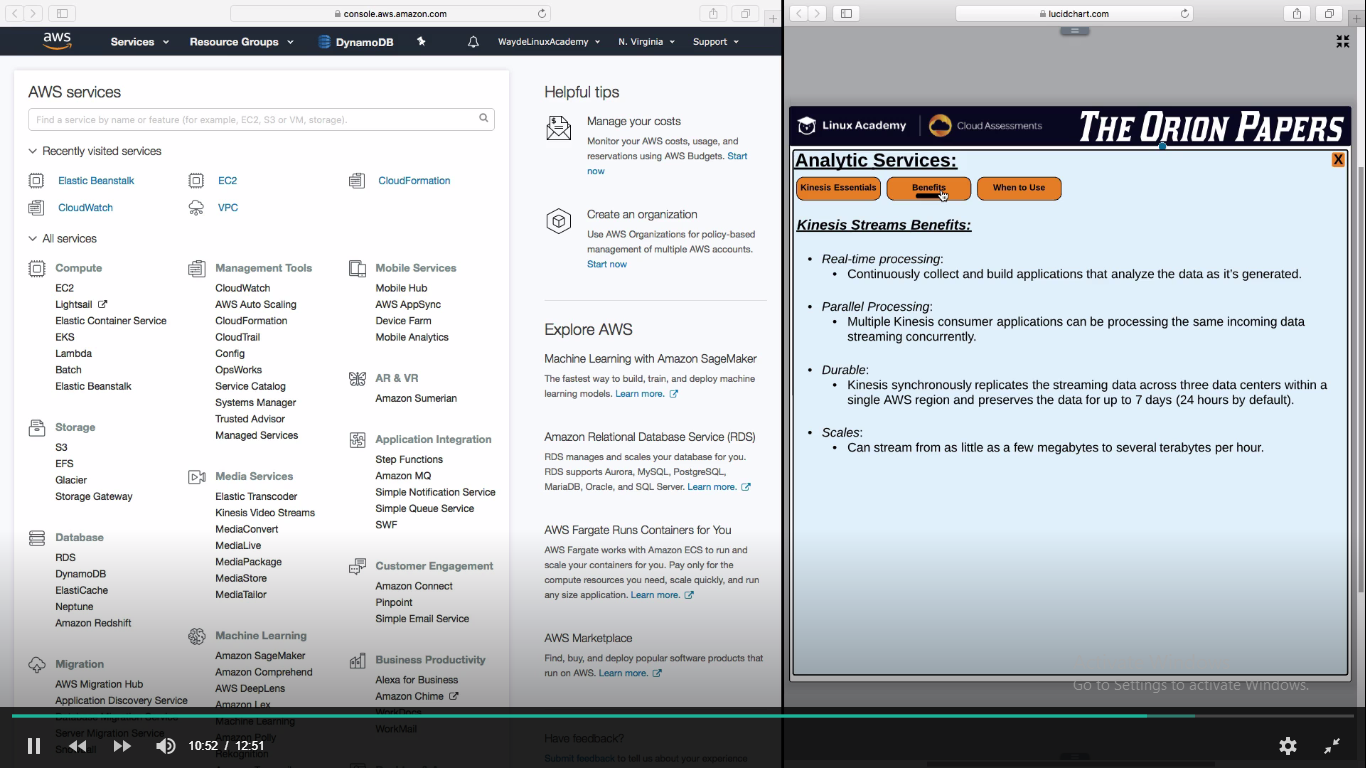
If multiple providers are providing data to kinesis and each data is specific for a consumer, we have to supply a partition id to the shard (Each block of data in called shard).

How to produce data and feed it to kenesis?





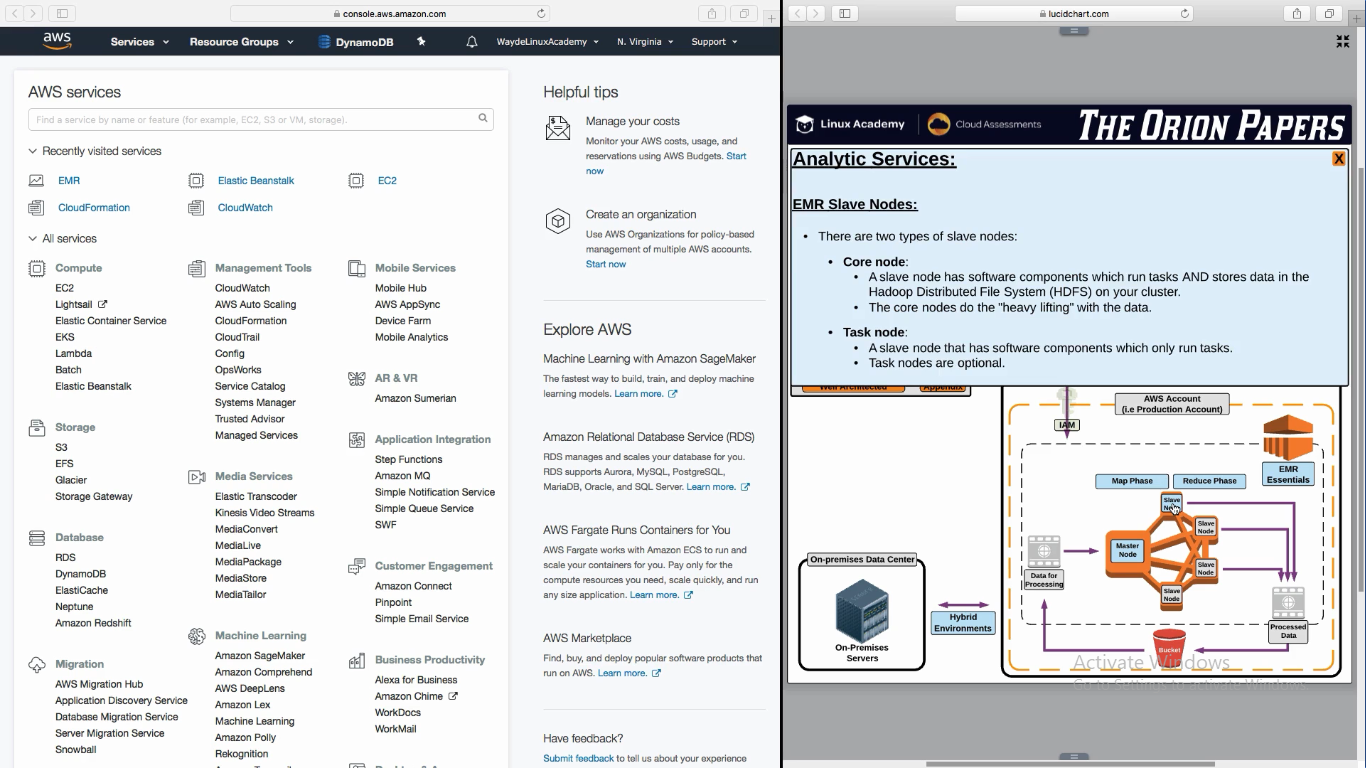
**Benefits of Kinesis Streams**

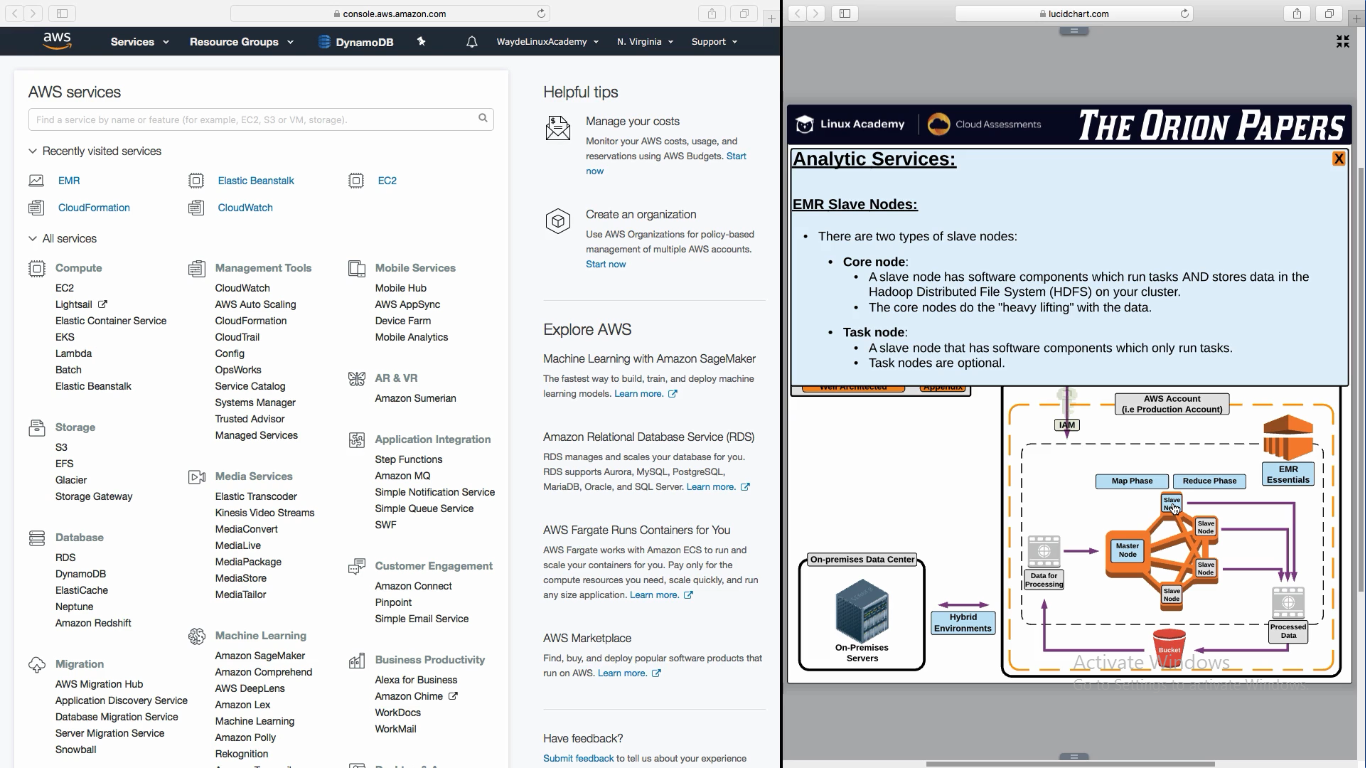


Reference:

For the latest stream and shard limits, see: <https://docs.aws.amazon.com/streams/latest/dev/service-sizes-and-limits.html>

**EMR**

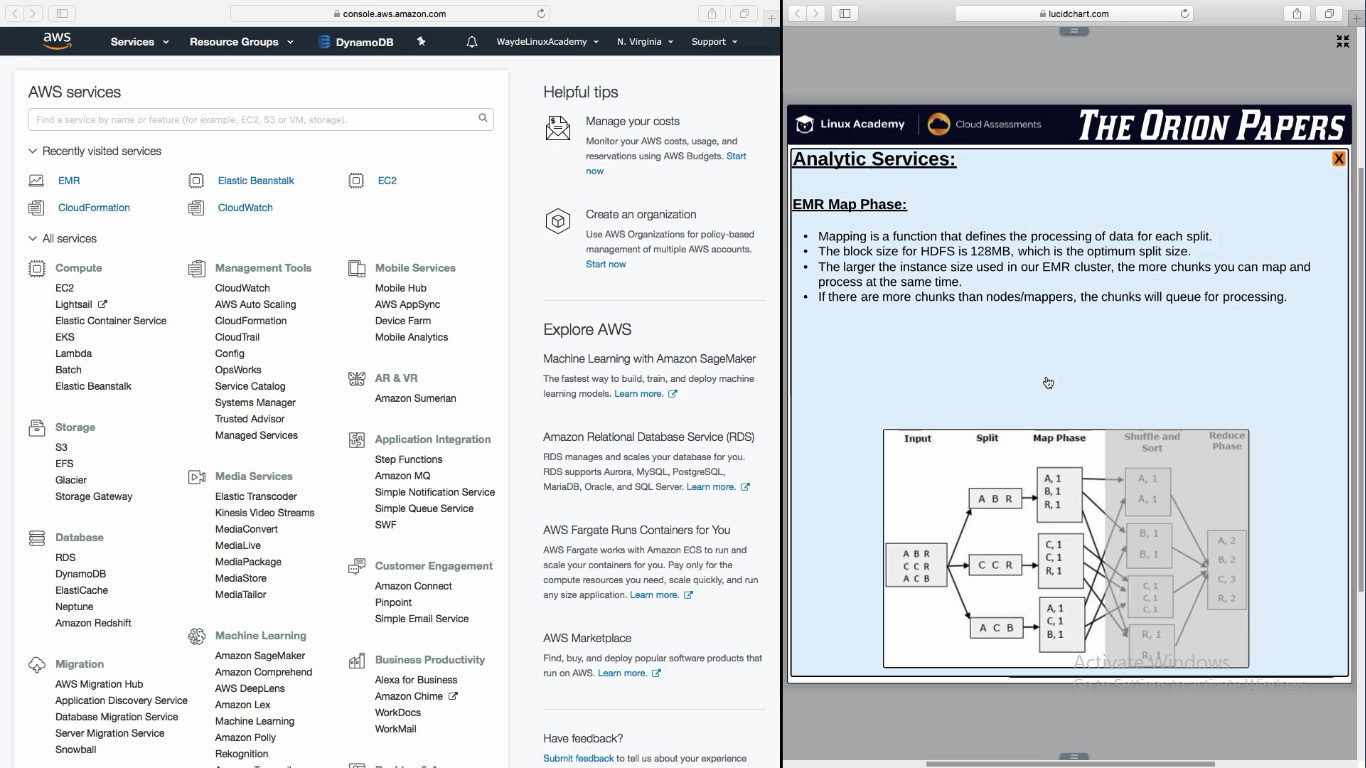
****

****

**Core nodes** have the HDFS to store files. We need atleast one core node

**Task nodes** have more cpu and ram. It is used when more compute is needed. It can be terminated if not needed

**Reduce Phase**

****