Best Practice for MapR cluster (100+ nodes)

In a large cluster control nodes need to live on their own nodes and not share the hardware to avoid any resource contention (100 nodes or more)

CLDB service :

Create CLDB-only nodes to ensure high performance. Setting up CLDB-only nodes involves restricting the CLDB volume to its own topology and making sure all other volumes are on and the non-CLDB path are children of the root topology path, new non-CLDB volumes are not guaranteed to keep off the CLDB-only nodes. This configuration also provides additiona balancing, fault tolerance, or high availability (HA).

CLDB nodes will serve just the CLDB data(mapr.cldb.internal volume). That data consists of a single container Cid 1, which in turn is stored on a single storage pool. Hence, all disks single storage pool, otherwise all the disks in the storage pools not containing the CLDB volume data will be always idle. During activities like bringing all of MapRFS online after a full c CLDB. Hence we should ensure there is enough disks to service the ops required based on cluster size (e.g. number of containers and number of nodes).

Using a 3 or 4 flash SSDs in one SP (20 GB SP) on the CLDB-only nodes should provide good performance. While 128 GB Memory and 32 cores with 10GbE Network would be suffi

ZK service :

Isolate the ZooKeeper on nodes that do not perform any other function. Isolating the ZooKeeper node enables the node to perform its functions without competing for resources with o similar to any typical node installation, but with a specific subset of packages. On ZK nodes Do not install the FileServer package in order to prevent MapR from using this node for dat

64 GB Memory and 32 cores with 10GbE Network would be sufficient while no disks to be given to MapR-FS.

On ZooKeeper nodes, dedicate 100 GB partition for the `/opt/mapr/zkdata` directory to avoid other processes filling that partition with writes and to reduce the possibility of errors d is used to store number of snapshots. Do not share the physical disk where `/opt/mapr/zkdata` resides with any MapR File System data partitions to avoid I/O conflicts that might le

RM nodes :

Isolate the RM's on nodes that do not perform any other function but resource management for the Yarn cluster.

6 * 1 TB SAS disks (Two 3 TB SP's) , 128 GB Memory and 32 cores with 10GbE Network would be sufficient.

Note :- Yarn behavior

- RM volume is a standard volume under /data.
- NM volumes are local volumes, they reside only on NM nodes.
- When a job is submitted, the job's needed dependencies are first copied over to RM volume, then localized to each NM volume. i.e Job client submit jar to RM Volume then NM copies the jar from RM volume.

RM data (mapr.resourcemanager.volume volume) volume topology .

There are two different best practices for large cluster, both involving creating RM topology:

1) If the cluster is homogeneous and each node has equivalent network capacity; do not move RM volume to RM only topology. Keep RM volume under /data so the load is distributed.
2) If the cluster is heterogeneous, RM volume needs to be on RM only topology to avoid job failure due to AM failing to write commit file on RM volume when mfs is busy; the RM topology nodes should have sufficient Network and Disk capacity to support the need of heavy data copying between RM and NM nodes.