Engineering Excellence J1

(NameNode HA Architecture)

TABLE OF CONTENT

Architecture

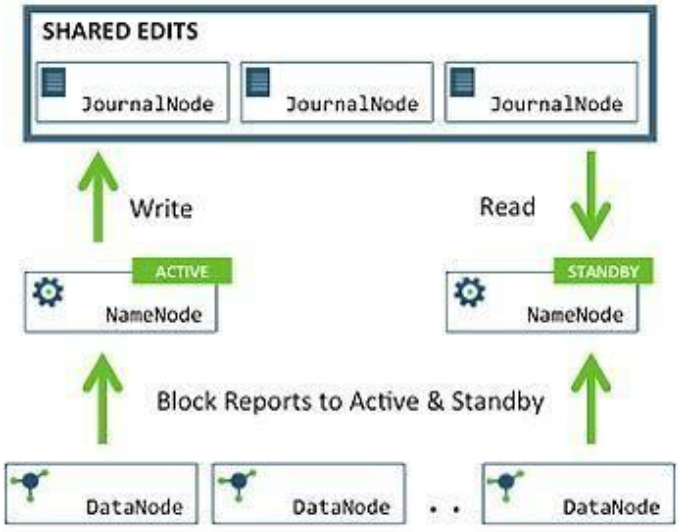
In a typical HA cluster, two separate machines are configured as NameNodes. In a working cluster, one of the NameNode machine is in the **Active** state, and the other is in the **Standby** state.

The Active NameNode is responsible for all client operations in the cluster, while the Standby acts as a slave. The Standby machine maintains enough state to provide a fast failover (if required).

In order for the Standby node to keep its state synchronized with the Active node, both nodes communicate with a group of separate daemons called **JournalNodes**(JNs). When the Active node performs any namespace modification, the Active node durably logs a modification record to a majority of these JNs. The Standby node reads the edits from the JNs and continuously watches the JNs for changes to the edit log. Once the Standby Node observes the edits, it applies these edits to its own namespace. When using QJM, JournalNodes acts the shared editlog storage. In a failover event, the Standby ensures that it has read all of the edits from the JounalNodes before promoting itself to the Active state. (This mechanism ensures that the namespace state is fully synchronized before a failover completes.)

Note: Secondary NameNode is not required in HA configuration because the Standby node also performs the tasks of the Secondary NameNode.

To provide a fast failover, it is also necessary that the Standby node have up-to-date information on the location of blocks in your cluster. To get accurate information about the block locations, DataNodes are configured with the location of both of the NameNodes, and send block location information and heartbeats to both NameNode machines.



It is vital for the correct operation of an HA cluster that only one of the NameNodes should be Active at a time. Failure to do so, would cause the namespace state to quickly diverge between the two NameNode machines thus causing potential data loss. (This situation is called a **split-brain scenario**.). To prevent the split-brain scenario, the JournalNodes allow only one NameNode to be a writer at a time. During failover, the NameNode, that is to chosen to become active, takes over the role of writing to the JournalNodes. This process prevents the other NameNode from continuing in the Active state and thus lets the new Active node proceed with the failover safely.