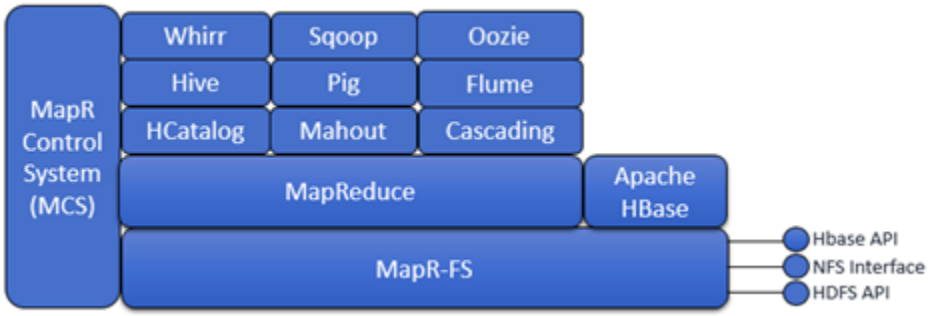
MapR Overview

MapR is a complete enterprise-grade distribution for Apache Hadoop. The MapR Distribution for Apache Hadoop has been engineered to improve Hadoop's reliability, performance, and ease of use. The MapR distribution provides a full Hadoop stack that includes the MapR File System (MapR-FS), MapReduce, a complete Hadoop ecosystem, and the MapR Control System user interface. You can use MapR with Apache Hadoop, HDFS, and MapReduce APIs.

The following image displays a high-level view of the MapR Distribution for Apache Hadoop:



The MapR distribution provides several unique features that address common concerns with Apache Hadoop:

Issue	Addressed by MapR Feature	Apache Hadoop
Data Protection	MapR Snapshots provide complete recovery capabilities. MapR Snapshots are rapid point-in-time consistent snapshots for both files and tables. MapR Snapshots make efficient use of storage and CPU resources, storing only changes from the point the snapshot is taken. You can configure schedules for MapR Snapshots with easy to use but powerful scheduling tools.	Snapshot-like capabilities are not consistent, require application changes to make consistent, and may lead to data loss in certain situations.
Security	With wire-level security, data transmissions to, from, and within the cluster are encrypted, and strong authorization mechanisms enable you to tailor the actions a given user is able to perform. Authentication is robust without burdening end-users. Permissions for users are checked on each file access.	Permissions for users are checked on file open only.
Disaster Recovery	MapR provides business continuity and disaster recovery services out of the box with mirroring that's simple to configure and makes efficient use of your cluster's storage, CPU, and bandwidth resources.	No standard mirroring solution. Scripts based on distcp quickly become hard to administer and manage. No enterprise-grade consistency.
Enterprise Integration	With high-availability Direct Access NFS, data ingestion to your cluster can be made as simple as mounting an NFS share to the data source. Support for Hadoop ecosystem projects like Flume or Sqoop means minimal disruptions to your existing workflow.	
Performance	MapR uses customized units of I/O, chunking, resync, and administration. These architectural elements allow MapR clusters to run at speeds close to the maximum allowed by the underlying hardware. In addition, the DirectShuffle technology leverages the performance advantages of MapR-FS to deliver strong cluster performance, and Direct Access NFS simplifies data ingestion and access. MapR-DB tables, available with the M7 license, are natively stored in the file system and support the Apache HBase API. MapR-DB tables provide the fastest and easiest to administer NoSQL solution on Hadoop.	Stock Apache Hadoop's NFS cannot read or write to an open file.

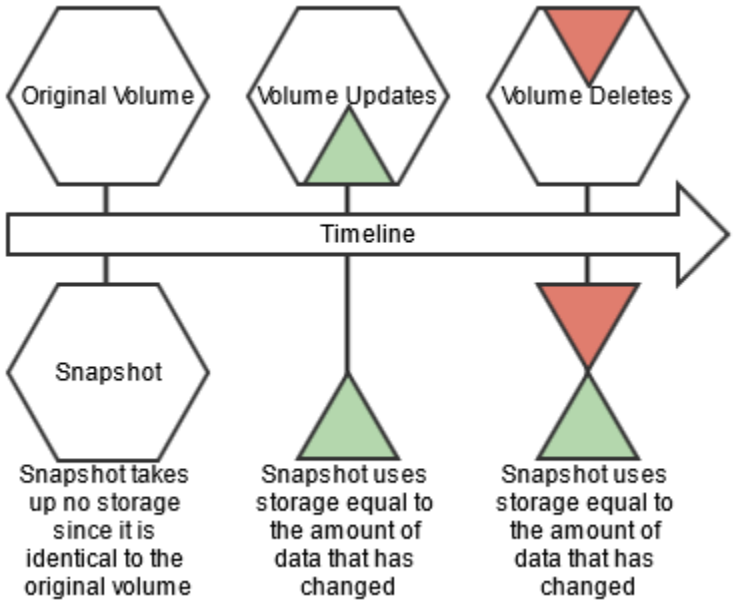
Scalable architecture without single points of failure	The MapR distribution for Hadoop provides High Availability for the Hadoop components in the stack. MapR clusters don't use NameNodes and provide stateful high-availability for the MapReduce JobTracker and Direct Access NFS. Works out of the box with no special configuration required.	NameNode HA provides failover, but no failback, while limiting scale and creating complex configuration challenges. NameNode federation adds new processes and parameters to provide cumbersome, error-prone file federation. The High-Availability JobTracker in stock Apache Hadoop does not preserve the state of running jobs. Failover for the JobTracker requires restarting all in-progress jobs and brings complex configuration requirements.

To learn more about MapR, including information about MapR partners, see the following sections:

Data Protection: Rolling Back with Snapshots

The MapR Distribution for Hadoop provides *snapshots*, which enable you to roll back to a known good data set. A snapshot is a read-only image of a volume that provides point-in-time recovery. Snapshots only store changes to the data stored in the volume, and as a result make extremely efficient use of the cluster's disk resources. Snapshots preserve access to historical data and protect the cluster from user and application errors. You can create a snapshot manually or automate the process with a schedule.

The following image represents a mirror volume and a snapshot created from a source volume:



Read the Snapshots documentation for details.

Keeping it Safe: Security Features

The 3.1 release of the MapR distribution for Hadoop provides authentication, authorization, and encryption services to protect the data in your cluster. MapR leverages Linux pluggable authentication modules (PAM) to support the main authentication protocols out of the box. A MapR cluster can authenticate users through Kerberos, LDAP/AD, NIS, or any other service that has a PAM module.

For authorization, MapR provides Access Control Lists (ACLs) for job queues, volumes, and the cluster as a whole. Because MapR supports POSIX permissions on files and directories, MapR-FS performs permission checks on each file access. Other Hadoop distributions only check permissions on file open.

MapR clusters also incorporate wire-level security (WLS) to encrypt data transmission for traffic within the cluster, as well as traffic between the cluster and client machines.

MapR leverages the Hadoop Fair Scheduler to ensure fair allocation of resources to different users, and includes support for SELinux.

Read the Security documentation for details.

Authorization with Volumes: Intelligent Policy Management

The MapR File System uses volumes as a unique management entity. A volume is a logical unit that you create to apply policies to a set of files, directories, tables, and sub-volumes. You can create volumes for each user, department, or project. Mirror volumes and volume snapshots, discussed later in this document, provide data recovery and data protection functionality.

Volumes can enforce disk usage limits, set replication levels, establish ownership and control permissible actions, and measure the cost generated by different projects or departments. When you set policies on a volume, all files contained within the volume inherit the same policies set on the volume. Other Hadoop distributions require administrators to manage policies at the file level.

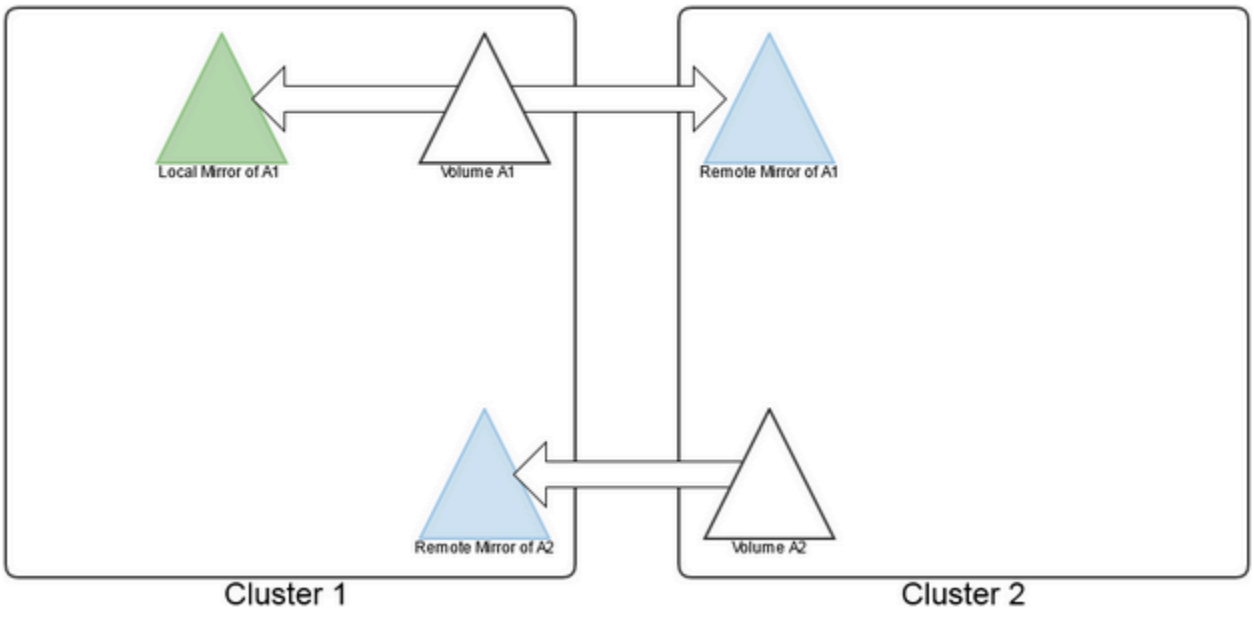
You can manage volume permissions through Access Control Lists (ACLs) in the MapR Control System or from the command line. You can also set read, write, and execute permissions on a file or directory for users and groups with standard UNIX commands, when that volume has been mounted through NFS, or using standard hadoop fs commands.

Read the Managing Data with Volumes documentation for details.

Disaster Recovery With Mirrors

The MapR Distribution for Hadoop provides built-in mirroring to set recovery time objectives and automatically mirror data for backup. You can create local or remote mirror volumes to mirror data between clusters, data centers, or between on-premise and public cloud infrastructures. Mirror volumes are read-only copies of a source volume. You can control the schedule for mirror refreshes from the MapR Control System or with the command-line tools.

The following image shows two clusters with mutual remote mirroring and a local mirror:



Read the Mirroring documentation for details.

For more information:

- Explore Data Protection scenarios

Simpler Data Flows with Direct Access NFS

The MapR direct access file system enables real-time read/write data ows using the Network File System (NFS) protocol. Standard applications and tools can directly access the MapR-FS storage layer using NFS. Legacy systems can access data and traditional file I/O operations work as expected in a conventional UNIX file system.

A remote client can easily mount a MapR cluster over NFS to move data to and from the cluster. Application servers can write log les and other data directly to the MapR cluster's storage layer instead of caching the data on an external direct or network-attached storage.

Read the NFS documentation for details.

Management Services

MapR provides high availability management and data processing services for automatic continuity throughout the cluster. You can use the MapR Control System, command-line interface, or REST API to start, stop, and monitor services at the node or cluster level.

MapReduce services such as the JobTracker, management services such as the ZooKeeper, and data access services such as NFS provide continuous service during any system failure.

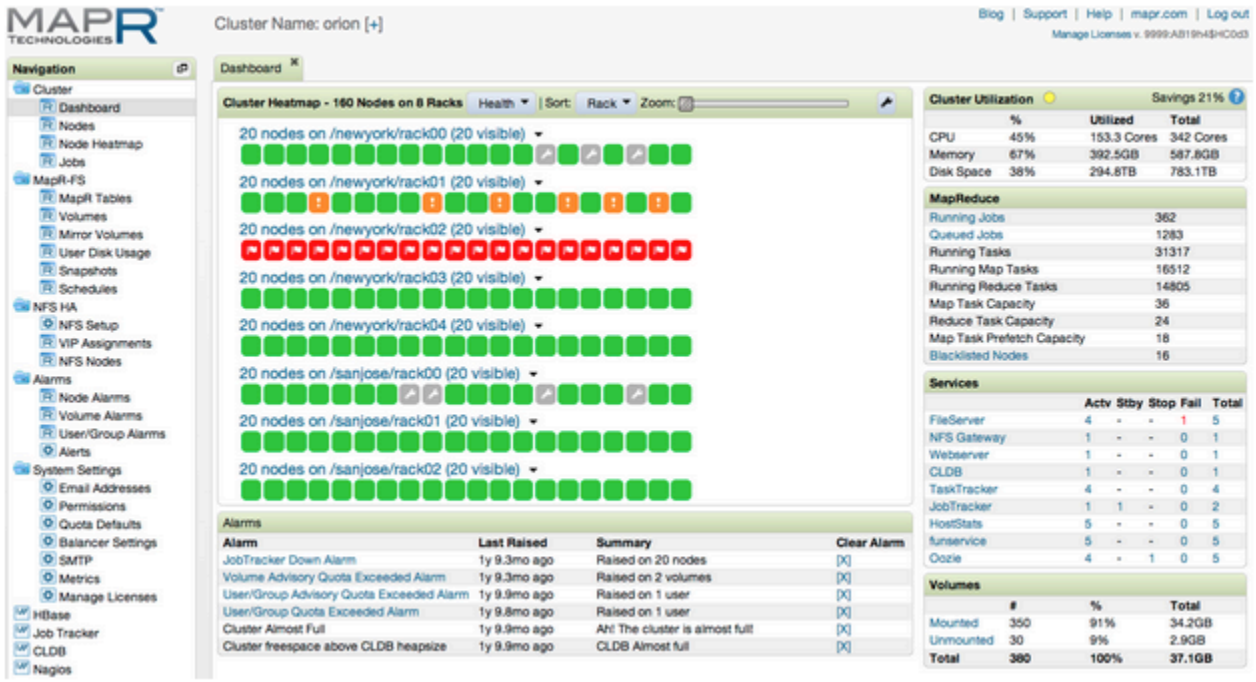
Read the Services documentation for details.

The MapR Control System (MCS)

The MapR Control System (MCS) provides a graphical control panel for cluster administration with all the functionality of the command-line or REST APIs. The MCS provides job monitoring metrics and helps you troubleshoot issues, such as which jobs required the most memory in a given week or which events caused job and task failures.

The MCS Dashboard provides a summary of information about the cluster including a cluster heat map that displays the health of each node; an alarms summary; cluster utilization that shows the CPU, memory, and disk space usage; services running across the cluster; the number of available, unavailable, and under replicated volumes; MapReduce jobs. Links in each pane provide shortcuts to more detailed information.

The following image shows the MCS Dashboard:



The MCS provides various views. You can navigate through views to monitor and configure your cluster. Select any of the following links to see

what each view in the MCS provides:

• Cluster • MapR-FS • NFS HA • Alarms • System Settings	• HBase • JobTracker • Nagios • Terminal

For more information:

- Take a look at the Heatmap
- Read about Analyzing Job Metrics and Node Metrics

MapReduce Performance

MapR provides performance improvements in the shuffle phase of MapReduce and adds high availability for all Hadoop services.

With MapR, you can configure Hadoop services to run on multiple nodes for failover. If one service node fails, another continues to perform the tasks related to that service without delaying the MapReduce job.

The shuffle phase of a MapReduce job combines the map output so that all the records from a given key/value pair's key go to one reduce task. This phase involves a great deal of copying and coordination between nodes in the cluster. Shuffling in MapR-FS is much faster than other Hadoop distributions because MapR uses highly optimized, efficient remote procedure call connections to transport data while other Hadoop distributions use HTTP connections.

Other Hadoop distributions keep map output on local disk, which creates competition for disk space between local and distributed storage. In MapR, any spilled data is stored in the distributed file system making it directly accessible.

Filesystem Storage for Tables: Keep your Data off the Performance Floor

A MapR cluster integrates NoSQL technology that stores tables natively in the filesystem layer. MapR-DB tables support the HBase API. The MapR distribution for Hadoop integrates files and tables to provide significant performance and administration benefits over other distributions. MapR clusters deliver a 2-10x throughput advantage and a 2-50x read latency decrease across different workloads compared to other distributions while significantly reducing latency variability. Tables stored in the MapR-FS layer benefit from the MapR distribution's high availability, automatic data protection, and disaster recovery with snapshots and mirrors.

There's no limit to the number of tables or files you can have in a MapR cluster. Tables can be managed by individual users, freeing cluster administrators from database administration overhead. With MapR-DB tables, cluster administrators do not have to manage RegionServers or daemons, and region splits are handled automatically. Node upgrades and other administrative tasks do not cause downtime for table storage.

HBase applications and MapReduce jobs can co-exist on the same nodes without disrupting cluster performance. MapR-DB tables support in-memory column families to speed inserts and updates. A MapR cluster supports mixed environments that use MapR-DB tables and Apache HBase as well as environments that use MapR-DB tables exclusively.

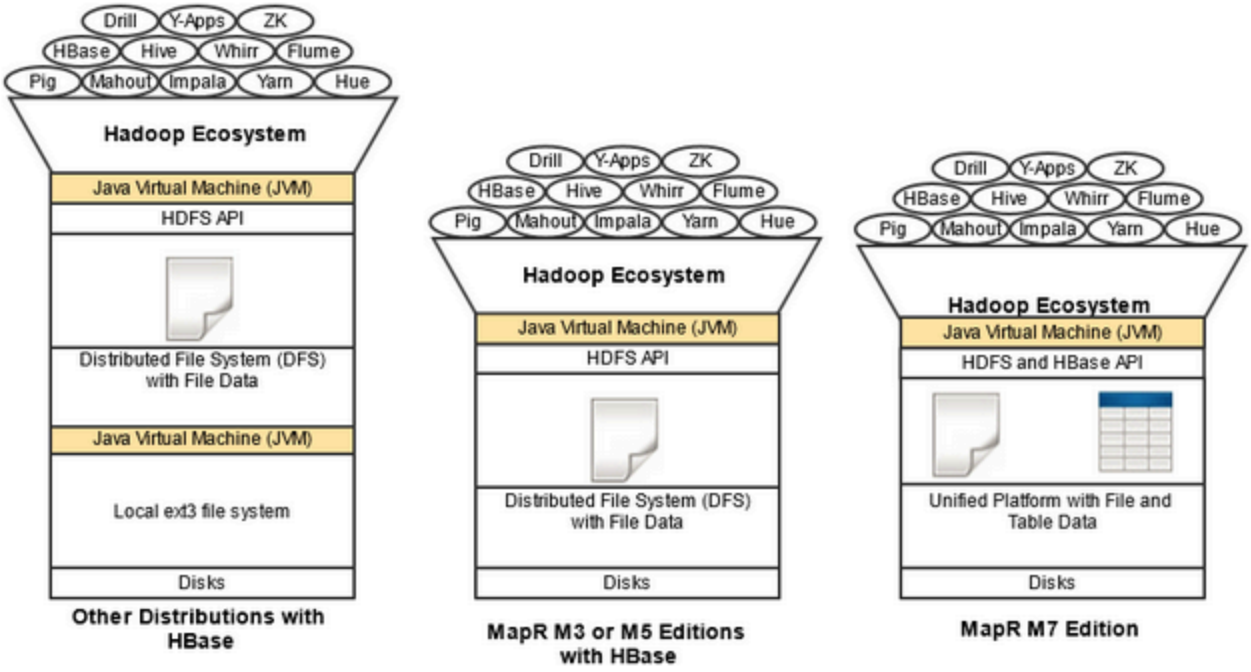
Read the MapR-DB for details.

The MapR-FS Layer: Performance on Disk

The MapR Filesystem, or MapR-FS, is a random read-write distributed file system that allows applications to concurrently read and write directly to disk. The Hadoop Distributed File System (HDFS), by contrast, has append-only writes and can only read from closed files.

Because HDFS is layered over the existing Linux file system, a greater number of input/output (I/O) operations decrease the cluster's performance.

The following image compares an Apache Hadoop stack to the MapR stack:



The storage system architecture used by MapR-FS is written in C/C++ and prevents locking contention, eliminating performance impact from Java garbage collection.

Expand Your Capabilities with Third-Party Solutions

MapR has partnered with Datameer, which provides a self-service Business Intelligence platform that runs best on the MapR Distribution for Apache Hadoop. Your download of MapR includes a 30-day trial version of Datameer Analytics Solution (DAS), which provides spreadsheet-style analytics, ETL and data visualization capabilities.

Other MapR partners include HParser, Karmasphere, and Pentaho.

MapR Editions

The edition of MapR that you use determines which features are available on the cluster.

MapR offers the following editions of the MapR distribution for Apache Hadoop:

Edition	Description
MapR Community Edition (formerly M3)	Free community edition
MapR Enterprise Edition (formerly M5)	Adds high availability and data protection, including multi-node NFS
MapR Enterprise Database Edition (formerly M7)	Adds structured table data natively in the storage layer and provides a flexible NoSQL database

Get Started

Now that you know a bit about how the features of MapR Distribution for Apache Hadoop work, take a quick tour to see for yourself how they can work for you:

- **MapR Sandbox for Hadoop** - Try out a single-node cluster that's ready to roll, right out of the box!
- **Advanced Installation Topics** - Learn how to set up a production cluster, large or small
- **Development Guide** - Read more about what you can do with a MapR cluster
- **Administrator Guide** - Learn how to configure and tune a MapR cluster for performance