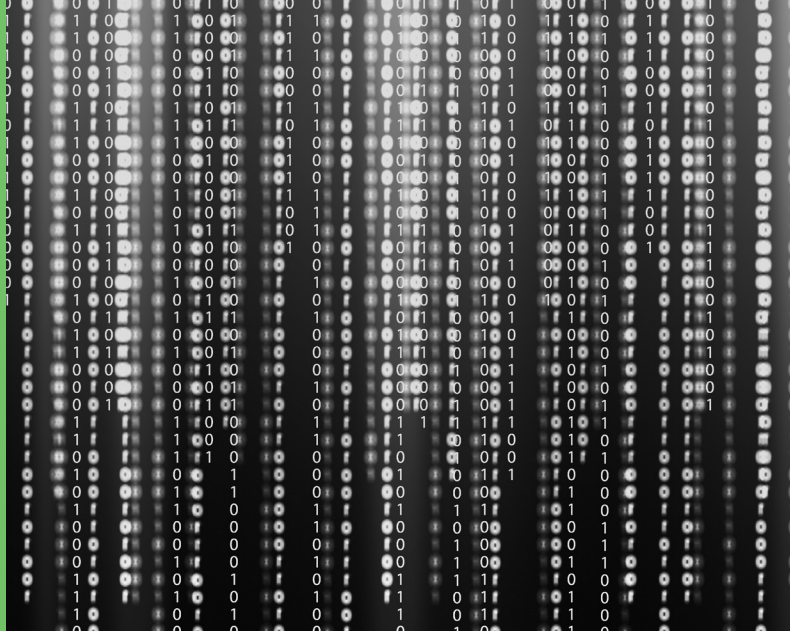
Platform Computing
an IBM Company

# Platform Symphony MapReduce

Platform Symphony MapReduce offers a proven distributed computing platform for enterprise-class Hadoop applications and shared services deployments



## Features

- HDFS NameNode, Job Tracker / TaskTracker automatic failover and job recovery
- Multiple file system and data source support, including HDFS, GPFS and additional HDFS 'plug in' distributed file systems.
- Priority based workload scheduler supporting up to 10,000 levels of prioritizations
- Application and job level predictability through sophisticated workload shceduler
- Supports up to 300 job trackers across a shared pool of resources
- Self-monitoring and trouble-shooting capabilities
- Dynamic resource allocation logic drives higher cluster utilization

## Benefits

- High resource utilization and Predictability
- Enterprise-class manageability and Security
- High availability
- Open architecture delivering greater compatibility with various applications and file systems
- Shared Services / Multi-application workload capable
- Guaranteed SLA requirements

## Overview

Platform Symphony MapReduce is an enterprise-class, best-of-breed distributed runtime engine for MapReduce applications. It is designed to deliver production ready capabilities such as high resource availability and predictability, multiple applications and file systems support, operational maturity, SLA policy control, and extremely high resource utilization for MapReduce applications. Built on Platform Computing's years of expertise in distributed workload scheduling and management – proven technologies that are powering many of today's Fortune500 companies for their mission critical, most demanding workloads, Platform Symphony MapReduce offers unprecedented distributed workload runtime services for your MapReduce applications.

## Product Capabilities

### Policy driven workload scheduler

The Platform Symphony MapReduce policy driven workload scheduler provides 10,000 levels of priority and support multiple MapReduce jobs running in parallel. This policy driven scheduler includes delivery of resource priority for preemptive jobs, as well as fair share scheduling of Mapper and Reducer jobs, all done at the job level to provide better granularity and control.

### High resource availability

Platform Symphony MapReduce guarantees uptime within the distributed runtime engine - there is no single points of failure. It provides job tracker /task tracker automatic failover and job recovery, without the need to restart jobs. For Hadoop file system, Platform Symphony MapReduce offers automatic failover of the NameNode within the Hadoop Distributed file system and provides file system recovery and dependent job recovery.

### Open architecture for application development and choice of the file system

Platform Symphony MapReduce is built on an open architecture to support multiple MapReduce applications including 100% Hadoop application compatibility for Java based MapReduce jobs. The Application Adapter technology built in the product delivers seamless application integration with Platform MapRedcue so that Jobs built with Hadoop MapReduce technology (Java, Pig, Hive and others) require no changes to the programming logic for execution on Platform

Symphony MapReduce. This open architecture also provides a method for leveraging multiple file system types as well as database architectures. Platform Symphony MapReduce fully supports HDFS, GPFS and other distributed file system types and data types. In addition, for MapReduce processes, the input data source file system type can be different from the output data source file system. This provides support for many uses, including extract, transformation, and load (ETL) workflow logic.

## Supporting multiple MapReduce applications and mixed types of workloads running on the same cluster

Platform Symphony MapReduce supports up to 300 separate applications (Job Trackers) for MapReduce workloads, as well as other types of distributed applications, simultaneously. This allows customers to leverage both existing and new resources and maximize their IT infrastructure while maintaining a single management interface.

## Support rolling upgrade

Platform Symphony MapReduce supports multiple versions of MapReduce applications running on the same clusters, there is no need to take down the entire cluster for software upgrade. The severs running upgraded applications can co-exist with the previous version of the product on other nodes and thus allow upgrades to be done incrementally over a set of servers without taking down the entire cluster.

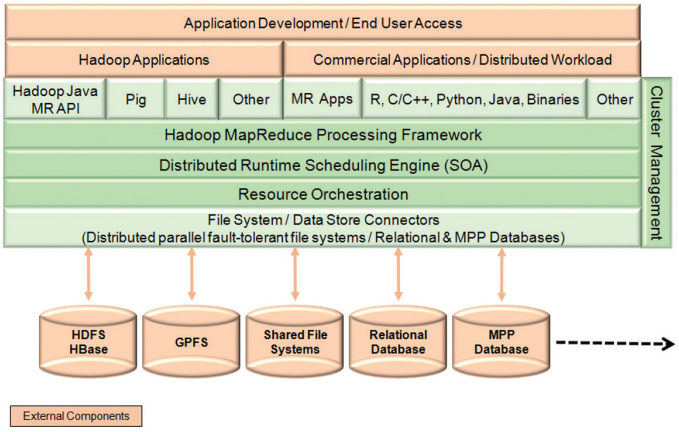
## Greater monitoring and troubleshooting capabilities

Platform Symphony MapReduce monitors CPU and memory utilization level and allocates resources accordingly. It provides the ability to pull log data from individual servers and manage them from a single interface.

## Platform Symphony MapReduce Data Affinity

Platform Symphony MapReduce includes powerful data affinity capabilities to significantly improve application performance and resource utilization by taking into account data locality when scheduling MapReduce workloads. Its data affinity solution virtually eliminates the time it takes to access large data volumes required by MapReduce applications. It increases overall application performance by up to 400% through faster file access.

**Platform Symphony MapReduce**
Architecture Layout




## System requirements

Two kinds of hosts are needed per distributed file system cluster for running Platform Symphony MapReduce:

1. Master hosts to run job tracker
2. Compute hosts to run task tracker

### Master Host Requirements

- Number of node >=1
- 2.4GHz of CPU
- 4GB of RAM
- 1GB of disk space for installation
- 30GB of free disk space

### Compute Host Requirements

- Number of nodes: >=1
- 2.4GHz of CPU
- 1GB of RAM
- 1GB of disk space for installation
- 10GB of free disk space plus additional capacity needed by user applications

### OS Requirements

- Red Hat Enterprise Linux AS 4, AS 5and AS 6 (Kernel 2.6.x compiled with glibc 2.3.x, both 32 bit and 64 bit)
- JDK 1.6 is installed and an OS environment variable "JAVA_HOME" is already set up.

Platform Computing
an IBM Company

Platform Computing, an IBM Company, is a leader in cluster, grid, and cloud management software - serving more than 2,000 of the world's most demanding organizations. Since 1992 its workload and resource management solutions have delivered optimized IT infrastructures, ease of management, and lower costs for enterprise, HPC, and technical computing clients. Visit www.platform.com. Twitter: @Platform_Tweets. For more information about IBM Technical Computing: http://www.ibm.com/deepcomputing.


**World Headquarters**
Platform Computing,
an IBM Company
3760 14th Avenue
Markham, Ontario
Canada L3R 3T7
Tel: +1 905 948 8448
Fax: +1 905 948 9975
Toll-free Tel: 1 877 528 3676
info@platform.com

**Sales - Headquarters**
Toll-free Tel: 1 877 710 4477
Tel: +1 905 948 8448

**North America**
New York: +1 212 888 6270
San Jose: +1 408 392 4900

**Europe**
Bramley: +44 (0) 1256 883756
London: +44 (0) 20 3206 1470
Paris: +33 (0) 1 41 10 09 20
Düsseldorf: +49 2102 61039 0

**Asia-Pacific**
Beijing: +86 10 82276000
Xi'an: +86 029 87607400
Tokyo: +81(0)3 6302 2901
Singapore: +65 6307 6590