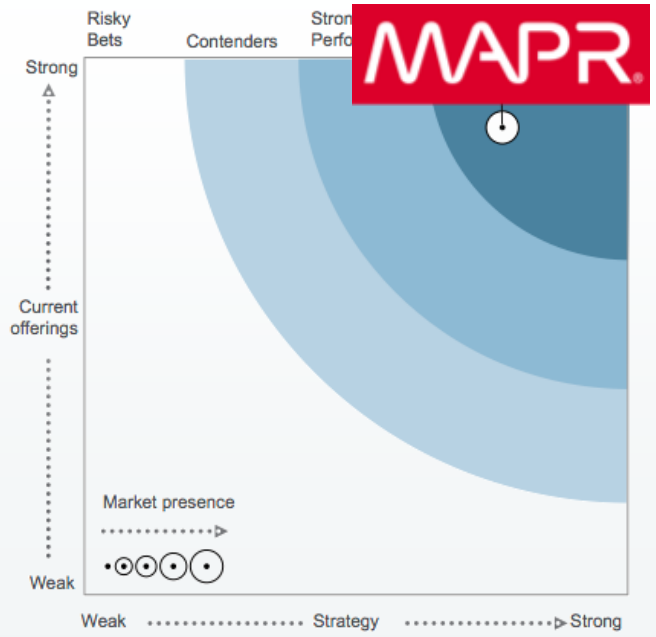
MapR Overview

Aurélien GOUJET - SE Director Southern Europe and Benelux

July 2014

MapR Enterprise Hadoop

Top Ranked



FORRESTER®

Cloud Leaders

amazon web services™

Google

500+ Customers

comScore

hp

URBAN AIRSHIP

LIVE NATION ENTERTAINMENT®

CISCO

rubicon PROJECT

Key MapR Advantage Partners

APPLICATIONS & OS

SAP
splunk>
vmware
redhat
ubuntu
UNIVA
LucidWorks

ANALYTICS & BUSINESS INTELLIGENCE

tableau SOFTWARE
sas
MicroStrategy
REVOLUTION ANALYTICS
Datameer Powerfully Simple™
platfora™
SKYTREE THE MACHINE LEARNING COMPANY

DATA WAREHOUSE & INTEGRATION

INTEGRATING BIG DATA... SMARTER syncsort
informatica
NETEZZA
VERTICA An HP Company
talend* open integration solutions
pentaho
TERADATA

INFRASTRUCTURE & CLOUD

CISCO
hp
IBM
amazon web services™
Google
CenturyLink Technology Solutions

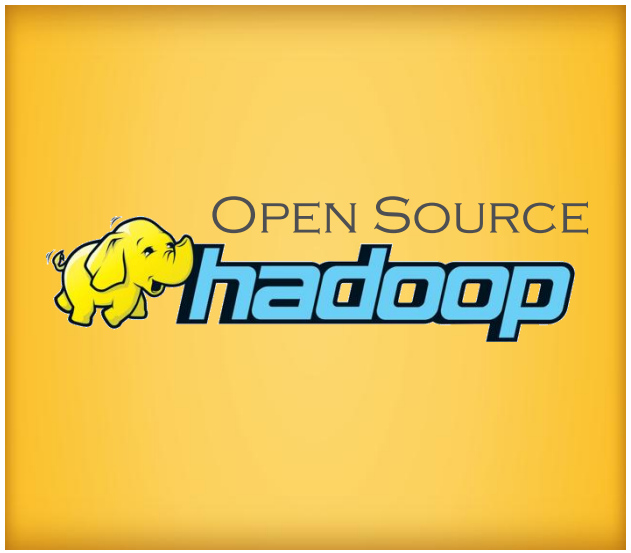
CONSULTANTS & INTEGRATORS

Ysance simplifions les projets informatiques
OCTO Technology
BULL
FastConnect
bluestone
Quantmetry expertise statistique
Affini-TECH
Umanis

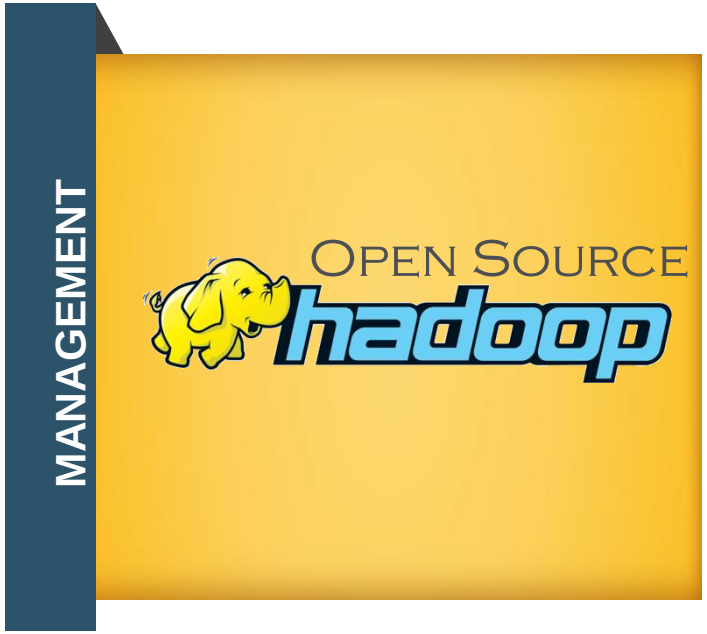
MapR Distribution for Hadoop

Hadoop Distributions

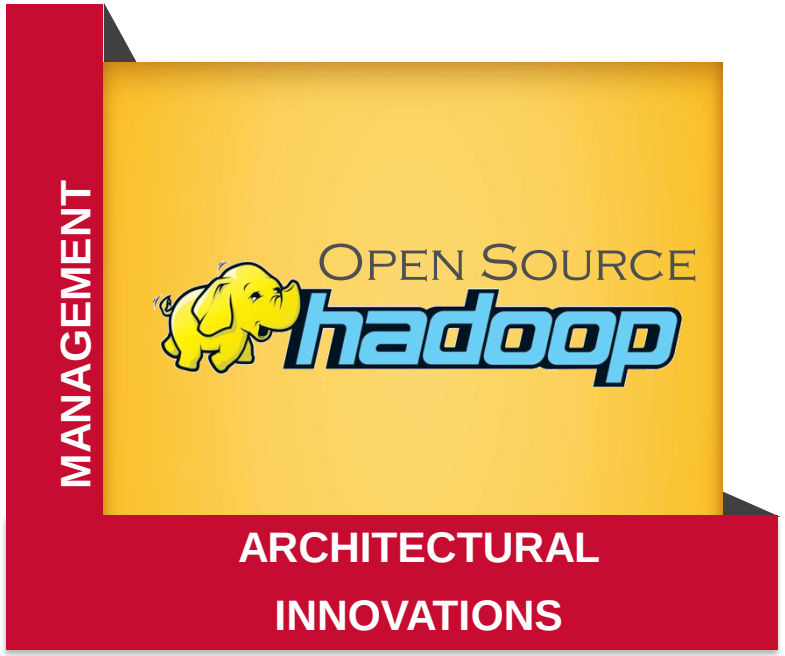
Distribution A



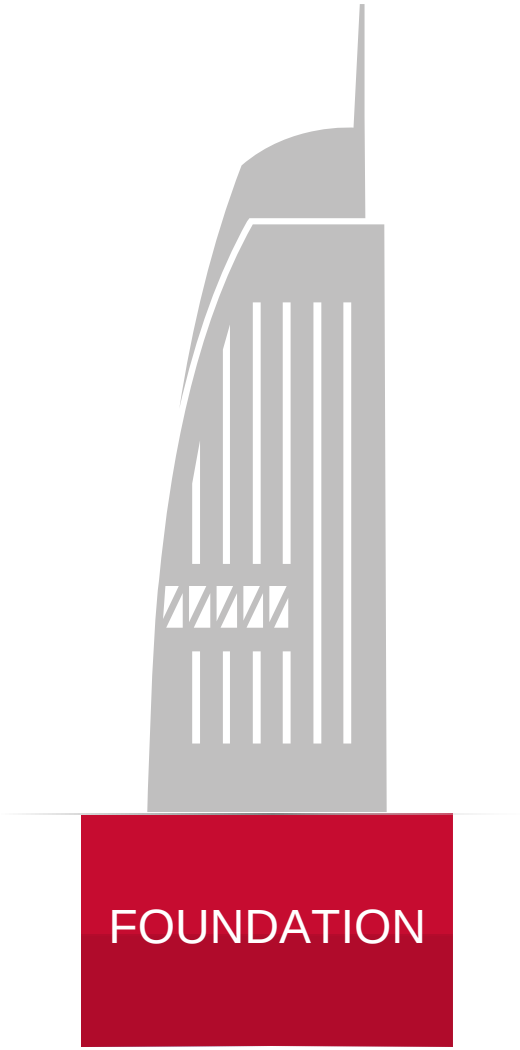
Distribution C



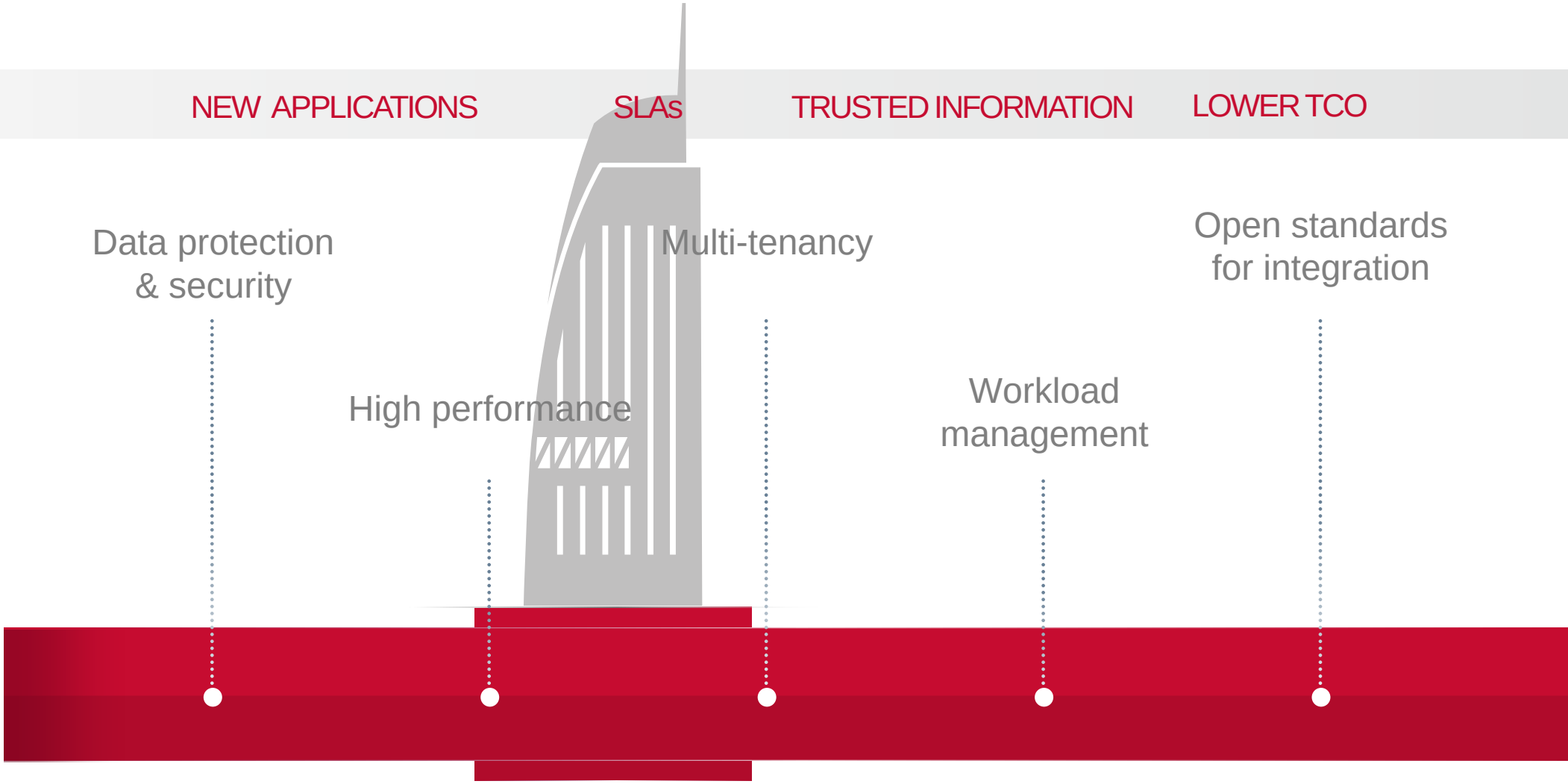
MAPR



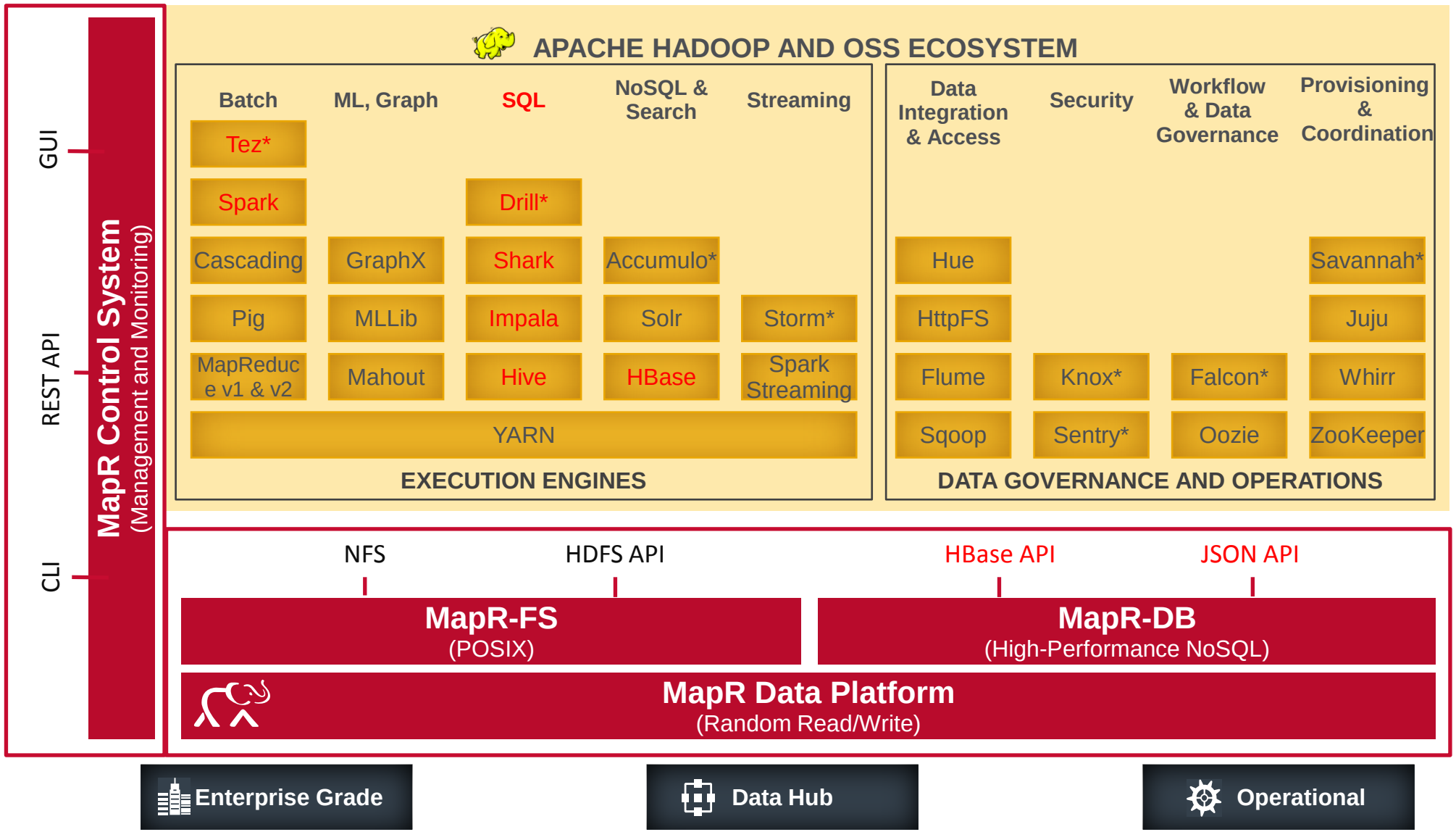
Architecture Matters for Success



Architecture Matters for Success



MapR Distribution for Apache Hadoop



Business Continuity

Business Continuity

What are your requirements?

What do you have for your enterprise storage, databases and data warehouses?

AVAILABILITY 1

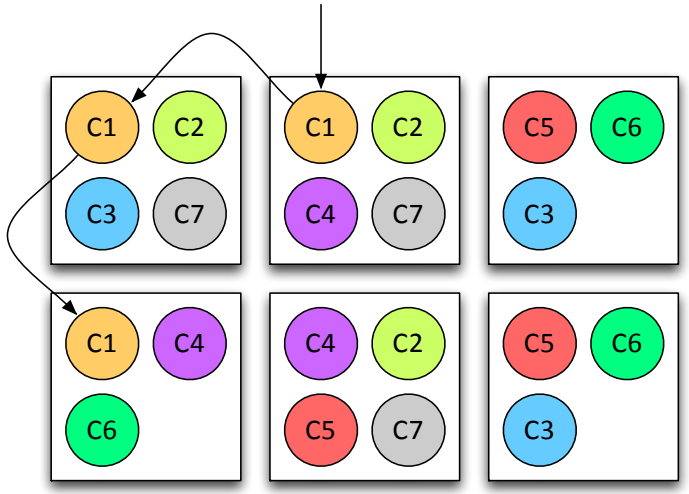
High Availability Everywhere

No NameNode architecture	• Distributed metadata can self-heal • No practical limit on # of files
MapReduce/YARN HA	• Jobs are not impacted by failures • Meet your data processing SLAs
NFS HA	• High throughput and resilience for NFS-based data ingestion, import/export and multi-client access
Instant recovery	• Files and tables are accessible within seconds of a node failure or cluster restart
Rolling upgrades	• Upgrade the software with no downtime
HA is built in	• No special configuration to enable HA • All MapR customers operate with HA

Data Protection: Replication and Snapshots

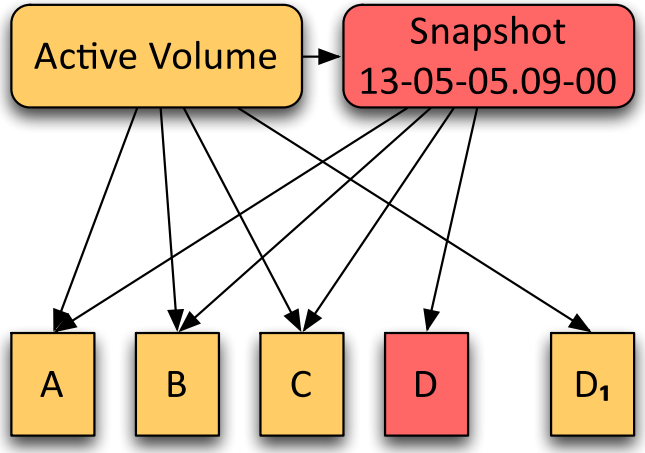
Replication

- Protect from hardware failures
- File chunks, table regions and metadata are automatically replicated (3x by default)
- At least one replica on a different rack



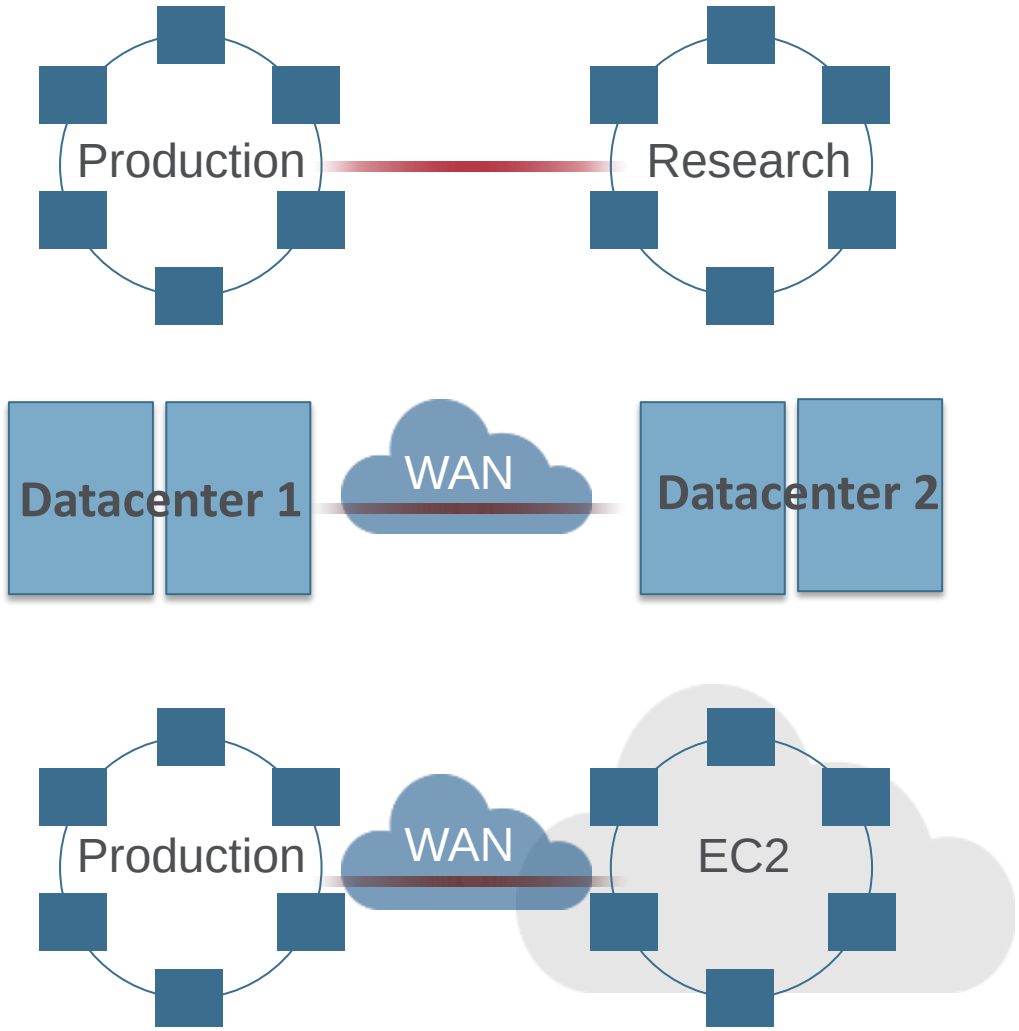
Snapshots

- Protect from user and application errors
- Point-in-time recovery
- No data duplication
- No performance or scale impact
- Read files and tables directly from snapshot

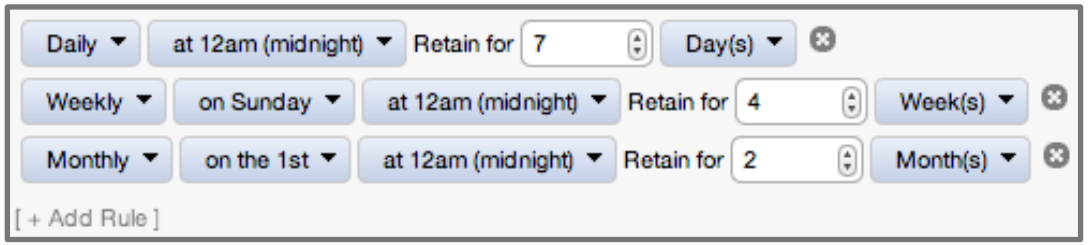


DISASTER RECOVERY 3

Disaster Recovery : MapR Mirroring



- Flexible
  - Choose the volumes/directories to mirror
  - You don't need to mirror the entire cluster
  - Active/active
- Fast
  - No performance impact
  - Block-level (8KB) deltas
  - Automatic compression
- Safe
  - Point-in-time consistency
  - End-to-end checksums
- Easy
  - Graceful handling of network issues
  - No third-party software
  - Takes less than two minutes to configure!



SQL

Interactive SQL-on-Hadoop: MapR Customers Have Options!

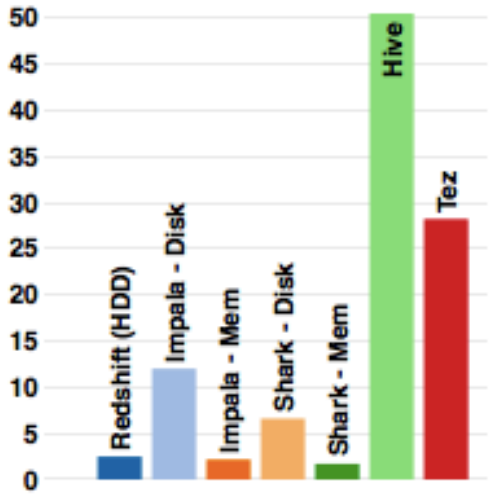
	Drill 1.0	Hive 0.13 with Tez	Impala 1.x	Presto 0.56	Shark 0.8	Vertica
Latency	Low	Medium	Low	Low	Medium	Low
Files	Yes (all Hive file formats)	Yes (all Hive file formats)	Yes (Parquet, Sequence, …)	Yes (RC, Sequence, Text)	Yes (all Hive file formats)	Yes (all Hive file formats)
HBase/M7	Yes	Yes	Various issues	No	Yes	No
Schema	Hive or schema-less	Hive	Hive	Hive	Hive	Proprietary or Hive
SQL support	ANSI SQL	HiveQL	HiveQL (subset)	ANSI SQL	HiveQL	ANSI SQL + advanced analytics
Client support	ODBC/JDBC	ODBC/JDBC	ODBC/JDBC	ODBC/JDBC	ODBC/JDBC	ODBC/JDBC, ADO.NET, …
Large joins	Yes	Yes	No	No	No	Yes
Nested data	Yes	Limited	No	Limited	Limited	Limited
Hive UDFs	Yes	Yes	Limited	No	Yes	No
Transactions	No	No	No	No	No	Yes
Optimizer	Limited	Limited	Limited	Limited	Limited	Yes
Concurrency	Limited	Limited	Limited	Limited	Limited	Yes

Benchmark Berkeley AmpLab

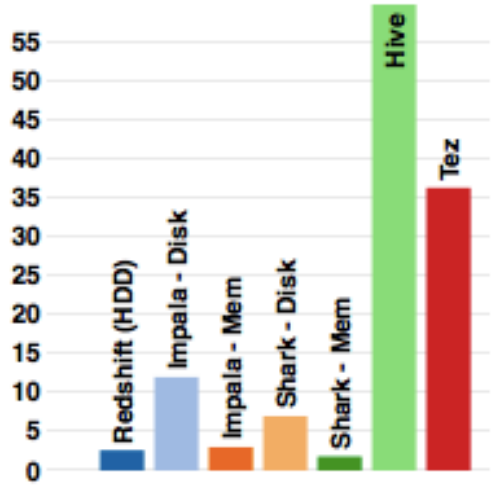
1. Scan Query

```
SELECT pageURL, pageRank FROM rankings WHERE pageRank > X
```

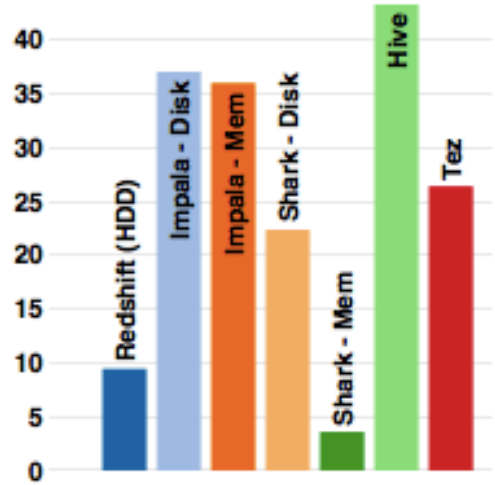
Query 1A
32,888 results



Query 1B
3,331,851 results



Query 1C
89,974,976 results



amplab UC BERKELEY

https://amplab.cs.berkeley.edu/benchmark/

Apache Open Source Community Projects

		Q1	Q2	Q3	Q4
Resource management	YARN 2.4	◆	◆	◆	◆
Batch	MapReduce 2.4	◆	◆	◆	◆
	Hive 0.12	◆	◆	◆	◆
	Pig 0.11		◆		◆
	Cascading 2.5	◆	◆	◆	◆
	Spark 0.9	◆	◆	◆	◆
Interactive SQL	Drill 1.0		◆	◆	◆
	Shark 0.9	◆	◆	◆	◆
	Impala 1.2.3	◆	◆	◆	◆
	Hive on Tez 0.13			◆	◆
Data integration	Flume 1.4.0	◆	◆	◆	◆
	Sqoop 1.4.4	◆		◆	
	HttpFS 1.0		◆		◆
Machine learning	Mahout 0.8	◆		◆	
	MLLib 0.9	◆	◆	◆	◆
	GraphX 0.9	◆	◆	◆	◆
Coordination	Oozie 3.3.2	◆	◆	◆	◆
	ZooKeeper	◆	◆	◆	◆
Streaming	Storm 0.9.0		◆	◆	◆
	Spark Streaming 0.9	◆	◆	◆	◆
Data management	Falcon 0.3			◆	◆
	Knox 0.3			◆	◆
	Sentry 1.2.0			◆	◆
GUI and provisioning	Hue 3.5		◆	◆	◆
	Savannah 0.4		◆	◆	◆
NoSQL and search	HBase 0.94	◆	◆	◆	◆
	Solr (LWS 2.6.1)	◆	◆	◆	◆

- Complete distribution with aggressive roadmap
  - 20 projects in the distribution
  - 10+ new projects in 2014
- Monthly update to projects
- Run multiple versions simultaneously
  - Adopt new project versions w/o upgrading cluster
  - Upgrade cluster without migrating all applications to new project versions

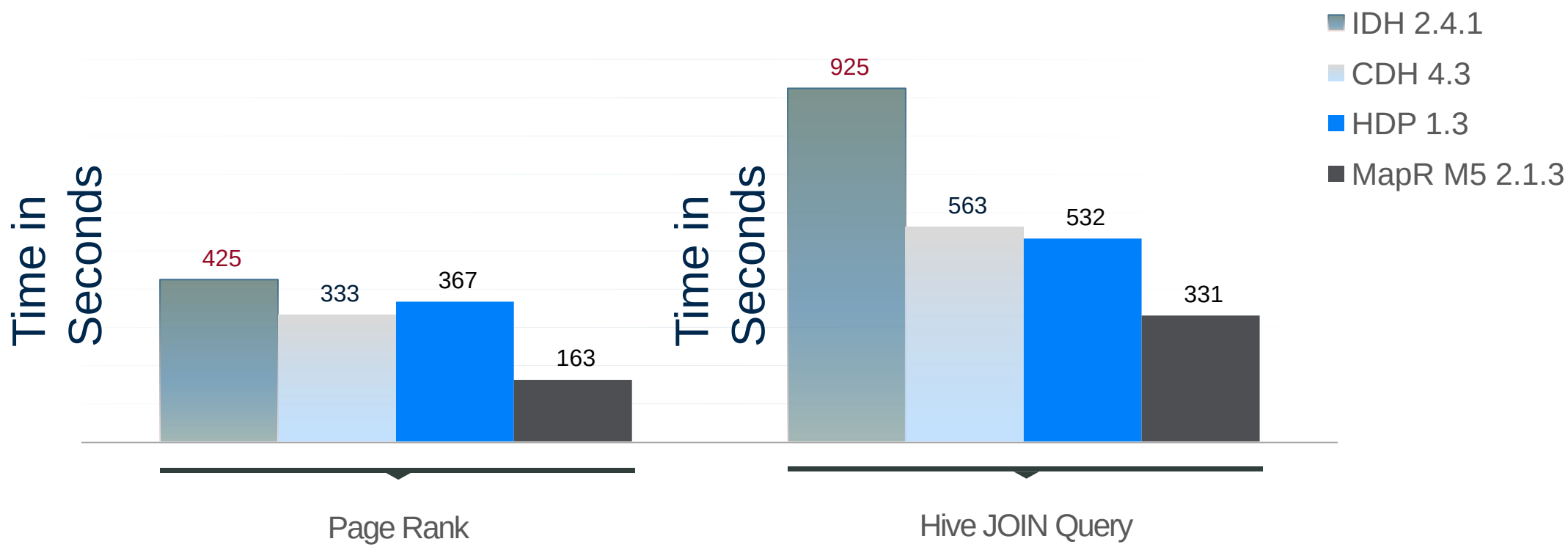
◆ Update
◆ New

Performance

Flux7: Comparative Study of Hadoop Distributions

Web Search and Data Analytics Benchmarks

Lower is Better

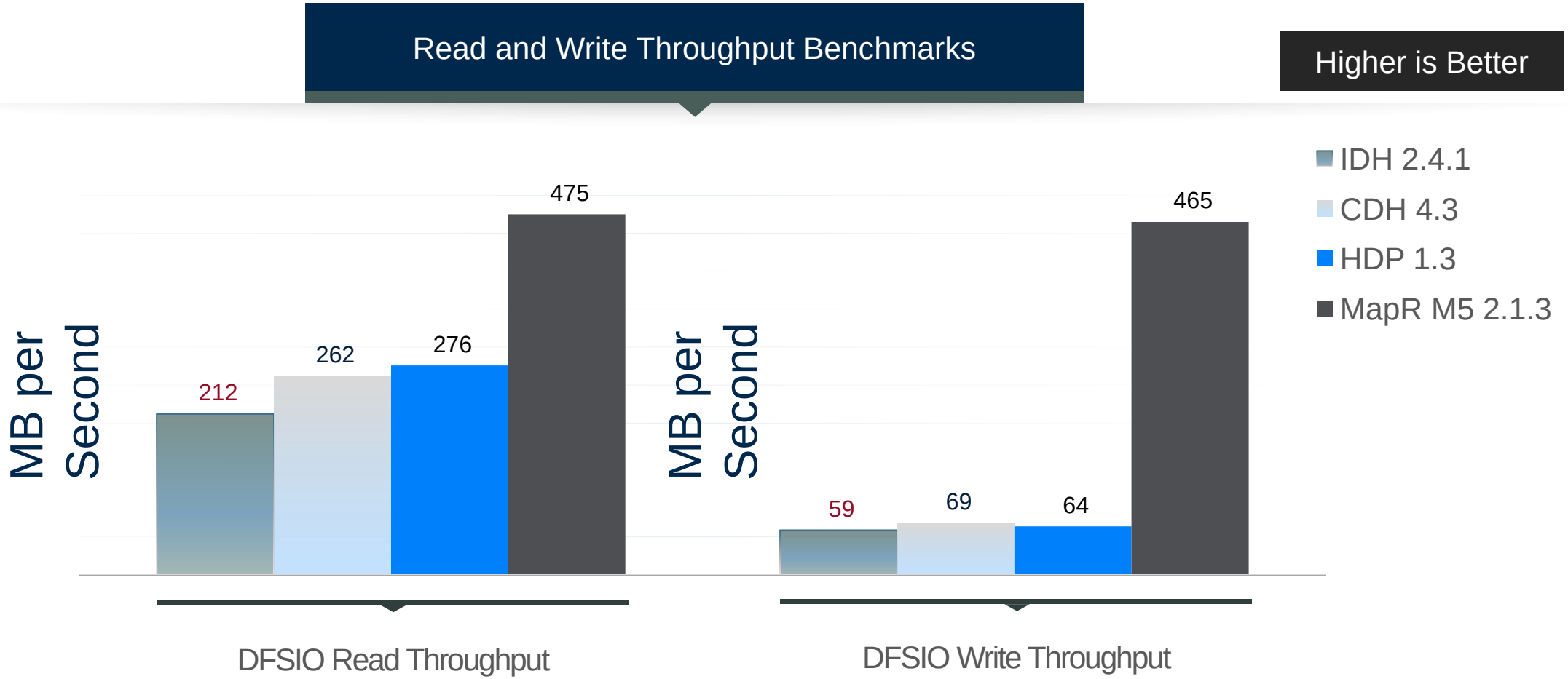


Source: Flux7 Labs Study, October 2013

Hardware Specs: EC2 on AWS

1 Master: m1.xlarge; 64-bit; 4 vCPU, 8 ECU; 15 GiB RAM; 4x420 GB Storage; 4x Intel ® Xeon ® CPU E5-2650 0 @ 2.00 GHz
4 Slaves: m1.large; 64-bit; 2 vCPU, 4 ECU; 7.5 GiB RAM; 2x420 GB Storage; 2x Intel ® Xeon ® CPU E5430 @ 2.66 GHz

Flux7: Comparative Study of Hadoop Distributions



Source: Flux7 Labs Study, October 2013

Hardware Specs: EC2 on AWS

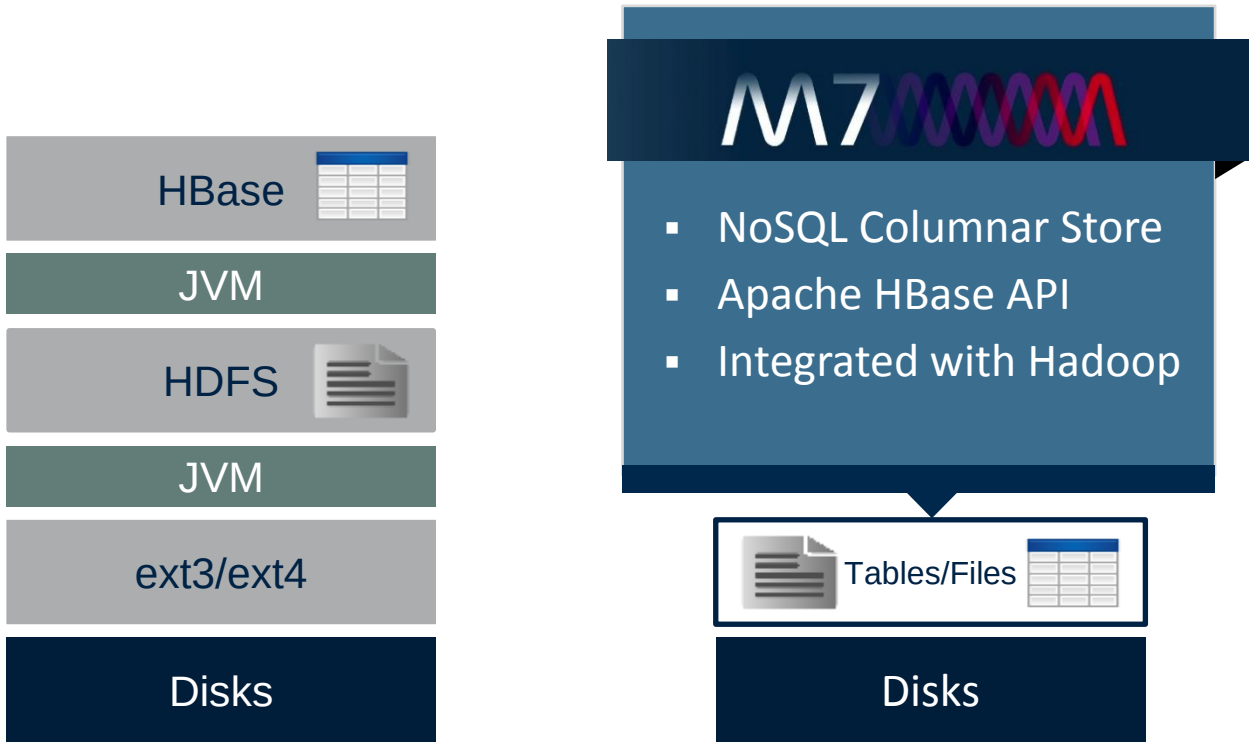
1 Master: m1.xlarge; 64-bit; 4 vCPU, 8 ECU; 15 GiB RAM; 4x420 GB Storage; 4x Intel ® Xeon ® CPU E5-2650 0 @ 2.00 GHz
4 Slaves: m1.large; 64-bit; 2 vCPU, 4 ECU; 7.5 GiB RAM; 2x420 GB Storage; 2x Intel ® Xeon ® CPU E5430 @ 2.66 GHz

M7: In-Hadoop Database

Operational Applications and Analytics Combined

MAPR

MapR M7: The Best In-Hadoop Database

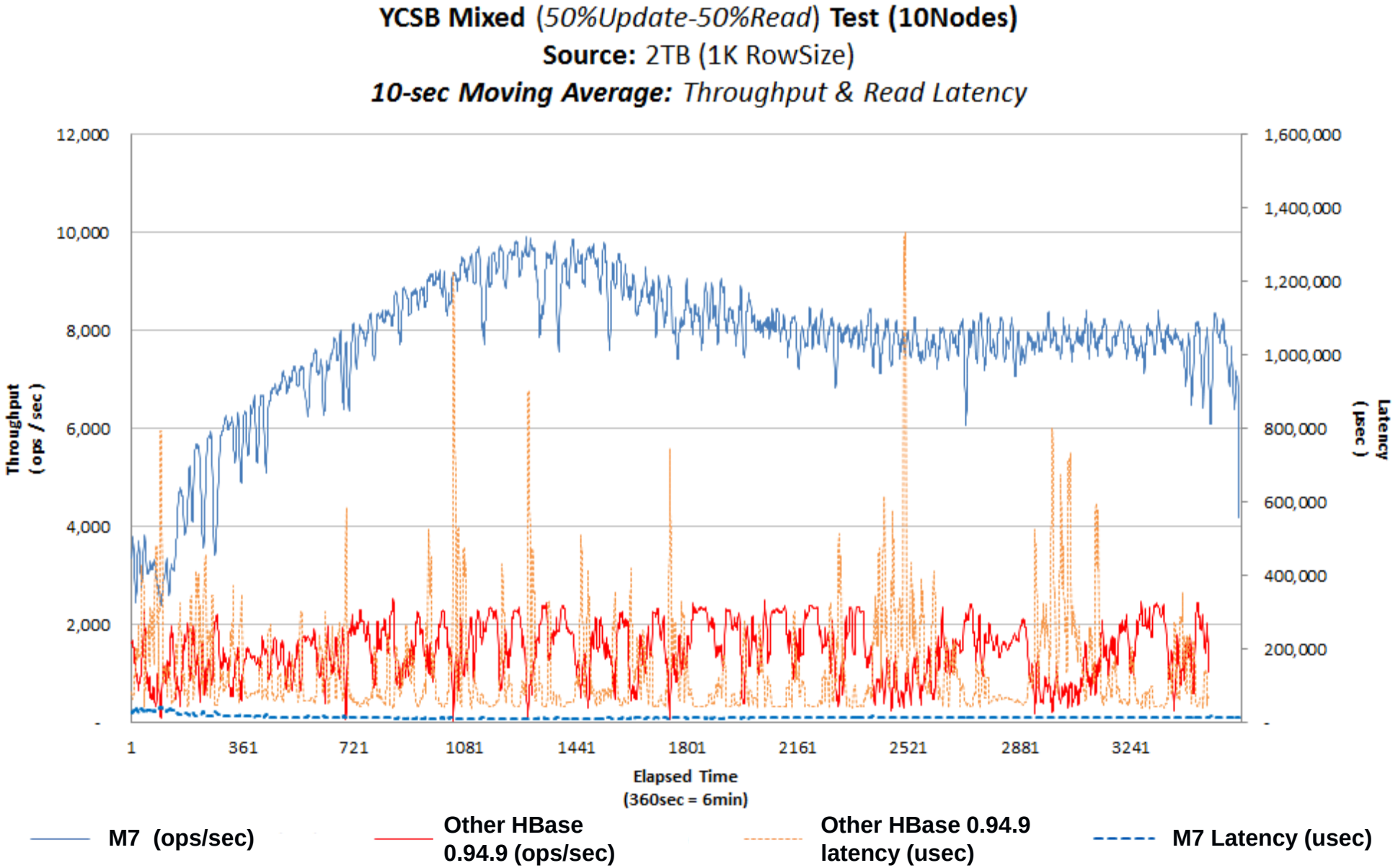


Other Distros

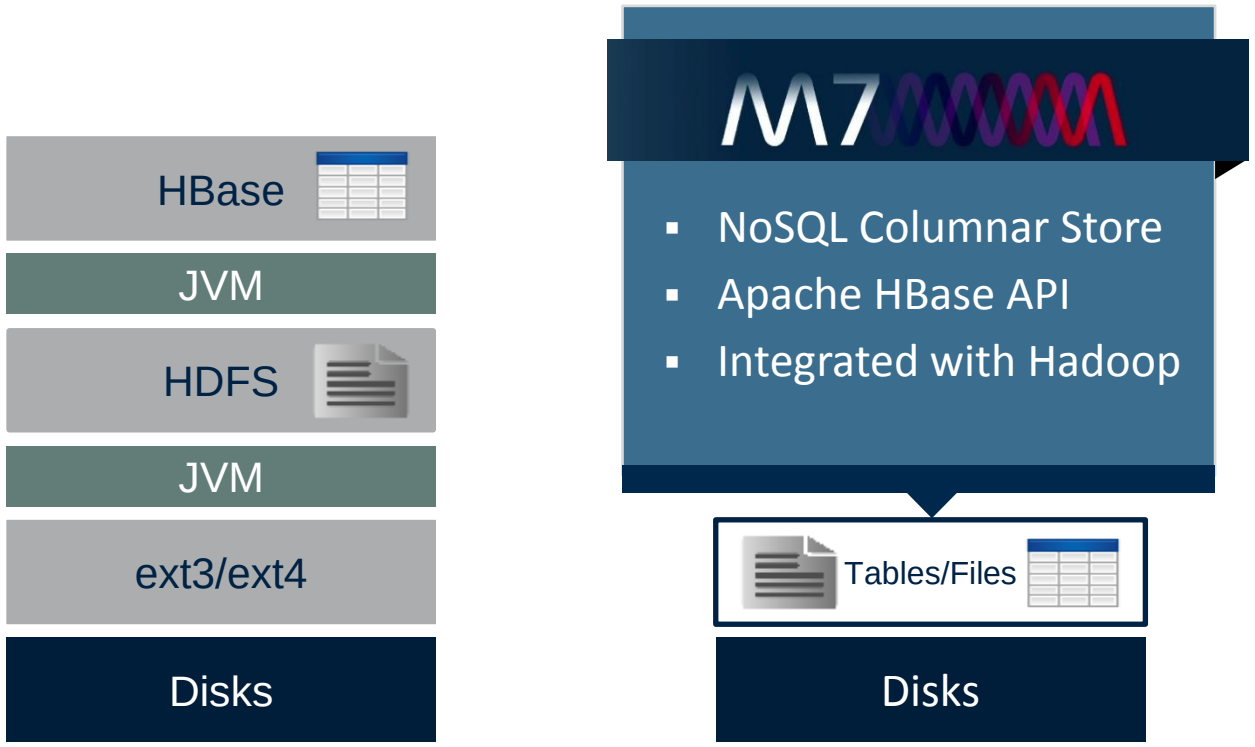
MapR M7

The most scalable, enterprise-grade,
NoSQL database that supports online applications and analytics

HBase Apps: High Performance with Consistent Low Latency



MapR M7: The Best In-Hadoop Database



Other Distros **MapR M7**

The most scalable, enterprise-grade,
NoSQL database that supports online applications and analytics

MapR Editions

M3

STANDARD EDITION

- Control System
- NFS Access
- Performance
- Unlimited Nodes
- Free

M5

ENTERPRISE EDITION

- Control System
- NFS Access
- Performance
- High Availability
- Snapshots & Mirroring
- 24 X 7 Support
- Annual Subscription

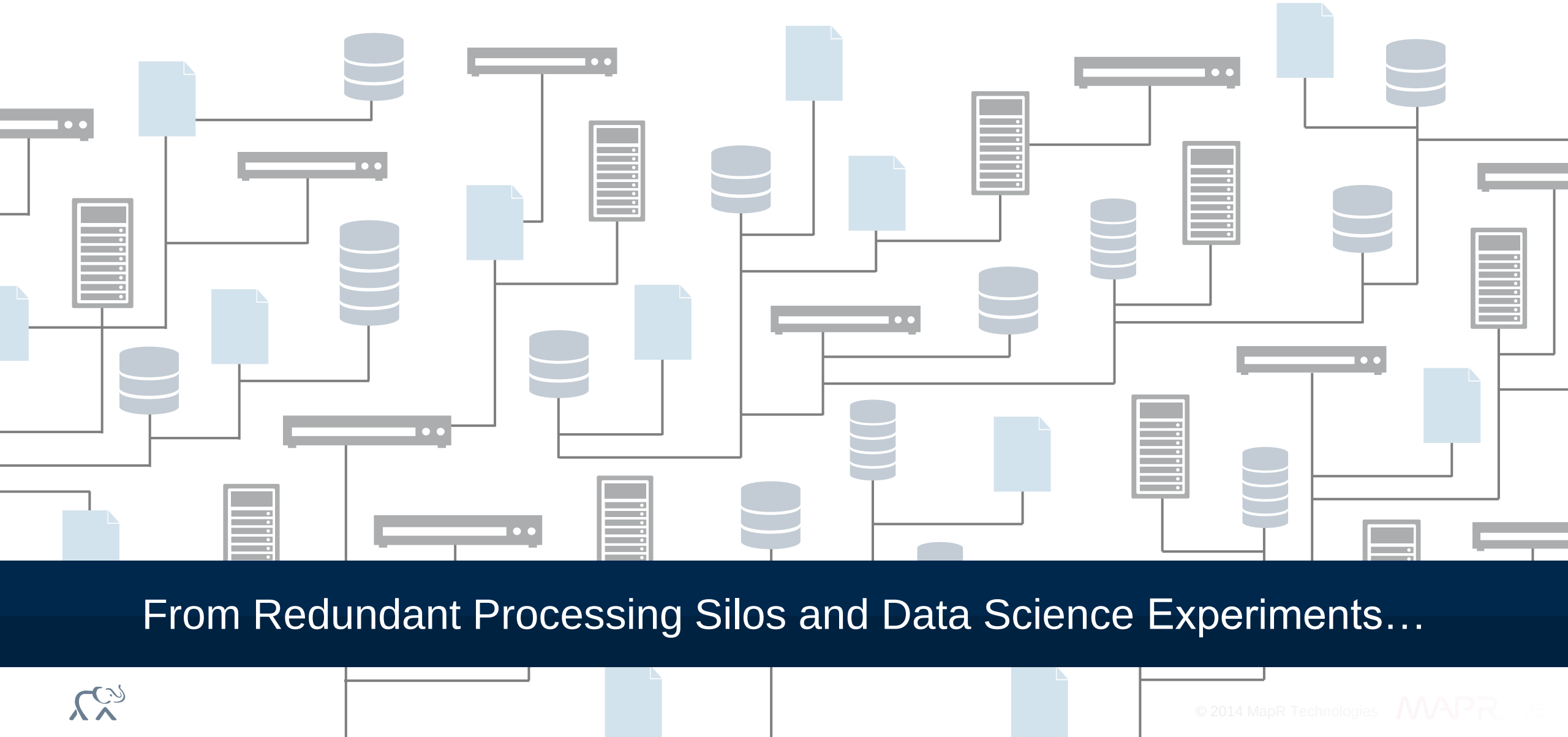
M7

ENTERPRISE DATABASE EDITION FOR HADOOP

- All the Features of M5
- Simplified Administration for HBase
- Increased Performance
- Consistent Low Latency
- Unified Snapshots, Mirroring

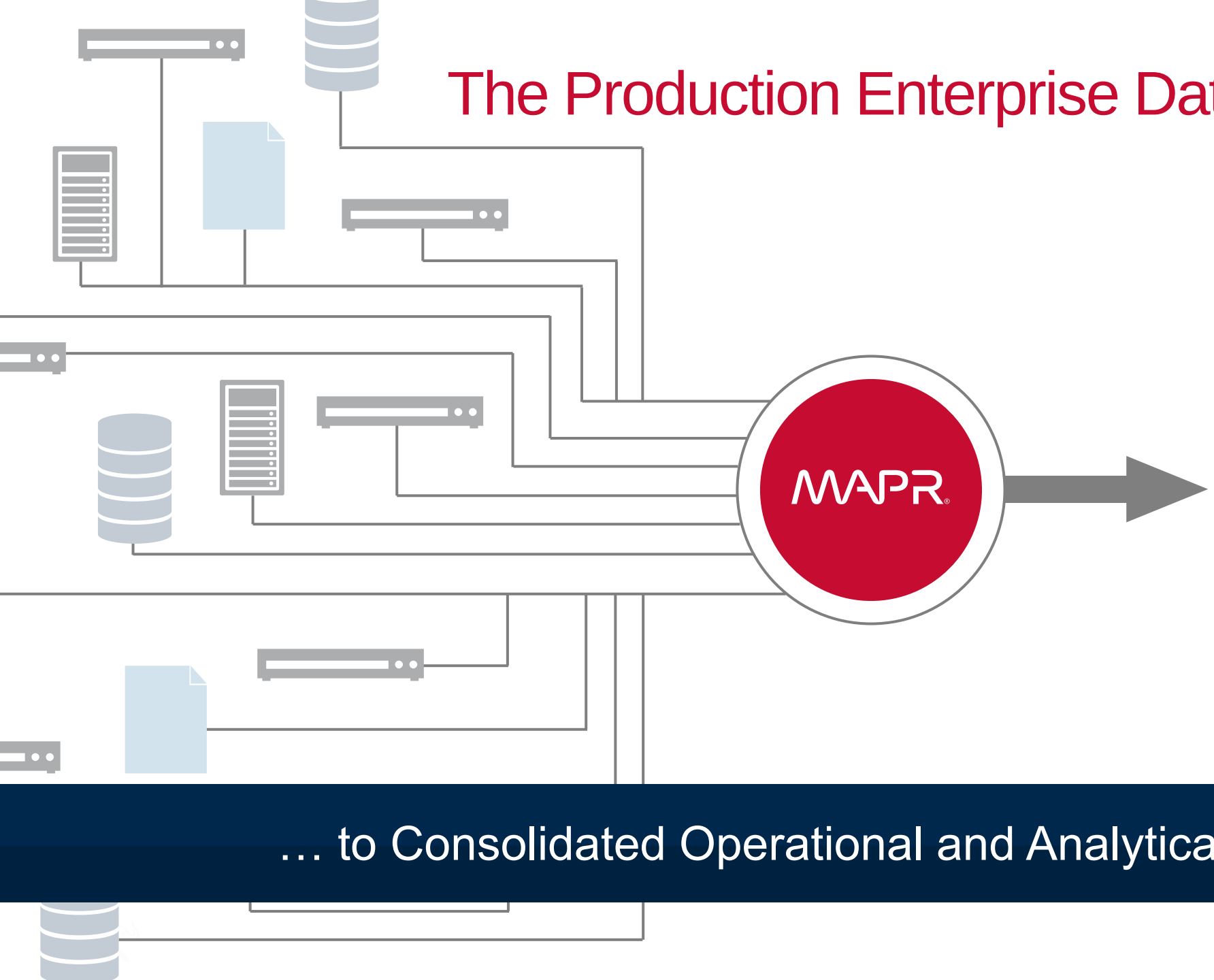
Fastest On-Ramp:
MapR Sandbox for Hadoop

Opportunity to Revolutionize Enterprise Data Architecture



From Redundant Processing Silos and Data Science Experiments…

The Production Enterprise Data Hub



… to Consolidated Operational and Analytical Workloads

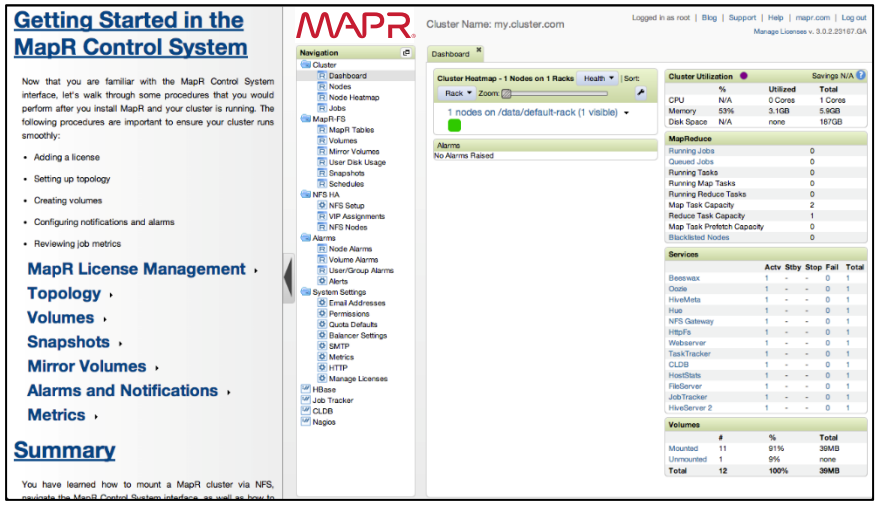
MapR Sandbox for Hadoop

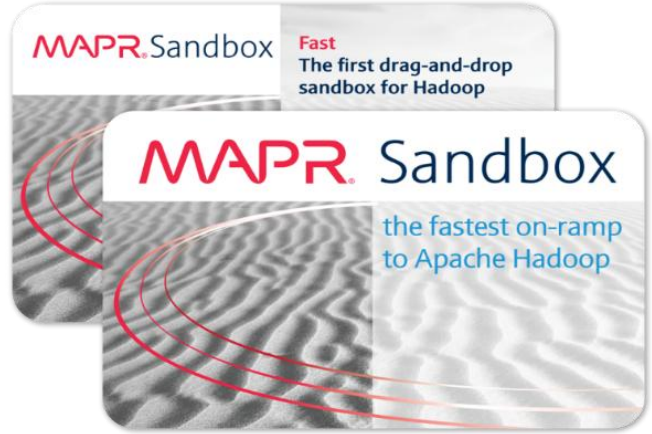
The Fastest On-Ramp to Hadoop

- Complete MapR distribution for Hadoop
  - Free download
  - Most advanced distribution

- Tutorials and advanced user interfaces
  - MapR Control System (MCS) for administrators
  - Hadoop User Experience (HUE) for developers
  - Point-and-click tutorials

- Fully configured in virtual machines
  - Supports VMware and VirtualBox
  - Drag-and-drop data movement





Q&A

Engage with us!

@mapr / @agoujet   maprtech

mapr-technologies   MapR

agoujet@mapr.com   maprtech