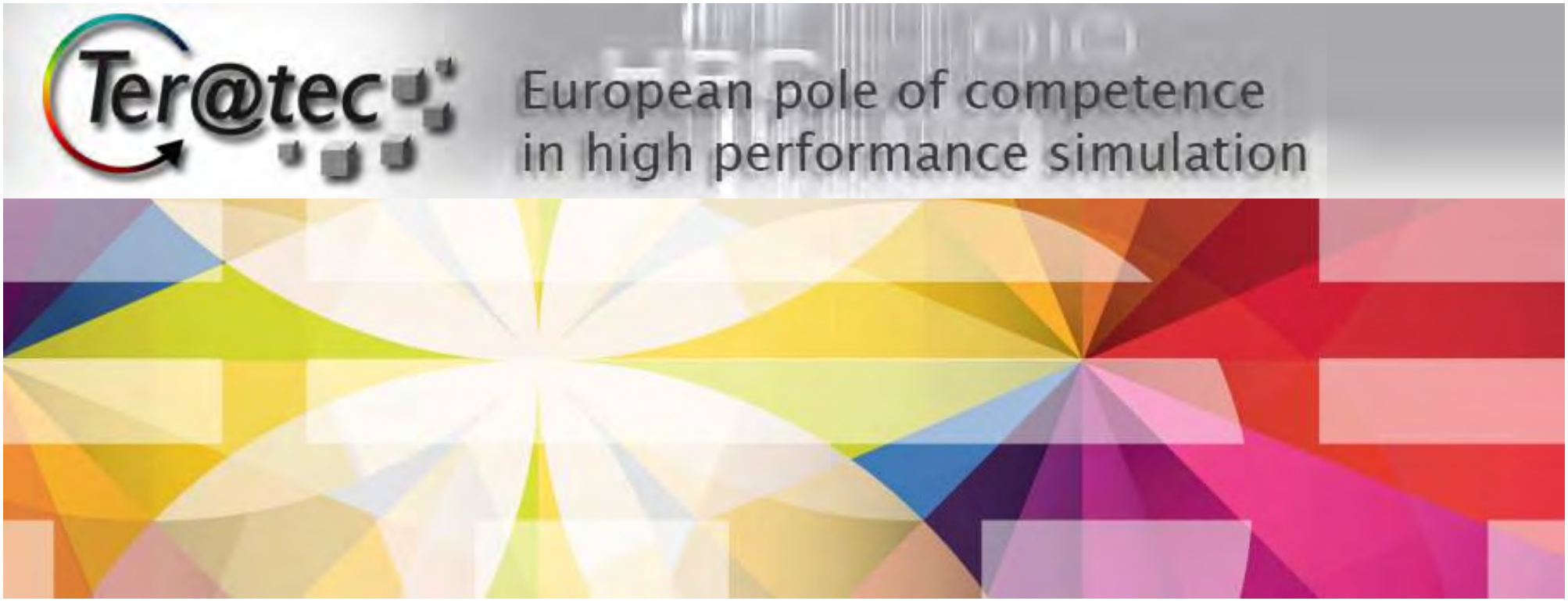
Addressing Open Source Big Data, Hadoop, and MapReduce limitations



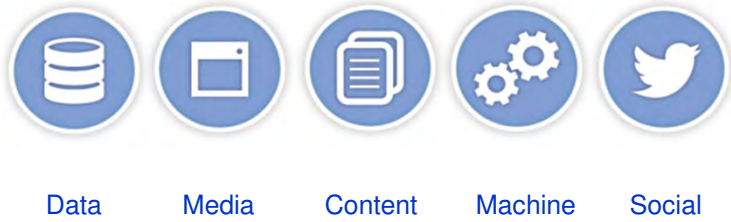
Agenda

- What is Big Data / Hadoop ?
- Limitations of the existing hadoop distributions
- Going enterprise with Hadoop

How Big are Data ?

90 % of data in the world has been created

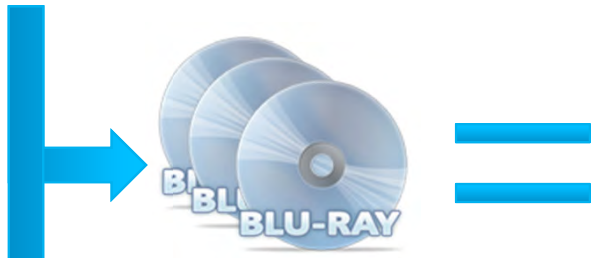
last 2 years



IDC says the digital universe will represent

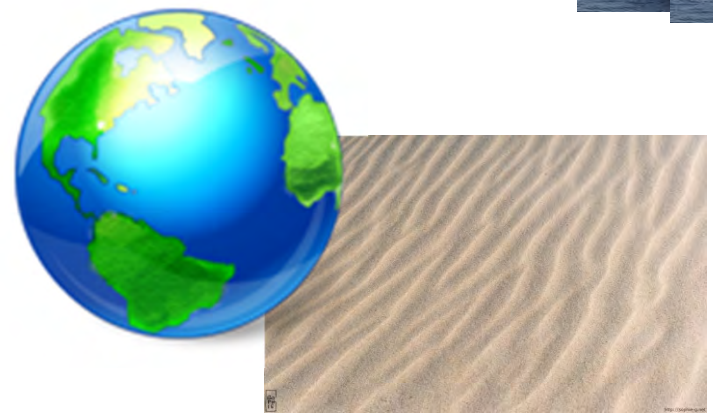
40 zettabytes in 2020

1 000 000 000 000 000 000 000



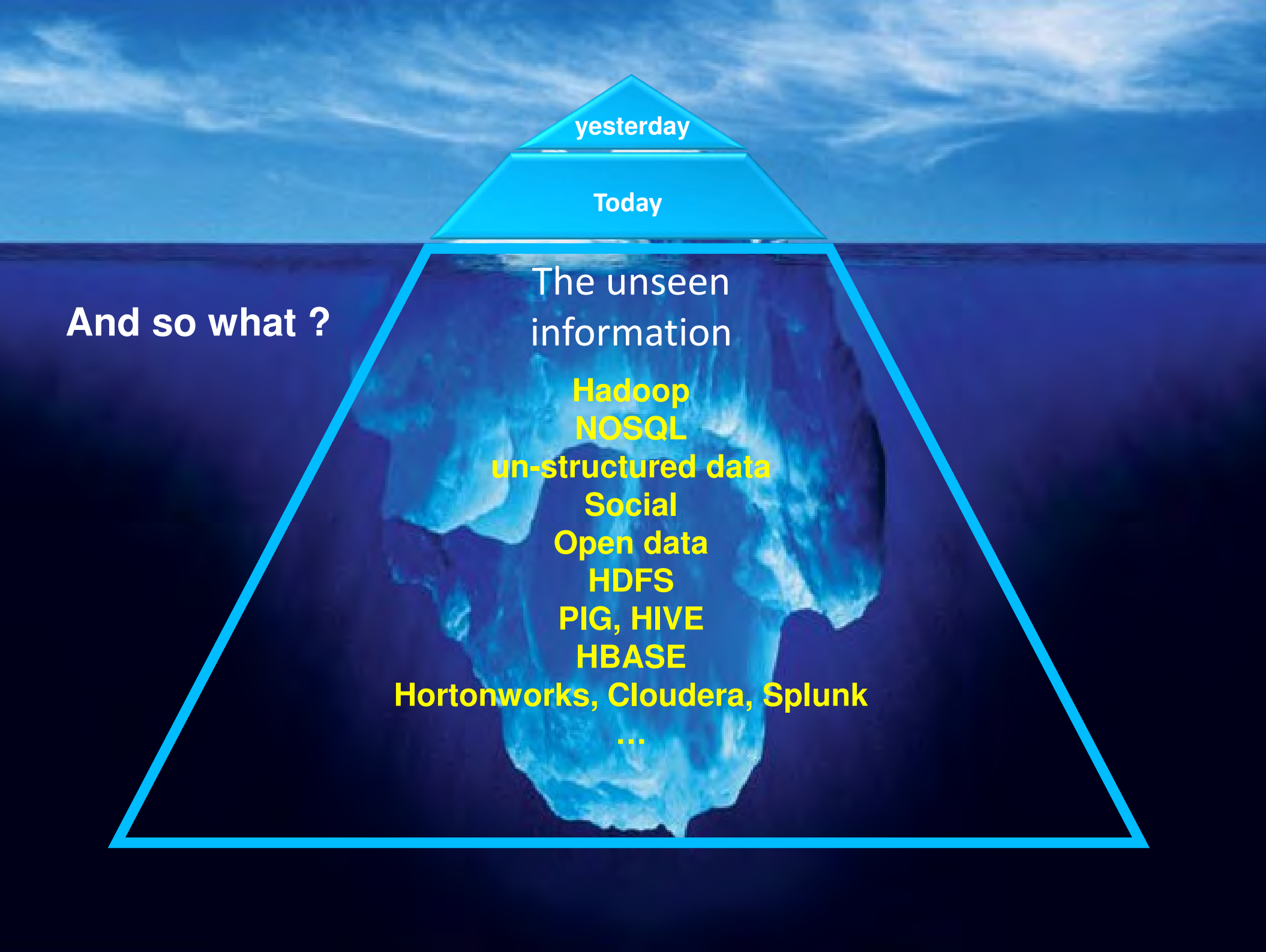
=

... × 424



× 4

1 byte


yesterday
Today
And so what ?
The unseen information
Hadoop
NOSQL
un-structured data
Social
Open data
HDFS
PIG, HIVE
HBASE
Hortonworks, Cloudera, Splunk
…

What can you do with big data?

Financial Services
- Fraud detection
- Risk management
- 360° View of the Customer

Utilities
- Weather impact analysis on power generation
- Transmission monitoring
- Smart grid management

Transportation
- Weather and traffic impact on logistics and fuel consumption

IT
- Transition log analysis for multiple transactional systems
- Cybersecurity

Health & Life Sciences
- Epidemic early warning system
- ICU monitoring
- Remote healthcare monitoring

Retail
- 360° View of the Customer
- Click-stream analysis
- Real-time promotions

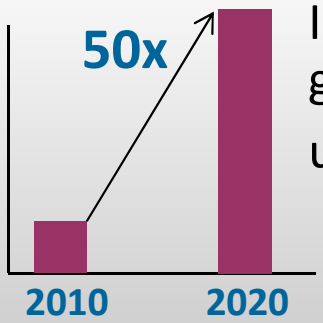
Telecommunications
- CDR processing
- Churn prediction
- Geomapping / marketing
- Network monitoring

Law Enforcement
- Real-time multimodal surveillance
- Situational awareness
- Cyber security detection

Big Data Landscape (Version 2.0)

Market definition

Volume



Information growing up to ZB

Velocity

Process information of 30 billions of RFID chips

Variety

80% of worldwide data are unstructured

Veracity

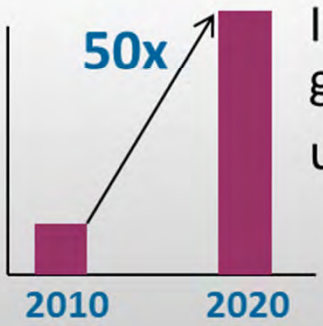
1/3 makers do not trust their information system to make decisions.

83% of CIO says analytics is the first path to competitiveness*

*IBM CIO Study

Hadoop scope

Volume



Information growing up to ZB

Velocity

Process information of 30 billions of RFID chips

Variety

80% of worldwide data are unstructured

Hadoop focus on processing huge volume of different type of data

Do not confuse Big Data and Hadoop

- Big data is a generic term:

Big data *is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications (http://en.wikipedia.org/wiki/Big_data)*

- Hadoop:

Apache Hadoop is an open source framework that allows for distributed processing of large data sets across computing clusters and is the most widely used technology for Big Data processing.
Hadoop framework has evolved into a set of tools and technologies to efficiently process, store and analyze huge amounts of varied data in a linear, scalable and reliable fashion.
It is important to understand that Hadoop is not a complete replacement for the traditional enterprise Data Warehousing and Business Intelligence tools, but is a complementary approach to solve some of its challenges

Hadoop

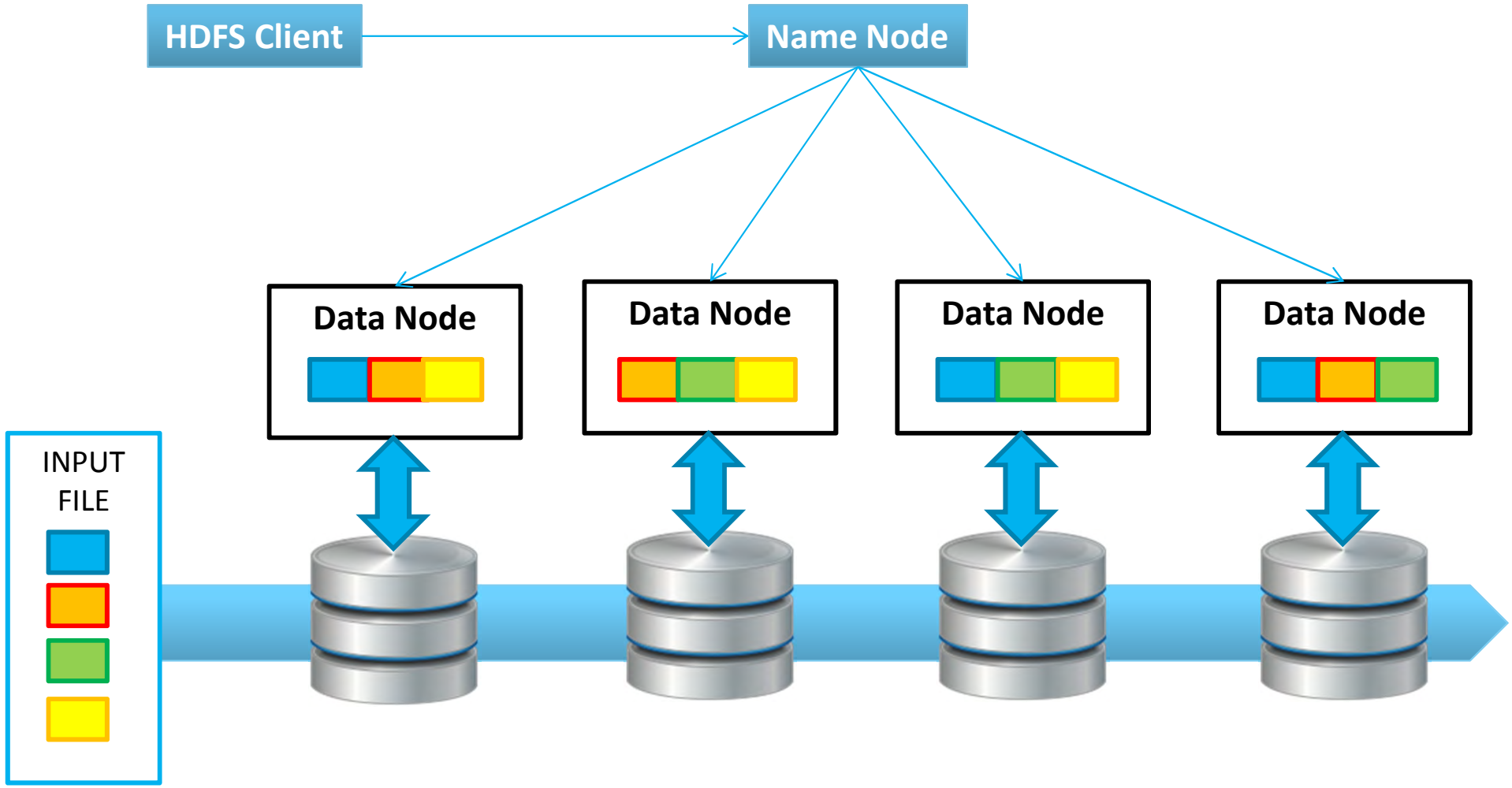
=

Distributed data processing

+

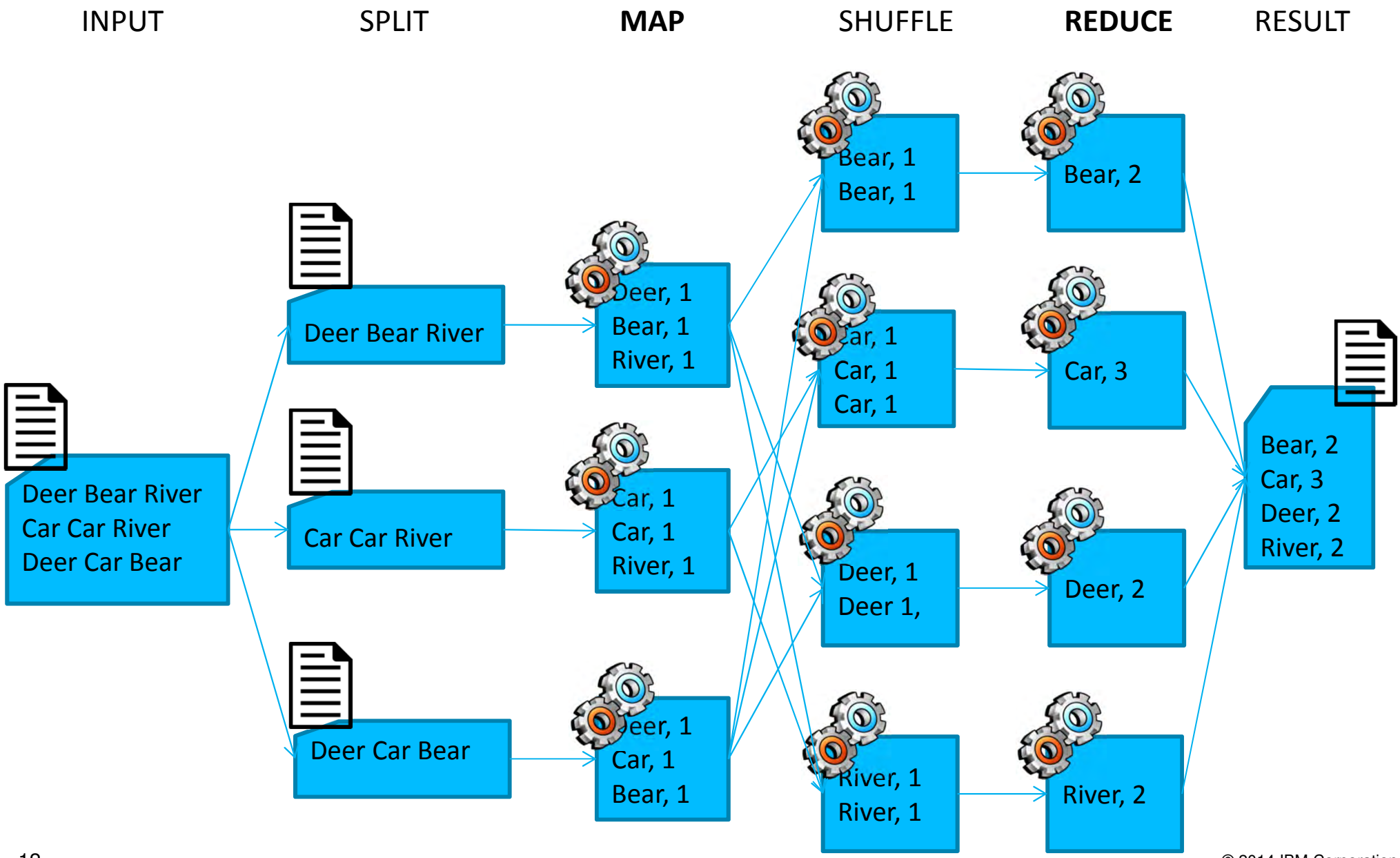
Distributed data storage

Hadoop HDFS



Elastic Storage can replace HDFS

Hadoop Mapreduce



Is it HPC 2.0 ?

It is HPDA…

Traditionally		Increasingly
GOVT, EDU & RESEARCH Exotic and expensive	→	**WIDESPREAD** Accessible to all Entertainment, Finance, Gaming, xSPs etc …
BATCH Command line driven workloads	→	**BATCH + REAL-TIME SOA** API driven – throughput and latency are key
DEDICATED Homogeneous Infrastructure dedicated to specific applications	→	**DEDICATED + SHARED** Heterogeneous infrastructure shared fluidly among groups and applications
STRUCTURED DATA problems involve mostly structured data	→	**STRUCTURED + UNSTRUCTURED** Data types unstructured or semi-structured – email, video, documents, logs

Common pain points

High-Availability

- Limited HA features in the workload engine
- HDFS NameNode lacks automatic failover logic

Not scalable

- Large performance overhead during job initiation
- Scalability concerns

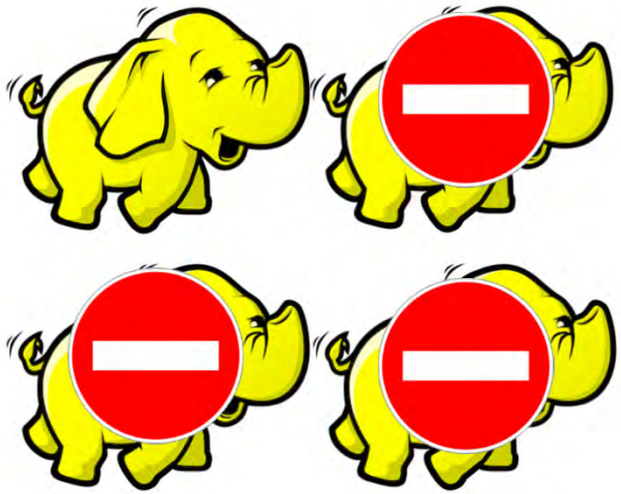
No resources sharing

- Resource silos associated with MapReduce applications
- No way to manage a shared services model tied to an SLA
- Single purpose clusters - under utilized resources

Common pain points

Lack of sophisticated scheduling engine

- Large jobs can still hog cluster resources
- Lack of real time resource monitoring
- Lack of granularity in priority management

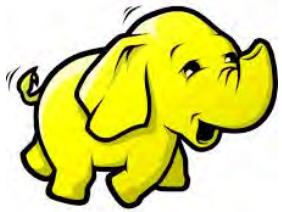


Lack of application life cycle / rolling upgrades

- No ability to run multiple Hadoop middleware in parallel
- No ability to manage/deploy multiple Hadoop versions

IBM Platform Symphony unique capabilities

Multi-tenant



100% Hadoop

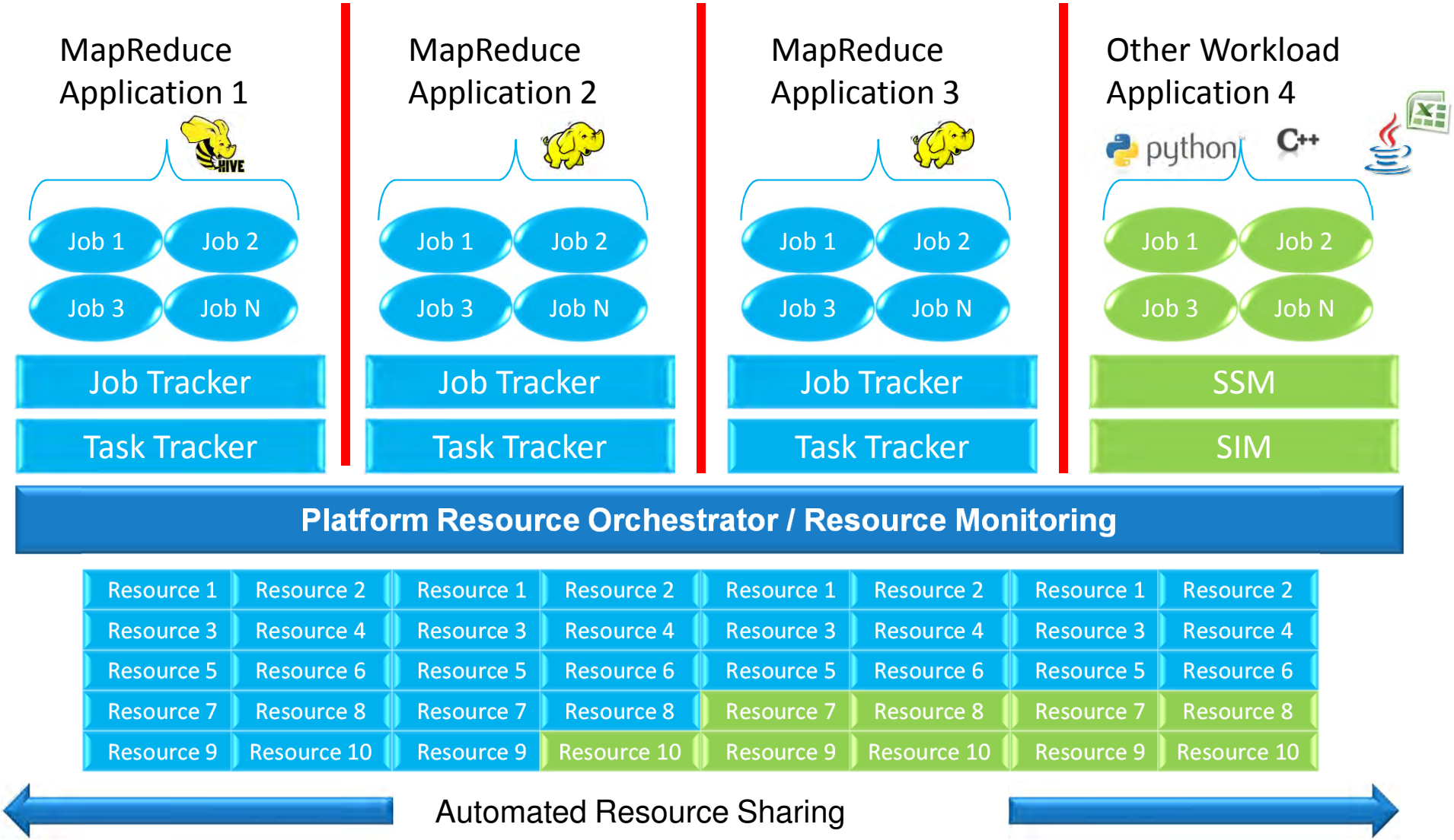
Performances

Scalable

Robustness

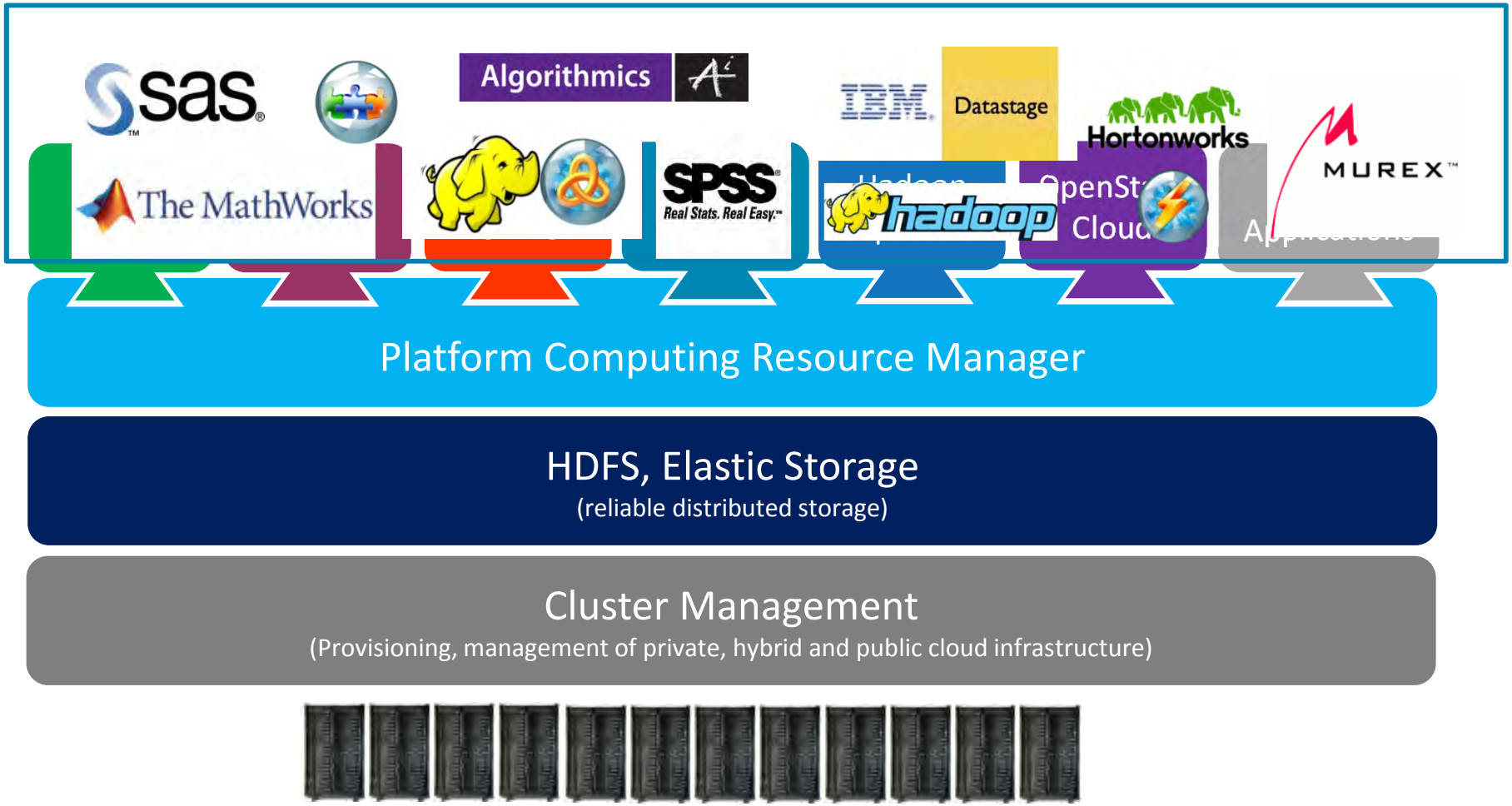
Production ready

Symphony brings unique capabilities to Big Data



Multitenancy in IBM Platform Symphony

Customers need to support diverse applications – both Hadoop and non-Hadoop



With sophisticated multitenancy, customers can share a broader set of application types and scheduling patterns on a common resource foundation

IBM InfoSphere BigInsights For Hadoop

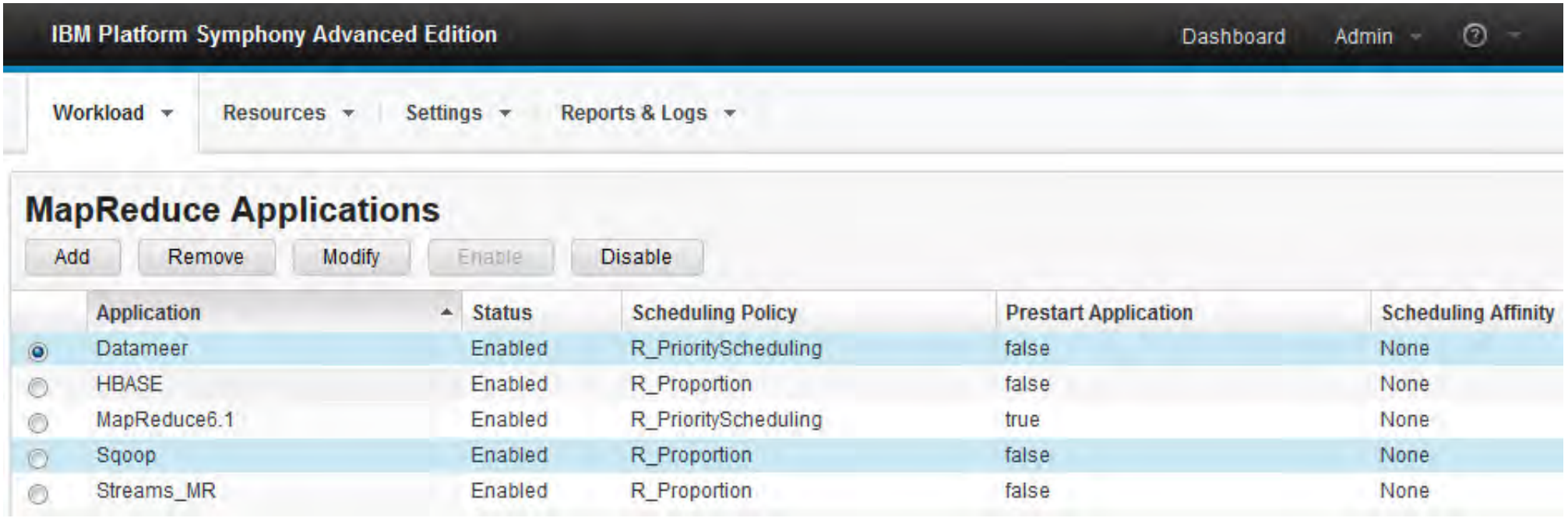
4x performance gain on average over open source Hadoop[1]

Audited STAC Report™
Securities Technology Analysis Center

1. 4x is approximate value. *See the* STAC Report™ *at http://www.stacresearch.com/node/15370. Testing involved the SWIM benchmark (https://github.com/SWIMProjectUCB/SWIM) and jobs derived from production workload traces. Testing was conducted in controlled laboratory conditions.*

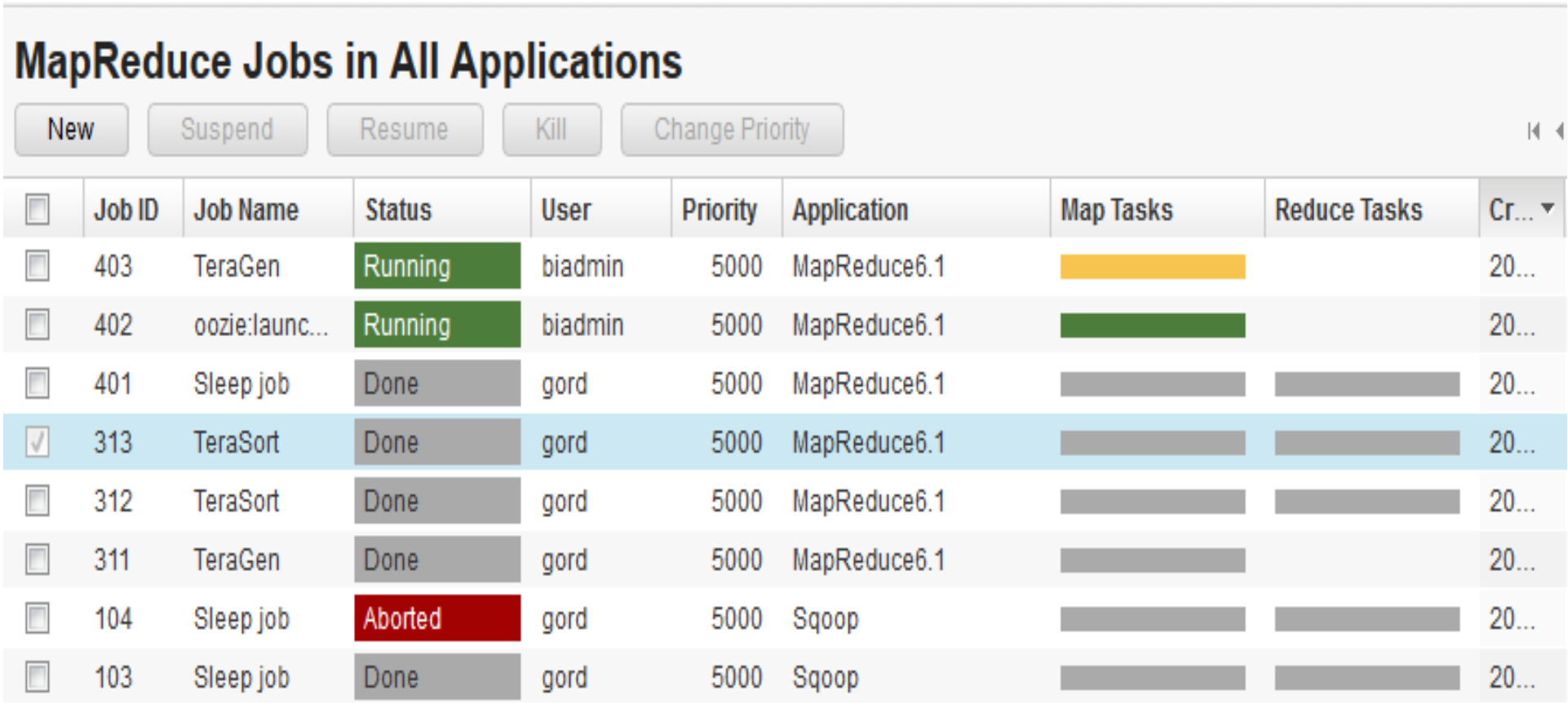
A Single Dashboard

Manage multiple tenants including BigInsights and third party workloads



Transparent access

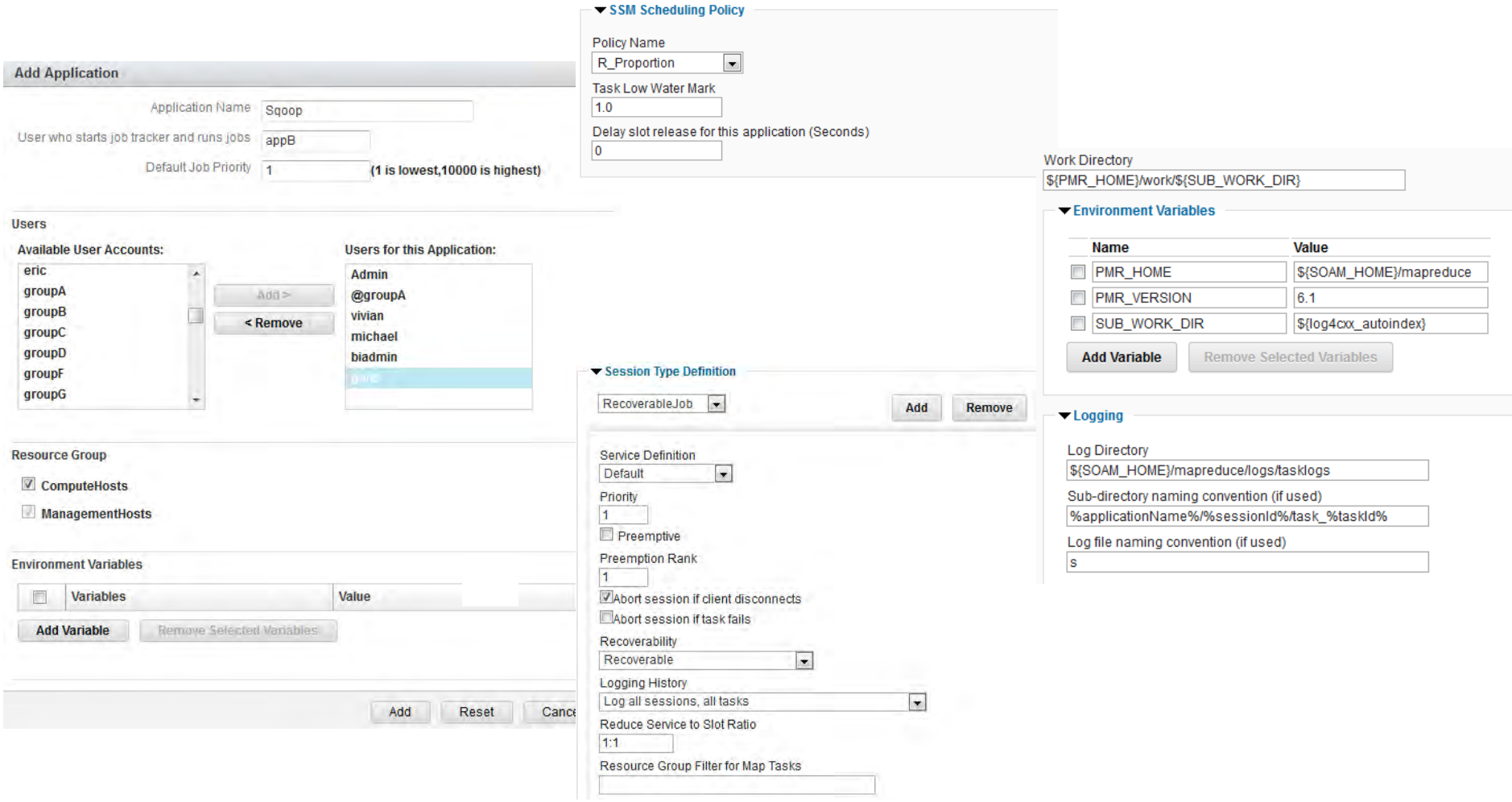
Applications think they have their own private cluster, but with configurable access to shared data



Job ID	Job Name	Status	User	Priority	Application	Map Tasks	Reduce Tasks	Cr...
403	TeraGen	Running	biadmin	5000	MapReduce6.1			20...
402	oozie:launc...	Running	biadmin	5000	MapReduce6.1			20...
401	Sleep job	Done	gord	5000	MapReduce6.1			20...
313	TeraSort	Done	gord	5000	MapReduce6.1			20...
312	TeraSort	Done	gord	5000	MapReduce6.1			20...
311	TeraGen	Done	gord	5000	MapReduce6.1			20...
104	Sleep job	Aborted	gord	5000	Sqoop			20...
103	Sleep job	Done	gord	5000	Sqoop			20...

Flexible configuration of tenant applications

Easily solve problems that normally would prevent workloads from sharing infrastructure



Configurable, Dynamic Resource Sharing

Establish configurable resource sharing policies that ensure service levels while maximizing utilization

Resource Plan

Resource Group: ComputeHosts | Time Intervals and Settings

▶ Slot allocation policy

00:00
24:00

Consumer	Model type: Ownership				Model type: Share	
	Owned Slots	Consumer Rank	Lend | Limit	Borrow | Limit	Share Ratio	Limit
▼PSMR61	192					
▸SymTesting	0	0			☑ 1	
▾SampleApplications	0	0			☑ 1	
SOASamples	0	0	☐	☐	☑ 1	
EclipseSamples	0	50	☐	☐	☑ 1	
Total	0	-	-	-	-	-
Balance	0	-	-	-	-	-
▸SymExec	0	0			☑ 1	
▾MapReduceConsumer	128	0			☑ 1	
MapReduce61	64	0	☑ 32 Details	☐	☑ 1	
Datameer	64	0	☑ 32 Details	☐	☑ 1	
HBase	0	0	☐	☐	☑ 1	
Streams_MR	0	0	☐	☐	☑ 1	
Sqoop	0	0	☐	☐	☑ 1	
Total	128	-	-	-	-	-
Balance	0	-	-	-	-	-
Total	128	-	-	-	-	-
Balance	64	-	-	-	-	-

▸ Expand All
▾ Collapse All

Apply | Revert | Import... | Export...

USAA deploys a multitenant, shared infrastructure

A shared environment for multiple Hadoop analytic applications

Business problem

- Multiple lines of business deploying new applications driving infrastructure cost
- Rapid growth in storage requirements, traditional DW too expensive
- Need to facilitate rapid development and deployment of new Hadoop applications
- obtaining an information advantage
- Investing in commercial and in-house Hadoop based solutions to enhance existing data warehouse
- Concerned about uncontrolled growth of Hadoop environments as multiple lines deploy the technology

Big Data Solution

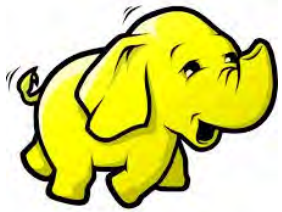
- InfoSphere BigInsights for Hadoop Services and Analytics
- IBM Platform Symphony for multi-tenancy
- IBM Elastic Storage (GPFS FPO)

Business Result

- Significant infrastructure cost avoidance – Approx 30 applications on a shared infrastructure
- Estimated 4x performance gain on average

IBM Platform Symphony unique capabilities

Multi-tenant



100% Hadoop

Performances

Scalable

Robustness

Production ready

Contact: Emmanuel Lecerf, Big Data expert EMEA

Emmanuel.lecerf@fr.ibm.com