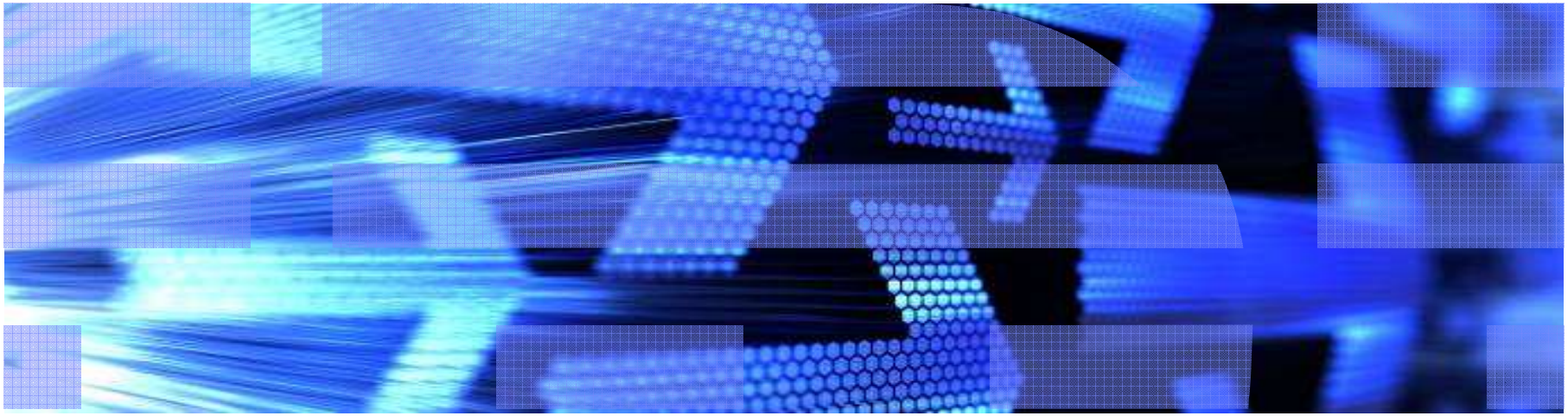
Pursuing Big Data with IBM Platform Symphony

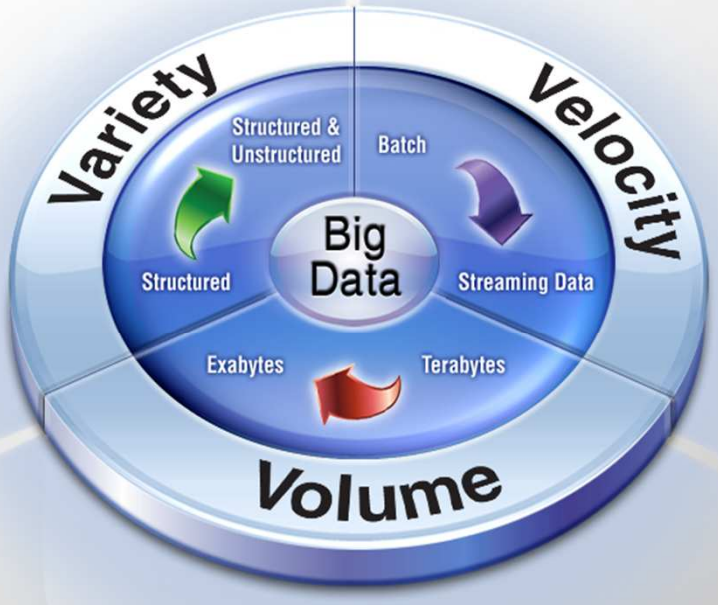
Philippe Bricard
Emmanuel Lecerf

Platform Computing
An IBM Company



Smarter Decisions for Optimized Performance

IBM Big Data = Volume, Variety and Velocity

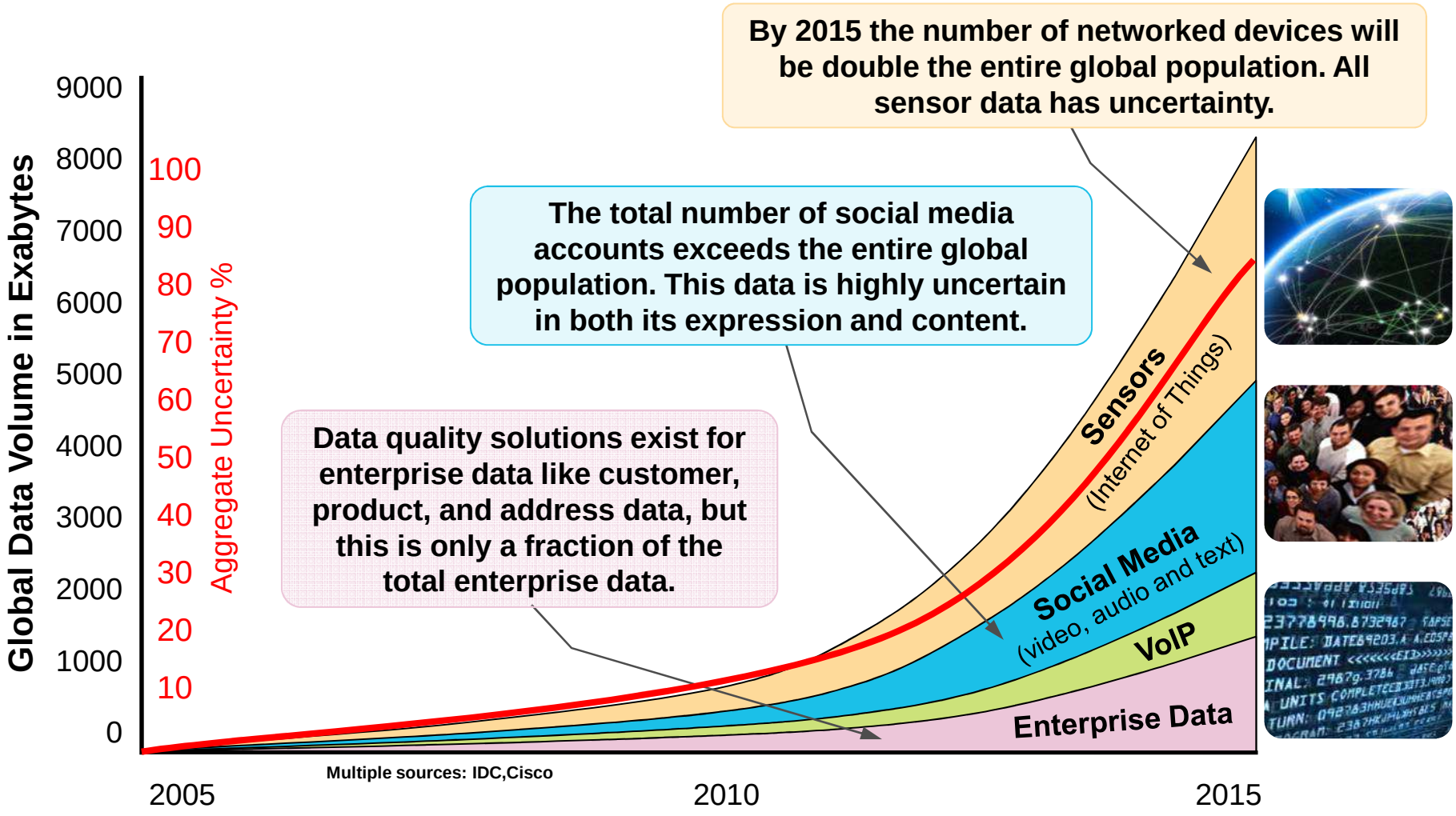


Volume: Scale from terabytes to zettabytes

Variety: Relational and non-relational data types from an ever-expanding variety of sources.

Velocity: Streaming data and large volume data movement

By 2015, 80% of all available data will be uncertain: Veracity



Big Data Customer Examples

Big Data for the CMO

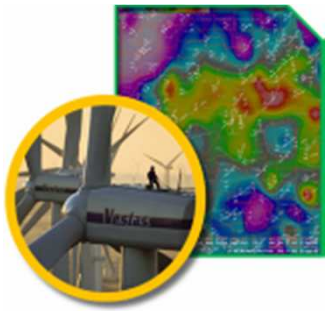
BBVA

”Listen to the voice of clients”

5.8 terabytes of Internet and Social Media

Fix negative opinions and build on positive ones

Big Data for Smarter Planet



Vestas
No. 1 in Modern Energy

Reduced modeling time by 97%

2.8 petabytes of public and private weather data

Modeling time reduced from weeks to hrs.

Big Data for Telco

bharti

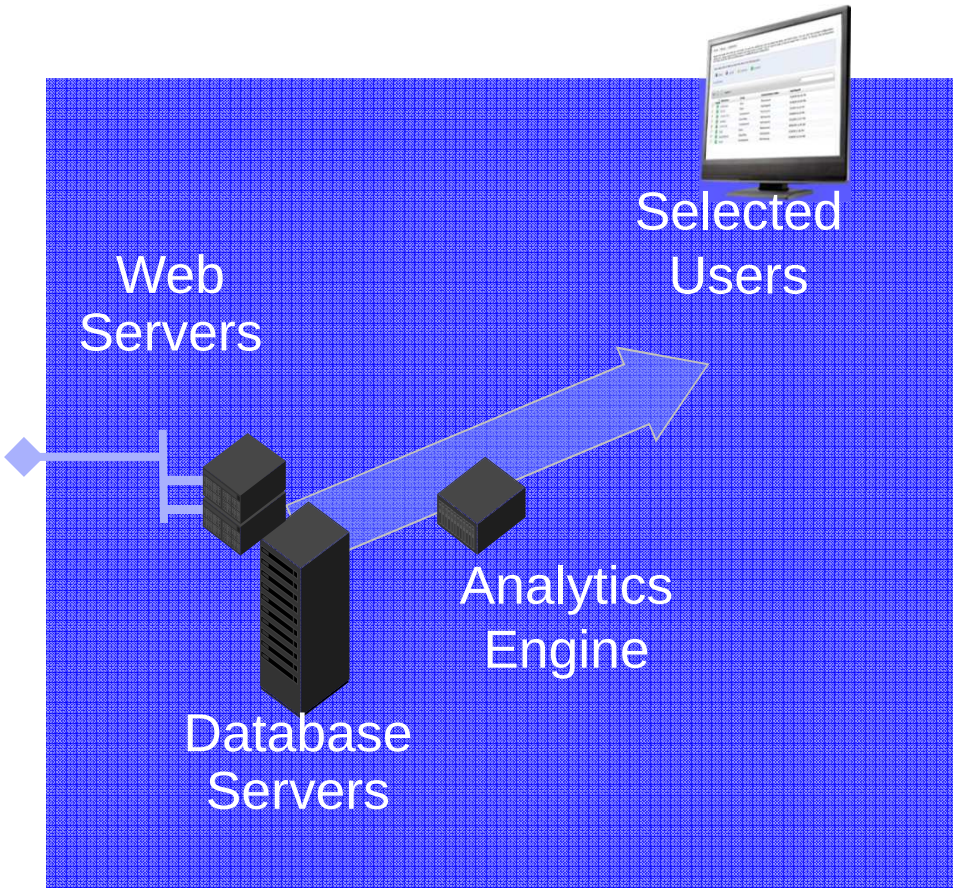
Latency reduced from 12 hrs to 1 sec

6 billion Call Detail Records per day

Personalized marketing to individual customers

Traditional Analytics Solution Architecture

- **Analytics** to examine past events and existing conditions
- **Web servers** provide minimal reference data, if any
- **Data Warehouses** and databases provide almost all of the information



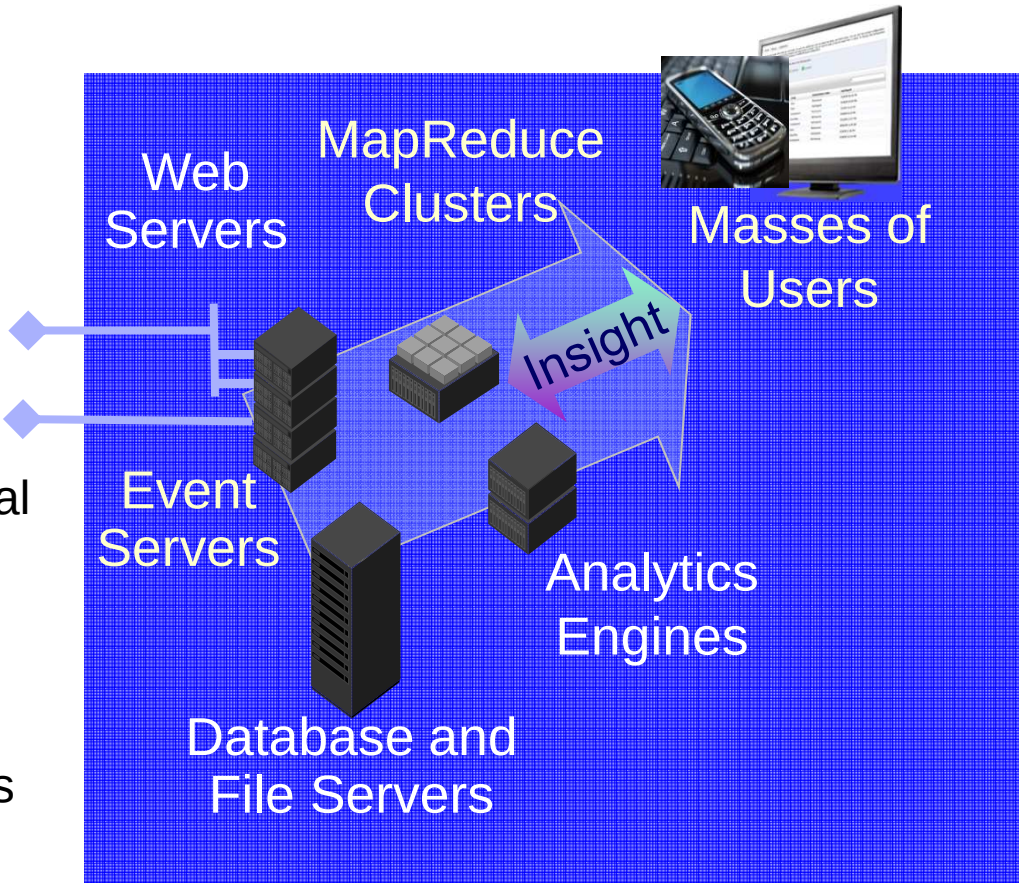
IBM Big Data Reference Architecture

- Multiple **Analytics Engines** predict outcomes and behavior
- **Web Servers** with crawlers and stream services gather vast amounts of external data
- **Data Warehouses**, Databases, and **File Servers** provide internal data
- **Map/Reduce Clusters** process unstructured data quickly

New!

- **Event Servers** manage streams from devices and networks

New!



A new infrastructure point of view

IBM

Text Analytics /Mining

Content Analytics
UIMA Based high volume unstructured management

Streams

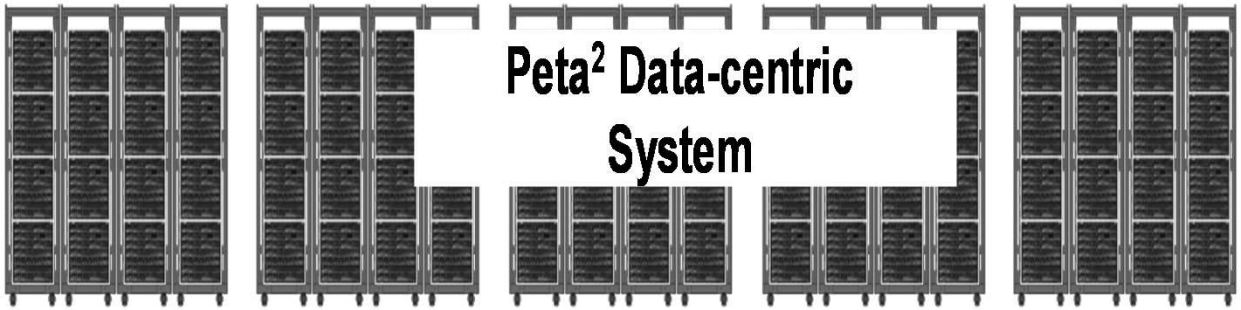
InfoSphere BigInsights
Hadoop-based low latency analytics for variety and volume

Hadoop

InfoSphere Streams
Low Latency Analytics for streaming data

Platform Computing
an IBM Company

GPFS



Committed to Open Source

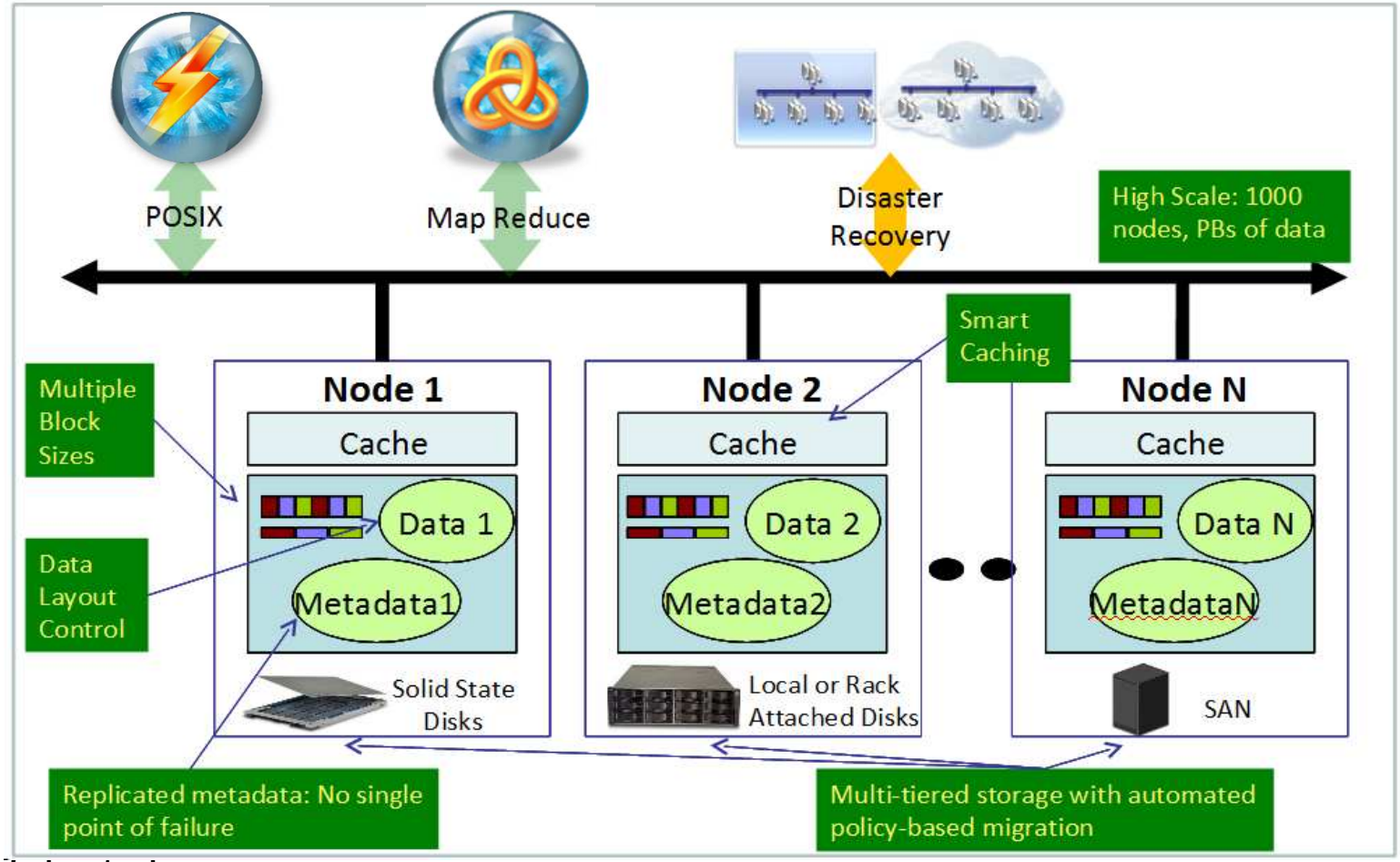
Committed to Innovation

Industry-grade Distributed File System

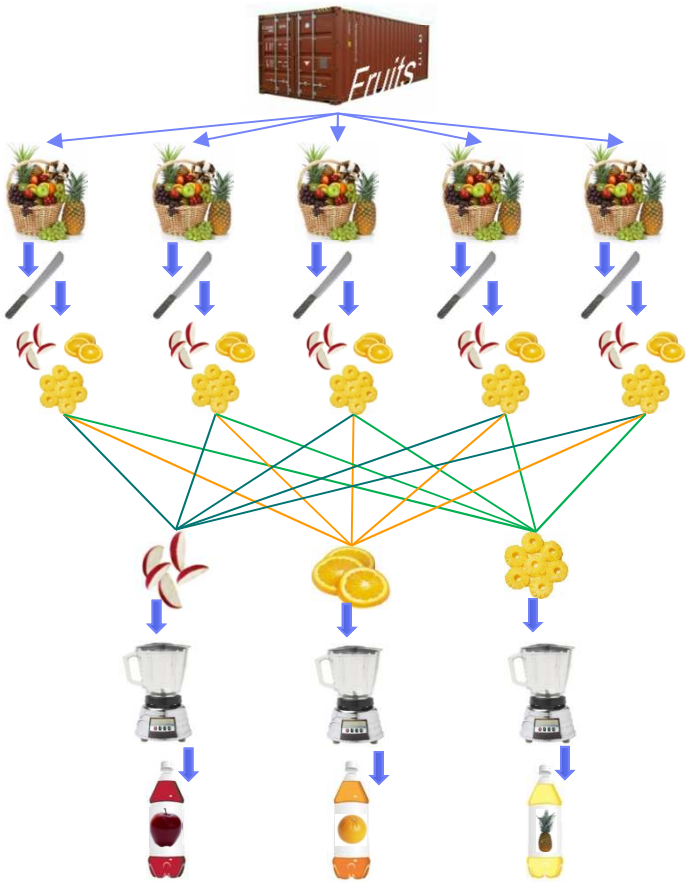
GPFS SNC (General Parallel File System for Shared Nothing Clusters)

- *GPFS-SNC* - scalable, high performance, and highly reliable file system
- Support applications on MapReduce and standard POSIX interfaces
- Optimized for online and batch processing applications

Beta



Map Reduce paradigm to support innovation



Each input to a map is a list of <key, value> pairs

(<a, > , <o, >, <p, > ..)

Each output of a map is a list of <key, value> pairs

(<a', > , <o', >, <p', > ,)

Grouped by key

Each input to a reduce is a <key, value-list>
(possibly a list of these, depending on the grouping/hashing mechanism)
e.g. <a', (...)>

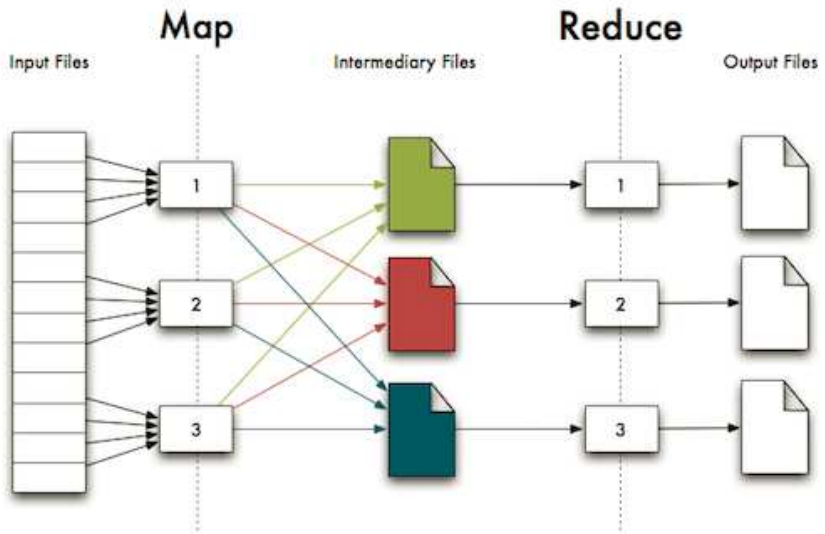
Reduced into a list of values

Three Logical Layers

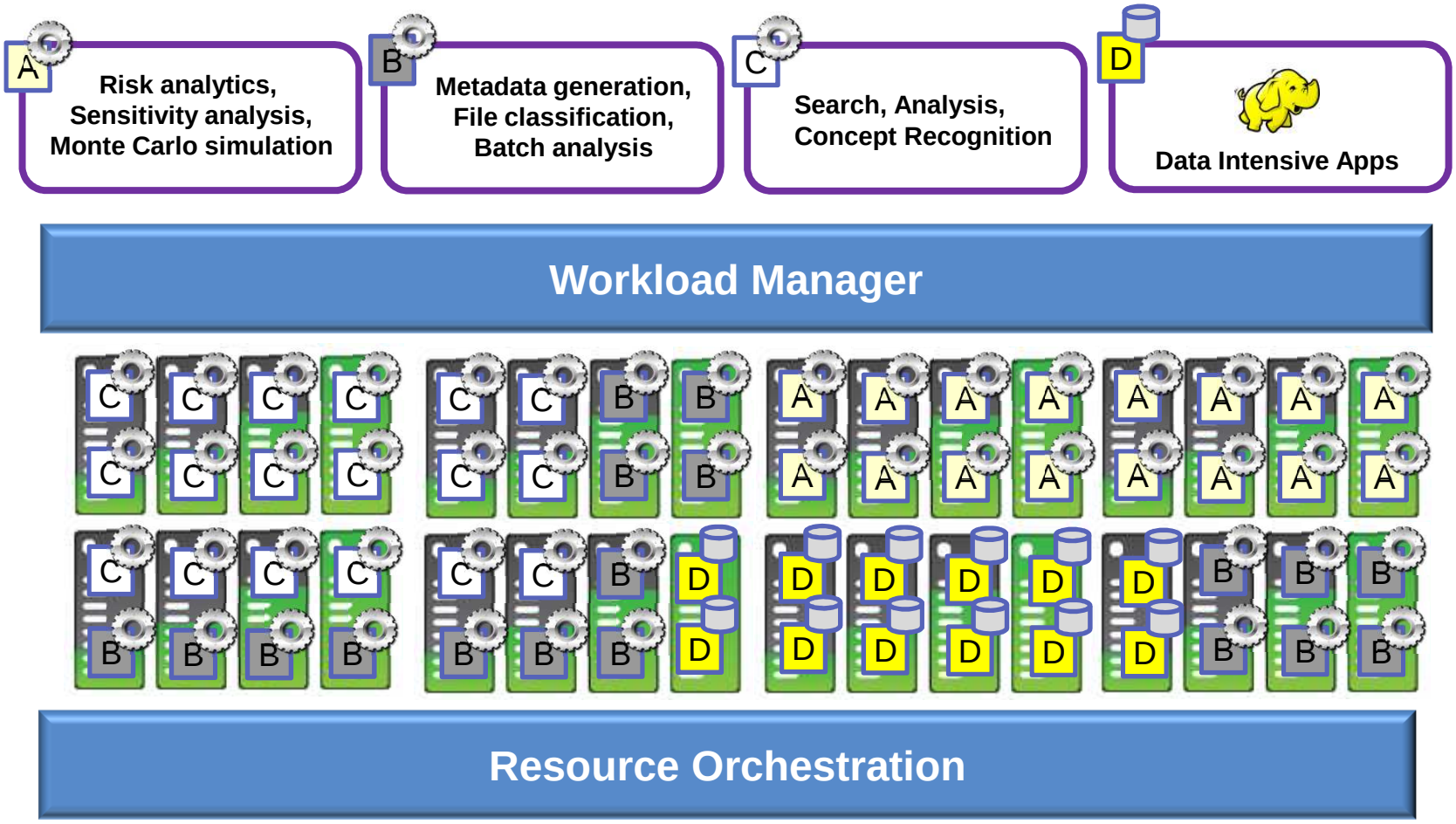
Applications or End User Access

MapReduce Workload Management

Distributed Parallel File Systems / Data Storage



Platform Symphony - Support for Diverse Workloads & Platforms



BigInsights & Platform Symphony

Benefits of an Integrated Solution

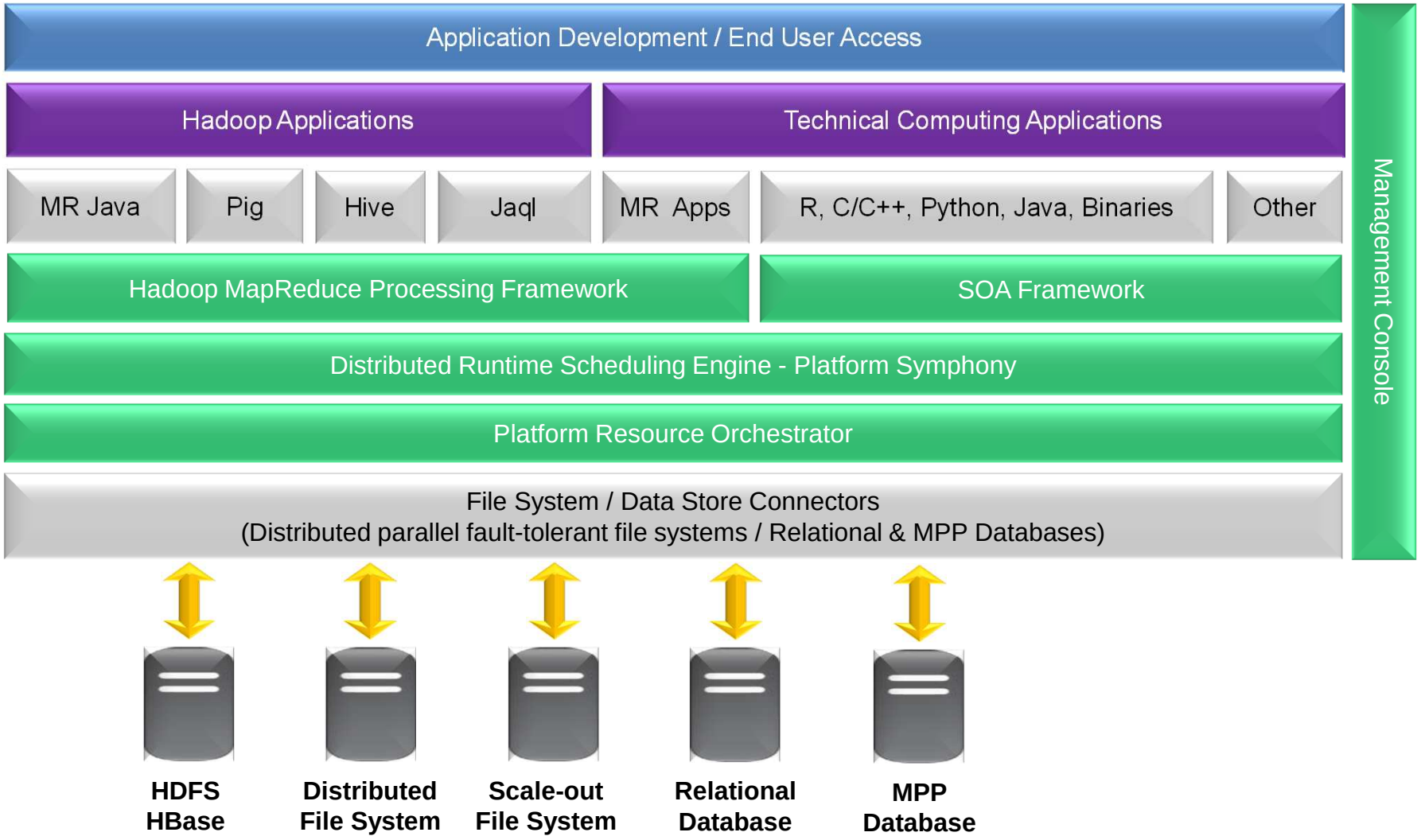
- **Low Latency Requirements**
  - Many simultaneous short running jobs
  - Job cycle measured in seconds or a few minutes
  - Thousands or millions of tasks

  Platform Symphony provides: A Service Oriented (SOA), low latency architecture and a sophisticated scheduling engine.

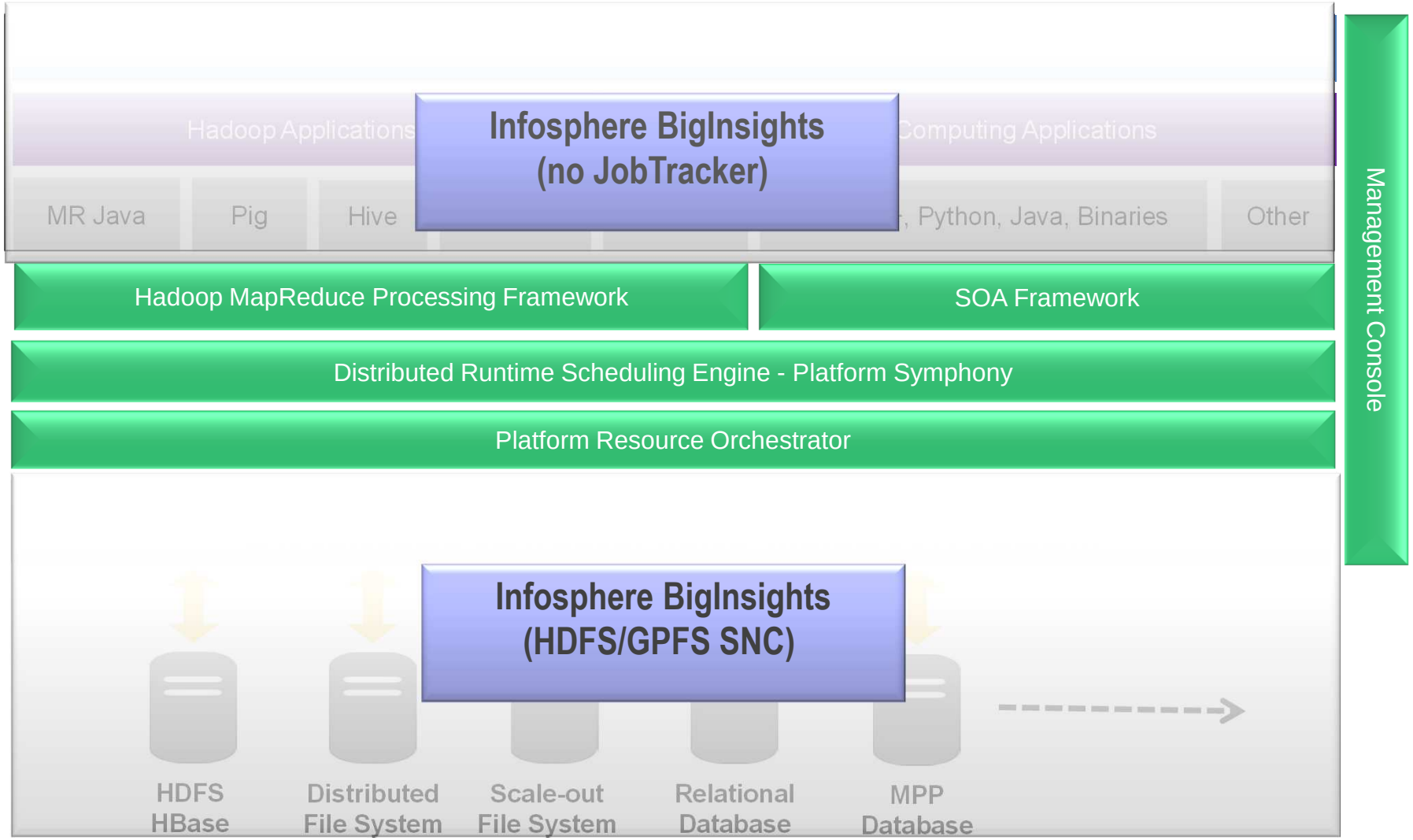
- **Heterogeneous Application Support on a Multi Tenant Grid**
  - Existing Platform Symphony customers extending grid for MapReduce
  - Combined compute & data intensive applications on a single grid
  - Support for multiple applications & job types
    - C#, C++, MapReduce, .NET, Python, etc.
    - SOA & MPI workloads
  - Lines of business sharing a common grid infrastructure

  Platform Symphony provides: Dynamic resource orchestration & sharing across application boundaries.

Application & Data Integration Architecture

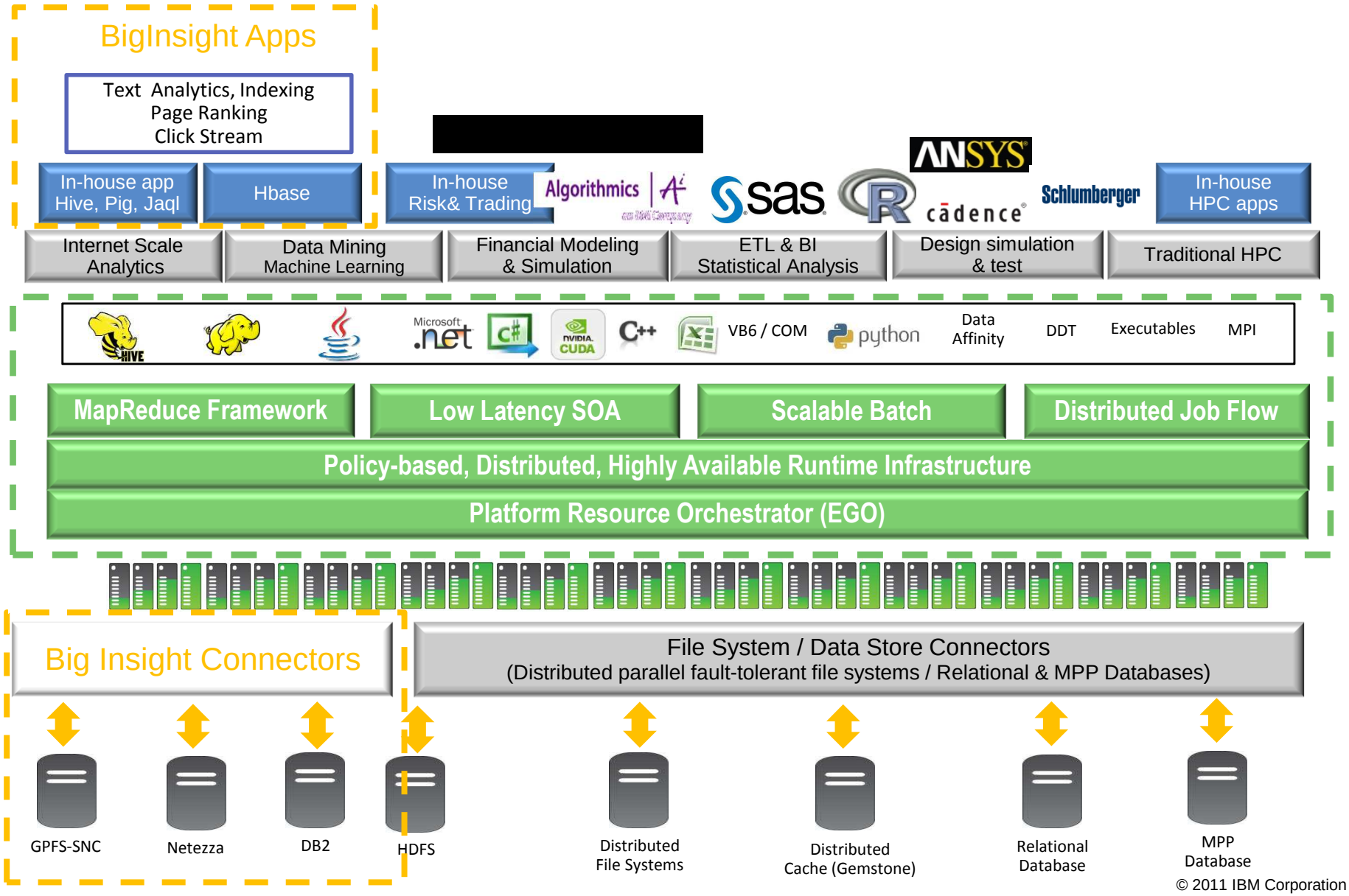


Application & Data Integration Architecture



Heterogeneous Applications Support

Platform Technology – (Green Boxes)



What makes Symphony unique?

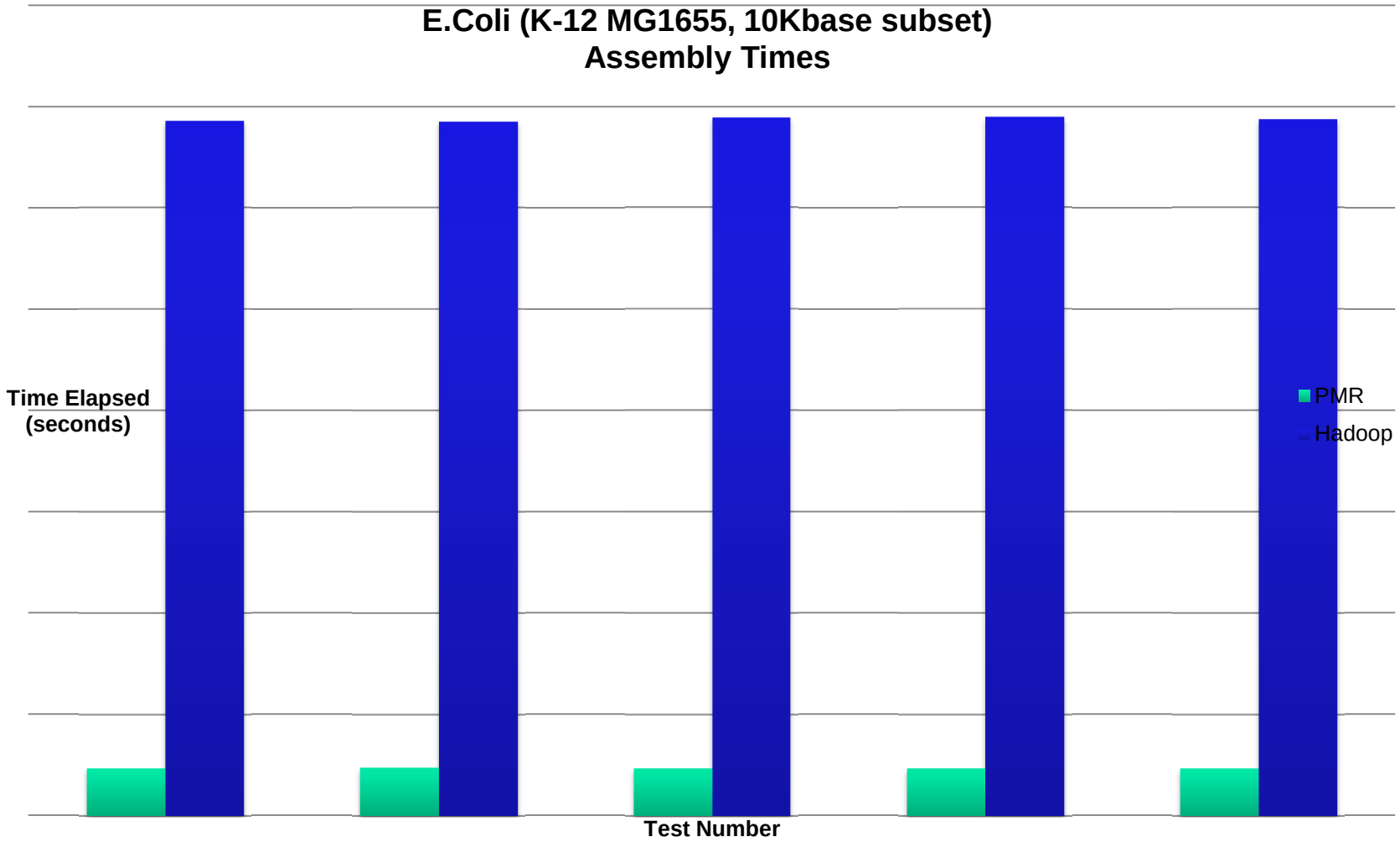
- **Higher quality results faster**
  - Starts & runs jobs the fastest
  - Scales the highest
- **Lower cost**
  - Uses infrastructure more efficiently
  - Easier to manage
  - Simplifies application integration
- **Better resource sharing**
  - Integrated compute + data services
  - Sophisticated hierarchical sharing model
  - Harvesting & multi-site sharing options
- **Smarter data handling**
  - Full MapReduce & HDFS implementation
  - Considers data locality when scheduling tasks
  - Adapts to multiple data sources
- **World-wide support & services**
  - Consulting, Customer Education, Comprehensive support services

Sophisticated Scheduling Engine

- **Fair Share Proportional Scheduling**
  - 10,000 Level of Prioritization
- **Priority Based Scheduling**
  - Higher priority consumes all resources
- **Pre-emptive Scheduling**
  - Interruptive or non-interruptive
- **Threshold Based Scheduling**
  - Resources dynamically monitored
  - Dynamic Open/Close Logic
  - Administrator sets limits
- **Task Reclaim Logic**
  - Automatic when resources fail or ‘hang’
- **Resource Draining**
  - Maintenance mode
- **Administrative Control of Running Jobs**
  - Suspend, Resume, Change Priority, Kill Jobs/Tasks, Monitor

Performance Comparison

Platform Symphony MapReduce versus Hadoop



FSS Use cases

- **Customer Correspondence Analytics Approaches**
  - Call Center Call Data and Agent Notes Analytics
  - Background Quality of Service Screening / Risk Mitigation
  - Cross-channel Behavior Correlation
  - Call Center Conversation “Health” Monitoring
- **Risk Platform and Analytics**
  - Trade Manipulation Monitoring
  - Faster and Expanded Trade History Analytics
  - Corporate Risk / Exposure Analytics
  - Debit and credit card fraud detection
  - Increased Automation of account opening and “know your customer” due diligence
- **IT enablement**
  - Holistic SOA Environment Analytics
  - Application Security and Action Auditing on DB2 on Z
  - Application & Server Log Clearinghouse
  - Low Latency Combined DB2 on Z and Distributed System Data
  - Cyber-Security Packet Investigation
- **Social Media Analytics**
  - Social Media Analytics Focused on High Value Retail Banking:
  - Social Media Monetization *
  - Social Media Analytics Focused on Private Banking:
  - Advisor Social Media Monitoring

A large US Investment Banking

- **Challenge:**
  - Expanding compute capacity while reducing costs
- **Needs**
  - Low latency (< 1 ms) pre-trade application
  - Homegrown overnight batch risk application
  - Trading anomalies detection
- **Solution:**
  - Built a grid infrastructure combining compute and data intensive risk systems