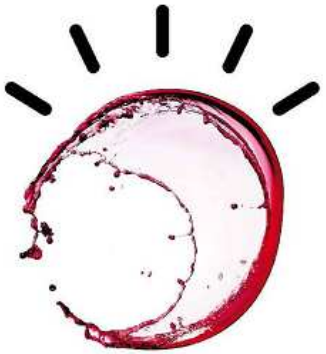
Degustare la crescita

Un percorso in 4 tappe alla scoperta delle soluzioni IBM

Customer Analytics:
la prospettiva dell'infrastruttura



Il paradigma delle "V"

Extracting insight from massive data collections. beyond what was previously possible.

Volume

Variety

Velocity

Veracity

Principi chiave dell'infrastruttura IT per Smarter Analytics

Align

Deploy an information and big data strategy that flows from your business architecture.

Anticipate

Leveraging and Integrating Business Analytics to deliver actionable insights

Act

Embed analytics into your processes and empower a culture of data-driven decision making

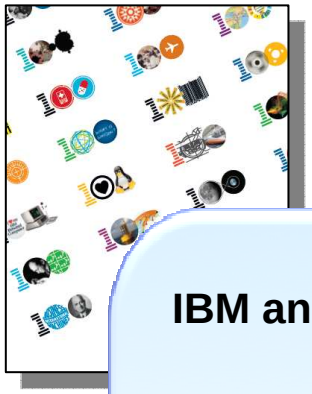
~

Creating a **scalable,** trusted information and systems foundation that improves IT economics and optimizes analytic workload performance using all available data and information.

Optimizing high performance parallel technologies to support complex decision making, spotting trends and anomalies, predicting business outcomes.

Deploying analytics throughout the organization, it's customers and suppliers using **resilient** architectures either on premise or in the cloud.

IBM Systems e Smarter Analytics



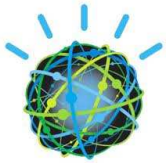
IBM and Analytics at a glance:

- More than **$16B** in Acquisitions of **25** companies **since 2005**
- More than **9.000** Technology Experts & Consultants worldwide
- More than **20.000** Client Engagements
- **Largest Math Department** in Private Industry
- **Nearly 600** Analytics Patents per year
- More than **27.000** Business Partner Certifications

IBM Systems is the market leader in support of analytical workloads

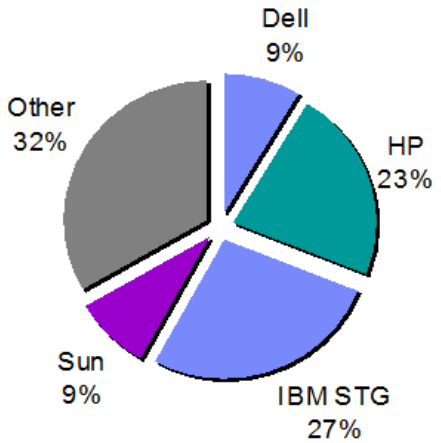
- **IBM Systems**
  - **Smarter Analytics HW/SW Optimizations**

- **Smarter Analytics Appliances**



- **IBM Watson**

IDC Server Workload study 2011



Workloads:
Data Analysis/Data Mining
Data Warehousing/Data Mart
Scientific/Engineering/Industrial R&D

Semplicità - ottimizzazione - flessibilità

IBM PureData System for Analytics

powered by Netezza

Easily load petabytes of data and run complex analytics and reports with minimal administration

IBM DB2 Analytics Accelerator for z/OS

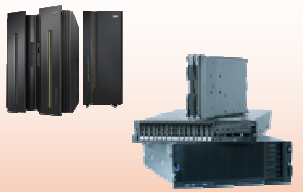
powered by Netezza

Boost DB2 for z/OS query performance maintaining transparency to applications

IBM PureData System for Operational Analytics

Run complex analytics while handling operational reads and writes for real-time decision making

IBM Systems for custom solutions



Scalable, resilient infrastructure that improves IT economics and optimizes analytic workload performance

IBM Infrastructure for analytics foundation

Simplicity ← ***Optimization*** → **Flexibility**

Align – Anticipate – Act

IBM Power Systems per SPSS & Cognos
Ottimizzati per le massime performances

Cognos BI optimized for maximum performance on POWER7

- **40% better** performance with Cognos Business Intelligence V10.1.1 on POWER7/AIX 7.1, over Windows 2008 on x86

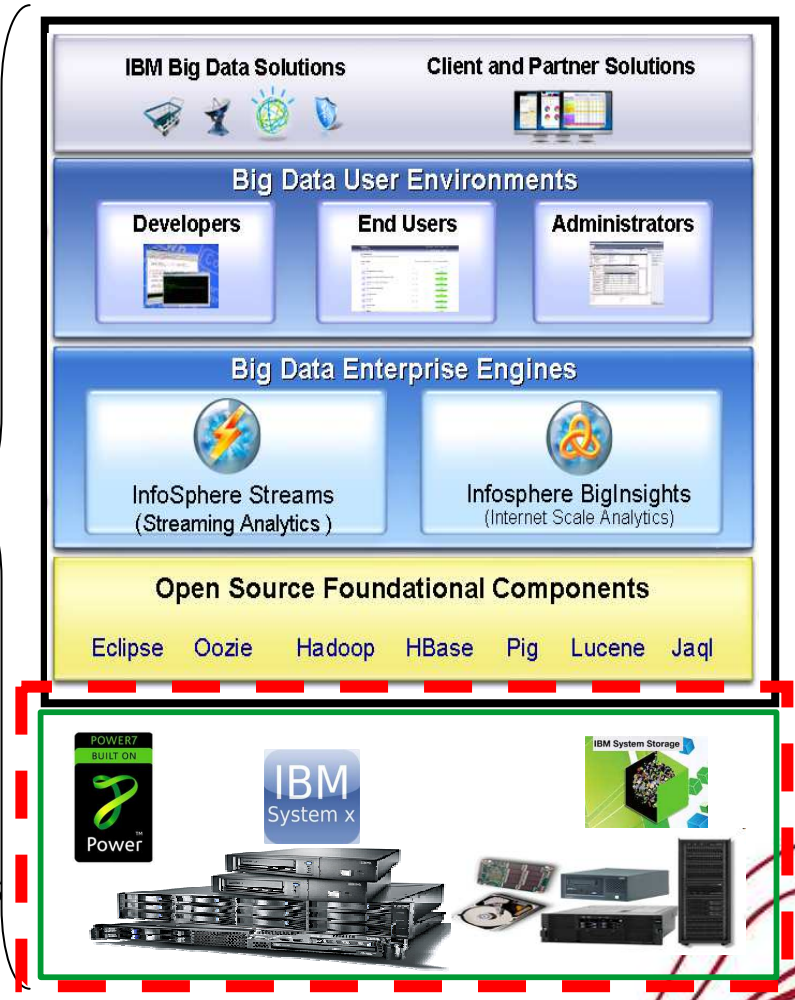
SPSS optimized for maximum performance on POWER7

- **22% better** performance for real-time scoring with SPSS Collaboration and Deployment Services V4.2 on POWER7/AIX 7.1, over Windows 2008 on x86
- **38 times better** performance for real-time scoring with IBM SPSS Collaboration and Deployment Services V4.2 optimized for POWER7 over default POWER7 environment configuration settings.

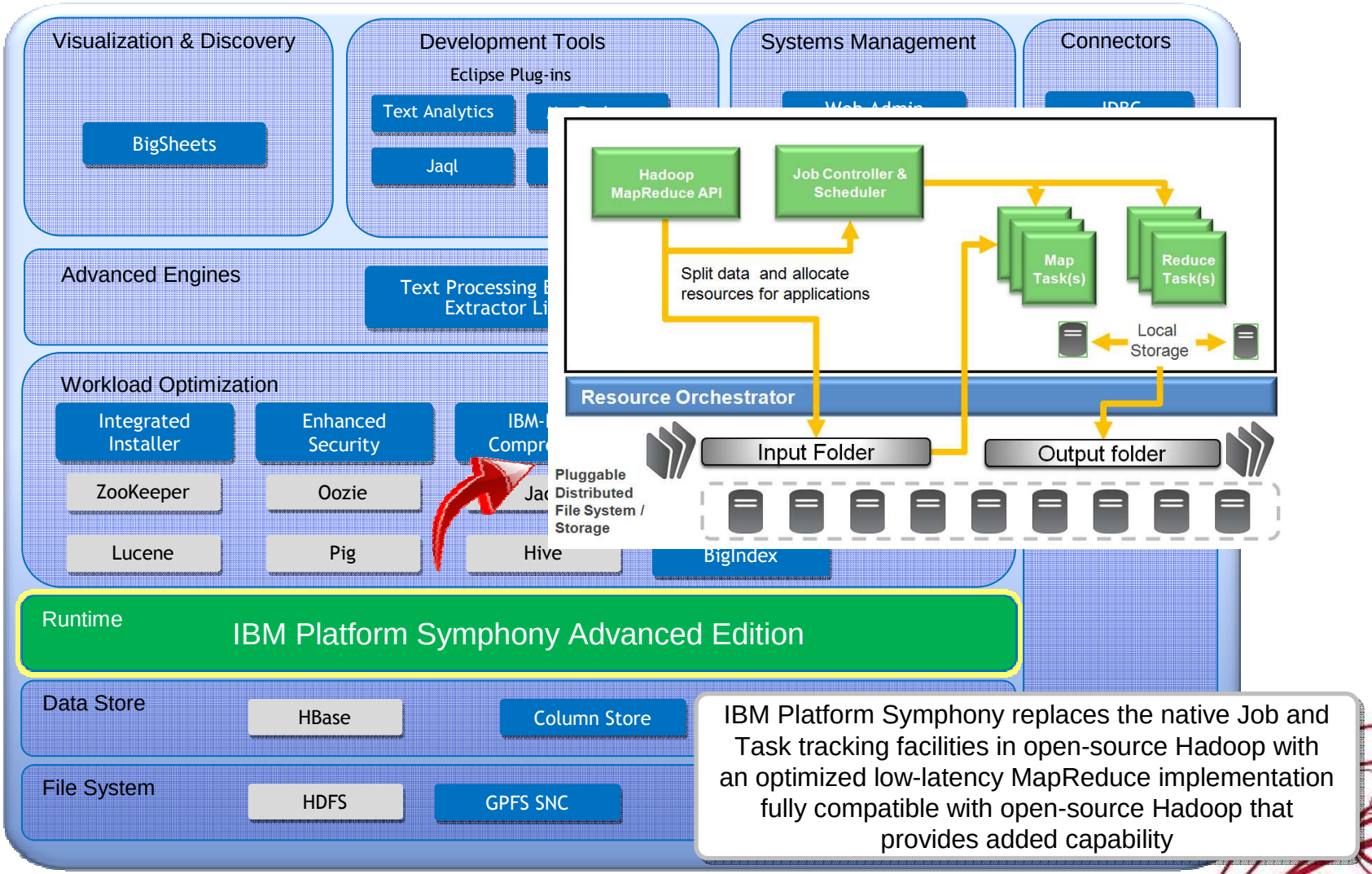
https://www.ibm.com/services/forms/signup.do?source=stg-web&S_PKG=us-en-po-wp-cognosbi&S_CMP=web-ibm-po-ws-analytics

IBM Systems per Big Data analytics

- **A complete infrastructure support for Big Data**
  - Other vendors require multi-vendor solutions
- **Enterprise-class focus**
  - Performance tested
  - Administrative and development tooling
  - Deep integration with information management software inside and outside IBM
  - Security and governance
  - High availability and backup
- **System X, POWER Systems**
  - Industry leading innovation and technology
  - Best in class reliability and availability
  - #1 in customer satisfaction
- **IBM Services, Consulting and Research**
  - Deep expertise in Hadoop and other applications of Big Data
  - **Platform Symphony**: extending the capabilities of IBM Infosphere BigInsights



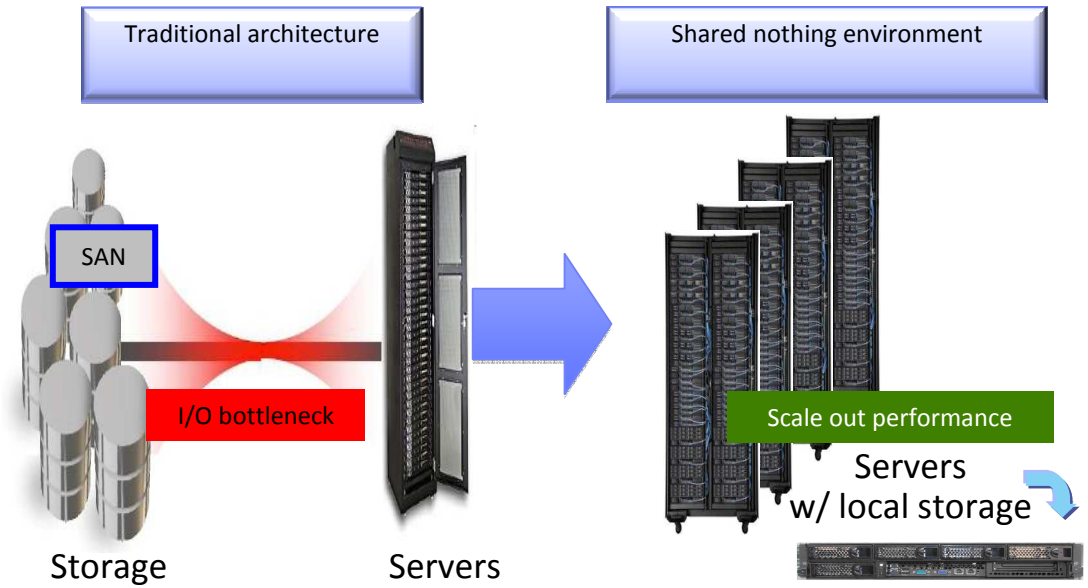
BigInsights Enterprise Edition & Platform Symphony

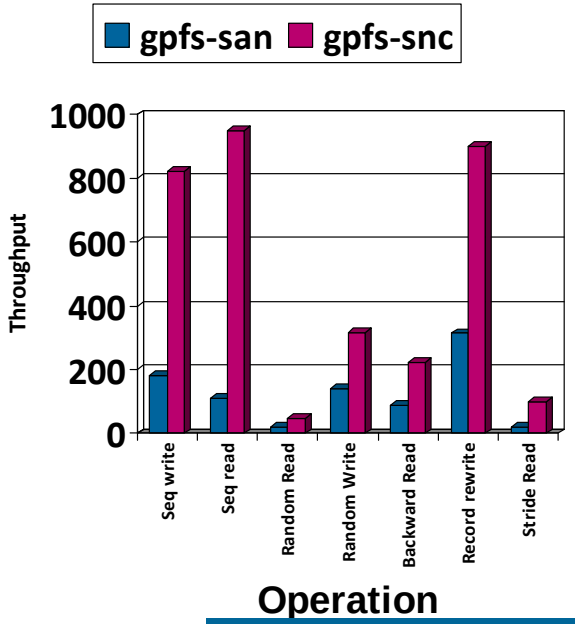


IBM Platform Symphony replaces the native Job and Task tracking facilities in open-source Hadoop with an optimized low-latency MapReduce implementation fully compatible with open-source Hadoop that provides added capability

IBM components Open source (IBM) Symphony

Shared Nothing Clusters: Foundation for Data Intensive Computing





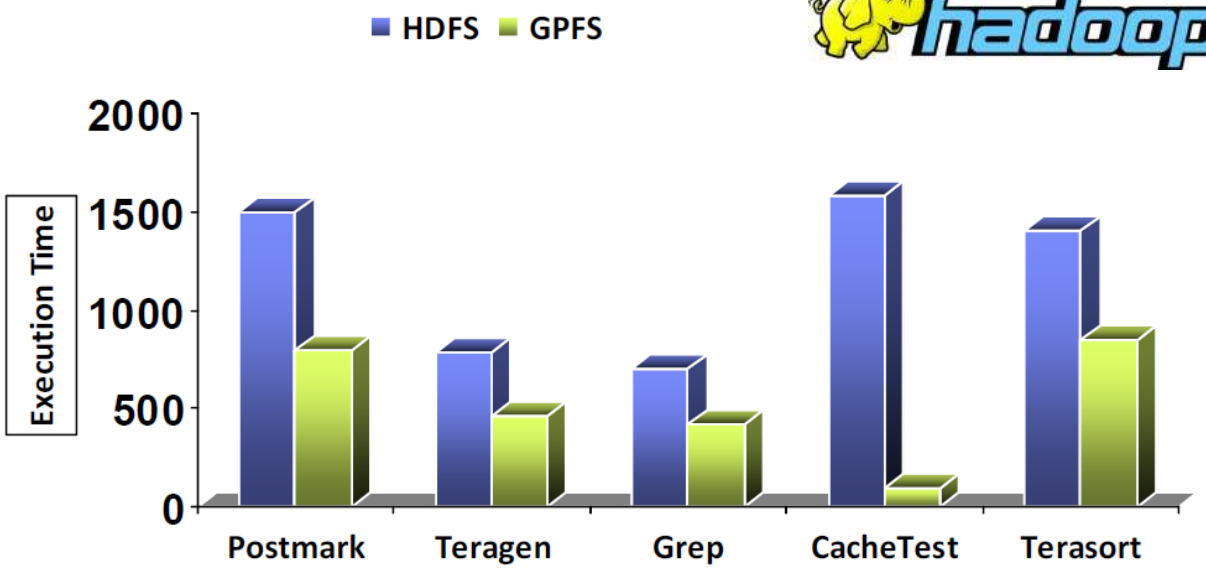
Hardware: 8-node cluster
SNC: locally attached disks
SAN: DS4800 with equiv disks
Reason for difference:
DS4800 outbound b/w

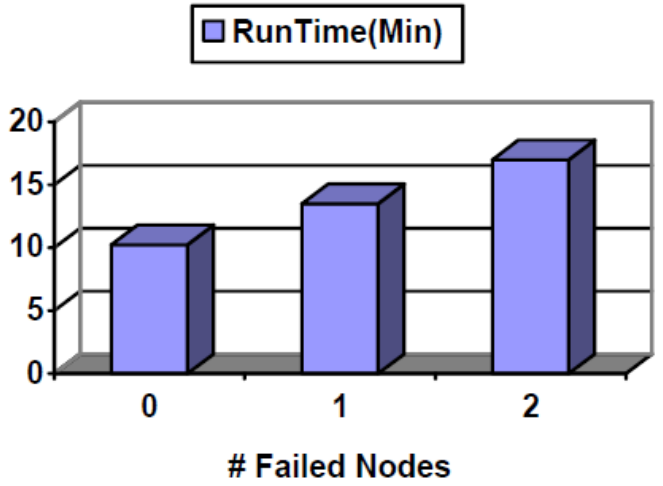
Needed:
A programming model and an enterprise file system to deliver scale out performance to data-intensive applications

Challenges:
Relatively lower network b/w → compute to data
Failures are common → replication

GPFS vs HDFS MapReduce Benchmark

- ✓ Has **better striping of data across disks** so that MapReduce threads read and write from disks in parallel in GPFS instead of large block allocations on a single disk in HDFS.
- ✓ **Efficient distributed metadata** for random block access in GPFS instead of a single metadata node in HDFS.
- ✓ **Client side caching** for locality in MapReduce applications in GPFS instead of no caching in HDFS.





Platform Symphony – Berkeley SWIM

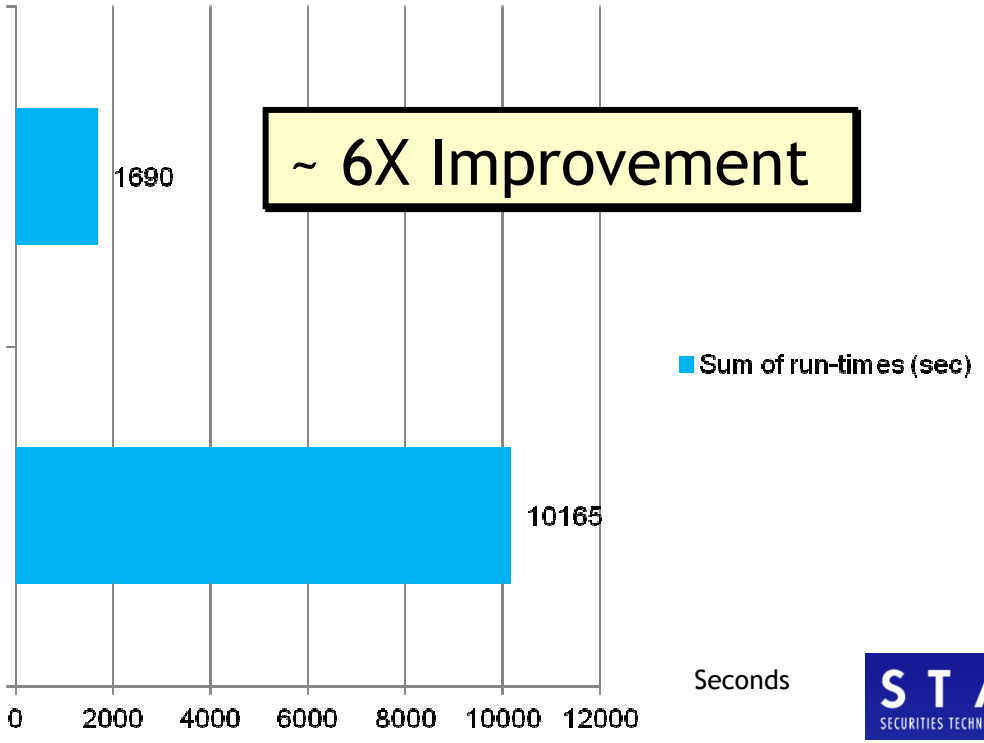
Facebook 2010 workload – sum of total job run-times in second. Workload represents first 20 minutes of Facebook workload, comprised of 302 jobs.

Hadoop 1.0.1 + Symphony

First 302 jobs of SWIM Facebook 2010 workload on 19 node 204 core cluster running IBM Platform Symphony 5.2

Hadoop 1.0.1 with native scheduling

First 302 jobs of SWIM Facebook 2010 workload on 19 node 204 core cluster using standard Hadoop scheduler

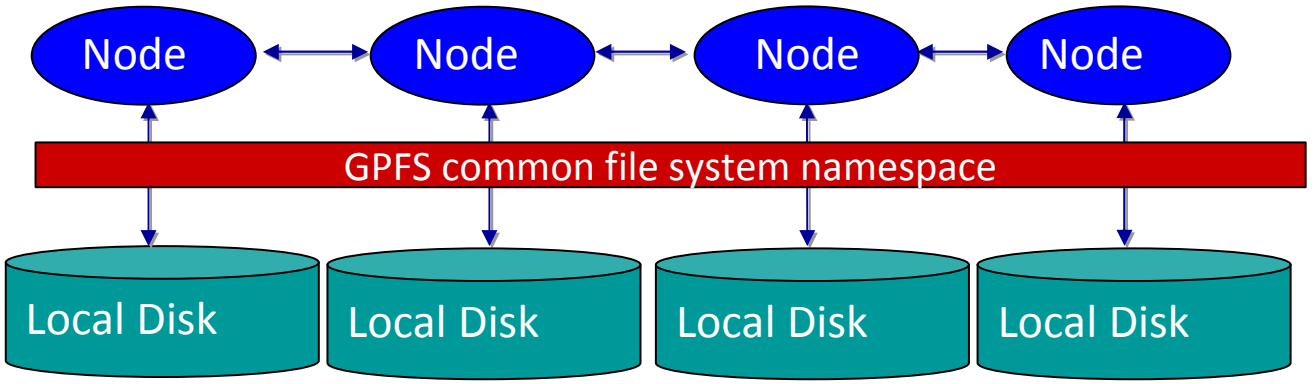


This audited result demonstrates clearly the dramatic impact that Symphony has on a “real” Hadoop workload. In both cases the cluster configuration was identical – 19 nodes, 17 compute hosts with slots per host. (204 cores) – Identical cluster, Hadoop and HDFS configurations.


ThankYou
Thanks
merci
Dankie
Grazie
saha
takk
dank
terima
hvala
rahmat
gracias
salamat
murakoze
mauruuru
kasih
mèsi
kié
wel
tak
dhanyavadalu
ačiū
Diakuiu
děkuji
çox
lava
fyri
akun
gratias
Gråces
aabhar
Barak
misaotra
ThankYouou
trugarez
dieuf
Tau
tanemirt
agimus
aguyjé
gràcies
spas
köszönöm
fiik
maith
wel
dank
díky
mulţumesc
dziękuję
Vinaka
danke
khob
paldies
ederim
akpé
kiitos
niżżik
Bayarlalaa
obrigado
toda
arigató
meharbani
sobodi
diolch
dankì
birepo
go
salamat
trugaré
asante
waita
marahaba
chokrane
Yekeniele
najis
blagodaram
dakujem
faleminderit
tesekkur
tashakor
mercé
nanni
asko
tuke
sulpáy
Nandri
po
tänan
sag
ajr
chnorakaloutioun
manana
cám
gratzias
kam
aabhari
dankon
zikomo
bedankt
agat
faafetai
leibh
grassie
raibh
chai
je
n
djere enkosi
stuutiyi tapadh
dhanyavad gha-ana
dhanyavadagalu
dankewol
mahalo
Allahu
taiku
ngiyabonga
shukriya
san
dhanyabaad
xièxie
barkal
tack
tanmirt
tak
namnida
motashakkeram
eskerrik asko
olun
gracie
sheun
JCBED
tenki
hamnida

BACKUP SLIDES

GPFS: A Scalable File-system for Shared Nothing Architectures



Cluster: thousands of nodes, fast reliable communication, common admin domain.

Shared disk: all data and metadata on disk accessible from any node, coordinated by distributed lock service.

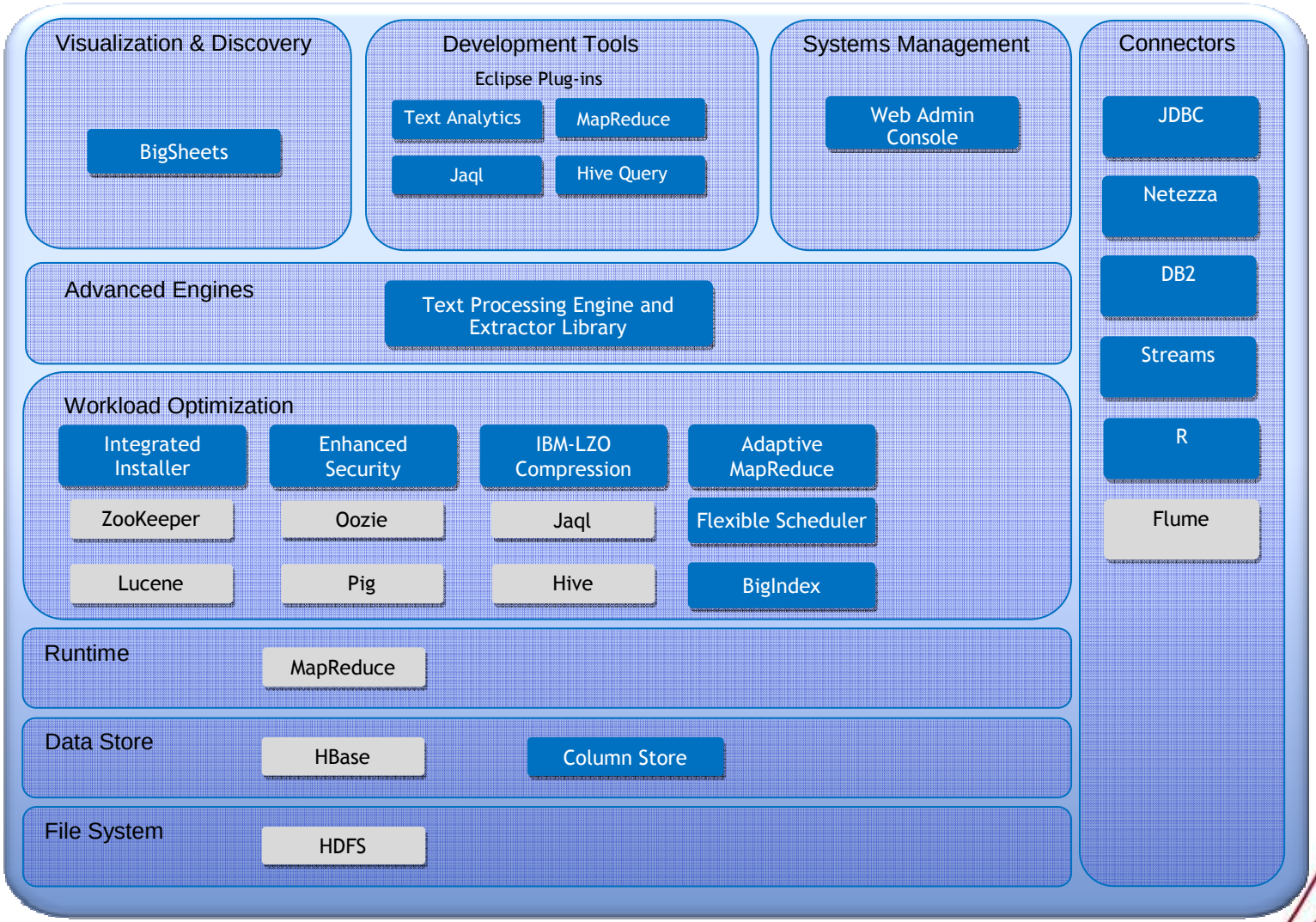
Parallel: data *and metadata* flow to/from all nodes from/to all disks in parallel; files striped across all disks.

Key Ideas:

- Locality
- Write Affinity
- Metablocks
- Pipelined replication
- Distributed recovery

NEW

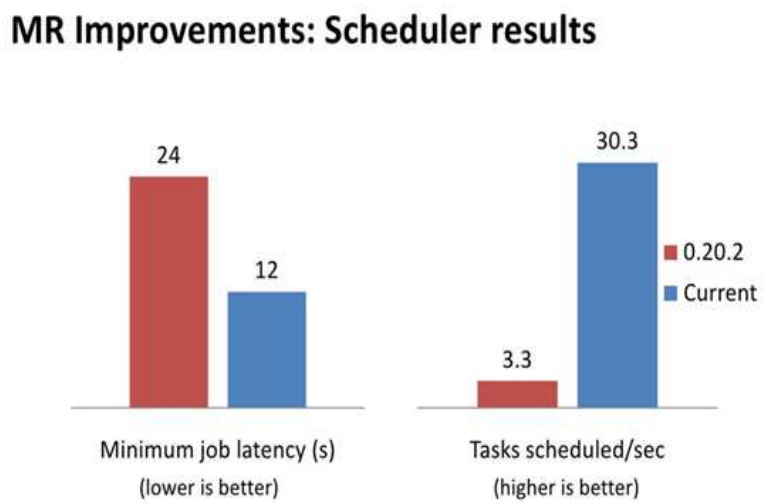
BigInsights Enterprise Edition Components



Hadoop sleep test – scheduling performance

Symphony vs. commercial and open source Hadoop distros

Hadoop - previous "state of the art" - Comparing Hadoop 0.20.2 to Cloudera CDH3 – results below shared by Cloudera at Hadoop World 2011

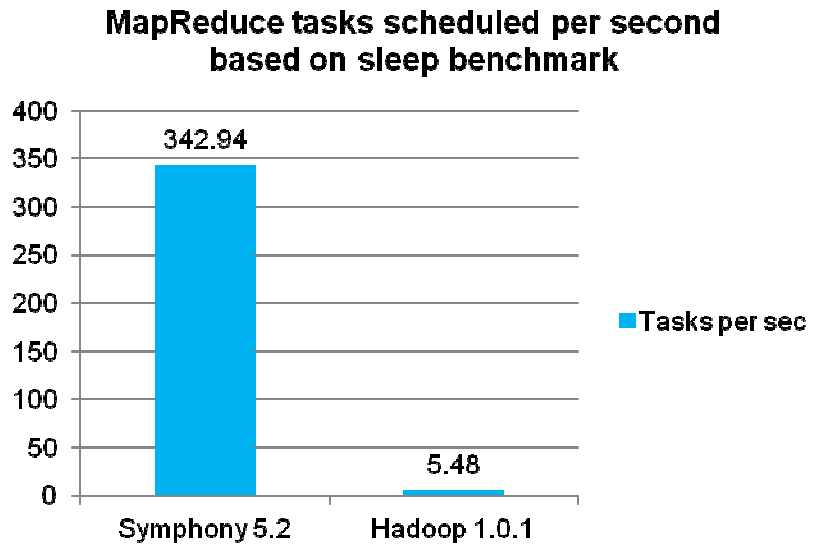


10 node cluster, 10 map slots per machine.
hadoop jar examples.jar sleep –mt 1 –rt 1 –m 5000 –r 1

13

http://www.slideshare.net/cloudera/hadoop-world-2011-hadoop-and-performance-todd-lipcon-yanpei-chen-cloudera

Platform Symphony 5.2 MapReduce scheduling engine performance

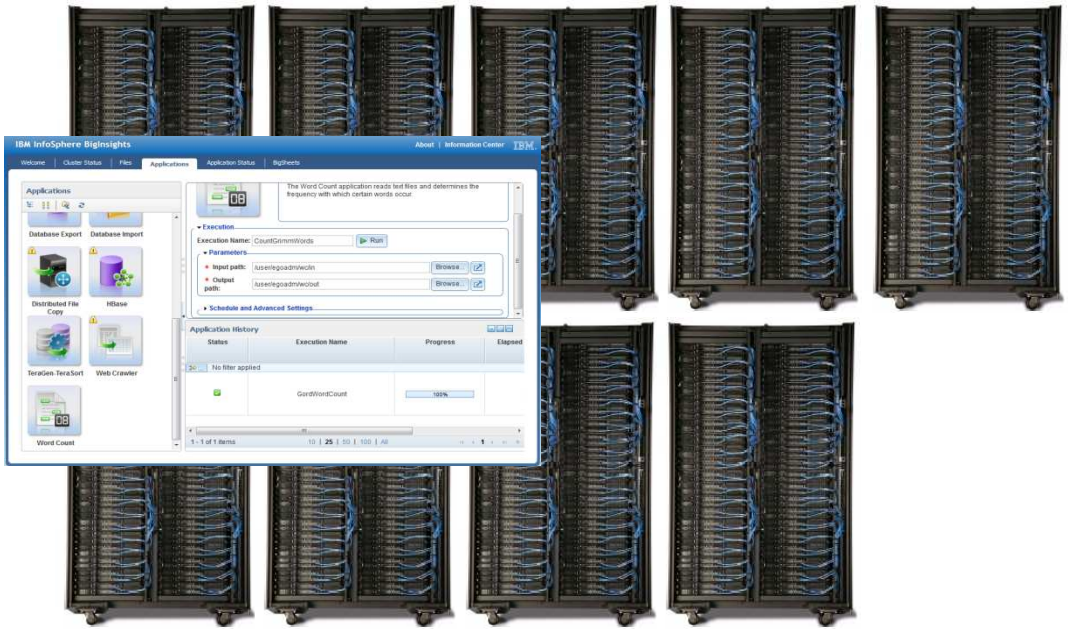


63X Improvement

This was a separate audited test, on a separate cluster, replicating the sleep test done by Cloudera on 204 slots. Note the similarity of the Hadoop results.

Proof point – Recent large scale terasort test

Tests conducted at IBM using BigInsights 1.3.0.1 and Platform Symphony 5.2 in August 2012.



Symphony Compute Hosts

- 250 x dx360 M3 nodes (not all used)
-120 GB memory per host
-12 spindles per host, 3 TB each
-RHEL 6.3 with KVM
-5 VMs configured per physical host
-Each VM with 16GB RAM, 2 vCPUs

Symphony Master Host

- dx360 M3 nodes
- Single large VM
- 100GB RAM, 10 vCPUs (cores)

Software

-IBM InfoSphere BigInsights 1.3.0.1
-IBM Platform Symphony AE 5.2
-BigInsights 1.3.0.1 integration patch

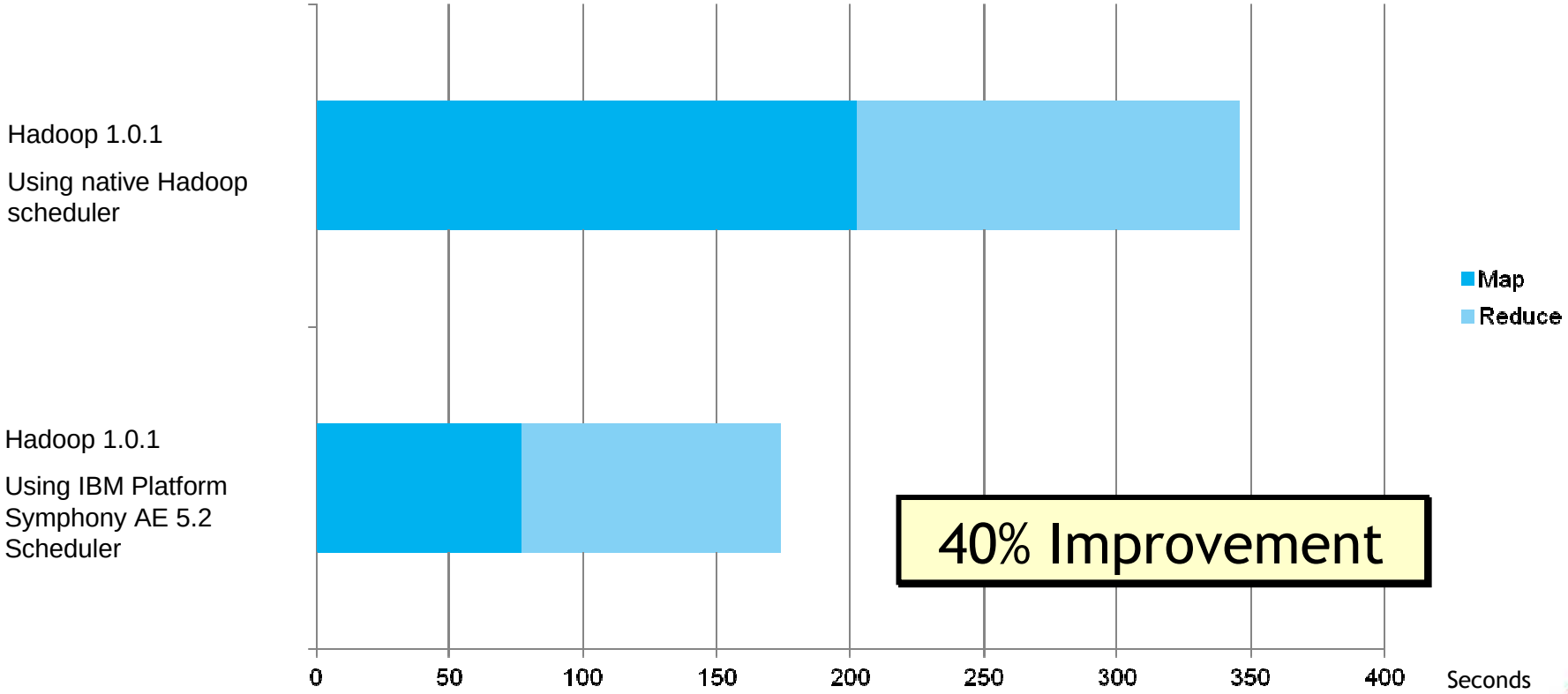
40% Improvement

World Class Result:

100 TB sort completed in 1,000 VMs and 200 nodes in *10,369 seconds. Exceeding a previous world record, but with one-tenth the hardware!*

Platform Symphony benchmark - terasort

Measuring the impact of IBM Platform Symphony on an identical cluster with an identical Terasort workload



In this (unaudited test), the absolute result is not significant – it is the comparative result that is important. Using the identical cluster hardware, software and HDFS environments, Platform Symphony reduced the total terasort run-time by 40%

Comparison of GPFS and HDFS

File System	GPFS	HDFS
Robust	No single point of failure	NameNode vulnerability
Data Integrity	High	Evidence of data loss[1,2]
Scale	Thousands of nodes	Thousands of nodes
POSIX Compliance	Full – supports a wide range of applications	Limited
Data Management	Security, Backup, Snapshot, Caching, Wide-area Replication	Limited
MapReduce Performance	Good	Good
Traditional Application Performance	Good	Poor performance with random reads and writes

[1] Care and Feeding of Hadoop Clusters, Marc Nicosia, Usenix 2009
[2] The Komos Distributed File System, Sriram Rao, Quantcast Inc. (Invited talk)

PureData System for Analytics Hardware Overview

- 8 Disk Enclosures
- 96 1TB SAS Drives (4 hot spares)
- RAID 1 Mirroring

- 2 Hosts (Active-Passive):
- 2 Quad-Core Intel 2.6 GHz CPUs
- 7x146 GB SAS Drives
- Red Hat Linux 5 64-bit

- 14 PureData for Analytics S-Blades™:
- 2 Intel Quad-Core 2+ GHz CPUs
- 4 Dual-Engine 125 MHz FPGAs
- 24 GB DDR2 RAM
- Linux 64-bit Kernel

Scales from
¼ Rack to 10 Racks

32 TB to 1.2 PB of
User Data

- User Data Capacity: 128 TB**
- Data Scan Speed: 145 TB/hr**
- Load Speed (per system): 5+ TB/hr

- Power Requirements: 7.6 kW
- Cooling Requirements: 7.8 kW

**: 4X compression assumed