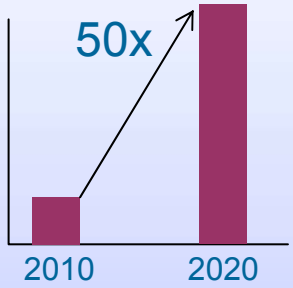
Smarter Computing:
What's Next. Ready Now.

Big Data on POWER Systems

Smarter Computing Briefing 2013

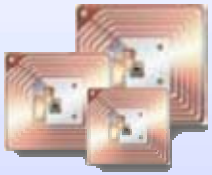
Big Data Characteristics

Cost effectively processing the growing **Volume**



35 ZB

Responding to the increasing **Velocity**



30 Billion RFID sensors and counting

Collectively analyzing the broadening **Variety**

80% of the worlds data is unstructured

Establishing the **Veracity** of big data sources

1 in 3 business leaders don't trust the information they use to make decisions

Where is the Big Data Coming From?

Text Documents

Blogs

Web Logs

Mfg. Equipment

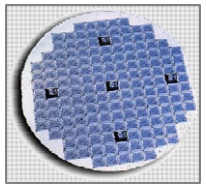
Email

Weather Data

Social Media

Stock Trades

Mfg. Equipment



Utility Meters

Medical Equip.

Call Data Records

Point of Sale Data

Video Cameras

Audio Devices

Oil Rigs

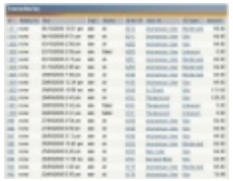
Data at rest

- Data is stored on disk
- Huge volumes of unstructured data
- No pre-defined schemas
- Too large for traditional tools to process in a timely manner

Data in motion

- Data is typically not stored
- Tremendous velocity
- Multiple data sources
- Huge volumes of unstructured data
- Ultra low latency required

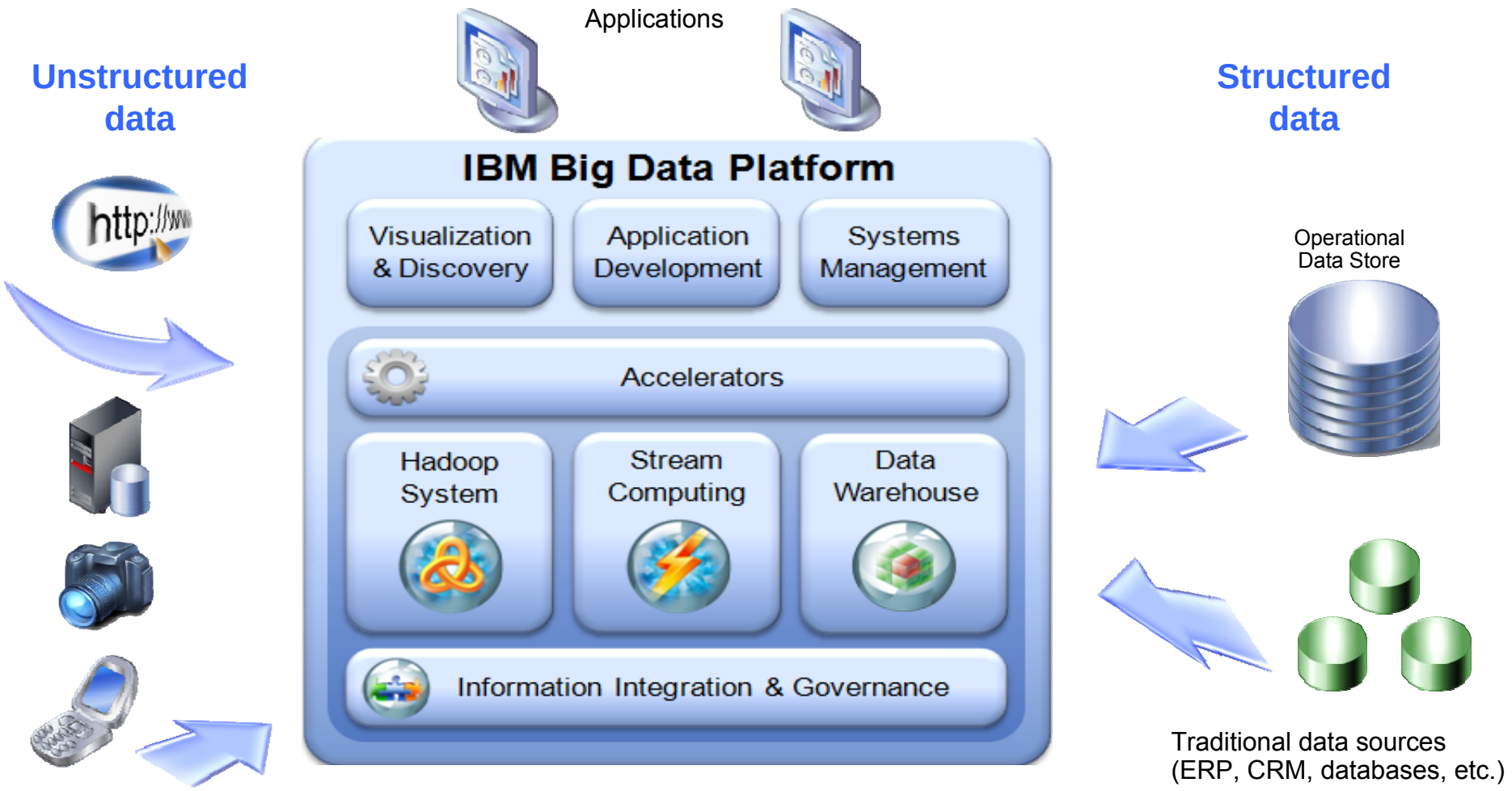
Gaining Value from Data at Rest

Data Source	Analysis	Business Value
Web Logs 	*Analyze online shopper behavior*	*Maximize retail web site sales*
Social Media 	*Analyze customer sentiment and experience*	*Attract and retain customers*
Weather Data 	*Analyze vast amounts of historical weather data*	*Determine optimal wind turbine placement*

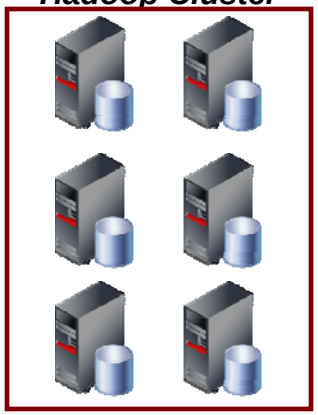
Gaining Value from Data in Motion

Data Source	Analysis	Business Value
Medical Equipment 	*Monitor various medical devices for anomalies*	*Detect life-threatening conditions in time to intervene*
Audio Devices 	*Analyze sound from audio sensors around buildings and plants*	*Detect intruders at vulnerable locations*
Point of Sale Data 	*Combine Point of Sale data with relational data about customers*	*Maximize up-sell opportunities*

Big Data Platform: Gain Value From Unstructured Data Sources And Structured Enterprise Data

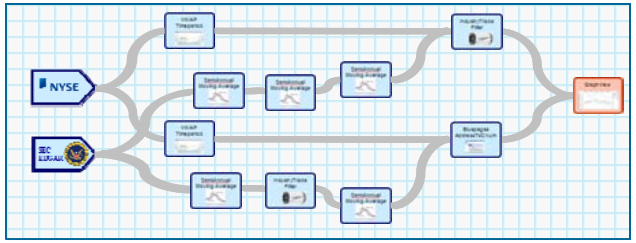


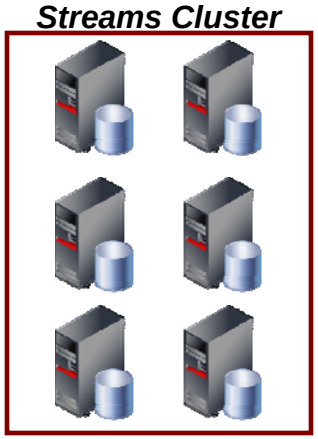
New Programming Models and Low Cost Hardware For Handling Unstructured Data



- Apache Hadoop and InfoSphere Streams
  - Proven frameworks to process large amounts of data
  - Hadoop for data at rest, Streams for data in motion
  - Enable applications to transparently work with large clusters of nodes in parallel

InfoSphere Streams





Service Oriented Finance Wants To Grow Their Business

We need to attract more customers…and retain the ones that we have.

Service Oriented Finance Marketing VP

You can easily find out what your competitors are doing right to attract and keep customers…

And what they are doing wrong to lose customers.

IBM

Sentiment Analysis - A Big Data Challenge But Also A Big Data Opportunity

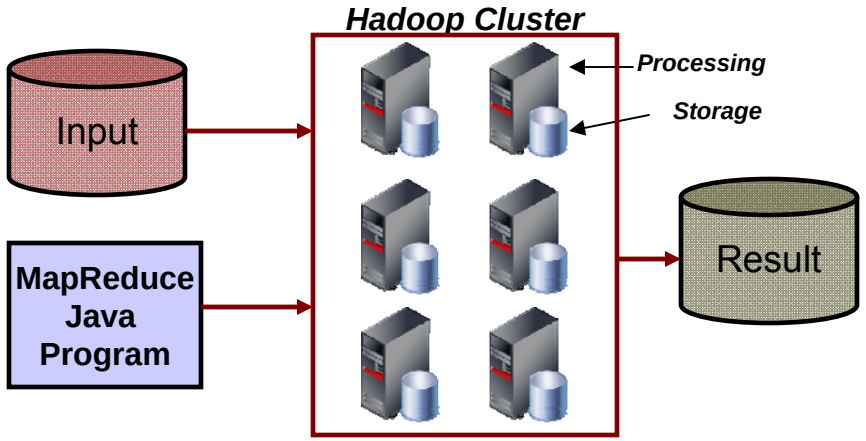
Huge volumes of unstructured data

Trying to determine…

- Product demand
- New product acceptance
- Competitive threats
- Threats to brand reputations
- Advertisement targets

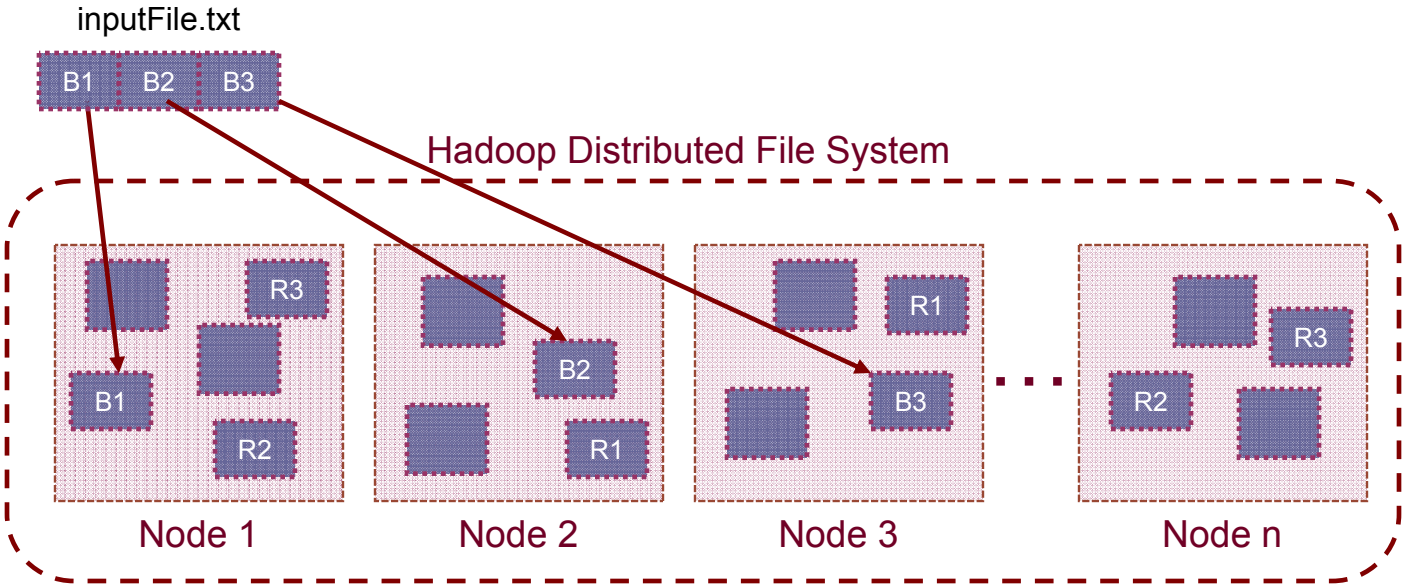
Finding sentiment from social media site data

Apache Hadoop



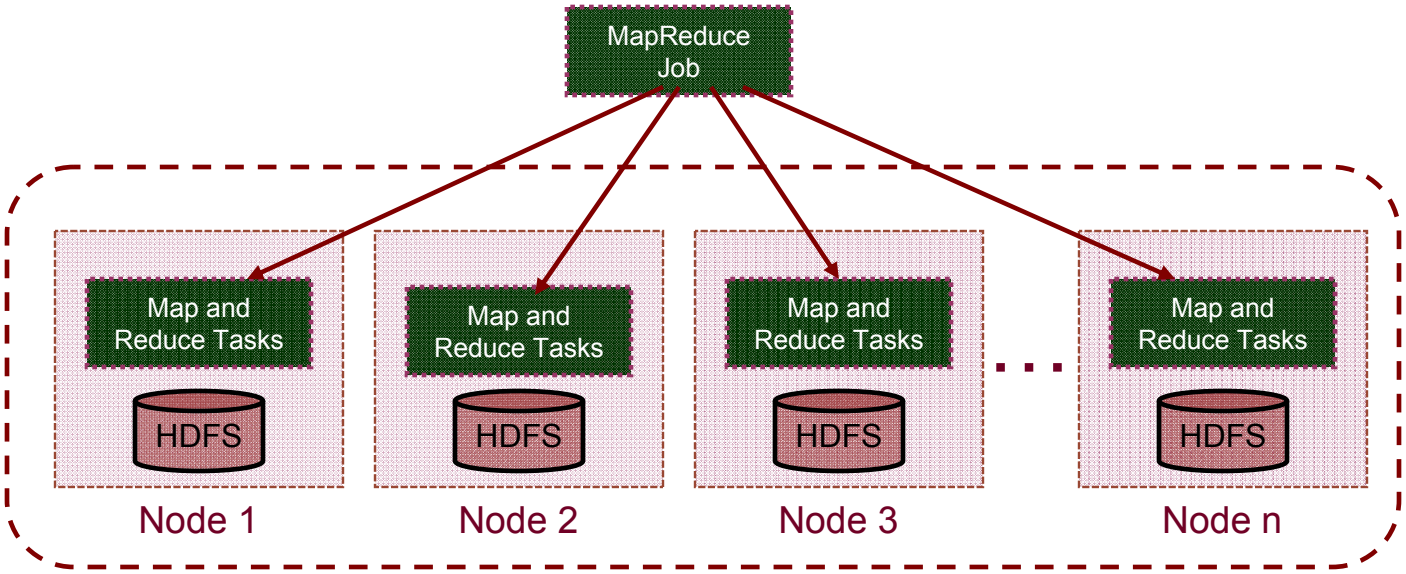
- Comprised of a cluster of inexpensive hardware
  - Nodes have processors, memory and disks
- Special file system – Hadoop Distributed File System (HDFS)
- Special programming model - MapReduce

Hadoop Distributed File System (HDFS)



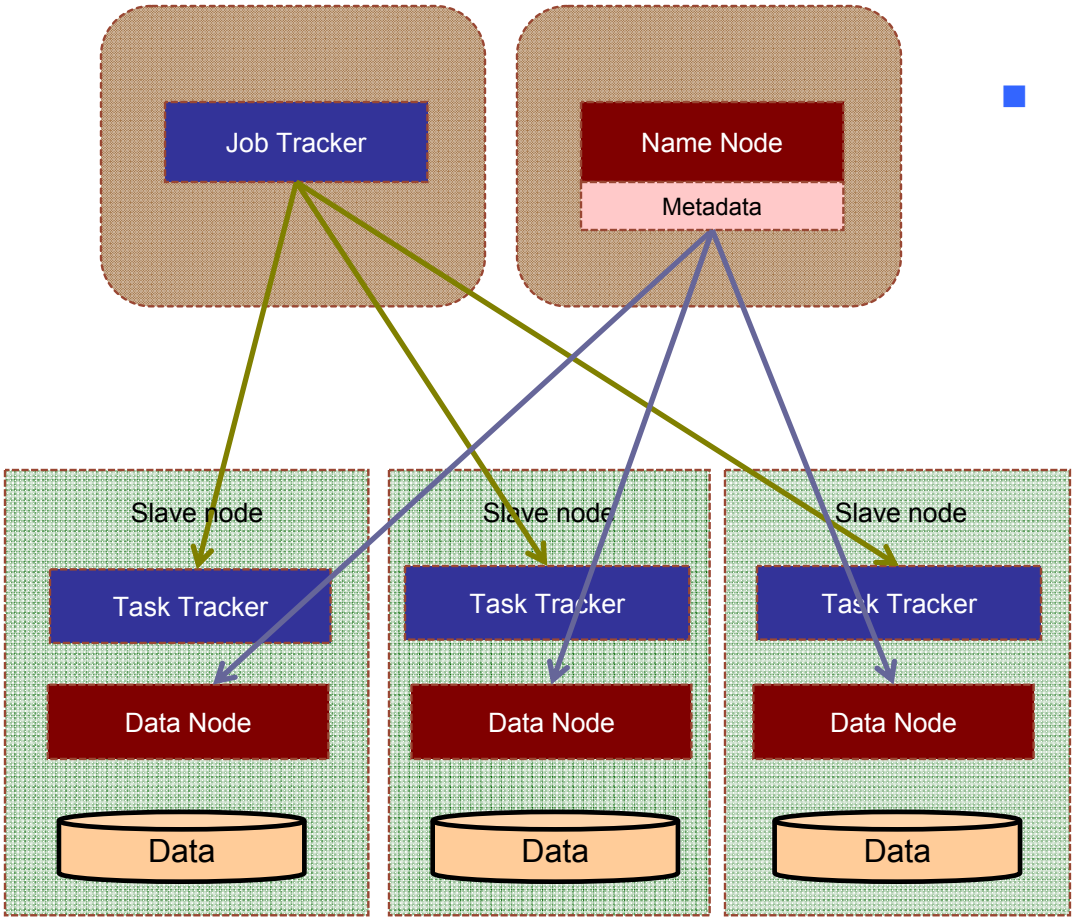
- A distributed file system that spans all the nodes in a Hadoop cluster
- Files are split automatically at load time into blocks and spread among Data Nodes
- Name Node is consulted for placement of blocks, then data is written directly to Data Nodes.
- Assumes nodes will fail - achieves reliability by replicating data across multiple nodes
- Elastically scalable

MapReduce Framework



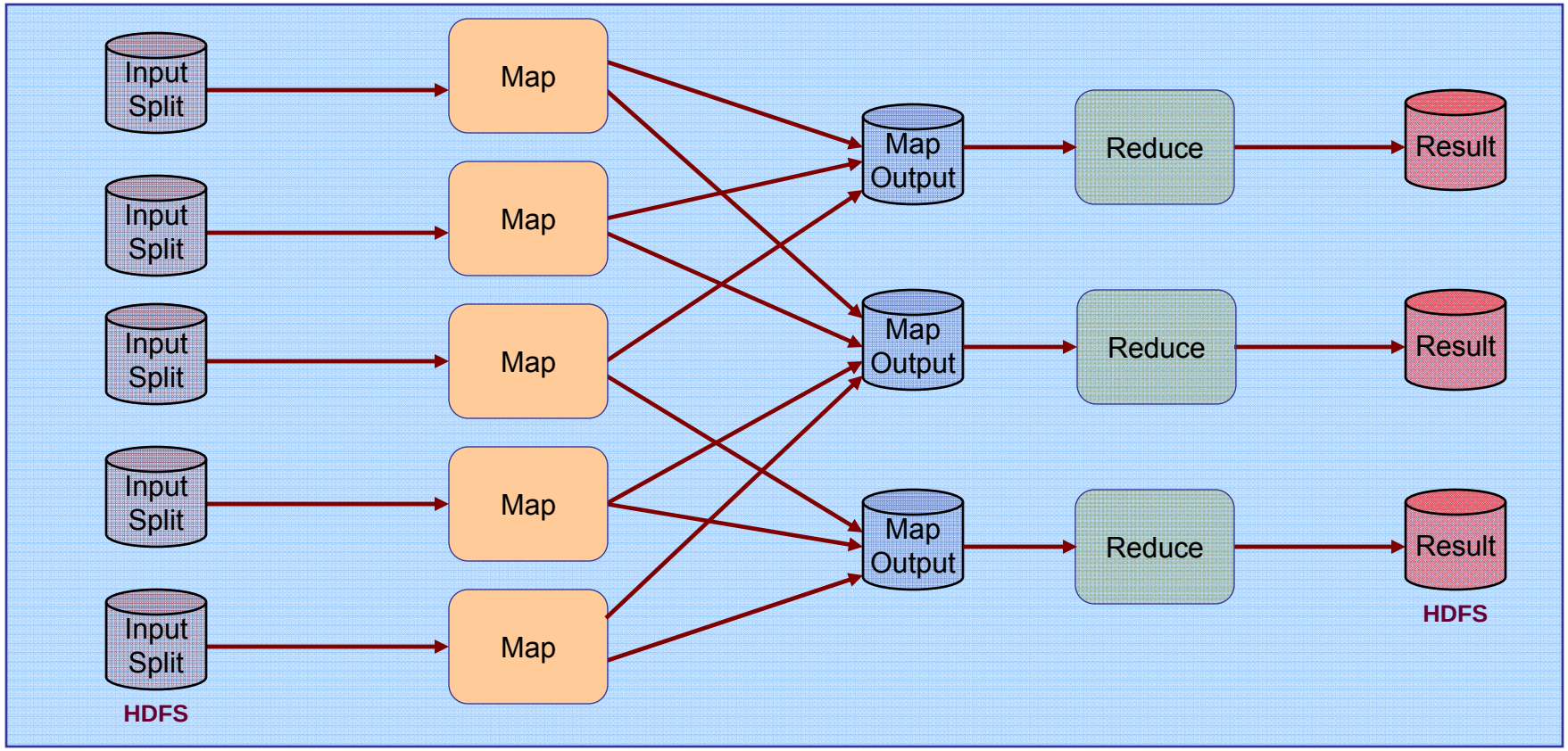
- MapReduce job is sent out to each node
- Map and Reduce steps/tasks are code that a programmer implements
- Map and Reduce tasks run in parallel across nodes
- The steps process key/value pairs in some way
- How the steps manipulate the pairs defines the solution
- Hadoop framework does a lot of the “heavy lifting”
  - e.g., moving data between map and reduce tasks

Hadoop Framework Processes



- Master processes
  - Job Tracker for MapReduce
  - Name Node for HDFS
  - Metadata is file placement information
- Slave processes
  - Task Trackers for MapReduce
  - Data Nodes for HDFS

Hadoop Framework



MapReduce Job Is Executed

1. Framework invokes Map steps with one row of data from split
2. Map steps execute in parallel
3. Map steps write out key/value pairs

Framework Shuffling Process

1. Hash code created to determine which reducer a key/value is sent to
2. When all keys arrive they are sorted
3. Keys are grouped and given to a reducer

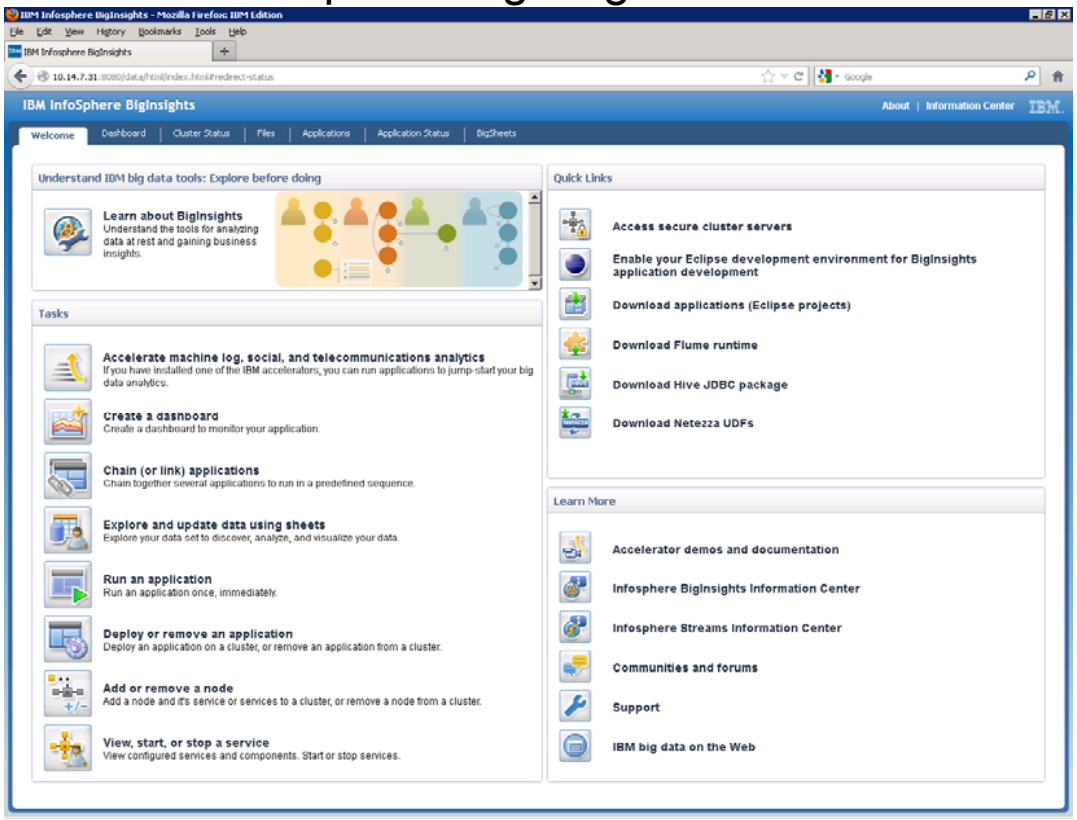
Reduce Process

1. Reduce steps are invoked with one key and all values for that key
2. Reduce steps write out final key/value pairs

BigInsights Makes It Easy

- Complete management of cluster
  - Monitor/start/stop components
  - Add/remove nodes
- Portal style dashboards
- Read/write access to HDFS
- Extensive views of jobs and workflows in system
- Application staging, launch and scheduling center
- Many built in accelerators
- Business user tools
  - No Java required
  - Spreadsheet style
  - Visualization

InfoSphere BigInsights Console



Demo: Using BigInsights To Determine What Customers Like/Dislike About A Competitor

Big Bank Customers: likes and dislikes

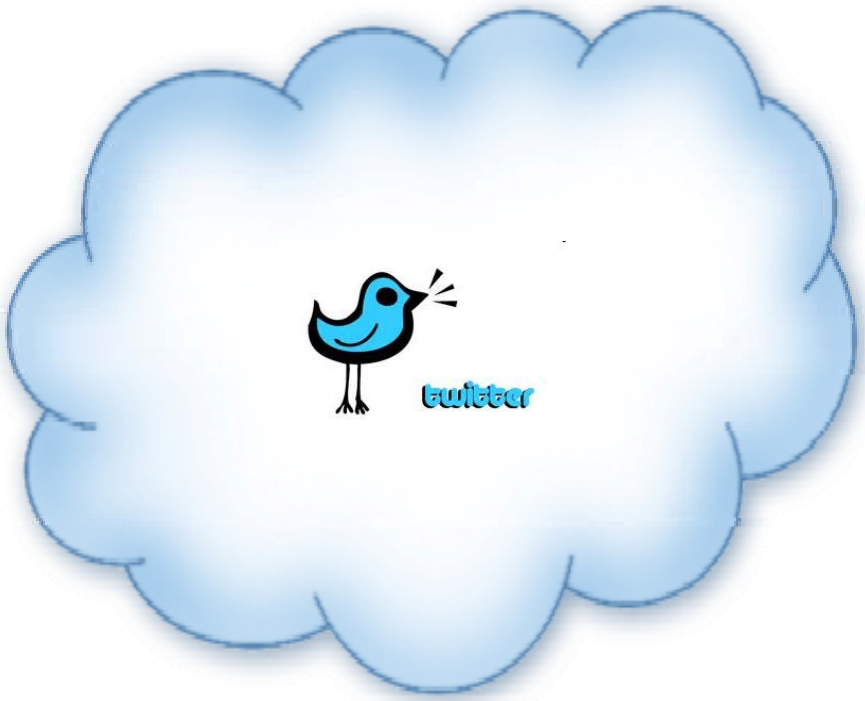
Likes

- Love the check guard feature
- Like the on-line bill pay feature
- Like that the ATMs are located all over the city
- Like the service representatives

Dislikes

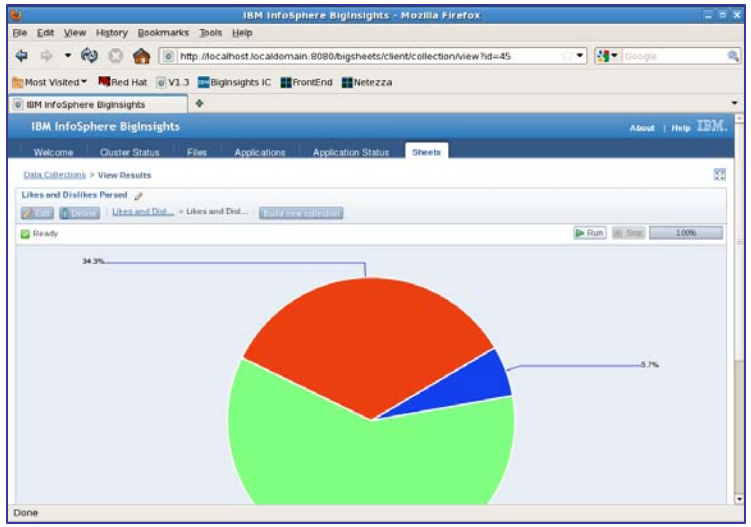
- Don't trust the on-line banking feature
- Don't like to wait in line for a long time
- Don't like the ATM fees
- Hate the overdraft fees

What You Just Saw In The Demo



Large volumes of raw, unstructured data

InfoSphere BigInsights



Valuable insights into customer sentiment

Service Oriented Finance Develops An Action Plan

Given this information, here's our action plan

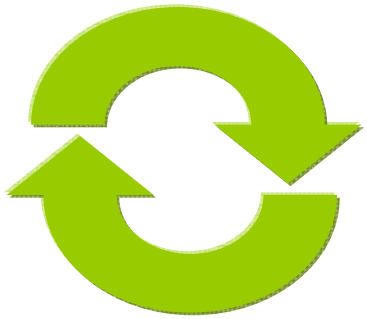
Service Oriented Finance Marketing VP

- Increase on-line banking usage by developing education materials
- Consider reducing ATM fees
- Consider offering check guard to more customers
- Consider reducing overdraft fees

BigInsights Eases The Pain Of Hadoop Setup – Standing Up An 18 Node Cluster

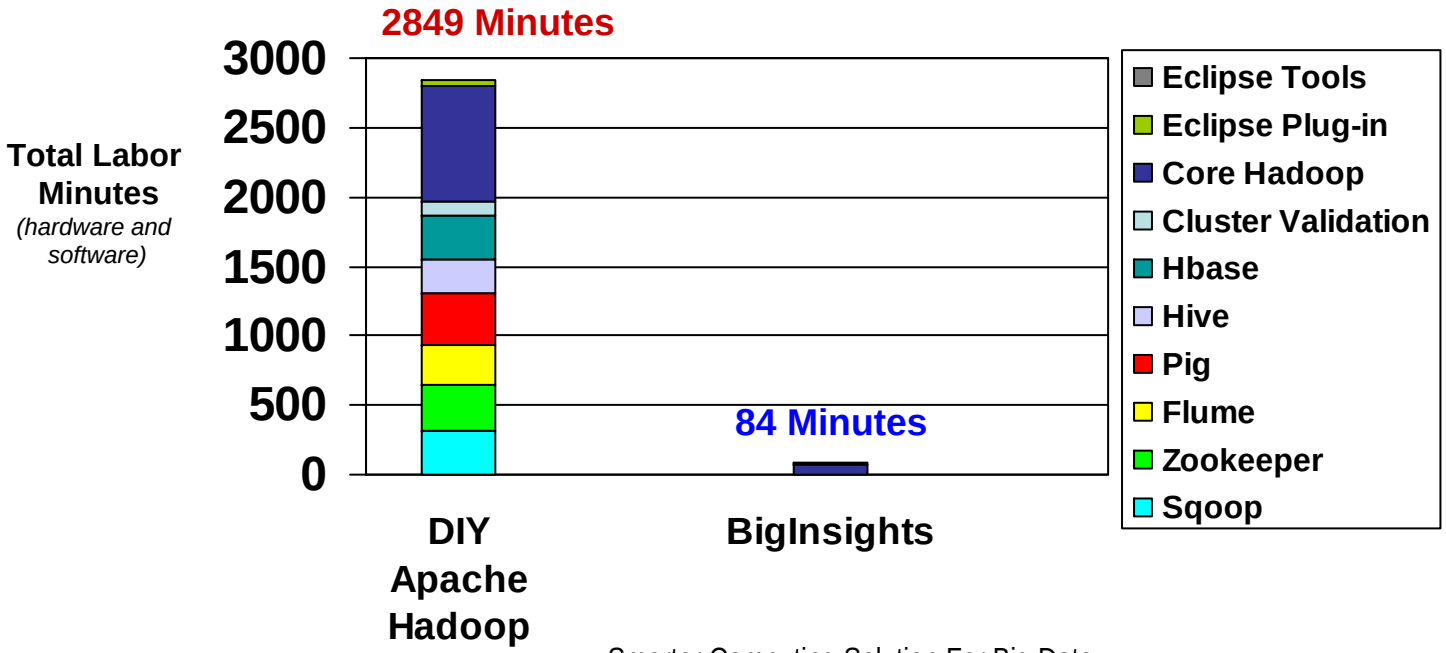
The Pain Of DIY Apache Hadoop

- Finding all the correct versions of the Hadoop subprojects
- Making sure they are all configured together properly
- Does not give a sense of a stable environment
- For every release of Hadoop, the entire process repeats
- Support only available from forums



Each upgrade of Hadoop will continue the cycle of pain

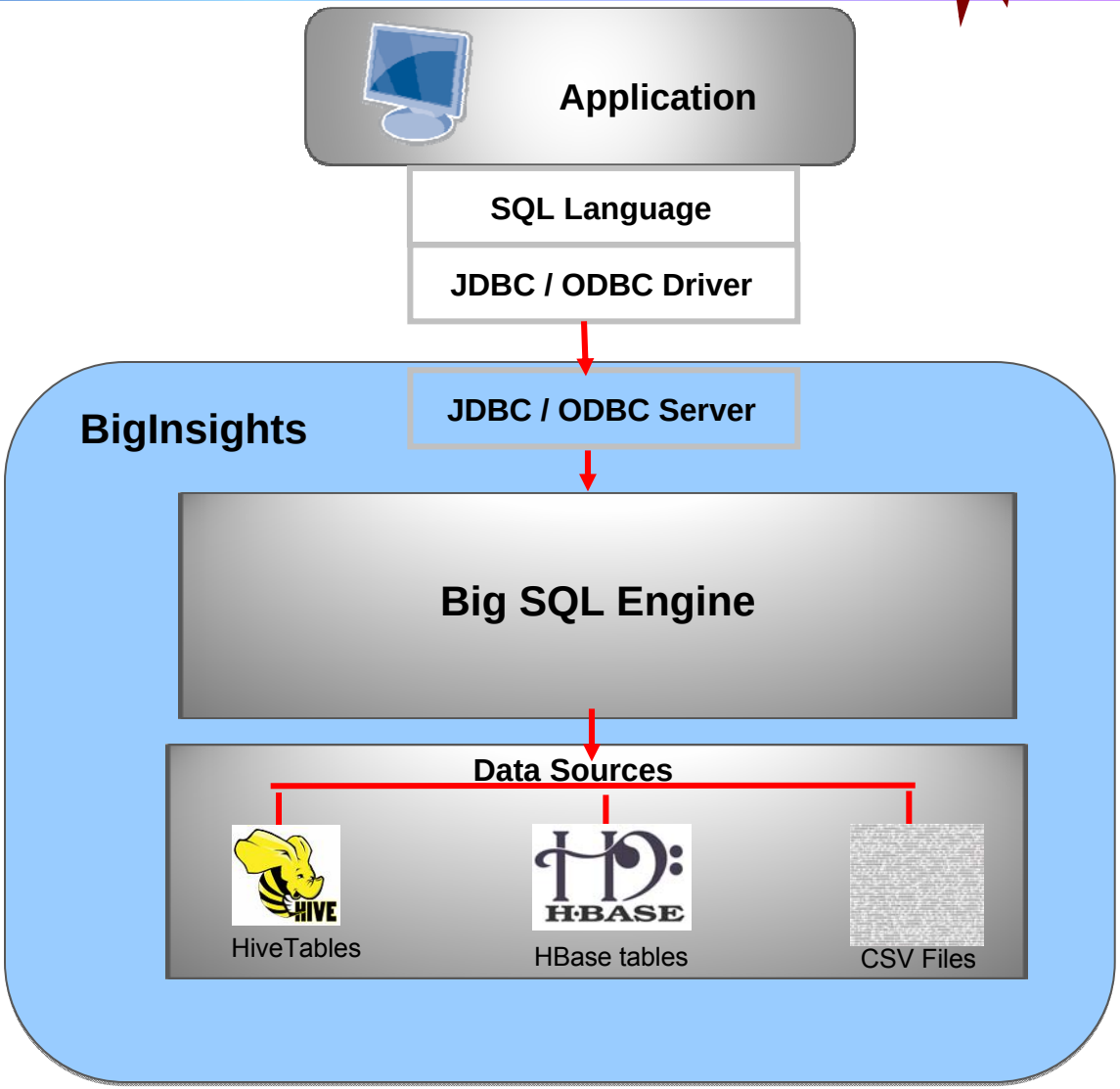
47.4 hours vs. 84 minutes



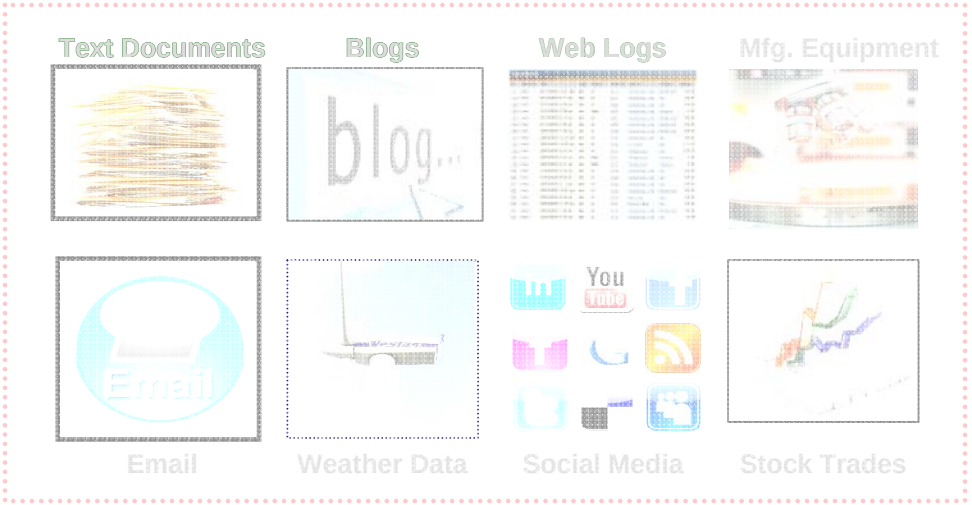
Introducing Big SQL

New

- Big SQL brings *robust* SQL support to the Hadoop ecosystem
- Standard SQL access to all data stored in BigInsights
- Use existing BI tools
- Opens up the data to much wider audience
  - Familiar and widely known syntax



Where Is The Big Data Coming From?



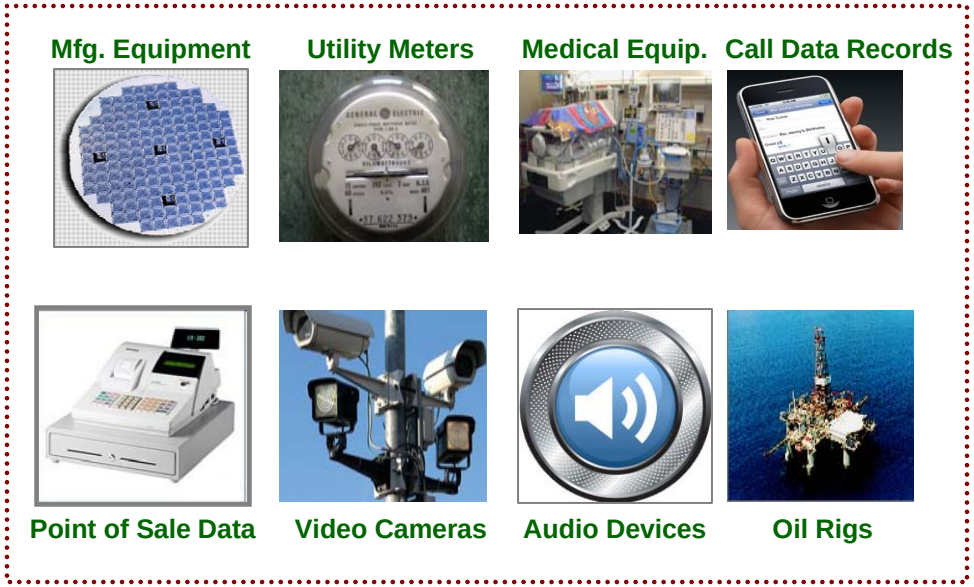
Data at rest

Data is stored on disk

Huge volumes of unstructured data

No pre-defined schemas

Too large for traditional tools to process in a timely manner



Data in motion

Data is typically not stored

Tremendous velocity

Multiple data sources

Huge volumes of unstructured data

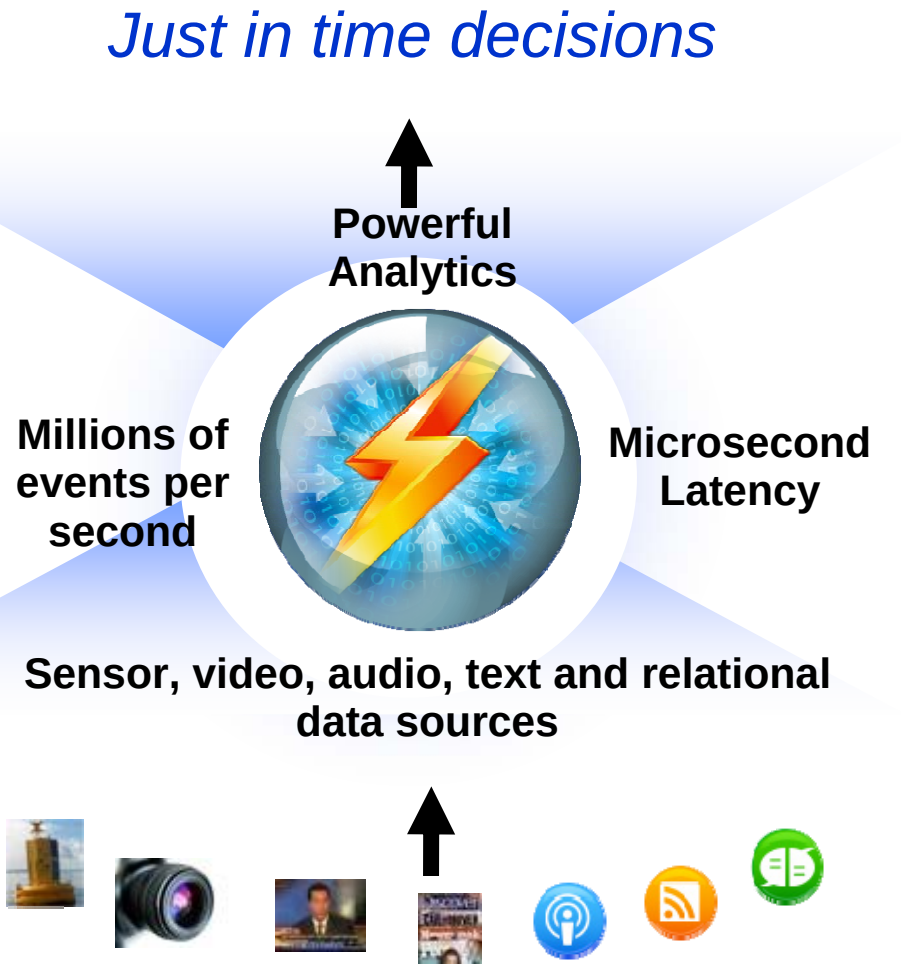
Ultra low latency required

What Is InfoSphere Streams?

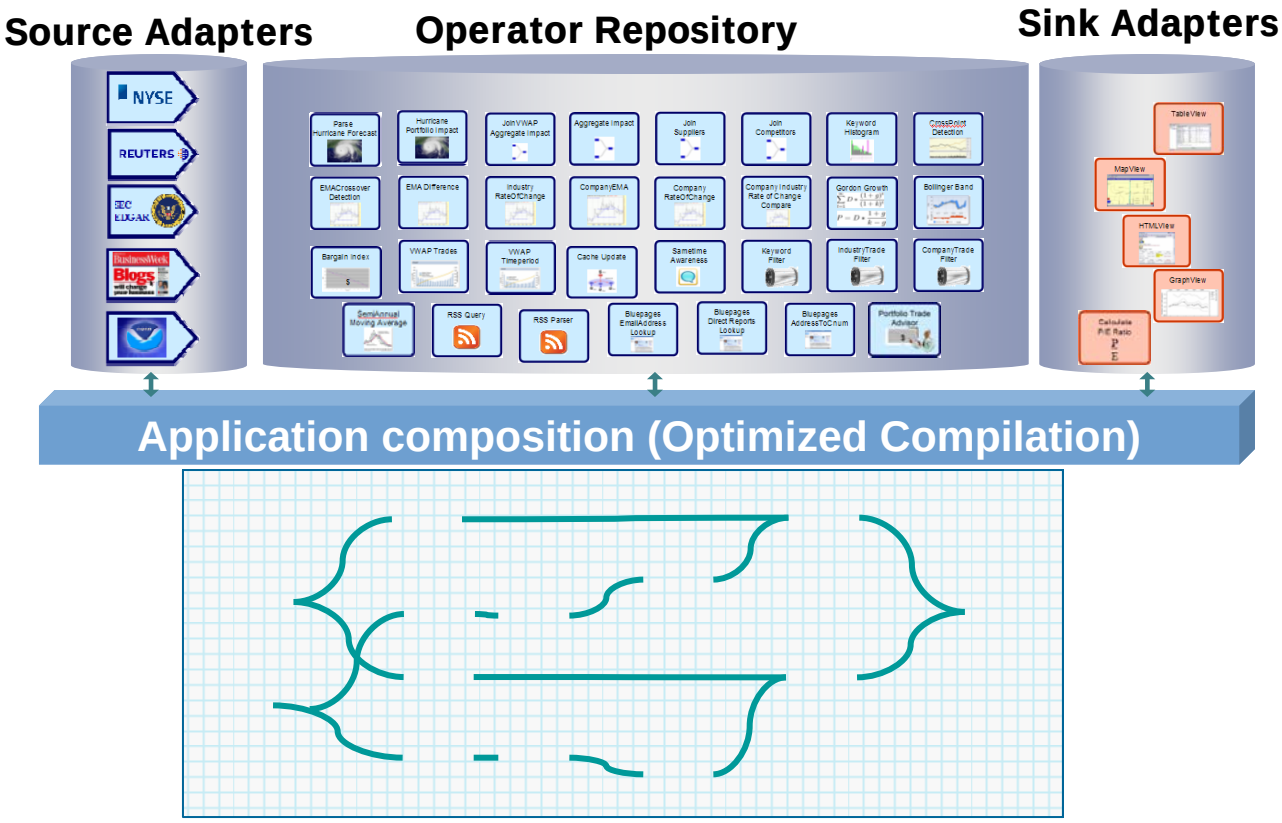
A platform for real-time analytics on BIG data

A Streams application has…

- Unique input requirements
  - Multiple sources, multiple varieties
- Demanding performance requirements
  - Millions of events per second
  - Process petabytes per day
  - Microsecond latency
  - May require multiple processors
- Sophisticated logic requirements
  - Correlations and computations between multiple input sources



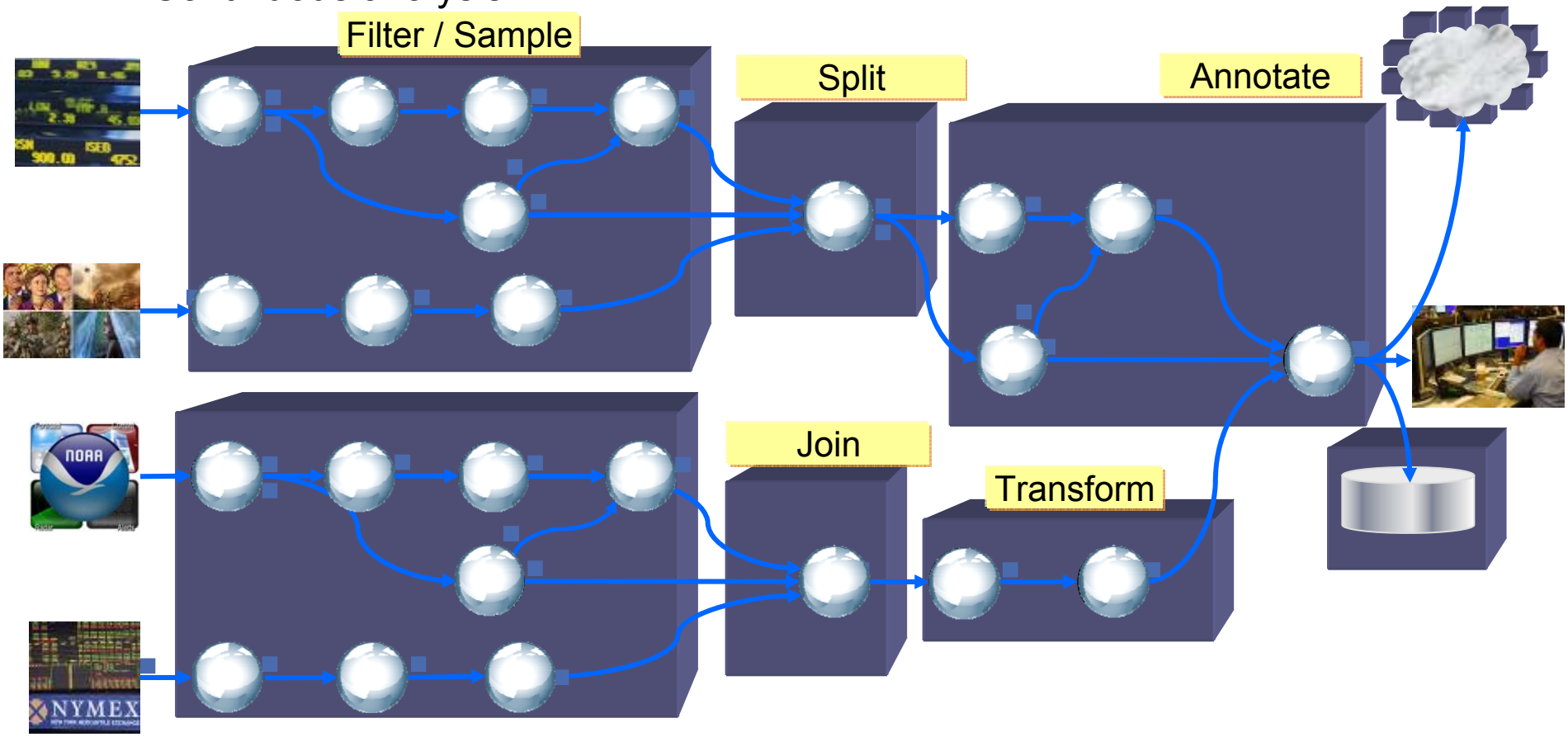
Streams Programming Model



How InfoSphere Streams Works

→ Continuous ingestion

→ Continuous analysis



Achieve scale:

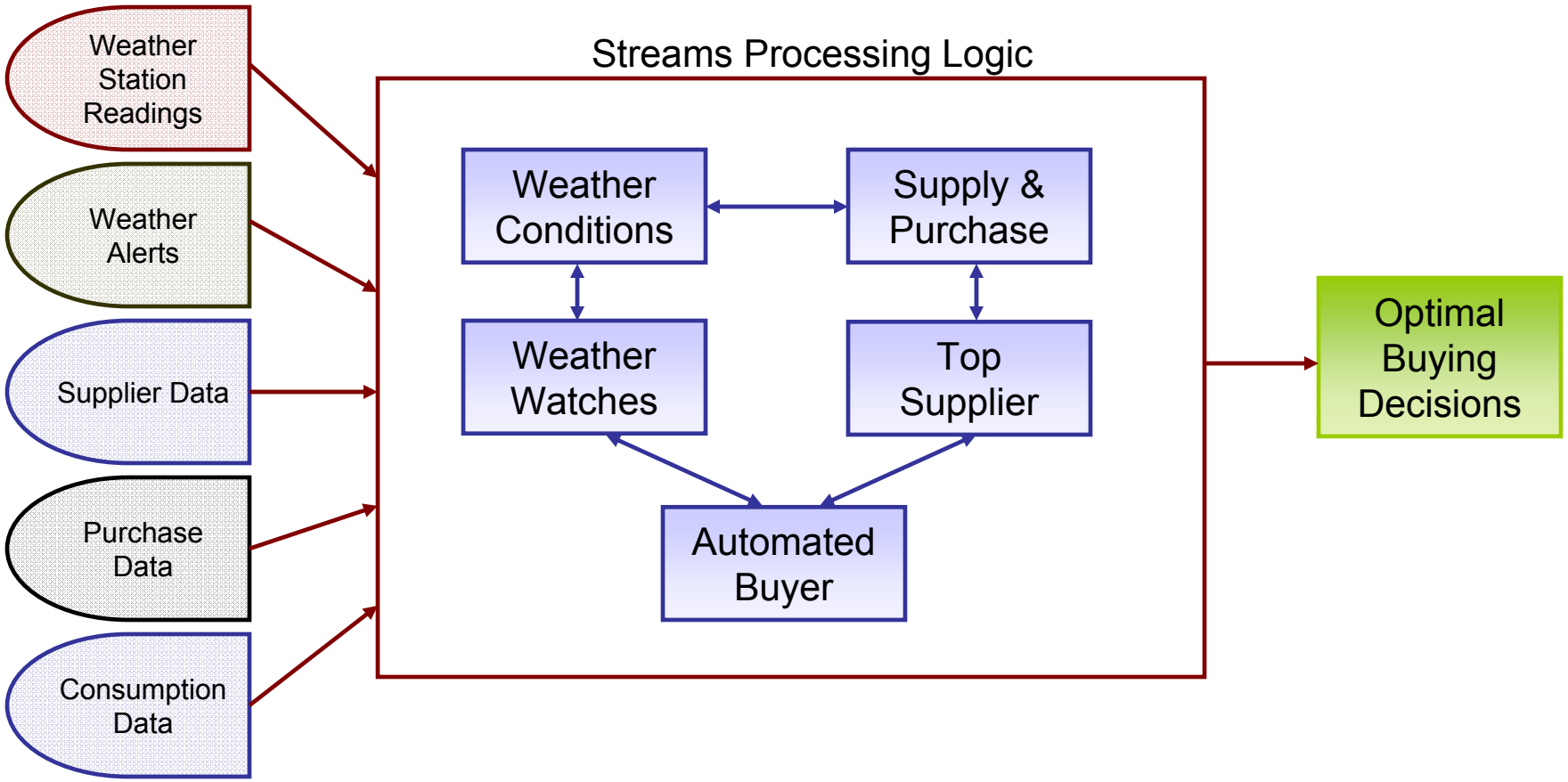
By partitioning applications into software components

By distributing across stream-connected hardware hosts

Friendly Juice Berry Purchasing Application

- Background information
  - Friendly Juice Co. makes a wide variety of healthy juice products
  - They need to purchase berries from suppliers around the United States
  - Weather plays an important role in the quality of the berries
    - Temperature, humidity, storms
- InfoSphere Streams solution
  - Analyze weather data from U.S. Weather Service
  - Take into account berry usage and supply
  - Provide for optimal purchase decisions
    - Automated purchasing when conditions are perfect
    - User initiated purchasing

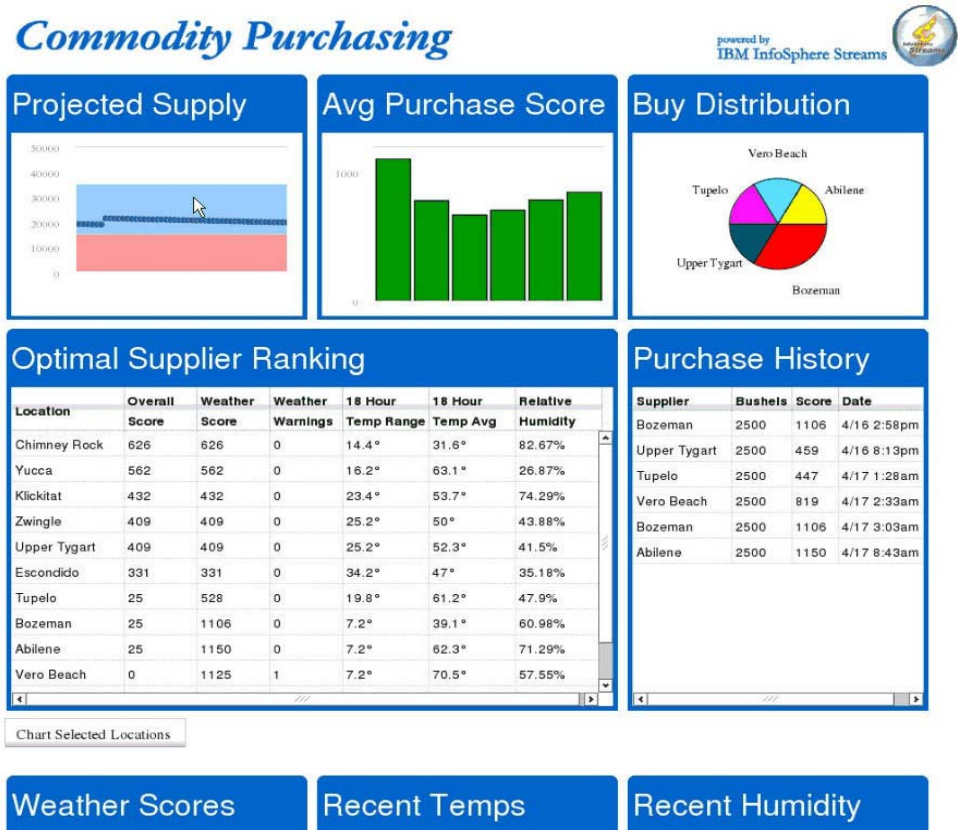
Purchasing Commodities Requires Analysis Of Multiple Real Time Data Sources And Sophisticated Logic



Demo: Berry Purchasing Application

What we will show in demo

- Application running in Eclipse development environment
- Application user interface
- New application requirement: Auditing purchases



Oracle NoSQL Requires 8X The Lines Of Code And Takes 3.5X Longer To Implement

InfoSphere Streams

IBM PowerLinux™ 7R2

63 Lines of code

4:10 Elapsed time

Oracle NoSQL Database

Oracle NoSQL Database

IBM System x3550 M3

499 Lines of code

14:15 Elapsed time

✓ **8X** **Lines of code**

✓ **3.5X** **Time to implement**

Source: IBM CPO internal studies

Big Data Analytics Needs Key Hardware Features Found In PowerLinux

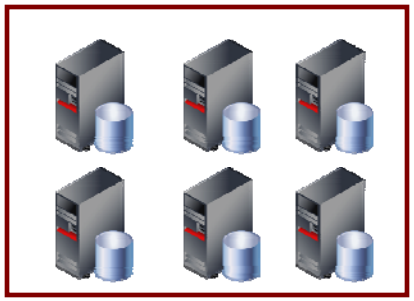
For Data At Rest

InfoSphere BigInsights

A storage-dense server is the right choice for Hadoop clients who need a storage-rich server that offers a low cost per terabyte.

Runs on

Server Cluster with Dense Local Storage



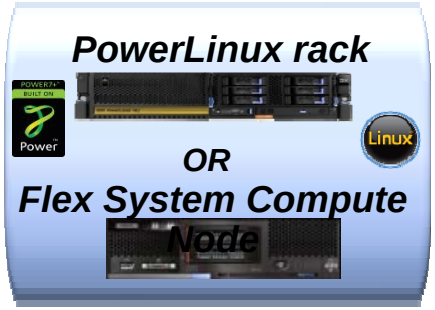
For Data in Motion

InfoSphere Streams

A memory-rich server is the right choice for running Streams to support large scale in-memory operations.

Runs on

Server Cluster with Lots of Main Memory



Up to 512 GB Mem
Up to 5.4 TB HDD

PowerLinux Costs Less Than Linux on x86

Linux only POWER7+
2U rack, one or two socket

PowerLinux 7R1

- 1 socket: 4-core @ 3.6 GHz
- 1 socket: 6 or 8-core @ 4.2 GHz

PowerLinux 7R2

- 2 sockets: 8-core @ 3.6 GHz
- 2 sockets: 8-core @ 4.2 GHz

redhat SUSE — **Operating Systems**

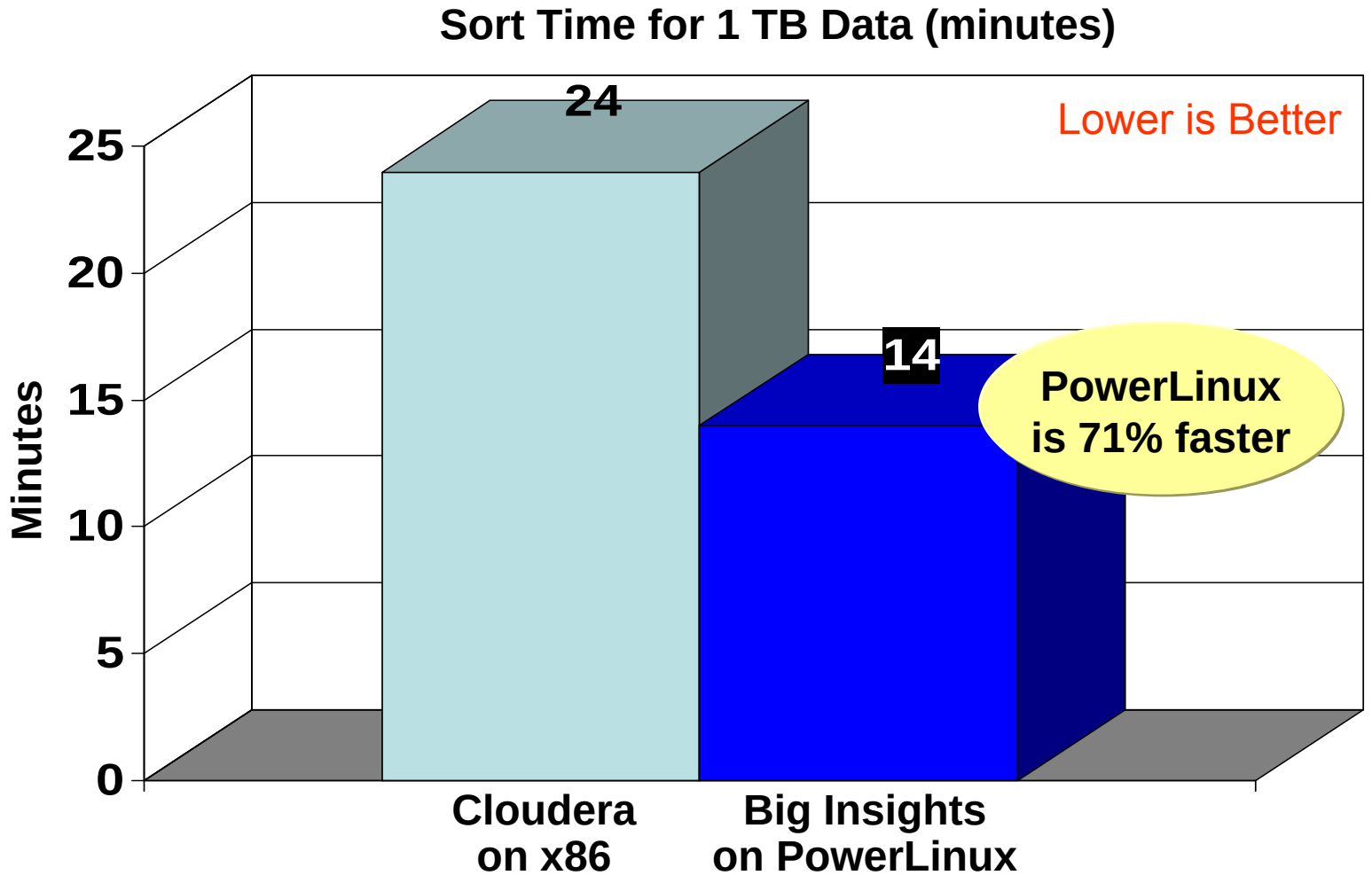
PowerVM — **Virtualization & Mgmt.**

Server Name	Dell R720	PowerLinux 7R1
Server list price*	**$6,206**	**$6,995**
Virtualization - OTC + 3yr. 9x5 SWMA	**$4,687** VMware vSphere Enterprise 5.1	**$3,920** PowerVM for IBM PowerLinux
Linux OS list price - RHEL, 1-2 sockets, unlimited guests, 9x5, 3 yr. sub./ supp.	**$5,697** Red Hat subscription and Red Hat support	**$4,489** Red Hat subscription and IBM support
Total list price:	**$16,590**	**$15,404**

Server model	Dell R720	IBM PowerLinux 7R1
Processor / cores	**8-core, 2.9 GHz Xeon E5-2690, Sandy Bridge**	**8-core, 4.2 GHz POWER7+**
Memory/HDD	32 GB mem, 2 x 147GB HDD, 1 Gb four port	Same memory, HDD, NIC

*Current as of Feb 5, 2013

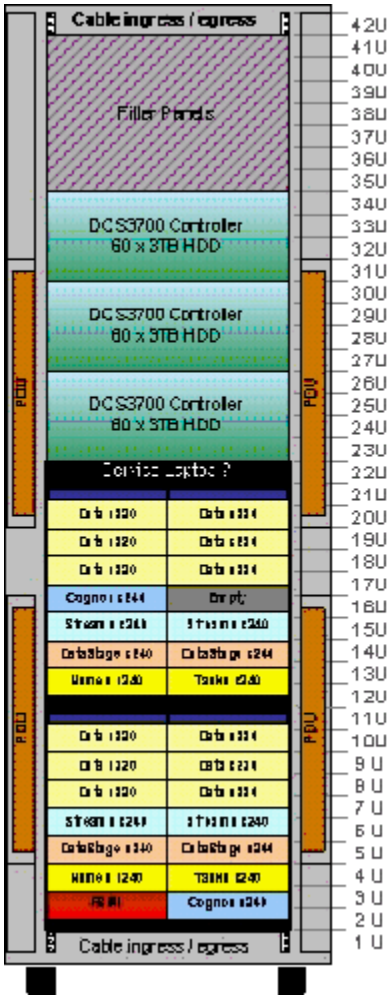
IBM On PowerLinux Performs Better Than Cloudera On Intel



*Both tests ran on 10-node Linux Server Cluster

* Based on results from publicly available sources to sort 1 terabyte per the Sort Benchmark 'rulebook': ttp://sortbenchmark.org/Yahoo2009.pdf
IBM source: http://domino.watson.ibm.com/library/CyberDig.nsf/1e4115aea78b6e7c85256b360066f0d4/f085753cf57c8c35852579e90050598f!OpenDocument&Highlight=0,rc25281
Cloudera source: http://www.slideshare.net/cloudera/hadoop-world-2011-hadoop-and-performance-todd-lipcon-yanpei-chen-cloudera

Examples Of Big Data Analytics Application By Industry



IBM Flex Enterprise

- PowerLinux p24L compute nodes
- DCS3700
- Hadoop
- GPFS or HDFS
- BigInsights and Streams
- Third Party Big Data SW
- Integration with DW (Netezza, SAS)
- Small, Medium, Large

- Energy and Utilities → Sensors and Smart Meters to parse unstructured Data in motion
- Health Care → patient data (ex Premature newborns), clinical data
- Every Industry → log correlation across the board, Customer sentiment
- Telco → Subscriber Data Management (real time promotions)
- Banking → Optimize Risk Management

PureData System For Operational Analytics
A Complete Solution For Structured Data

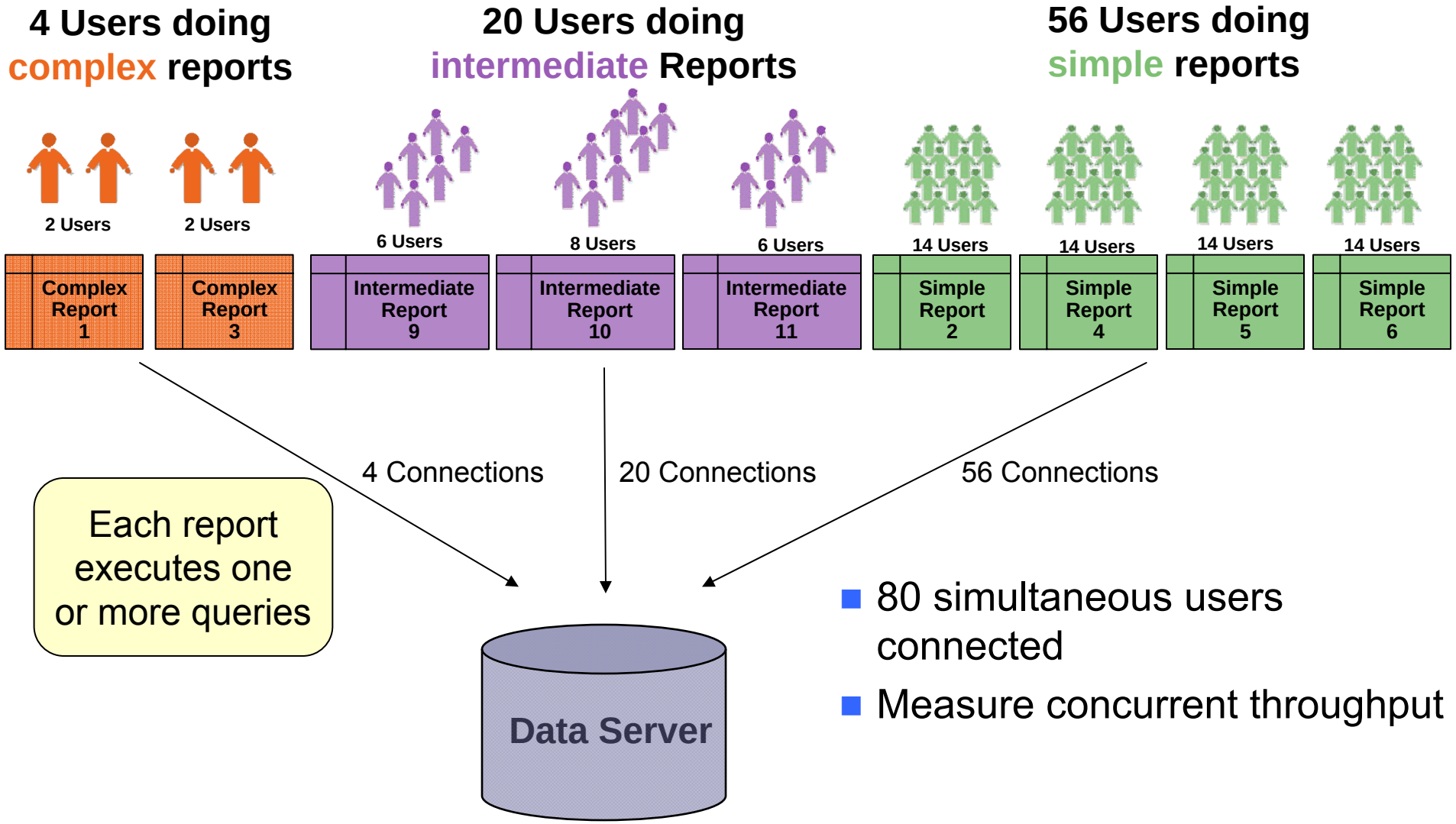
- Hardware
  - Power Systems servers
  - AIX v7.1
  - Storwize V7000 storage
  - EXP30 Ultra SSD
- Software
  - InfoSphere Warehouse v10.1
  - Tivoli Automation*
  - Optim Performance Manager
- Analytics
  - Cognos 10.1.1

- IBM POWER7 P740 & P730 16 Core servers @ 3.55GHz
- IBM Storwize® V7000 with 900GB drives
- Ultra SSD I/O Drawers, each with six 387GB SSD
- Blade Network Technologies 10G and 1G Ethernet switches
- Brocade SAN switches (SAN48B-5)

* For Failover Orchestration

More Info: IBM PureData System for Operational Analytics Brochure
http://public.dhe.ibm.com/common/ssi/ecm/en/wad12351usen/WAD12351USEN.PDF

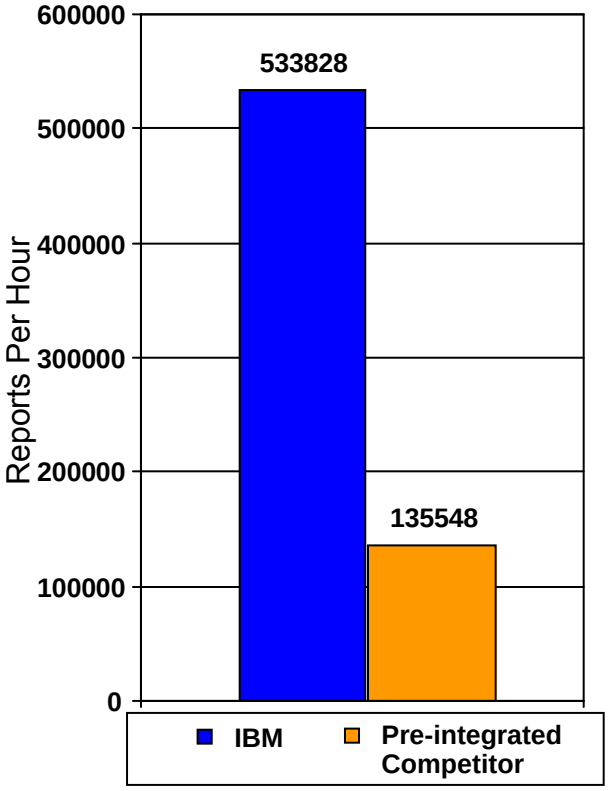
Operational Analytics - BI Day Workload Measures High Levels Of Concurrently Executing Workloads



Note: Distribution of complex, intermediate, and simple workloads based on Forrester Research, Profiling the Analytic End User for Business Intelligence, 2004

IBM Operational Analytics Delivers More Throughput For Concurrent Operational Reports

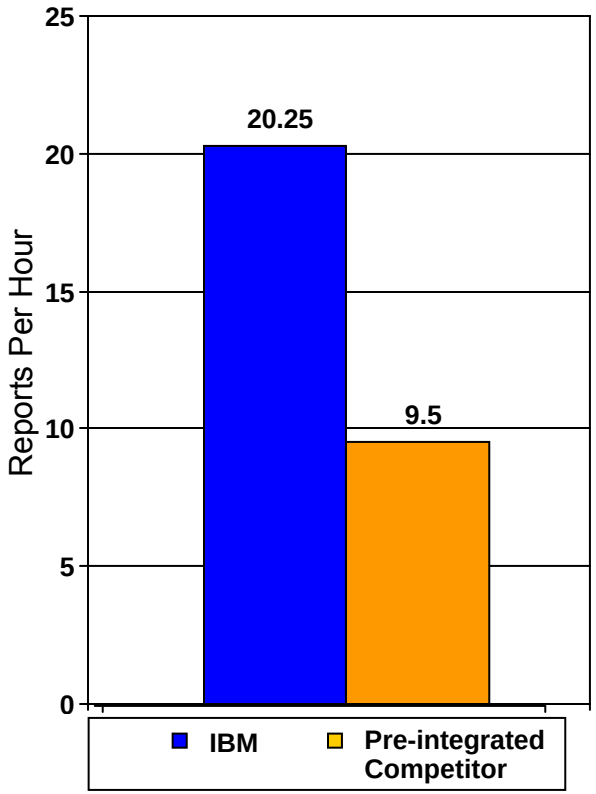
3.9X More Simple Reports



Reports Per Hour at 10 TB data size

(Higher is Better)

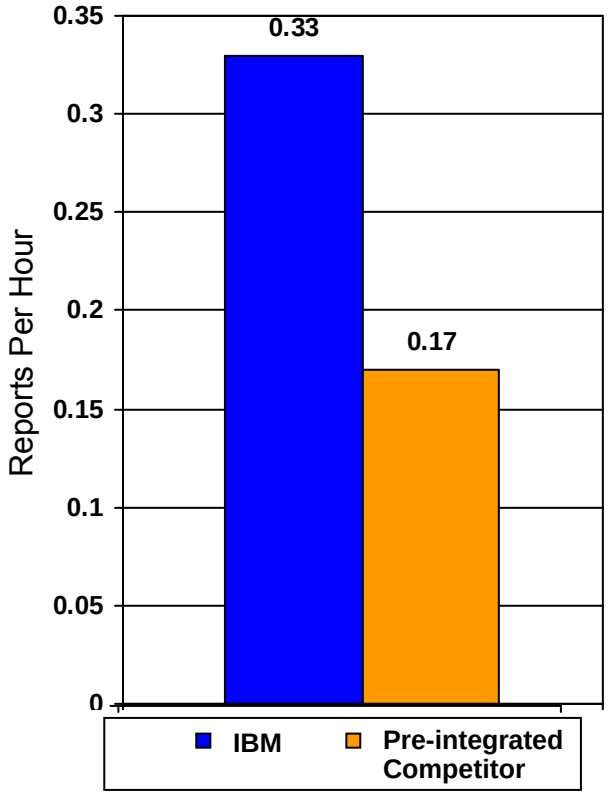
2X More Intermediate Reports



Reports Per Hour at 10 TB data size

(Higher is Better)

1.9x More Complex Reports

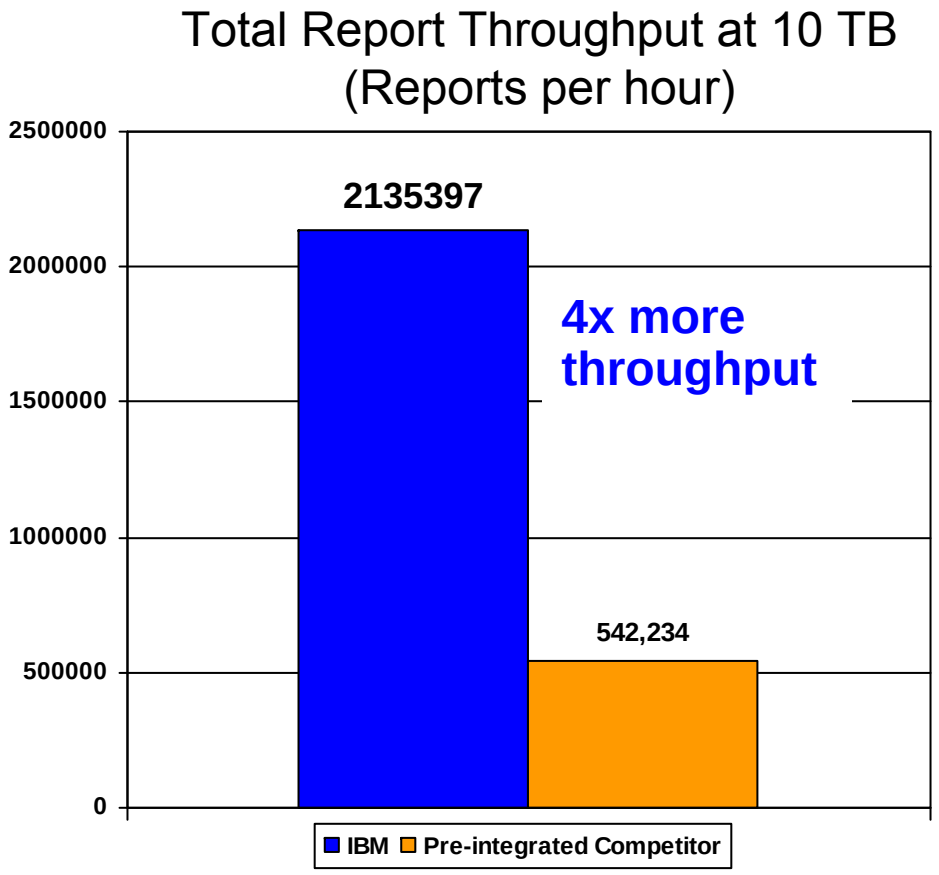


Reports Per Hour at 10 TB data size

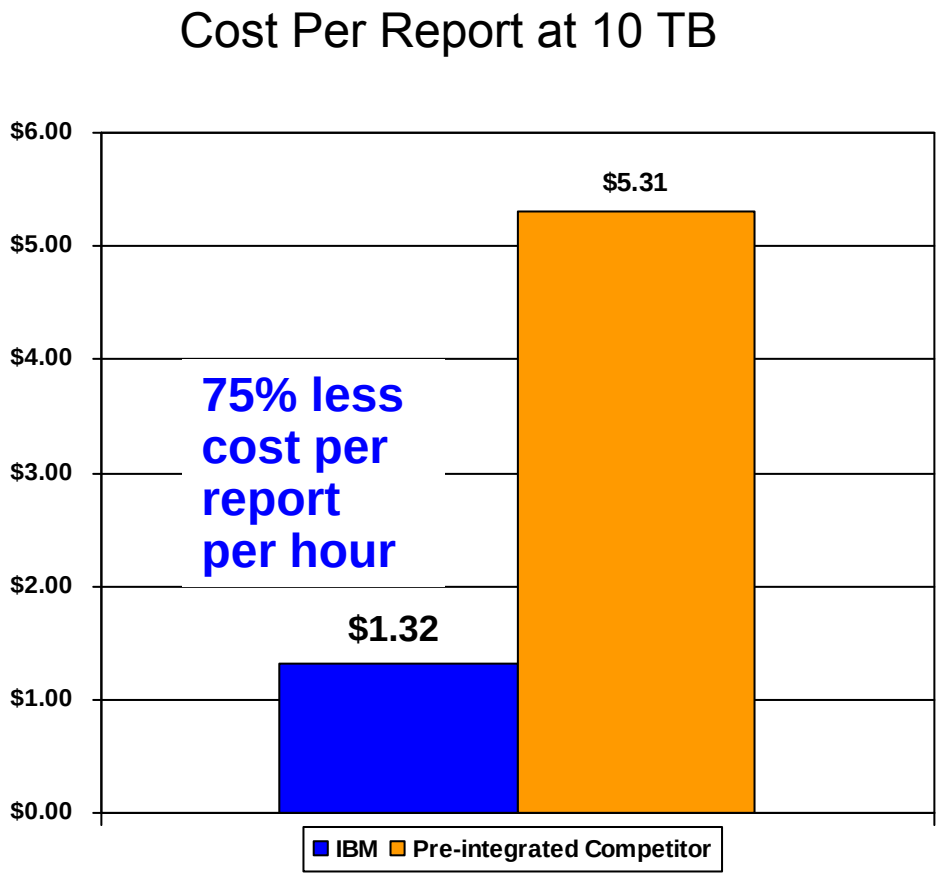
(Higher is Better)

Performance numbers may vary based on workload profiles.

IBM Operational Analytics Delivers More Throughput For Concurrent Operational Reports



(Higher Throughput is Better)

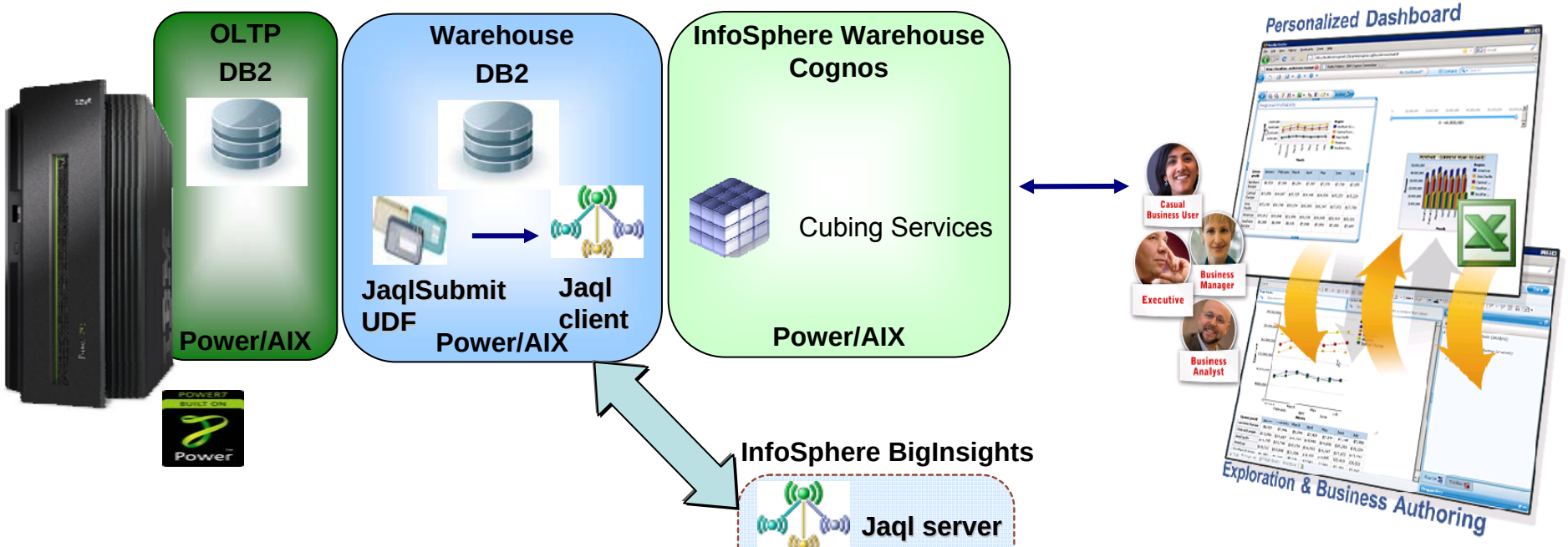


(Lower Cost is Better)

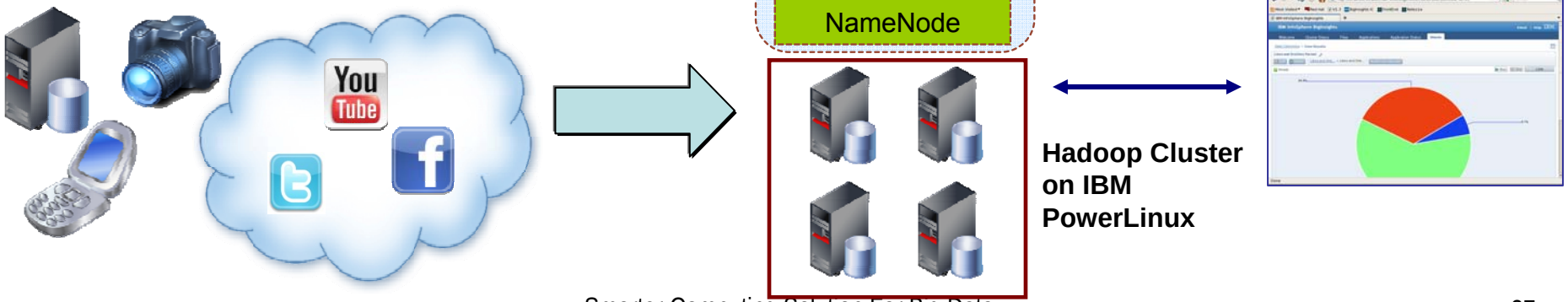
Performance numbers may vary based on workload profiles. 3 year total cost of acquisition includes hardware, software, service & support. Based on US list prices, prices will vary by country.

Integrate Structured And Unstructured Data On POWER Systems To Derive Insights

Structured data on Power



Unstructured data

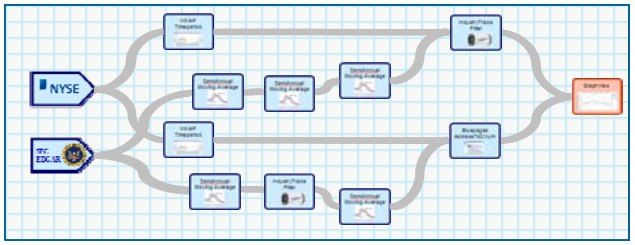


IBM Can Help You Solve Big Data Problems

InfoSphere BigInsights

Data at Rest

InfoSphere Streams



Data in Motion

PureData for Operational Analytics

Structured Data

- Big Data problems dealing with new unstructured data require new algorithms running on large clusters of low cost servers
  - Hadoop and InfoSphere Streams are proven frameworks for these problems
  - Problems that could not be solved before
- Forrester: “"IBM has the deepest Hadoop platform and application portfolio"
- BigInsights on PowerLinux performs better than Cloudera on x86
- InfoSphere Streams is far more productive and requires much less code than using Oracle NoSQL DB for a streaming application
- IBM PureData for Operational Analytics provides a complete solution for dealing with structured data
  - Higher concurrent throughput and lower cost per report than the competition