7 Kasım 2012 - Çırağan Palace Kempinski

IBMConnected 2012 Istanbul

Learn. Collaborate. Innovate.

A Smarter Computing Solution for Big Data

Jim Williams
IBM

Where is the Big Data Coming From?

Text Documents

Blogs

Web Logs

Mfg. Equipment

Email

Weather Data

Social Media

Stock Trades

Data at rest

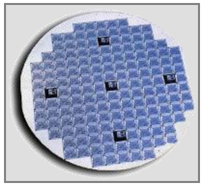
Data is stored on disk

Huge volumes of unstructured data

No pre-defined schemas

Too large for traditional tools to process in a timely manner

Mfg. Equipment



Utility Meters

Medical Equip.

Call Data Records

Point of Sale Data

Video Cameras

Audio Devices

Oil Rigs

Data in motion

Data is typically not stored

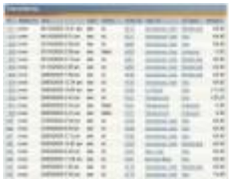
Tremendous velocity

Multiple data sources

Huge volumes of unstructured data

Ultra low latency required

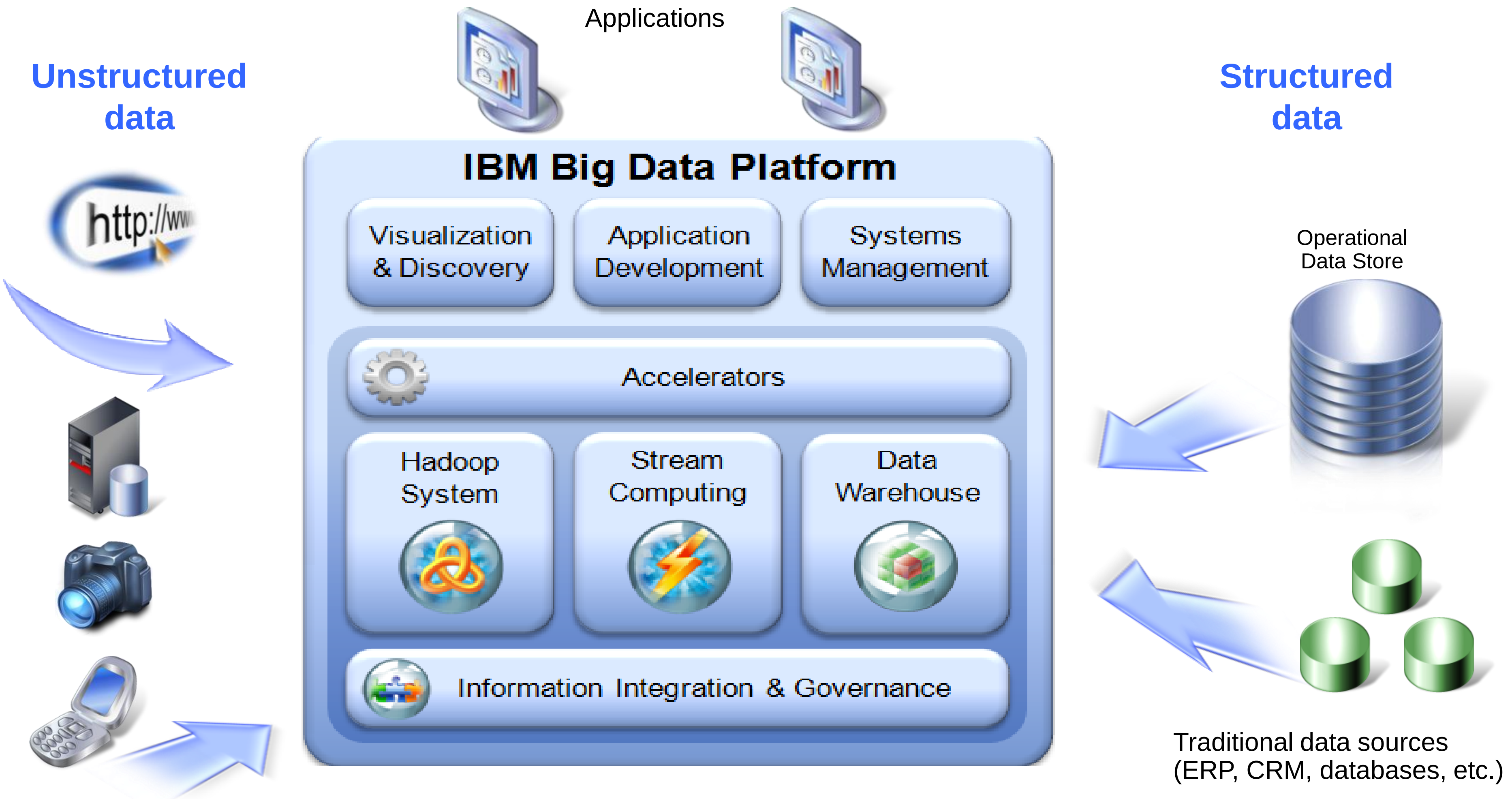
Gaining Value from Data at Rest

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



Applications

Unstructured data

Structured data

IBM Big Data Platform

Visualization & Discovery

Application Development

Systems Management

Accelerators

Hadoop System

Stream Computing

Data Warehouse

Information Integration & Governance

Operational Data Store

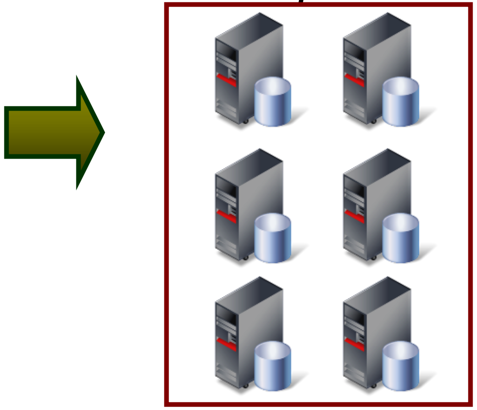
Traditional data sources
(ERP, CRM, databases, etc.)

Source data (Web, sensors, logs, media, etc.)

New Programming Models and Low Cost Hardware For Handling Unstructured Data

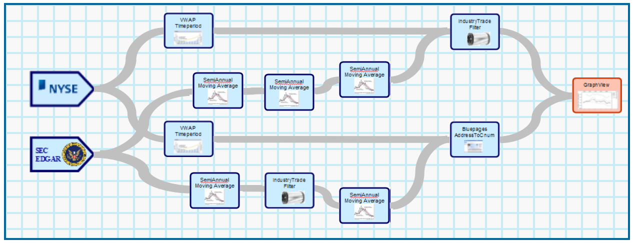
- Apache Hadoop and InfoSphere Streams
  - Proven frameworks to process large amounts of data
  - Hadoop for data at rest, Streams for data in motion
  - Enable applications to transparently work with large clusters of nodes in parallel

Hadoop Cluster

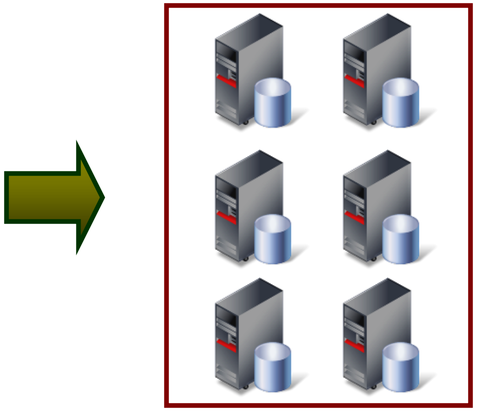


Clusters of low cost **PowerLinux** servers that are ideal for Hadoop and Streams

InfoSphere Streams


NYSE
SEC EDGAR
VWAP Timeperiod
SemiAnnual Moving Average
SemiAnnual Moving Average
SemiAnnual Moving Average
IndustryTrade Filter
VWAP Timeperiod
SemiAnnual Moving Average
IndustryTrade Filter
SemiAnnual Moving Average
Bluepages AddressToCnum
GraphView

Streams Cluster

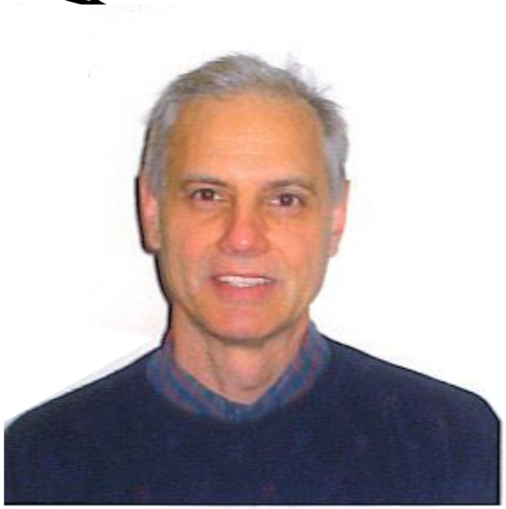


Service Oriented Finance Wants To Grow Their Business

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

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.



Service Oriented Finance Marketing VP

IBM

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

Feelings - Attitudes

Emotions - Opinions

Thoughts - Desires

Huge volumes of unstructured data

Trying to determine…

Product demand

New product acceptance

Competitive threats

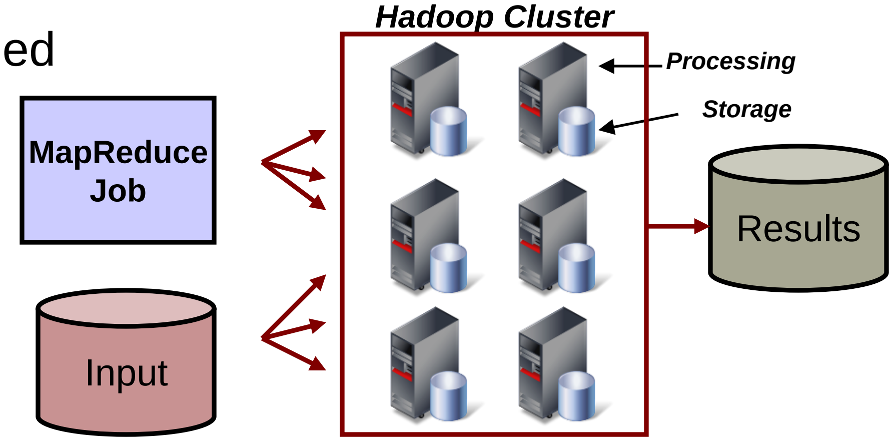
Threats to brand reputations

Advertisement targets

Finding sentiment from social media site data

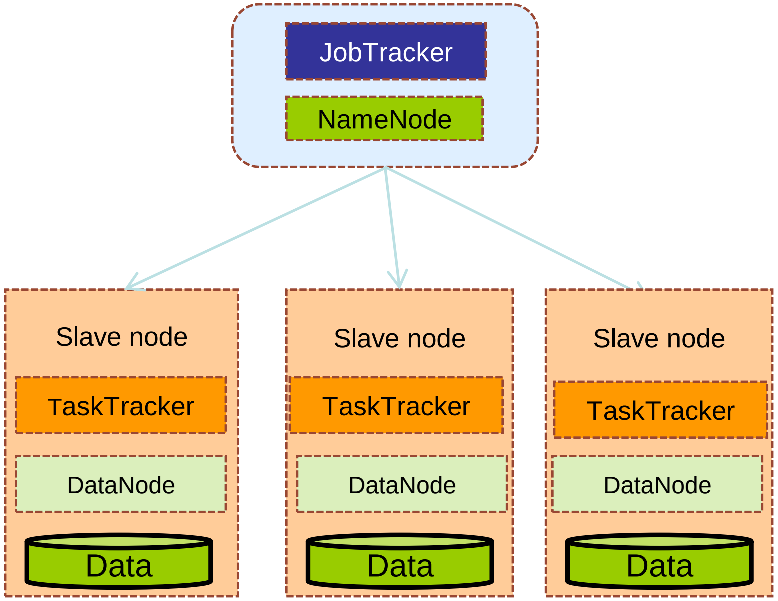
Apache Hadoop

- Open source framework for data-intensive applications
  - Proven approach to processing Big Data
  - Inspired by Google (MapReduce)
- Enables applications to transparently work with thousands of nodes and petabytes of data in a highly parallel, cost effective manner
  - Hadoop “node” is a processor and disks
  - Nodes can be combined into clusters
  - Original data is parceled out to nodes
  - MapReduce jobs are sent to nodes
  - Results from each node are assembled



Key Aspects Of Hadoop

- 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
  - Elastically scalable
  - Assumes nodes will fail - achieves reliability by replicating data across multiple nodes

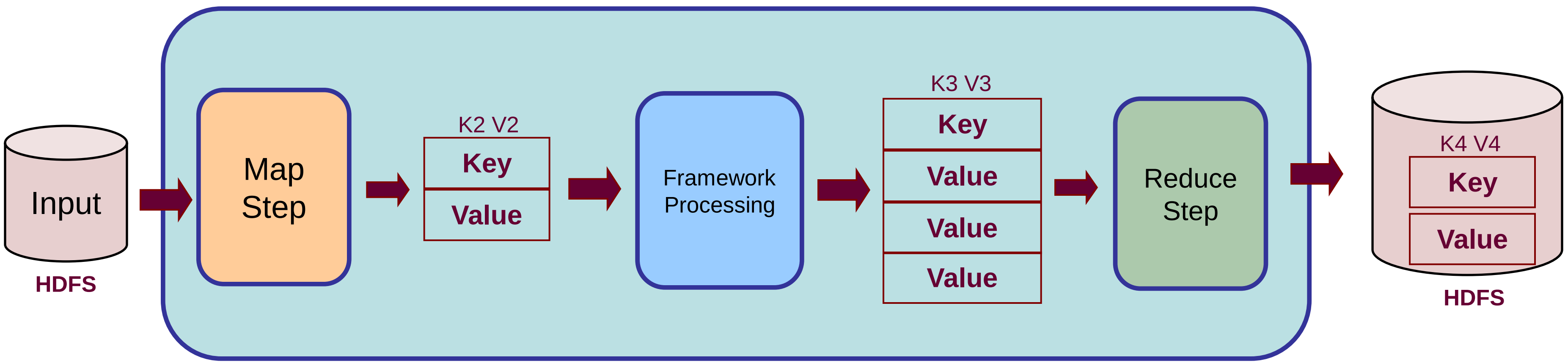


- MapReduce framework
  - A processing technique that produces results despite each node working independently on a portion of the data
  - MapReduce job is cloned and sent out to the nodes - jobs run in parallel
  - Framework handles
    - Shuffling data to correct nodes
    - Monitoring with heartbeats

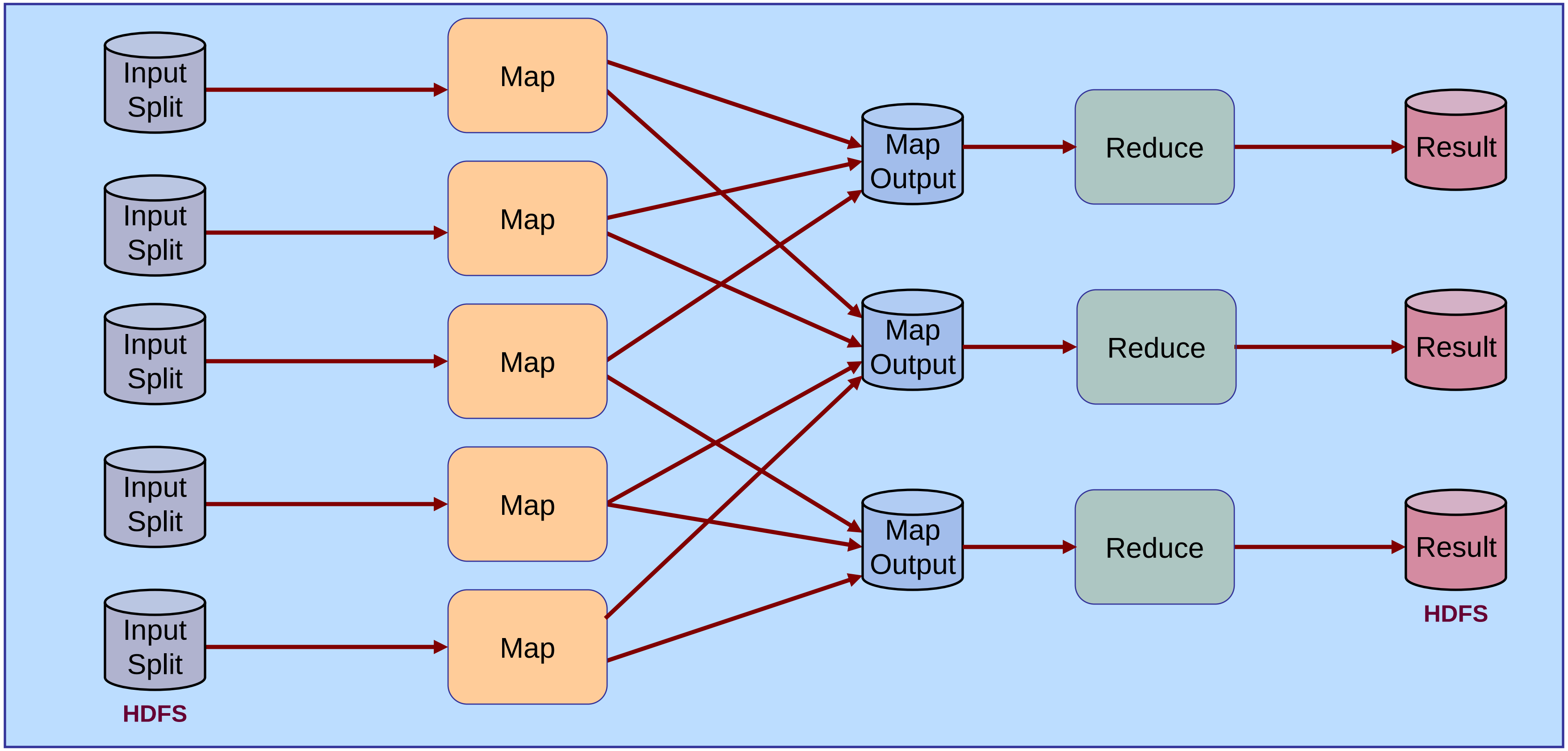
MapReduce Basics

- Map and Reduce are steps in the framework that a programmer implements
- Hadoop framework orchestrates Map and Reduce steps
- MapReduce jobs are sent out to each node to run
- MapReduce jobs run in parallel across nodes
- The steps process key/value pairs in some way
- How the steps manipulate the pairs defines the solution

View Inside One MapReduce Job



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

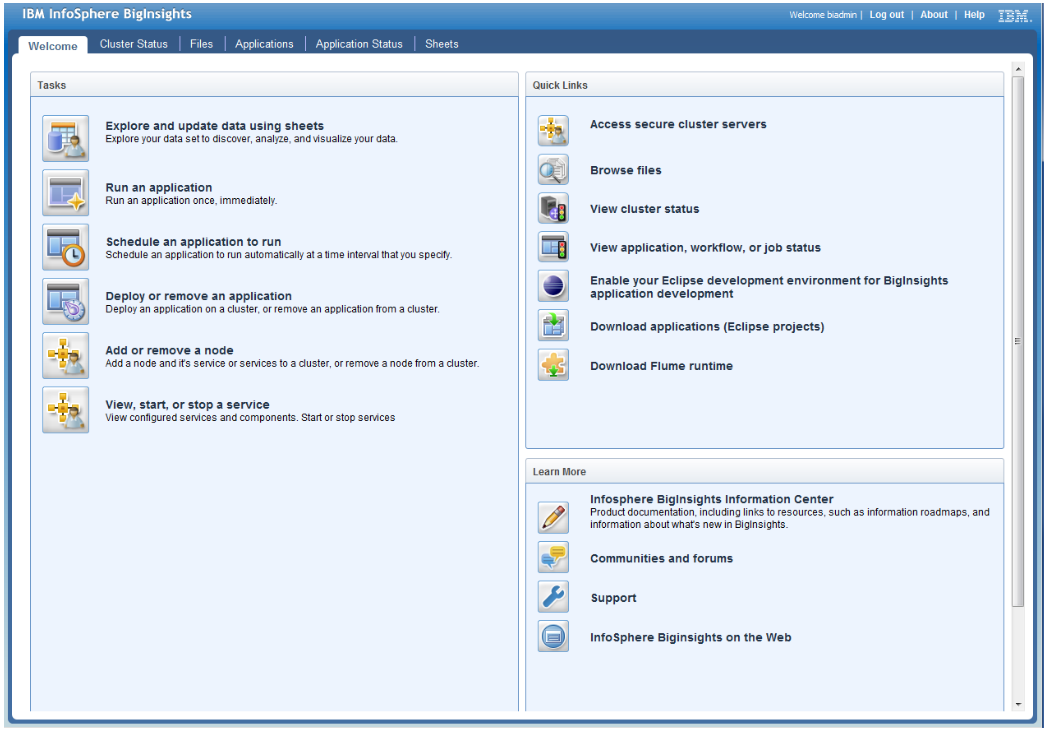
1. Reduce steps are invoked with one key and all values for that key

2. Reduce steps write out final key/value pairs

BigInsights – Makes It Easy

- Web based management console
- Security enhancements
  - LDAP authentication
- Administrator enhancements
  - Installation and configuration
  - Data import/export tools
  - Monitoring tools
- Developer enhancements
  - Eclipse tools
  - Job management tools
- Integration enhancements
  - Database/warehouse integration
- Business user enhancements
  - Spreadsheet style tool for users without Java skills

InfoSphere BigInsights Console


IBM InfoSphere BigInsights
Welcome biadmin | Log out | About | Help
IBM.
Welcome
Cluster Status
Files
Applications
Application Status
Sheets

Tasks

Explore and update data using sheets
Explore your data set to discover, analyze, and visualize your data.

Run an application
Run an application once, immediately.

Schedule an application to run
Schedule an application to run automatically at a time interval that you specify.

Deploy or remove an application
Deploy an application on a cluster, or remove an application from a cluster.

Add or remove a node
Add a node and it's service or services to a cluster, or remove a node from a cluster.

View, start, or stop a service
View configured services and components. Start or stop services

Quick Links

Access secure cluster servers

Browse files

View cluster status

View application, workflow, or job status

Enable your Eclipse development environment for BigInsights application development

Download applications (Eclipse projects)

Download Flume runtime

Learn More

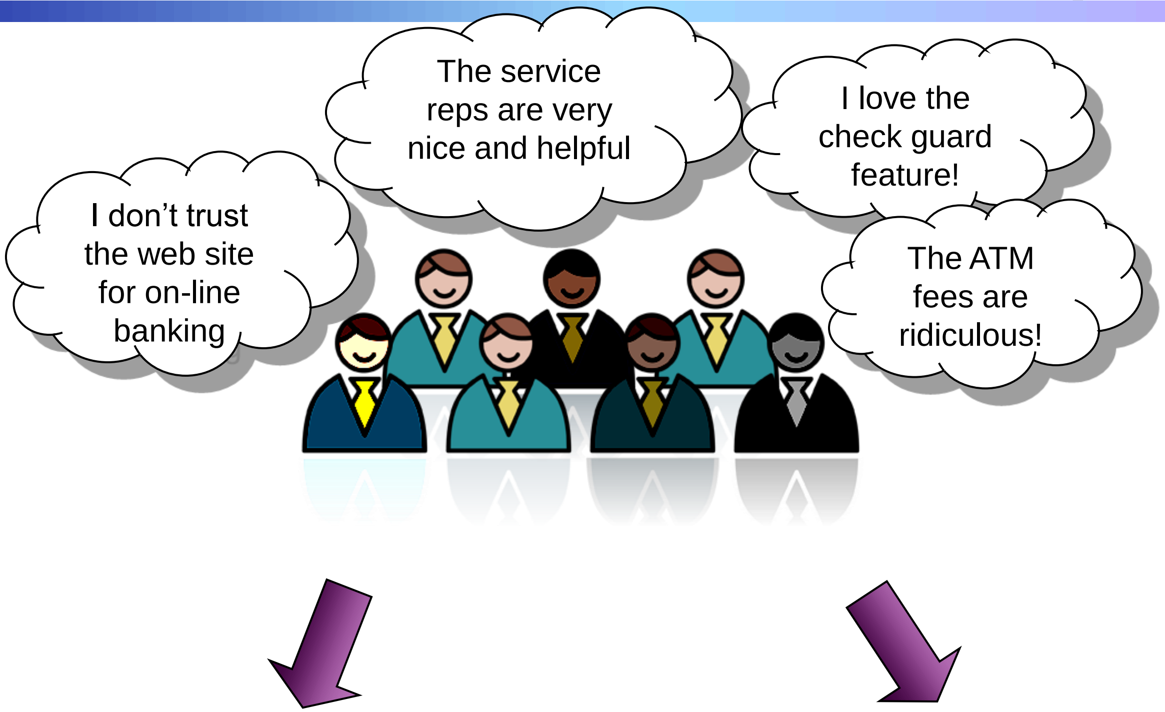
Infosphere BigInsights Information Center
Product documentation, including links to resources, such as information roadmaps, and information about what's new in BigInsights.

Communities and forums

Support

InfoSphere Biginsights on the Web

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


I don't trust the web site for on-line banking
The service reps are very nice and helpful
I love the check guard feature!
The ATM fees are ridiculous!

Big Bank Customers: likes and dislikes

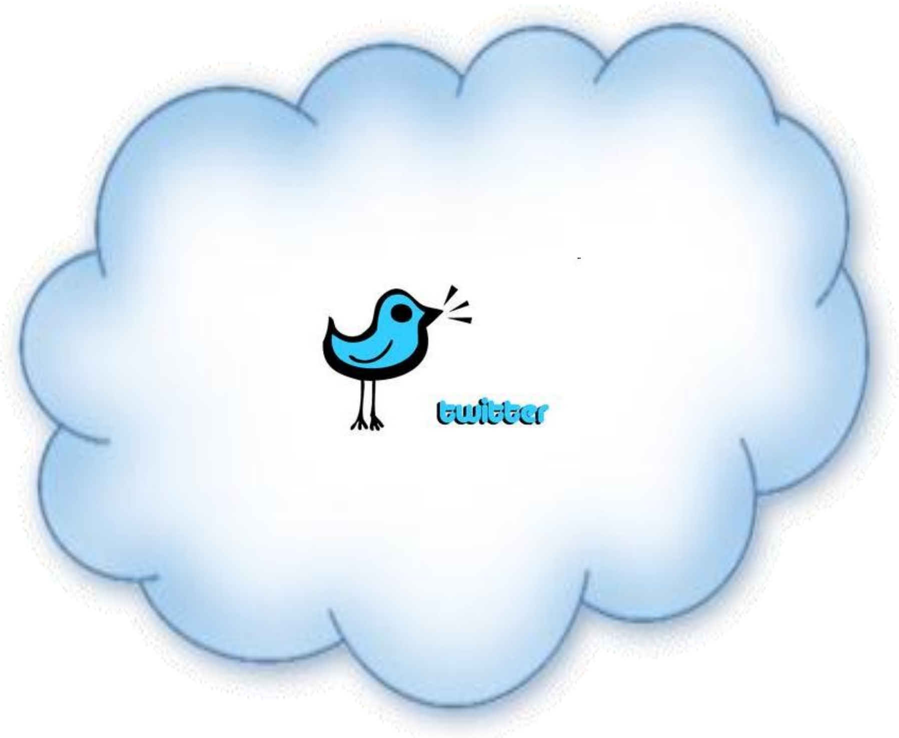
Likes

- 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

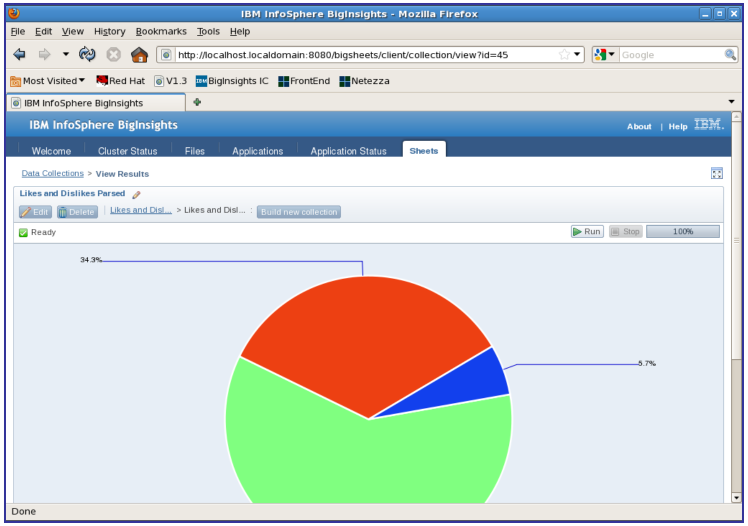
- Don't trust the on-line banking feature
- Don't like to wait in line for a long time
- Don't like the ATM fees
- Hate the overdraft fees

What You Just Saw In The Demo



Large volumes of raw, unstructured data

InfoSphere BigInsights


IBM InfoSphere BigInsights - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost.localdomain:8080/bigsheets/client/collection/view?id=45
Google
Most Visited
Red Hat
V1.3
BigInsights IC
FrontEnd
Netezza
IBM InfoSphere BigInsights
IBM InfoSphere BigInsights
About | Help
IBM.
Welcome
Cluster Status
Files
Applications
Application Status
Sheets
Data Collections > View Results
Likes and Dislikes Parsed
Edit
Delete
Likes and Disl... > Likes and Disl... :
Build new collection
Ready
Run
Stop
100%
34.3%
5.7%
Done

Valuable insights into customer sentiment

New PowerLinux Servers Ideal For Big Data

Key Benefits

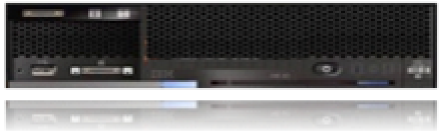
- Up to 17% lower power/cooling costs than x86 rack servers
- Industry standard (Redhat & SUSE) Linux only servers, optimized for POWER architecture
- Competitively priced compared to x86 Linux
- BigInsights on PowerLinux runs 71% faster than Cloudera on x86

IBM PowerLinux 7R1

- **Linux only POWER7**
- **2U rack, 1 or 2 socket**

IBM PowerLinux 7R2

IBM Flex System p24L

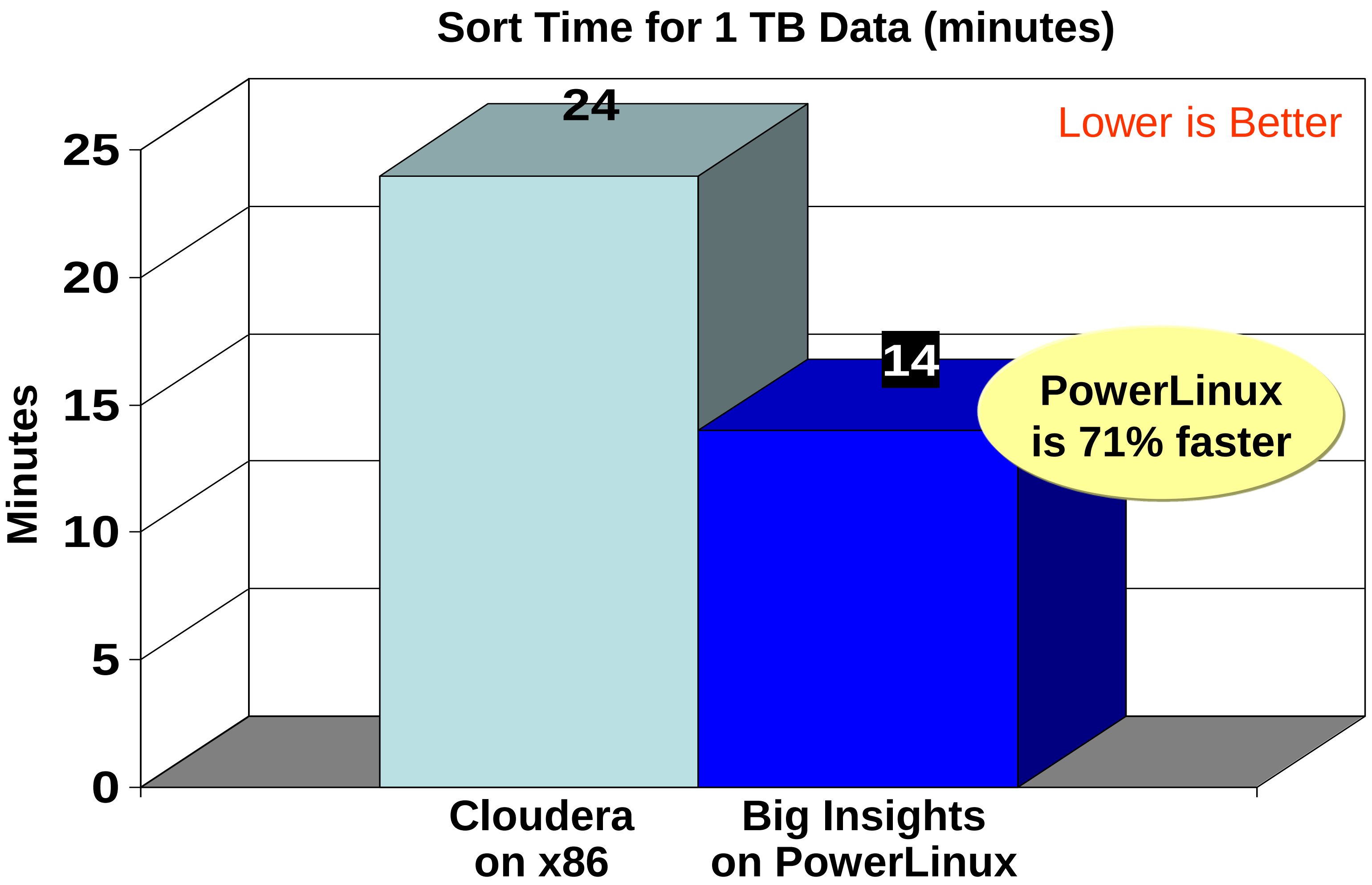


PowerLinux Compute Node

Linux redhat suse

More Info: http://www.ibm.com/systems/power/software/linux/powerlinux/bigdata.html

IBM On PowerLinux Performs Better Than Cloudera On Intel



Lower is Better

PowerLinux is 71% faster

*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

Where Is The Big Data Coming From?

Text Documents

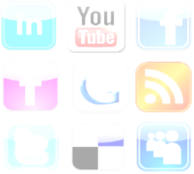
Blogs

Web Logs

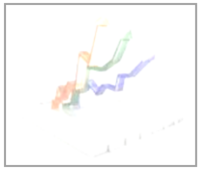
Mfg. Equipment

Email

Weather Data



Social Media



Stock Trades

Data at rest

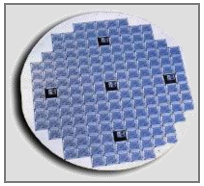
Data is stored on disk

Huge volumes of unstructured data

No pre-defined schemas

Too large for traditional tools to process in a timely manner

Mfg. Equipment



Utility Meters

Medical Equip.

Call Data Records

Point of Sale Data

Video Cameras

Audio Devices

Oil Rigs

Data in motion

Data is typically not stored

Tremendous velocity

Multiple data sources

Huge volumes of unstructured data

Ultra low latency required

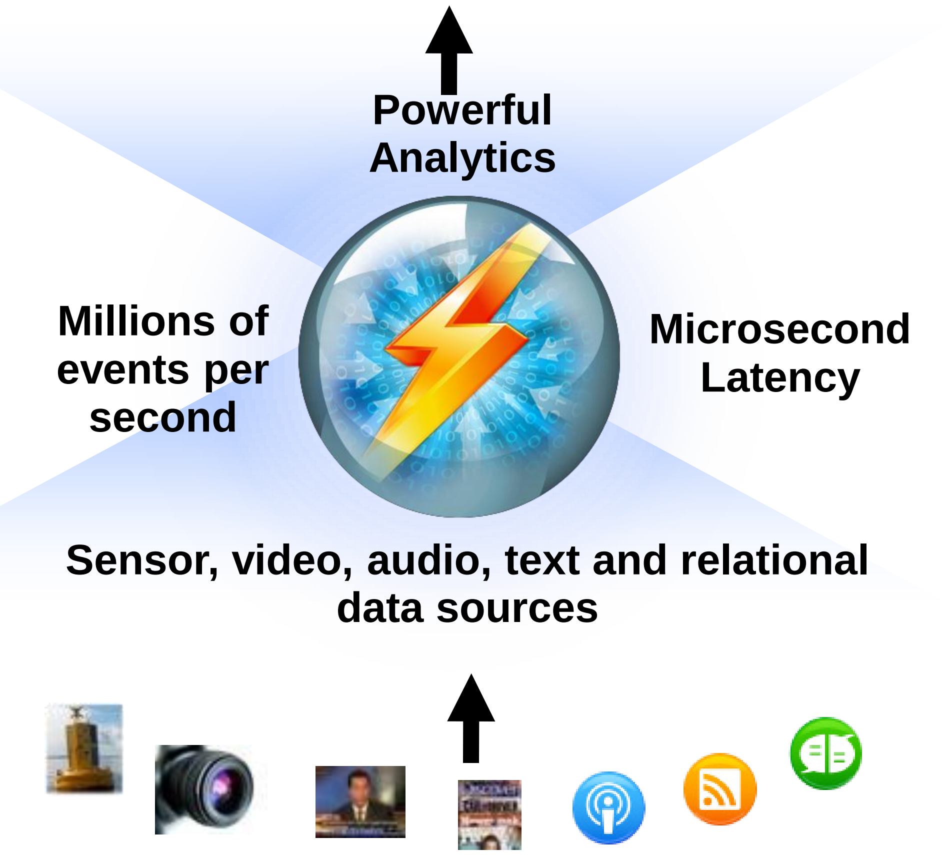
What Is InfoSphere Streams?

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

Just in time decisions



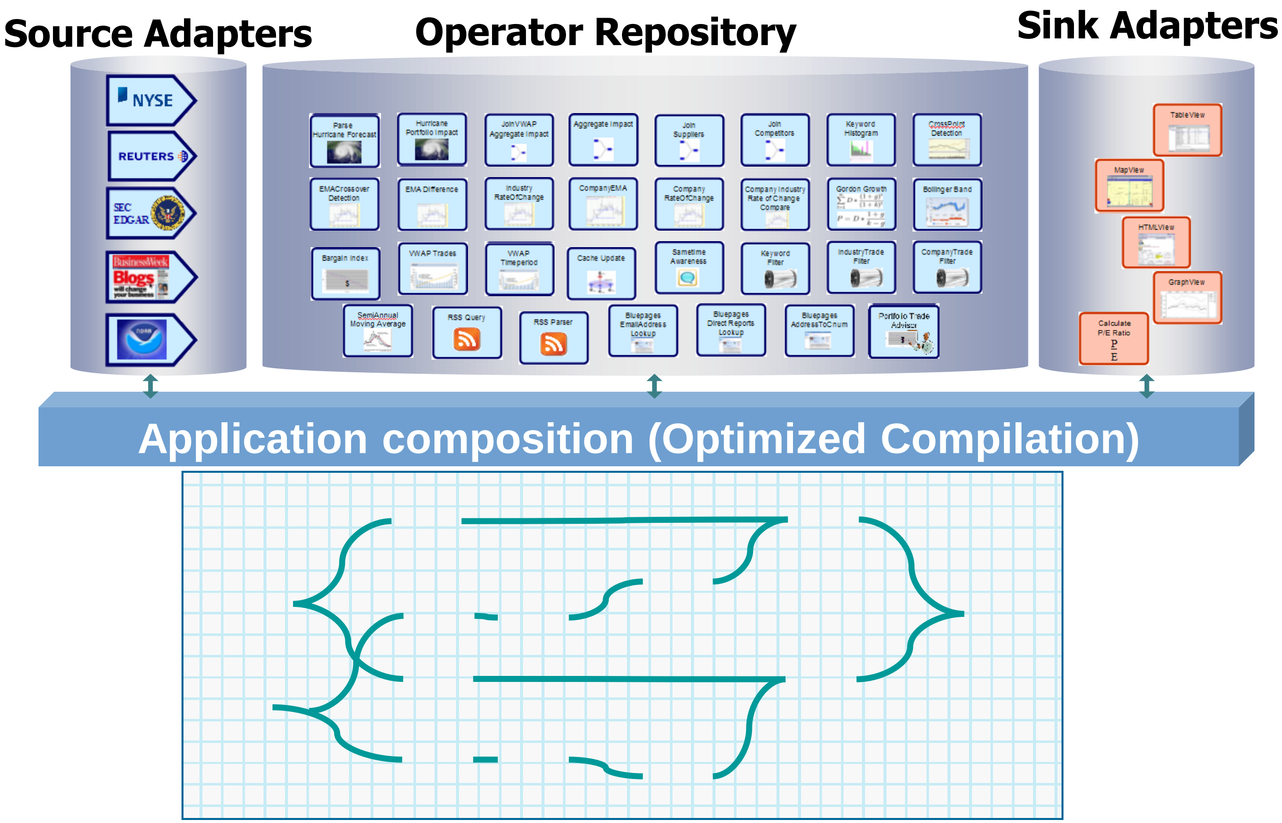
Powerful Analytics

Millions of events per second

Microsecond Latency

Sensor, video, audio, text and relational data sources

Streams Programming Model Uses the Streams Programming Language - SPL



Streams Versus Oracle NoSQL Productivity Study

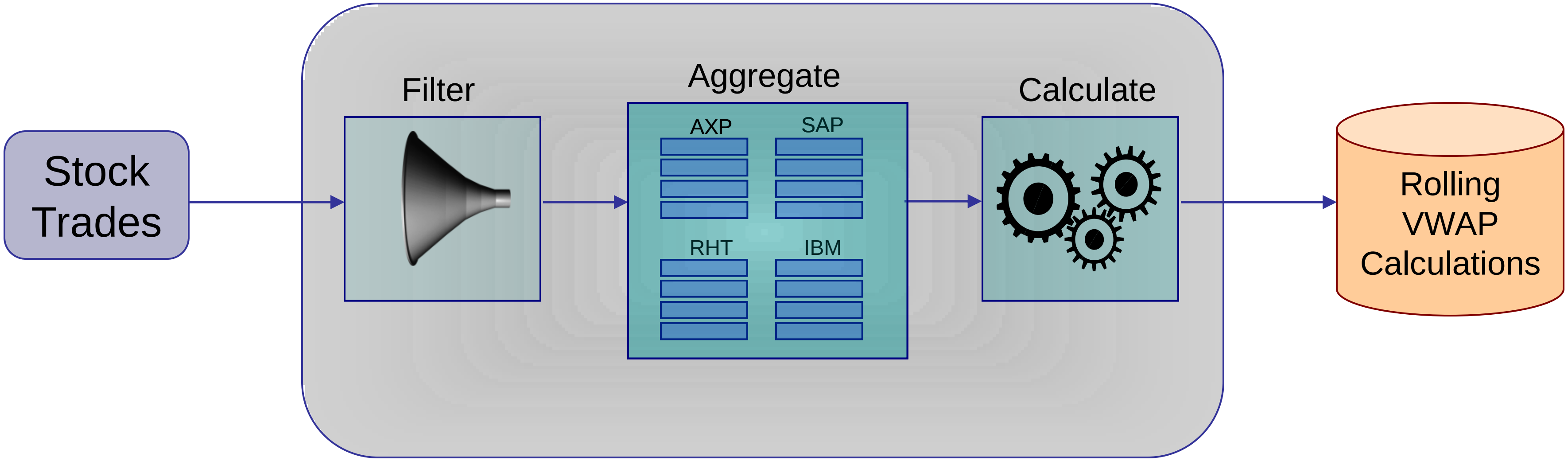
InfoSphere Streams

VS.

Oracle NoSQL Database

Use Case: Stock Trade Processing

Rolling Volume Weighted Average Price Calculation



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

PureData System For Operational Analytics

A Complete Solution For Structured Data

PureData System for Operational Analytics

Optimized for a mix of interactive and analytic queries

- ***Built-in expertise***
- ***Integration by design***
- ***Simplified experience***

Based on Power Systems

- **Simplicity**
  - Automatic, policy-based data placement and workload management
  - Integrated management and support
- **Speed**
  - Handles 1000+ concurrent operational queries[1]
  - Continuous ingest of operational data
  - MPP analytics (Massively Parallel Processing)
- **Scalability**
  - Available in multiple sizes with up to a Petabyte of data capacity[2]
- **Smart**
  - In-database analytics for leading applications
  - Supports DB2 applications unchanged and Oracle Database apps with minimal change
  - Clients have experienced cases of 10x storage space savings via Adaptive Compression[3]

1. Based on internal tests of prior generation system,, and on system Design for normal operation under expected typical workload . Individual Results may vary.
2. Total raw data capacity based on 1 XLarge configuration with five full rack data expansion add-ons.
3. Based on client testing in the DB2 10 Early Access Program.

PureData System For Operational Analytics

- 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

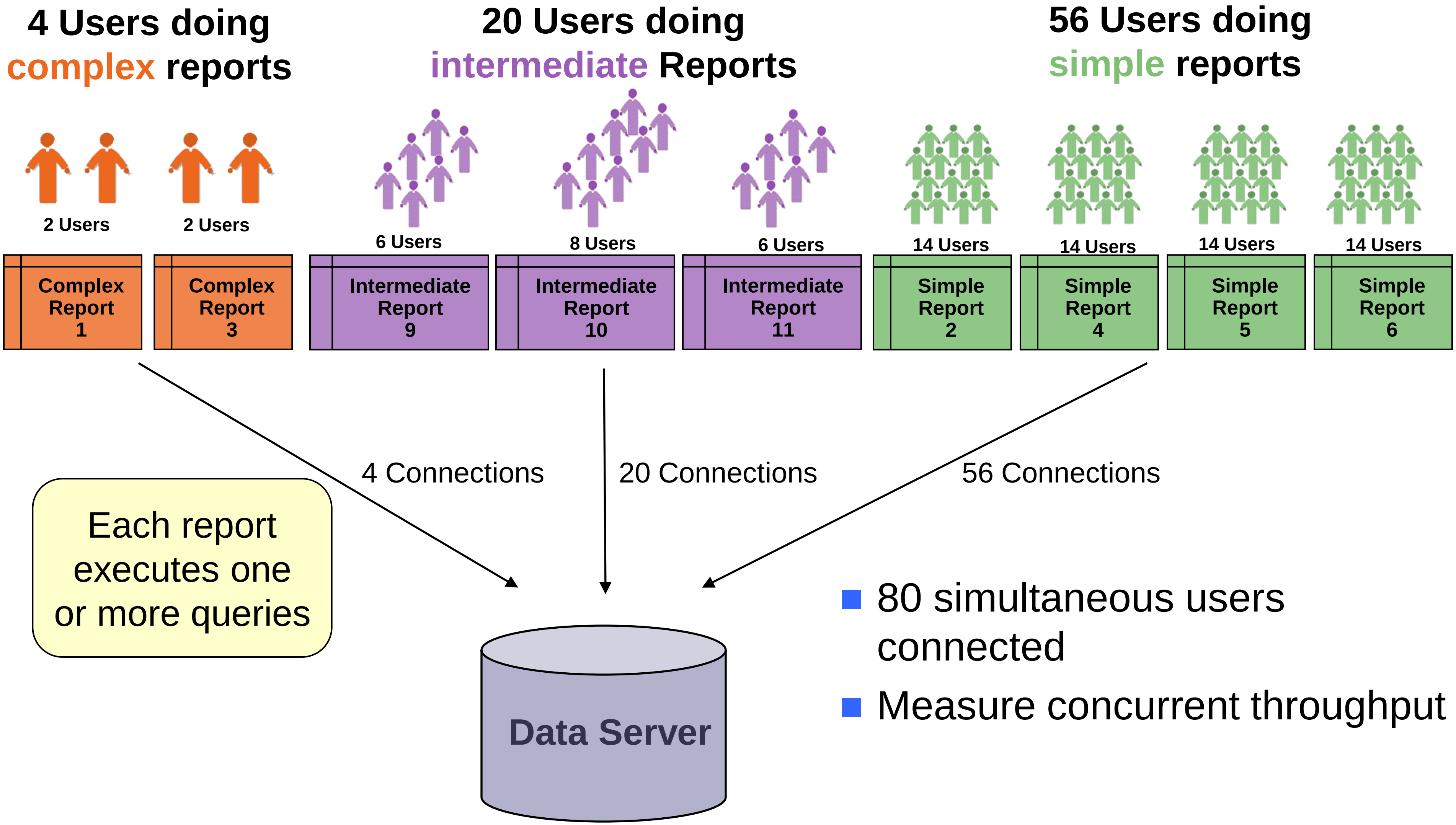
- IBM POWER7 P740 & P730 16 Core servers @ 3.55GHz

- 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

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



4 Users doing complex reports

20 Users doing intermediate Reports

56 Users doing simple reports

2 Users	2 Users	6 Users	8 Users	6 Users	14 Users	14 Users	14 Users	14 Users
Complex Report 1	**Complex Report 3**	**Intermediate Report 9**	**Intermediate Report 10**	**Intermediate Report 11**	**Simple Report 2**	**Simple Report 4**	**Simple Report 5**	**Simple Report 6**

4 Connections

20 Connections

56 Connections

Each report executes one or more queries

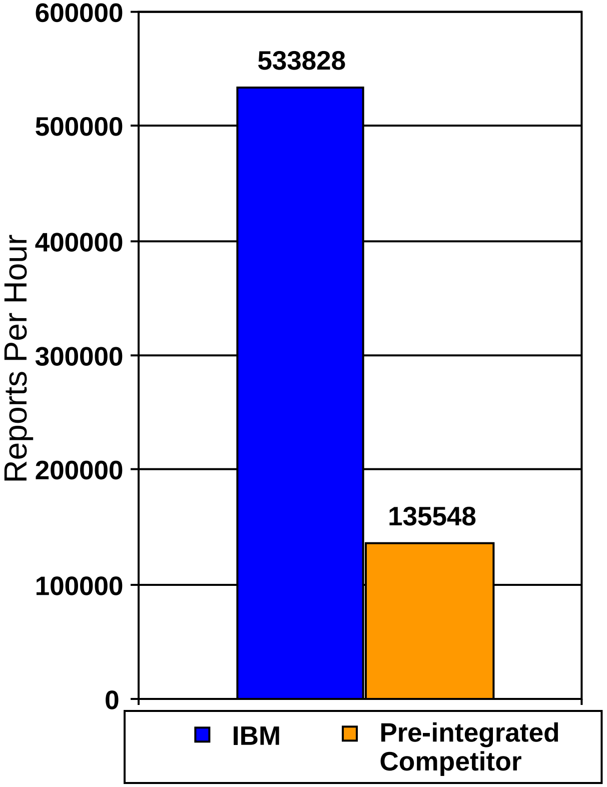
Data Server

- 80 simultaneous users connected
- Measure concurrent throughput

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

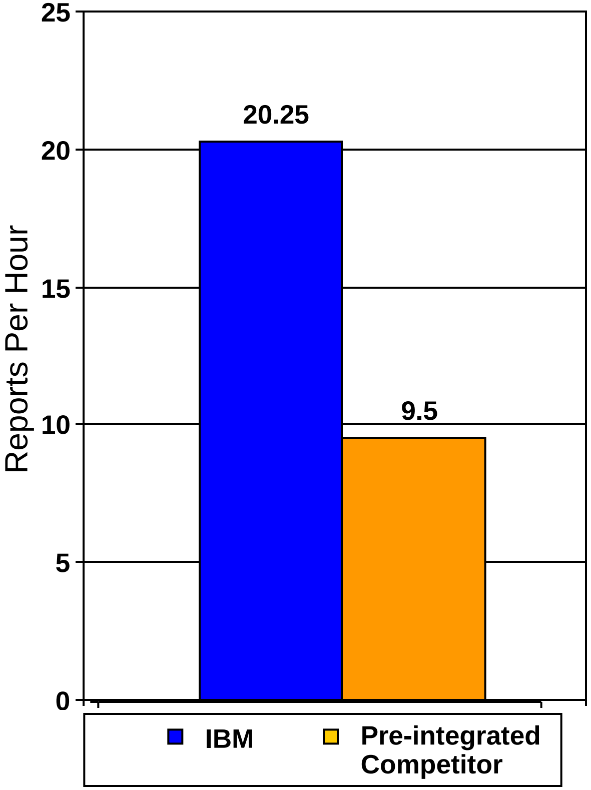
3.9X More
Simple Reports



Reports Per Hour
at 10 TB data size

(Higher is Better)

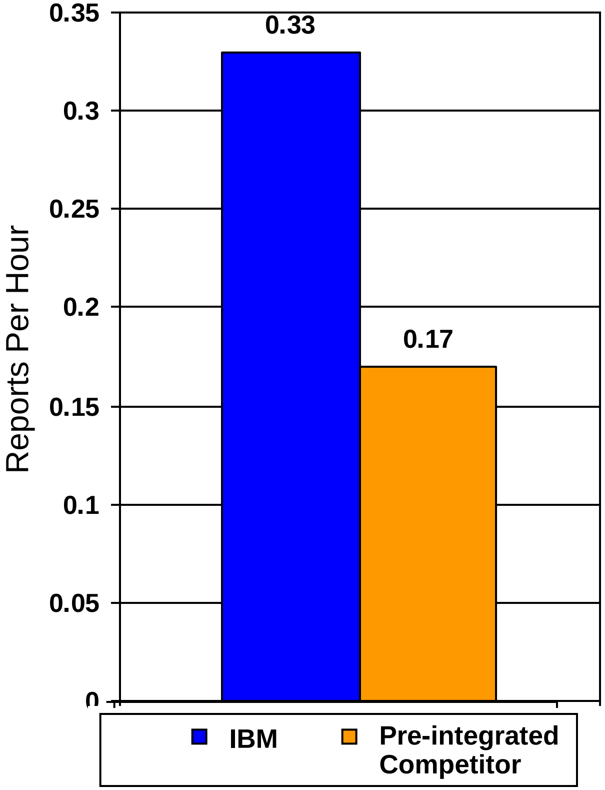
2X More
Intermediate Reports



Reports Per Hour
at 10 TB data size

(Higher is Better)

1.9x More
Complex Reports



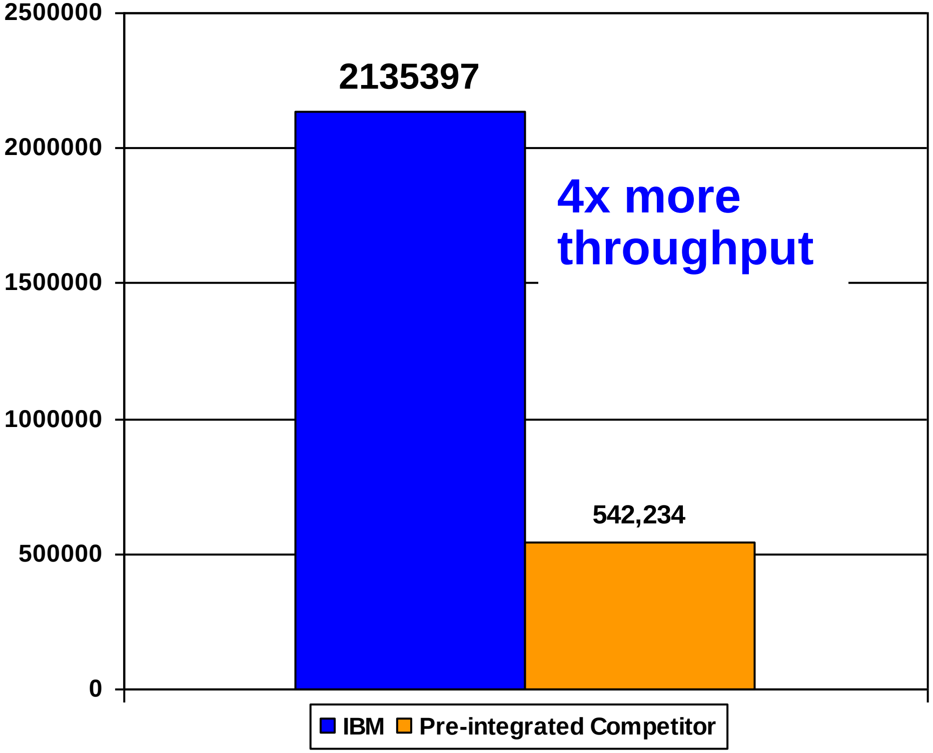
Reports Per Hour
at 10 TB data size

(Higher is Better)

Performance numbers may vary based on workload profiles.

IBM Operational Analytics Delivers More Throughput For Concurrent Operational Reports

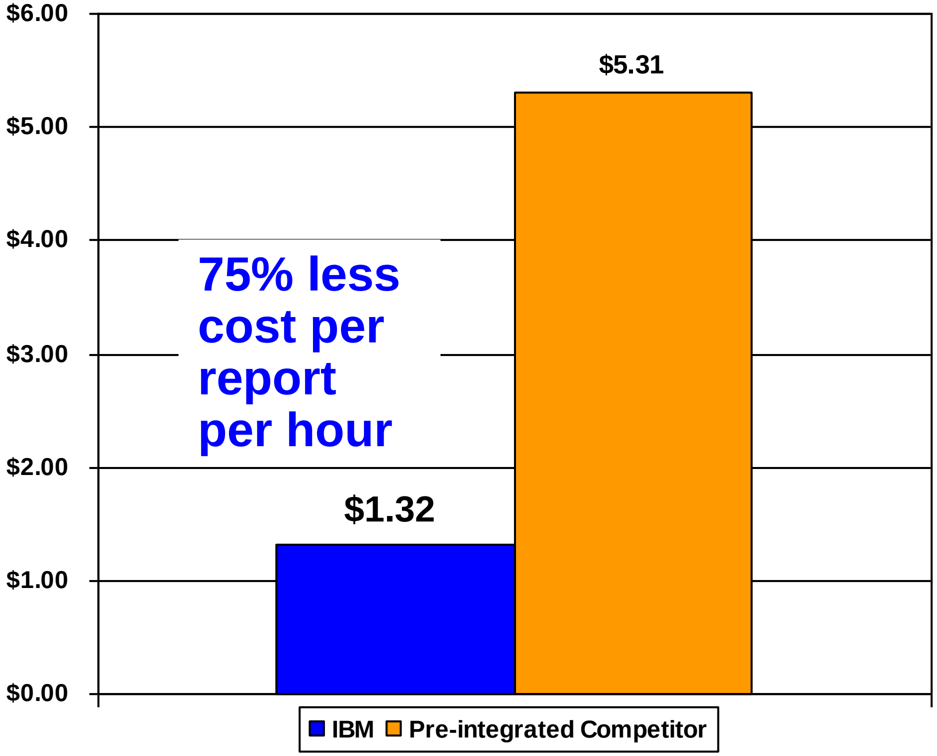
Total Report Throughput at 10 TB (Reports per hour)



4x more throughput

(Higher Throughput is Better)

Cost Per Report at 10 TB



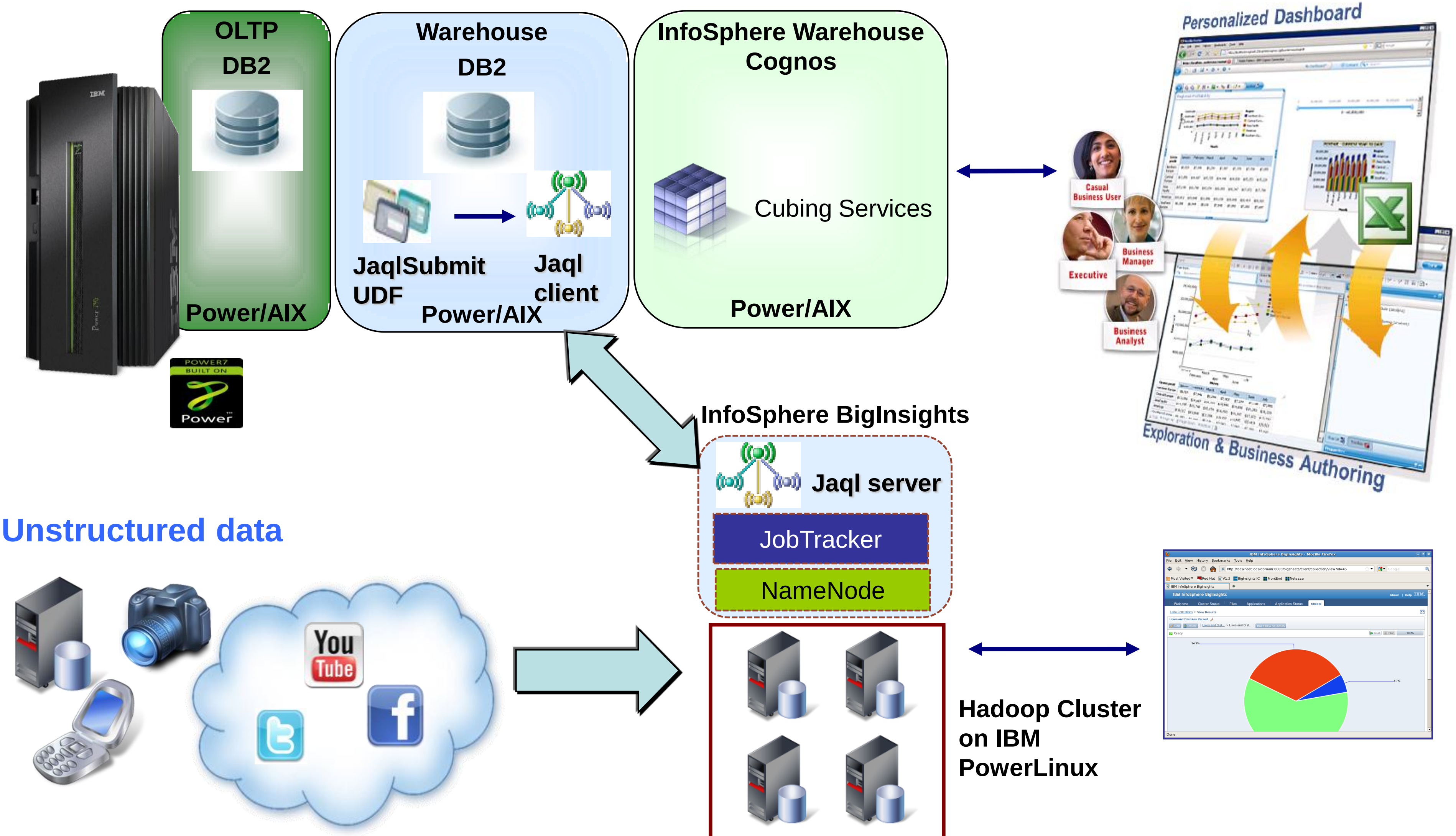
75% less cost per report per hour

(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



OLTP
DB2
Power/AIX

Warehouse
DB2
JaqlSubmit UDF
Jaql client
Power/AIX

InfoSphere Warehouse
Cognos
Cubing Services
Power/AIX

InfoSphere BigInsights
Jaql server
JobTracker
NameNode

Hadoop Cluster on IBM PowerLinux

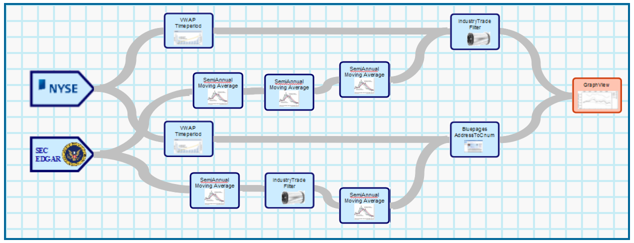
Unstructured data

IBM Can Help You Solve Big Data Problems

InfoSphere BigInsights

Data at Rest

InfoSphere Streams


NYSE
SEC EDGAR
VWAP Timeperiod
SemiAnnual Moving Average
SemiAnnual Moving Average
SemiAnnual Moving Average
IndustryTrade Filter
VWAP Timeperiod
Bluepages AddressToChurn
SemiAnnual Moving Average
IndustryTrade Filter
SemiAnnual Moving Average
GraphView

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