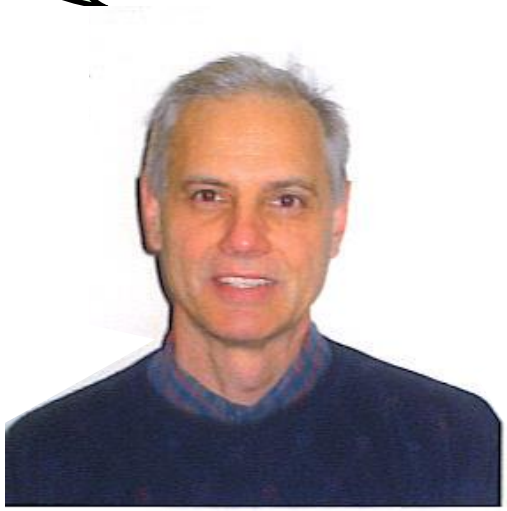
Harnessing and Capitalizing on New Sources of Big Data

It’s More Than Just Hadoop!

Big Data is More than Just Hadoop

What can you tell me about Big Data?

I want to know all about Hadoop.



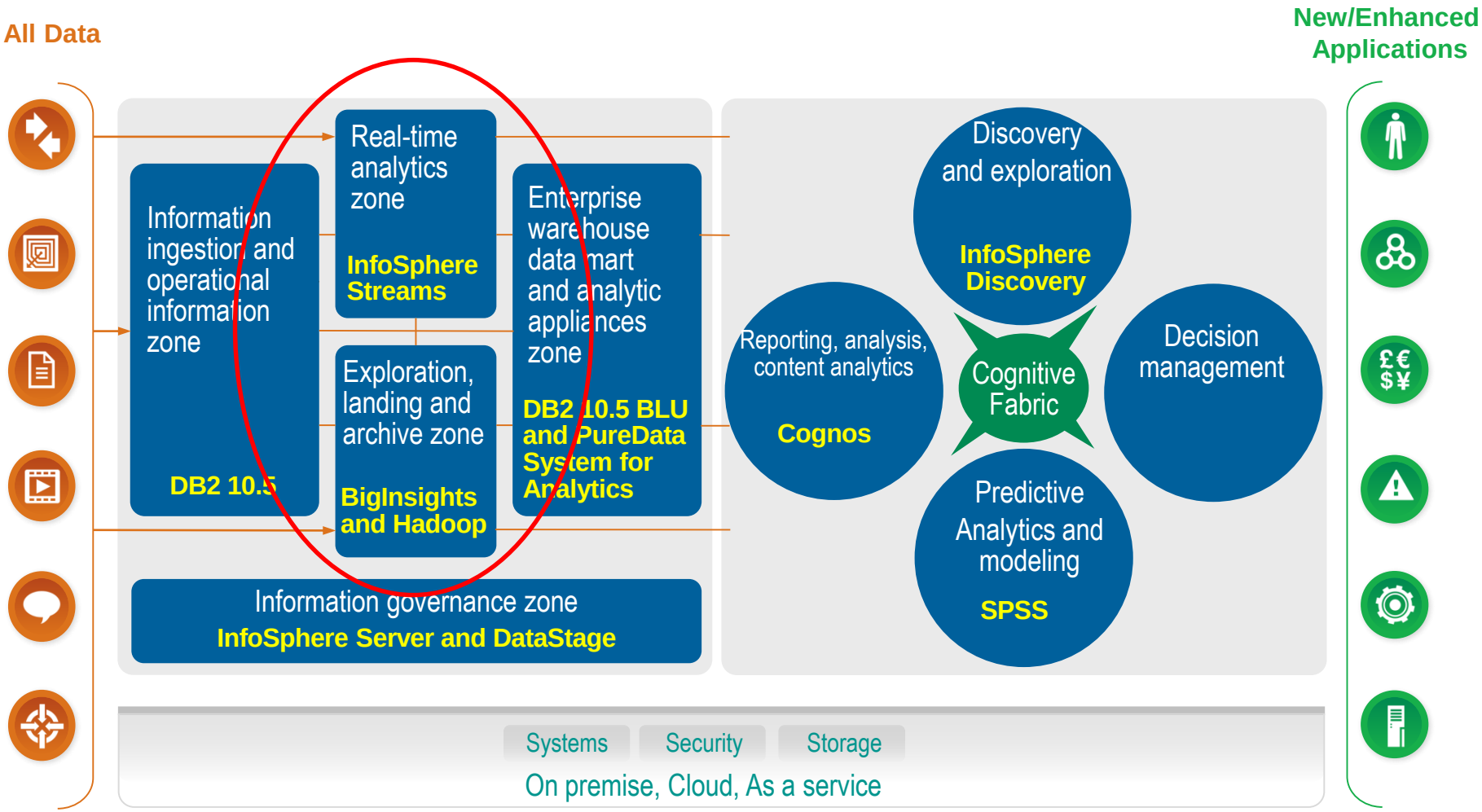
Service Oriented Finance CMO

Big Data is a lot more than Hadoop!

And our competitors don't understand this – they cannot deliver value on an entire set of Big Data use cases.

IBM

IBM Big Data and Analytics Platform



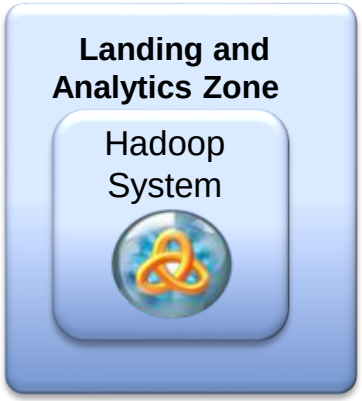
Analyze all data, from any source, with the right technology

There are Two Main Types of Big Data

Data in motion

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

Our competitors do not address both of these!

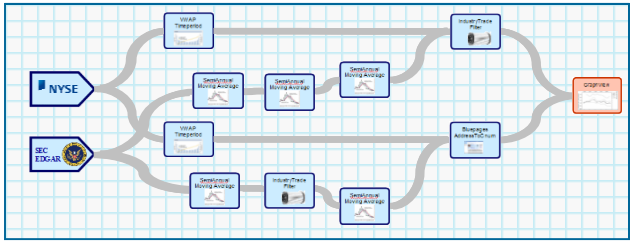


Data at rest

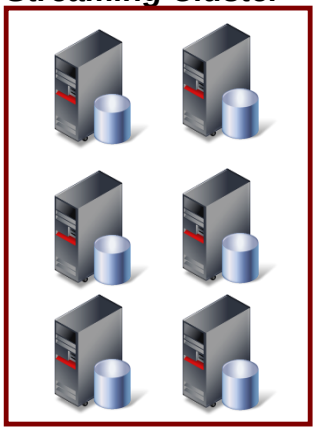
- Data stored on disk
- Huge volumes of unstructured data
- No pre-defined schemas
- Too large for traditional tools to process in a timely manner

New Programming Models and Low Cost Hardware Solve Big Data Problems

Streaming Application



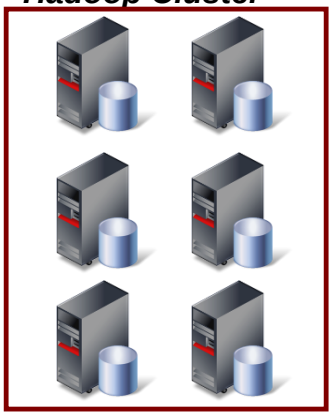
Streaming Cluster



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

Clusters of low cost Power8 servers are ideal for Streaming and Hadoop applications

Hadoop Cluster



Gaining Value from Data in Motion

Use Case	Analysis	Business Value
Real Time Marketing 	***Monitoring current events, cultural happenings or real time customer activities***	*Marketing effectiveness, customer satisfaction, customer retention*
Next Best Action 	***Monitoring customer interests, desires and needs*** *+ organization's objectives and offerings*	*Provide targeted and relevant offers to help reduce churn, increase sales*
Upsell opportunity 	***Monitoring point of Sale*** *data in real time + relational data about customers*	*Maximize up-sell opportunities for products in context of current purchases*

Service Oriented Finance Wants to Gain a Competitive Advantage from Big Data

This application will give our market managers a real advantage!

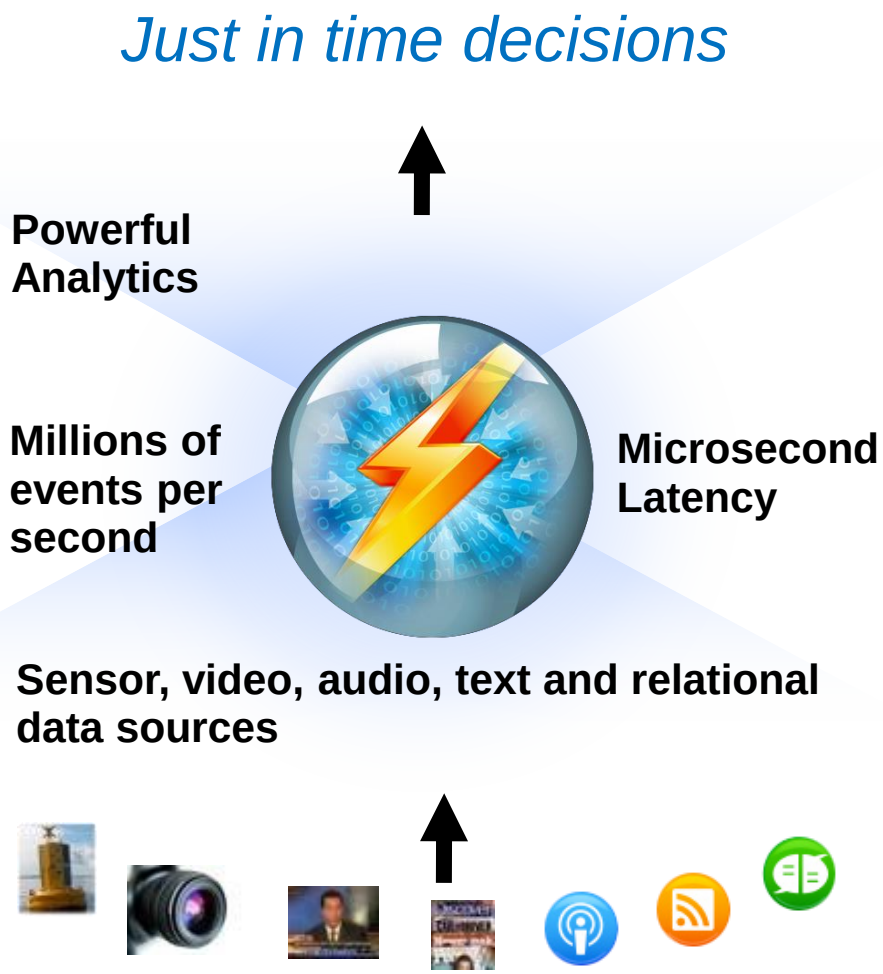
Service Oriented Finance Market Manager

Service Oriented Finance wants to deploy a stock trading application with the following requirements

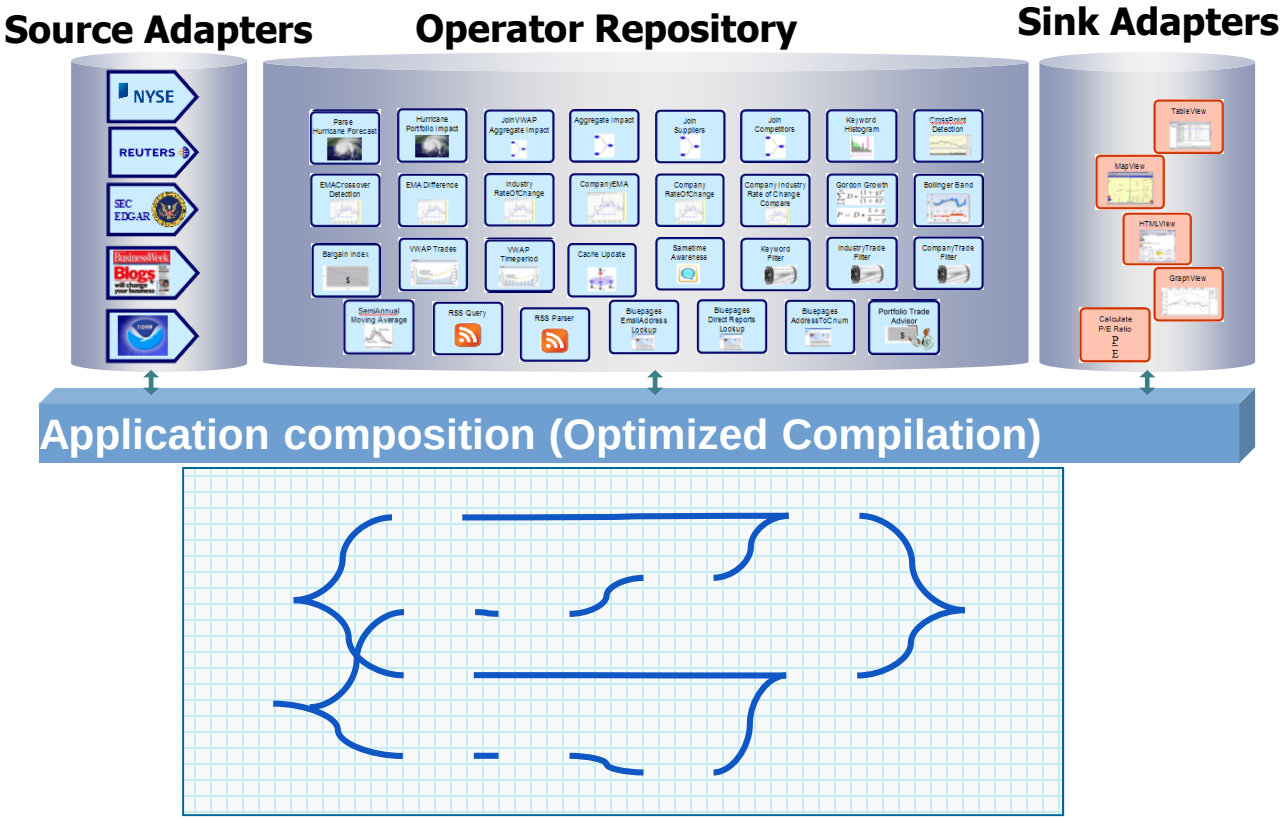
- Process millions of trades per second
  - Application must scale
- Constant flow of input data
- Microsecond latency
- Unstructured trade data input
- Sophisticated analytics logic

InfoSphere Streams is a Platform for Real Time Analytics

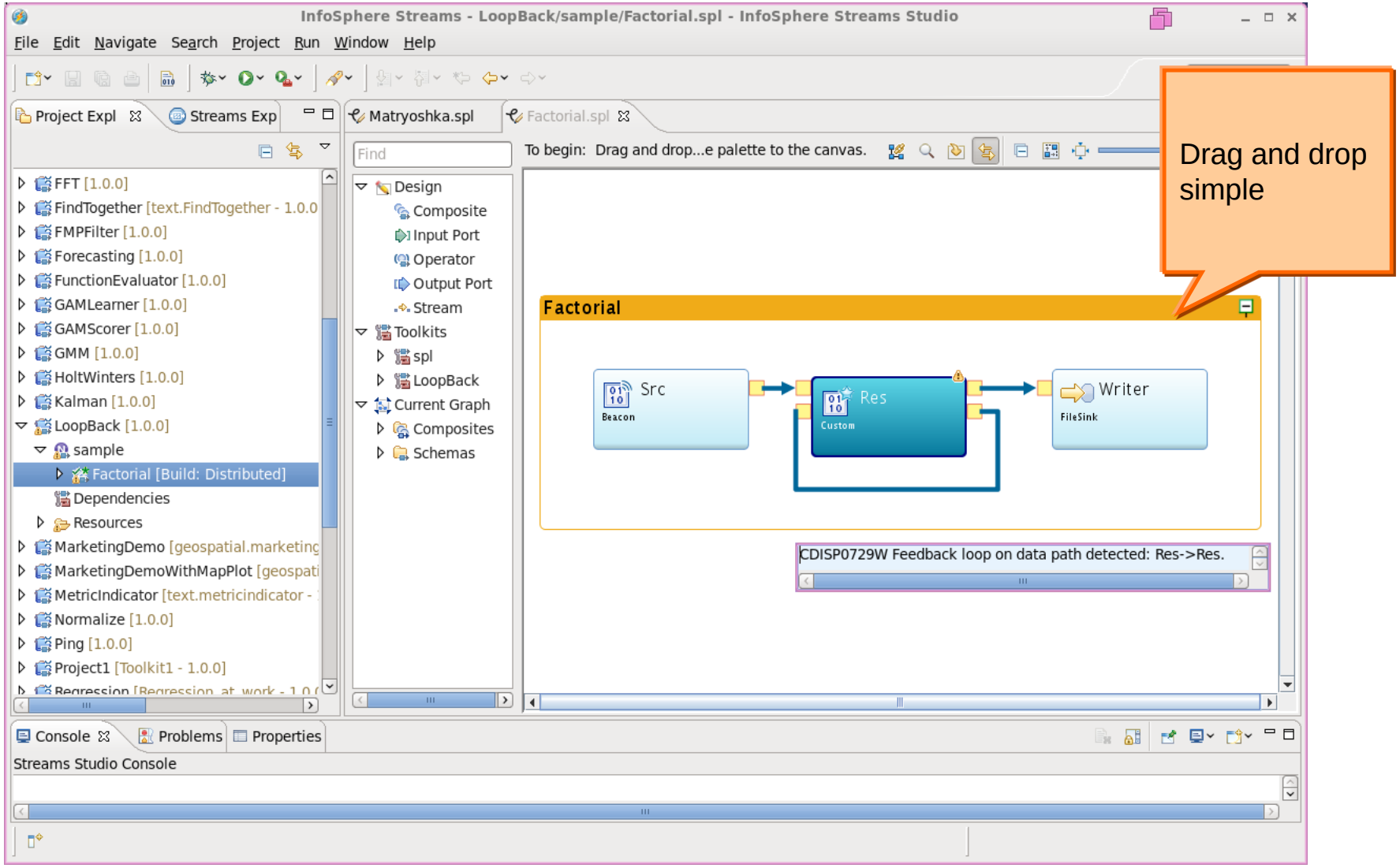
- Built to analyze data in motion
  - Multiple concurrent input streams
  - Massively scalable
- Process and analyze a variety of data
  - Structured, unstructured, video, audio, network logs
  - Advanced analytics operators built in
- Productive tools from development to deployment
  - Eclipse based development
  - Advanced visualization



Streams Programming is Drag and Drop Simple



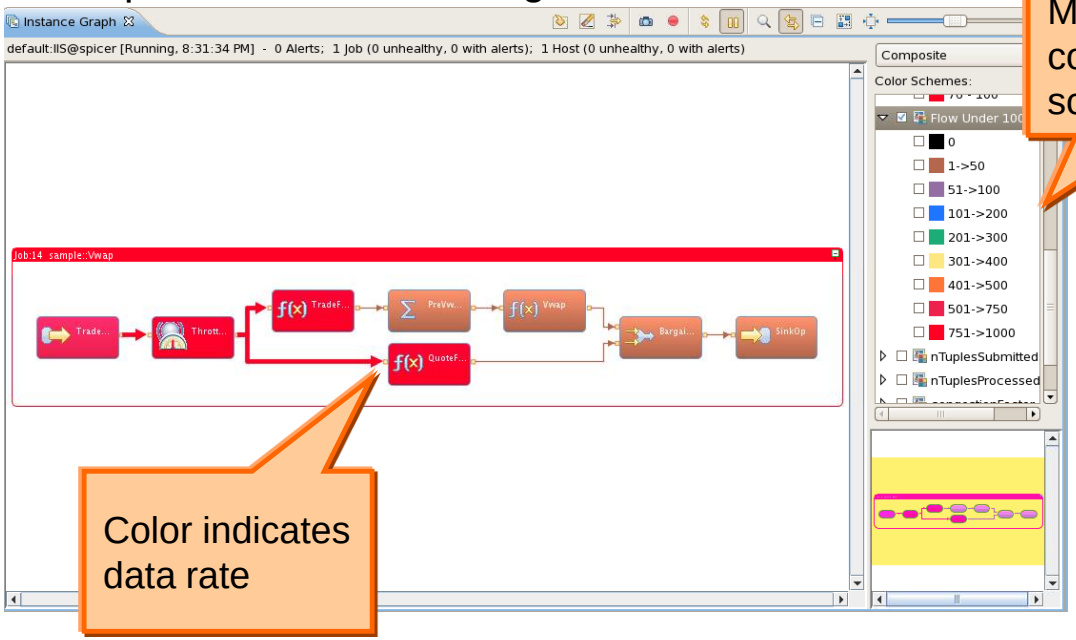
Streams Studio Provides a Rich Set of Eclipse Based Tools



Drag and drop simple

Visual Application Monitoring Provides a Clear View of Your Running Applications

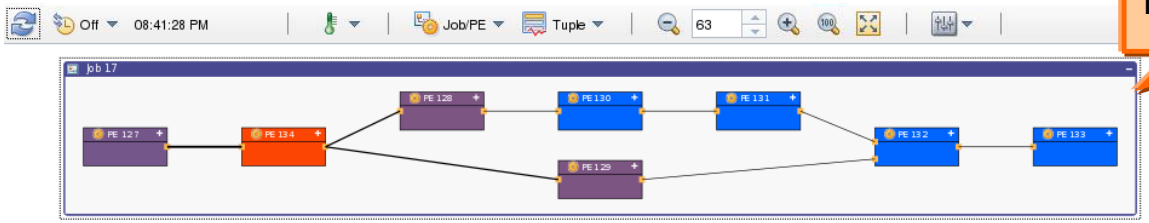
Development time monitoring with Streams Studio



Metric-based coloring schemes

Color indicates data rate

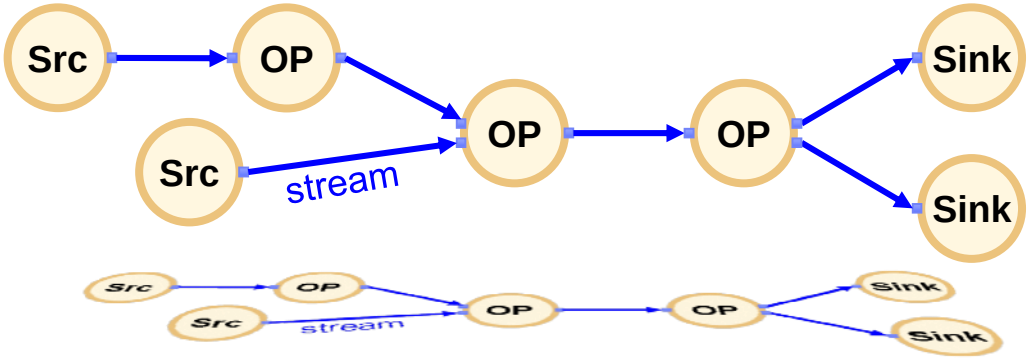
Production monitoring from the Streams Console



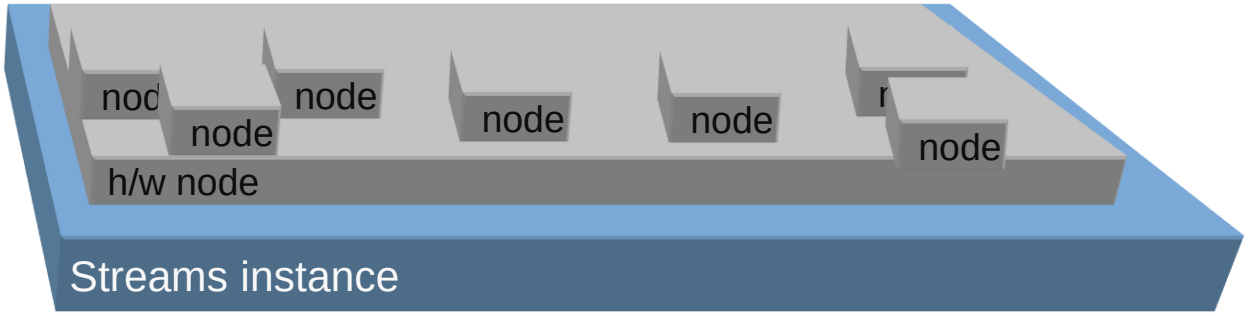
Visually monitor application health

Streams Jobs Get Deployed to a Single Node or a Cluster of Nodes

- Streams job
  - A collection of operators
  - Connected by streams



- Jobs are deployed to a Streams runtime environment, known as a Streams Instance (or simply, an instance)
- An instance can include a single processing node (hardware)
- Or multiple processing nodes



Every Industry can Leverage Real Time Analytics

Banking
- Optimizing Offers and Cross-sell
- Customer Service and Call Center Efficiency
- Fraud Detection & Investigation
- Credit & Counterparty Risk

Insurance
- 360˚ View of Domain or Subject
- Catastrophe Modeling
- Fraud & Abuse
- Producer Performance Analytics
- Analytics Sandbox

Telco
- Pro-active Call Center
- Network Analytics
- Location Based Services

Energy & Utilities
- Smart Meter Analytics
- Distribution Load Forecasting/Scheduling
- Condition Based Maintenance
- Create & Target Customer Offerings

Media & Entertainment
- Business process transformation
- Audience & Marketing Optimization
- Multi-Channel Enablement
- Digital commerce optimization

Retail
- Actionable Customer Insight
- Merchandise Optimization
- Dynamic Pricing

Travel & Transport
- Customer Analytics & Loyalty Marketing
- Predictive Maintenance Analytics
- Capacity & Pricing Optimization

Consumer Products
- Shelf Availability
- Promotional Spend Optimization
- Merchandising Compliance
- Promotion Exceptions & Alerts

Government
- Civilian Services
- Defense & Intelligence
- Tax & Treasury Services

Healthcare
- Measure & Act on Population Health Outcomes
- Engage Consumers in their Healthcare

Automotive
- Advanced Condition Monitoring
- Data Warehouse Optimization
- Actionable Customer Intelligence

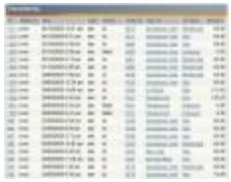
Chemical & Petroleum
- Operational Surveillance, Analysis & Optimization
- Data Warehouse Consolidation, Integration & Augmentation
- Big Data Exploration for Interdisciplinary Collaboration

Aerospace & Defense
- Uniform Information Access Platform
- Data Warehouse Optimization
- Airliner Certification Platform
- Advanced Condition Monitoring (ACM)

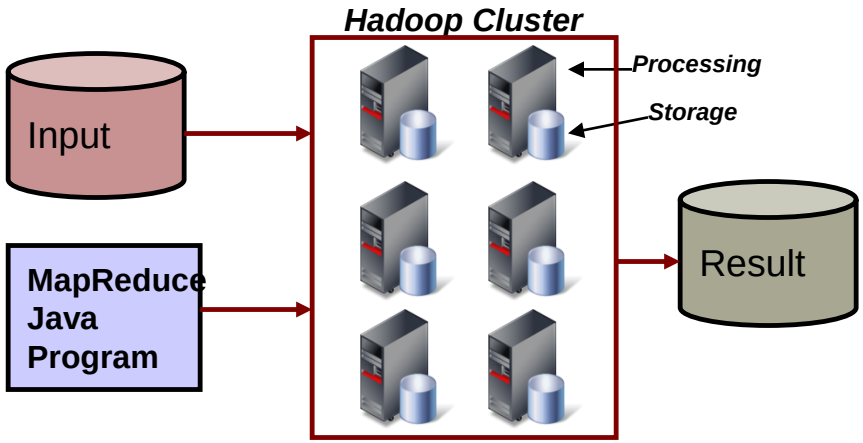
Electronics
- Customer/ Channel Analytics
- Advanced Condition Monitoring

Life Sciences
- Increase visibility into drug safety and effectiveness

Gaining Value from Data at Rest

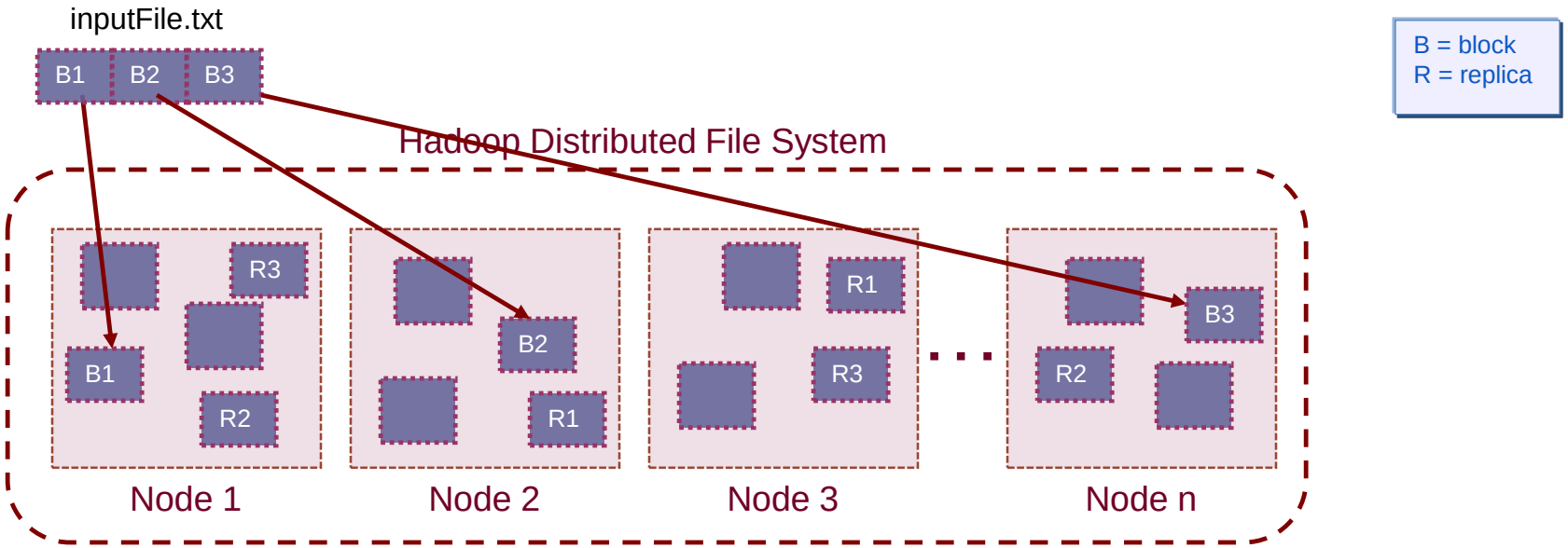
Data Source	Analysis	Business Value
Web Logs 	*Analyze online shopper behavior from e-commerce site*	*Maximize retail web site sales*
Social Media	*Analyze customer sentiment and experience*	*Attract and retain customers*
Weather Data 	*Analyze vast amounts of historical weather data*	*Determine optimal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

Process Data at Rest using Hadoop: InfoSphere BigInsights



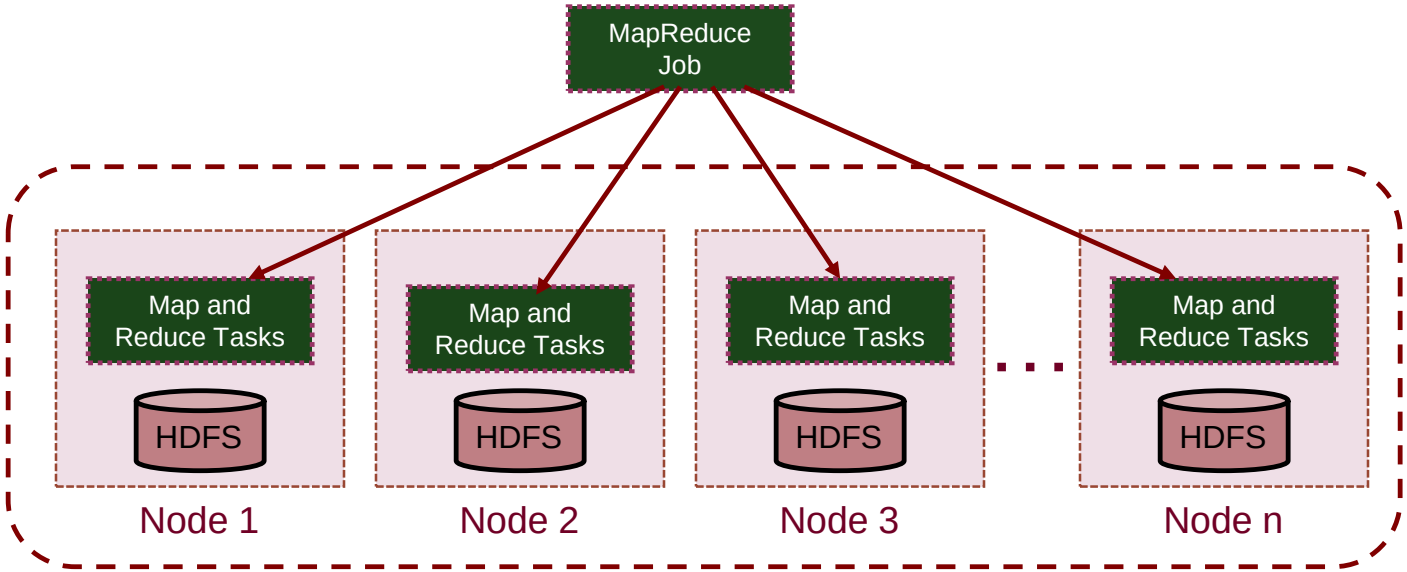
- Comprised of a cluster of inexpensive hardware
  - Nodes have processors, memory and disks
- Special file system – Hadoop Distributed File System (HDFS)
- Special programming model – MapReduce

The Hadoop Distributed File System (HDFS) Distributes Data Across a Hadoop Cluster



- 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
- System assumes nodes will fail
  - Achieves reliability by replicating data across multiple nodes
- Elastically scalable

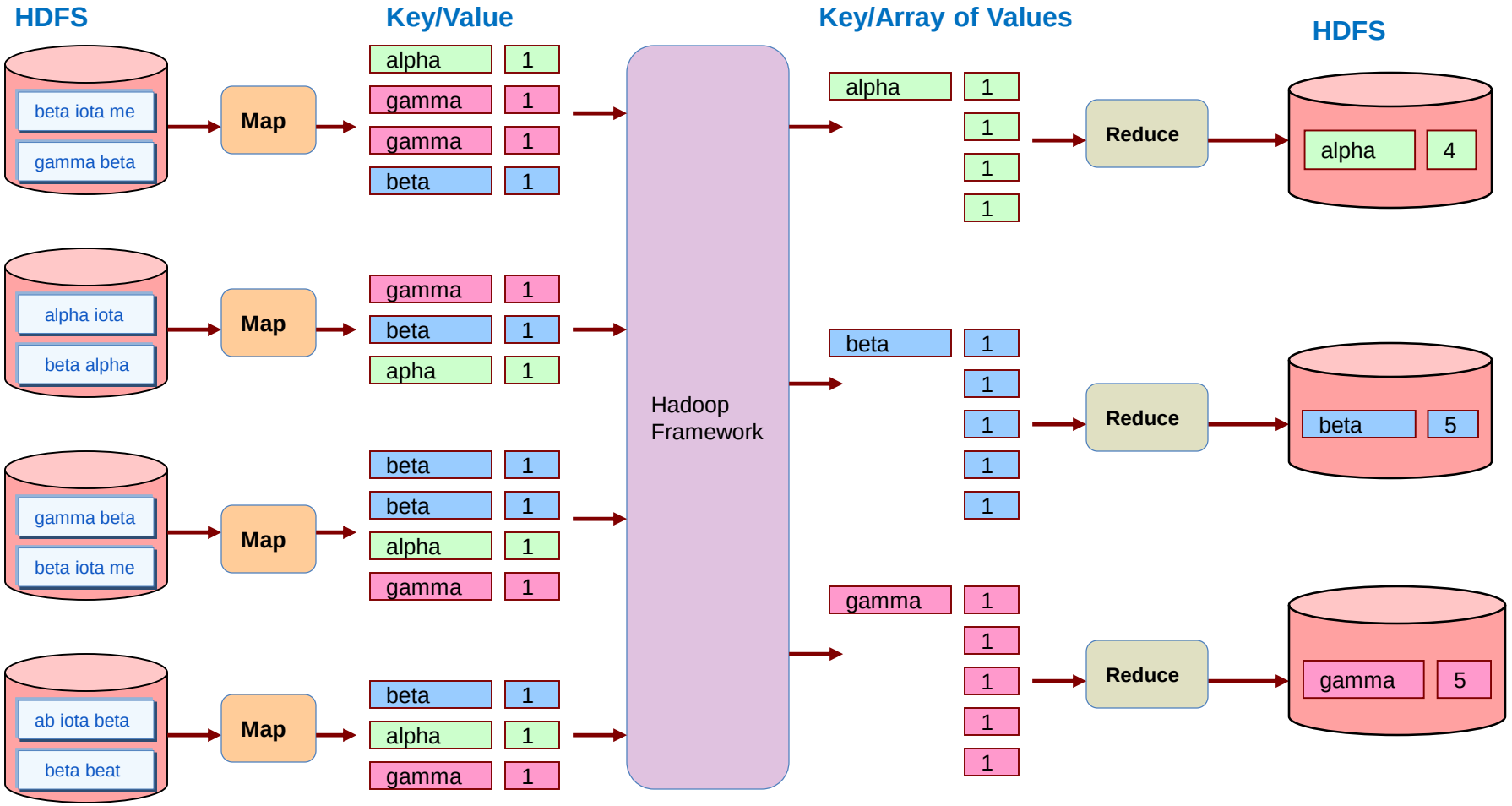
The MapReduce Framework Sends Programs Out to the Data



- MapReduce job is sent out to each node
- Map and Reduce tasks run in parallel across nodes
- Hadoop framework does a lot of the “heavy lifting”
  - e.g., moving data between map and reduce tasks

Simple MapReduce Example of Counting Occurrences of Strings in Text

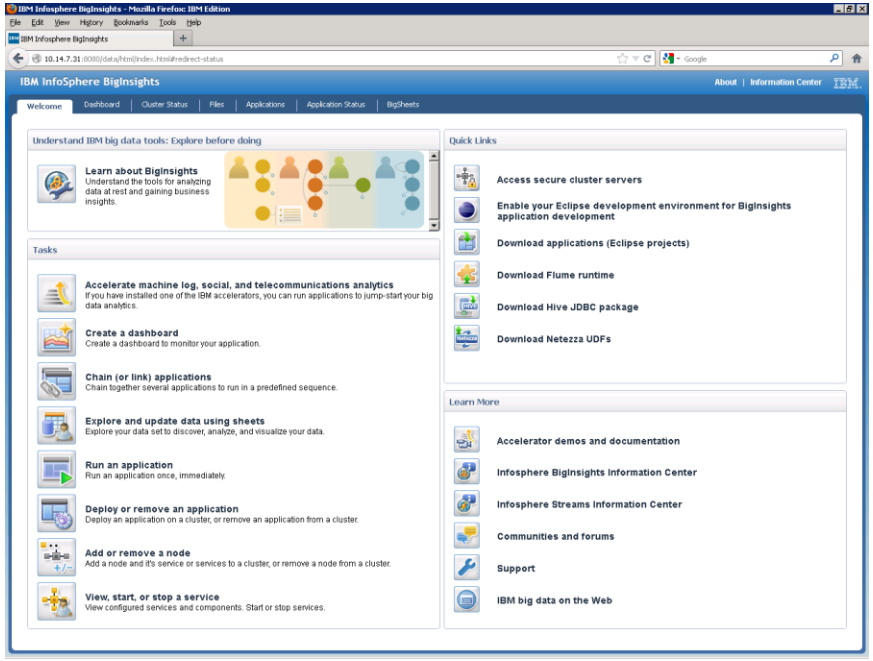
Goal: Count the number of occurrences of alpha, beta, and gamma in a text file.



BigInsights Makes it Easy for All Big Data Roles

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

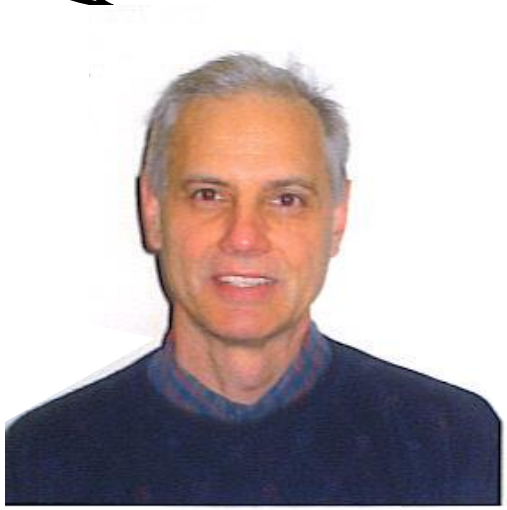
InfoSphere BigInsights Console



Service Oriented Finance Wants to Analyze Customer Complaints

We need to know what our customers are complaining about.

We can help you do that with sentiment analysis using BigInsights



Service Oriented Finance CMO

IBM

Sentiment Analysis - A Big Data Challenge but Also a Big Data Opportunity

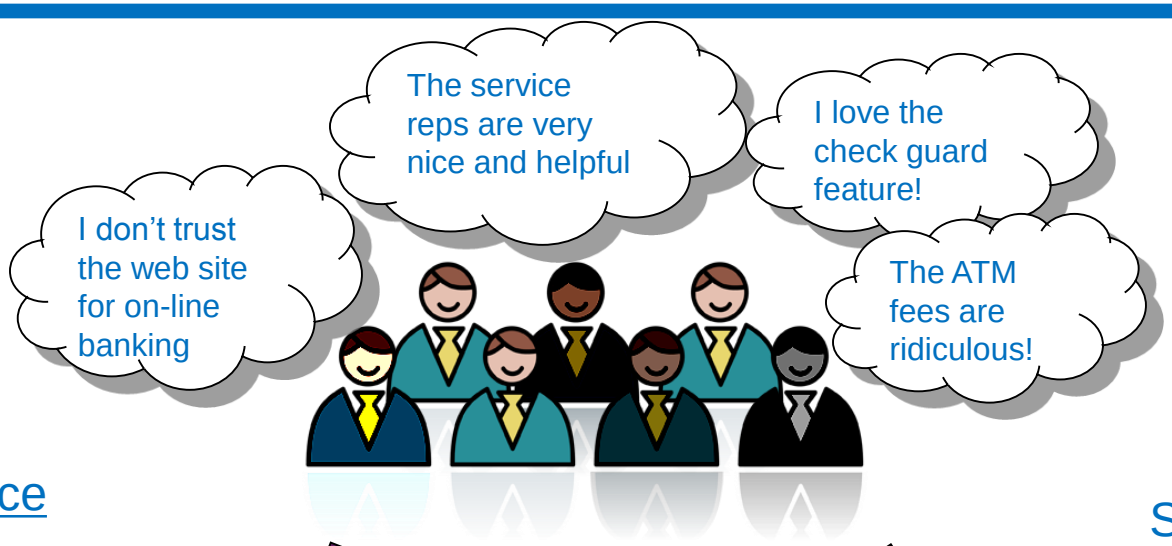
Huge volumes of unstructured data

Trying to determine…

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

Finding sentiment from social media data

DEMO: Using BigInsights to Analyze Negative Sentiment on Twitter



Data Source
Twitter

Topic
Service Oriented Finance

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

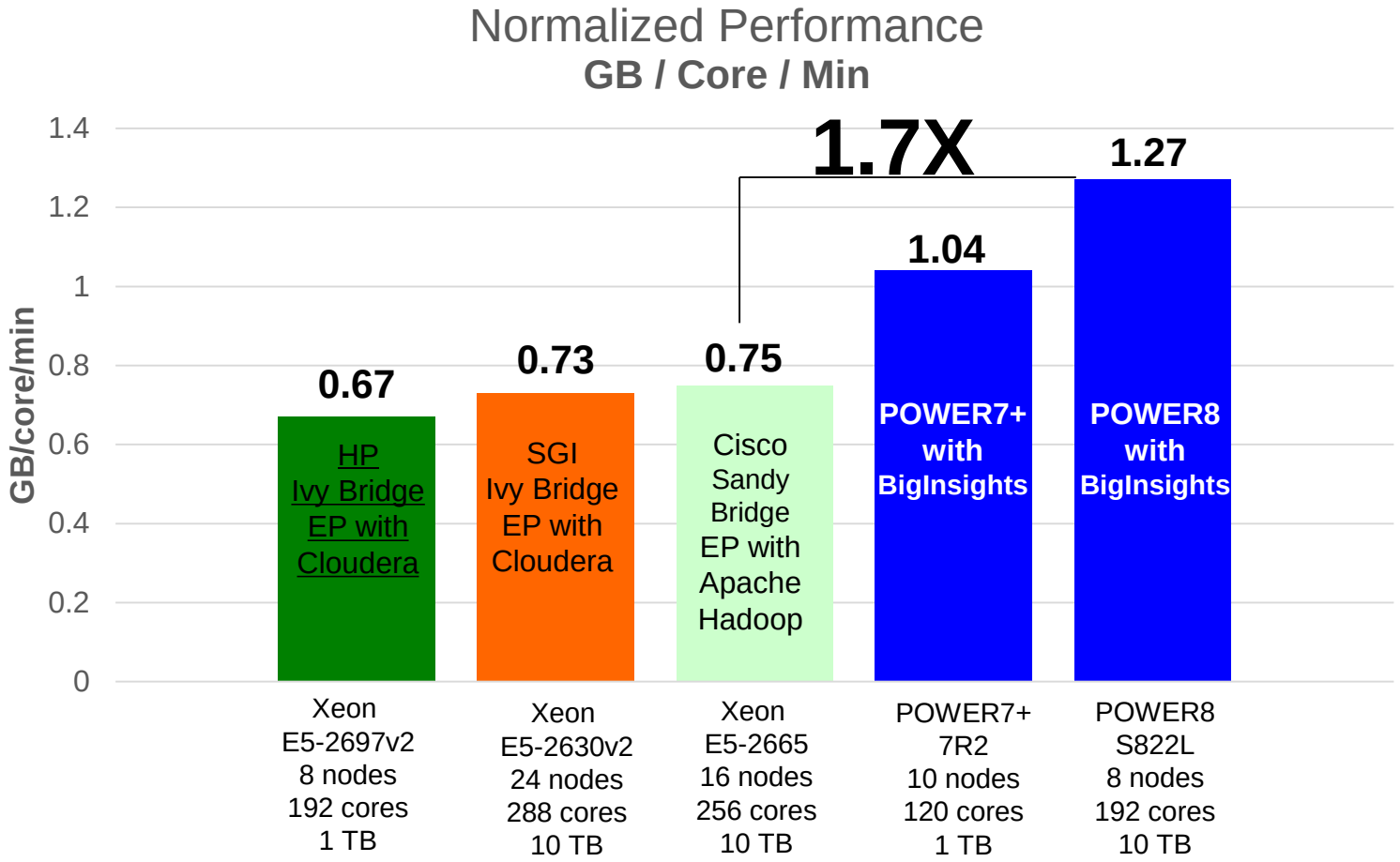
Dislikes

- 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

BigInsights has Capabilities Other Hadoop Distributions Lack

- Two powerful processing engines
  - Massively parallel batch processing with MapReduce
  - Fully ANSI compliant SQL engine with Big SQL
- Performance and Optimization
  - Adaptive MapReduce
  - Advanced Scheduler
  - BigIndex for large scale indexing
  - Fast, splittable compression
- Optim Development Studio
  - Eclipse based IDE for Java
- Big Data Integration
  - Information Server, InfoSphere Streams, Netezza, DB2
- Analytic Accelerators
  - BigSheets spreadsheet and visualization
  - Machine Data
  - Social Media
  - Advanced Text Analytics
  - JAQL query language

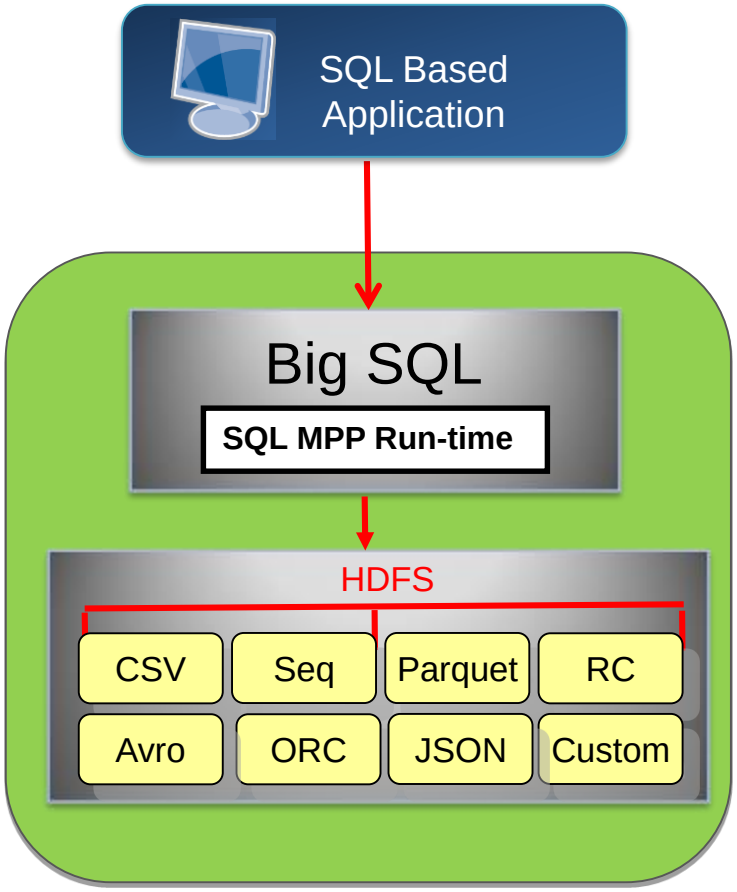
BigInsights on POWER beats the competition with TeraSort benchmark



Cisco Paper - http://www.cisco.com/c/dam/en/us/solutions/collateral/borderless-networks/advanced-services/le_tera.pdf
SGI Paper - http://www.sgi.com/pdfs/4440.pdf
HP Ivy Bridge with Cloudera was tested in the IBM laboritories
POWER7+ and POWER8 with BigInsights was tested in IBM laboratories

Big SQL V3.0: Bringing SQL on Hadoop to the Next Level

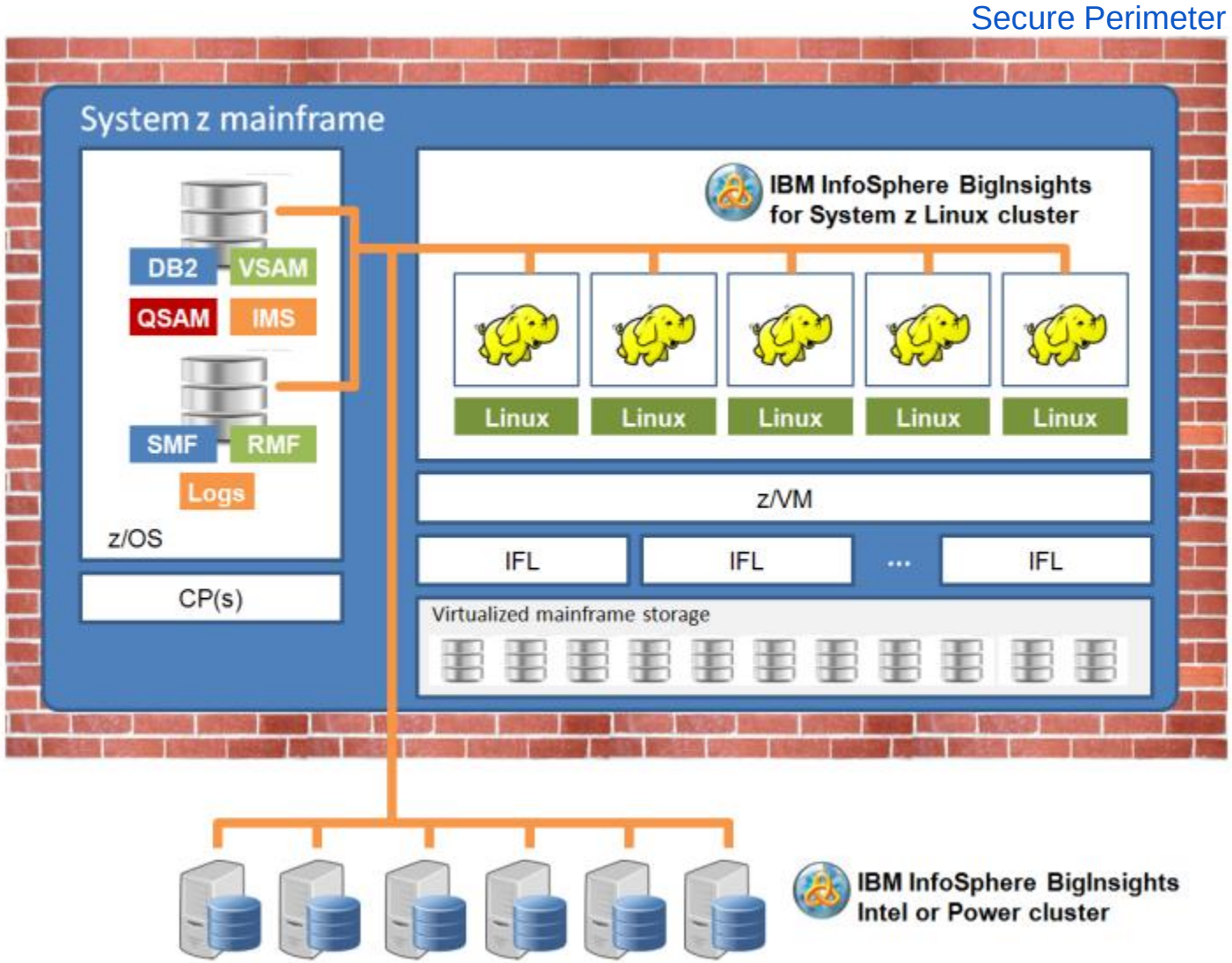
- Massively parallel SQL engine on Hadoop
  - Architected from the ground up for low latency and high throughput
- Comprehensive SQL support
  - The same SQL you use on your data warehouse should run with few or no modifications
  - Full support for sub-queries
  - All standard join operations
  - Stored procedures / User defined functions
- Supports all modern file formats



Big SQL V3.0: Why Do We Want to Use SQL on Hadoop Data?

- MapReduce programming is difficult
  - MapReduce Java API requires programming expertise
- Hadoop/MapReduce are new technologies
  - Expertise is in limited supply
- Unfamiliar languages (such as Pig) also require special skills
- SQL support opens the data to a much wider audience
  - Easy on-ramp to Hadoop for SQL professionals
- SQL support opens the data to the many SQL tools available
  - Cognos, JDBC, ODBC

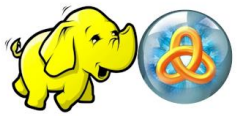
InfoSphere BigInsights for Linux on System z



IBM BigInsights Brings Unique Software Capabilities to Hadoop

Reduce time to market, increase customer value

Software Capabilities	Other Hadoop	BigInsights
Open Source Hadoop	✓	✓
Rich SQL on Hadoop – Big SQL	*Some capability*	✓
Tools for business users – BigSheets	-	✓
Advanced Text Analytics	-	✓
In-Hadoop Analytics	-	✓
Rich Developer tools	-	✓
Enterprise-grade workload & storage mgmt.	-	✓
Comprehensive suite	-	✓



“ IBM has the strongest strategy, most compelling roadmap.

-- Forrester Wave

Q1 2014

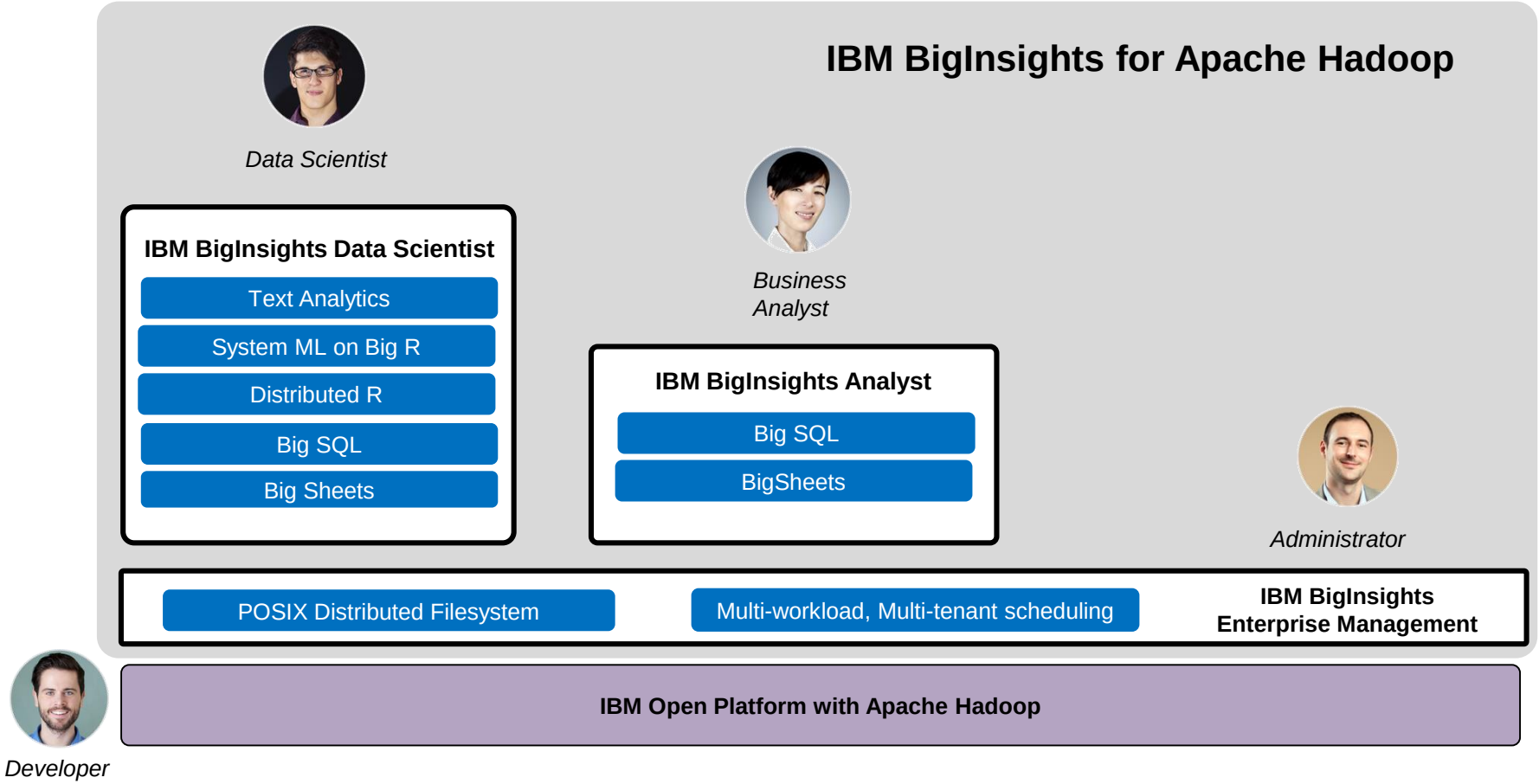
IBM is Founder Member in Open Data Platform Initiative

The Open Data Platform Initiative (ODP) is a shared industry effort focused on promoting and advancing the state of Apache Hadoop and Big Data technologies for the enterprise.

- **Test, Certify, and Standardize** the core components of a new "Open Data Platform" of select Apache Software Foundation (ASF) projects to provide a foundation for which Big Data solutions providers can build upon.
- Initially **Apache Hadoop** (HDFS, YARN, MapReduce) and **Apache Ambari** (Provisioning, Management, and Monitoring), **Spark**
- **Support f**or community development and outreach activities

IBM BigInsights for Apache Hadoop

Three new user-centric modules founded on an Open Data Platform



Streams is a Powerful Tool for High Velocity Real-Time Analytics

- Drag and drop simple development
- Extensive visualization capabilities
- Built-in integration tools
- Tools for all roles
  - Developer, administrator, business user

2.6X - 12.3X
More Throughput

5.5X – 14.2X
Less CPU Time

BigInsights Extends Hadoop Into an Enterprise Class Big Data Platform

- Built-in accelerators
- Built-in text analytics tools
- Built-in visualization tools
- Built-in integration tools
- Big SQL engine
- R language support
- Tools for all roles
  - Developer, administrator, business user

34X Faster
Standing up Cluster

2X Faster
Building Hadoop Applications

1.7X Faster
Running Terasort

Agenda for Istanbul

9:30 – 10:00	Leveraging Big Data to Deliver Immediate Value
10:00 – 10:45	Breakthrough Analytics Performance With BLU
10:45 – 11:15	Bringing Big Data and Analytics Together for Greater Insight
11:15 – 11:30	Break
11:30 – 12:15	Harnessing and Capitalizing on New Sources of Big Data
12:15 – 12:30	Efficiently Integrating All Your Data
12:30 – 12:50	Accelerating Time To Value with Smarter Deployment
12:50 – 13:00	Summary

Including comparisons with Oracle Exadata and SAP HANA