Information Management

IBM’s InfoSphere BigInsights: Smart Analytics for Big Data

Claus Samuelsen
csa@dk.ibm.com

November 7, 2011

IBM Disclaimer

Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Agenda

- **The “Big Data” challenge: smarter analytics for a smarter planet**
- **IBM’s approach**
  - The big picture
  - Details on BigInsights
  - How BigInsights fits in your software stack (with data warehouses, DBMSs, streams, etc.)
- **How IBM can help you get off to a quick start**

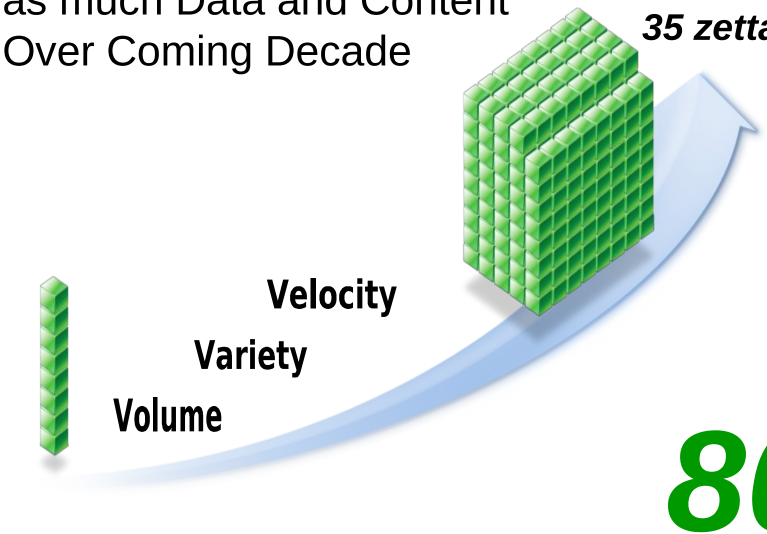
The “Big Data” Challenge

November 7, 2011

Information is at the Center of a New Wave of Opportunity…

44x

as much Data and Content Over Coming Decade



80%

Of world’s data is unstructured

… And Organizations Need Deeper Insights

1 in 3 Business leaders frequently make decisions based on information they don’t trust, or don’t have

1 in 2 Business leaders say they don’t have access to the information they need to do their jobs

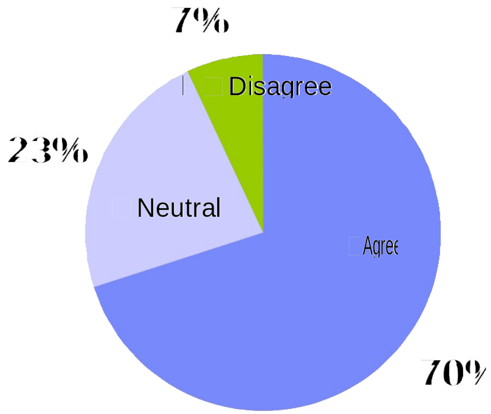
83% of CIOs cited “Business intelligence and analytics” as part of their visionary plans to enhance competitiveness

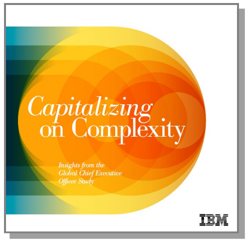
60% of CEOs need to do a better job capturing and understanding information rapidly in order to make swift business decisions

Example: The Perception Gap Surrounding Social Media . . . .

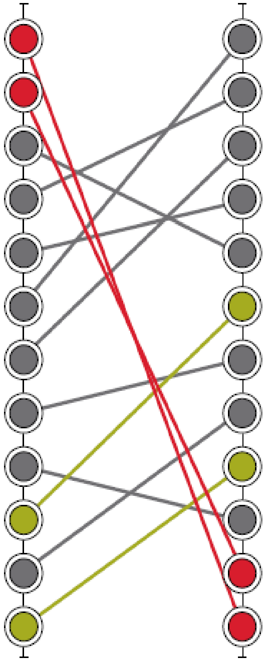
- ***IBM 2010 CEO Study: 88 percent of CEOs said "getting closer to customers" was top priority over next 5 years and viewed social media as a core part of that strategy***
- ***However, a March 2011 IBM study identified that companies fail to understand what customers want from social advertising and outreach***

Social media and social networking will increase customer advocacy?

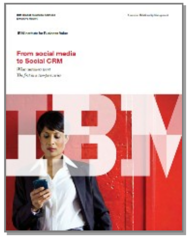



Capitalizing on Complexity
Insights from the Global Chief Executive Officer Study
IBM

Source: "Capitalizing on complexity, Insights from the Global Chief Executive Office Study," IBM Institute for Business Value, 2010



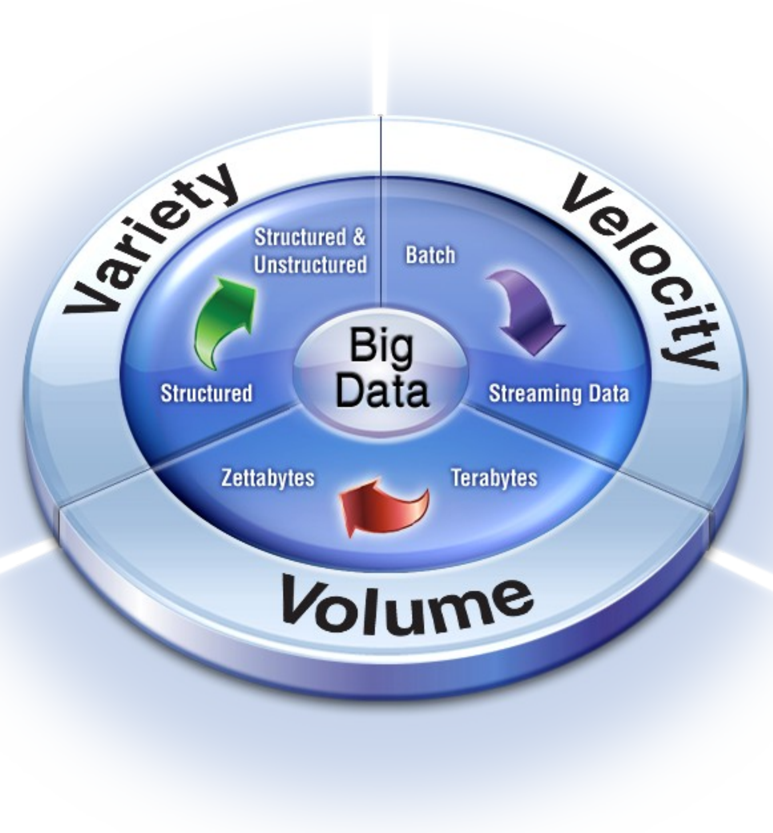
Consumers' ranking: The reasons they interact with companies via social sites	Perception gap	***Businesses' ranking:*** Why they think consumers follow them via social sites
(61%) Discount		Learn about new products (73%)
(55%) Purchase		General information (71%)
(53%) Reviews and product rankings		Submit opinion on current products
(53%) General information		Exclusive information (68%)
(52%) Exclusive information		Reviews and product rankings (67%)
(51%) Learn about new products		Feel connected (64%)
Submit opinion on current products/services		Customer service (63%)
(37%) Customer service		Submit ideas for new products/services
(34%) Event participation		Be part of a community (61%)
(33%) Feel connected		Event participation (61%)
(30%) Submit ideas for new products/services		Purchase (60%)
(22%) Be part of a community		Discount (60%)



"What Customers Want"
First in a two-part series
IBM Institute for Business Value
Published March 2011

Big Data Presents Big Opportunities

Extract insight from a high volume, variety and velocity of data in a timely and cost-effective manner



Variety: Manage and benefit from diverse data types and data structures

Velocity: Analyze streaming data and large volumes of persistent data

Volume: Scale from terabytes to zettabytes

What we hear from customers . . . .

- Lots of potentially valuable data is dormant or discarded due to size/performance considerations
- Large volume of unstructured or semi-structured data is not worth integrating fully (e.g. Tweets, logs, . . .)
- Not clear what should be analyzed (exploratory, iterative)
- Information distributed across multiple systems and/or Internet
- Some information has a short useful lifespan
- Volumes can be extremely high
- Analysis needed in the context of existing information (not stand alone)

Merging the Traditional and Big Data Approaches

Traditional Approach

Structured & Repeatable Analysis

Business Users

Determine what question to ask

IT

Structures the data to answer that question

Monthly sales reports
Profitability analysis
Customer surveys

Big Data Approach

Iterative & Exploratory Analysis

IT

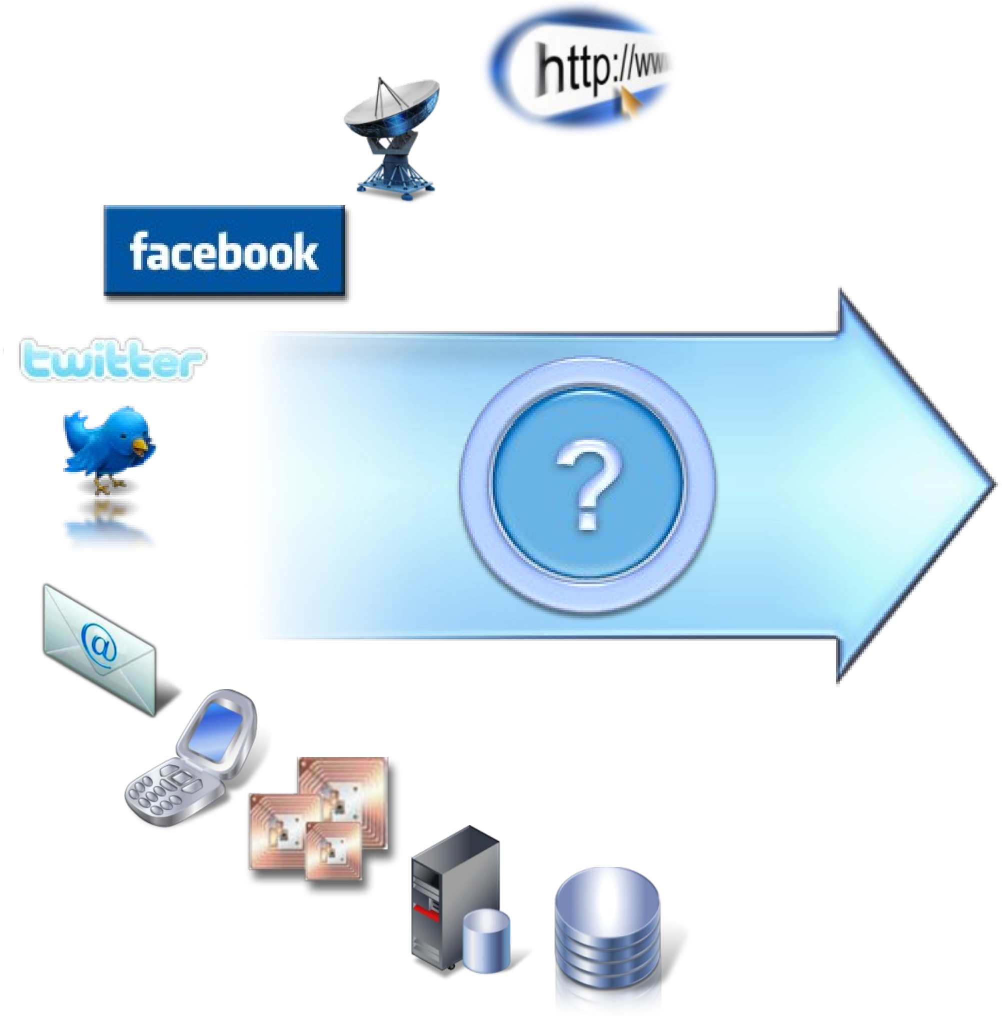
Delivers a platform to enable creative discovery

Business

Explores what questions could be asked

Brand sentiment
Product strategy
Maximum asset utilization

Big Data Scenarios Span Many Industries



Multi-channel customer sentiment and experience a analysis

Detect life-threatening conditions at hospitals in time to intervene

Predict weather patterns to plan optimal wind turbine usage, and optimize capital expenditure on asset placement

Make risk decisions based on real-time transactional data

Identify criminals and threats from disparate video, audio, and data feeds

Vestas (European Energy Company)

Business Challenge

- Analyze large volumes of public and private weather data for alternative energy business
- Existing high-performance computing hardware, limited staff

Project objectives

- Leverage large volume (2+ PB) of weather data to optimize placement of turbines.
- Reduce modeling time from weeks to hours.
- Optimize ongoing operations.

The benefits

- Reliability, security, scalability, and integration needs fulfilled
- Standard enterprise software support
- Single-vendor solution for software, hardware, storage, support

Solution Components:

- IBM InfoSphere BigInsights Enterprise Edition:
  - Scalability (data volumes)
  - Jaql (query support and extensibility)
  - IBM-provided file system (support existing hardware & apps)

- Strong runtime performance

- IBM xSeries hardware

http://www.ibm.com/developerworks/wiki/biginsights: videos/interviews

Global Technology Firm

Business challenge

- Analyze & correlate log records across to improve service
- Detect & predict failure patterns; initiate automated or manual preventive actions

Project objectives

- Process variety of logs generated by multiple systems, devices in distinct formats (XML, text, …)
- Accommodate large data volumes growing at ~1 TB /day
- Parse logs, identify/extract entities of interest, index as needed, cluster data by sessions, detect & visualize patterns through GUI
- Report on Top X, Bottom X patterns; support exploratory queries

The benefits

- IBM analytics and tooling simplify development and speed time-to-value.
- *"You have done in 2 weeks what I have been trying to generalize for the past 6 months."* -- Customer project leader

Solution Components:

- IBM InfoSphere BigInsights Enterprise Edition including:
  - Spreadsheet data discovery and visualization
  - Text analytics runtime and tooling
  - Flexible query support
  - Scalability
- IBM InfoSphere Streams

Global Media Firm

Business challenge

- Identify unauthorized content streaming (piracy)
- Quantify annual revenue loss, analyze trends
- Monitor social media sites (e.g., Twitter, Facebook) to identify dissemination of pirated content. Time sensitive!

Project objectives:

- Analyze high variety of data. Volumes unclear.
- Start with social media data for 1 year. Use text analytics to
  - Qualify & classify info of interest (complex, custom set of rules)
  - Search for URLs with live streaming of target data, sentiment, ….
- Future potential for video analysis

The benefits

- Improved understanding of business exposures through advanced analytics
- Improved decision-making process
- Scalable, flexible infrastructure for handling future analytic needs

Solution Components:

- IBM InfoSphere BigInsights Enterprise Edition including:
  - Text analytics runtime and tooling
  - Custom text annotators
  - Flexible query support
  - Scalability

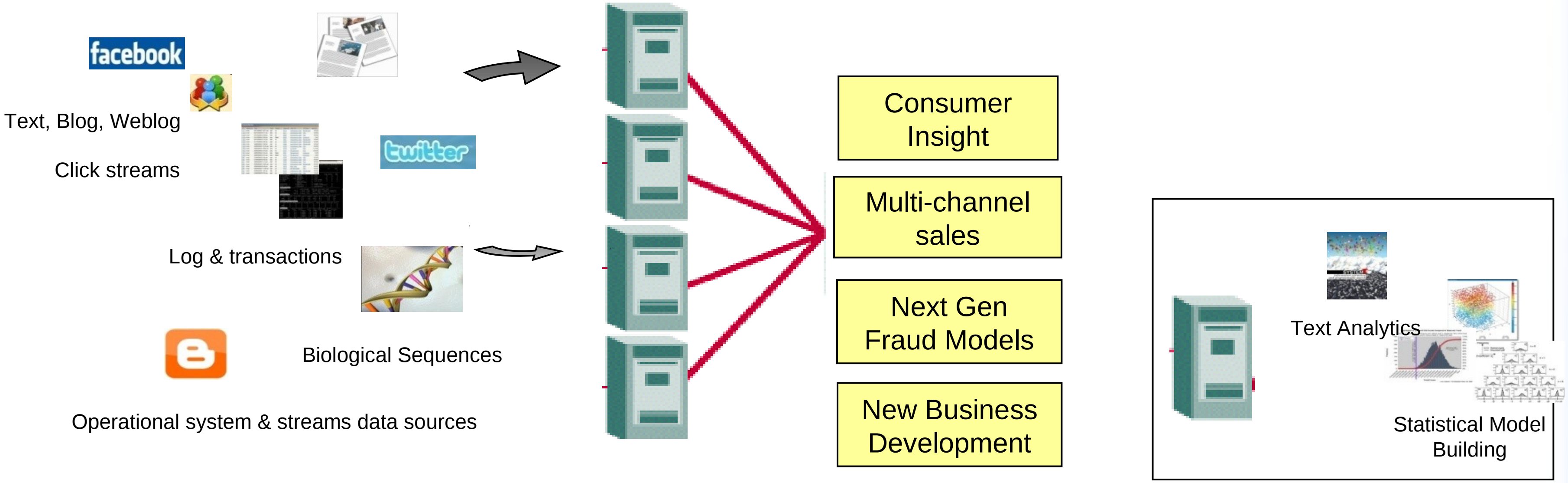
Customer Engagements

Use patterns

- Customer sentiment analysis (cross-sell, up-sell, campaign management)
- Integrated retail and web customer behavior modeling
- Predictive modeling (credit card fraud)
- System log analytics (reduce operational risk)

Common requirements

- Extract business insight from large volumes of raw data (often outside operational systems)
- Integrate with other existing software
- Ready for enterprise use

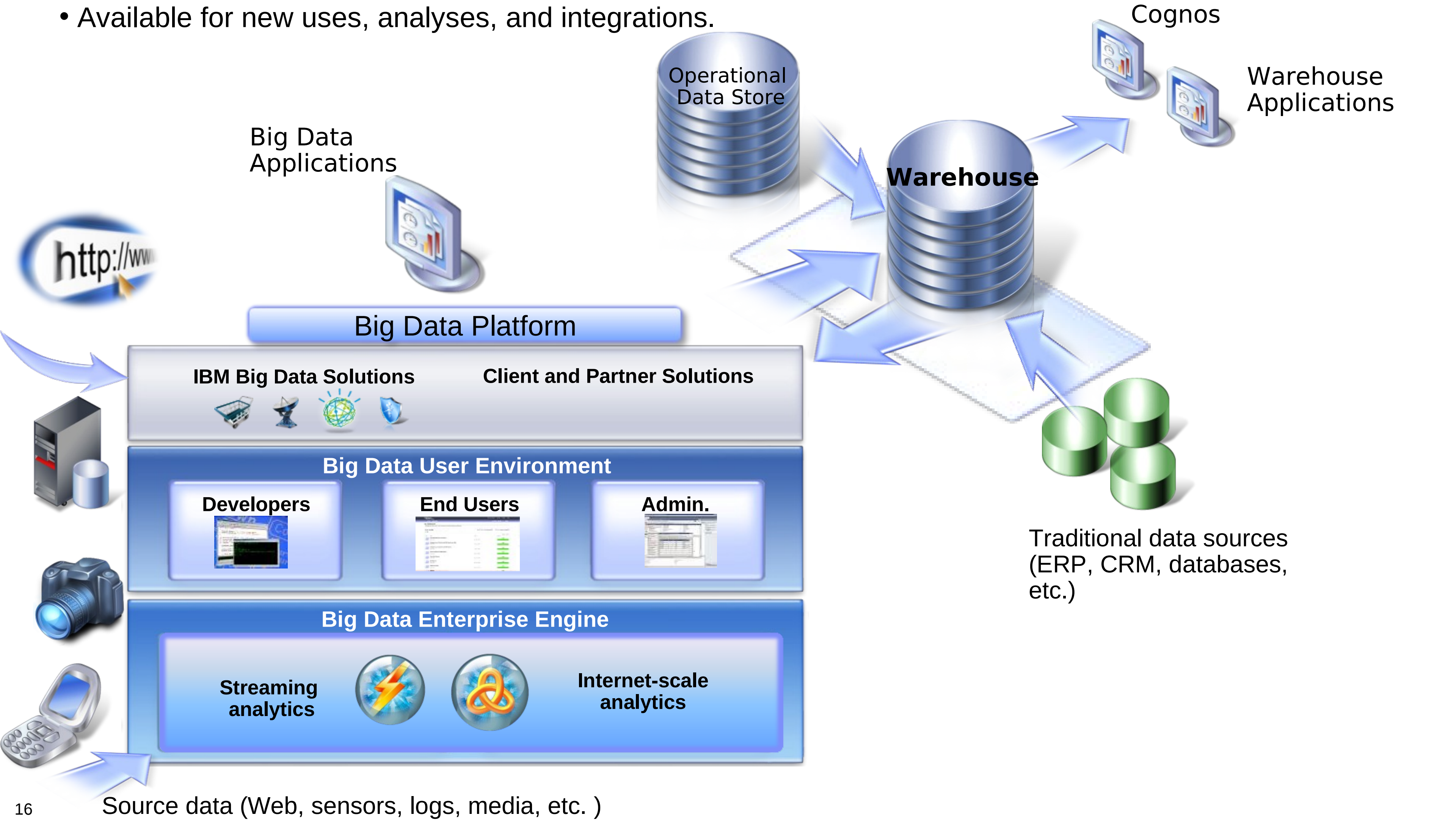


IBM’s approach

November 7, 2011

Big Data: an integral part of an enterprise data platform

- Manage Big Data from the instant it enters the enterprise
- High fidelity – no changes to original format
- Available for new uses, analyses, and integrations.



IBM's Platform Addresses Key Requirements

1. Platform for V³ – Variety, Velocity, Volume

- **Variety -** manage data & content "As Is"
- **Handle any velocity -** low-latency streams and large volume batch
- **Volume -** huge volumes of at-rest or streaming data

2. Analytics for V³

- **Analyze Sources in their native format -** text, data, rich content
- **Analyze all of the data -** not just a subset
- **Dynamic analytics -** automatic adjustments and actions

3. Ease of Use for Developers and Users

- **Developer UIs, common languages & automatic optimization**
- **End-user UIs & visualization**

4. Enterprise Class

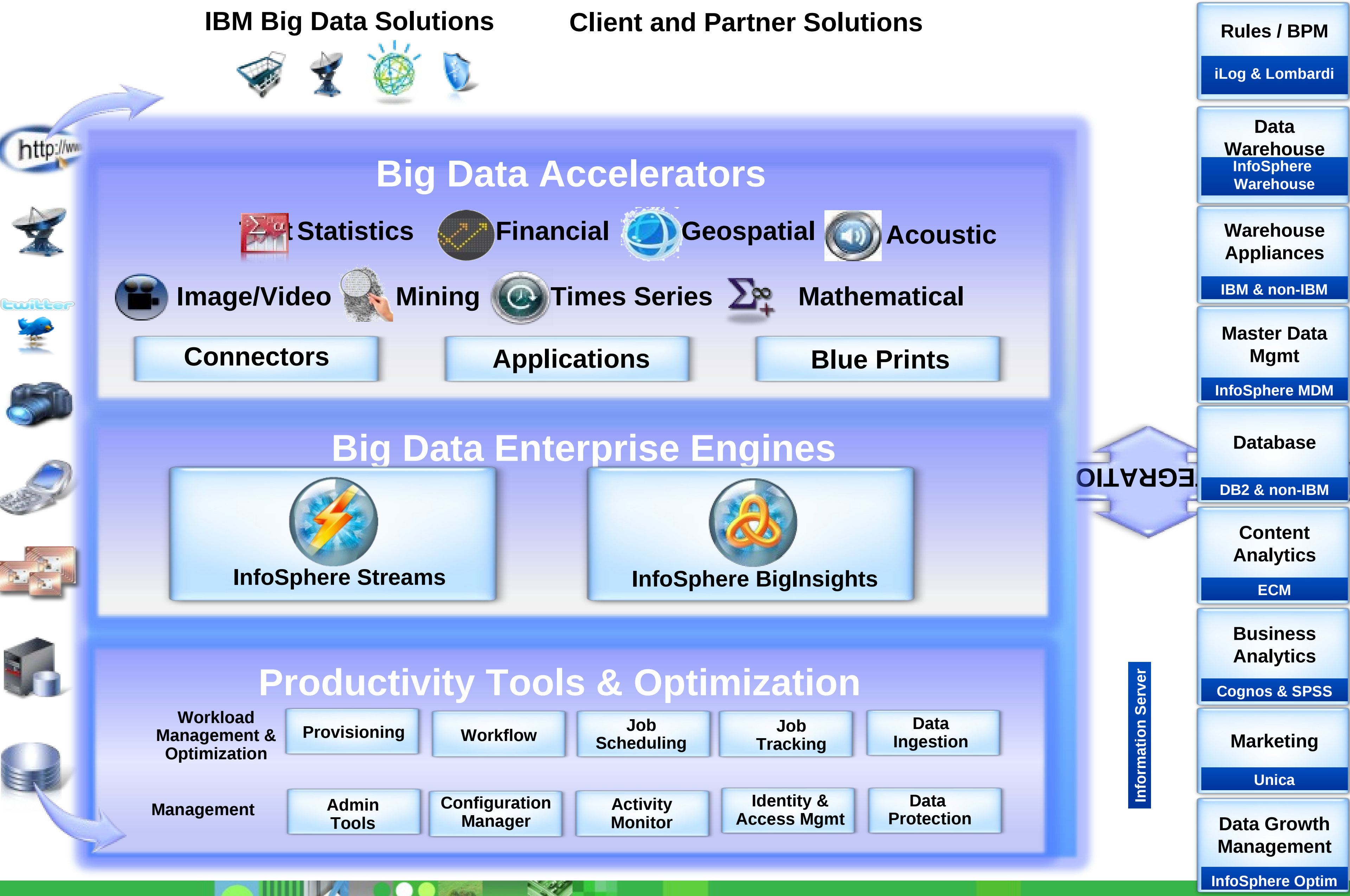
- **Failure tolerance, Security and Privacy**
- **Scale Economically**

5. Extensive Integration Capabilities

- **Integrate wide variety of sources**
- **Leverage enterprise integration technologies**

Big Data Platform

Platform Vision



IBM Big Data Solutions

Client and Partner Solutions

Big Data Accelerators

Statistics | Financial | Geospatial | Acoustic

Image/Video | Mining | Times Series | Mathematical

Connectors | Applications | Blue Prints

Big Data Enterprise Engines

InfoSphere Streams

InfoSphere BigInsights

Productivity Tools & Optimization

Workload Management & Optimization: Provisioning | Workflow | Job Scheduling | Job Tracking | Data Ingestion

Management: Admin Tools | Configuration Manager | Activity Monitor | Identity & Access Mgmt | Data Protection

Information Server

Rules / BPM – iLog & Lombardi

Data Warehouse – InfoSphere Warehouse

Warehouse Appliances – IBM & non-IBM

Master Data Mgmt – InfoSphere MDM

Database – DB2 & non-IBM

Content Analytics – ECM

Business Analytics – Cognos & SPSS

Marketing – Unica

Data Growth Management – InfoSphere Optim

BigInsights Summary

- **BigInsights = analytical platform for persistent "Big Data"**
  - Based on open source & IBM technologies
  - Managed like a start-up . . . . Emphasis on deep customer engagements, product plan flexibility

- **Distinguishing characteristics**
  - Built-in analytics . . . . *Enhances business knowledge*
  - Enterprise software integration . . . . *Complements and extends existing capabilities*
  - Production-ready platform . . . . *Speeds time-to-value; simplifies development and maintenance*

- **IBM advantage**
  - Combination of software, hardware, services and advanced research

InfoSphere BigInsights

Platform for volume, variety, velocity -- V^3

- Enhanced Hadoop foundation

Analytics for V^3

- Text analytics & tooling

Usability

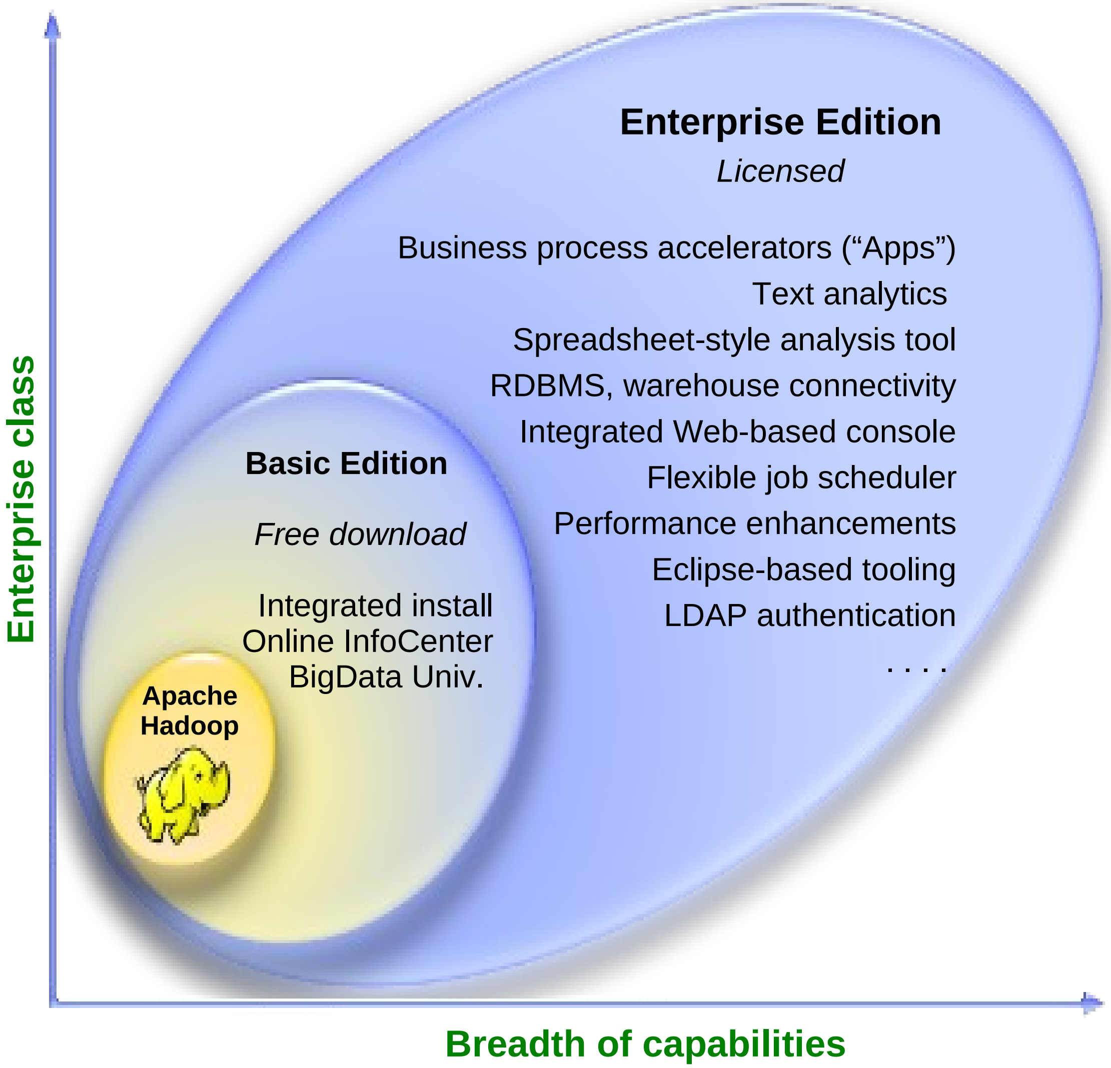
- Web console
- Integrated install
- Spreadsheet-style tool
- Ready-made "apps"

Enterprise Class

- Storage, security, cluster management

Integration

- Connectivity to DB2, Netezza



Enterprise Edition

Licensed

Business process accelerators ("Apps")
Text analytics
Spreadsheet-style analysis tool
RDBMS, warehouse connectivity
Integrated Web-based console
Flexible job scheduler
Performance enhancements
Eclipse-based tooling
LDAP authentication
. . . .

Basic Edition

Free download

Integrated install
Online InfoCenter
BigData Univ.

BigInsights Content

Function	Version	Basic Edition	Enterprise Edition
Integrated Install*		Inc	Inc
Hadoop (including common utilities, HDFS, MapReduce framework)	0.20.2	Inc	Inc
Jaql (programming / query language)	0.5.2	Inc	Inc
Pig (programming / query language)	0.7	Inc	Inc
Flume (data collection/aggregation)	0.9.1	Inc	Inc
Hive (data summarization/querying)	0.5	Inc	Inc
Lucene (text search)*	3.1.0	Inc	Inc
Zookeeper (process coordination)	3.2.2	Inc	Inc
Avro (data serialization)*	1.5.1	Inc	Inc
HBase (real time read/write)	0.20.6	Inc	Inc
Oozie (workflow/ job orchestration)	2.2.2	Inc	Inc
Online documentation		Inc	Inc
Capability to integrate with JDBC sources through general-purpose Jaql module*		Inc	Inc
Capability to integrate with DB2, InfoSphere Warehouse (DB2 UDF samples to submit jobs, and read results from BigInsights)		Inc	Inc

**New or upgraded in 1.2*

BigInsights Content (cont’d)

Function	Version	Basic Edition	Enterprise Edition
Capability to integrate with R (Jaql module to invoke R statistical capabilities from BigInsights)		n/a	**Inc**
Capability to integrate with Netezza, DB2 LUW with DPF from Jaql		n/a	**Inc**
LDAP Authentication		n/a	**Inc**
Integrated Web Console*		n/a	**Inc**
Integrated workflow capabilities		n/a	**Inc**
Integrated flexible scheduler		n/a	**Inc**
Platform performance enhancements (Adaptive MapReduce, efficient processing of compressed files)*		n/a	**Inc**
Text analytics capability		n/a	**Inc**
Eclipse support for text analytic development, Jaql, Hive, Java*		n/a	**Inc**
Spreadsheet-like analytical tool (BigSheets)*		n/a	**Inc**
IBM Optim Development Studio V2.2.1.0		n/a	**Inc**

**New or upgraded*

Announcing BigInsights V1.3

Enhanced Web Console:

- **Administration tools**
  - –View cluster health
  - –Manage cluster access
  - –Manage/install cluster instances.
- **Tools for big data – Web tools to:**
  - –Run big data applications
  - –View progress
  - –Graph results
  - –Integrate with BigSheets
  - –Manage and schedule workflows, jobs, tasks, and files

Greater Efficiency:

- **Adaptive MapReduce – Improve performance for small jobs (without altering how jobs are created)**
- **Compression – Decrease disk space & storage infrastructure requirements.**

Better Manageability:

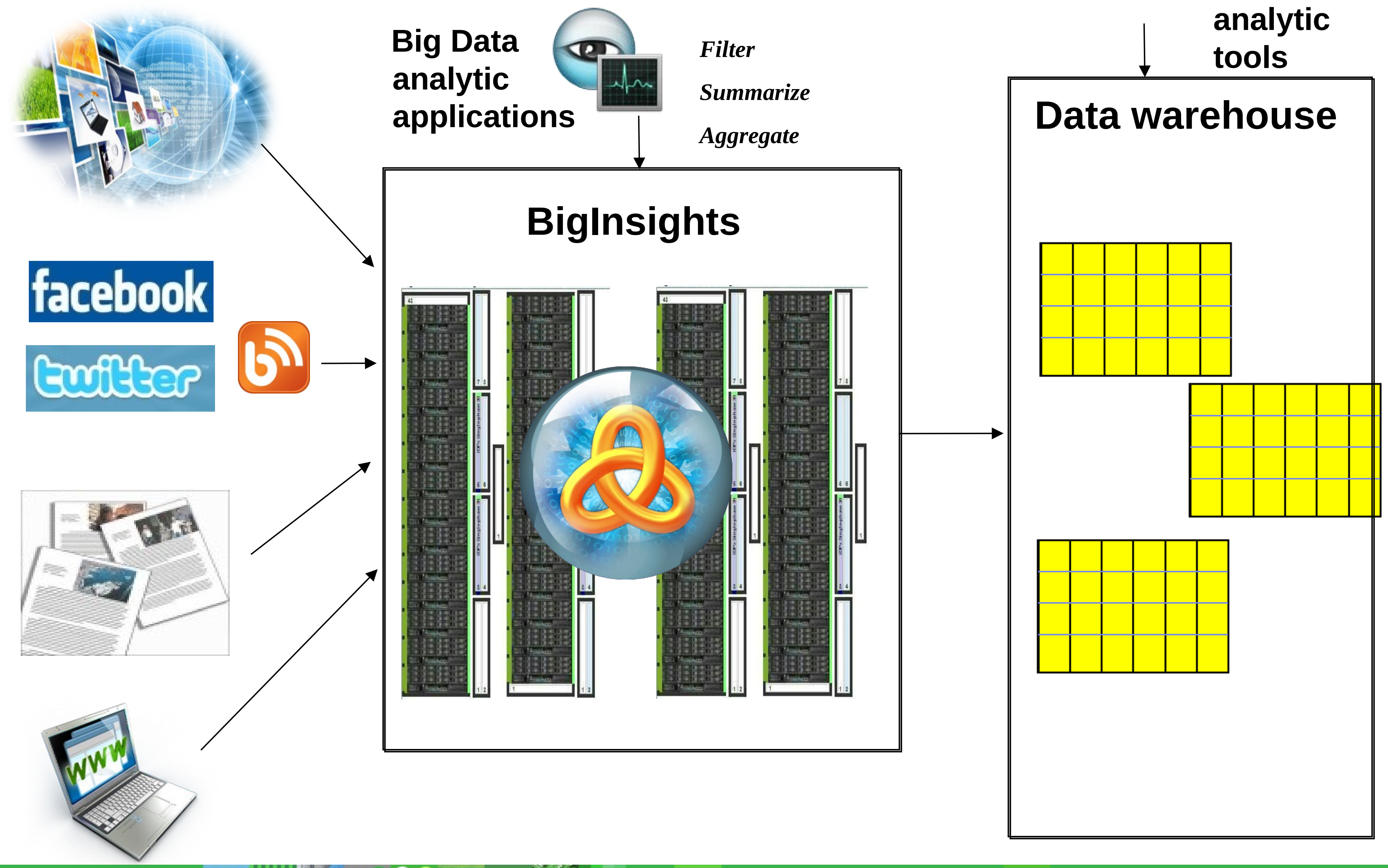
- **Development tools for:**
  - –Text analytics
  - –Java map reduce development
  - –Cluster file browsing
  - –Job submission
  - –Jaql and Hive development
  - –Developing and publishing applications to the web console
- **Web Secure online REST access to cluster to automatically leverage applications and access data**
- **Web applications for:**
  - –Securely importing and exporting data with relational databases
  - –Importing and export files to the cluster
  - –Importing data from web crawlers and social media.

BigInsights: Value Beyond Open Source

- **Technical differentiators**
  - Built-in analytics
    - Text processing engine, annotators, Eclipse tooling
    - Interface to project R (statistical platform)
  - Enterprise software integration (DBMS, warehouse)
  - Simplified programming / query interface (Jaql)
  - Integrated installation of supported open source and IBM components
  - Web-based management console
  - Platform enrichment: additional security, job scheduling options, performance features, . . .
  - Standard IBM licensing agreement and world-class support
  - More to come in future releases!

- **Business benefits**
  - Quicker time-to-value due to IBM technology and support
  - Reduced operational risk
  - Enhanced business knowledge with flexible analytical platform
  - Leverages and complements existing software assets

BigInsights and the data warehouse



Traditional analytic tools

Big Data analytic applications

Filter

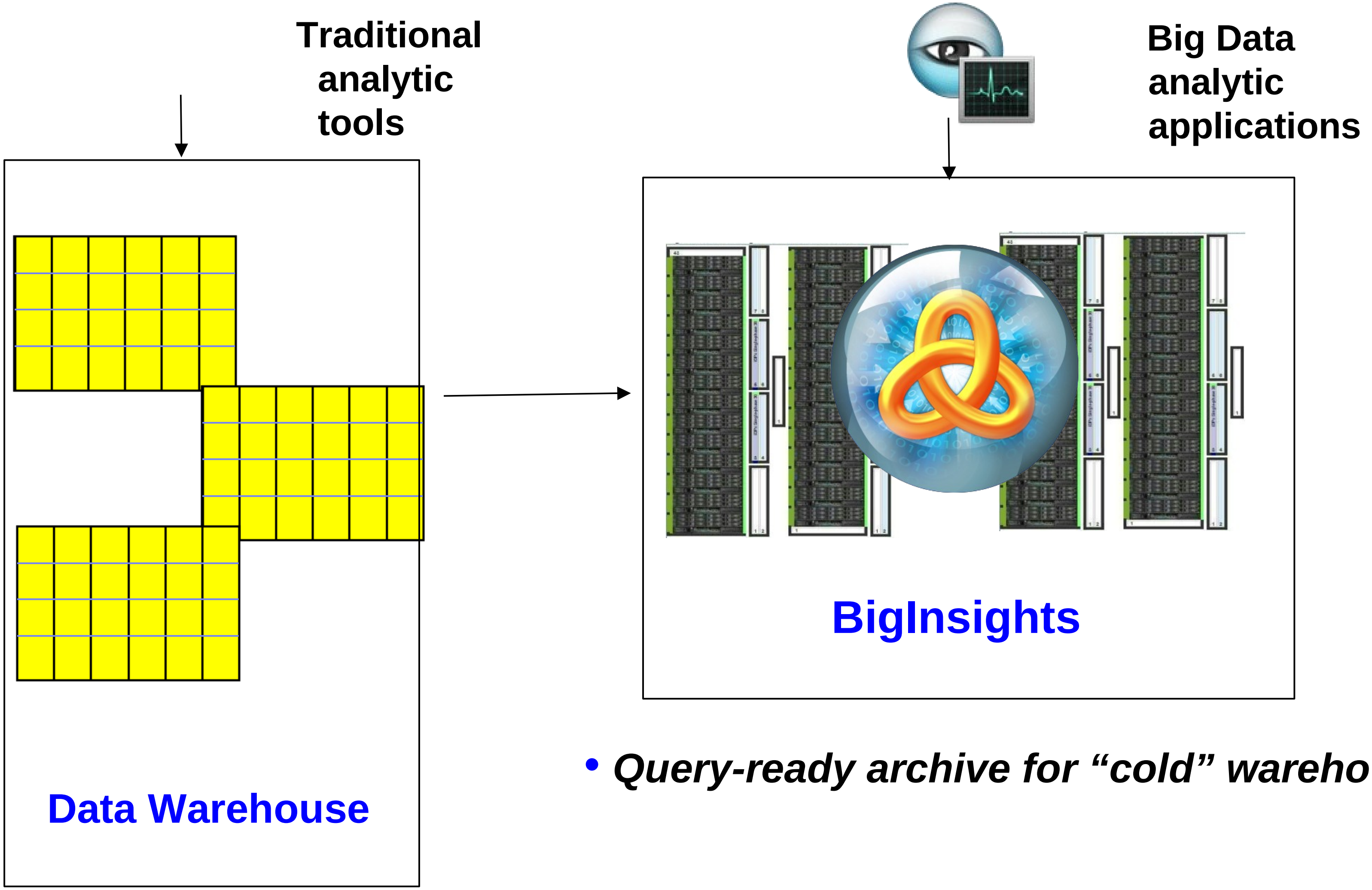
Summarize

Aggregate

BigInsights

Data warehouse

BigInsights and the data warehouse



Traditional analytic tools

Big Data analytic applications

Data Warehouse

BigInsights

- ***Query-ready archive for “cold” warehouse data***

Growing Ecosystem of Solutions

IBM BigInsights Solutions

Cognos Consumer Insights
Social media analytics solution that uses BigInsights. Available now.

IBM Content Analytics
Unlock valuable business insight from unstructured data. Proof of technology completed. Production offering due soon.

Partner Solutions

KARMASPHERE

Kitenga
Reinventing Information

GOTOMETRICS
metrics for the masses

DIGITAL REASONING
SYSTEMS

PivotLink
Intelligence on demand

JASPERSOFT

Datameer
Powerfully Simple™

. . . with more to come

IBM Big Data User Environments

IBM Big Data Platform

A Closer Look at BigInsights . . . .

About the BigInsights Platform

- **Flexible, enterprise-class support for processing large volumes of data**
  - Based on Google’s MapReduce technology
  - Inspired by Apache Hadoop; compatible with its ecosystem and distribution
  - Well-suited to batch-oriented, read-intensive applications
  - Supports wide variety of data

- **Enables applications to work with thousands of nodes and petabytes of data in a highly parallel, cost effective manner**
  - CPU + disks = “node”
  - Nodes can be combined into clusters
  - New nodes can be added as needed without changing
    - Data formats
    - How data is loaded
    - How jobs are written

The MapReduce Programming Model

- **"Map" step:**
  - Input split into pieces
  - Worker nodes process individual pieces in parallel (under global control of the Job Tracker node)
  - Each worker node stores its result in its local file system where a reducer is able to access it
- **"Reduce" step:**
  - Data is aggregated ('reduced" from the map steps) by worker nodes (under control of the Job Tracker)
  - Multiple reduce tasks can parallelize the aggregation

Logical MapReduce Example: Word Count

```
map(String key, String value):
// key: document name
// value: document contents
for each word w in value:
  EmitIntermediate(w, "1");

reduce(String key, Iterator values):
// key: a word
// values: a list of counts
int result = 0;
for each v in values:
  result += ParseInt(v);
Emit(AsString(result));
```

Content of Input Documents

Hello World Bye World

Hello IBM

Map 1 emits:
< Hello, 1>
< World, 1>
< Bye, 1>
< World, 1>

Map 2 emits:
< Hello, 1>
< IBM, 1>

Reduce (final output):

< Bye, 1>
< IBM, 1>
< Hello, 2>
< World, 2>

MapReduce Processing

Input Documents

```
Hello World Bye
World

Hello IBM
```

Map 1 emits:

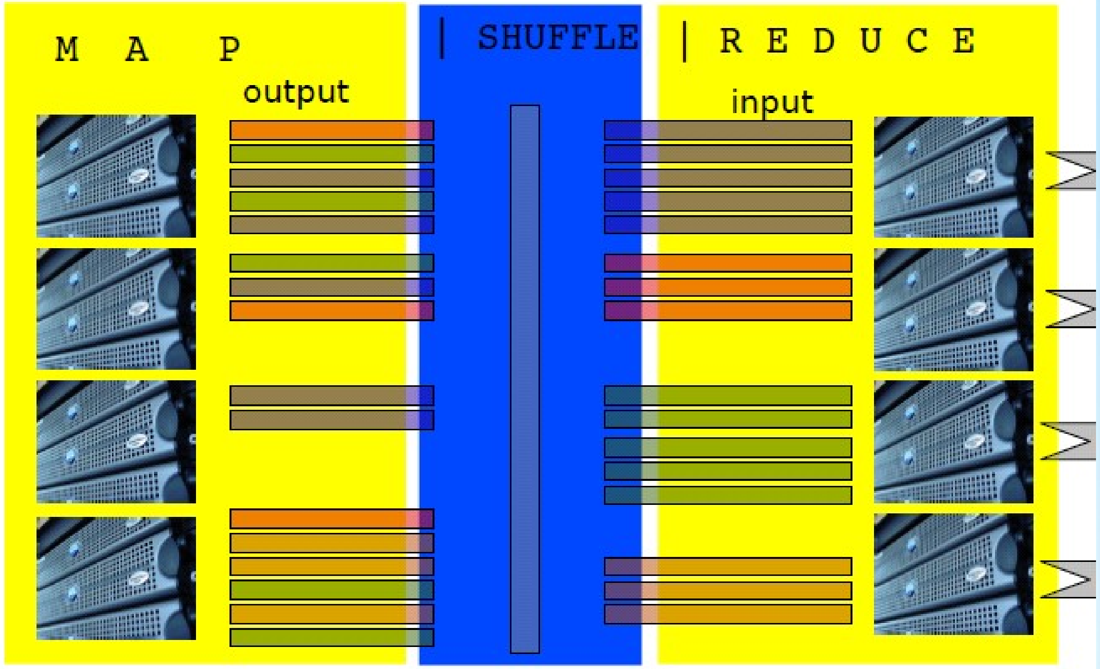
```
< Hello, 1>
< World, 1>
< Bye, 1>
< World, 1>
```

Map 2 emits:

```
< Hello, 1>
< IBM, 1>
```

Reduce (final output):

```
< Bye, 1>
< IBM, 1>
< Hello, 2>
< World, 2>
```



Local to data.
Outputs a lot less data.
Output can cheaply move.

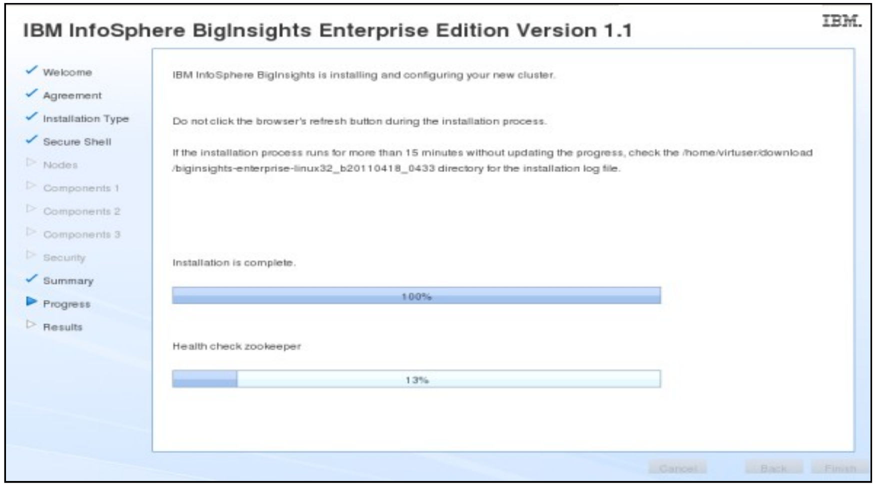
Shuffle sorts input by key.
Reduces output significantly.

So What Does This Result In?

- Easy To Scale
- Fault Tolerant and Self-Healing
- Data Agnostic
- Extremely Flexible

Web-based Installation, Management Consoles

- **Integrated installation**
  - Seamless process for single node and cluster environments
  - Post-install validation of IBM and open source components



IBM InfoSphere BigInsights Enterprise Edition Version 1.1

IBM InfoSphere BigInsights is installing and configuring your new cluster.

Do not click the browser's refresh button during the installation process.

If the installation process runs for more than 15 minutes without updating the progress, check the /home/virtuser/download /biginsights-enterprise-linux32_b20110418_0433 directory for the installation log file.

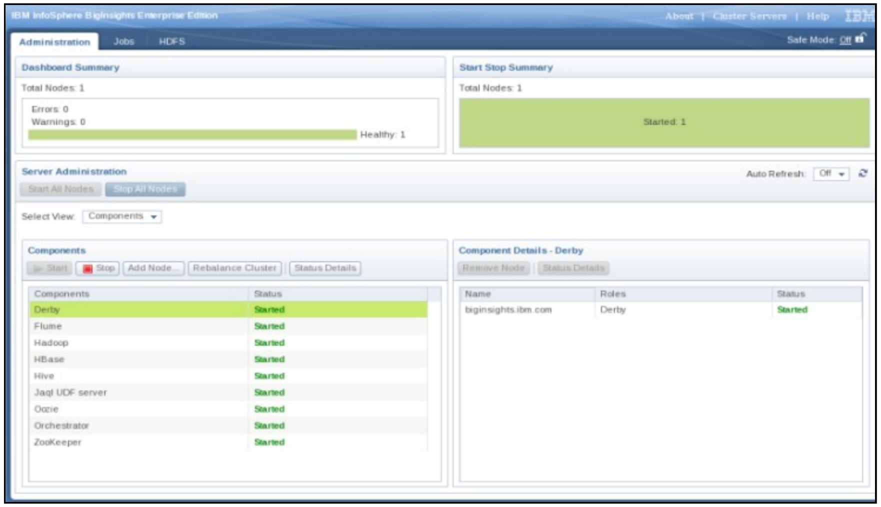
Installation is complete.

100%

Health check zookeeper

13%

- **Integrated management console**
  - System health management
  - Add / drop nodes
  - Start / stop services
  - Run / monitor jobs (applications)
  - Explore / modify file system
  - . . .



IBM InfoSphere BigInsights Enterprise Edition

Safe Mode: Off

Dashboard Summary

Total Nodes: 1

Errors: 0
Warnings: 0
Healthy: 1

Start Stop Summary

Total Nodes: 1

Started: 1

Server Administration

Auto Refresh: Off

Select View: Components

Components

Components	Status
Derby	Started
Flume	Started
Hadoop	Started
HBase	Started
Hive	Started
Jaql UDF server	Started
Oozie	Started
Orchestrator	Started
ZooKeeper	Started

Component Details - Derby

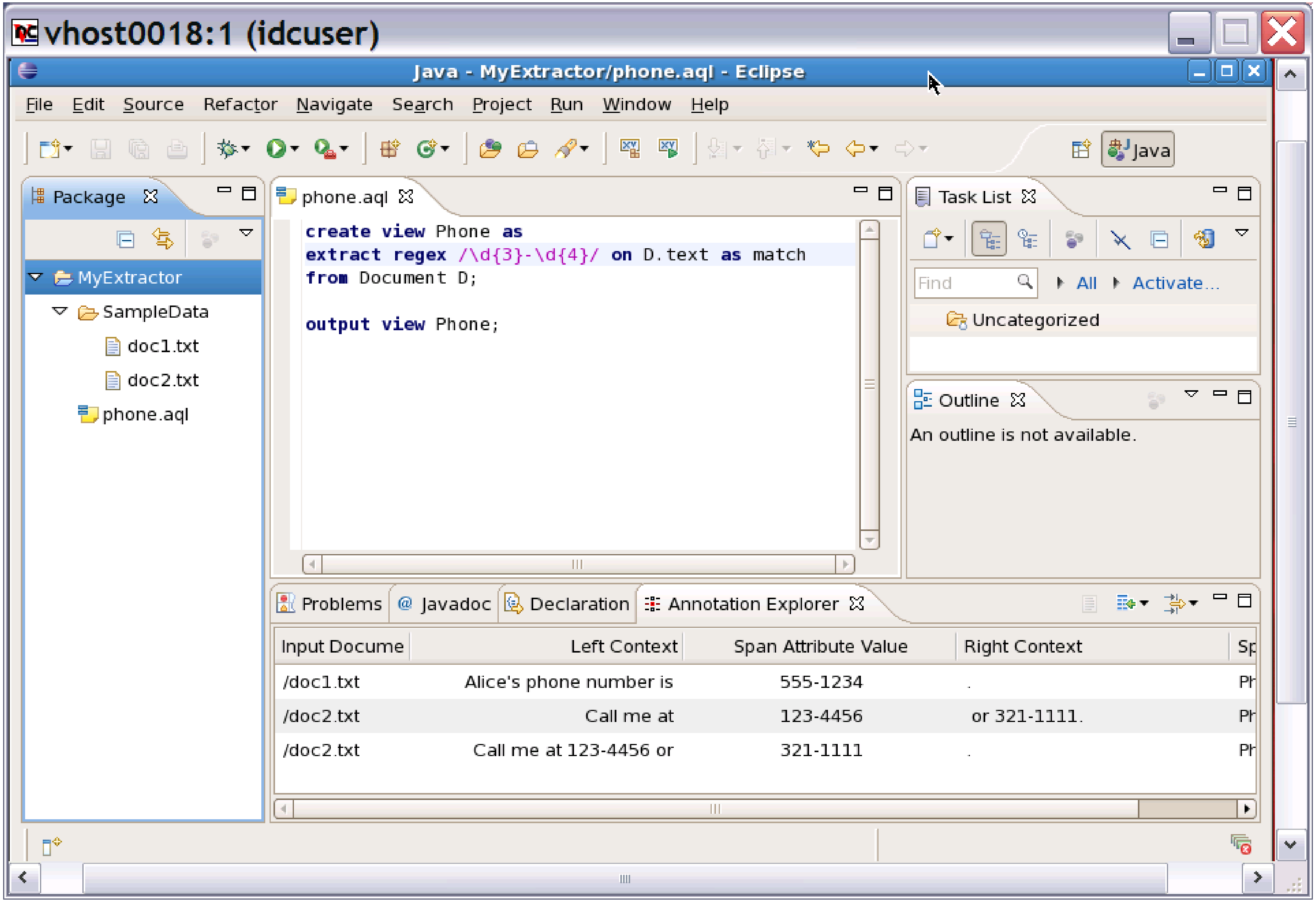
Name	Roles	Status
biginsights.ibm.com	Derby	Started

BigInsights and Text Analytics

- Distill structured info from unstructured data
  - Sentiment analysis
  - Consumer behavior
  - Illegal or suspicious activities
  - . . .
- Pre-built library of text annotators for common business entities
- Rich language and tooling to build custom annotators
- Support for Western languages (English, Dutch/Flemish, French, German, Italian, Portuguese, or Spanish) plus select Asian languages (Japanese, Chinese)

"Acquisition"
"Address"
"Alliance"
"AnalystEarningsEstimate"
"City"
"CompanyEarningsAnnouncement"
"CompanyEarningsGuidance"
"Continent"
"Country"
"County"
"DateTime"
"EmailAddress"
"JointVenture"
"Location"
"Merger"
"NotesEmailAddress"
"Organization"
"Person"
"PhoneNumber"
"StateOrProvince"
"URL"
"ZipCode"

BigInsights Text Analytics Development



```
create view Phone as
extract regex /\d{3}-\d{4}/ on D.text as match
from Document D;

output view Phone;
```

Input Docume	Left Context	Span Attribute Value	Right Context	Sp
/doc1.txt	Alice's phone number is	555-1234	.	Ph
/doc2.txt	Call me at	123-4456	or 321-1111.	Ph
/doc2.txt	Call me at 123-4456 or	321-1111	.	Ph

Example Analysis : Extraction from Twitter messages

Extract intent, interests, life events and micro segmentation attributes

Monetizable Intent

I had an iphone, but it's dead @JoaoVianaa. (I've no idea where it's) ! *Want* a *blackberry* now !!!

Relocation

@rakonturmiami* *im* *moving to miami* *in 3 months. i look foward to the new lifestyle

Name, Birth Day

@silliesylvia* *good!!! U shouldnt! Think about the important stuff, like ur birthday ;) btw* *happy birthday Sylvia* *;)

Location

I'm at Mickey's Irish Pub Downtown* *(206 3rd St, Court Ave, Des Moines)* *w/ 2 others http://4sq.com/gbsaYR

While accounting for less relevant messages

Subtle Spam, Advertising

I think that @justinbieber deserves his 2 AMAZING songs in top ten!!! Buy them on itunes

http://Cell-Pones.com Looking to buy a phone? WiFi Cell Phones, Windows Mobile

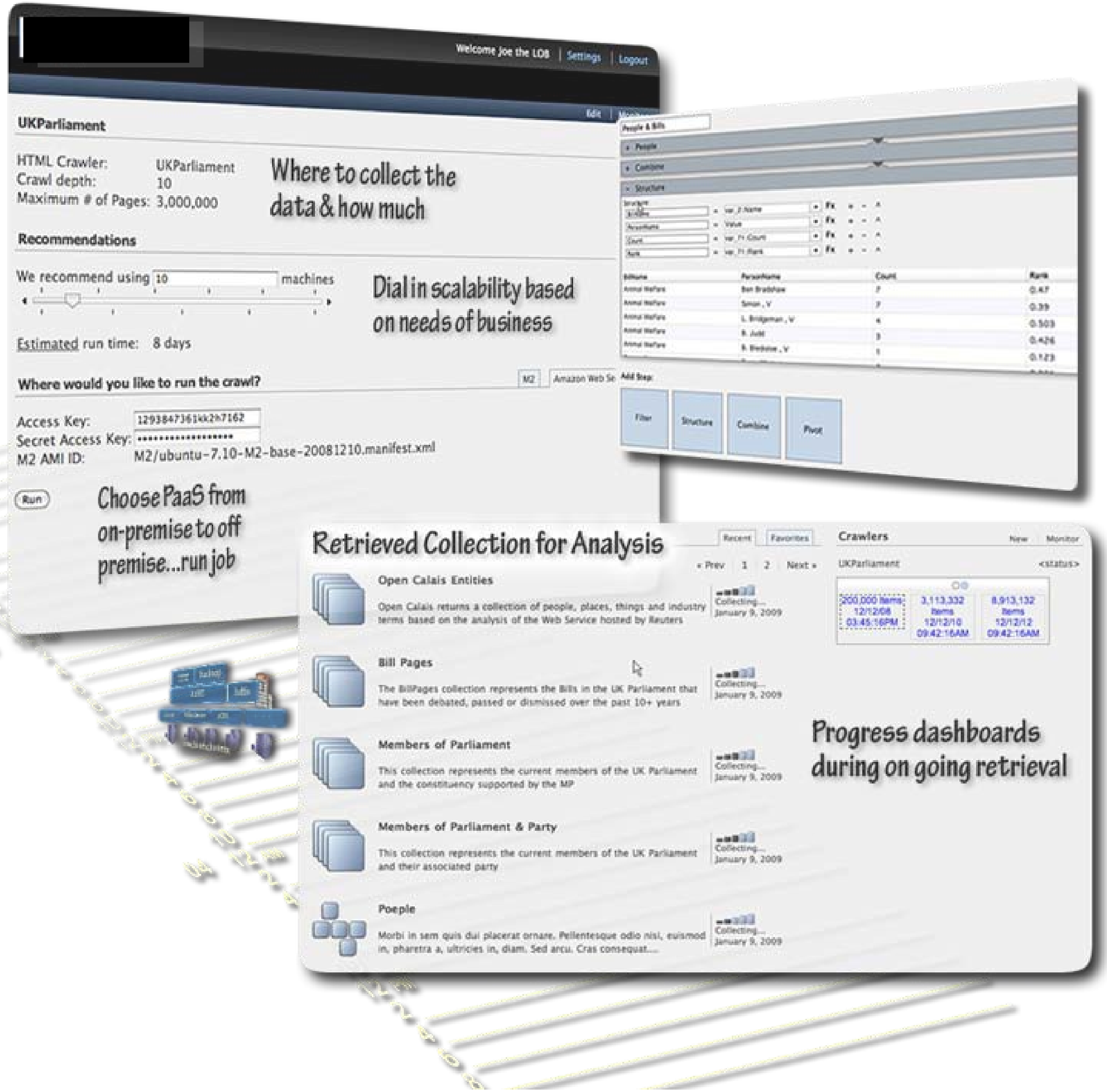
Sarcasm, Wishful Thinking

@purplepleather Gotta do more research my Versace term paper 2day. Before I die, I want a versace purple diamond tiara. Im just savin&gt;lol

had so much fun today! I want to buy a million dollar house with a wrap around porch … … wading river on the long island sound, ha i wish!

Spreadsheet-like Analysis Tool

- **Web-based analysis and visualization tool**
- **Spreadsheet-like interface**
  - Define and manage long running data collection jobs
  - Analyze content of the text on the pages that have been retrieved



[redacted]

Welcome Joe the LOB | Settings | Logout

Edit | Monitor

UKParliament

HTML Crawler: UKParliament
Crawl depth: 10
Maximum # of Pages: 3,000,000

Where to collect the data & how much

Recommendations

We recommend using 10 machines

Estimated run time: 8 days

Dial in scalability based on needs of business

Where would you like to run the crawl?

M2 | Amazon Web Se

Access Key: 1293847361kk2h7162
Secret Access Key: ••••••••••••••••••••
M2 AMI ID: M2/ubuntu-7.10-M2-base-20081210.manifest.xml

Run

Choose PaaS from on-premise to off premise...run job

People & Bills

People

Combine

Structure

Structure

BillName = var_2 Name
PersonName = Value
Count = var_11 Count
Rank = var_11 Rank

BillName	PersonName	Count	Rank
Animal Welfare	Ben Bradshaw	7	0.47
Animal Welfare	Simon , V	7	0.39
Animal Welfare	L. Bridgeman , V	4	0.503
Animal Welfare	B. Judd	3	0.426
Animal Welfare	B. Bledisloe , V	1	0.123

Add Step:

Filter | Structure | Combine | Pivot

Retrieved Collection for Analysis

Recent | Favorites

« Prev 1 2 Next »

Open Calais Entities

Open Calais returns a collection of people, places, things and industry terms based on the analysis of the Web Service hosted by Reuters

Collecting... January 9, 2009

Bill Pages

The BillPages collection represents the Bills in the UK Parliament that have been debated, passed or dismissed over the past 10+ years

Collecting... January 9, 2009

Members of Parliament

This collection represents the current members of the UK Parliament and the constituency supported by the MP

Collecting... January 9, 2009

Members of Parliament & Party

This collection represents the current members of the UK Parliament and their associated party

Collecting... January 9, 2009

Poeple

Morbi in sem quis dui placerat ornare. Pellentesque odio nisi, euismod in, pharetra a, ultricies in, diam. Sed arcu. Cras consequat....

Collecting... January 9, 2009

Crawlers New | Monitor

UKParliament <status>

200,000 Items 12/12/08 03:45:16PM	3,113,332 Items 12/12/10 09:42:16AM	8,913,132 Items 12/12/12 09:42:16AM

Progress dashboards during on going retrieval

Business Process Accelerators (“apps”)

- **Resuable software assets based on customer engagements**
  - Useful for starting point for various applications
  - Can be customized by BigInsights application developers as needed
  - Accessible through Eclipse, Web console
- **Available assets**
  - Data import/export (from relational DBMS, files)
  - Web crawler
  - Boardreader.com support (Web forum search engine)

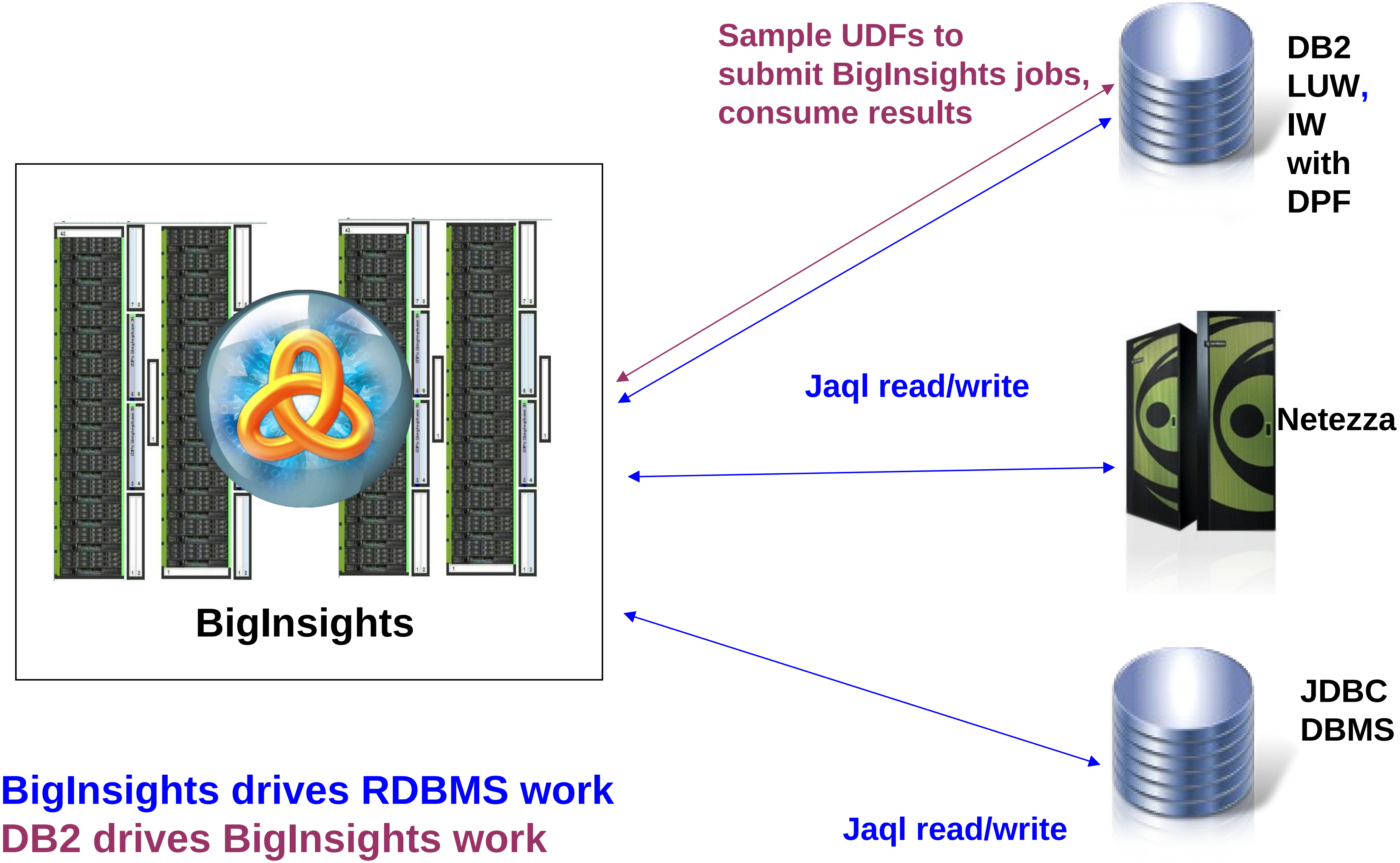
Performance enhancements

- **Flexible job scheduler option**
  - Optimize response time for small jobs
  - Available in addition to FAIR, FIFO scheduling

- **Adaptive MapReduce**
  - Speeds up a class of jobs (e.g., jobs that process small files)
  - Accomplished by changing how certain MapReduce tasks executed
    - Mappers can decide at runtime to take on more work (until it doesn’t make sense anymore). Communication via ZooKeeper.
  - Enabled through Jaql option, MapReduce job property setting

- **Efficient processing of compressed text data**
  - Use multiple Map tasks (vs Hadoop default of 1) for processing compressed text files
  - Enabled through BigInsights LZO-based compression technology
  - Automatic with Jaql; programming option with Java MapReduce

BigInsights Connectivity to DBMS / Warehouse



Sample UDFs to submit BigInsights jobs, consume results

DB2 LUW, IW with DPF

Jaql read/write

Netezza

BigInsights

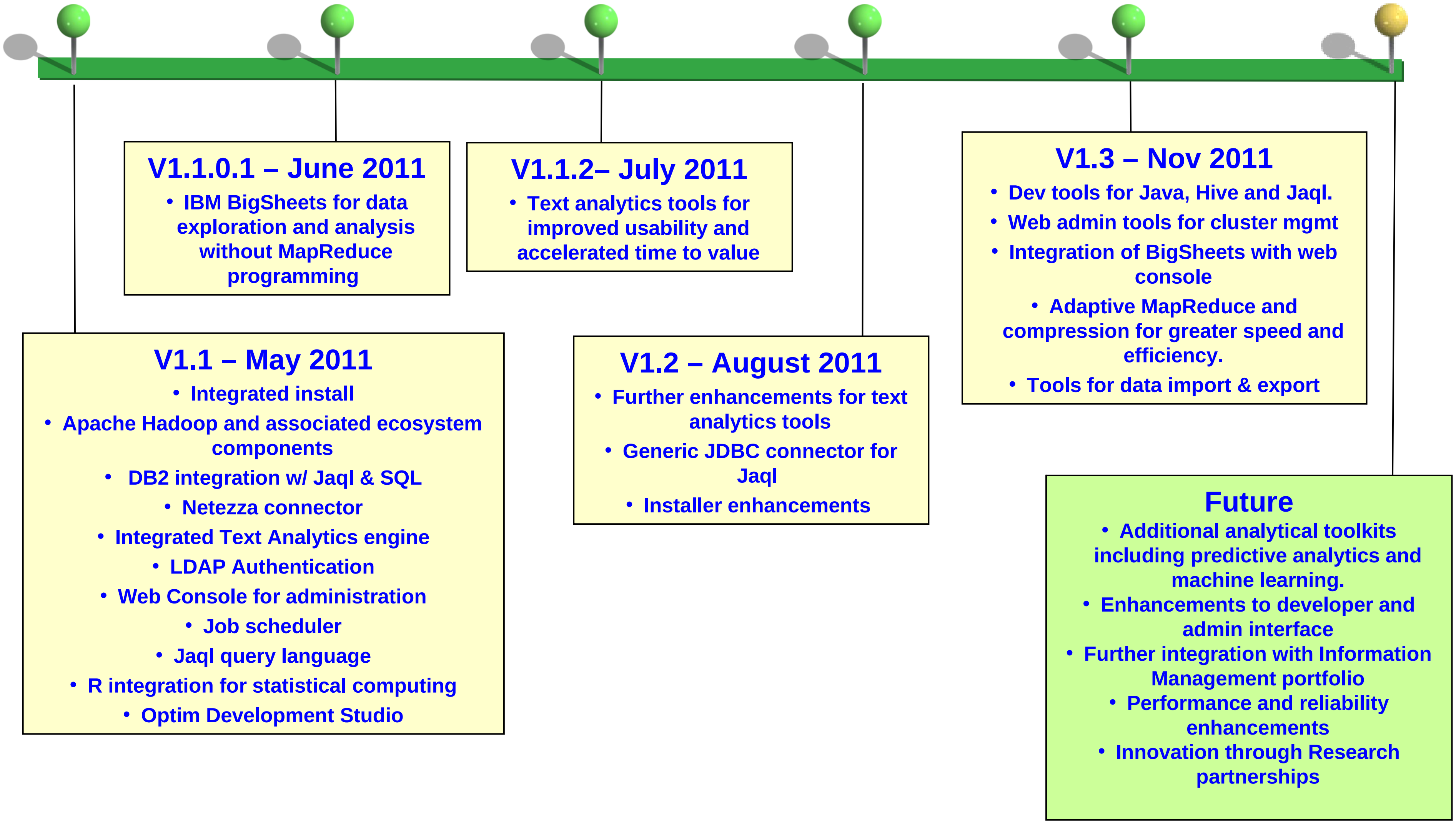
JDBC DBMS

Jaql read/write

- BigInsights drives RDBMS work
- DB2 drives BigInsights work

(As of October 2011)

InfoSphere BigInsights Roadmap



V1.1 – May 2011

- Integrated install
- Apache Hadoop and associated ecosystem components
- DB2 integration w/ Jaql & SQL
- Netezza connector
- Integrated Text Analytics engine
- LDAP Authentication
- Web Console for administration
- Job scheduler
- Jaql query language
- R integration for statistical computing
- Optim Development Studio

V1.1.0.1 – June 2011

- IBM BigSheets for data exploration and analysis without MapReduce programming

V1.1.2– July 2011

- Text analytics tools for improved usability and accelerated time to value

V1.2 – August 2011

- Further enhancements for text analytics tools
- Generic JDBC connector for Jaql
- Installer enhancements

V1.3 – Nov 2011

- Dev tools for Java, Hive and Jaql.
- Web admin tools for cluster mgmt
- Integration of BigSheets with web console
- Adaptive MapReduce and compression for greater speed and efficiency.
- Tools for data import & export

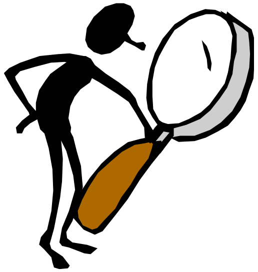
Future

- Additional analytical toolkits including predictive analytics and machine learning.
- Enhancements to developer and admin interface
- Further integration with Information Management portfolio
- Performance and reliability enhancements
- Innovation through Research partnerships

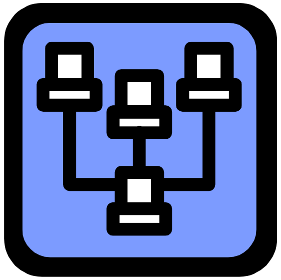
Trends and directions

- **Enterprise software integration**
  - Data warehouses, RDBMSs (IBM and non-IBM)
  - ETL platforms (e.g., DataStage)
  - Business intelligence tools (e.g., Cognos, SPSS)
  - Applications (e.g., Coremetrics, IBM partners)
  - . . .

- **Diverse range of analytics**
  - Text
  - Image / video (e.g., content-based user profiling)
  - Predictive modeling (e.g., ranking and classification based on machine learning)
  - . . .



- **Sophisticated, scalable infrastructure and tooling for processing massive data volumes**
  - High-performance file system with POSIX compliance, granular security
  - Fully recoverable and restartable workflows
  - Parallel, distributed indexing for text (“BigIndex”)
  - Tooling for administrators, programmers, analysts
  - Pre-built business process accelerators (“apps”)
  - . . .



About Big Data and BigInsights . . .

- **Big Data is a strategic initiative for IBM**
  - Significant investments across software, hardware and services.

- **InfoSphere BigInsights**
  - Enables firms to exploit growing variety, velocity, and volume of data
  - Delivers diverse range of analytics
  - Leverages and extends open source
  - Provides enterprise-class features and supporting services
  - Complement existing software investments and commercial offerings
  - Available in basic (free) and enterprise editions

- **IBM advantage**
  - Full solution spanning software, hardware & services
  - Rapid technology advances through partnerships with IBM Research
  - Global reach

Getting Off to a Fast Start with IBM

BigInsights – Try Before You Buy

- **In the Cloud**
  - Via RightScale, or directly on Amazon, Rackspace, IBM Smart Enterprise Cloud, or on private clouds.
  - Pay only for the resources used.
- **In the Virtual Classroom**
  - Free Hadoop Fundamentals training course @ www.bigdatauniversity.com
- **On Your Cluster**
  - Download Basic Edition from ibm.com.
- **In the Classroom**
  - Enroll in the InfoSphere BigInsights Essentials course.

Start your career on Hadoop with this FREE course

Learn with hands-on exercises on a Hadoop cluster

Why register?

Easy and Affordable Learning Hadoop and other Big Data technologies has never been more affordable! Many courses are FREE!

Latest industry trends Aquire valuable skills and get updated about industry's latest trends right here. Today!

Learn from the Experts! Big Data University offers education about Hadoop and other technologies by the industry's best!

Learn at your Own Pace! Find everything right here when you need it and from wherever you are.

View sample completion document!

sign up now for a chance to receive: FREE Books

Study Made Easy!

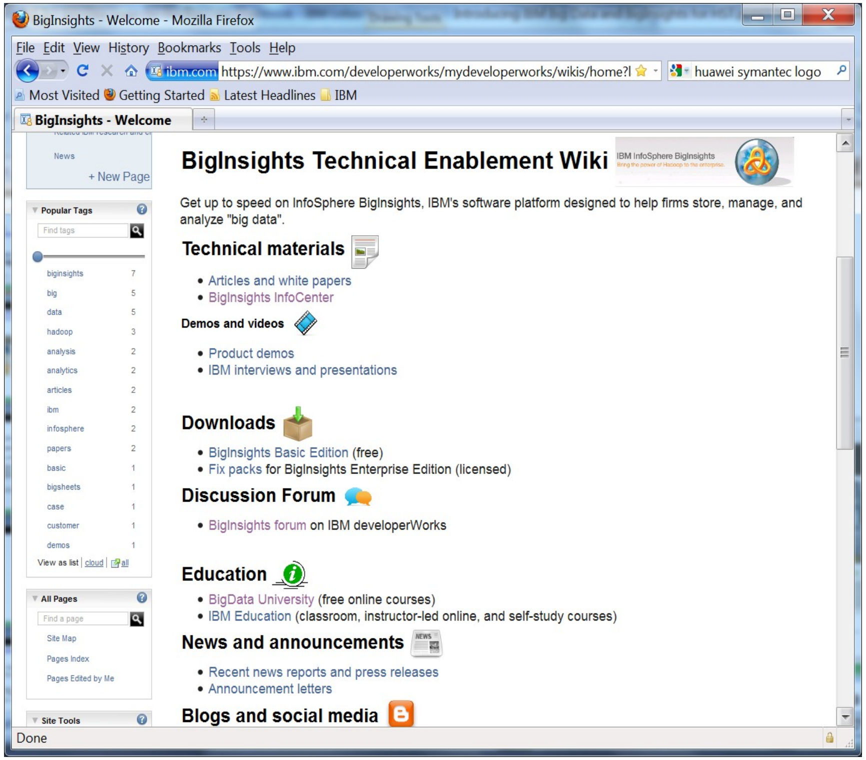
Student Testimonials

Balázs (USA)
The training material has short, easy to digest videos, with separately available transcripts allowing to execute the same commands after watching the videos. Each command line in the transcript can be cut and pasted making the examples easy to reproduce. There is very good support for Windows users both on 32bit and 64bit. Everything works right out of the box as described in the course materials. Online support on the Course Forums is excellent, most questions were answered before even I encountered the issue. The curriculum associates the chapters form the "Getting Started with

Visit the BigInsights technical portal . . . .

- **Free links to papers, demos, discussion forum, and more**

- **http://www.ibm.com/developerworks/wiki/biginsights/**



BigInsights - Welcome - Mozilla Firefox

File Edit View History Bookmarks Tools Help
ibm.com https://www.ibm.com/developerworks/mydeveloperworks/wikis/home?l
huawei symantec logo
Most Visited Getting Started Latest Headlines IBM
BigInsights - Welcome
News
+ New Page
Popular Tags
Find tags

biginsights	7
big	5
data	5
hadoop	3
analysis	2
analytics	2
articles	2
ibm	2
infosphere	2
papers	2
basic	1
bigsheets	1
case	1
customer	1
demos	1

View as list | cloud | all
All Pages
Find a page
Site Map
Pages Index
Pages Edited by Me
Site Tools
IBM InfoSphere BigInsights
Bring the power of Hadoop to the enterprise.
Done

BigInsights Technical Enablement Wiki

Get up to speed on InfoSphere BigInsights, IBM's software platform designed to help firms store, manage, and analyze "big data".

Technical materials

- Articles and white papers
- BigInsights InfoCenter

Demos and videos

- Product demos
- IBM interviews and presentations

Downloads

- BigInsights Basic Edition (free)
- Fix packs for BigInsights Enterprise Edition (licensed)

Discussion Forum

- BigInsights forum on IBM developerWorks

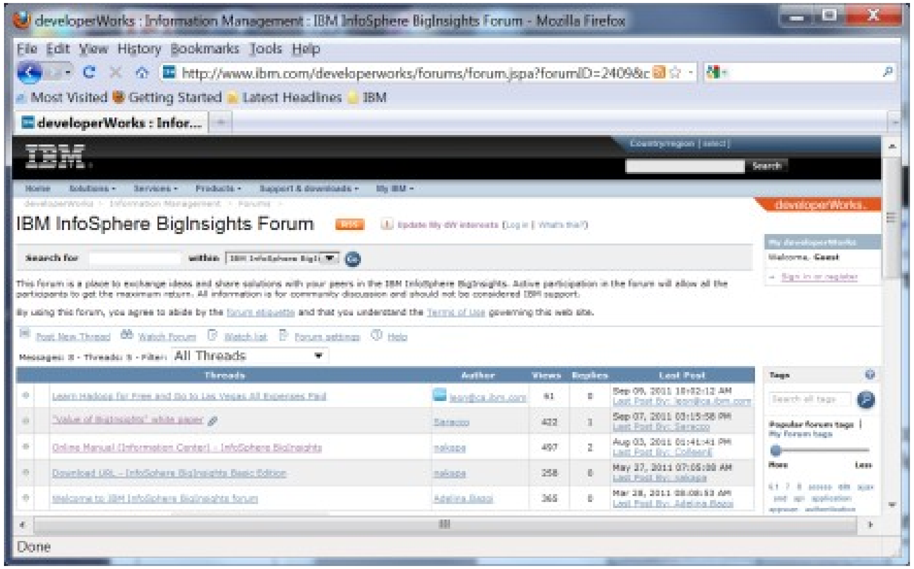
Education

- BigData University (free online courses)
- IBM Education (classroom, instructor-led online, and self-study courses)

News and announcements

- Recent news reports and press releases
- Announcement letters

Blogs and social media



developerWorks : Information Management : IBM InfoSphere BigInsights Forum - Mozilla Firefox

File Edit View History Bookmarks Tools Help
http://www.ibm.com/developerworks/forums/forum.jspa?forumID=2409&c
Most Visited Getting Started Latest Headlines IBM
developerWorks : Infor...
IBM
Country/region [select]
Search
Home Solutions Services Products Support & downloads My IBM
developerWorks > Information Management > Forums >
developerWorks.
RSS
Update My dW interests (Log in | What's this?)
Search for within IBM InfoSphere BigI
My developerWorks
Welcome, Guest
Sign in or register
Post New Thread Watch Forum Watch list Forum settings Help
Messages: 8 - Threads: 5 - Filter: All Threads
Tags
Search all tags
Popular forum tags | My forum tags
More Less
Done

IBM InfoSphere BigInsights Forum

This forum is a place to exchange ideas and share solutions with your peers in the IBM InfoSphere BigInsights. Active participation in the forum will allow all the participants to get the maximum return. All information is for community discussion and should not be considered IBM support.

By using this forum, you agree to abide by the forum etiquette and that you understand the Terms of Use governing this web site.

Threads	Author	Views	Replies	Last Post
Learn Hadoop for Free and Go to Las Vegas All Expenses Paid	leon@ca.ibm.com	61	0	Sep 09, 2011 10:02:12 AM Last Post By: leon@ca.ibm.com
"Value of BigInsights" white paper	Saracco	422	1	Sep 07, 2011 03:15:58 PM Last Post By: Saracco
Online Manual (Information Center) - InfoSphere BigInsights	nakaga	497	2	Aug 03, 2011 01:41:41 PM Last Post By: Colleen
Download URL - InfoSphere BigInsights Basic Edition	nakaga	258	0	May 27, 2011 07:05:00 AM Last Post By: nakaga
Welcome to IBM InfoSphere BigInsights forum	Adelina Bogo	365	0	Mar 28, 2011 08:08:53 AM Last Post By: Adelina Bogo

IBM big data • IBM big data • IBM big data

IBM big data • IBM big data

THINK

BIG

IBM big data • IBM big data

IBM big data • IBM big data • IBM big data

Supplemental

Sampling of MapReduce Use Cases

- **Extracted from public Web sites . . . .**
  - AOL: advanced algorithms for doing behavioral analysis and targeting
  - Detikcom (Indonesian portal): analyze search log, generate Most Viewed News
  - eBay: Search optimization, research
  - Facebook: store copies of internal log and dimension data sources and use it as a source for reporting/analytics and machine learning.
  - Financial institutions: determine credit worthiness for loan applicants – review changes in buying behaviors, etc.
  - LinkedIn: determine “People you may know”
  - Tennessee Valley Authority: Analyze electrical power sensor data to better predict power failures
  - Web advertisers: analyze historical click stream data, determine better ad choices

Sample Scenarios for Internet-Scale Analytics

Financial Services

- Improved risk decisions
- Customer sentiment analysis
- AML

Utilities

- Weather impact analysis on power generation
- Smart meter data analysis

Transportation

- Weather and traffic impact on logistics and fuel consumption

IT

- Transition log analysis for multiple transactional systems

Call Centers

- Voice-to-text mining for customer behavior understanding

E Commerce

- Analyze internet behavior and buying patterns
- Digital asset piracy

Telecommunications

- Operations and failure analysis from device, sensor, and GPS inputs

Multi-channel Integration

- Integrated customer behavior modeling

What is Hadoop?

- **Apache Hadoop = free, open source framework for data-intensive applications**
  - Inspired by Google technologies (MapReduce, GFS)
  - Well-suited to batch-oriented, read-intensive applications
  - Originally built to address scalability problems of Nutch, an open source Web search technology

- **Enables applications to work with thousands of nodes and petabytes of data in a highly parallel, cost effective manner**
  - CPU + disks of commodity box = Hadoop “node”
  - Boxes can be combined into clusters
  - New nodes can be added as needed without changing
    - Data formats
    - How data is loaded
    - How jobs are written

Two Key Aspects of Hadoop

- **MapReduce framework**
  - How Hadoop understands and assigns work to the nodes (machines)
- **Hadoop Distributed File System = HDFS**
  - Where Hadoop stores data
  - A file system that spans all the nodes in a Hadoop cluster
  - It links together the file systems on many local nodes to make them into one big file system

What is the Hadoop Distributed File System?

- **HDFS stores data across multiple nodes**
- **HDFS assumes nodes will fail, so it achieves reliability by replicating data across multiple nodes**
- **The file system is built from a cluster of *data nodes*, each of which serves up blocks of data over the network using a block protocol specific to HDFS.**

How To Create MapReduce Jobs

- **MapReduce development in Java**
- **Pig**
  - Open source language / Apache sub-project
- **Hive**
  - Open source language / Apache sub-project
  - Provides a SQL-like interface to Hadoop
- **Jaql**
  - IBM Research Invented query language
  - Very useful for loosely structured data
- **. . .**

Limitations with Apache Hadoop (examples)

- **Need to "roll your own" or "deal with multiple suppliers"**
  - Iteratively install, configure, and test Hadoop and complementary projects
  - Verify software pre-requisites and project versions for compatibility
  - Add-your-own analytics
- **Pig/Hive (Languages)**
  - Limited support for nested objects, multi-level hierarchies
  - No built-in connectivity to commercial DBMSs
- **Storage: Hadoop Distributed File System (HDFS)**
  - NameNode = single point of failure
  - Limited POSIX compliance. Cannot run other applications on node.
  - Poor security at the file system
  - Poor performance with random reads and writes
- **Technical support via open source community (or your own experts)**

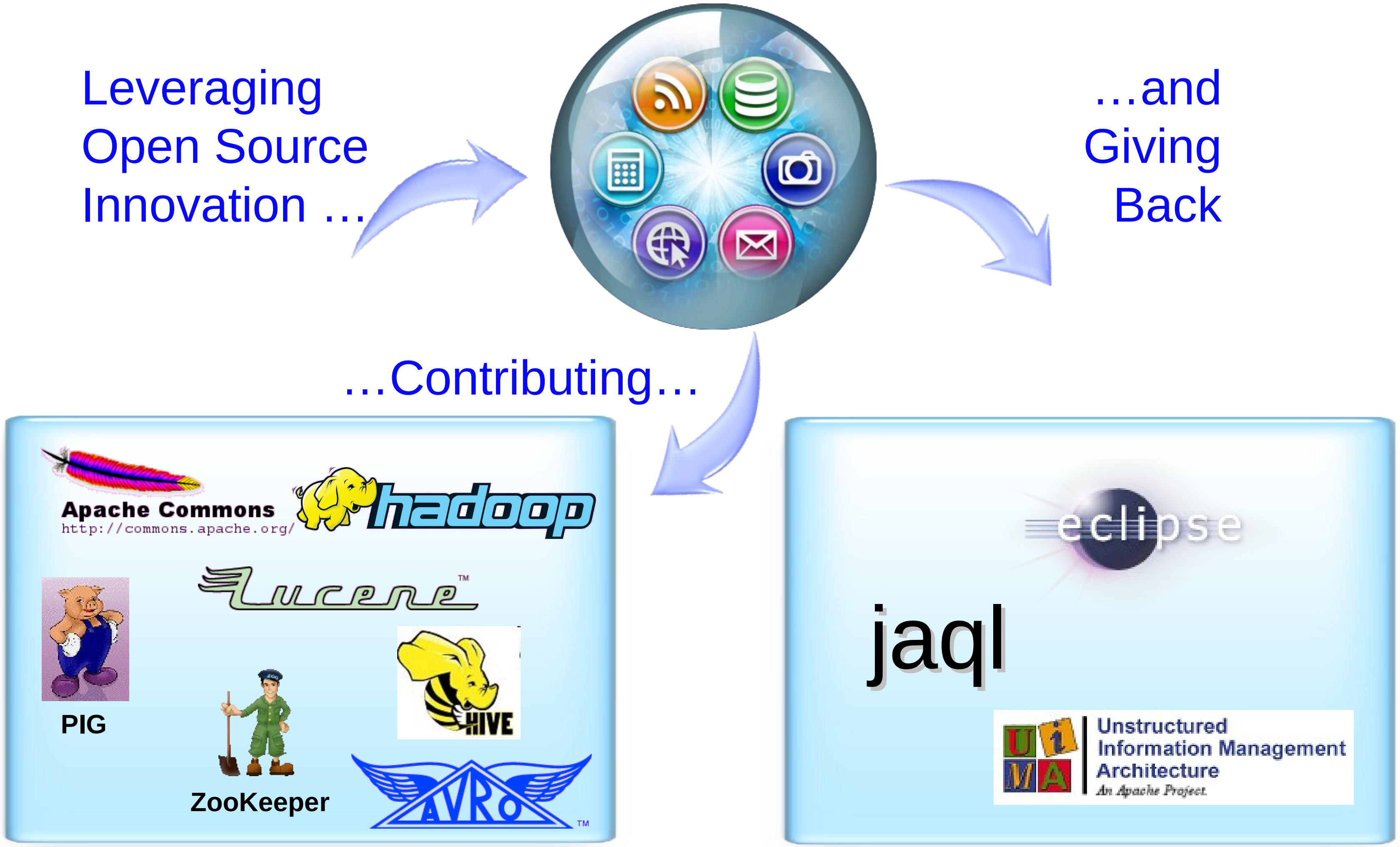
IBM: Building with the Open Source Community

Big Data Platform

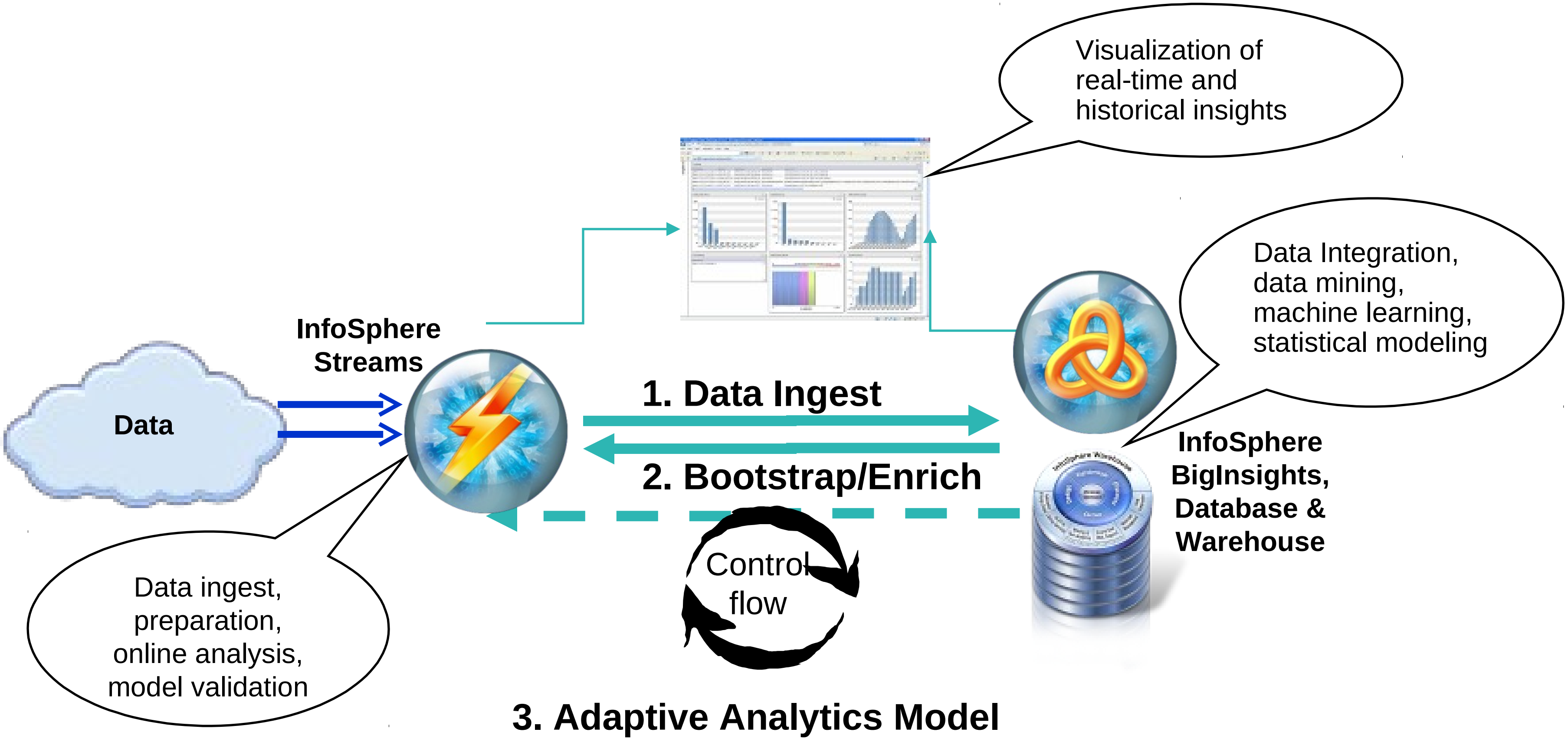
Leveraging
Open Source
Innovation …

…and
Giving
Back

…Contributing…



Streams and BigInsights - Integrated Analytics on Data in Motion & Data at Rest



IBM Watson

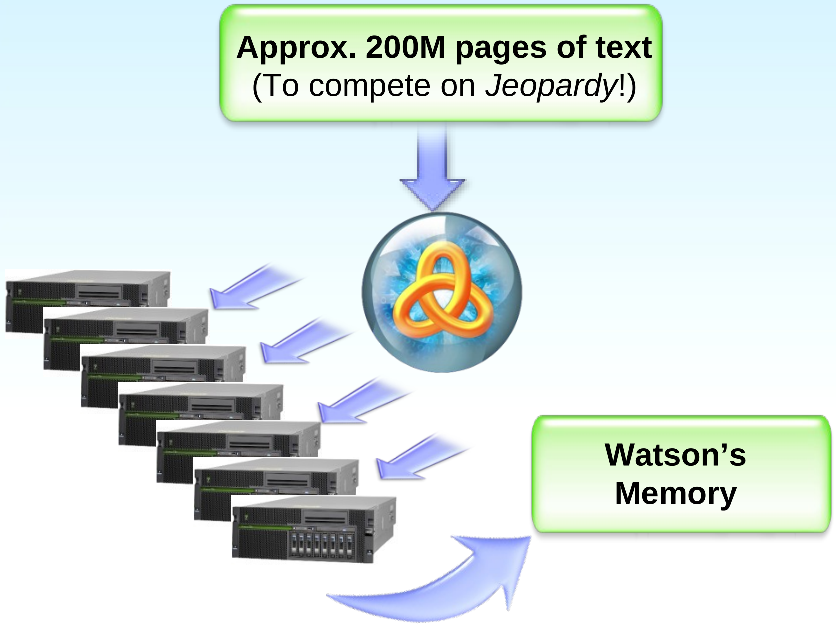
IBM Watson is a breakthrough in analytic innovation, but it is only successful because of the quality of the information from which it is working.

Big Data and Watson

Big Data technology is used to build Watson's knowledge base

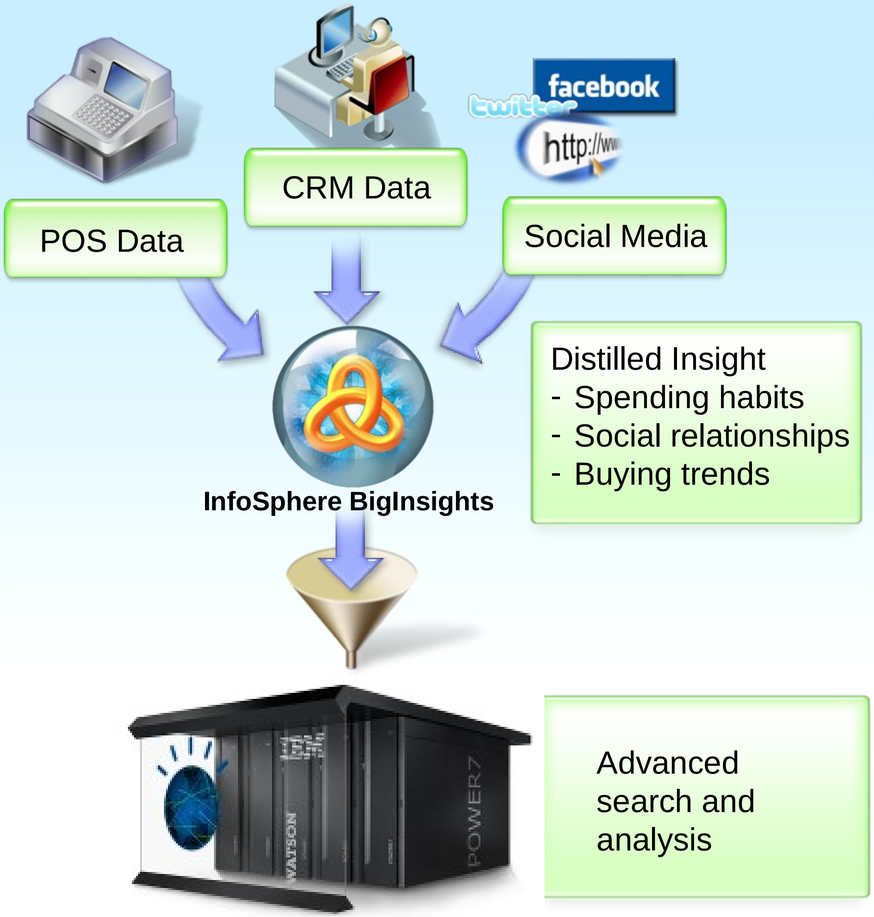
Watson technology offers great potential for advanced business analytics

Watson uses the Apache Hadoop open framework to distribute the workload for loading information into memory.



Approx. 200M pages of text
(To compete on *Jeopardy*!)

Watson's Memory



POS Data

CRM Data

Social Media

InfoSphere BigInsights

Distilled Insight
- Spending habits
- Social relationships
- Buying trends

Advanced search and analysis

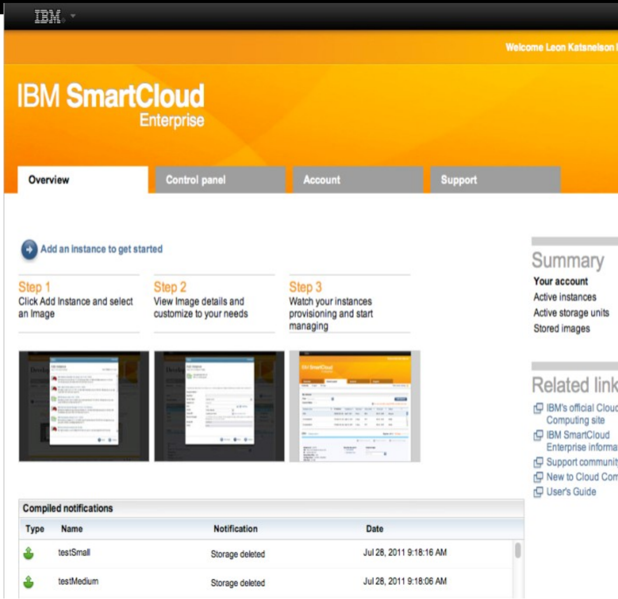
Sample hardware configuration

- **Hardware requirements vary by customer workload**

- **Reference hardware configuration for storage-dense or data-intensive workloads**
  - IBM System x3630
  - Two 6-core processors with 24TB local attached storage, 24GB RAM, Gigabit network
  - Replication factor of 3 for fault tolerance and distributed processing

BigInsights and the Cloud

- IBM SmartCloud Enterprise
- Amazon, Rackspace clouds through RightScale.com
- Low hourly charges



Welcome Leon Katsnelson

IBM SmartCloud Enterprise

Add an instance to get started

Step 1
Click Add Instance and select an Image

Step 2
View Image details and customize to your needs

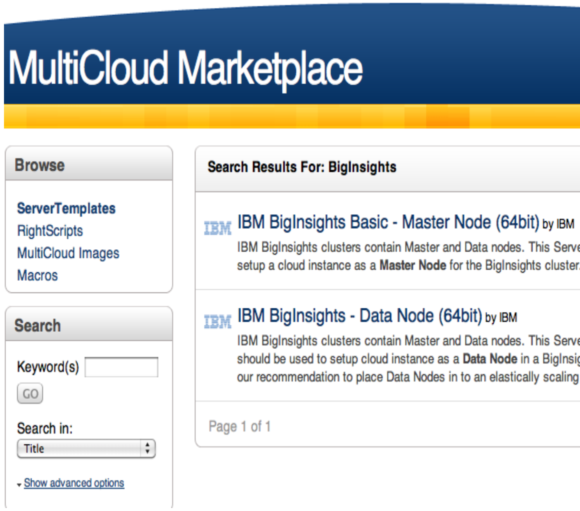
Step 3
Watch your instances provisioning and start managing

Summary
Your account
Active instances
Active storage units
Stored images

Related link
IBM's official Cloud Computing site
IBM SmartCloud Enterprise informa
Support communit
New to Cloud Com
User's Guide

Compiled notifications

Type	Name	Notification	Date
	testSmall	Storage deleted	Jul 28, 2011 9:18:16 AM
	testMedium	Storage deleted	Jul 28, 2011 9:18:06 AM



MultiCloud Marketplace

Browse
ServerTemplates
RightScripts
MultiCloud Images
Macros

Search
Keyword(s)

Search in: Title

Search Results For: BigInsights

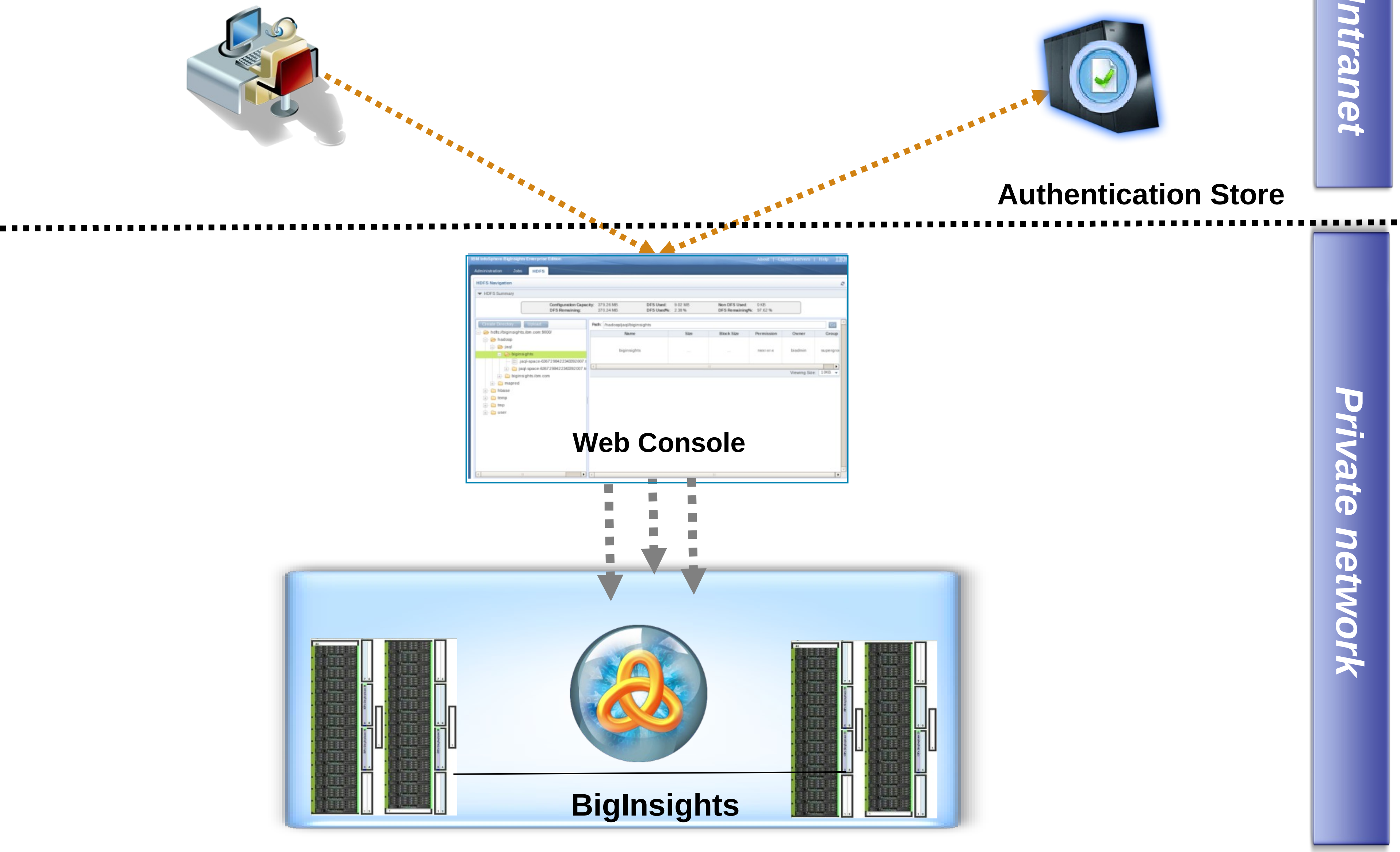
IBM BigInsights Basic - Master Node (64bit) by IBM
IBM BigInsights clusters contain Master and Data nodes. This Serve setup a cloud instance as a **Master Node** for the BigInsights cluster.

IBM BigInsights - Data Node (64bit) by IBM
IBM BigInsights clusters contain Master and Data nodes. This Serve should be used to setup cloud instance as a **Data Node** in a BigInsig our recommendation to place Data Nodes in to an elastically scaling

Page 1 of 1

IM Cloud Computing Center of Competence IMcloud@ca.ibm.com

BigInsights Secure Architecture



Authentication Store

Web Console

BigInsights

Intranet

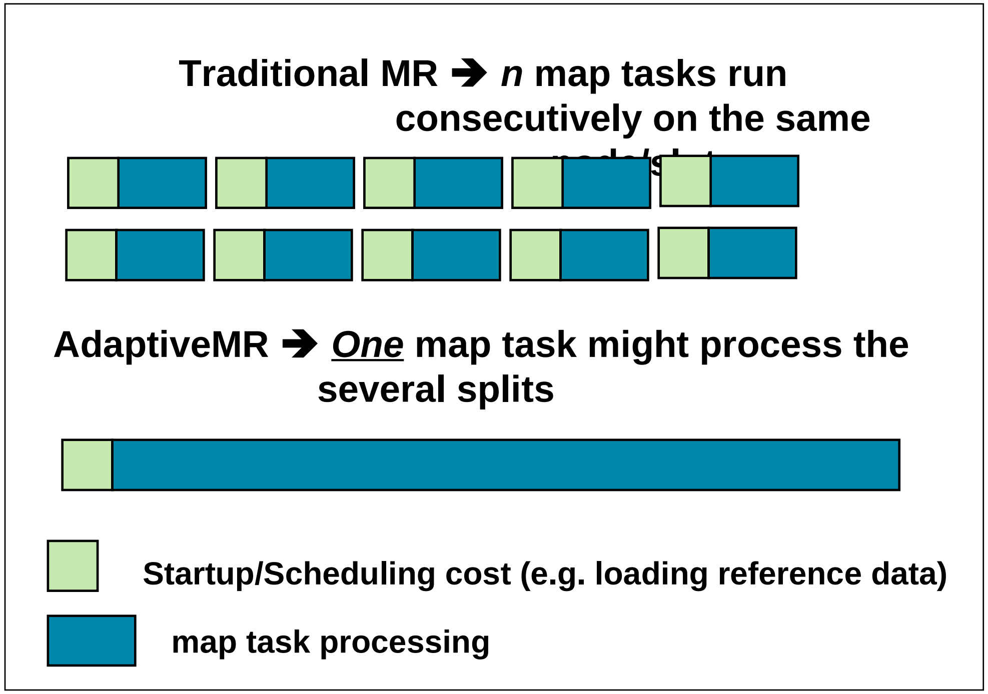
Private network

Security enhancements

	Apache Hadoop	BigInsights
Secure access	▪ Over 30 ports open ▪ Ports on every node open ▪ Clients outside cluster must have the same level of Hadoop libraries for RPC access	▪ Clients outside cluster use REST HTTP access ▪ Console serves as cluster gateway – Secure access via LDAP authenticated HTTP – Open up minimal external ports (8080 for Console). Block off other ports – Reverse proxy support: Console retrieves resources from cluster servers. Details hidden from requesting clients.
Authentication	▪ Dependency on client authentication. ▪ Easy to impersonate other users via config file (hadoop-policy.xml) or parms (hadoop.job.ugi)	▪ Offers LDAP authentication ▪ Users guided through config process by GUI-based installer
Authorization	▪ Unix-like file permissions	▪ Adds 4 built-in roles with distinct privileges ▪ Enforced through web console access

Design Goals of Adaptive MapReduce . . . .

- Balance workload across Map tasks
- Minimize startup and scheduling costs
  - Relatively high when operating on small files or splits
- Promote greater local aggregation
- Allow Map tasks to take on additional work until it doesn’t make sense anymore



Traditional MR ➔ *n* map tasks run consecutively on the same node/slot

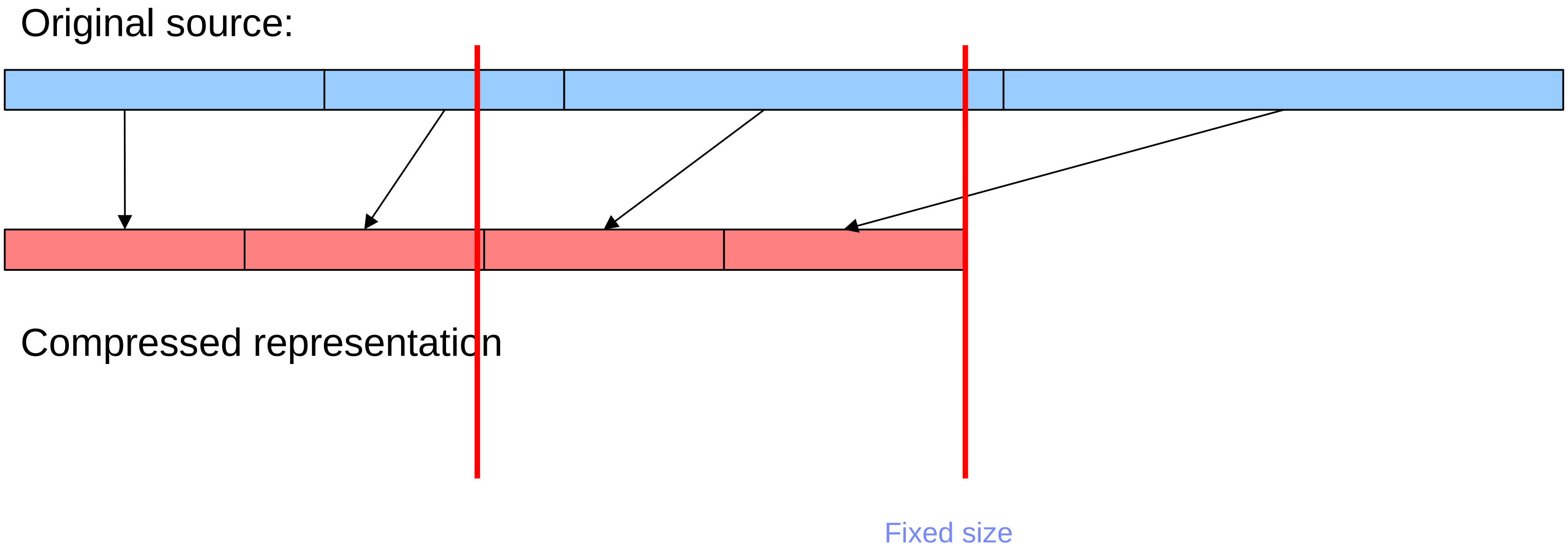
AdaptiveMR ➔ *__One__* map task might process the several splits

🟩 Startup/Scheduling cost (e.g. loading reference data)

🟦 map task processing

About IBM’s LZO-based compression

- Similar to GNU-based LZO compression, but no index needed
- Fixed-size compression blocks automatically created



Comparison of general compression technologies (conducted by third party)

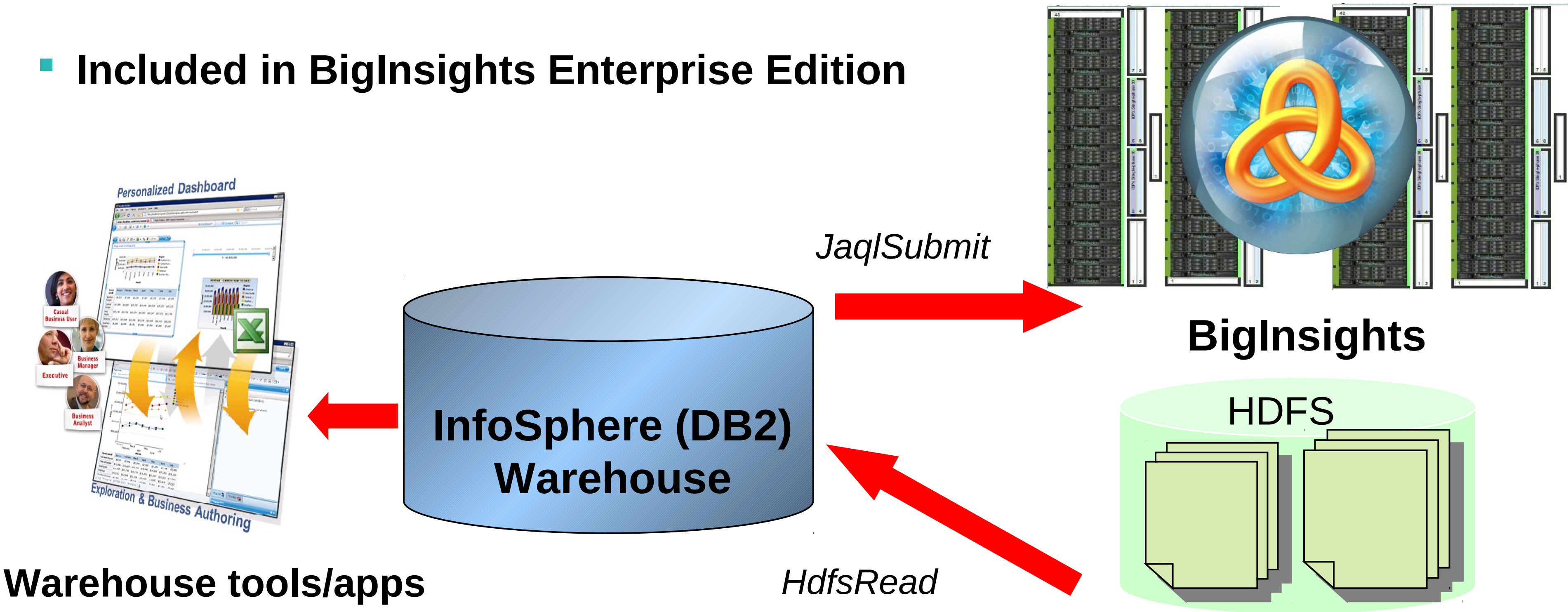
	Size (Mbytes)	Compression speed (sec)	Compression memory used (MBytes)	Decompression speed	Decompression memory used (Mbytes)
uncompressed	96				
gzip	23	10	.7	1.3	.5
bzip2	19	22	8	5	4
lzo	36	1	1	.6	0
lzm	18	63	14	3	1.8

Approximate values from

http://stephane.lesimple.fr/wiki/blog/lzop_vs_compress_vs_gzip_vs_bzip2_vs_lzma_vs_lzma2-xz_benchmark_reloaded

DB2 / InfoSphere Warehouse and BigInsights

- **Two sample DB2 UDFs to submit JAQL jobs, consume results**
  - *JaqlSubmit* to (pass parameters from DB2) and run analysis on BigInsights using Jaql
  - *HdfsRead* to transfer analysis results from BigInsights into DB2
  - Synchronous operations
- **Support provided for DB2 LUW**
- **Included in BigInsights Enterprise Edition**



JaqlSubmit

BigInsights

InfoSphere (DB2) Warehouse

HdfsRead

Warehouse tools/apps

About DBMS connectivity and Jaql. . . .

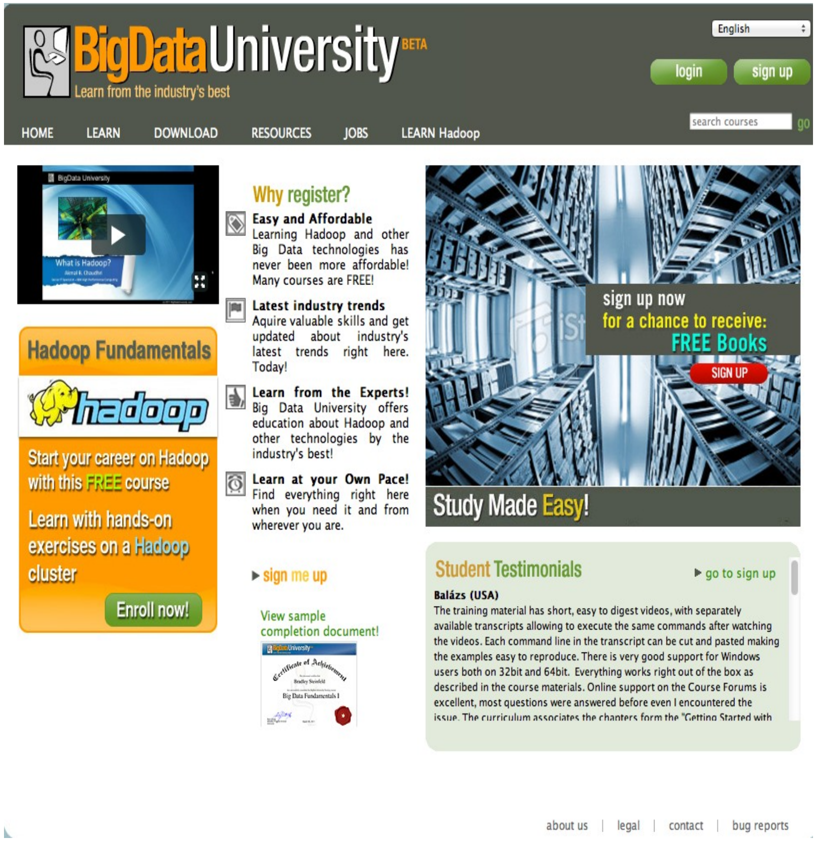
- **Jaql JDBC modules**
  - DB2 LUW and InfoSphere Warehouse
  - Netezza
  - Generic JDBC data source

- **Parallelism**
  - Multiple JDBC connections per job (one connection per Map task)
  - Native partitioning leveraged with Netezza
  - Distribution key column look up for DB2 LUW with DPF (automatic)

- **Writing data from BigInsights to RDBMS**
  - No ACID guarantee (no transactions in MapReduce)
  - Failed MapReduce tasks automatically restart – could lead to duplicate rows (repeated inserts) in DBMS
  - Consider writing BigInsights data to DBMS temp table. If no Map restarts in the job, copy temp table contents to target DBMS table

BigInsights on the Cloud: Free Education

- Flexible on-demand delivery @ your pace
- Free study materials, labs
- Cloud-based sandbox for hands-on work – no setup
- 8500+ registered students as of Oct. 2011



Start your career on Hadoop with this FREE course

Learn with hands-on exercises on a Hadoop cluster

Why register?

Easy and Affordable
Learning Hadoop and other Big Data technologies has never been more affordable! Many courses are FREE!

Latest industry trends
Aquire valuable skills and get updated about industry's latest trends right here. Today!

Learn from the Experts!
Big Data University offers education about Hadoop and other technologies by the industry's best!

Learn at your Own Pace!
Find everything right here when you need it and from wherever you are.

Student Testimonials

Balázs (USA)
The training material has short, easy to digest videos, with separately available transcripts allowing to execute the same commands after watching the videos. Each command line in the transcript can be cut and pasted making the examples easy to reproduce. There is very good support for Windows users both on 32bit and 64bit. Everything works right out of the box as described in the course materials. Online support on the Course Forums is excellent, most questions were answered before even I encountered the issue. The curriculum associates the chapters form the "Getting Started with

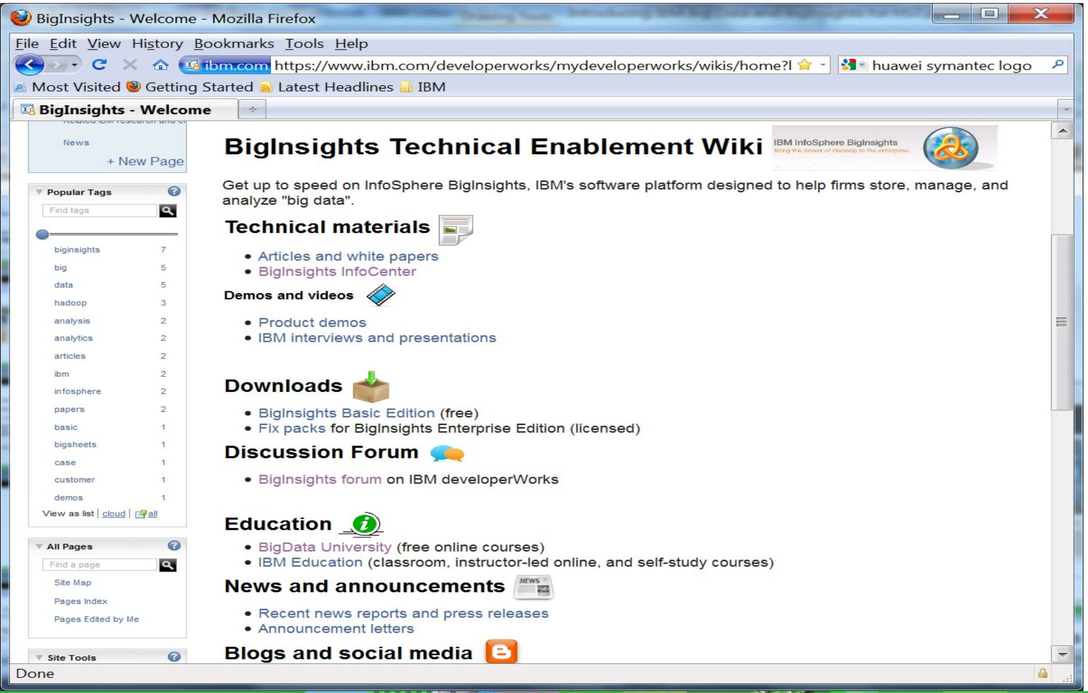
about us | legal | contact | bug reports

IM Cloud Computing Center of Competence IMcloud@ca.ibm.com

Technical portal for BigInsights

One-stop source for technical materials with links to papers, downloads, demos, education, discussion forum, and more

http://www.ibm.com/developerworks/wiki/biginsights/



BigInsights - Welcome - Mozilla Firefox

https://www.ibm.com/developerworks/mydeveloperworks/wikis/home?l

BigInsights Technical Enablement Wiki

Get up to speed on InfoSphere BigInsights, IBM's software platform designed to help firms store, manage, and analyze "big data".

Technical materials

- Articles and white papers
- BigInsights InfoCenter

Demos and videos

- Product demos
- IBM interviews and presentations

Downloads

- BigInsights Basic Edition (free)
- Fix packs for BigInsights Enterprise Edition (licensed)

Discussion Forum

- BigInsights forum on IBM developerWorks

Education

- BigData University (free online courses)
- IBM Education (classroom, instructor-led online, and self-study courses)

News and announcements

- Recent news reports and press releases
- Announcement letters

Blogs and social media

Popular Tags

biginsights	7
big	5
data	5
hadoop	3
analysis	2
analytics	2
articles	2
ibm	2
infosphere	2
papers	2
basic	1
bigsheets	1
case	1
customer	1
demos	1

View as list | cloud | all

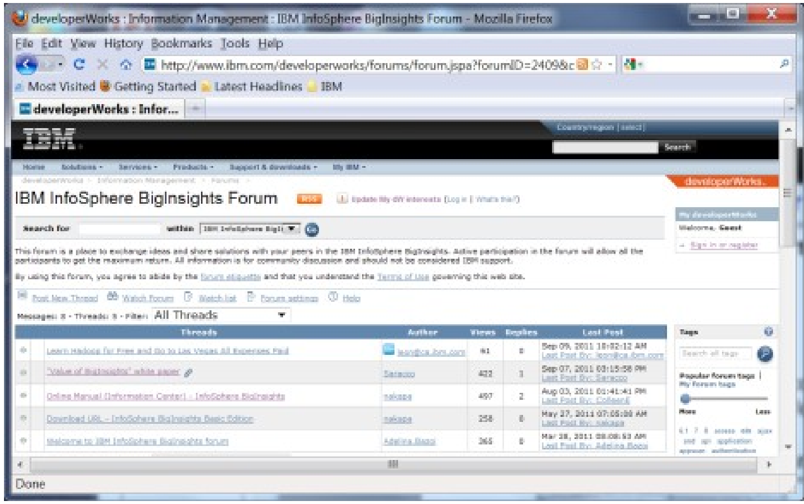
All Pages

Site Map
Pages Index
Pages Edited by Me

Site Tools

Done

File Edit View History Bookmarks Tools Help
huawei symantec logo
Most Visited Getting Started Latest Headlines IBM
BigInsights - Welcome
IBM InfoSphere BigInsights
News
+ New Page
Find tags
Find a page



developerWorks : Information Management : IBM InfoSphere BigInsights Forum - Mozilla Firefox

http://www.ibm.com/developerworks/forums/forum.jspa?forumID=2409&c

IBM InfoSphere BigInsights Forum

This forum is a place to exchange ideas and share solutions with your peers in the IBM InfoSphere BigInsights. Active participation in the forum will allow all the participants to get the maximum return. All information is for community discussion and should not be considered IBM support.

By using this forum, you agree to abide by the forum etiquette and that you understand the Terms of use governing this web site.

Messages: 8 · Threads: 5 · Filter: All Threads

Threads	Author	Views	Replies	Last Post
Learn Hadoop for Free and Go to Las Vegas All Expenses Paid	jeon@ca.ibm.com	61	0	Sep 09, 2011 10:02:12 AM Last Post By: jeon@ca.ibm.com
"Value of BigInsights" white paper	Saracco	422	1	Sep 07, 2011 03:15:58 PM Last Post By: Saracco
Online Manual (Information Center) - InfoSphere BigInsights	nakapa	497	2	Aug 03, 2011 01:41:41 PM Last Post By: Colleen
Download URL - InfoSphere BigInsights Basic Edition	nakapa	258	0	May 27, 2011 07:05:00 AM Last Post By: nakapa
Welcome to IBM InfoSphere BigInsights forum	Adelina Bagci	365	0	Mar 28, 2011 08:08:53 AM Last Post By: Adelina Bagci

Done

File Edit View History Bookmarks Tools Help
Most Visited Getting Started Latest Headlines IBM
developerWorks : Infor...
IBM
Country/region [select]
Search
Home Solutions Services Products Support & downloads My IBM
developerWorks > Information Management > Forums
developerWorks.
RSS
Update My dW interests (Log in | What's this?)
Search for
within IBM InfoSphere BigI
My developerWorks
Welcome, Guest
Sign in or register
Post New Thread
Watch Forum
Watch list
Forum settings
Help
Tags
Search all tags
Popular forum tags
My forum tags
More
Less