

# Big Data



**Paul Zikopoulos, BA, MBA**

Director, IBM Information Management Technical Professionals,

WW Competitive Database, and Big Data

IBM Certified Advanced Technical Expert (Clusters and DRDA)

IBM Certified Customer Solutions Expert (DBA and BI)

paulz\_ibm@msn.com

 @BigData\_paulz



**Paul C. Zikopoulos, B.A., M.B.A.,** is the Director of Technical Professionals for IBM Software Group's Information Management

division and additionally leads the World Wide Database Competitive and Big Data SWAT teams. Paul is an award-winning writer and speaker with more than 18 years of experience in Information Management. Paul has written more than 350 magazine articles and 15 books including *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, *DB2 pureScale: Risk Free Agile Scaling*, *Break Free with DB2 9.7: A Tour of Cost Saving Features*; *Information on Demand: Introduction to DB2 9.5 New Features*; *DB2 Fundamentals Certification for Dummies*; *DB2 for Dummies*; and more. Paul is a DB2 Certified Advanced Technical Expert (DRDA and Clusters) and a DB2 Certified Solutions Expert (BI and DBA). In his spare time, he enjoys all sorts of sporting activities, including running with his dog Chachi, avoiding punches in his MMA training, and trying to figure out the world according to Chloë—his daughter. You can reach him at: [paulz\\_ibm@msn.com](mailto:paulz_ibm@msn.com).

@BigData\_paulz



# Why Big Data How We Got Here







**In 2005 there were 1.3 billion RFID tags in circulation...**



**...by the end of 2011, this was about 30 billion and growing even faster**





An increasingly sensor-enabled and instrumented business environment generates **HUGE** volumes of data with **MACHINE SPEED** characteristics...

**1 BILLION** lines of code  
**EACH** engine generating 10 TB every 30 minutes!



**350B**  
Transactions/Year

**Meter Reads**  
every 15 min.

120M – meter reads/month

3.65B – meter reads/day







Read the full story here: <http://nyti.ms/917h>

- In August of 2010, Adam Savage, of “Myth Busters,” took a photo of his vehicle using his smartphone. He then posted the photo to his Twitter account including the phrase “Off to work.”
- Since the photo was taken by his smartphone, the image contained metadata revealing the exact geographical location the photo was taken
- By simply taking and posting a photo, Savage revealed the exact location of his home, the vehicle he drives, and the time he leaves for work

# The Social Layer in an Instrumented Interconnected World

**12+ TBs**  
of tweet data  
every day



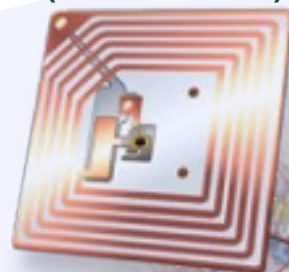
? TBs of  
data every day



**25+ TBs** of  
log data  
every day



**30 billion** RFID  
tags today  
(1.3B in 2005)



**4.6 billion**  
camera  
phones  
world  
wide



**100s of millions**  
of GPS  
enabled  
devices  
sold  
annually



**76 million** smart  
meters in 2009...  
200M by 2014

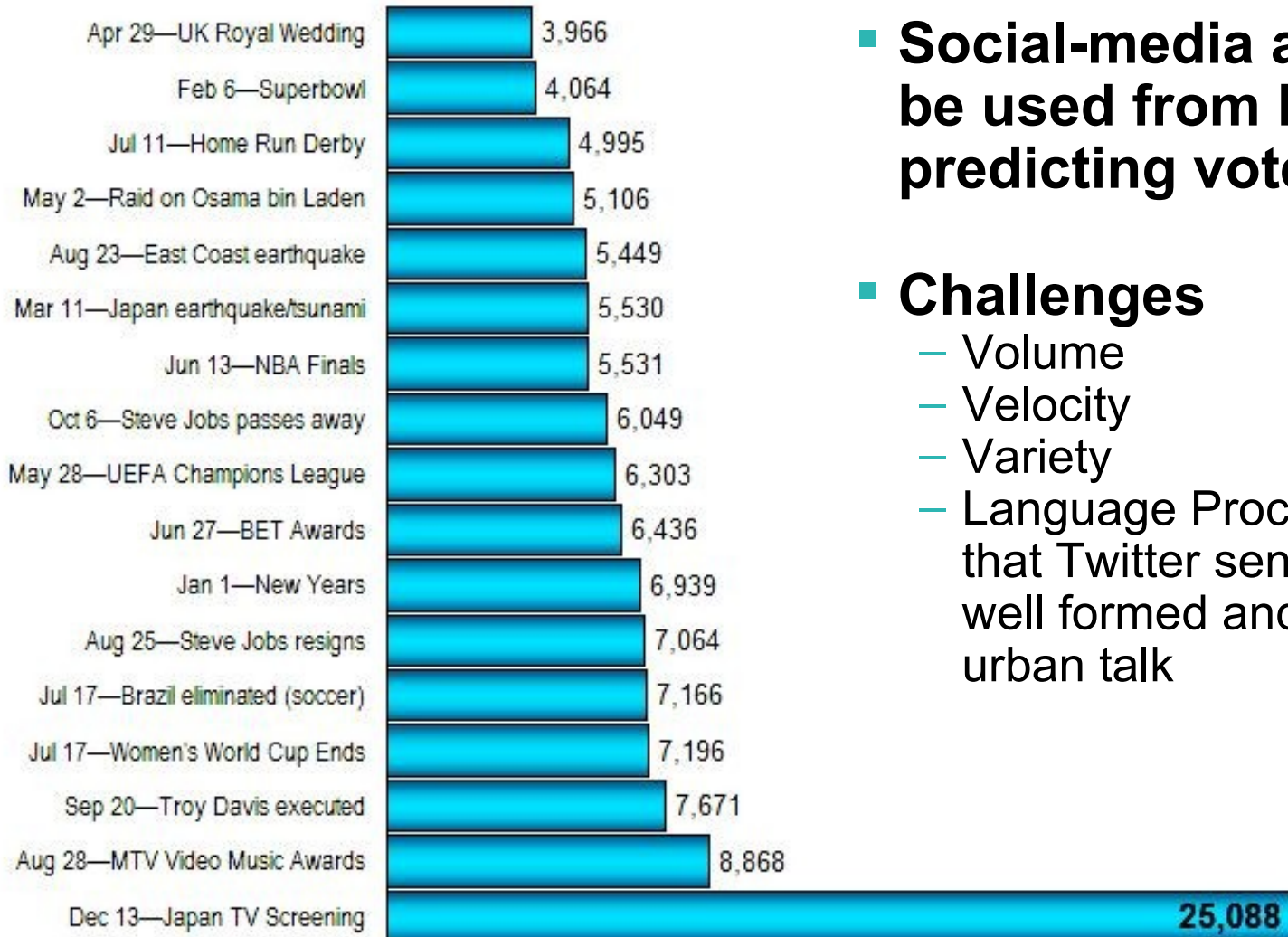
http



**2+ billion**  
people  
on the  
Web by  
end 2011



# Twitter Tweets per Second Record Breakers of 2011



- **Social-media analytics can be used from healthcare to predicting votes**
- **Challenges**
  - Volume
  - Velocity
  - Variety
  - Language Processing: consider that Twitter sentences are not well formed and often use urban talk

# Can a Social Media Persona be Monetized?



THE WALL STREET JOURNAL | ARTS & ENTERTAINMENT

DECEMBER 28, 2011, 7:00 PM ET

## Man Leaves Job, Takes Twitter Followers, Gets Sued

*"This will establish precedent in the online world, as it relates to ownership of social media accounts. We've actually been waiting to see such a case as many of our clients are concerned about the ownership of social media accounts vis-a-vis their branding."*



# Extract Intent, Life Events, Micro Segmentation Attributes

A screenshot of a Facebook interface. On the left is a navigation menu with 'FAVORITES' (News Feed, Messages, Other, Events) and 'APPS' (Pokes, Photos, Apps and Games). The main area shows four profile pictures with corresponding labels in black boxes with teal text:

- Profile picture of a woman wearing a yellow hat: **Name, Birthday, Family**
- Profile picture of a black and white dog: **Not Relevant - Noise**
- Profile picture of a family of four: **Monetizable Intent**
- Profile picture of an Angry Bird character: **Not Relevant - Noise**

A screenshot of a Twitter interface header. It includes the Twitter logo, a search bar, and navigation tabs for 'Home', 'Profile', 'Messages', and 'Who To Follow'.

What's happening?

**Location**

What's happening?

**Wishful Thinking**

What's happening?

**Relocation**

What's happening?

**SPAMbots**



**John Rill, Steve Alexander and Bill Hitchon like Tough Mudder.**



**Tough Mudder**  
Like

**IKEA Canada asked: My #1 source for bathroom reno ideas is:**

- Friends and family
- Reality reno shows
- World Wide Wonder We

Like This Page

**Christina Steenberg likes Travel Alberta.**



**Travel Alberta**  
Like

**Mercedes-Benz Canada**  
You like this



Like- Michelle Maria Codner likes this.

**Stephen Michael O'Grady likes Buffalo.**



**Buffalo**  
Like

**Stephen Michael O'Grady and Helen Stoumbos like Wine Country Ontario.**



**Wine Country Ontario**  
Like

**Aiichiro Noma likes The Boeing Store.**



**The Boeing Store**  
Like

**Lorraine Evans likes Taco Bell Canada.**



**Taco Bell Canada**  
Like

**1 donor can save 8 lives.**



Ontarians are waiting for an organ transplant. Register your consent now to become an organ and tissue donor.

1,478 people like Trillium Gift of Life Network.

2011

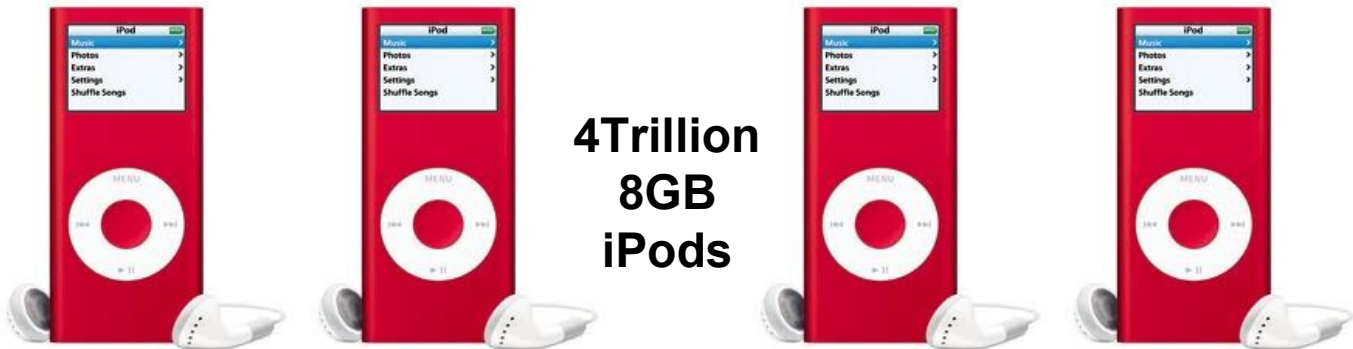
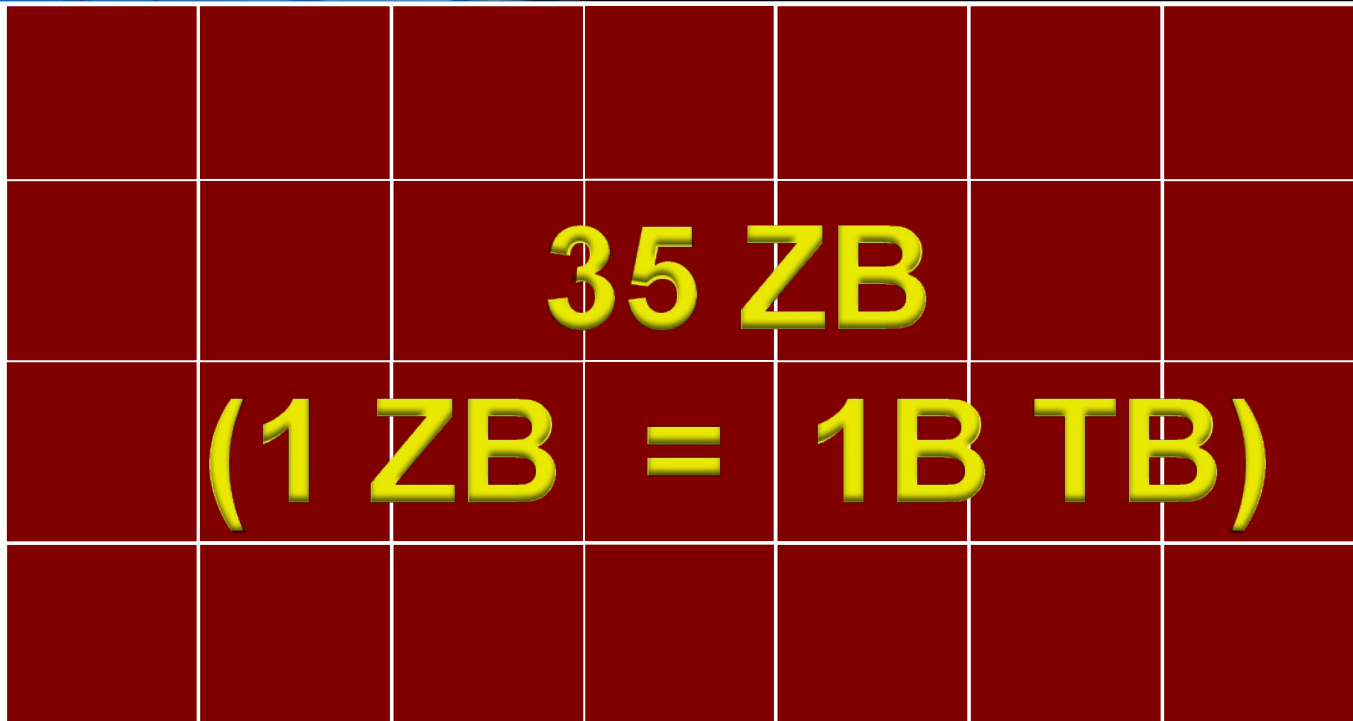
1.8 ZB

2009

1 ZB

1 ZB=1T GB

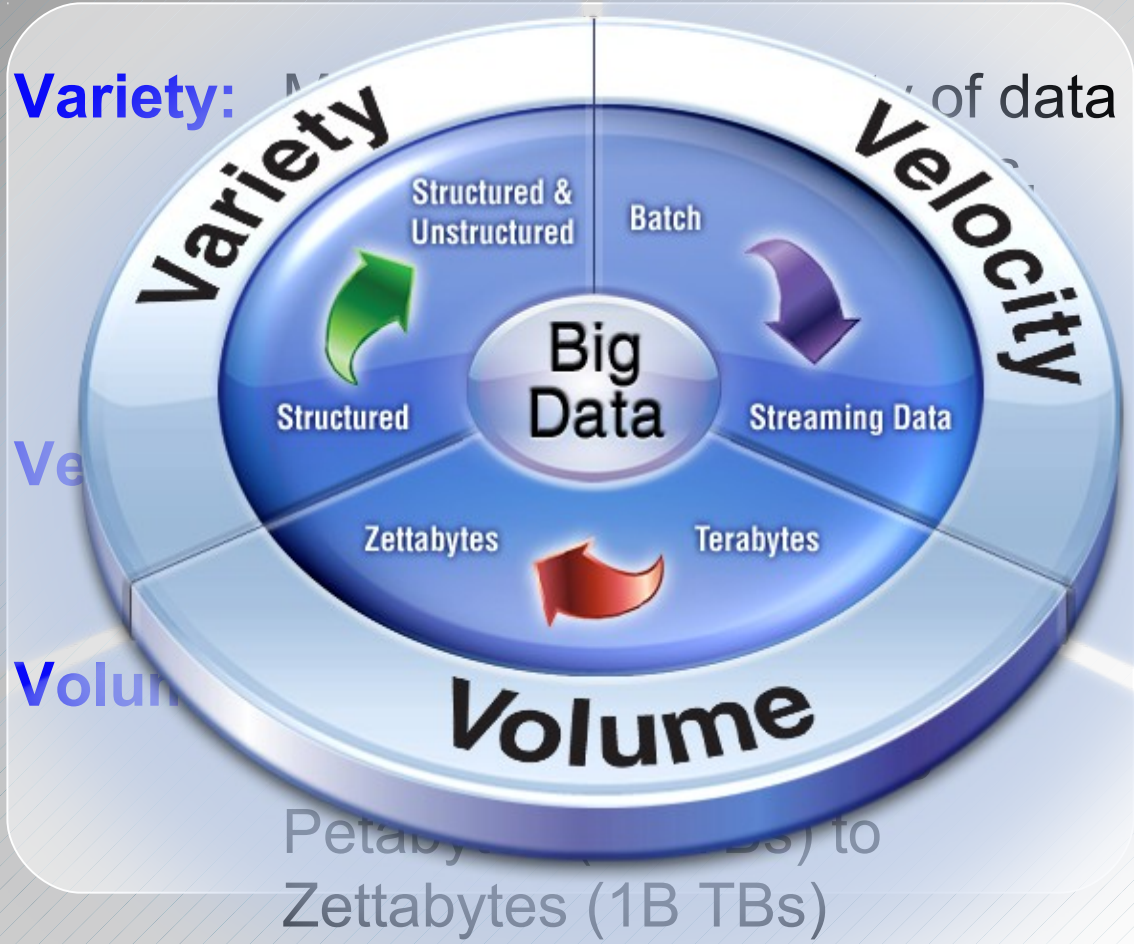
2010  
2020





# The Big Data Opportunity

*Extracting insight from an immense volume, variety and velocity of data, in context, beyond what was previously possible.*



# Data In Motion

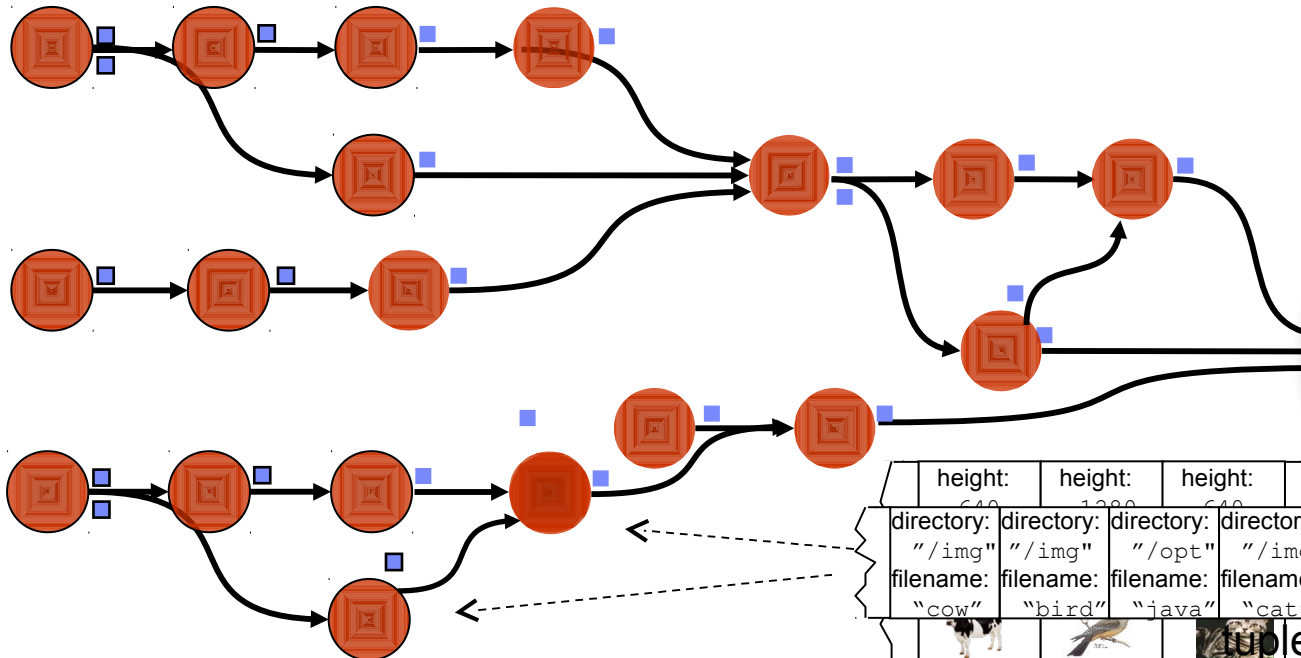
Algorithmic Trading *Forum*

Analyzes and correlates 5M+ market messages/sec to execute algorithmic option trades with average latency of 30 micro-secs.

ipdr

500K/sec, 6B+ IPDRs analyzed per day on more than 4 PBs/yr. sustaining 1GBps.

## Consider: Data that is never stored, never has to be subjected to retention policies: COST SAVINGS



# The Big Data Conundrum

- The economies of deletion have changed....
  - Leading us into new opportunities and challenges
- The percentage of available data an enterprise can analyze is decreasing proportionately to the available to that enterprise
- Quite simply, this means as enterprises, we are getting “more naive” about our business over time

Data AVAILABLE to  
an organization

Signals  
and  
Noise



Data an organization  
can PROCESS



# Applications for Big Data Analytics

Smarter Healthcare



Multi-channel



Finance



Log Analysis



Homeland Security



Traffic Control



Telecom



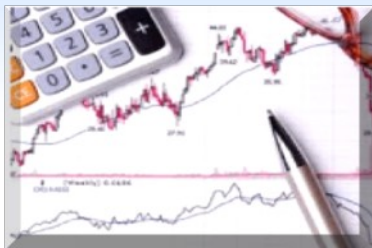
Search Quality



Manufacturing



Trading Analytics



Fraud and Risk



Retail: Churn, NBO



# Bigger and Bigger Volumes of Data

- **Retailers collect click-stream data from Web site interactions and loyalty card-drive transaction data**
  - This traditional POS information is used by retailer for shopping basket analysis, inventory replenishment, +++
  - But data is being provided to suppliers for customer buying analysis
- **Healthcare has traditionally been dominated by paper-based systems, but this information is getting digitized**
- **Science is increasingly dominated by big science initiatives**
  - Large-scale experiments generate over 15 PB of data a year and can't be stored within the data center; then sent to laboratories
- **Financial services are seeing larger volumes through smaller trading sizes, increased market volatility, and technological improvements in automated and algorithmic trading**
- **Improved instrument and sensory technology**
  - Large Synoptic Survey Telescope's GPixel camera generates 6PB+ of image data per year or consider Oil and Gas industry

# A Simple Log File

```
Jun 29 04:03:37 192.168.190.24 nagios: GLOBAL SERVICE EVENT HANDLER: ast-oral-
vip;remedy_oracle_db_latch-contention; (null); (null); (null); handle-all-
critical-event
```

```
DATE      TIME      IP ADDRESS  MONITOR
Jun 29 04:03:57 192.168.190.24 nagios: SERVICE ALERT: ast-oral-
vip;remedy_oracle_db_soft-prase-ratio;CRITICAL;HARD;3;CRITICAL - Soft parse
ratio
86.63%
```

```

                                         TYPE      COMPONENT
Jun 29 07:50:57 192.168.190.24 nagios: SERVICE ALERT: ast-oral
vip;remedy_oracle_db_latch-waiting; OK;SOFT;2;OK - SGA latch xssinfo freelist
(#350)
sleeping 0.000000% of the time
```

```
Jun 29 07:50:57 192.168.190.24 nagios: GLOBAL SERVICE EVENT HANDLER: ast-oral-
vip;remedy_oracle_db_latch-waiting; (null); (null); (null); handle-all-
critical-event
```



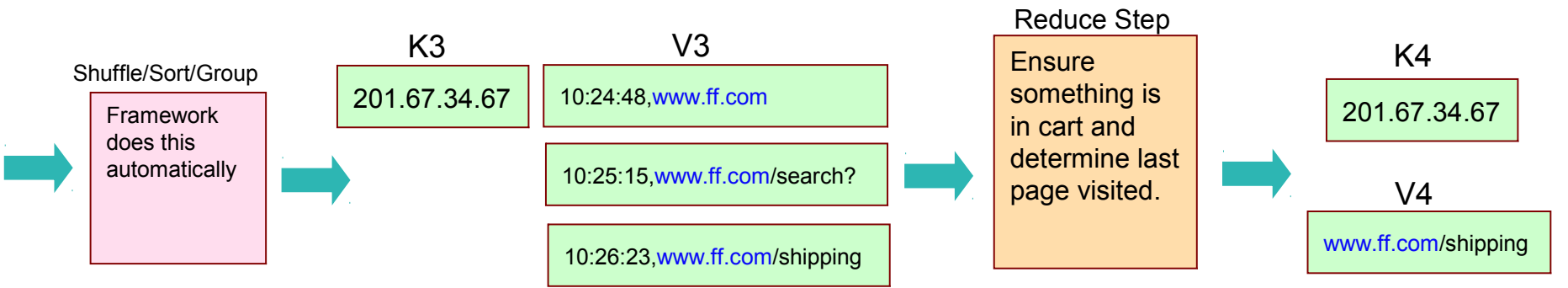
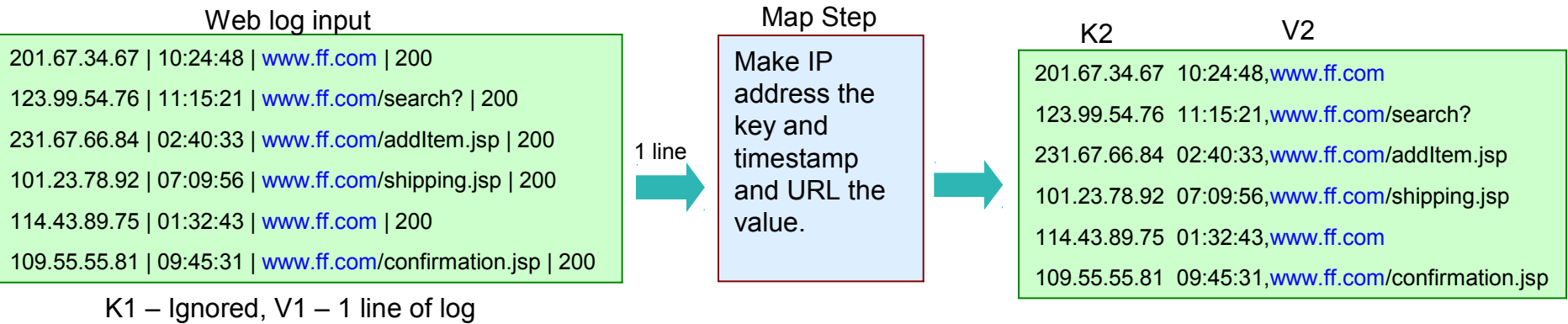


# Click Stream Analysis Use Case

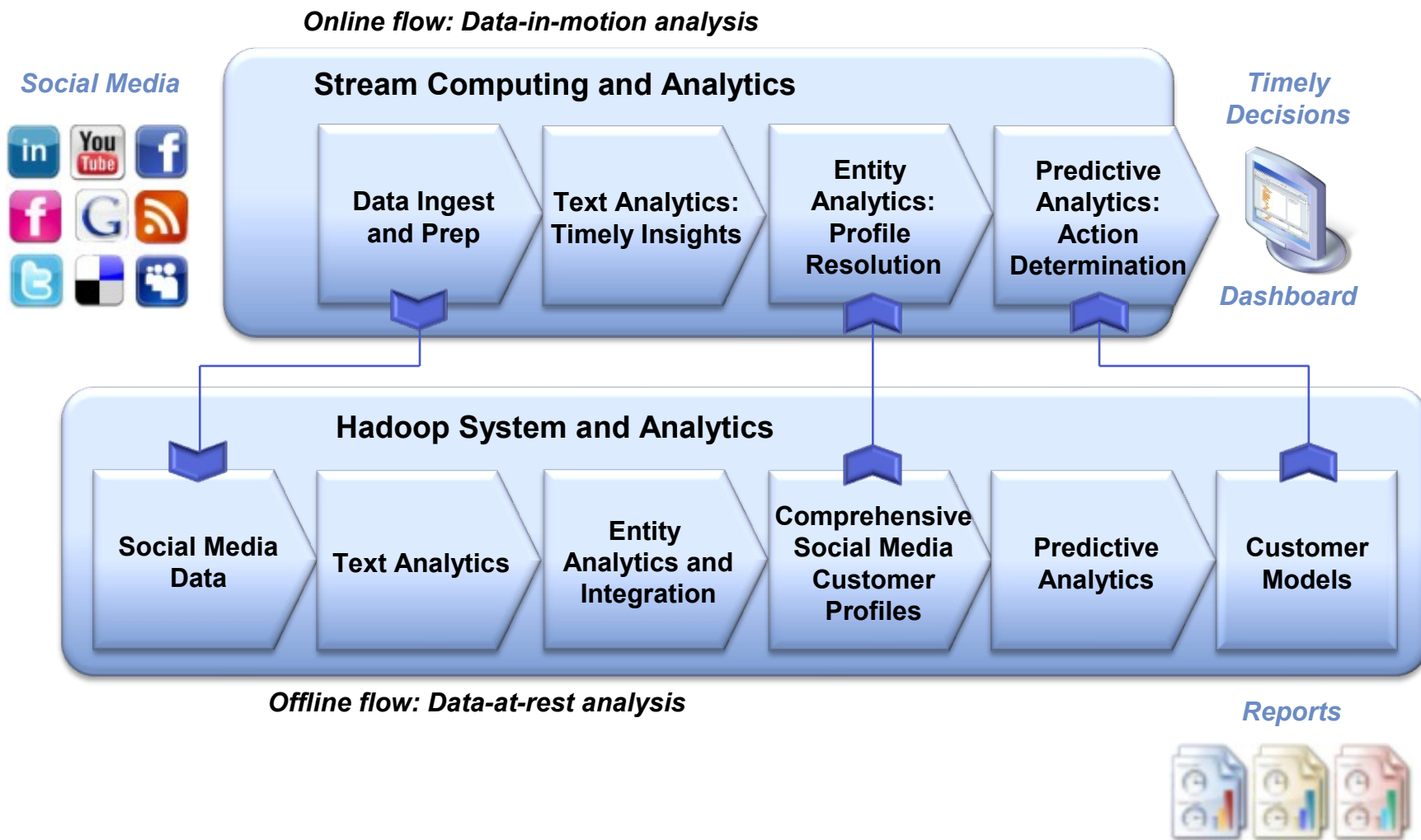
Goal: Determine how many abandoned shopping carts there are and where they were abandoned.

Input: Web log data (IP address, timestamp, URL, HTTP return codes).

Output: List of IP addresses and last page visited by user.



# Solution Architecture – Social Media Analytics for Media and Entertainment



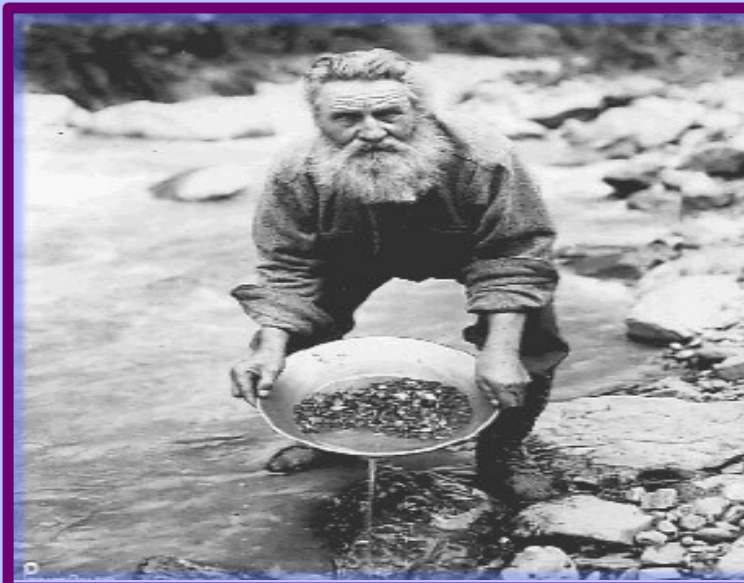


# Most Requested Uses of Big Data

- Log Analytics & Storage
- Smart Grid / Smarter Utilities
- RFID Tracking & Analytics
- Fraud / Risk Management & Modeling
- 360° View of the Customer
- Warehouse Extension
- Email / Call Center Transcript Analysis
- Call Detail Record Analysis
- ++++



# Why Didn't We Use All of the Big Data Before?



**Vestas**<sup>®</sup>

***“IBM gave us an opportunity to turn our plans into something that was very tangible right from the beginning. IBM had experts within data mining, Big Data, and Apache Hadoop and it was clear to use from the beginning we wanted to improve our business, not only today, but also prepare for the challenges we will face in three to five years, we had to go with IBM.”***

*– Lars Christian Christensen VP Plant Siting & Forecasting*





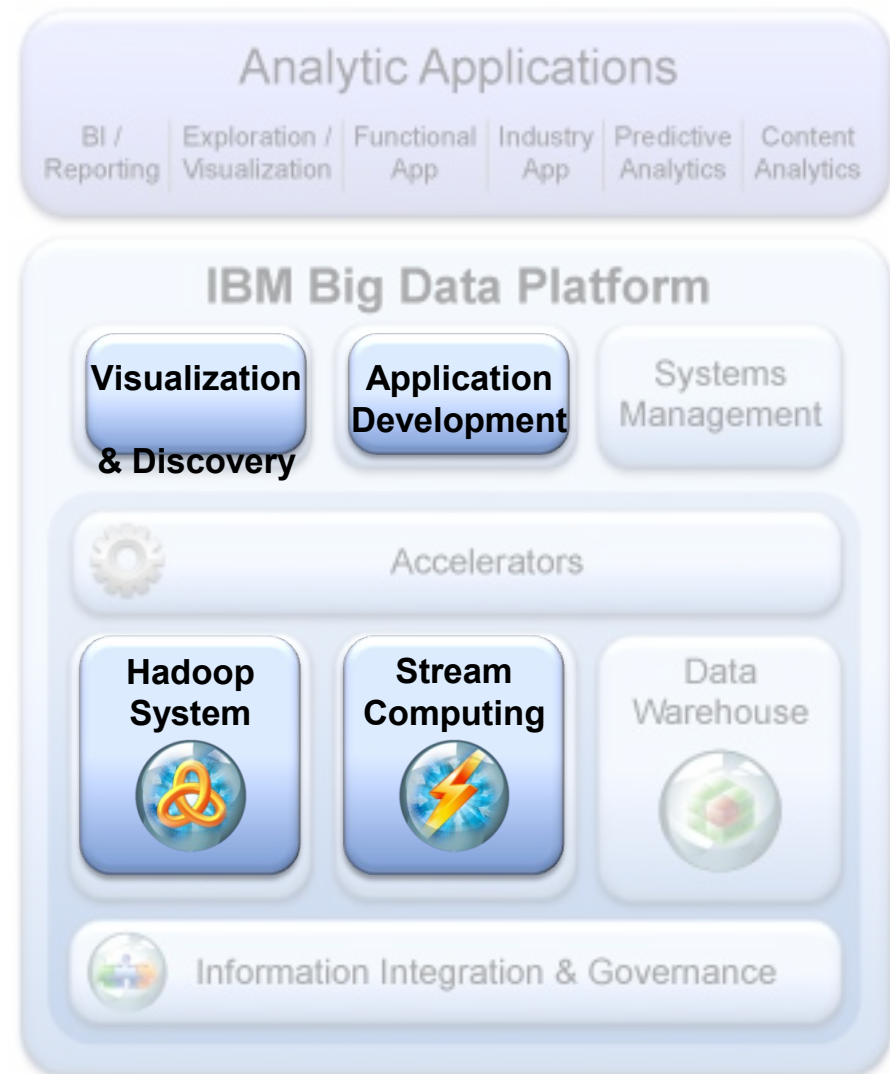




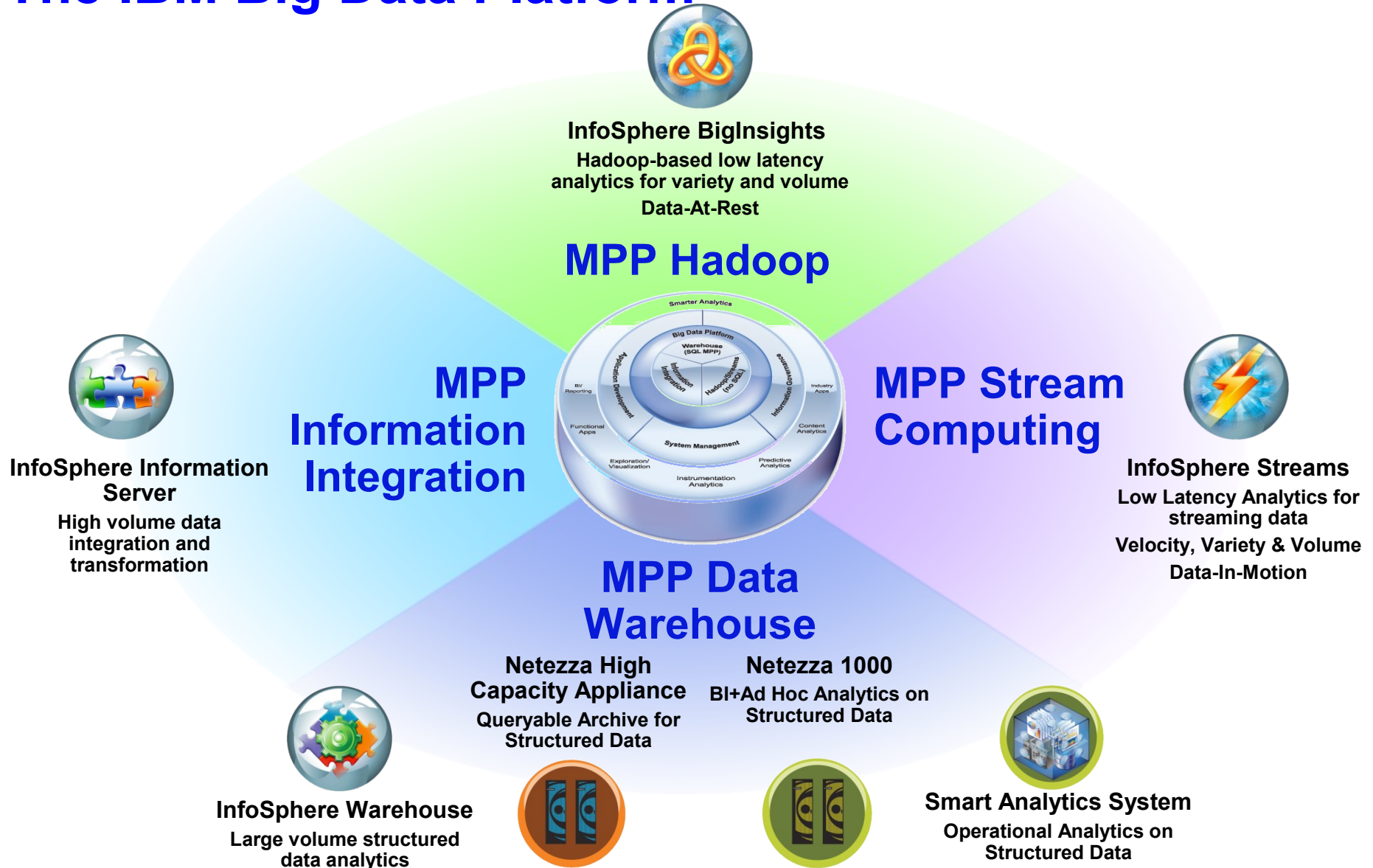
**Watson's advanced analytic capabilities can sort through the equivalent of 200 MILLION pages of data to uncover an answer in 3 SECONDS.**

# IBM Big Data Strategy: Move the Analytics Closer to the Data

- **New analytic applications drive the requirements for a big data platform**
  - Integrate and manage the full variety, velocity and volume of data
  - Apply advanced analytics to information in its native form
  - Visualize all available data for ad-hoc analysis
  - Development environment for building new analytic applications
  - Workload optimization and scheduling
  - Security and Governance



# The IBM Big Data Platform





# Deep Analytics Appliance – Revolutionized Analytics

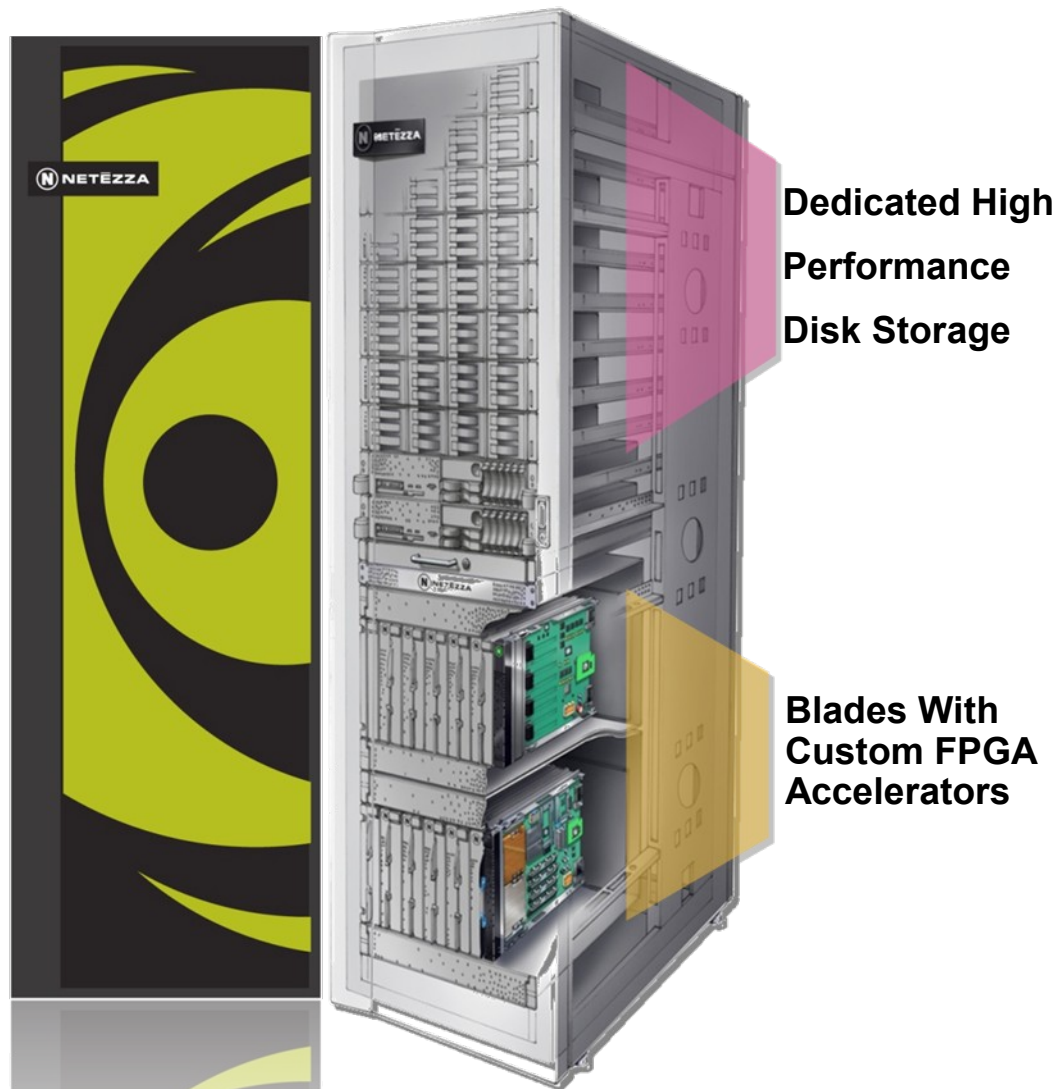
## Purpose-built analytics appliance

**Speed:** 10-100x faster than traditional systems

**Simplicity:** Minimal administration and tuning

**Scalability:** Peta-scale user data capacity

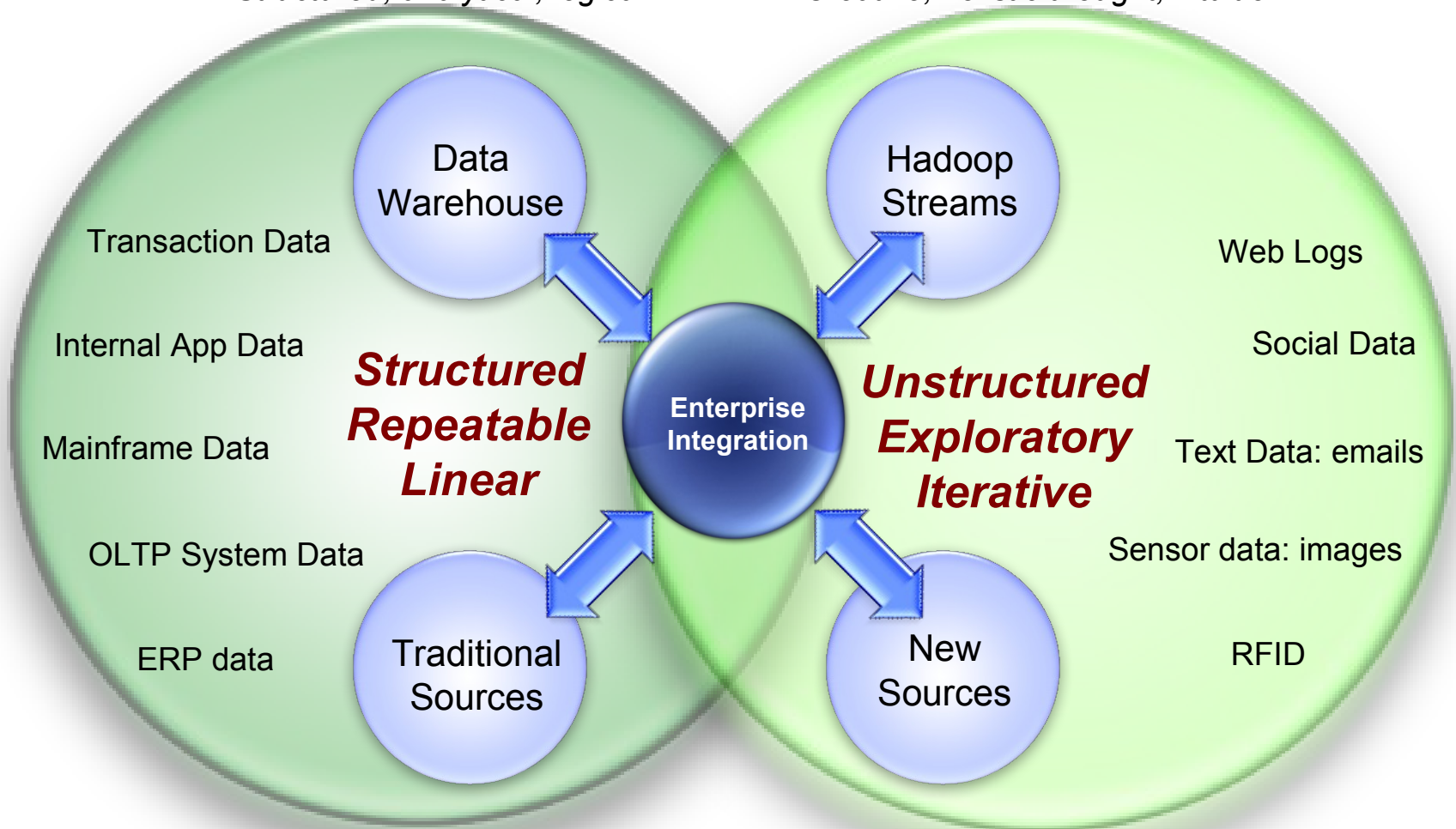
**Smart:** High-performance advanced analytics



# Complementary Analytics

**Traditional Approach**  
*Structured, analytical, logical*

**New Approach**  
*Creative, holistic thought, intuition*

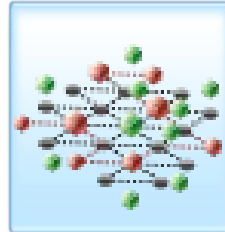


# Most Client Use Cases Combine Multiple Technologies



## Pre-processing

Ingest and analyze unstructured data types and convert to structured data



## Combine structured and unstructured analysis

Augment data warehouse with additional external sources, such as social media



## Combine high velocity and historical analysis

Analyze and react to data in motion; adjust models with deep historical analysis



## Reuse structured data for exploratory analysis

Experimentation and ad-hoc analysis with structured data



# Data Governance

## SQL

Sometimes termed  
*"Schema First"*

- **Separation of Duties**
  - Users can't delete their audit logs
- **Privilege Users**
  - Monitoring authorized and privilege user activities
- **Sensitive Data (in tables)**
  - Protecting and blocking unauthorized access to the system and the data
- **Workflow for audit and compliance controls to validate proper security procedures (satisfies regulations!)**

## noSQL

Sometimes termed  
*"Schema Later"*

- **Separation of Duties**
  - No authorized access
- **Privilege Users**
  - Monitoring authorized and privilege user activities
- **Sensitive Data (in the file system)**
  - Protecting and blocking unauthorized access to the system and the data
- **Workflow for audit and compliance controls to validate proper security procedures (satisfies regulations!)**

# Data Security Design Goals

## SQL

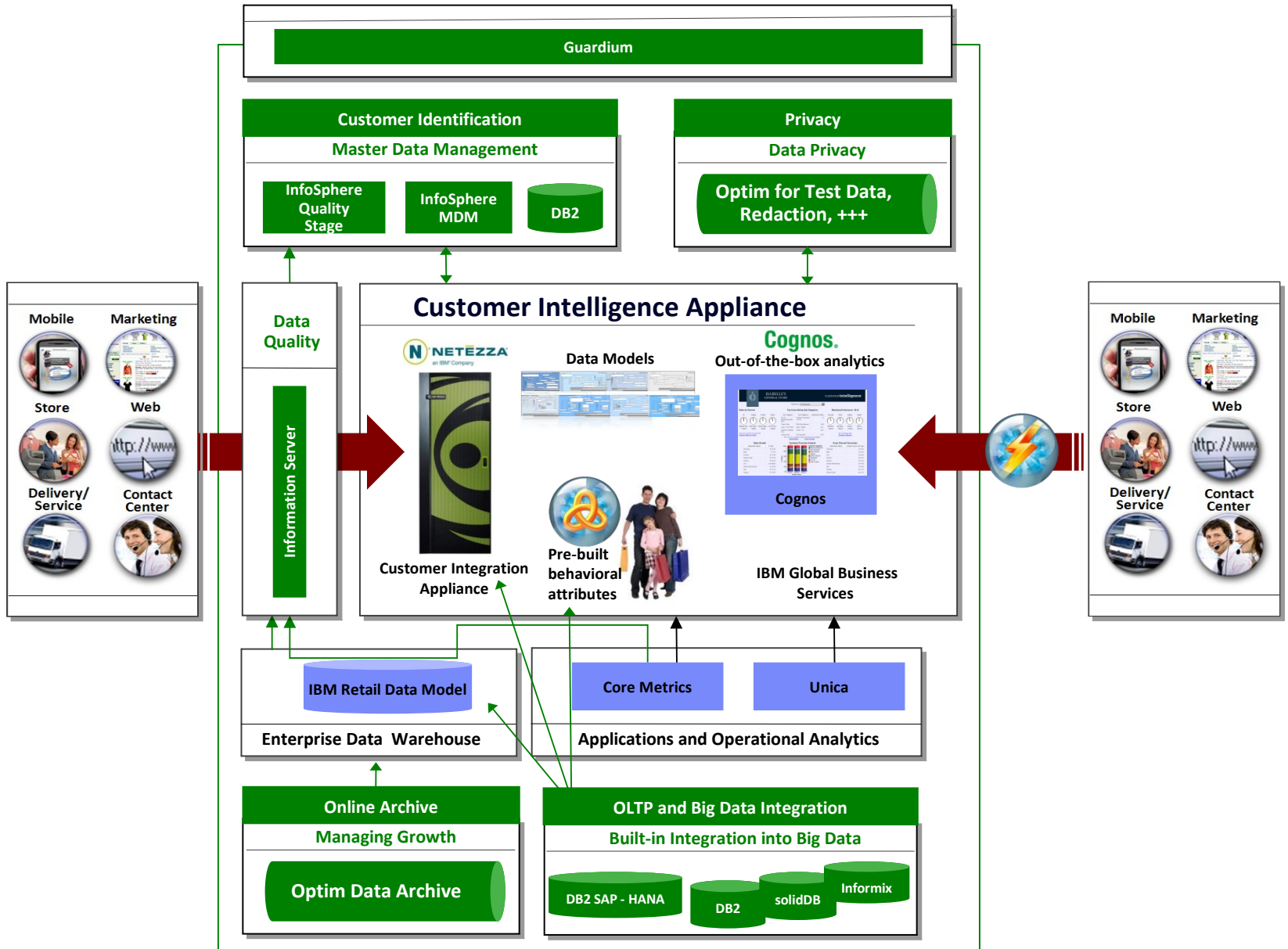
Sometimes termed  
*"Schema First"*

- Minimal impact to the **Database** server resources
- No **Database** configuration changes
- Separation of Duties: Preventing privilege **users** from doing malicious activities
- Audit trail with very granular details of **Database** activity
- Real-time alerting and blocking
- Minimal impact to the network
- 100% **Database** activity visibility
- Heterogeneous support

## noSQL

Sometimes termed  
*"Schema Later"*

- Minimal impact to the **Big Data** server resources
- No **Big Data** configuration changes
- Separation of Duties: Preventing privilege **users/jobs** from doing malicious activities
- Audit trail with very granular details of **Big Data** activity
- Real-time alerting and blocking
- Minimal impact to the network
- 100% **Big Data** activity visibility
- Heterogeneous support





# IBM Flattens the Time to Value Big Data Curve

IBM + IBM + IBM + IBM + IBM + IBM + IBM

Velocity

Hadoop

Harden File System

ETL

Dev. Tooling

Text/ML Analytics

Visualization



Velocity

Hadoop

Harden File System

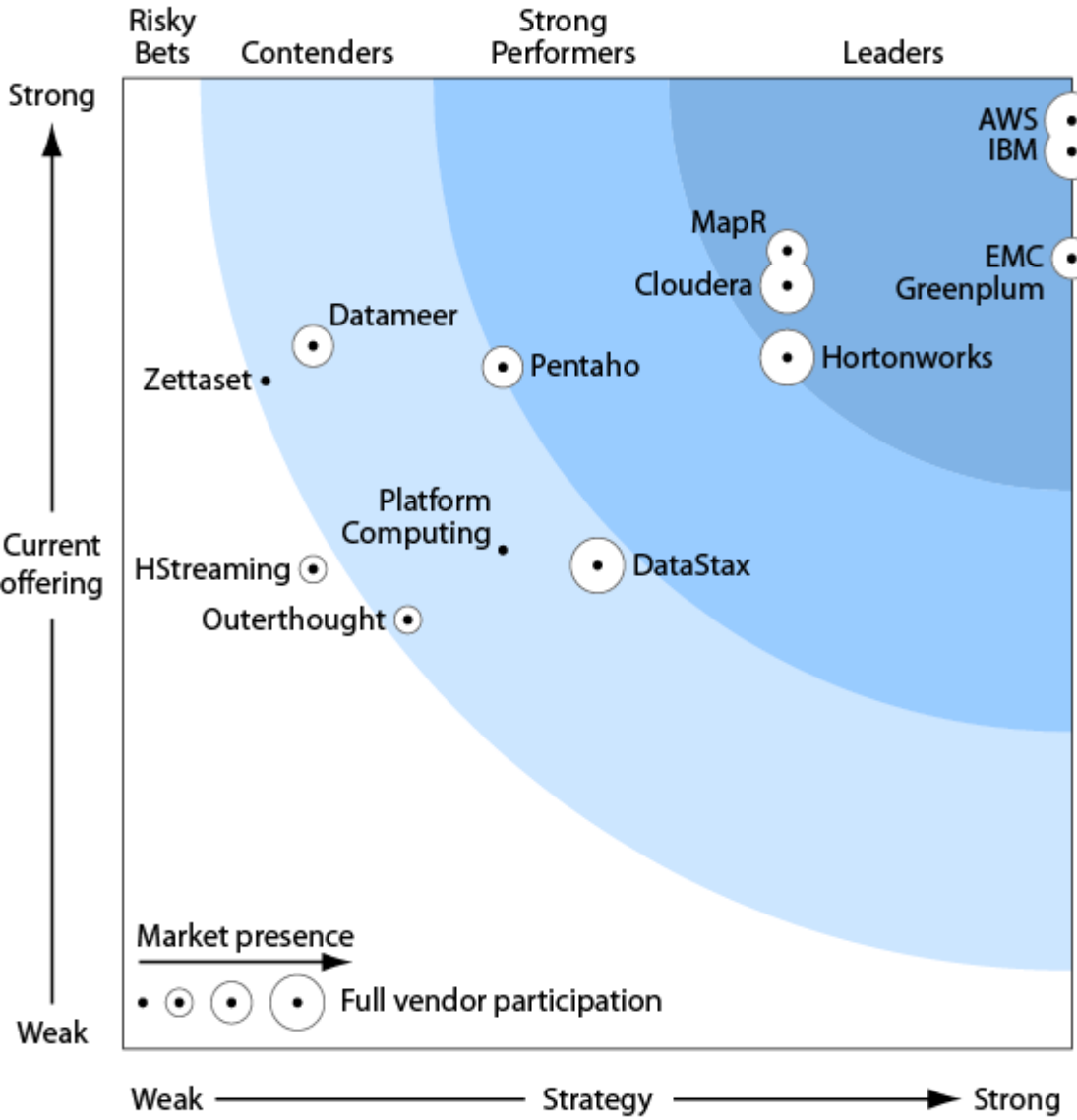
100+ Big Data Business Partners Signed

Text/ML Analytics

Visualization



# First Ever Forrester Wave on Big Data



*“IBM has the deepest Hadoop platform and application portfolio. IBM, an established EDW vendor, has its own Hadoop distribution; an extensive professional services force working on Hadoop projects; extensive R&D programs developing Hadoop technologies; connections to Hadoop from its EDW.”*

–The Forrester Wave™: Enterprise Hadoop Solutions, 1Q12

# Open Source Technology Behind Hadoop





# A Difference of Processing Models

- **SETI@home is a Computational Processing Model**

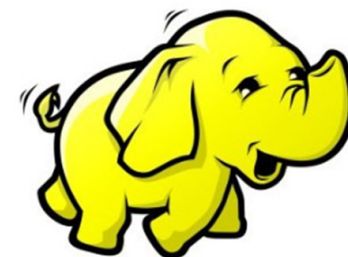
- Service for Extraterrestrial Intelligence (SETI) uses unused desktop CPU processing power to perform wide-spread analysis of radio telescope data
- Pushes data to the program for processing
- **Data to function**

SETI@HOME  
Needs your Help  
Donate to SETI@home

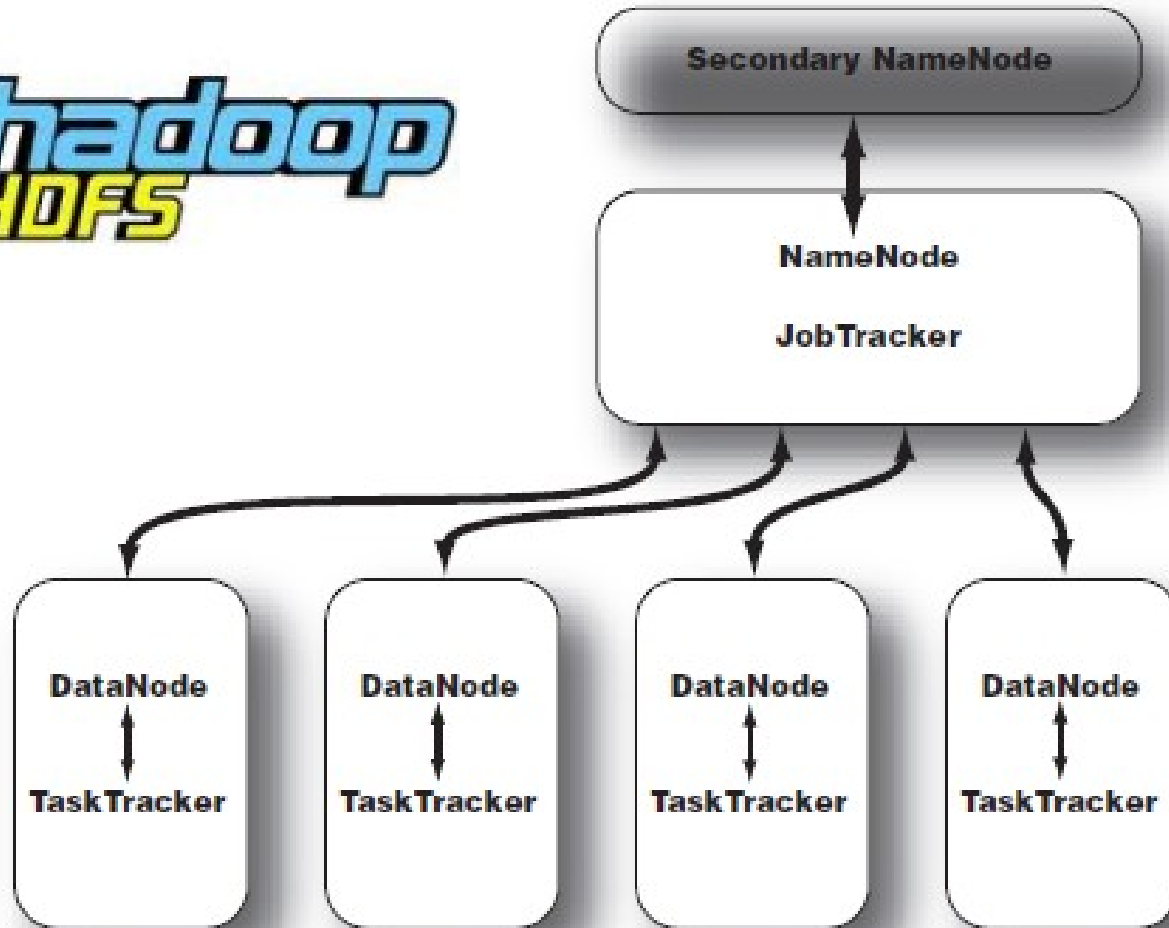


- **MapReduce is a Data Processing Model**

- Data processing primitives of Mappers and Reducers
- Can be complex to write MapReduce programs, but a very **simple configuration change** makes it scale to 1000s of nodes
- **Function to data**



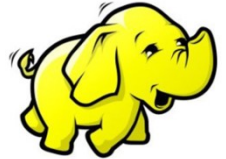
# Building Blocks of Hadoop



# Hadoop Framework

## ■ Hadoop Common

- A utility layer that provides access to the HDFS and projects



## ■ HDFS

- Data storage platform for Hadoop framework
- Can scale to massive size when distributed over multiple computers



## ■ MapReduce

- Framework to process data across a node clusters
- MAP process splits work by first mapping the input across the control nodes of the cluster, then splitting the workload into even smaller data sets and distributing it further throughout the computing cluster
  - Allows MPP – think DB2 DPF ‘like’
- REDUCE collects and combines the nodes’ answers to deliver a result





# Hadoop Explained

## ■ Hadoop computation model

- Data stored in a distributed file system spanning many inexpensive computers
- Bring function to the data
- Distribute application to the compute resources where the data is stored

## ■ Scalable to thousands of nodes and petabytes of data

```

public static class TokenizerMapper
  extends Mapper<Object,Text,Text,IntWritable> {
  private final static IntWritable
    one = new IntWritable(1);
  private Text word = new Text();
  public void map(Object key, Text val, Context
    StringTokenizer itr =
    new StringTokenizer(val.toString());
    while (itr.hasMoreTokens()) {
    word.set(itr.nextToken());
    context.write(word, one);
    }
  }
}

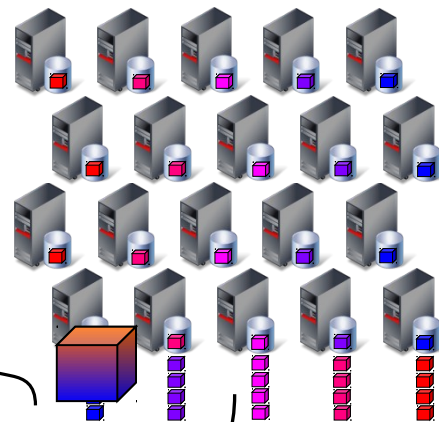
public static class IntSumReducer
  extends Reducer<Text,IntWritable,Text,IntWritable> {
  private IntWritable result = new IntWritable();
  public void reduce(Text key,
    Iterable<IntWritable> vals, Context context){
    int sum = 0;
    for (IntWritable v : vals) {
    sum += v.get();
    }
  }
}

```

MapReduce Application

Distribute map  
tasks to cluster

Hadoop Data Nodes



Shuffle

Result Set

Return a single result set

- 1. Map Phase**  
(break job into small parts)
- 2. Shuffle**  
(transfer interim output for final processing)
- 3. Reduce Phase**  
(boil all output down to a single result set)

# Growing the Hadoop Environment

- **Avro: Data serialization system that converts data into a fast, compact binary data format**
  - Avro data can be stored in a file with versioned schemas
- **Chukwa: Large-scale monitoring system that provides insights into the Hadoop distributed file system and MapReduce**
- **HBase is a scalable, column-oriented distributed database modeled after Google's BigTable distributed storage system**
- **Hive is a data warehouse infrastructure that provides ad hoc query and data summarization for Hadoop**
  - Hive utilizes SQL-like query language call HiveQL
  - HiveQL also used by programmers to run custom MapReduce jobs.



# Growing the Hadoop Environment

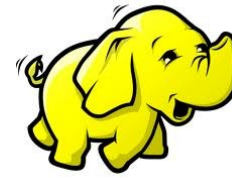
- **Mahout is a data mining library designed to work on the Hadoop framework**
  - Mahout delivers a core set of algorithms designed for clustering, classification and batch-based filtering.
- **Pig is a high-level programming language and execution framework for parallel computation**
  - Pig works within the Hadoop and MapReduce frameworks
- **ZooKeeper provides coordination, configuration and group services for distributed applications working over the Hadoop stack**





# Committed to Open Source

- **Decade of lineage and contributions to the open source community**
  - Apache Hadoop and Jaql, Apache Derby, Apache Geronimo, Apache Jakarta, +++
  - Eclipse: founded by IBM
  - Significant Lucene contributions via IBM Lucene Extension Library (ILEL)
  - DRDA, XQuery, SQL, XML4J, XERCES, HTTP, Java, Linux, +++
  
- **IBM products built on open source**
  - WebSphere: Apache
  - Rational: Eclipse and Apache
  - InfoSphere: Eclipse and Apache, +++
  
- **IBM's BigInsights (Hadoop) is 100% open source compatible with no forks**



Jakarta



eclipse



JAVA™



W3C®

# Learn Hadoop: At Your Place, At Your Pace

## Making Learning Hadoop Easy and Fun

- **Flexible** on-line delivery allows learning @your place and @your pace
- **Free** courses, **free** study materials
- **Cloud-based** sandbox for exercises – **zero setup**
- **8500+** registered students
- **Hadoop Programming Challenge** is sending 3 students to IOD 2011 Conference, **all expenses paid**

The screenshot displays the BigDataUniversity website interface. At the top, the logo reads "BigDataUniversity BETA" with the tagline "Learn from the industry's best". Navigation links include HOME, LEARN, DOWNLOAD, RESOURCES, JOBS, and LEARN Hadoop. A search bar is present on the right.

The main content area features a video player titled "What is Hadoop?" and a prominent orange box for "Hadoop Fundamentals" with the text: "Start your career on Hadoop with this FREE course. Learn with hands-on exercises on a Hadoop cluster. Enroll now!".

Below this, a "Why register?" section lists benefits:
 

- Easy and Affordable:** Learning Hadoop and other Big Data technologies has never been more affordable! Many courses are FREE!
- Latest industry trends:** Acquire 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.

A "sign me up" button is visible. To the right, a banner for "Study Made Easy!" promotes "FREE Books" with a "SIGN UP" button. Below that, a "Student Testimonials" section features a quote from Balázs (USA) praising the training material's quality and support.

At the bottom right, navigation links for "about us", "legal", "contact", and "bug reports" are provided.

# Why IBM for Big Data The Solution Side



**Paul Zikopoulos, BA, MBA**

Director, IBM Information Management Technical Professionals,  
WW Competitive Database, and Big Data

IBM Certified Advanced Technical Expert (Clusters and DRDA)

IBM Certified Customer Solutions Expert (DBA and BI)

[paulz@ca.ibm.com](mailto:paulz@ca.ibm.com)



# A Big Data Platform

## Analytics Excellence

Text Analytics Toolkit  
 Machine Learning Toolkit  
 Industry Accelerators  
 Development Tooling  
 Visualization Tooling  
 Deployment Tooling (“App Store”)  
 \$14B in 5 yrs. on Analytics  
 +++

## In-Motion Operational Excellence

Unrivalled....  
**Semi-structured data**

## At-Rest Operational Excellence

Harden Hadoop - GPFS  
 Surface Area Lock Down  
 Policy Driven Retention & Immutability  
 Role-Based Security  
 Adaptive MapReduce  
 Workload Manager  
 Fast Splittable CMX Compression  
 REST-exposed Administration  
 +++

*Embed... extend*



## In-Motion

Analyze extreme amounts of data in milliseconds  
 Uses same analytics as BigInsights  
 Data can be analyzed on the way into the enterprise for earlier pattern detection

## At-Rest

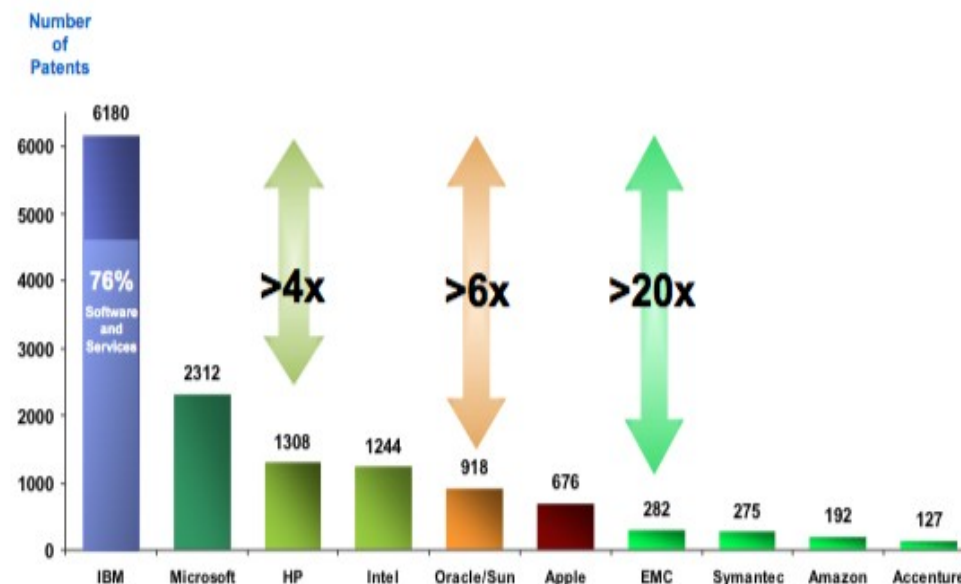
Beyond traditional structured data  
 BigInsights uses same analytics as Streams  
 No forked, not ported: Hadoop Extended with operational excellence and security  
 Netezza for in-database MapReduce  
 MPP Data Warehouses

## Committed to Innovation

- \$100M new investment into analytics announced 2Q11
- IBM spent \$14+ billion in 24 analytics acquisitions in 5 years
- IBM has the **largest commercial research organization on Earth**
  - 200+ mathematicians developing breakthrough analytics
- Largest patent portfolio in the industry



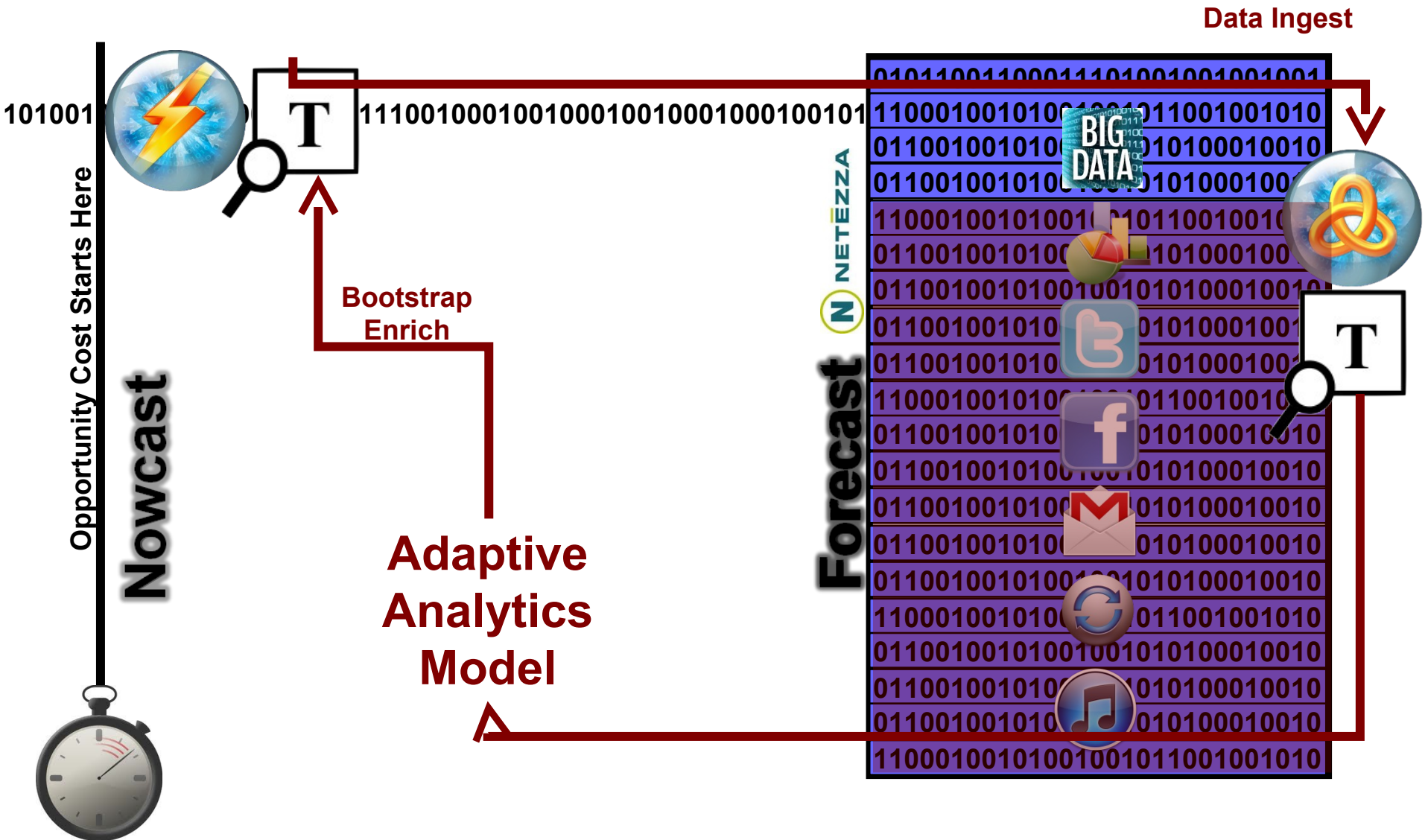
Over the last five years, the company spent \$14 billion on the acquisition of two dozen data tools companies. IBM believes its future relies on helping customers manage and learn from the large amount of data available today. The company is currently working on integrating its system, Watson, into the health care field as a physician's assistant by feeding it medical specific domain information.



For **19** consecutive years IBM inventors have received the most U.S. Patents

More than **6,000** patents in a single year

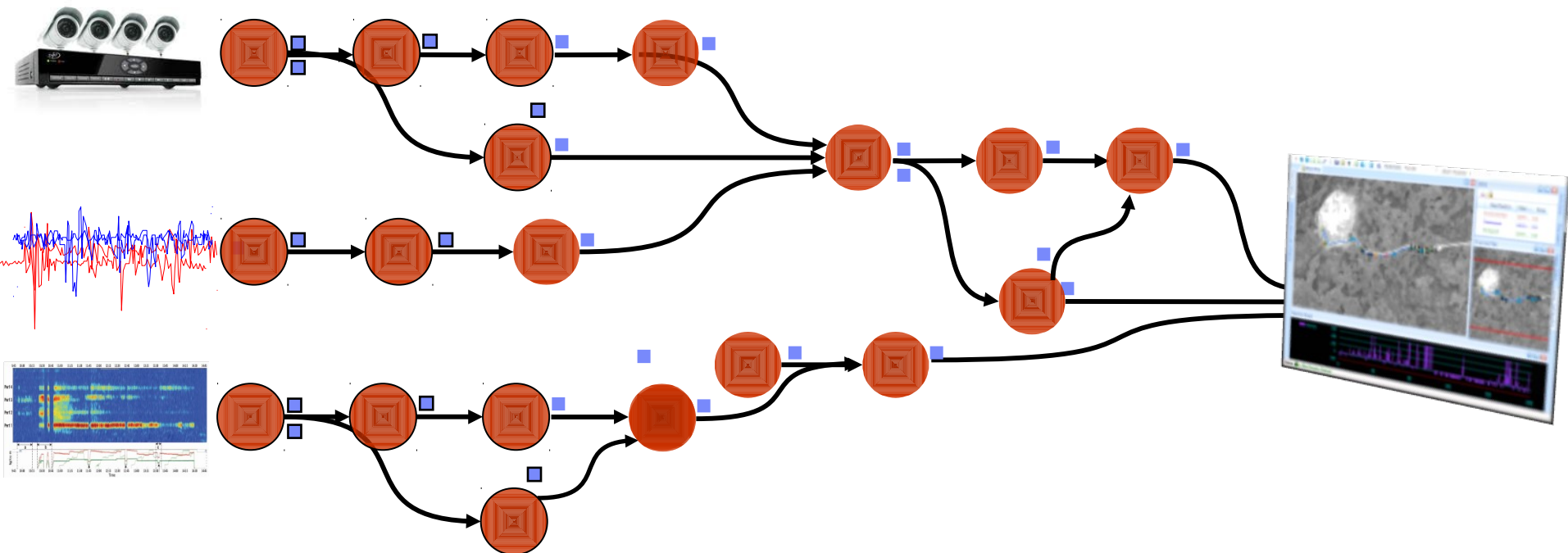
# A Big Data Platform for Data In-Motion and At-Rest



# Data In Motion



- Hear what's going on miles away to optimize perimeter displacements
- **Perspective: Try to find the word "Zero" in a 1000 MP3 song library in a fraction of a second**
  - Figure out the difference between the sound of a human whisper and the wind





Average cluster size is **100** nodes

one  
What insight could you gain if you had  
hour  
full use of a **100-node** Hadoop  
What if of this 100-node  
cluster for an hour?  
cluster cost

**\$34**

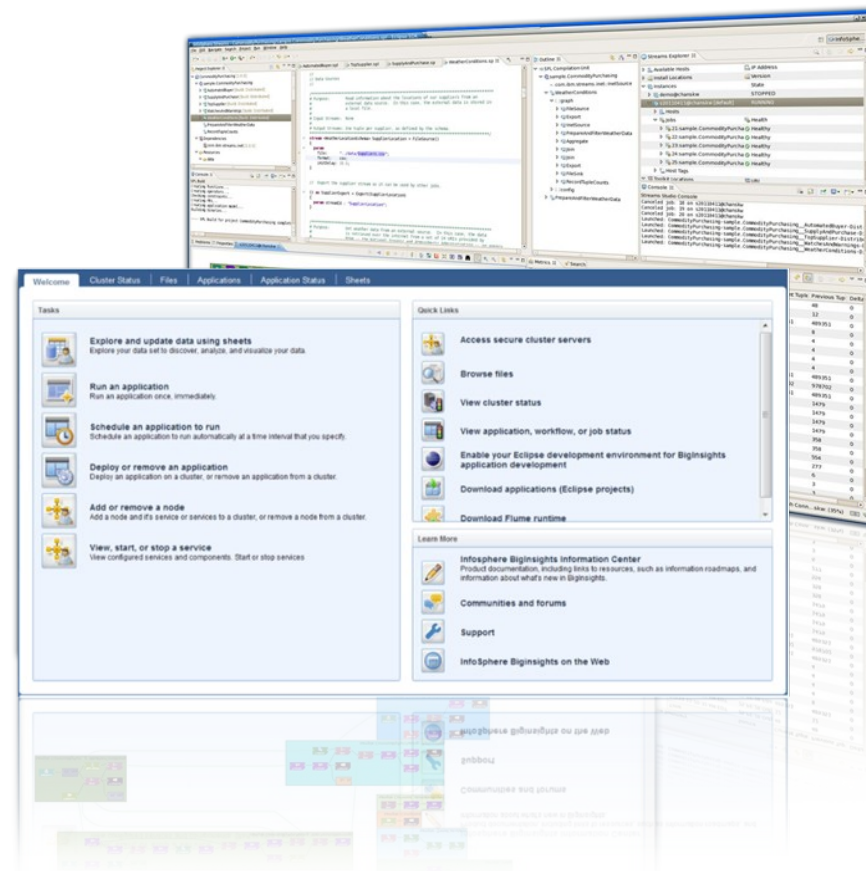


# Ease of Use for Developers and Users



## End-user Visualization

Data exploration, crawling, and analytics



## Development Environment

Familiar coding and tooling environment, testing, and optimization

# What is BigSheets?

## Browser-based Big Data analytics tool for business users

### Big Data Challenges...

- Business users need a no programming approach for analyzing Big Data
- Extremely difficult to find actionable business insights in data from multiple sources with different formats
- Translating untapped data into actionable business insights is a common requirement that requires visualization

### How can BigSheets help?

- Spreadsheet-like discovery interface lets business users easily analyze Big Data with **ZERO PROGRAMMING**
- **BUILT-IN “readers”** can work with data in several common formats
  - JSON arrays, CSV, TSV, Web crawler output, . . .
- Users can **VISUALLY** combine and explore various types of data to identify “hidden” insights



Data Collections > View Results > Create

Unnamed Collection(1)

Save Exit

fx

	A	B
	EMPNO	FIRSTNAM
1	10	Jennifer
2	20	Pablo
3	30	Patricia
4	50	Sanderson
5	60	Franco
6	70	Hedi
7	90	Coleen
8	100	Ramesh
9	110	Andrew
10	120	Robert
11	130	Heidi
12	140	Peggy
13	150	Jay
14	160	Jun

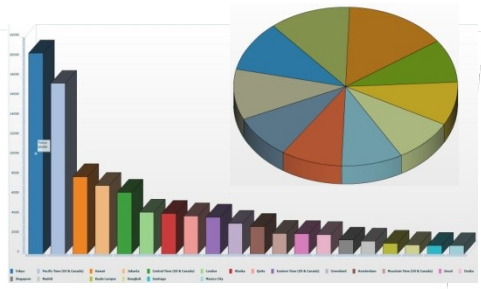
ANALYST **CLERK**

**DESIGNER** FIELDREP

MANAGER OPERATOR

SALESREP

DESIGNER (10 occurrences)



Select a type of sheet:

- Filter
- Macro
- Load
- Pivot
- Combine
- Union
- Limit
- Distinct
- Copy
- Formula**

Add Sheet using by entering a formula

Data Collections > View Results

### How many folks "feel" Positive or Negative towards which Brands?

[Edit](#) [Delete](#) | [Twitter Data](#) > ... > [What Brands ar...](#) > [How many folks...](#) : [Build new collection](#) ▾

Ready 0% | [Close](#)

[Edit](#)  
sentiment  
Click to select,  
Ctrl-Click: multiple  
Shift-Click: range

- com.ibm.en.Positi
- com.ibm.en.Negat

**8% spread  
change  
affecting  
brand**

**Over 1M  
tweets  
analyzed**

**2nd most  
Tweets/sec  
run rate as  
of 1Q12**

[Search>>](#)



Label

Color

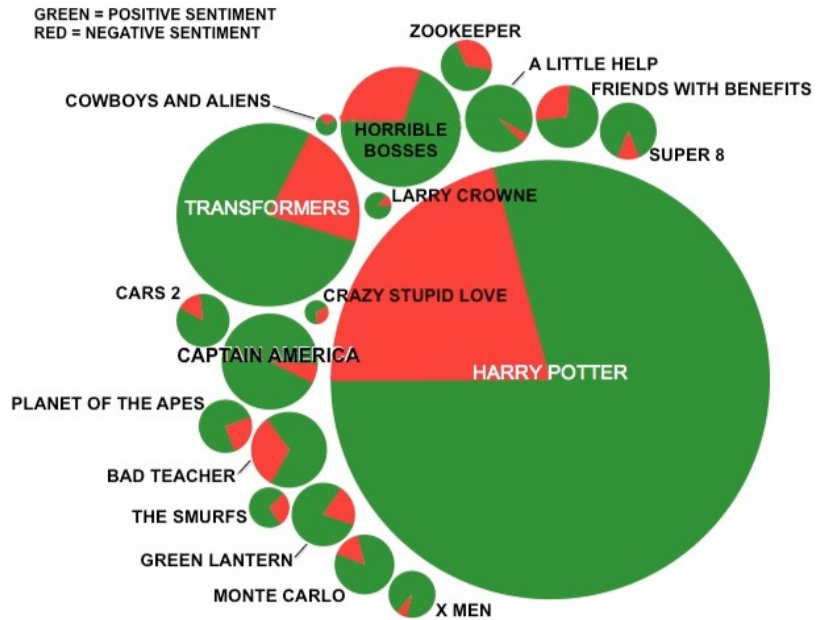
[Add Chart](#) | [Result](#) [Bubble](#)

► Details:

► Related Collections:

# Big Data Made Easy for the Little Guy

- USC's Film Forecaster correctly predicted a clamor for "Hangover 2" that resulted in \$100 million opening over Memorial Day weekend
  - Looked at 250K-500K Tweets and broke down positive and negative messages using a lexicon of 1700 words



*The Film Forecaster sounds like a big undertaking for USC, but it really came down to one communications masters student who **learned Big Sheets in a day**, then pulled in the tweets and analyzed them - Ryan Kim*

# Running Applications from the Web Console

**Applications**

Ad hoc Hive query

Ad hoc Jaql query

Ad hoc Pig query

Boardreader

Database Export

Database Import

Distributed File Copy

TeraGen-TeraSort

Web Crawler

Word Count

**Name:** Word Count Undo Deploy Delete Refresh

**Description**

The Word Count application reads text files and determines the frequency with which certain words occur.

**Execution**

Execution Name:  Run

**Parameters**

Input path:  Browse... ▶

Output path:  Browse... ▶

**Advanced Settings**

Update Sheets Collection

Schedule Job [View Schedule Configuration](#)

**Application History** Refresh Print Close

Status	Execution Name	Progress	Start Time	Elapsed Time	Output	Details
No filter applied						
✓	test2	100%	Dec 1, 2011 2:11:48 PM	55(sec)		
✓	test2	100%	Nov 30, 2011 10:36:47 AM	45(sec)		
✗	Default Execution	100%	Nov 28, 2011 11:46:10 AM	19(sec)	N/A	
✓	Default Execution	100%	Nov 28, 2011 11:21:03 AM	45(sec)		
✓	test2	100%	N/A	N/A		

1 - 5 of 5 items 10 | 25 | 50 | 100 | All 1



# A Rich Management Big Data Tool

**Tasks**

- explore and update data using sheets**  
run an application and explore the resulting data in a sheets. Rerun the application and explore the result.
- 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 an applications**  
deploy an application on a cluster.
- add a node**  
add a new node and service to a cluster. add a service to an existing node.
- start or stop a service**  
start or stop a service

**Where and how to begin performing common administrative or analytical tasks**

**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

**Quick links to common functions**

**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

**Learn more through external Web resources**

# A Rich Management Big Data Tool

**Add Nodes** ✕

Service: DataNode/TaskTracker

Start IP/Host:

Number of nodes: 1

End IP:

The rack can be specified as an arbitrary path in the following format: /top-switch-name/rack-name. Example: /default-rack

Rack:

The root user will be used to create the Biginsights administrative user and di

Root password:

Confirm\_Root password:

**Nodes**

Welcome Cluster Status Files Applications

Welcome biadmin | [Log out](#) | [About](#) | [Help](#)

Welcome
Cluster Status
Files
Applications
Application Status
Sheets

**Application Status > Workflow Summary**

Status	External Status	ID	External ID	Type	Start Time	End Time	Details
✔	RUNNING	0000006-111021082208237-oozie-hado-W@nutch	job_201110210808_0023	java	Oct 21, 2011 2:32:27 PM	N/A	▶

1 - 1 of 1 items 10 | 25 | 50 | 100 | All

**Workflow Information**

<b>Status:</b> RUNNING	<b>Start Time:</b> Oct 21, 2011 2:32:27 PM
<b>Workflow ID:</b> 0000006-111021082208237-oozie-hado-W	<b>End Time:</b> N/A
<b>Name:</b> nutch-wf	<b>Created:</b> Oct 21, 2011 2:32:27 PM
<b>Path:</b> hdfs://svlhdev03.svl.ibm.com:9000/user/applications/15311683-6264-4b74-8a6c-9e9aba2e5874/workflow/workflow.xml	<b>Last Modified:</b> Oct 21, 2011 2:32:29 PM

**Workflow Configuration**

**Workflow Log**

```

2011-10-21 14:32:27,690 WARN ActionStartCommand:96 - USER[biadmin] GROUP[supergroup] TOKEN[] APP[nutch-wf] JOB[0000006-111021082208237-oozie-hado-W] ACTION[-] [***0000006-111021082208237-oozie-hado-W@nutch***]In call()...status=PREP
2011-10-21 14:32:29,690 INFO JavaActionExecutor:84 - USER[biadmin] GROUP[supergroup] TOKEN[] APP[nutch-wf] JOB[0000006-111021082208237-oozie-hado-W] ACTION[0000006-111021082208237-oozie-hado-W@nutch] checking action, external ID [job_201110210808_0023] status [RUNNING]
2011-10-21 14:32:29,692 WARN ActionStartCommand:96 - USER[biadmin] GROUP[supergroup] TOKEN[] APP[nutch-wf] JOB[0000006-111021082208237-oozie-hado-W] ACTION[0000006-111021082208237-oozie-hado-W@nutch] [***0000006-111021082208237-oozie-hado-W@nutch***]Action status=RUNNING
2011-10-21 14:32:29,695 WARN ActionStartCommand:96 - USER[biadmin] GROUP[supergroup] TOKEN[] APP[nutch-wf] JOB[0000006-111021082208237-oozie-hado-W] ACTION[0000006-111021082208237-oozie-hado-W@nutch] [***0000006-111021082208237-oozie-hado-W@nutch***]Action updated in DB!
                
```

**Nodes** ✔ 1

Map/Reduce ✔ Running

Hive ✔ Running

JAQL Server ✘ Unavailable

Flume ✔ Running

Zookeeper ✔ Running

Distributed File System ✔ Running

HBase ✔ Running

Oozie ✔ Running

Catalog ✔ Running

# Rich Big Data Development Environment

- Eclipse base Development tools
  - For JQAL, Hive, PIG, Java MapReduce and Text Analytics

The screenshot displays the Eclipse IDE interface. At the top, a 'New' wizard dialog is open with 'Select a wizard' and a search filter. The Project Explorer on the left shows a project structure with 'WordCount.java' selected under 'myPackage'. The main editor shows the Java code for 'WordCount.java', which is a MapReduce application. A context menu is open over the code, with 'Run on Hadoop' selected. At the bottom, the 'Map/Reduce Locations' table shows the execution status of the job across different nodes.

Location	Master node	State	Status
svlhdev18.svl.ibm.com	svlhdev18.svl.ibm.com		
job_201105101647_0014		SUCCEEDED	Maps : 2/2 (1.0) Reduces : 0/0 (1.0)
job_201105101647_0015		SUCCEEDED	Maps : 2/2 (1.0) Reduces : 8/8 (1.0)
job_201105101647_0016		SUCCEEDED	Maps : 2/2 (1.0) Reduces : 8/8 (1.0)
svltest385.svl.ibm.com	svltest385.svl.ibm.com		

# The Path to Efficiency: Declarative Languages

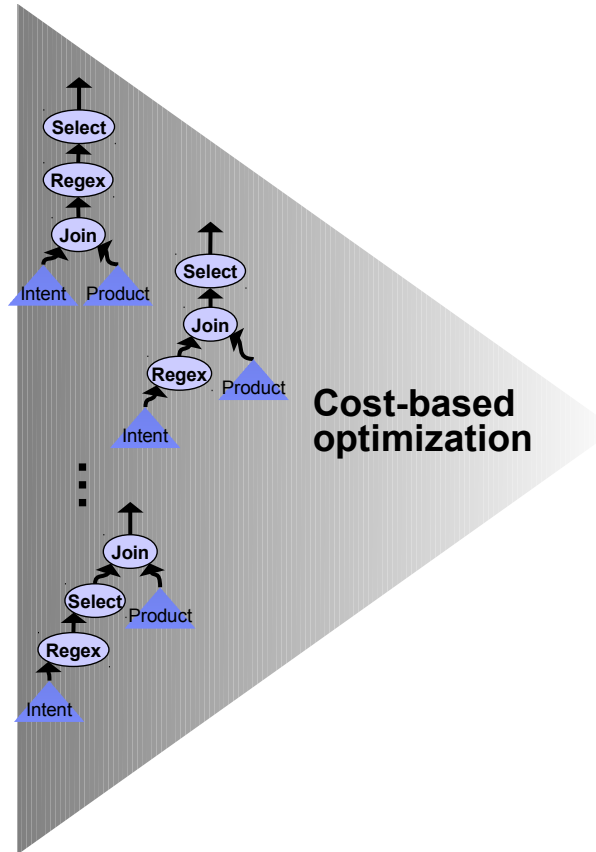
## Offline

### Development Environment

#### Declarative Language

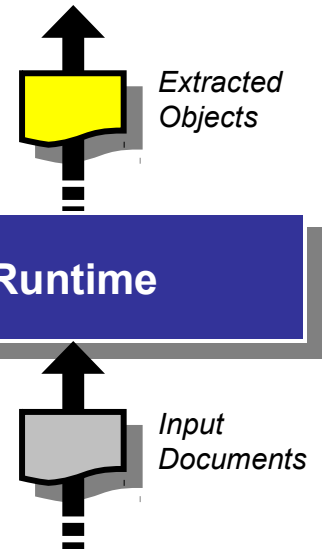
```
create view MonetizableIntent as
select P.name as product,
       I.clue as strength
from Intent I, Product P
where Follows(I.clue, P.name, 0, 20)
and Not(ContainsRegex(/\b(not)\b/,
LeftContext(I.clue, 10)));
```

- Streams, Text Analytics, Machine Learning, SQL all are declarative, simple to learn languages
- All have strong development tooling and accelerators
- IBM has been optimizing declarative languages for decades, **IN FACT, IBM INVENTED IT!**



## Runtime

### Runtime



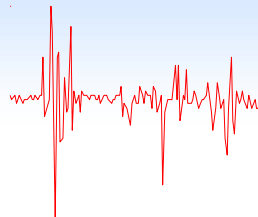
- High-throughput
- Small memory footprint
- Optimizes for tasks:
  - Example, Text Analytics needs CPU optimization



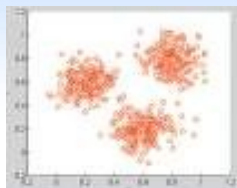
# Analytic Accelerators Designed for Variety

**Text**  
(listen, verb),  
(radio, noun)

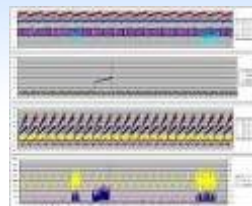
**Simple &  
Advanced Text**



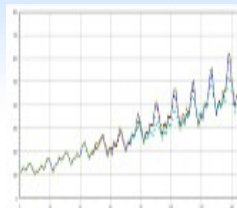
**Acoustic**



**Mining in  
Microseconds**



**Advanced  
Mathematical Models**



**Predictive**

$$\sum_{\text{population}} R(s_t, a_t)$$

**Statistics**



**GeoSpatial**



**Image & Video**

# Accelerators Improve Time to Value



## Telecommunications

CDR streaming analytics  
Deep Network Analytics



## Retail Customer Intelligence

Customer Behavior and Lifetime Value Analysis



## Finance

Streaming options trading  
Insurance and banking DW models



## Social Media Analytics

Sentiment Analytics, Intent to purchase



## Public Transportation

Real-time monitoring and routing optimization



## Data Mining

Streaming statistical analysis



Over 100 sample applications



User Defined Toolkits



Standard Toolkits



Industry Data Models

Banking, Insurance, Telco, Healthcare, Retail

# Telecommunications CDR Analytic Accelerator

## Analyze Call Detail Records in Real Time

### ■ Streaming Analytic Accelerators

- CDR dropped call analysis
- Determine VIP customers with service issues – proactive alerts
- CDR Adapters – ASN.1, Binary, ASCII
- Analytic Operators – CDR de-duplication, dropped call detection, termination reason, customer importance
- Visualization – real-time KPI dashboard

### ■ Data Warehouse Appliance

- Integrated network, devices, customer, and services model
- Telecom model, KPIs, and KQIs

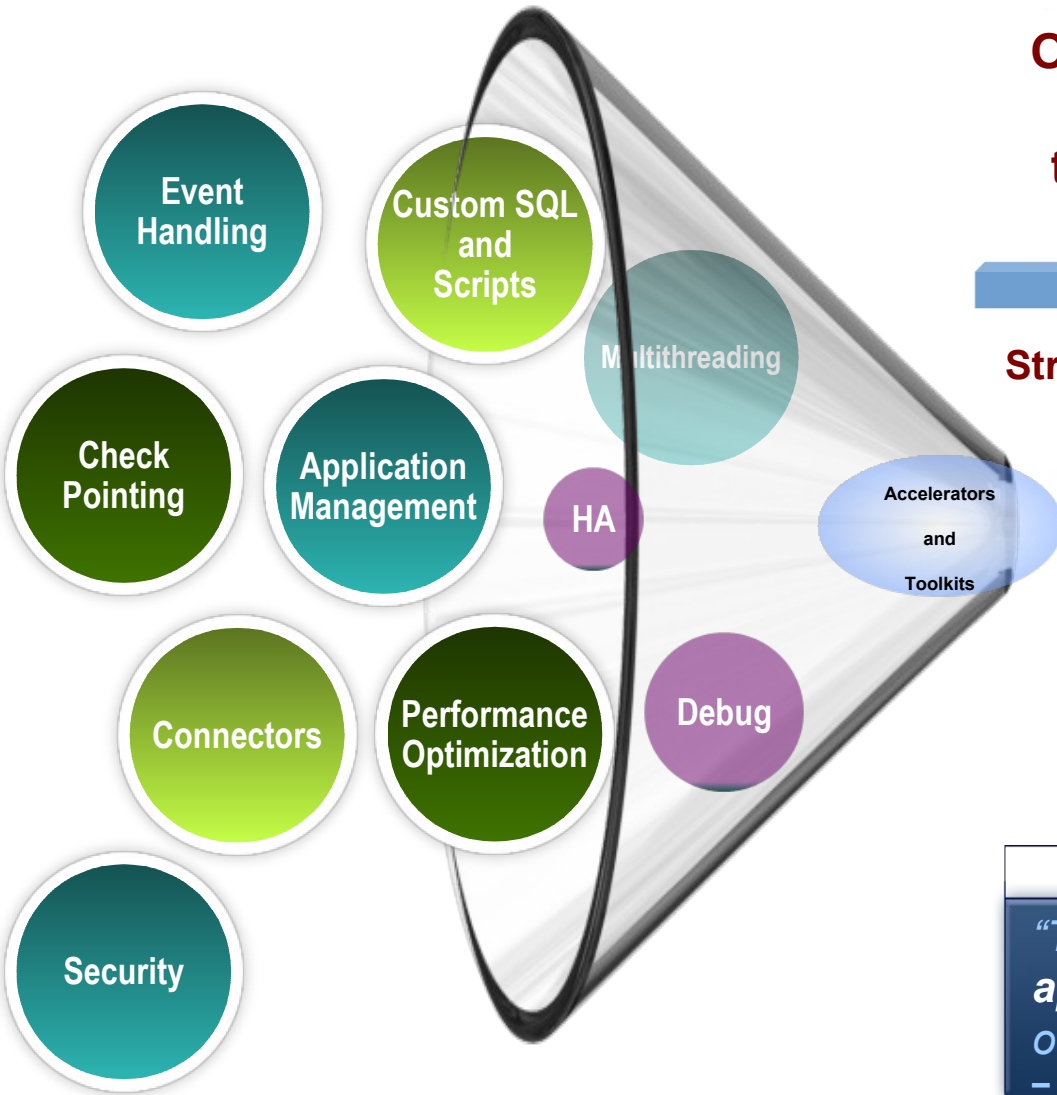




# Without a Big Data Platform You Code...

# IBM Big Data Platform

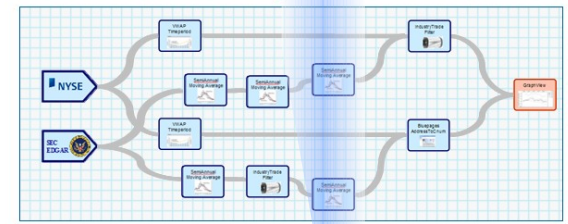
Over 100 sample applications and toolkits with industry focused toolkits with 300+ functions and operators!



Streams Processing Language

Streams provides development, deployment, runtime, and infrastructure services

Platform optimized compilation



*“TerraEchos developers can deliver applications 45% faster due to the agility of Streams Processing Language...”*  
 – Alex Philip, CEO and President



# Streams Runtime Illustrated

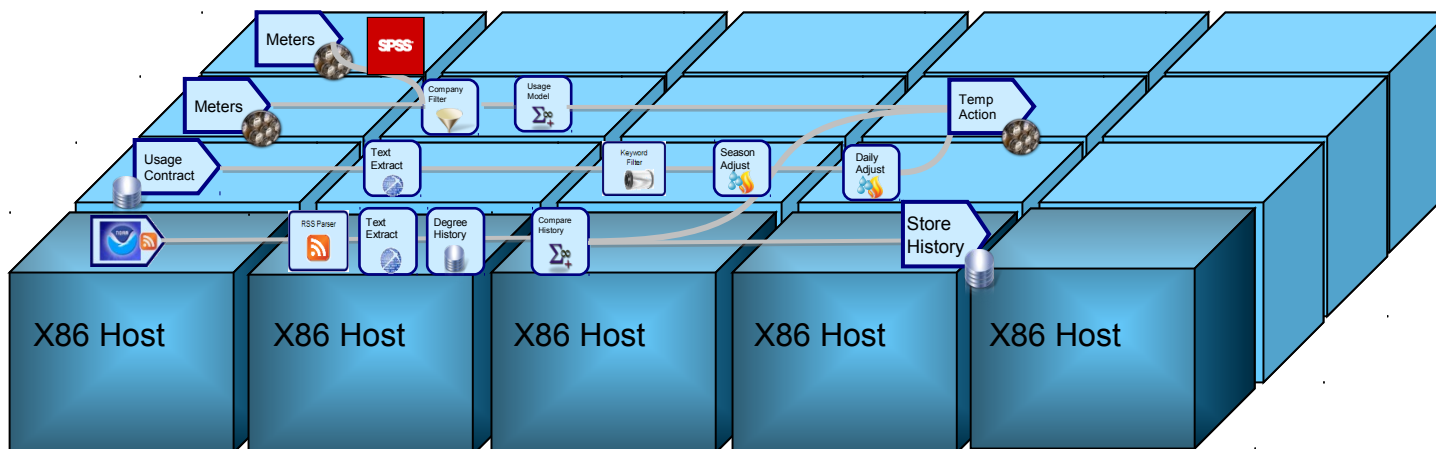


Optimizing scheduler assigns PEs to nodes, and continually manages resource allocation

Dynamically add nodes and jobs

Commodity hardware – laptop, blades or high performance clusters

Add in SPSS jobs in the flow

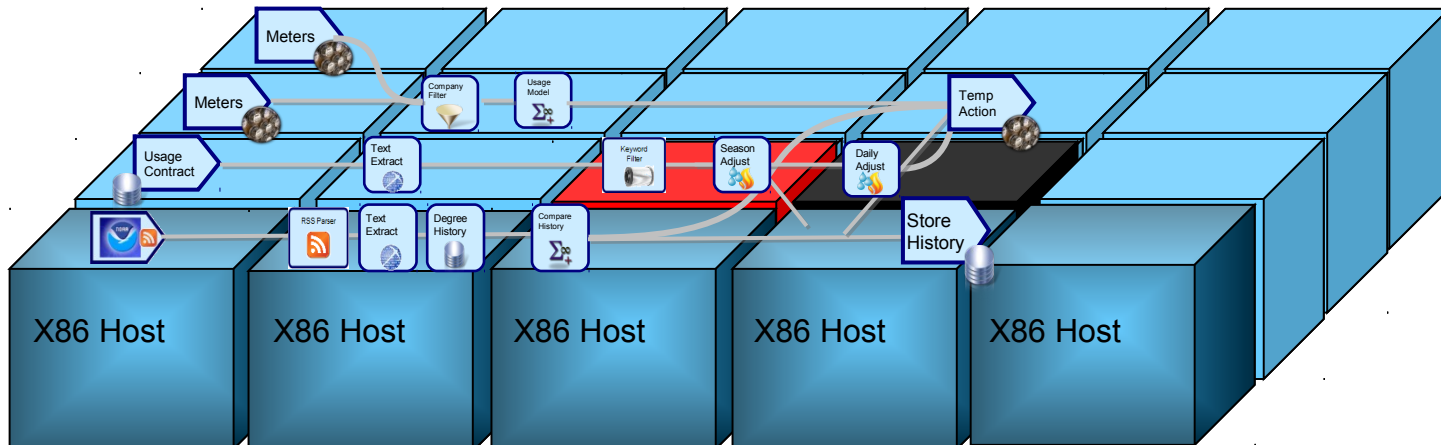


# Streams Runtime Illustrated

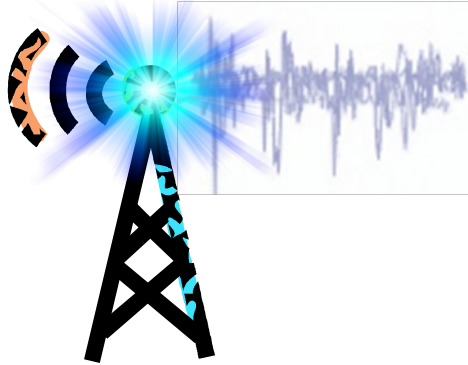


PEs on busy nodes, can be moved manually by the Streams administrator

PEs on failing nodes can be moved automatically, with communications re-routed



# How Text Analytics Works



Football **World Cup 2010**, one team distinguished themselves well, losing to the eventual champions 1-0 in the Final. Early in the second half, **Netherlands' striker, Arjen Robben**, had a breakaway, but the **keeper for Spain, Iker Casilas** made the save. **Winger Andres Iniesta** scored for **Spain** for the win.

## World Cup 2010 Highlights

Name	Position	Country
<b>Arjen Robben</b>	<b>Striker</b>	<b>Netherlands</b>
<b>Iker Casilas</b>	<b>Keeper</b>	<b>Spain</b>
<b>Andres Iniesta</b>	<b>Winger</b>	<b>Spain</b>

# What's Wrong with Text Analytics Today

- **Current alternative approaches and infrastructure for text analytics present challenges for analysts**
  - They tend to **perform poorly** (in terms of accuracy and speed)
  - They are **difficult** to use
- **These alternative approaches rely on the raw text flowing only forward through a system of extractors and filters**
  - **Inflexible and inefficient** approach, often resulting in **redundant processing**
- **Existing toolkits are also limited in their expressiveness**
  - Analysts having to **develop custom code**
  - Programmer  $\leftrightarrow$  Analyst (think Java Developer  $\leftrightarrow$  DBA struggles)
  - Leads to more delays, complexity, and difficulties getting it right
  - Biggest factor **hurting analyst productivity** is the difficulty in determining how the system produced a certain result



# Extracting Person and Phone Relationships



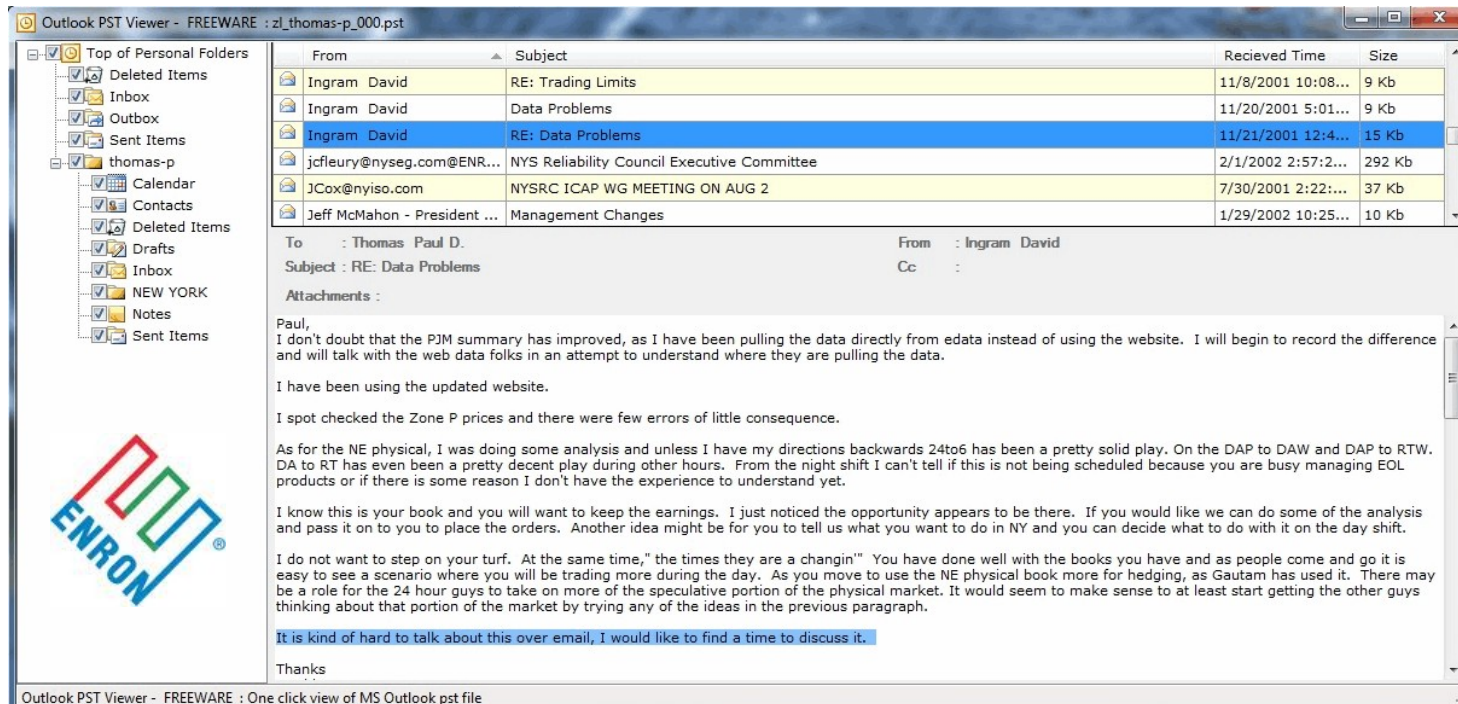
## Testing



- Write complex expressions to identify syntactic features (Phone numbers and capitalized words), collecting dictionaries (lists of common first and last names), rule to combine features into larger concepts (full names)
- First Name Rule: "A match of a dictionary first name followed by an immediate match of an expression identifying capitalized words"
- Annotators executed on collection of documents
- Developer manually examines 1000s of extraction results (annotators) to determine correctness and missing rules
- Developer seeks to understand the causes of the mistakes.
- Morgan Stanley: Remove or add names, create new dictionaries, test again, and again, and again...

# Text Analytic Toolkit Example

- **Semantically search Enron's emails to support queries such as "Find Tom's phone" in order to find his actual phone number**
  - As opposed to emails containing words Tom and Phone
- **Need to be able to accurately extract email entities of type Person and Phone and the relationship between them**



Outlook PST Viewer - FREEMWARE : z:\thomas-p\_000.pst

Top of Personal Folders

- Deleted Items
- Inbox
- Outbox
- Sent Items
- thomas-p
  - Calendar
  - Contacts
  - Deleted Items
  - Drafts
  - Inbox
  - NEW YORK
  - Notes
  - Sent Items

From	Subject	Received Time	Size
Ingram David	RE: Trading Limits	11/8/2001 10:08...	9 Kb
Ingram David	Data Problems	11/20/2001 5:01...	9 Kb
Ingram David	RE: Data Problems	11/21/2001 12:4...	15 Kb
jcifleury@nyseg.com@ENR...	NYS Reliability Council Executive Committee	2/1/2002 2:57:2...	292 Kb
JCox@nyiso.com	NYSRC ICAP WG MEETING ON AUG 2	7/30/2001 2:22:...	37 Kb
Jeff McMahon - President ...	Management Changes	1/29/2002 10:25...	10 Kb

To : Thomas Paul D.  
From : Ingram David  
Subject : RE: Data Problems  
Cc :

Attachments :

Paul,  
I don't doubt that the PJM summary has improved, as I have been pulling the data directly from edata instead of using the website. I will begin to record the difference and will talk with the web data folks in an attempt to understand where they are pulling the data.

I have been using the updated website.

I spot checked the Zone P prices and there were few errors of little consequence.

As for the NE physical, I was doing some analysis and unless I have my directions backwards 24to6 has been a pretty solid play. On the DAP to DAW and DAP to RTW. DA to RT has even been a pretty decent play during other hours. From the night shift I can't tell if this is not being scheduled because you are busy managing EOL products or if there is some reason I don't have the experience to understand yet.

I know this is your book and you will want to keep the earnings. I just noticed the opportunity appears to be there. If you would like we can do some of the analysis and pass it on to you to place the orders. Another idea might be for you to tell us what you want to do in NY and you can decide what to do with it on the day shift.

I do not want to step on your turf. At the same time, "the times they are a changin'" You have done well with the books you have and as people come and go it is easy to see a scenario where you will be trading more during the day. As you move to use the NE physical book more for hedging, as Gautam has used it. There may be a role for the 24 hour guys to take on more of the speculative portion of the physical market. It would seem to make sense to at least start getting the other guys thinking about that portion of the market by trying any of the ideas in the previous paragraph.

It is kind of hard to talk about this over email, I would like to find a time to discuss it.

Thanks

Outlook PST Viewer - FREEMWARE : One click view of MS Outlook pst file

# The Enron Email Example

- Start with a naïve set of rules to identify a `PersonPhone` relationship built using some sort of editor
  - Build out an extractor that know how to find a person's name: Lorraine Smith
  - Build out an extractor that knows how to find a phone number: 607-205-4493

Person
PersonPhone
PhoneNumber

**View: PersonPhone**

person: Span over Doc.text	phone: Span over Doc.text	personphone: Span over Doc.text
Doc.text[429-437]: 'Lorraine'	Doc.text[445-457]: '607)205-4493'	Doc.text[429-457]: 'Lorraine Smith (607)205-4493'
Doc.text[478-484]: 'Morgan'	Doc.text[499-507]: '205-4493'	Doc.text[478-507]: 'Morgan Stanley, fax: 205-4493'

**Document Text**

Show Full Text Provenance

...  
 have project questions, please call Lorraine Smith (607)205-4493.  
 When done, send to Morgan Stanley, fax: 205-4493, then call  
 Emma, x33650.

607.205.4493

Euro spacing

xt. ext. ...

North American spacing

+0119054132112

Globalstar GSP-1600

911, 411

+++

# The Tricky Thing About Sentiment...



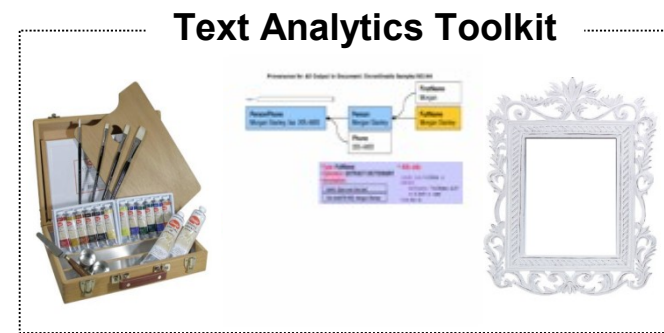
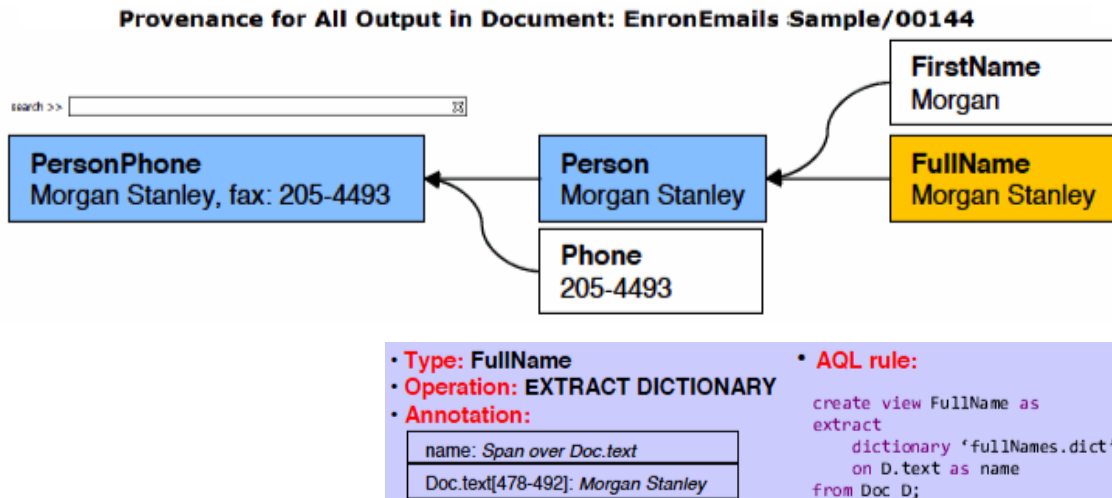


# Text Analytics Toolkit

- **System T text analytics engine previously only embedded in IBM products and hidden from end users**
  - Found in Lotus Notes, IBM e-discovery Analyzer, CCI, InfoSphere Warehouse,+++
  - Almost a decade since initial release
- **BigInsights is the first time IBM opens up the Text Analytics Engine technology for customization and development**
- **BigInsights Text Analytic Toolkit provides developer tools, an easy to use text analytics language, and a set of extractors for fast adoption**
  - **Multilingual support**, including support for DBCS languages
- **BigInsights includes Annotator Query Language (AQL): SQL-like!**
  - **Fully declarative** text analytics language
  - **No “black boxes”** or modules that can’t be customized.
  - **Tooling for easy customization** because you are abstracted from the programmatic details
  - Competing solutions make use of locked up **black-box modules that cannot be customized**, which **restricts flexibility** and are **difficult to optimize for performance**

# Accelerating Analytics – Explainability

- Every annotation's provenance can be visualized
- Provenance of `Morgan Stanley Fax: 205-4493` shows the `FullName` rule is responsible for generating incorrect annotation
  - Solutions?
    - Remove Morgan Stanley from `FullName` dictionary?
    - Create new dictionary for `CompanyName`?

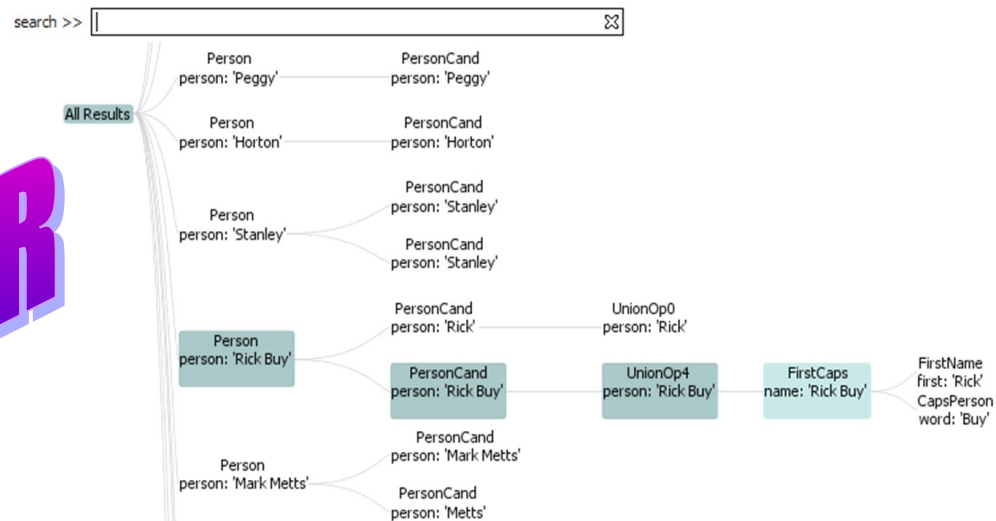


# Text Analytics Toolkit Provenance Viewer

- Major challenge for analysts is determining the lineage of changes that have been applied to text
  - REALLY difficult to discern which extractors need to be adjusted to tweak the resulting annotations
- Provenance Viewer for interactive visualizations to display output annotations for regression, debugging, version enhancements, +++
  - Reduce development of extractors by days to weeks

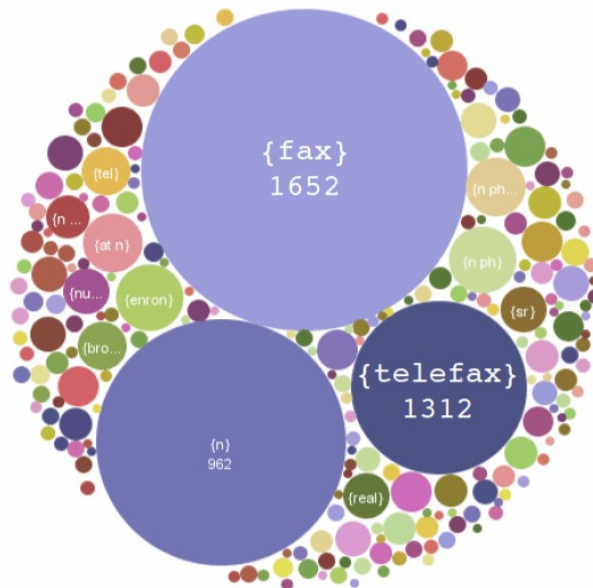


OR

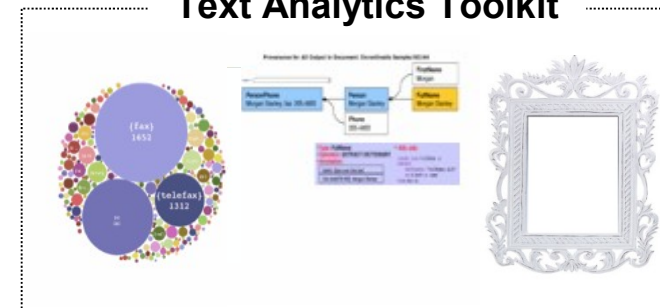


# Accelerating Analytics – discovery and Pattern Matching

- Contextual Clues discoverer cluster the context surrounding annotation in order to detect frequently occurring patterns
- Illustrate clustering results between incorrect PersonPhone pairs
  - Visualize frequent occurrences (bubble size) of clues: i.e. FAX and TELEFAX
  - Developer improves precision of annotator by adding a rule that filters out PersonPhone pairs if the Phone is preceded by a FAX clue
  - Developer finds `call at text` as key clue for rule extraction for PersonPhone



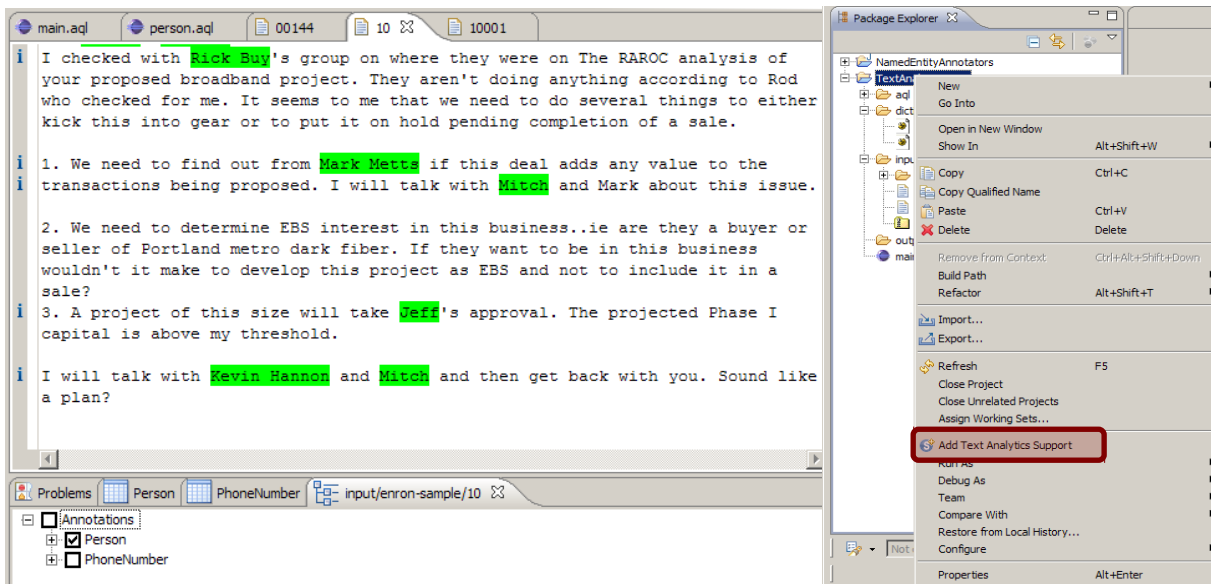
## Text Analytics Toolkit



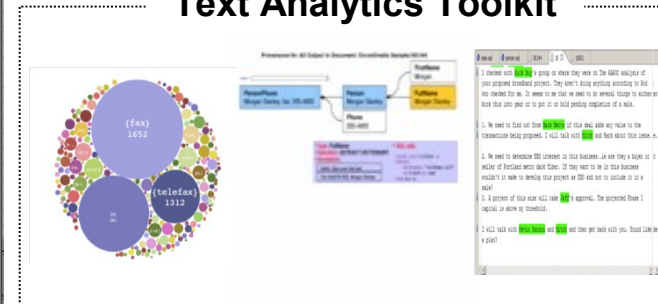


# Accelerating Analytics – Assisted Development

- IDE that exposes sophisticated techniques to assist rule developers throughout all states of the development cycle
  - Promotes agile development because it unifies the business analyst and the developer with a common toolset which fosters understanding
    - Business Analyst helps mature and validates extractor results
    - Developer understands visually that AQL enhancements needed to refine rules



## Text Analytics Toolkit



# IBM Text Analytics Toolkit - Development Toolset

The screenshot displays the Eclipse IDE interface for the IBM Text Analytics Toolkit. Several help panels are overlaid on the workspace:

- Assisted tasks:**
  - Create an extractor in a new BigInsights project
  - Organize an existing extractor within a BigInsights project
  - Import sample text analytics code
- Step-by-step task: Create and run an extractor:**
  - Create a BigInsights Project.** Create a BigInsights project to contain the files needed for a text analytics extractor, including the main annotation query language (AQL) file, dependent AQL files, dictionary (.dict) files, and UDF JAR files.
  - Create an AQL file.** Create the AQL file that is required for each extractor, which the optimizer uses as input to construct an execution plan for the extractor.
  - Set required extractor properties.** Follow the instructions in the help to provide the locations of the AQL file and the other project files for the extractor to run successfully.
  - Run the extractor.** Create launch settings in a configuration, and run the extractor.
- Perform More Tasks by Using Text Analytics Help:**
  - Visualize extractor results
  - Interpret the lineage of extractor results
  - Generate regular expressions
  - Discover patterns in the data
  - View differences in the results of two extractor runs
  - Evaluate extractor quality
  - Move extractor code to a cluster
- Procedural Help:** A sidebar on the left showing steps: Step 3: Develop Extractor, Step 4: Test Extractor, Step 5: Profile Extractor, and Step 6: Export Extractor.
- Context Sensitive Help:** A yellow box highlighting the 'output view' command in the AQL editor.
- Outline View:** A panel on the right showing a tree structure of project files and folders.
- Hyperlink Navigation:** A red dashed arrow points from the 'output view' command in the AQL editor to the 'create view' command in the 'phone-including-extensions.aql' file.

The main AQL editor shows the following code:

```
include 'phone-including-extensions.aql';
output view PhoneNumber;
```

The 'phone-including-extensions.aql' file shows the following code:

```
create view PhoneNumber as
select P.num as number
from
(
extract
regexes / (\ ) ? (\d{3}) ? (\ ) | - ? ( ) ? (\d{3}) ? (-) ? \d(4,5) /
on D.text as num
from Document D
) P;
```

# BigInsights Text Analytics Development

**Step 1 : Select Document Collection**  
**Step 2 : Label Examples and Clues**

**a. Label Example Snippets of Int**

From the open document(s), select example text snippets that you want to extract. To label an example, right-click the selected text and choose 'Add Example with New Label' or 'Label Example As'.

[Example](#)

Tip: Labeled examples are shown in the Extraction Plan.

**b. Label Extraction Clues**

From within the example snippet(s) or nearby text, define other labels to use as clues for extraction.

[Inside Clue Example](#)  
[Outside Clue Example](#)

**Step 3 : Develop Extractor**  
**Step 4 : Test Extractor**  
**Step 5 : Profile Extractor**  
**Step 6 : Export Extractor**

**Procedural Help**

```

53
54 output view acquisitiona;
55
56 Content Assist
57
58 -- Consolidation:
59 -- For now, we just remo
60 -- contained within othe
61
62 create view StateOrCount
63     (select S.match as m
64     union all
65     (select C.match as m

```

**Difference Viewer**

FOR ASSISTANCE  
 713) 853-1411 Enron Resolution Center

FOR ASSISTANCE  
 713) 853-1411 Enron Resolution Center

FOR ASSISTANCE  
 713) 853-5536  
 713) 284-3757 [Pager]  
 713) 327-3893 [Pager]  
 713) 853-9797 OR (888) 853-9797

FOR ASSISTANCE  
 713) 853-5536  
 713) 284-3757 [Pager]  
 713) 327-3893 [Pager]  
 713) 853-9797 OR (888) 853-9797

**Regular Expression Builder**

Regular Expression:  
 ((xX)?(-)?\d{4,5})

Match	Samples
YES	x-1981
YES	x9834
YES	x4926
YES	x67852

**Context Sensitive Help**

**Outline View**

**Hyperlink Navigation**

```

include 'phone-including-extensions.aql';
output view PhoneNumber;
-- Find dictionary matches for all title initials
create view Salutation as
extract dictionary 'SalutationDict'
on D.text as salutation
from Document D;
-- Dictionary of common greetings
create dictionary GreetingDict as
(
'regards', 'regds', 'hello', 'hi', 'thanks', 'bes
);

```

**Design Time Validation**

```

create view PhoneNumber as
select P.num as number
from
(
extract
regexes / (\() ? (\d{3}) ? (\)|-)? ( ) ? (\d{3}) ? (-)? \d{4,5} /
on D.text as num
from Document D
) P;

```



# Text Analytics Toolkit – Development Accelerator

- Eclipse plug-ins enhance analyst productivity
  - When writing AQL code, the editor features syntax highlighting, and automatic detection of syntax error at **design time** not runtime
  - Pre-built extractors and sample test tools, +++
- Reduce coding time and debugging by **30-50%+!**



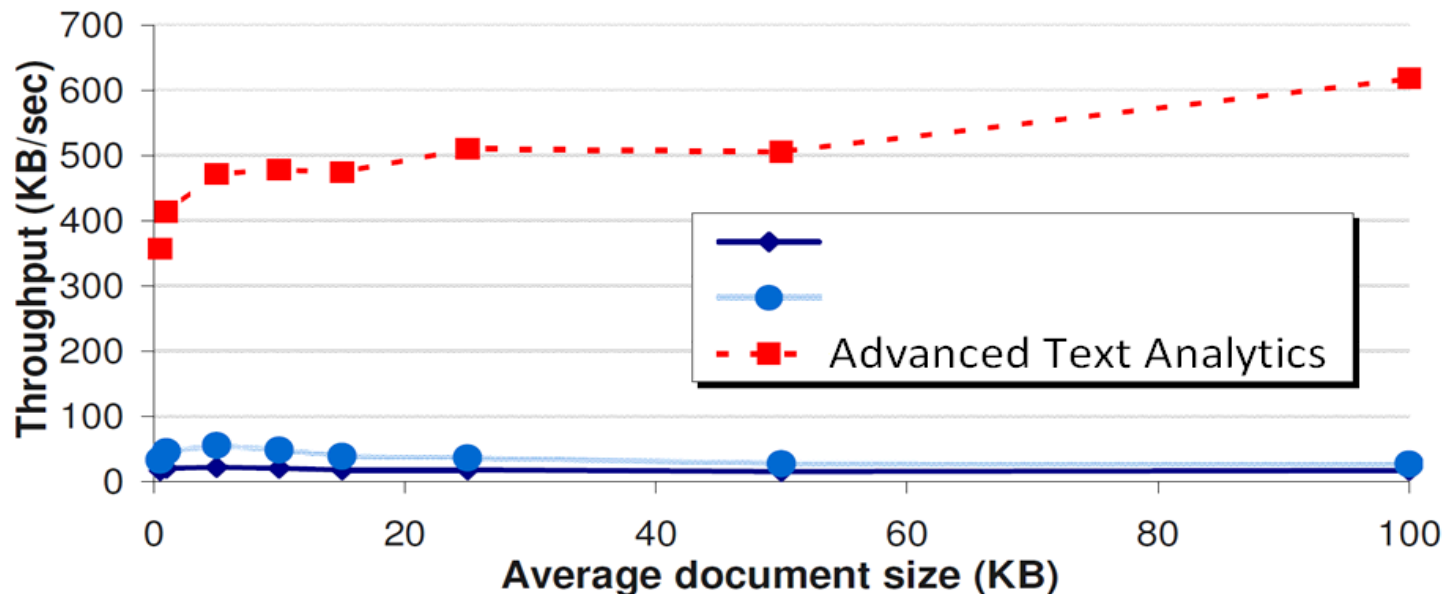
The screenshot displays the Eclipse IDE with the Bigsights Text Analytics project. The main editor shows AQL code with syntax highlighting and error detection. Several tool windows are open, providing assistance and debugging features:

- Content Assist:** Provides suggestions for AQL keywords and functions.
- Difference Viewer:** Compares two versions of a document to highlight changes.
- Regular Expression Builder:** Allows users to create and test regular expressions for text extraction.
- Context Sensitive Help:** Provides help topics relevant to the current code context.
- Hyperlink Navigation:** Enables navigation between related code elements.
- Design Time Validation:** Checks for errors in the AQL code before execution.
- Procedural Help:** Provides step-by-step guidance for development tasks.



# Text Analytics Toolkit – Performance Accelerator

- **AQL code is highly optimized for MapReduce** because of its declarative nature
- **Unlike other frameworks**, AQL optimizer determines order of execution of the extractor instructions for maximum efficiency
- Deliver analysis up to **10x faster** than other leading alternative frameworks running the same extractors



When comparing solutions, people always  
talk about how

**FAST**

Did anyone the answer was returned ever ask if the answer

was

**CORRECT?**



# Business Task: Find the Pictures that Have Cats





# BUT! Your Application Returns the Following...

## PRECISION

(a measure of exactness)

# 2/4

## RECALL

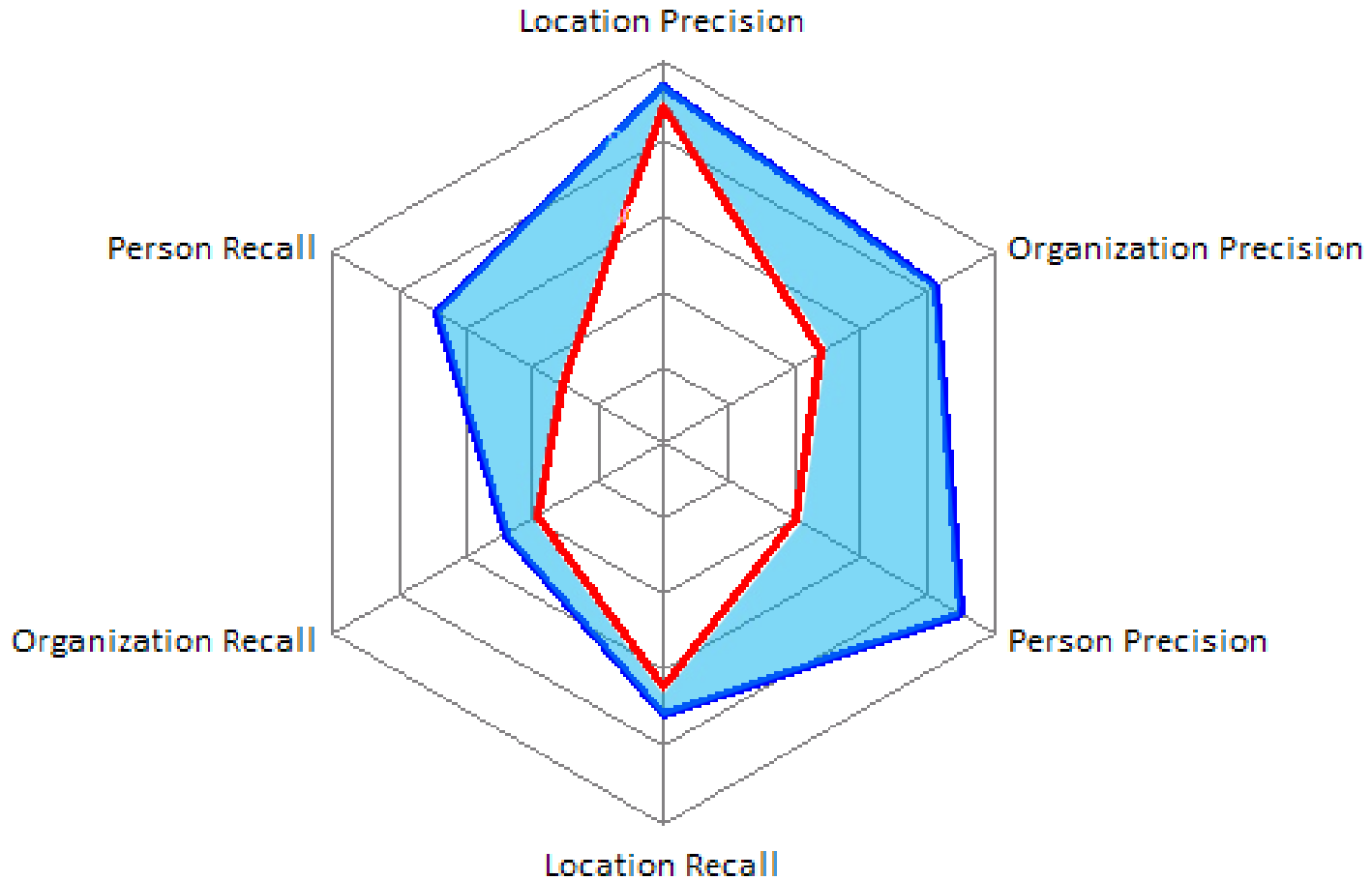
(a measure of coverage)

# 2/5





# IBM Finds RIGHT Answers Better Than Anyone



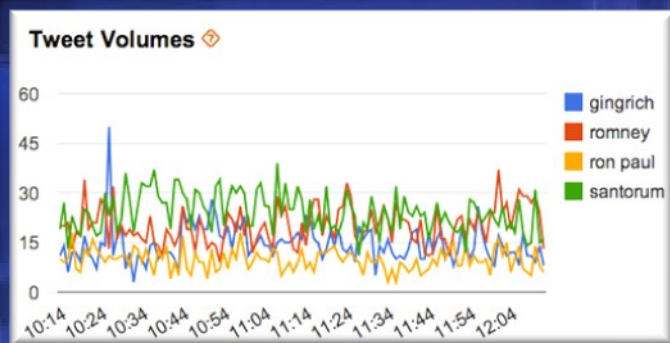
# IBM's Hadoop System Provides Unique Business Value

- **Optimized beyond open source Hadoop**
  - Workload optimization, security, +++
- **Integration with enterprise systems**
  - Connectors for multiple data sources
- **Accelerators reduce development and implementation times**
  - Industry & application accelerators
  - Analytic accelerators
- **Visualization tools enable business users to explore Big Data**



# University of Southern Cal Political Debate Monitoring

- Solution to measure public sentiment during the Republican Primary and Presidential Debates
- Examines trends, volume, and content of millions of public Twitter messages in real-time
- Analytic accelerators to understand sentiment (positive, negative, neutral)
  - Stream Computing and visualization
- Benefits
  - Real-time display of public sentiment as candidates respond to questions
  - Debate winner prediction based on public opinion instead of solely political analysts





# Cisco turns to IBM big data for intelligent infrastructure management

- **Optimize building energy consumption with centralized monitoring and control of building monitoring system**
- **Automates preventive and corrective maintenance of building corrective systems**
- **Uses Streams, InfoSphere BigInsights and Cognos**
  - Log Analytics
  - Energy Bill Forecasting
  - Energy consumption optimization
  - Detection of anomalous usage
  - Presence-aware energy mgt.
  - Policy enforcement





# Stream Computing Provides Unique Business Value

- **Real-time answers = low latency insight**
  - Better outcomes for time sensitive applications (e.g. fraud detection, network management)
- **Solution when data is too large or expensive to store**
  - Analyze data as it comes to you
  - Persist data of interest for deeper analysis
- **Insights derived across multiple streams**
  - Fuse streams for new insights





## Asian telco reduces billing costs and improves customer satisfaction

### Capabilities:

Stream Computing  
Analytic Accelerators

Real-time mediation and analysis of  
**6B CDRs per day**

Data processing time reduced from  
**12 hrs to 1 sec**

**Hardware cost reduced to 1/8<sup>th</sup>**

Proactively address issues  
(e.g. dropped calls) impacting  
customer satisfaction.

# Data Warehousing Provides Unique Business Value

- **Consolidate, manage and reconcile data for enterprise business intelligence**
- **Establish trust, quality and governance where necessary**
  - Financial data
  - Credit card data
  - Healthcare
- **Combine deep and operational analytics**
- **Maintain history for trending and historical reporting**





# Pacific Northwest Smart Grid Demonstration Project

## Capabilities:

**Stream Computing – real-time control system**

**Deep Analytics Appliance – analyze massive data sets**

**Demonstrates scalability from 100 to 500K homes while retaining 10 years' historical data**

**60k metered customers in 5 states**

**Accommodates ad hoc analysis of price fluctuation, energy consumption profiles, risk, fraud detection, grid health, etc.**





# Information Integration Provides Unique Business Value

- **Movement of large data sets in batch and real time**
  - Parallel processing engine for efficient data movement
- **Governance and trust for Big Data**
  - Lineage and meta data of new Big Data data sources
  - Profile sources to determine trust
- **Data Quality**
  - Standardize and transform data



# Marketing Services Leader integrates big data for customer intelligence

## Capabilities Utilized:

Information Integration – data  
quality, ETL

Deep Analytics Appliance

Complex customer data  
integration for

**54M records/hour**

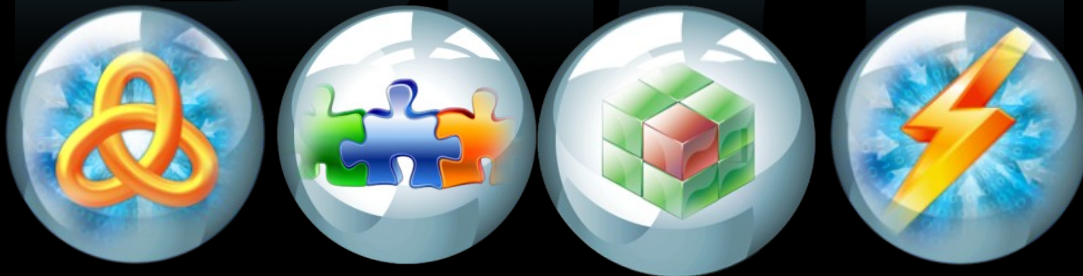
Processing

**5B simultaneous records**



# THINK

# BIG



## Understanding Big Data

**Analytics for Enterprise Class  
Hadoop and Streaming Data**

- Learn how IBM hardens Hadoop for enterprise-class scalability and reliability
- Gain insight into IBM's unique in-motion and at-rest Big Data analytics platform
- Learn tips and tricks for Big Data use cases and solutions
- Get a quick Hadoop primer

CHRIS EATON  
TOM DEUTSCH

DIRK DEROOS  
GEORGE LAPIS

PAUL ZIKOPOULOS

PDF <http://tinyurl.com/88auawg>  
Hardcopy <http://tinyurl.com/7ef9ubs>