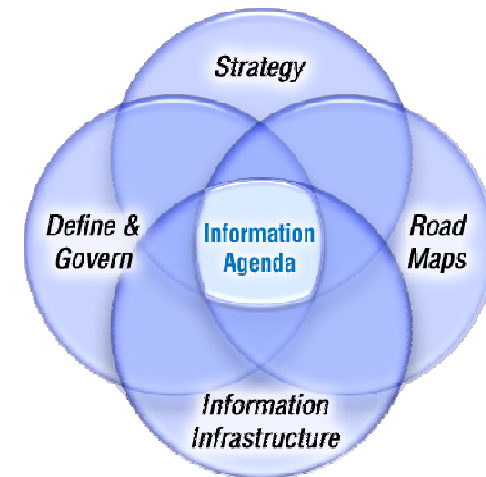Wolfgang Nimfuehr– Dipl. Ing.

Big Data Enterprise Class

Information Agenda Executive Consultant
Big Data Tiger Team
IBM Software Group Europe
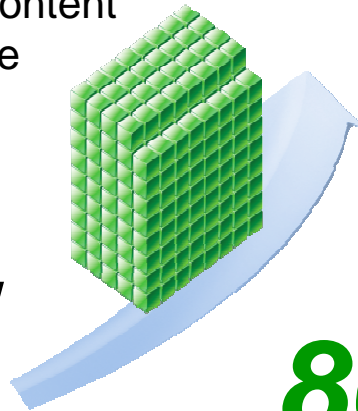
The Information Explosion in Data and Real World Events

44x

as much Data and Content Over Coming Decade

2020
35 zettabytes

2009
800,000 petabytes

Velocity

Variety

Volume

80%

Of world's data is unstructured

SI decimal prefixes		Binary usage
Name (Symbol)	Value	
kilobyte (kB)	10^3	2^{10}
megabyte (MB)	10^6	2^{20}
gigabyte (GB)	10^9	2^{30}
terabyte (TB)	10^{12}	2^{40}
petabyte (PB)	10^{15}	2^{50}
exabyte (EB)	10^{18}	2^{60}
zettabyte (ZB)	10^{21}	2^{70}
yottabyte (YB)	10^{24}	2^{80}

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

Organizations Need Deeper Insights

Challenge
Study a Large Volume and Variety of Data to Find New Insights

Multi-channel customer sentiment and experience a analysis

Support medical diagnostics Detect life-threatening conditions

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

Make risk decisions and frauds detection based on real-time transactional data

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

Leveraging Big Data Analytics

How do you address the challenges presented by empowered market participants generating mountains of data?

People and devices are talking more than ever...

30 billion RFID tags embedded into our world

6 billion mobile phone subscriptions

1 billion transistors per human

2 billion people on the Internet

big data

15 petabytes of new information per day

7 trillion text messages

1 Can you capture data generated by these interactions?

2 Can you do it in real-time?

3 Can you turn that data into insights to predict customer / competitive / market behavior?
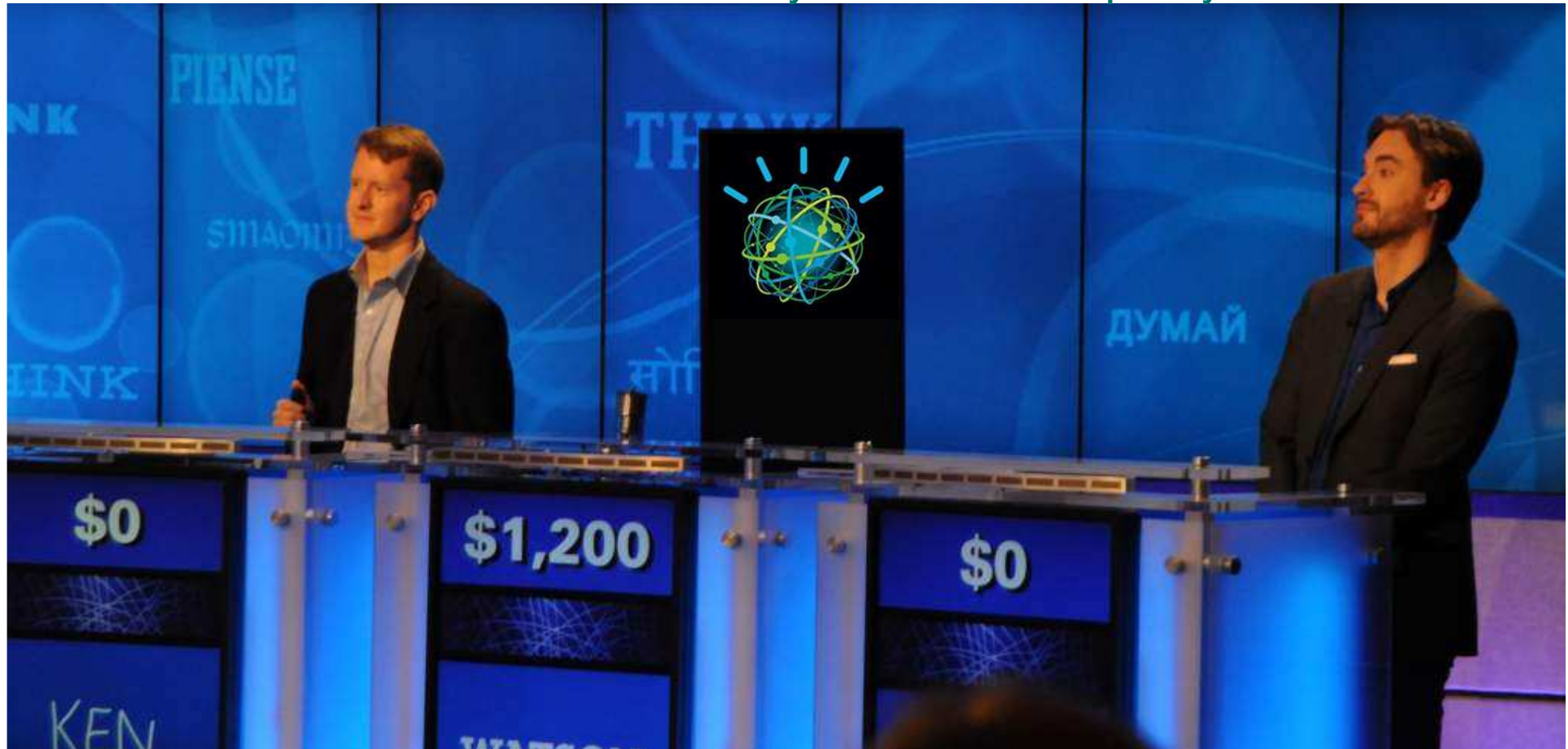
Source:
1 – Barrera, Clod and Wojtowecz. "Cloud Leads Five Storage Trends for 2011." CIO. January 27, 2011 .
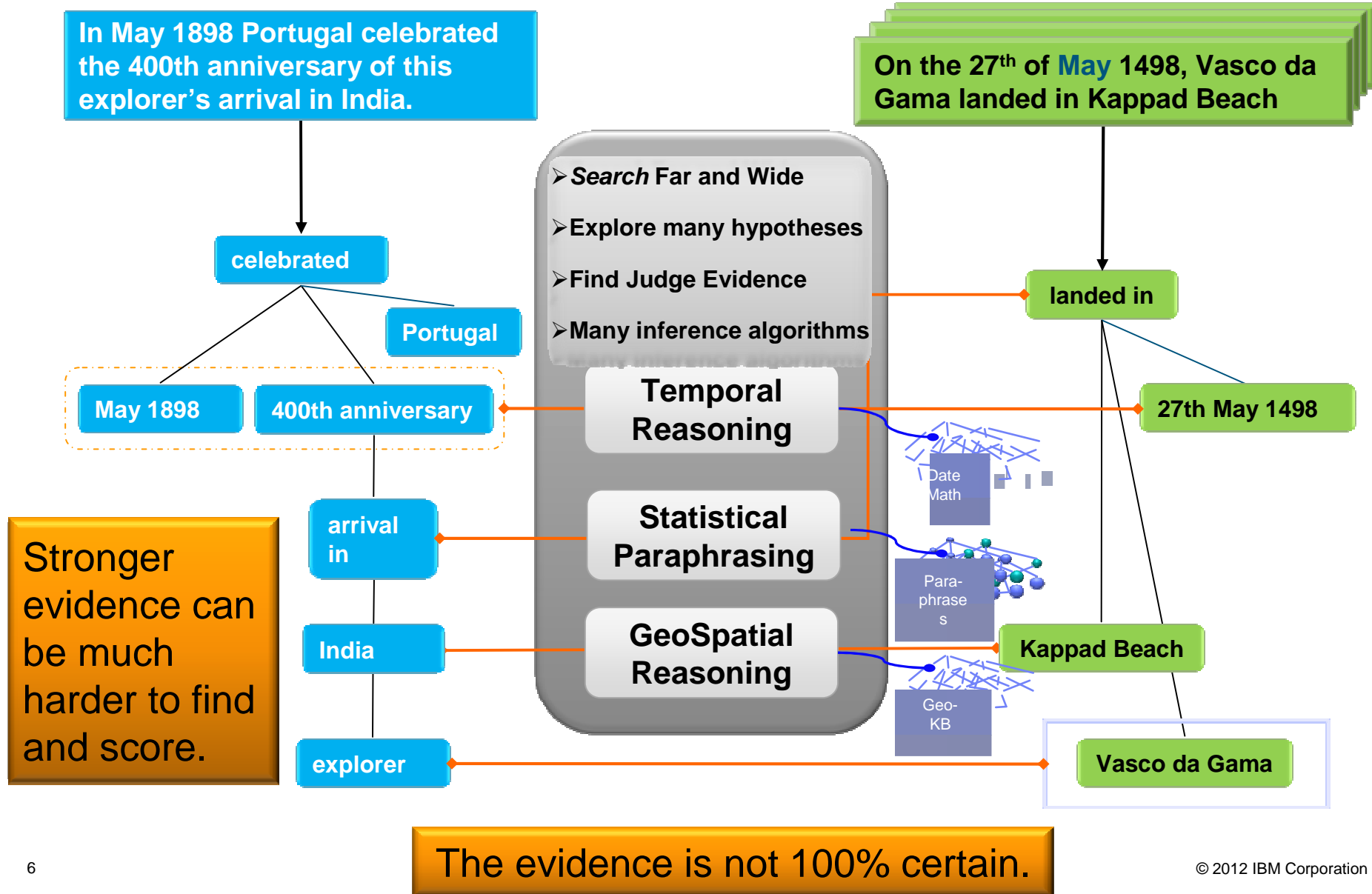2 – http://www.internetworldstats.com/stats.htm.
3 – http://www.abiresearch.com/press/3584-More+than+Seven+Trillion+SMS+Messages+Will+Be+Sent+in+2011

4

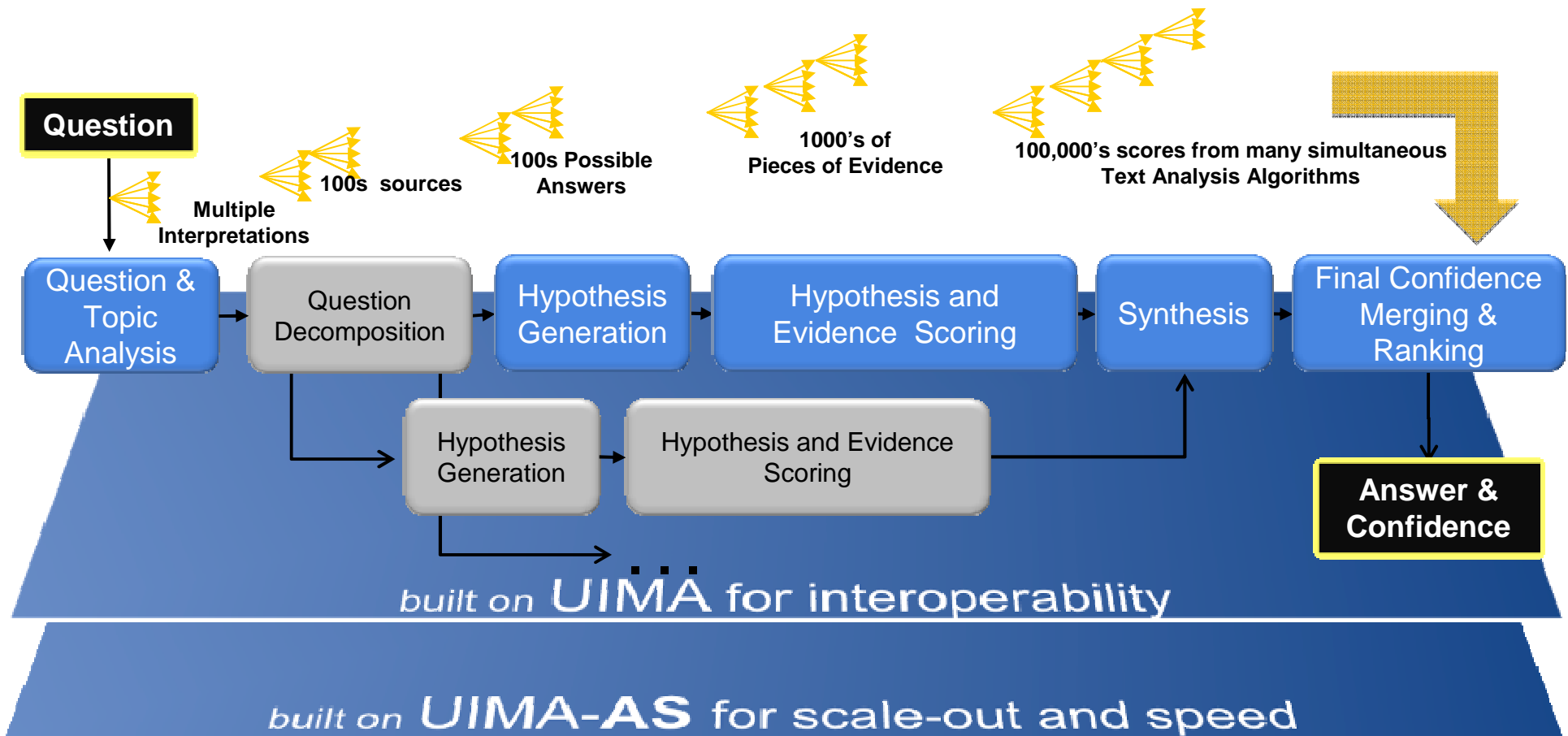## On Feb 16 2011 the IBM Watson system won Jeopardy!

Can we design a computing system that rivals a human's ability to answer questions posed in natural language, interpreting meaning and context and retrieving, analyzing and understanding vast amounts of information in real-time?

Different Types of Evidence: Deeper Evidence

In May 1898 Portugal celebrated the 400th anniversary of this explorer's arrival in India.

On the 27th of May 1498, Vasco da Gama landed in Kappad Beach

celebrated

Portugal

> *Search* Far and Wide
>
> Explore many hypotheses
>
> Find Judge Evidence
>
> Many inference algorithms

landed in

May 1898

400th anniversary

Temporal Reasoning

27th May 1498

Date Math

Statistical Paraphrasing

Para-phrases

arrival in

Stronger evidence can be much harder to find and score.

India

GeoSpatial Reasoning

Kappad Beach

Geo-KB

explorer

Vasco da Gama

The evidence is not 100% certain.

6

© 2012 IBM Corporation

Deep QA
Massively Parallel Probabilistic Evidence-Based Architecture

Question

Multiple Interpretations

100s sources

100s Possible Answers

1000's of Pieces of Evidence

100,000's scores from many simultaneous Text Analysis Algorithms

| Question & Topic Analysis | Question Decomposition | Hypothesis Generation | Hypothesis and Evidence Scoring | Synthesis | Final Confidence Merging & Ranking |

Hypothesis Generation → Hypothesis and Evidence Scoring

Answer & Confidence

built on UIMA for interoperability

built on UIMA-AS for scale-out and speed

Big Data use cases across all industries

Financial Services
- Fraud detection
- Risk management
- 360° View of the Customer

Utilities
- Weather impact analysis on power generation
- Transmission monitoring
- Smart grid management

Transportation
- Weather and traffic impact on logistics and fuel consumption

IT
- Transition log analysis for multiple transactional systems
- Cybersecurity

Health & Life Sciences
- Epidemic early warning system
- ICU monitoring
- Remote healthcare monitoring

Retail
- 360° View of the Customer
- Click-stream analysis
- Real-time promotions

Telecommunications
- CDR processing
- Churn prediction
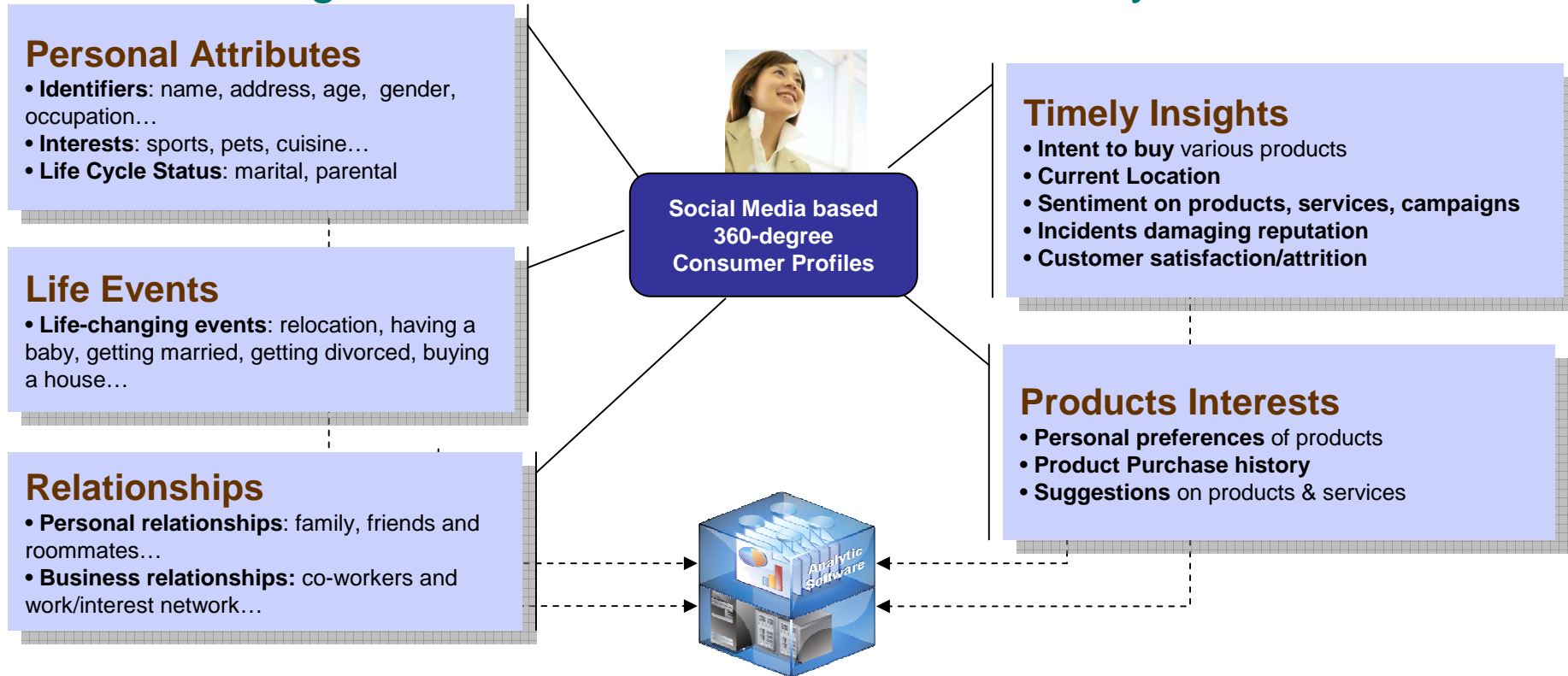- Geomapping / marketing
- Network monitoring

Law Enforcement
- Real-time multimodal surveillance
- Situational awareness
- Cyber security detection
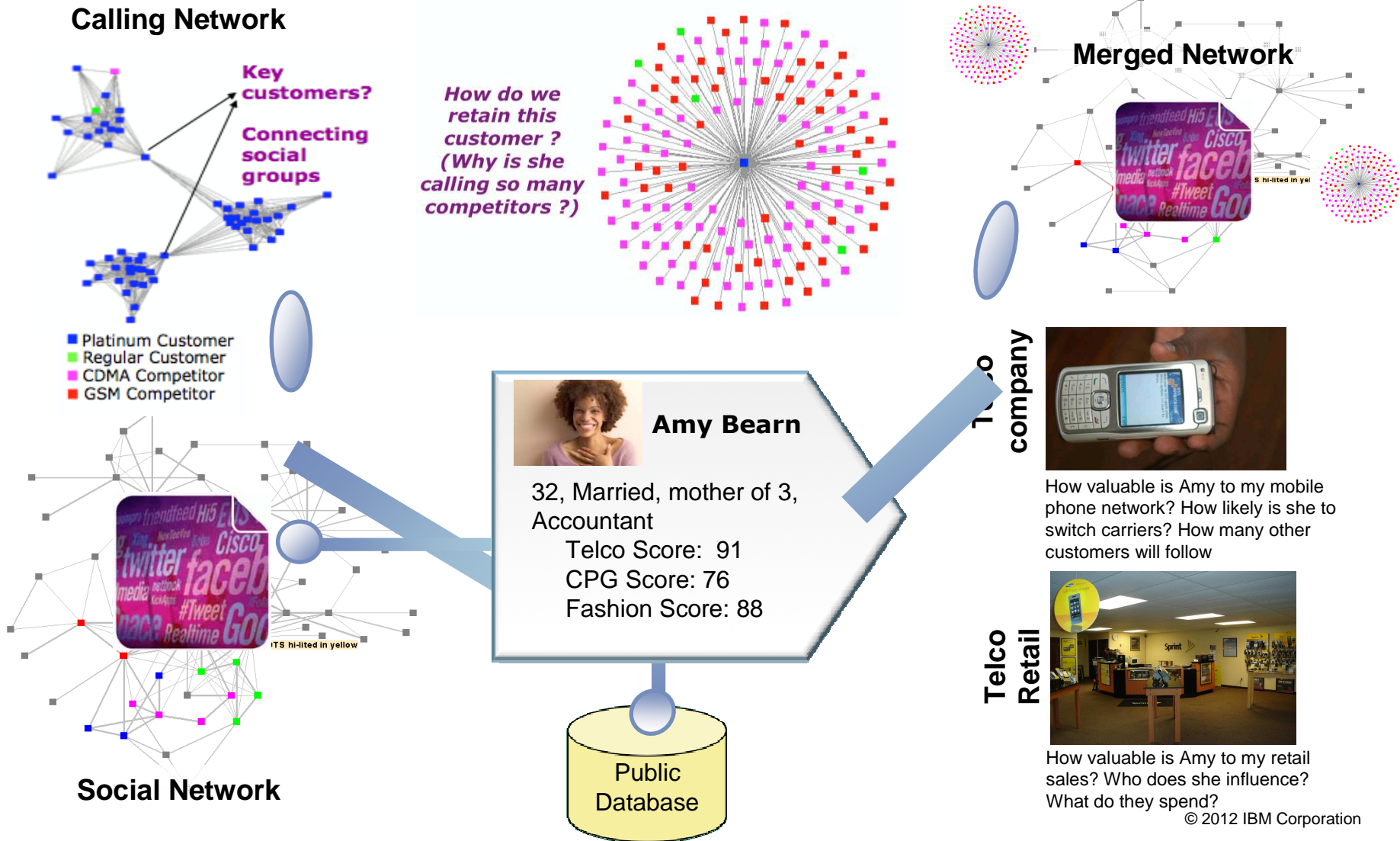
Use Case: Big Data 360° Profile / Sentiment Analysis

Personal Attributes
- **Identifiers**: name, address, age, gender, occupation…
- **Interests**: sports, pets, cuisine…
- **Life Cycle Status**: marital, parental

Life Events
- **Life-changing events**: relocation, having a baby, getting married, getting divorced, buying a house…

Relationships
- **Personal relationships**: family, friends and roommates…
- **Business relationships**: co-workers and work/interest network…

Social Media based 360-degree Consumer Profiles

Timely Insights
- **Intent to buy** various products
- **Current Location**
- **Sentiment on products, services, campaigns**
- **Incidents damaging reputation**
- **Customer satisfaction/attrition**

Products Interests
- **Personal preferences** of products
- **Product Purchase history**
- **Suggestions** on products & services

Analytic Software

Monetizable intent to buy products

I need a new digital camera for my food pictures, any recommendations around 300?

What should I buy?? A mini laptop with Windows 7 OR a Apple MacBook!??!

Location announcements

I'm at Starbucks Parque Tezontle http://4sq.com/fYReSj

Life Events

College: Off to Stanford for my MBA! Bbye chicago!

Looks like we'll be moving to New Orleans sooner than I thought.

Intent to buy a house

I'm thinking about buying a home in Buckingham Estates per a recommendation. Anyone have advice on that area? #atx #austinrealestate #austin
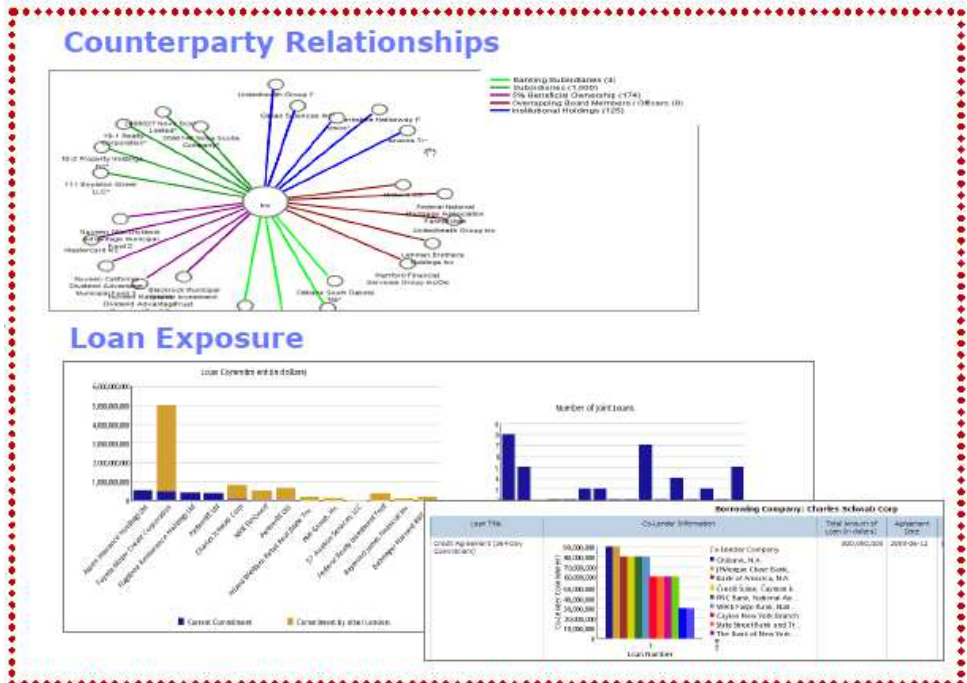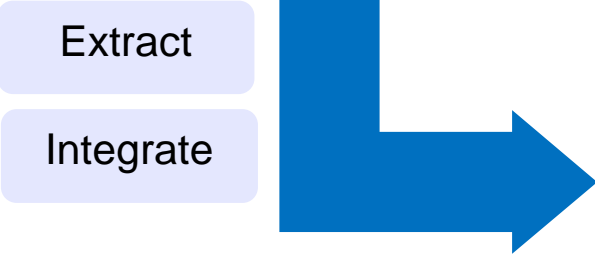
Use Case: Institutional Risk Application
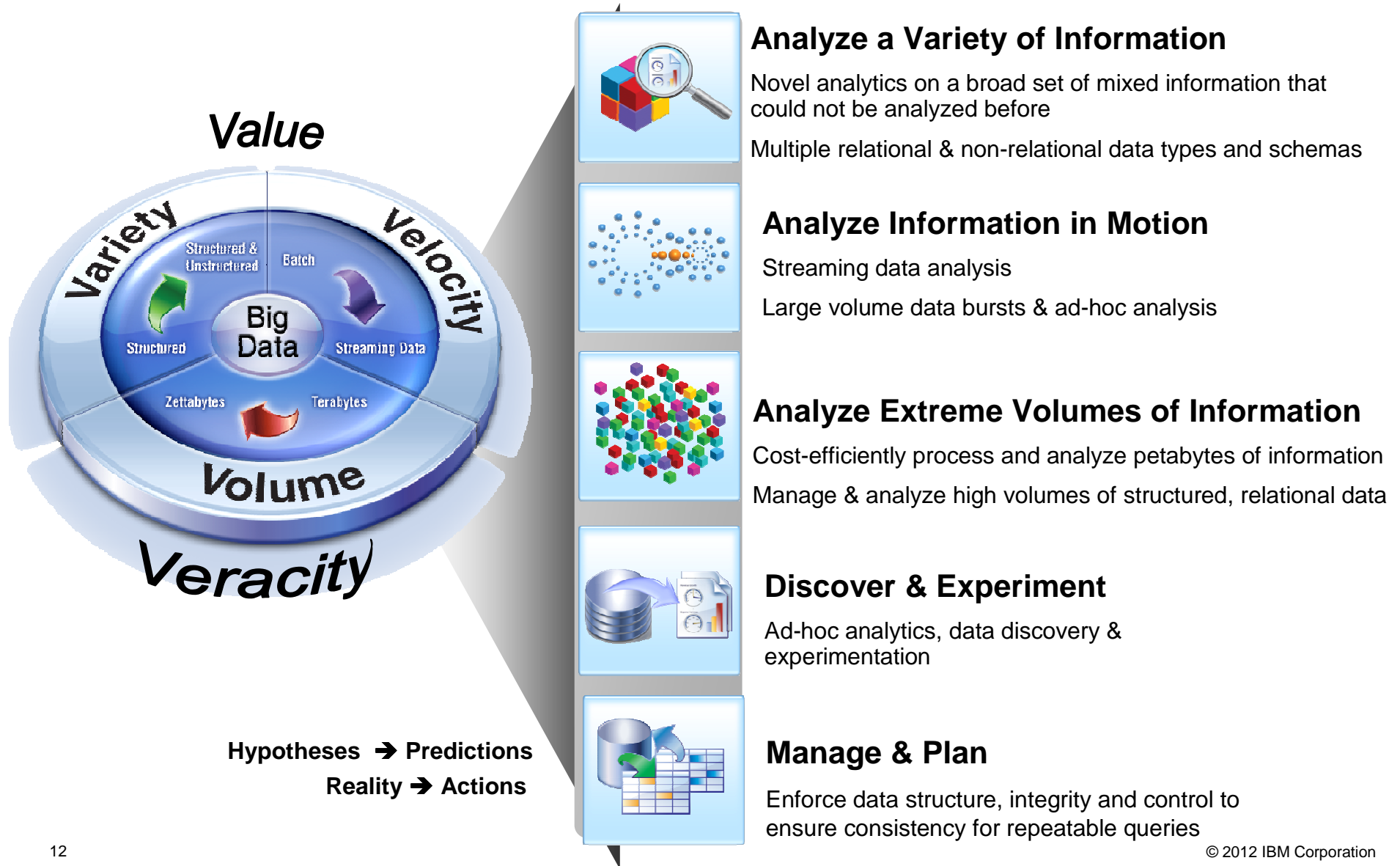# Comprehensive view of publicly traded companies and related people based on regulatory filings

Requirements for a Big Data Solution Platform

Value

Variety *Velocity*

Structured & Unstructured | Batch

Big Data

Structured | Streaming Data

Zettabytes | Terabytes

Volume

Veracity

Hypotheses ➔ Predictions
Reality ➔ Actions

Analyze a Variety of Information

Novel analytics on a broad set of mixed information that could not be analyzed before

Multiple relational & non-relational data types and schemas

Analyze Information in Motion

Streaming data analysis

Large volume data bursts & ad-hoc analysis

Analyze Extreme Volumes of Information

Cost-efficiently process and analyze petabytes of information

Manage & analyze high volumes of structured, relational data

Discover & Experiment

Ad-hoc analytics, data discovery & experimentation

Manage & Plan

Enforce data structure, integrity and control to ensure consistency for repeatable queries

Big Data Analytics harnesses the value of the new mix of information

IBM Data Big Analytics

BI / Reporting	Exploration / Visualization	Functional App	Industry App	Predictive Analytics	Content Analytics

IBM Big Data Platform

Visualization & Discovery	Application Development	Systems Management

Accelerators

Hadoop System	Stream Computing	Data Warehouse

Information Integration & Governance
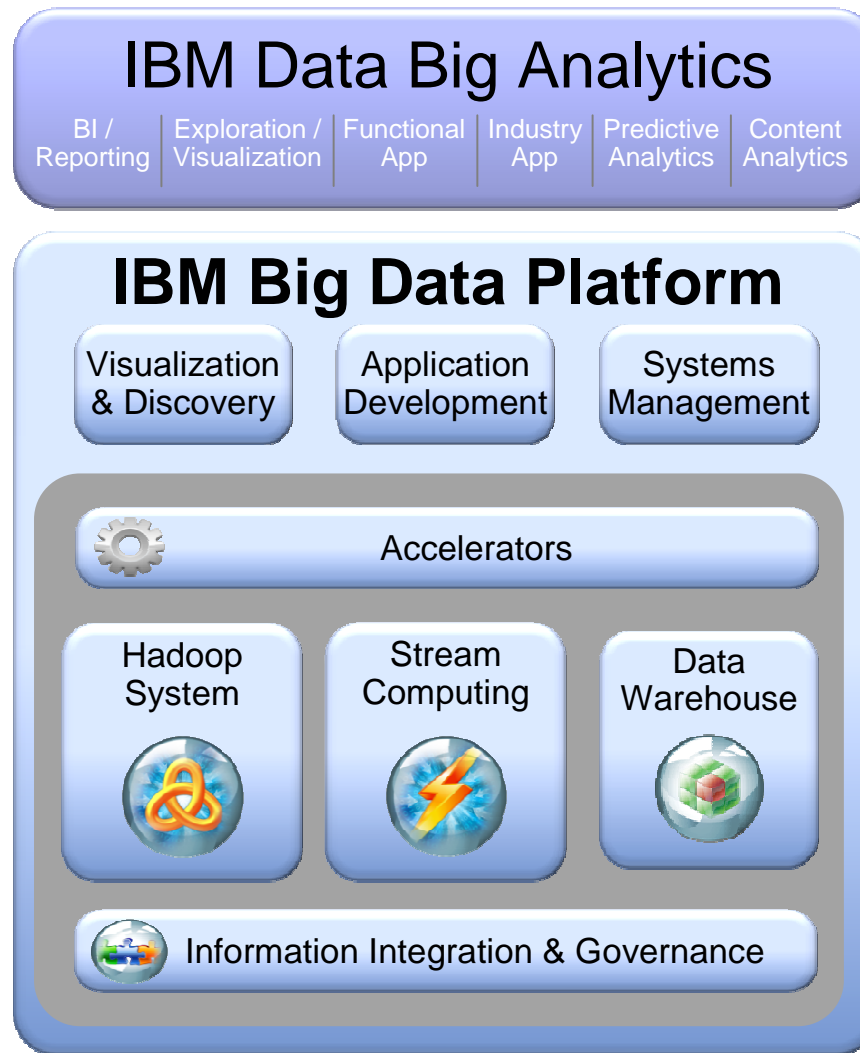
IBM Vivissimo Velocity
Unique Federated Discovery and Navigation Technology

Advanced On-the-Fly Analytics

- State-of-the-art real-time text and meta-data analytics

Accuracy

- More relevant results due to position-based indexing

Security

- Respects the security rights of underlying systems

Scalability

- Scales to trillions of records

Integration

- Rich connectors to a variety of data sources

IBM Big Data Analytics

| BI / Reporting | Exploration / Visualization | Functional App | Industry App | Predictive Analytics | Content Analytics |

IBM Big Data Platform

Visualization & Discovery | Application Development | Systems Management

Accelerators

Hadoop System | Stream Computing | Data Warehouse

Information Integration & Governance
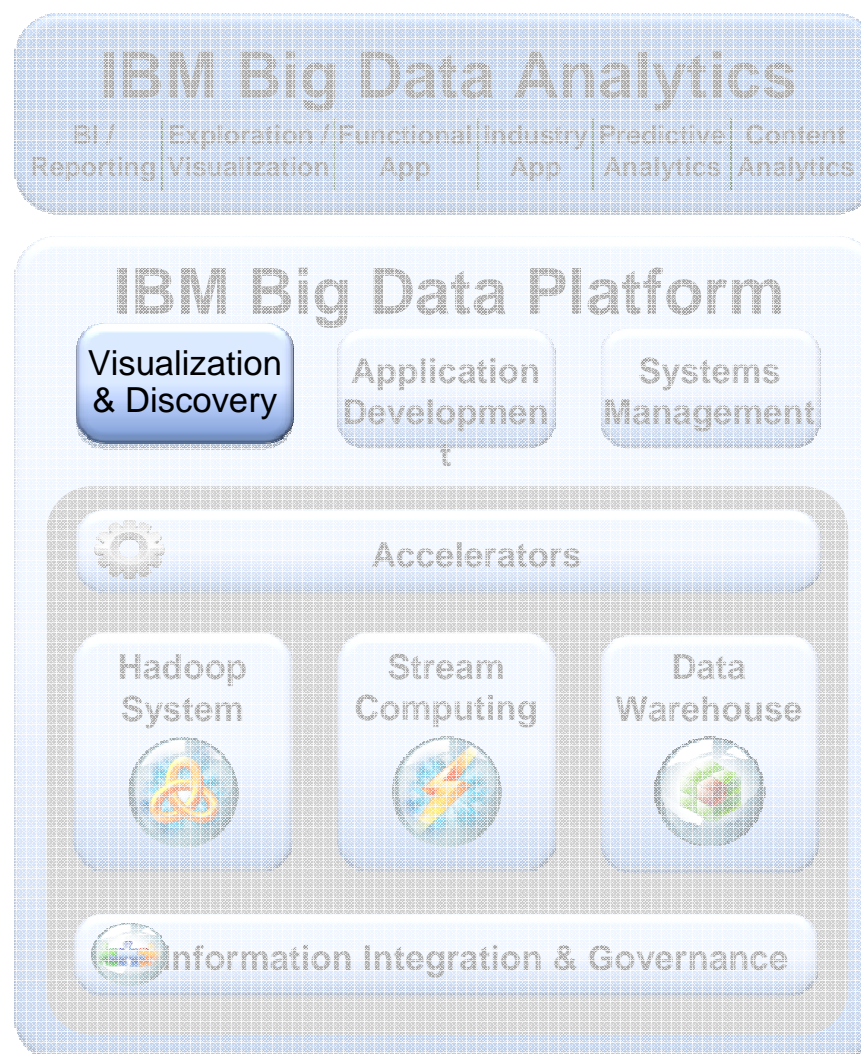
Vivissimo Velocity
Navigating Enterprise Class Information across different sources

InfoSphere BigInsights
Analytical platform for Big Data at-rest

Based on open source & IBM technologies
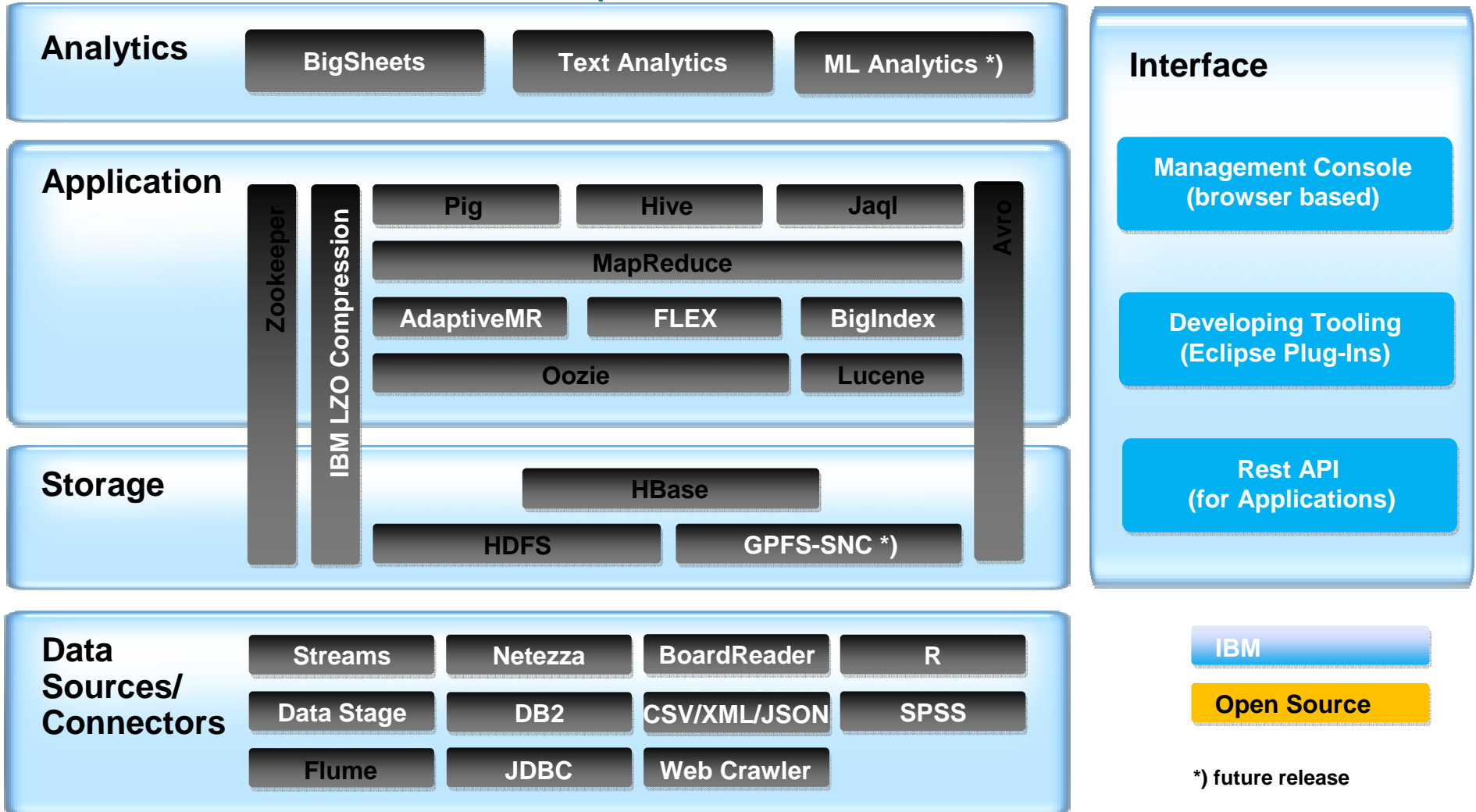
Distinguishing characteristics

- Built-in analytics enhances business knowledge

- Enterprise software integration complements and extends existing capabilities

- Production-ready platform with tooling for analysts, developers, and administrators speeds time-to-value and simplifies development/maintenance

IBM advantage

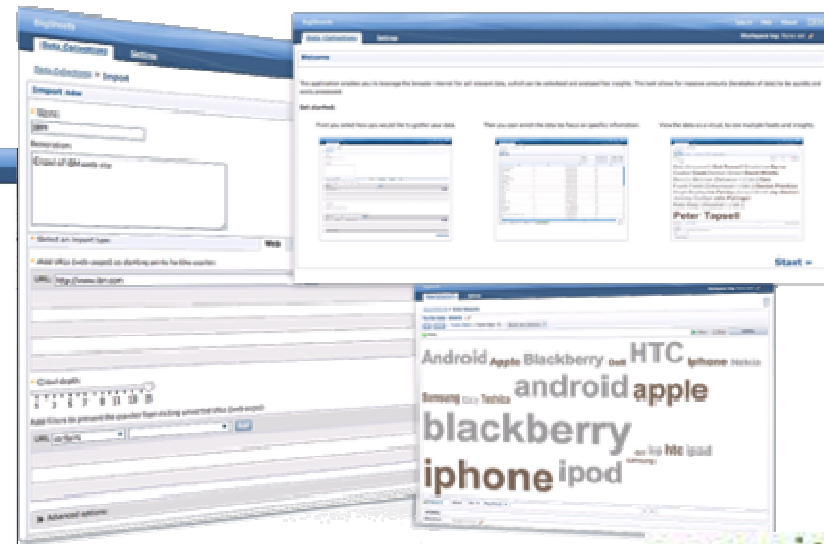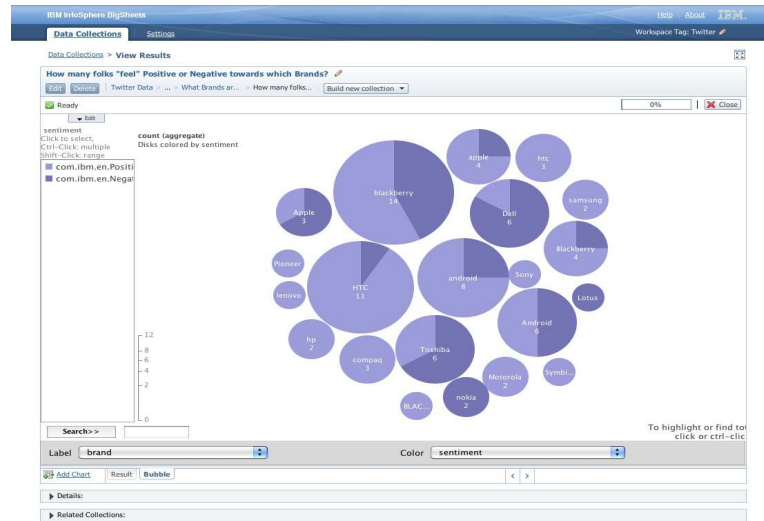- Combination of software, hardware, services and advanced research

IBM Big Data Analytics

| BI / Reporting | Exploration / Visualization | Functional App | Industry App | Predictive Analytics | Content Analytics |

IBM Big Data Platform

Visualization & Discovery   Application Development   Systems Management

Accelerators

Hadoop System   Stream Computing   Data Warehouse

Information Integration & Governance

InfoSphere BigInsights
Embrace and Extend Hadoop

Analytics

| BigSheets | Text Analytics | ML Analytics *) |

Interface

Application

Zookeeper

IBM LZO Compression

| Pig | Hive | Jaql |

MapReduce

| AdaptiveMR | FLEX | BigIndex |

| Oozie | Lucene |

Avro

**Management Console
(browser based)**

**Developing Tooling
(Eclipse Plug-Ins)**

Storage

HBase

HDFS   GPFS-SNC *)

**Rest API
(for Applications)**

**Data
Sources/
Connectors**

Streams	Netezza	BoardReader	R
Data Stage	DB2	CSV/XML/JSON	SPSS
Flume	JDBC	Web Crawler

IBM

Open Source

*) future release
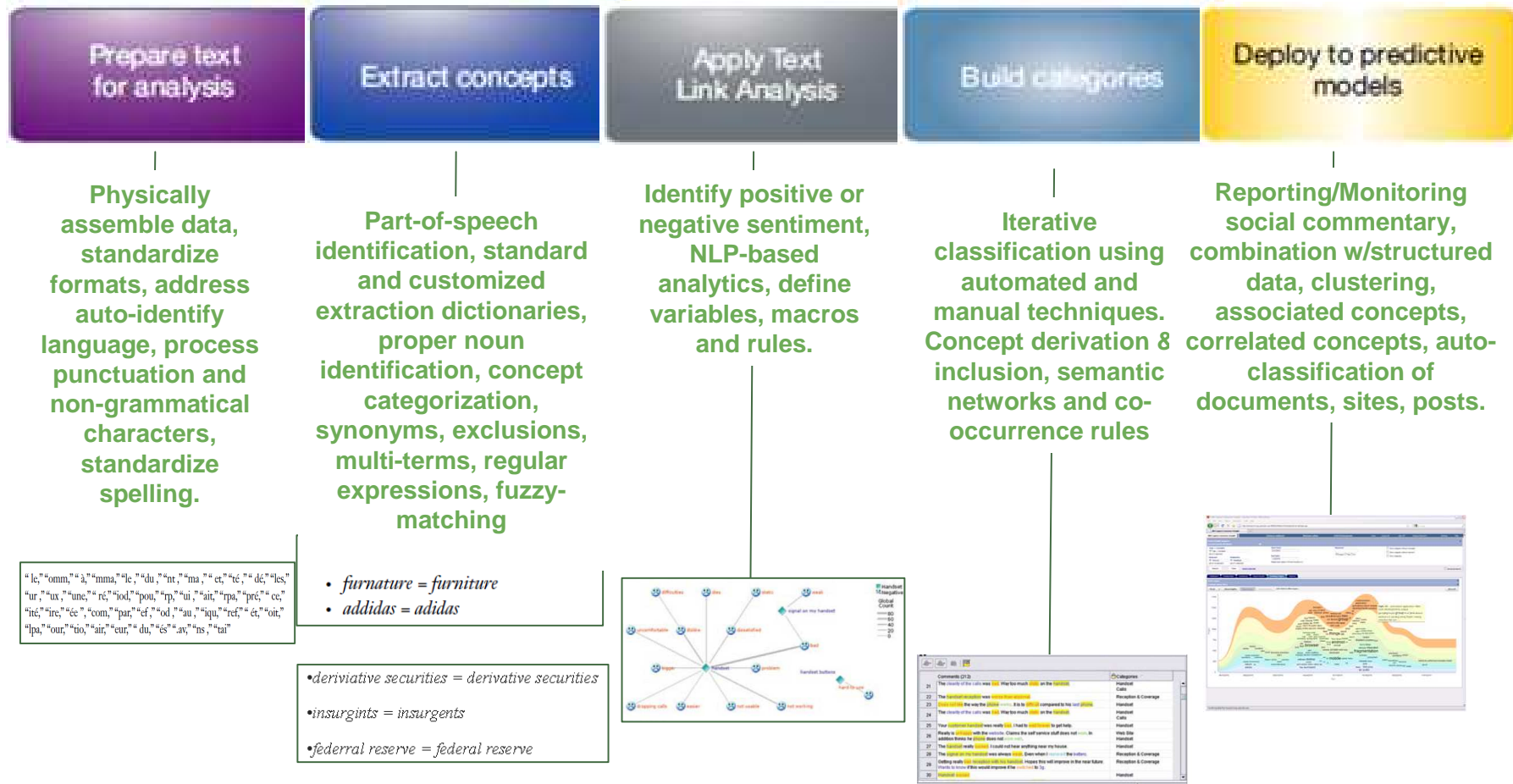
BigSheets (Many Eyes)
A visual tool for data manipulation and prototyping

- Ad-hoc analytics for LOB user

- Analyze a variety of data - unstructured and structured

- Spreadsheet metaphor for exploring/ visualizing data

- Browser-based

Text Analytics
Turns disparate words into measurable insights

Prepare text for analysis	Extract concepts	Apply Text Link Analysis	Build categories	Deploy to predictive models
Physically assemble data, standardize formats, address auto-identify language, process punctuation and non-grammatical characters, standardize spelling.	**Part-of-speech identification, standard and customized extraction dictionaries, proper noun identification, concept categorization, synonyms, exclusions, multi-terms, regular expressions, fuzzy-matching**	**Identify positive or negative sentiment, NLP-based analytics, define variables, macros and rules.**	**Iterative classification using automated and manual techniques. Concept derivation & inclusion, semantic networks and co-occurrence rules**	**Reporting/Monitoring social commentary, combination w/structured data, clustering, associated concepts, correlated concepts, auto-classification of documents, sites, posts.**

" le," "omm," " à," "mma," "le ," "du ," "nt ," "ma ," " et," "té ," " dé," "les," "ur ," "ux ," "une," " ré," "iod," "pou," "rp," "ui ," "ait," "rpa," "pré," " ce," "ité," "ire," "ée ","com," "par," "ef ,""od ," "au ," "iqu," "ref," " ét," "oit," "lpa," "our," "tio," "air," "eur," " du," "és" ".av," "ns ," "tai"

- *furnature = furniture*
- *addidas = adidas*

- *deriviative securities = derivative securities*
- *insurgints = insurgents*
- *federral reserve = federal reserve*

Pre-configured text annotators ready for distributed processing on Big Data

Support for native languages including double-byte

Vestas optimizes capital investments based on **2.5 Petabytes** of information

Capabilities Utilized

Hadoop System

Results

▪ Model the weather to optimize placement of turbines, maximizing power generation and longevity.

▪ Reduce time required to identify placement of turbine from weeks to hours.

▪ Incorporate 2.5 PB of structured and semi-structured information flows. Data volume expected to grow to 6 PB.

Reduced turbine placement identification from weeks to hours

Vestas

InfoSphere Streams
Analytical platform for Big Data in-motion

Built to analyze data in motion

- Multiple concurrent input streams
- Massive scalability

Process and analyze a variety of data

- Structured, unstructured content, video, audio
- Advanced analytic operators

IBM Big Data Analytics

BI / Reporting	Exploration / Visualization	Functional App	Industry App	Predictive Analytics	Content Analytics

IBM Big Data Platform

Visualization & Discovery

Application Development

Systems Management

Accelerators

Hadoop System

Stream Computing

Data Warehouse

Information Integration & Governance

InfoSphere Streams
Massively Scalable Stream Analytics

Linear Scalability

- Clustered deployments – unlimited scalability

Automated Deployment

- Automatically optimize operator deployment across clusters

Performance Optimization

- JVM Sharing – minimize memory use

- Fuse operators on same cluster

- Telco client – 25 Million messages per second

Analytics on Streaming Data

- Analytic accelerators for a variety of data types

- Optimized for real-time performance

Deployments

Source Adapters	Analytic Operators	Sync Adapters

Streams Studio IDE

Automated and Optimized Deployment

Streaming Data Sources

Streams Runtime

Visualization
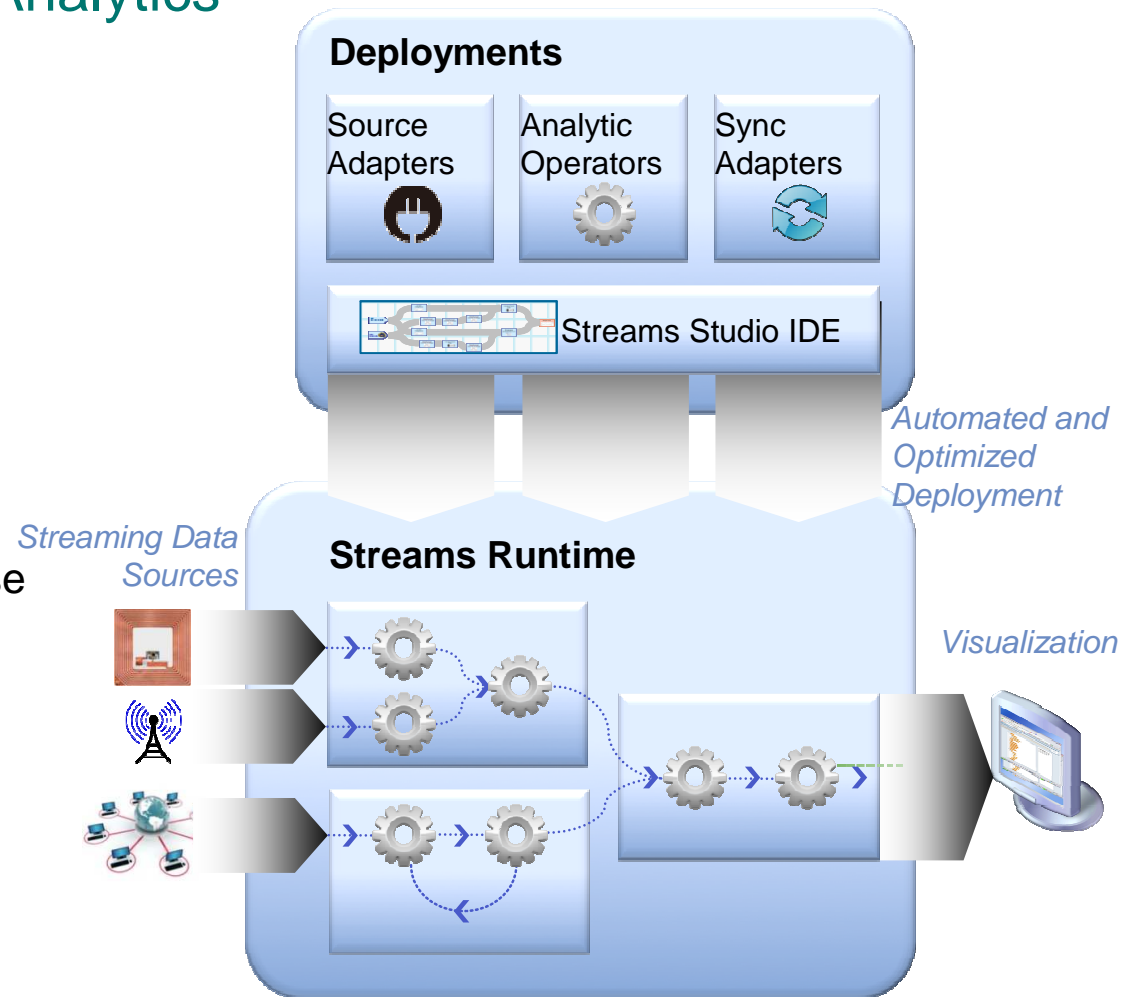
Streams approach illustrated

tuple

directory: "/img" filename: "farm"
directory: "/img" filename: "bird"
directory: "/opt" filename: "java"
directory: "/img" filename: "cat"

height: 640 width: 480 data:
height: 1280 width: 1024 data:
height: 640 width: 480 data:

Search and Explore
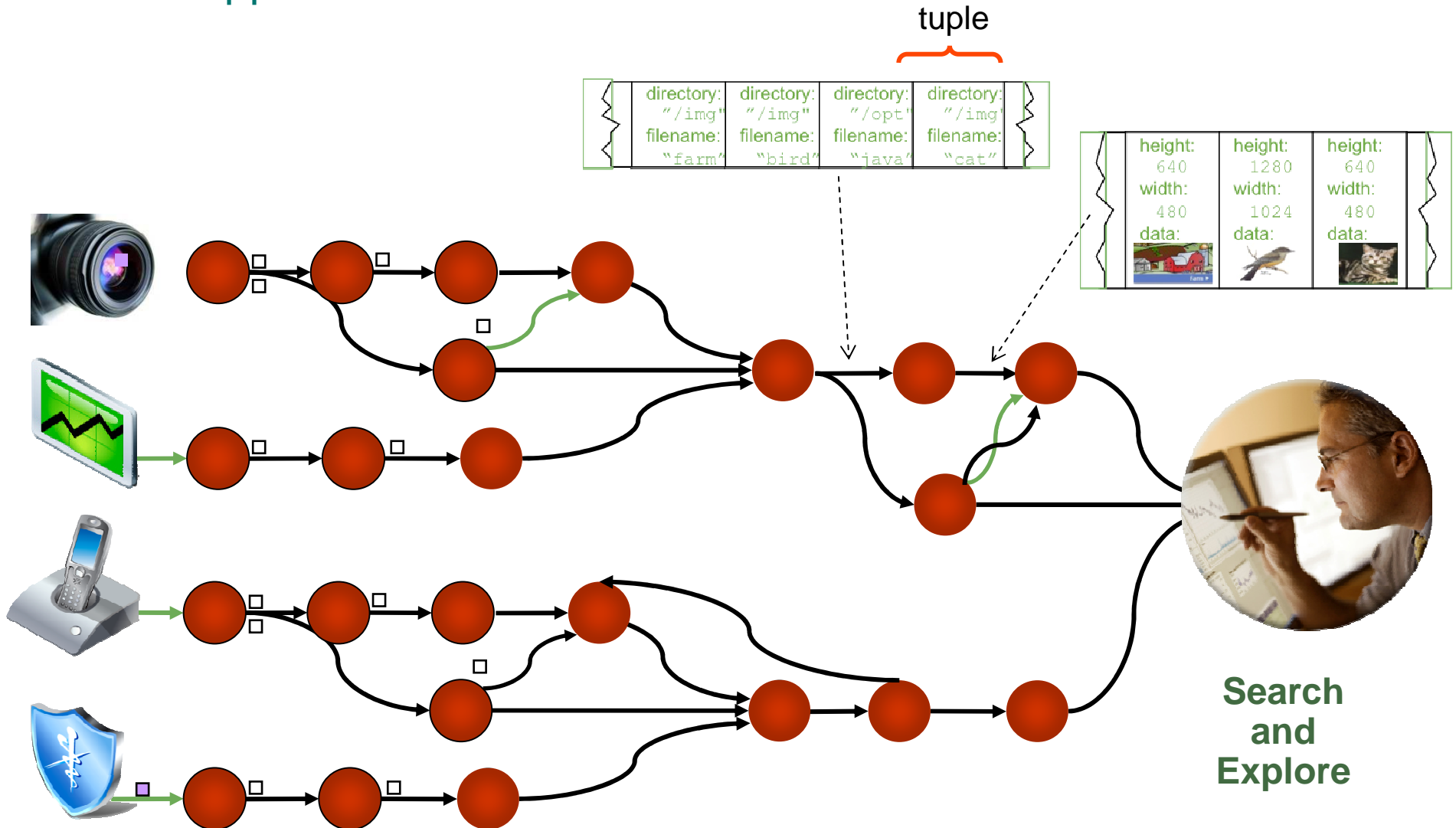
Use Case Automatic Recognition of Scenes of Interest

Automatic scene recognition, clustering, duplicate and near-duplicate detection from image and video content
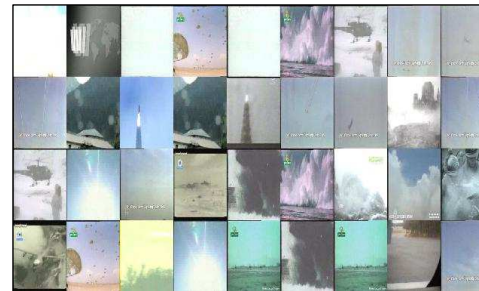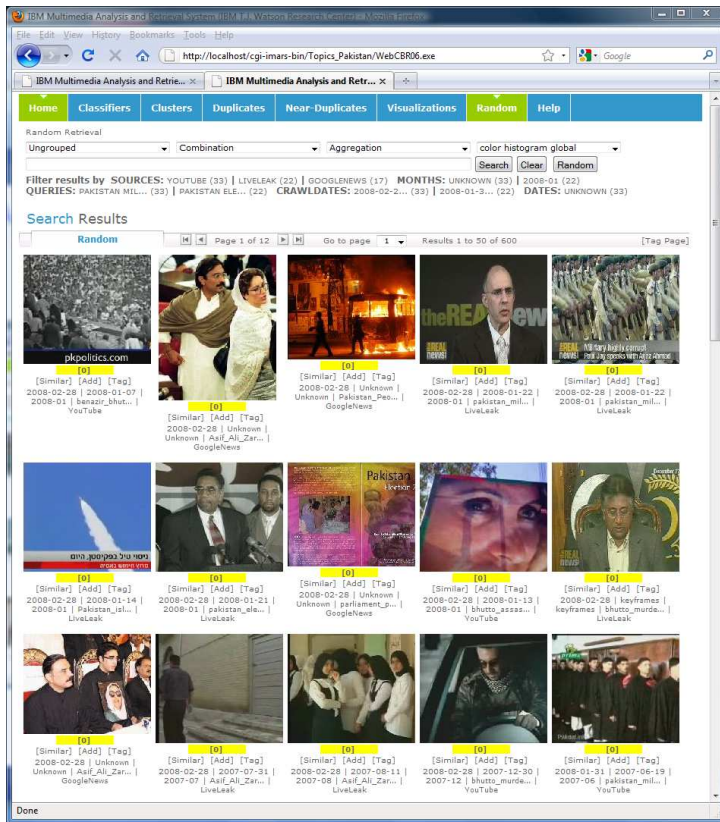
Air Attack

Explosion

Demonstration

Individual

Wreckage

Military

24

IBM Corporation

University of Ontario Institute of Technology (UOIT) Detects Neonatal Patient Symptoms Sooner

Capabilities Utilized

Stream Computing

- Performing real-time analytics using physiological data from neonatal babies

- Continuously correlates data from medical monitors to detect subtle changes and alert hospital staff sooner

- Early warning gives caregivers the ability to proactively deal with complications

Results

- Helps detect life threatening conditions up to 24 hours sooner

- Lower morbidity and improved patient care

"Helaps detect life threatening conditions up to 24 hours sooner"

UOIT
CHALLENGE INNOVATE CONNECT

Dublin City Centre Increases Bus Transportation Performance

Capabilities Utilized

Stream Computing

- Public transportation awareness solution improves on-time performance and provides real-time bus arrival info to riders

- Continuously analyzes bus location data to infer traffic conditions and predict arrivals

- Collects, processes, and visualizes location data of all bus vehicles

- Automatically generates transportation routes and stop locations

Results

- Monitoring 600 buses across 150 routes
- Analyzing 50 bus locations per second
- Anticipated to Increase bus ridership

Real-time visualization and visibility into the arrival times of their 1000 buses on 150 routes

Dublin City
Baile Átha Cliath

26

Identifies and classifies potential security threats miles away

TerraEchos Turns to IBM Big Data for Low Latency Surveillance Data Analysis

Capabilities Utilized

Stream Computing

- Deployed security surveillance system to detect, classify, locate, and track potential threats at highly sensitive national lab

- Stream computing collects and analyzes acoustic data from fiber-optic sensor arrays

- Analyzed acoustic data fed into TerraEchos intelligence platform for threat detection, classification, prediction & communication

Results

- Enables Terraechos solution to analyze and classify streaming acoustic data in real-time

- Provides lab & security staff with holistic view of potential threats & non-issues

- Enables a faster and more intelligent response to any threat

TerraEchos

IBM Netezza
The True Data Warehouse Appliance

**Simpler, Faster, More accessible
     Enterprise Class Analytics**

Purpose-built analytics engine

Integrated database, server and storage

Standard interfaces

Low total cost of ownership

Speed: 10-100x faster than traditional system

Simplicity: Minimal administration and tuning

Scalability: Peta-scale user data capacity

Smart: High-performance advanced analytics
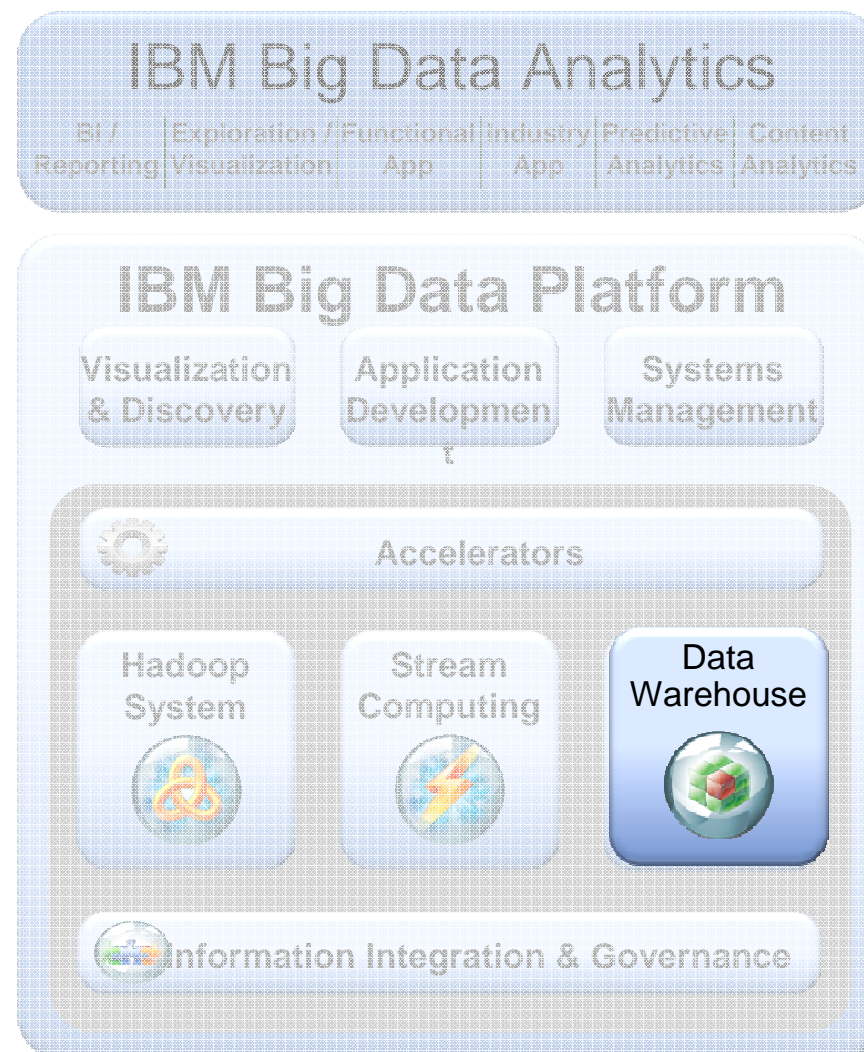
Barnes & Noble helps suppliers track sales and inventory in real time

Capabilities Utilized:

Data Warehouse Appliance

Need

- Publishers absorb the losses on returns when they print too many books, and have stock-outs when they print too few

- To provide publishers with the ability to get real-time insight into sales and view inventory trends over time

Benefits

- Decreased time to run queries from weeks to seconds and enabled an 80% reduction in time to run compared to its previous system

- Reduced inventory levels and inventory carrying costs

80% reduction in run time

BARNES&NOBLE
BOOKSELLERS

Without a Big Data Platform You Code…

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

Event Handling

Custom SQL and Scripts

Check Pointing

Application Management

HA

Connectors

Performance Optimization

Debug

Security

Accelerators and Toolkits

Input   **Process**   **Output**

Streams Processing Language
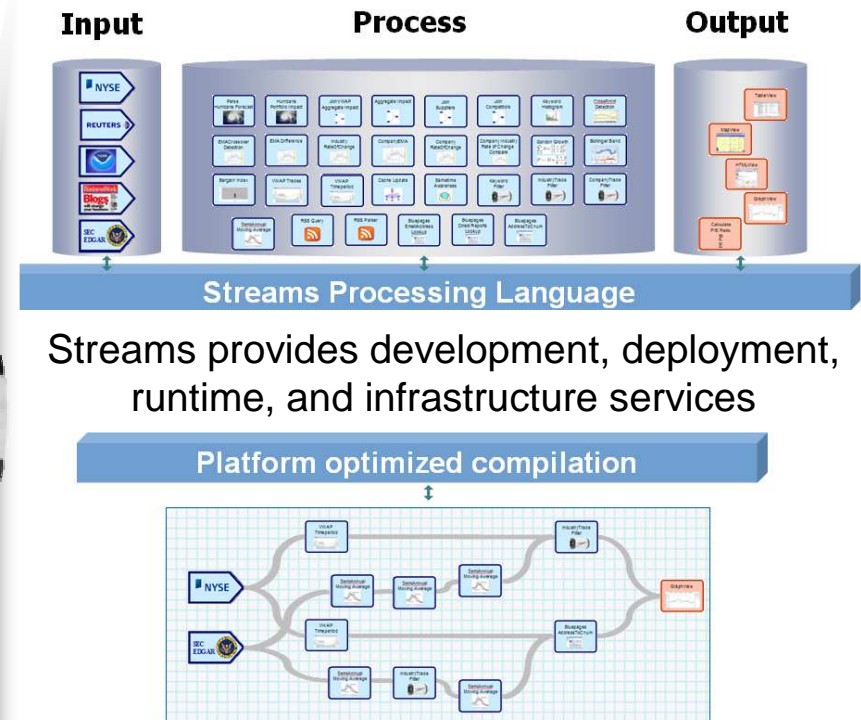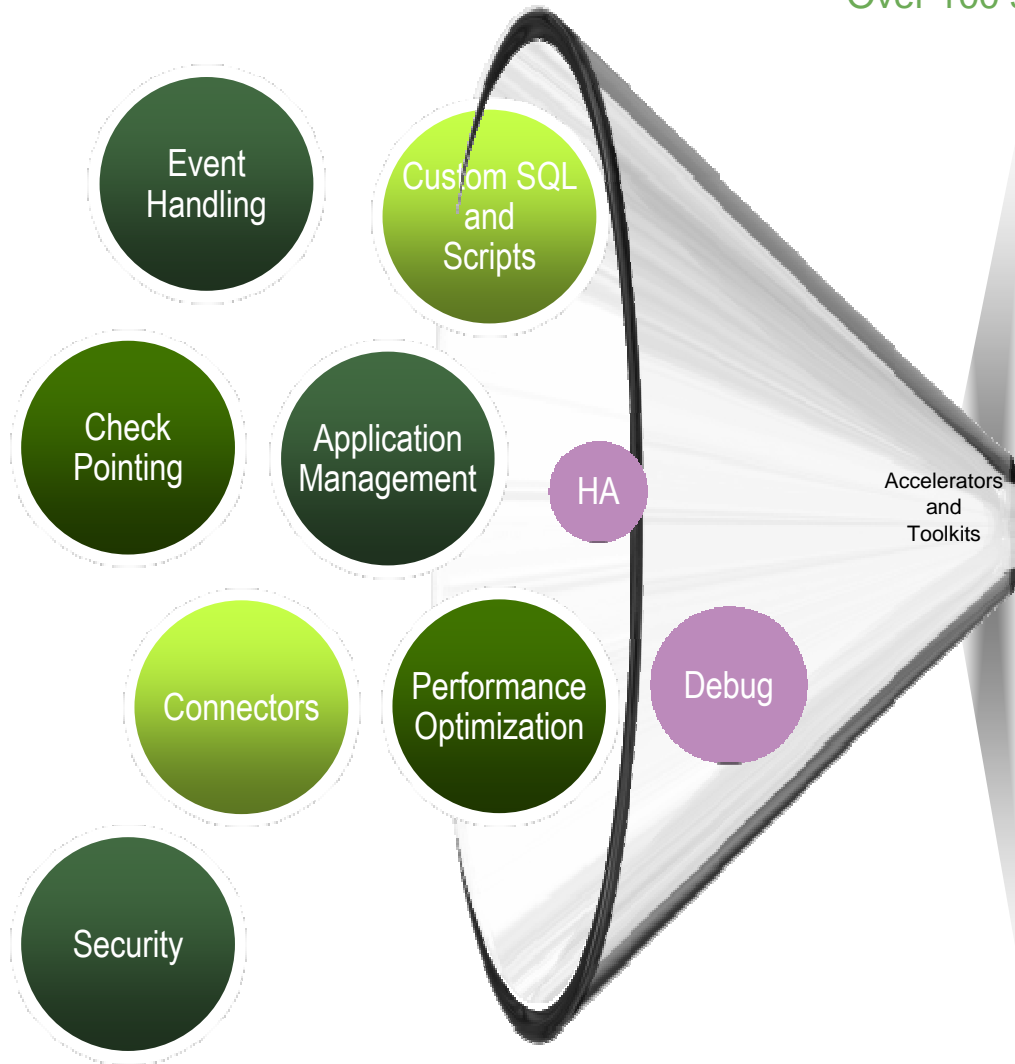
Streams provides development, deployment, runtime, and infrastructure services

Platform optimized compilation

"TerraEchos developers can deliver applications 45% faster due to the agility of Streams Processing Language…"
– Alex Philip, CEO and President, TerraEchos

IBM is Committed to Innovation

2012

IBM Resarch
Almaden
Austin
Melbourne
Sao Paulo
Beijing
Haifa
Delhi
Ireland
Yamato
Watson
Zurich

Selected SW Acquisitions

- **$16B+** in acquisitions since 2005
- **10,000+** technical professionals
- **~8000** dedicated consultants
- **27,000+** business partner certifications
- **8** Analytics Solutions Centers
- **100** analytics-based research assets; almost **300** researchers

i2
VARICENT
Vivísimo
Emptoris
DemandTec
Q1 Labs
Total Security Intelligence
Algorithmics A²
NETEZZA
CLARITY SYSTEMS
tealeaf
PSS
unica
OPEN PAGES
Datacap
Coremetrics
Sterling Commerce
Initiate
Guardium
SPSS
ILOG
princeton softech
solid.
COGNOS
FileNet
Ascential

"Watson is going to revolutionize many, many industries and it will fundamentally change the way we interact with computers & machines."

John Kelly, SVP & Head of IBM Research

2005

* TeaLeaf, Varicent Vivismo pending acquisition close

More Information …

bigdatauniversity.com/

ibm.com/software/data/infosphere/biginsights
ibm.com/software/data/infosphere/streams

ibm.com/software/data/bigdata/

youtube.com/user/ibmbigdata

Questions & Answers

Dipl.Ing.
Wolfgang Nimführ

Information Agenda
Executive Consultant
Big Data Tiger Team
IBM Software Group Europe

IBM Austria
Obere Donaustrasse 95
A1020 Vienna

Tel +43-664-618-5389
wolfgang.nimfuehr@at.ibm.com

Legal Disclaimer