Smarter Decision Making
Leverage Big Data to Gain New Actionable Insights

George Lapis

The World is Changing and Becoming More…
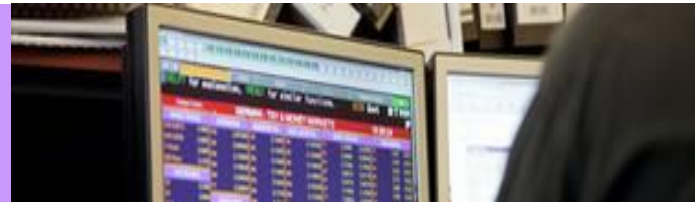
INSTRUMENTED

INTERCONNECTED

INTELLIGENT

The resulting explosion of information creates a need for a new kind of intelligence

…to help build a Smarter Planet

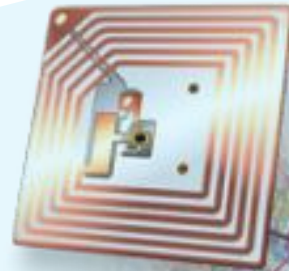# The Social Layer in an Instrumented Interconnected World

30 billion RFID tags today (1.3B in 2005)

4.6 billion camera phones world wide

12+ TBs of tweet data every day

? TBs of data every day

100s of millions of GPS enabled devices sold annually

25+ TBs of log data every day

2+ billion people on the Web by end 2011

76 million smart meters in 2009… 200M by 2014

35 ZB
(1 ZB = 1B TB)

2011

1.8 ZB

2010

2009

2020

1 ZB
1
ZB=1T

4Trillion
8GB
iPods
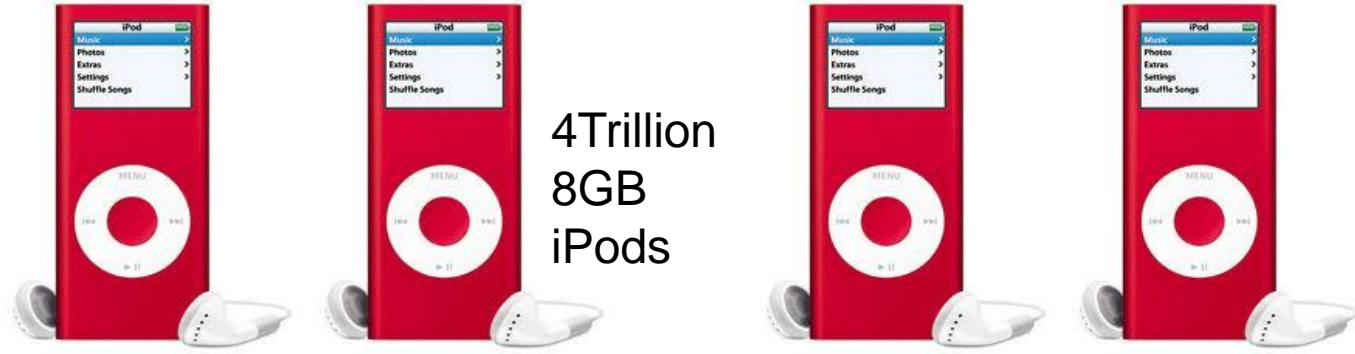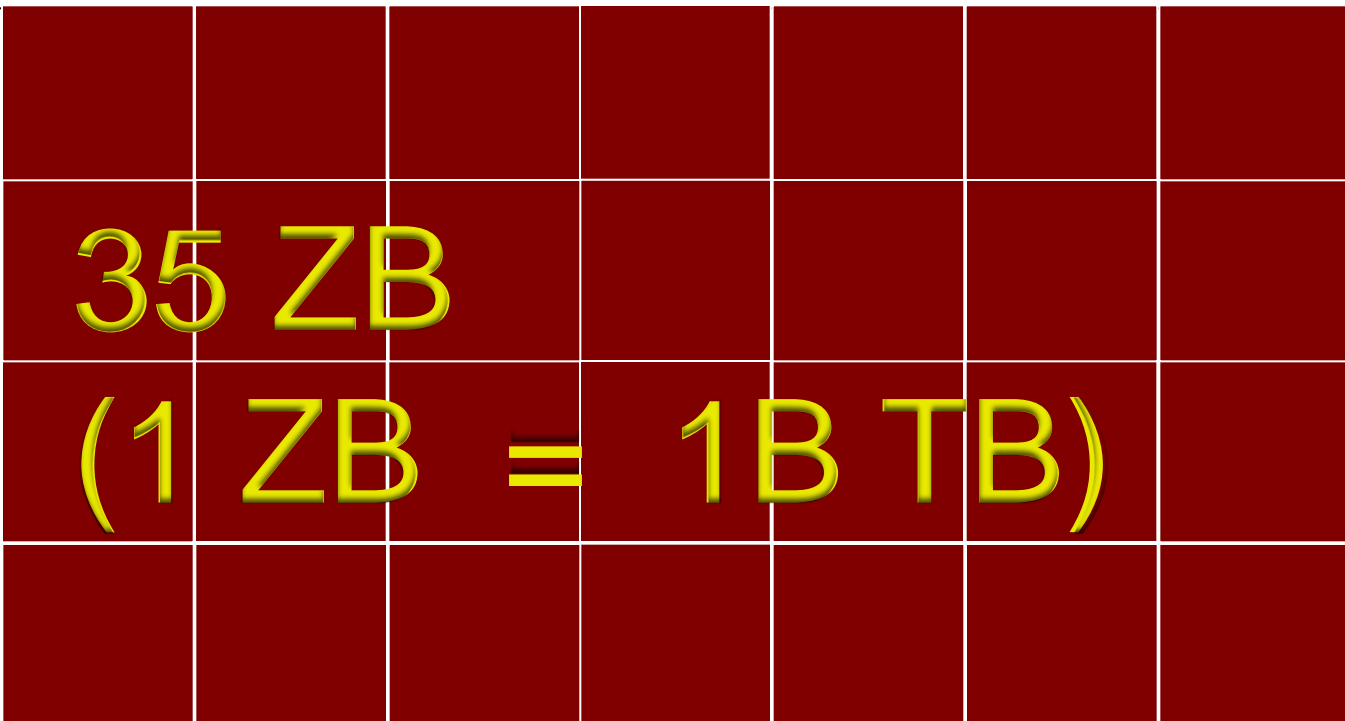
The Big Data Opportunity

Extracting insight from an immense volume, variety and velocity of data, in context, beyond what was previously possible.
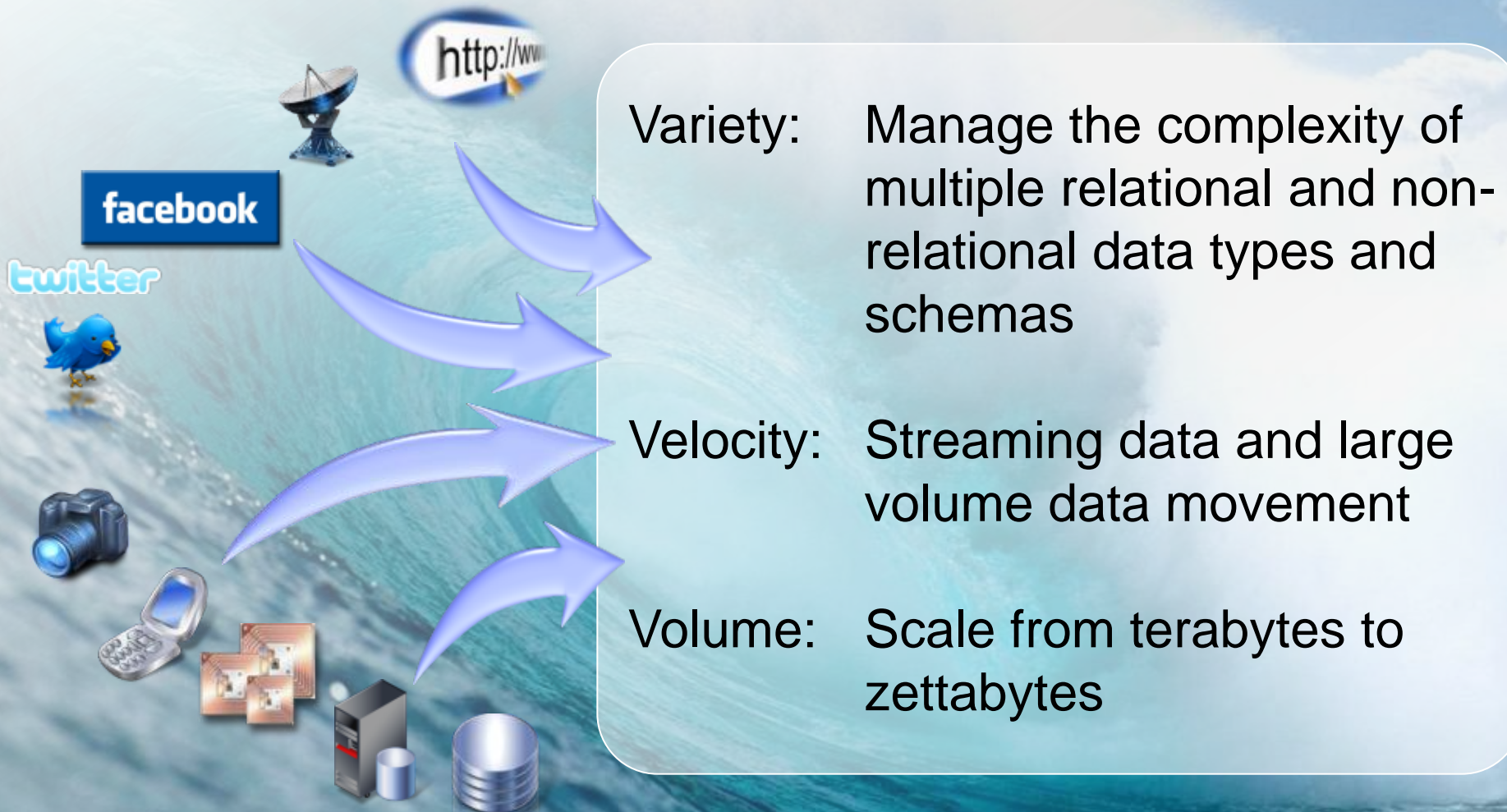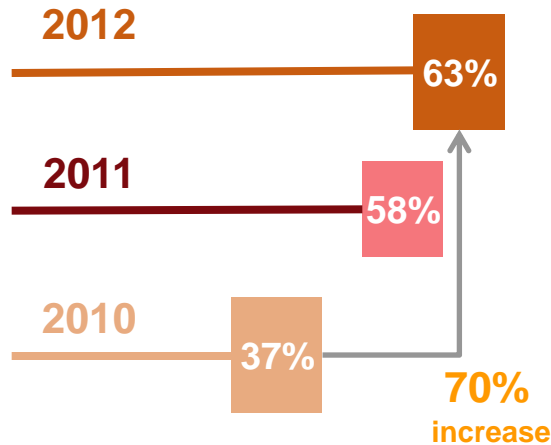
Variety: Manage the complexity of multiple relational and non-relational data types and schemas

Velocity: Streaming data and large volume data movement

Volume: Scale from terabytes to zettabytes

Benchmark of Global Big Data Activities (Oct 2012)

Realizing a competitive advantage
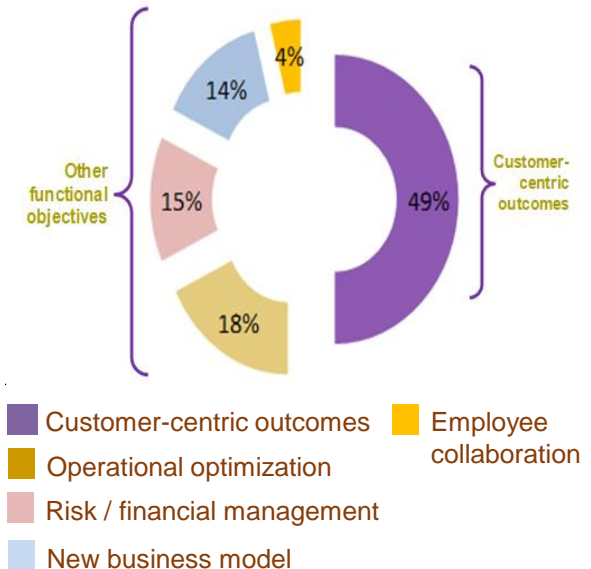
2012 — 63%

2011 — 58%

2010 — 37%

70% increase

Nearly two out of three respondents reports realizing a **competitive advantage** from information and analytics

Big data activities

28% Pilot and implementation of big data activities

24% Have not begun big data activities

47% Planning big data activities
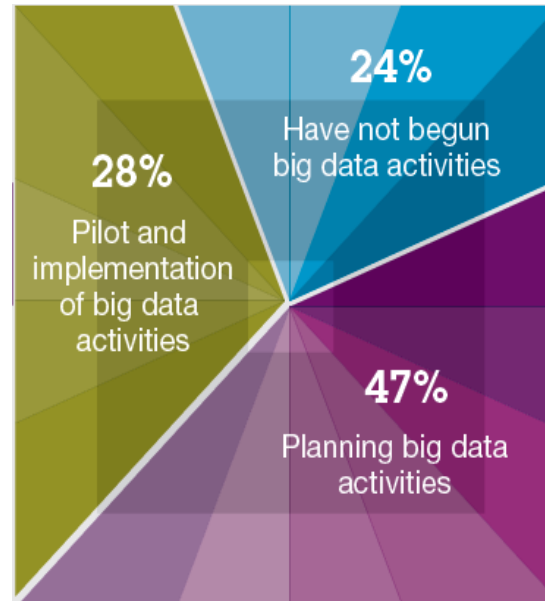
Three out of four organizations have **big data activities** underway; and one in four are either in **pilot or production**

Big data objectives

Other functional objectives

14%

4%

15%

18%

49% — Customer-centric outcomes

- ■ Customer-centric outcomes
- ■ Operational optimization
- ■ Risk / financial management
- ■ New business model
- ■ Employee collaboration

Improving *customer experience* by better understanding behaviors drives almost half of all active big data efforts followed by *Operational Optimization*

www.ibm.com/2012bigdatastudy

IBM Institute for Business Value and the University of Oxford Saïd Business School

6

The new era of analytics delivers value across the enterprise

Network Operations
...identify network bottlenecks in real-time for faster resolution

External Data
Google, facebook, YouTube, twitter, GPS

Customer Service Representatives
...offer personalized price promotions to different customer segments in real-time

Executive Leaders
...get real-time reports and analysis based on data inside as well as outside the enterprise (web, social media etc.)

Business Analysts
... analyze social media buzz for the new services/offerings to gauge initial success and any course correction needed

Big Data

Finance
...analyze all Call Detail Records (CDRs) to identify and reduce revenue leakage due to unbilled / underbilled CDRs

Business Development
... find and deliver new mechanisms to monetize network traffic and partner with upstream content providers

Marketing
... analyze subscriber usage pattern in real-time and combine that with the profile for delivering promotional or retention offers

7

Big Data enabled doctors from University of Ontario to apply neonatal infant monitoring to predict infection in ICU 24 hours in advance

**IBM Data Baby
youtube.com**

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

- 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.

Vestas.

Cisco turns to IBM big data for intelligent infrastructure management

- Optimize building energy consumption with centralized monitoring
- Automate preventive and corrective maintenance

Capabilities Utilized:
- Streaming Analytics
- Hadoop System
- Business Intelligence

Applications:
- Log Analytics
- Energy Bill Forecasting
- Energy consumption optimization
- Detection of anomalous usage
- Presence-aware energy mgt.
- Policy enforcement

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

Dublin City
Baile Átha Cliath

Asian telco reduces billing costs and improves customer satisfaction.
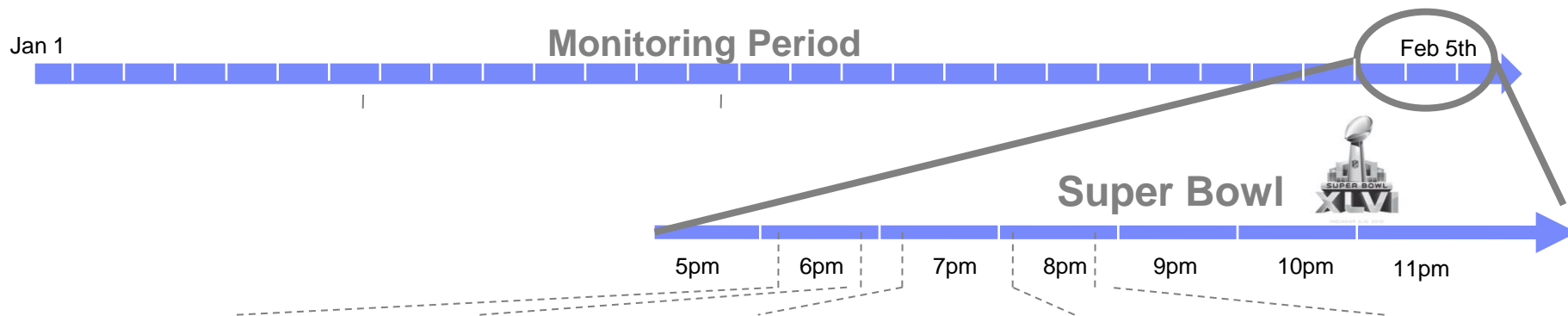
Capabilities:

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 1/8th

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

To-the-minute and historical product insight

Jan 1 — **Monitoring Period** — Feb 5th

Super Bowl

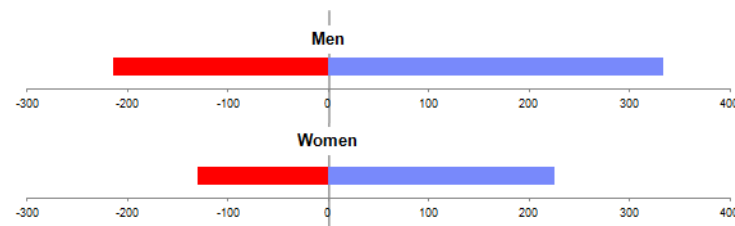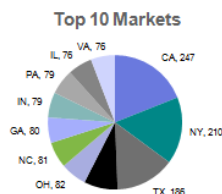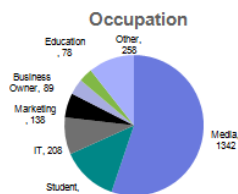5pm 6pm 7pm 8pm 9pm 10pm 11pm
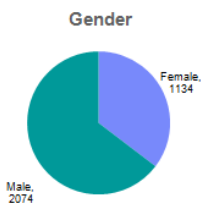
Data Set

- 1.1B tweets
- 5.7M blog and forum posts
- 3.5M relevant messages
- 97K referencing Product_A
- 18K referencing Product B

Information extracted

- Buzz and sentiment
- Gender, Location and Occupation
- Fans
- Intent to in purchase
- Specific attributes of products

Gender
- Female, 1134
- Male, 2074

Occupation
- Education, 78
- Other, 258
- Business Owner, 89
- Marketing, 138
- IT, 208
- Student, 318
- Media, 1342

Top 10 Markets
- VA, 76
- IL, 76
- PA, 79
- IN, 79
- GA, 80
- NC, 81
- OH, 82
- FL, 105
- TX, 186
- NY, 210
- CA, 247

Men

Women

Enhancing Fraud Detection for Banks and Credit Card Companies

Scenario

- Build up-to-date models from transactional to feed real-time risk-scoring systems for fraud detection

Requirement

- Analyze volumes of data with response times that are not possible today
- Apply analytic models to individual client, not just client segment.

Build Faster Real-time Trading Systems

Scenario
- Identify and execute trades
- Process over 5M events per second with average latency of 150 microseconds

Requirement
- Consuming, analyzing and acting on market data while maintaining sub-millisecond response time under extreme data loads
- Incorporate content feeds, news text, audio, video, to establish greater context for better decisions

Transaction Analysis for Banking Industry

Scenario
- Analyze transaction issues from federated systems and applications to provide up-to-date account status with less turnaround time
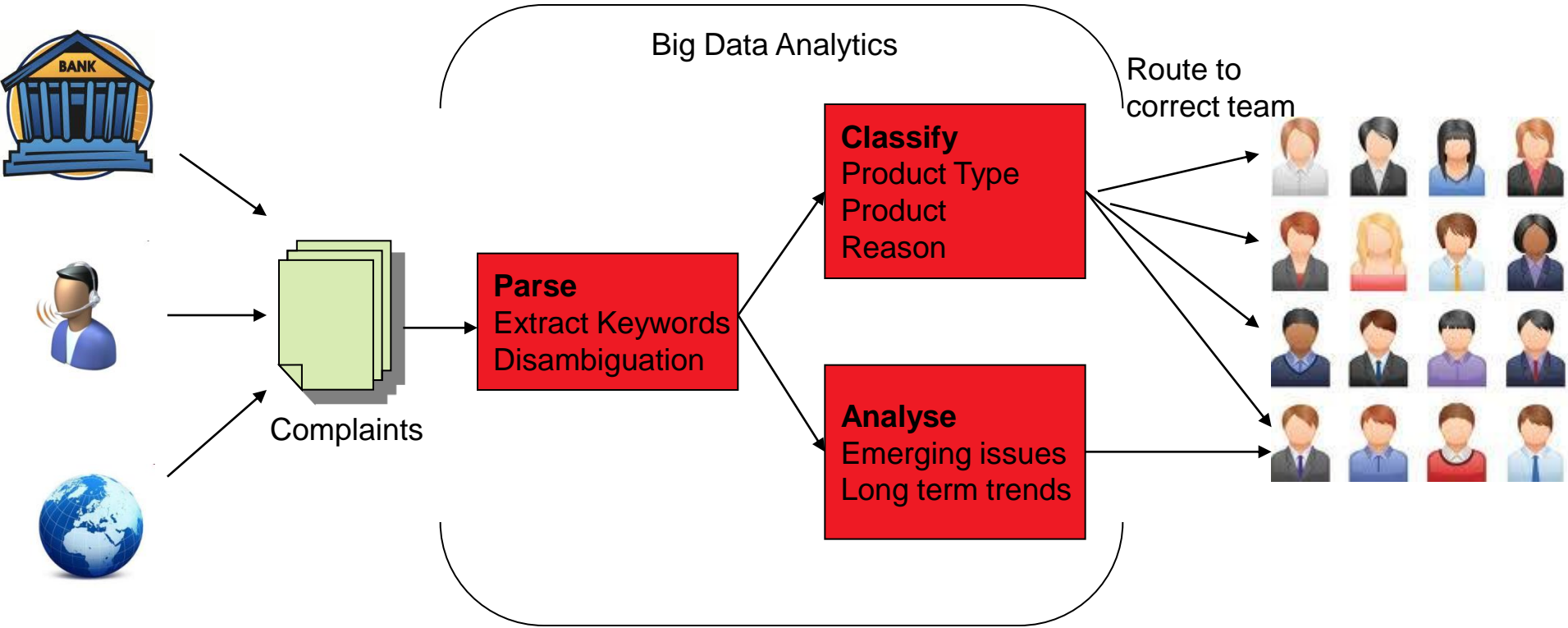
Requirement
- Collect, aggregate, and analyze log data from various application systems
- Handle logs in different formats and correlating errors across applications
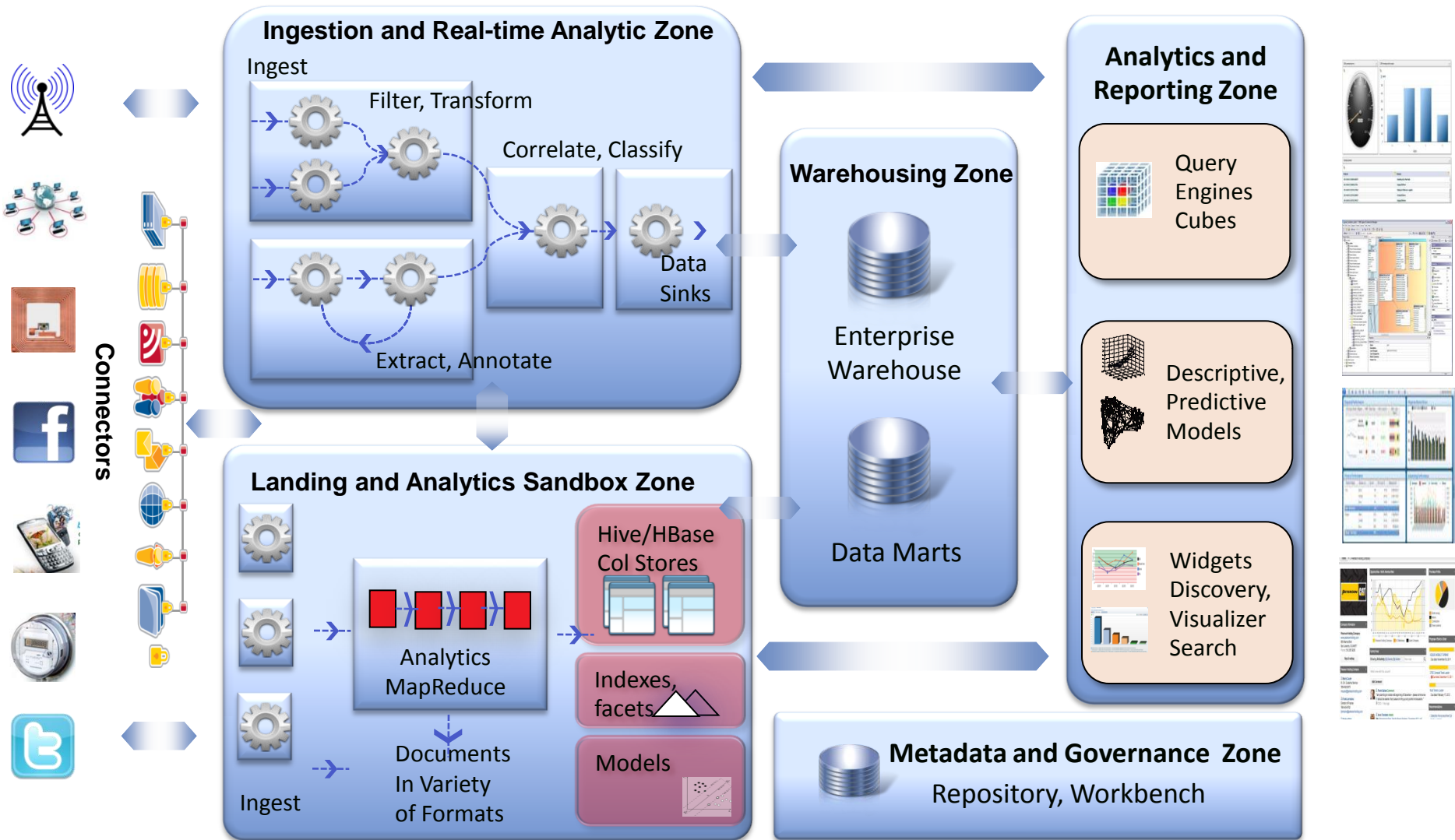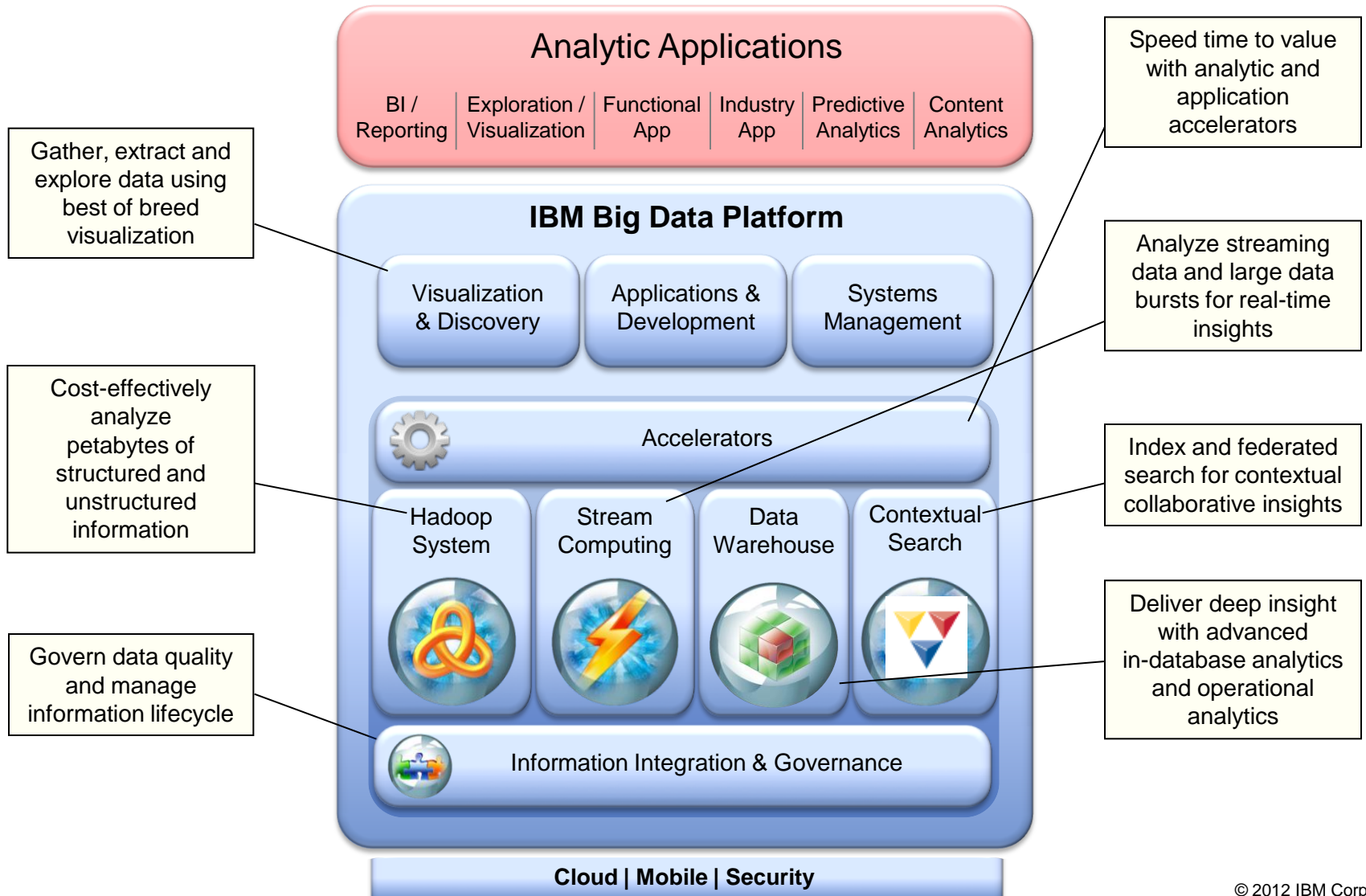- Reduce response time to less than 2 minutes

CRM Automatic Complaints Classification

Big Data Analytics

Route to correct team

Classify
Product Type
Product
Reason

Parse
Extract Keywords
Disambiguation

Complaints

Analyse
Emerging issues
Long term trends

Emerging Pattern of Big Data Implementation

Connectors

Ingestion and Real-time Analytic Zone

Ingest

Filter, Transform

Correlate, Classify

Data Sinks

Extract, Annotate

Landing and Analytics Sandbox Zone

Analytics MapReduce

Hive/HBase Col Stores

Indexes, facets

Models

Documents In Variety of Formats

Ingest

Warehousing Zone

Enterprise Warehouse

Data Marts

Analytics and Reporting Zone

Query Engines Cubes

Descriptive, Predictive Models

Widgets Discovery, Visualizer Search

Metadata and Governance Zone

Repository, Workbench

Big Data Platform and Application Framework

Analytic Applications

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

Speed time to value with analytic and application accelerators

Gather, extract and explore data using best of breed visualization

IBM Big Data Platform

Visualization & Discovery	Applications & Development	Systems Management

Analyze streaming data and large data bursts for real-time insights

Accelerators

Cost-effectively analyze petabytes of structured and unstructured information

Hadoop System	Stream Computing	Data Warehouse	Contextual Search

Index and federated search for contextual collaborative insights

Deliver deep insight with advanced in-database analytics and operational analytics

Govern data quality and manage information lifecycle

Information Integration & Governance

Cloud | Mobile | Security

Big Data Platform – Hadoop Distribution-Agnostic

Enterprise Capabilities

- Visualization & Exploration
- Development Tools
- Advanced Engines
- Connectors
- Workload Optimization
- Administration & Security

Open source components

IBM-certified Apache Hadoop *or* Other Hadoop Distributions *or* …

Platform Capabilities

- Built-in analytics
  - Text analytics engine, annotators, Eclipse tooling
  - Interface to project R (statistical platform)
- Deep integration with enterprise software stack
- Analytical tool for analysts
- Ready-made business process accelerators
- Integrated installation of supported open source and other components
- Web Console for admin and application access
- Platform enrichment: additional security, performance features, . . .
- World-class support
- Full open source compatibility

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

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
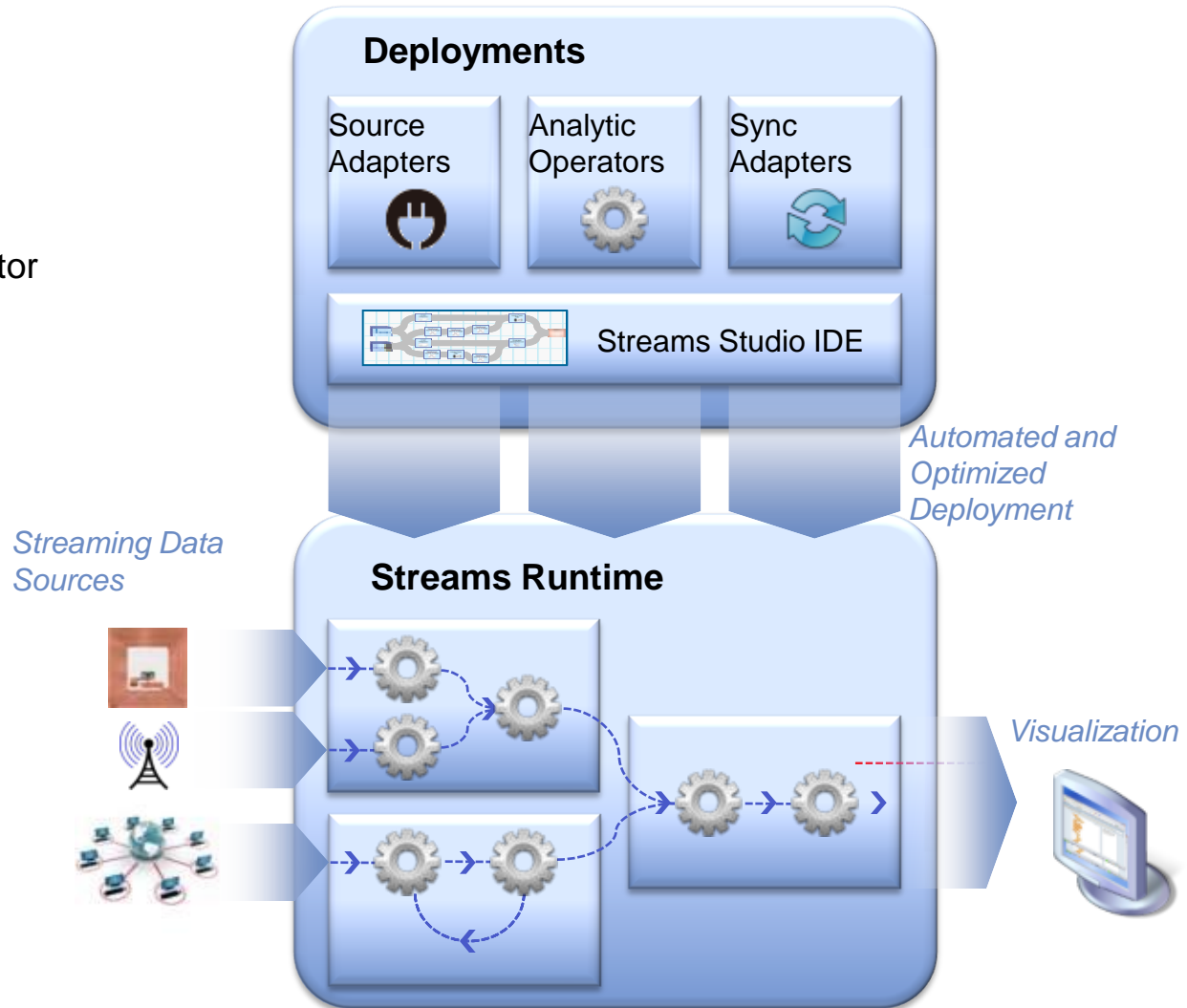
Big Data Accelerators Improve Time to Value

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

Media and Entertainment
Sentiment Analytics, Intent to purchase, micro segmentation,

Public transportation
Real-time monitoring and routing optimization
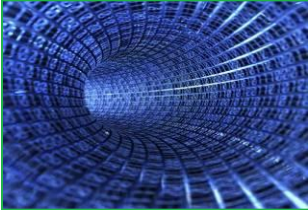
Oil and Gas
Streaming statistical analysis, new energy source discovery

Over 100 sample applications

User Defined Toolkits

Standard Toolkits
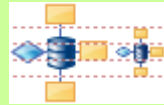
Industry Data Models
Banking, Insurance, Telco, Healthcare, Retail…

Stream Computing

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:
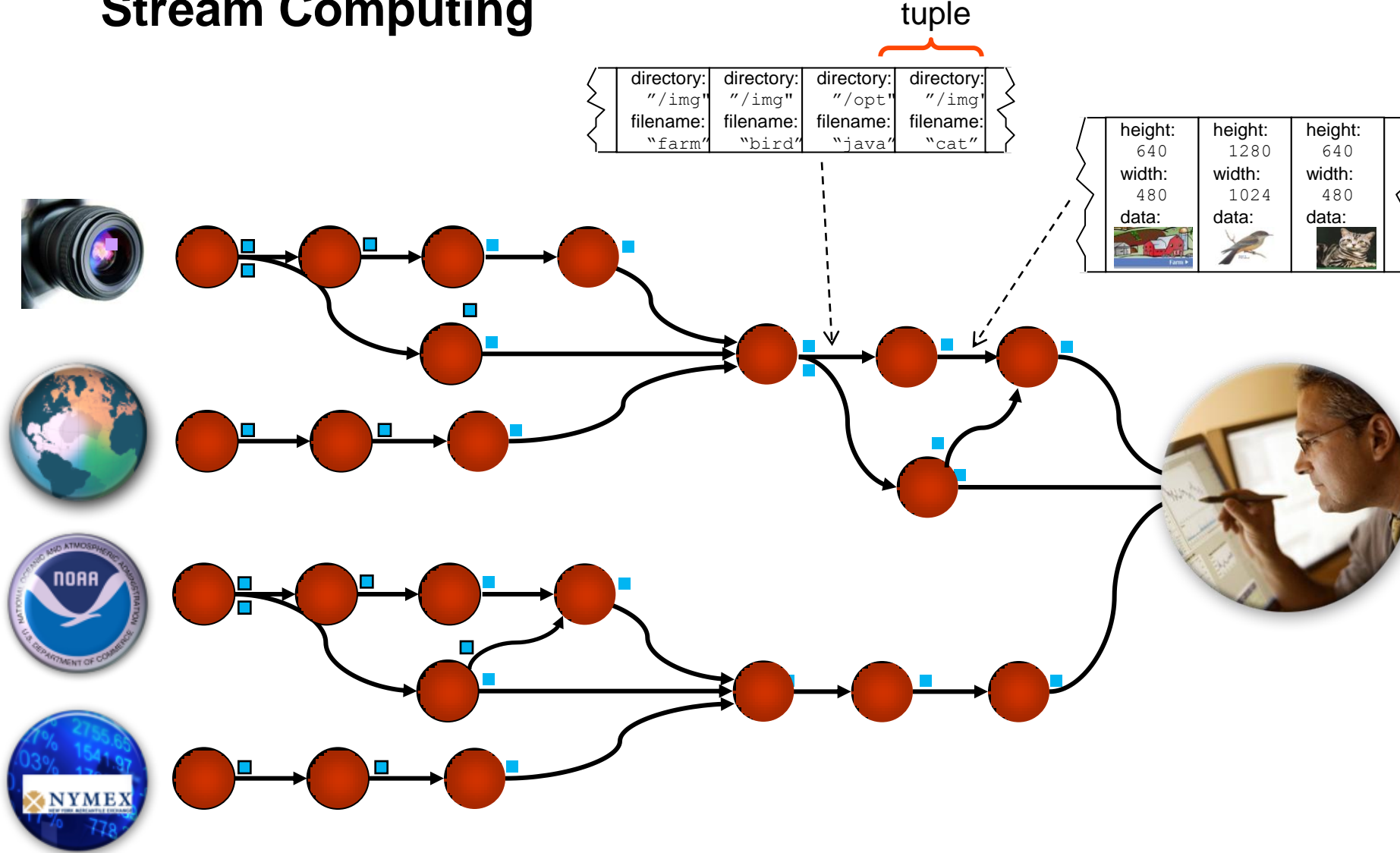
Real-time Stream Analytics

→ *continuous ingestion*
    → *continuous analysis*

infrastructure provides services for

scheduling analytics across h/w nodes

establishing streaming connectivity

Filter

Transform   **Annotate**

Correlate

Classify
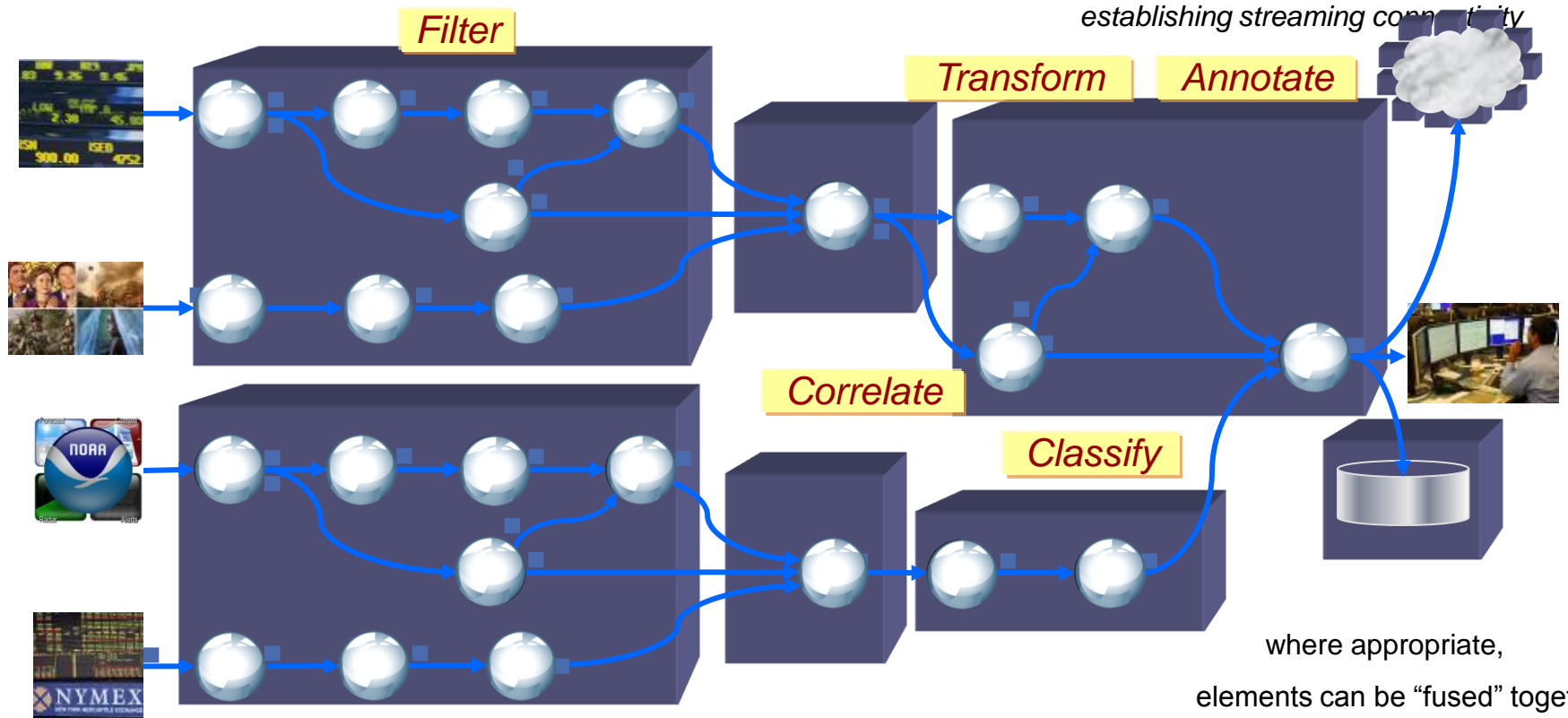
achieve scale

by partitioning applications into components

by distributing across stream-connected hardware nodes

where appropriate,

elements can be "fused" together

for lower communication latencies

IBM InfoSphere BigInsights

Visualization & Discovery
- BigSheets
- Dashboard & Visualization

Applications & Development
- Apps
- Text Analytics
- MapReduce
- Workflow
- Pig & Jaql
- BigSQL & Hive

Systems Management
- Admin Console
- Monitoring

Integration
- JDBC
- Netezza
- DB2
- Streams
- DataStage
- Flume
- Sqoop

Advanced Analytic Engines
- Machine Learning (ML)
- Text Processing Engine & Extractor Library (AQL+HIL)
- BigSQL
- R
- Mahout

Workload Optimization
- Integrated Installer
- Enhanced Security
- Splittable Text Compression
- Adaptive MapReduce
- ZooKeeper
- Oozie
- Jaql
- Flexible Scheduler
- Lucene
- Pig
- Hive
- Index

Runtime
- MapReduce

Data Store
- HBase
- Column Store

File System
- HDFS
- GPFS (EAP)

Management
- Security
- Audit & History
- Lineage
- HCatalog

New classes of applications for end-users

Real Time Analytics

Internet Scale Analytics

In-Database Analytics

Enterprise Data Connectors

Federated Discovery

Navigation and Visualization

Application Framework

IBM's Big Data Business Partner Ecosystem

230+ Big Data Business Partners

© 2012 IBM Corporation

Forrester Wave™: Enterprise Hadoop Solutions, Q1 '12

FORRESTER®

"*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