Stream Computing & BIG Data

Evolution or Revolution?

Mark McConnell
Enterprise Data Management
IBM Asia Pacific

IBM Software
ExecutiveSummit2011
A Premier Leadership Exchange

We are living in a world of explosive information growth

Volume

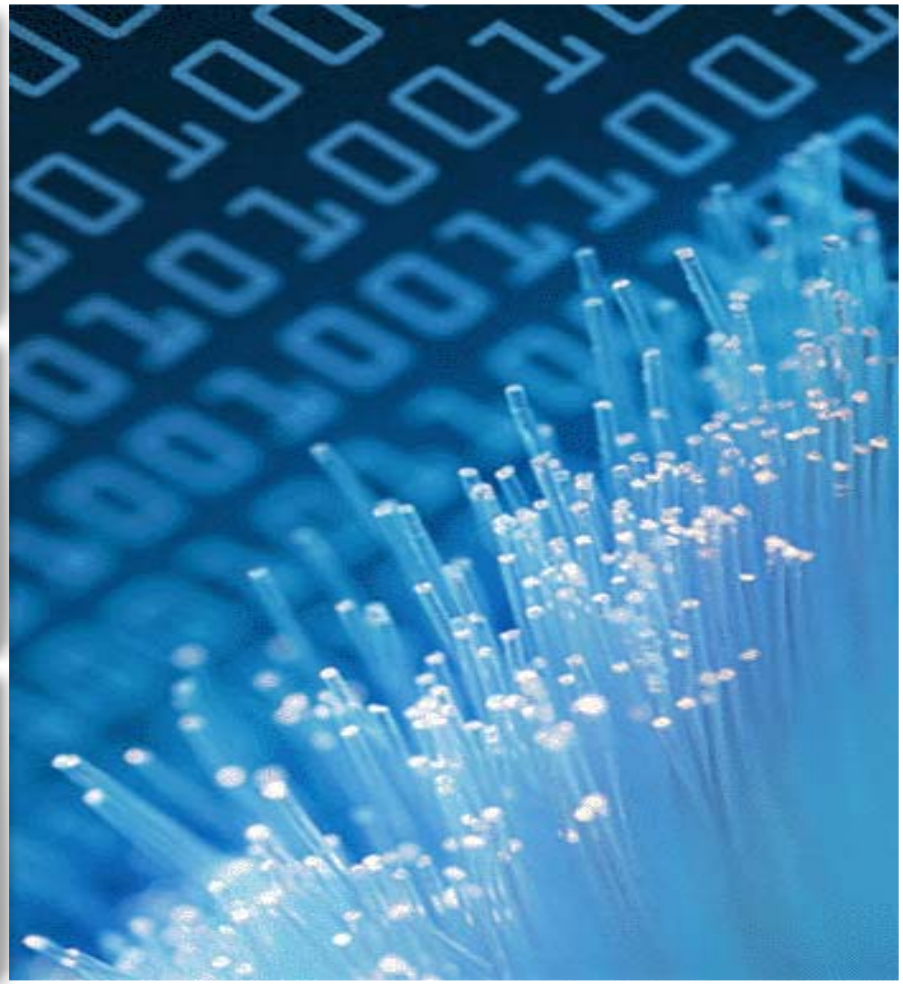
- Every day, **15 petabytes** of new information are being generated.

Variety

- **80%** of new data growth is unstructured content, generated largely by email, images and video

Velocity

- An average company with 1,000 employees spends **$5.3 million** a year to find its own information.



Data and Real World Events are growing …

2 Billion **Internet users by 2011**

1.3 Billion RFID tags in 2005
30 Billion **RFID tags by 2010**

4.6 Billion
Mobile Phones World Wide

Capital market data volumes grew ***1,750%*****, 2003-06**

Twitter process ***7 terabytes*** **of data every day**

World Data Centre for Climate
- ***220 Terabytes*** **of Web data**
- ***9 Petabytes*** **of additional data**

Facebook process ***10 terabytes*** **of data every day**

IBM Software
ExecutiveSummit2011

Streams processing in the context of BIG data

- Its not just about “Bigness” or Volume
- There are two other dimension to the data deluge…
  - Variety
  - Velocity
- Any comprehensive approach to big data needs a strategy for V^3.
- Does size matter?
  - Bigger than you’ve got today
  - Transcends your ability to manage it with traditional Database tools and techniques especially from a economic viability perspective.
- Teasing Value from data considered to amorphous to yield value
  - Considered too expensive to clean, normalise and manage

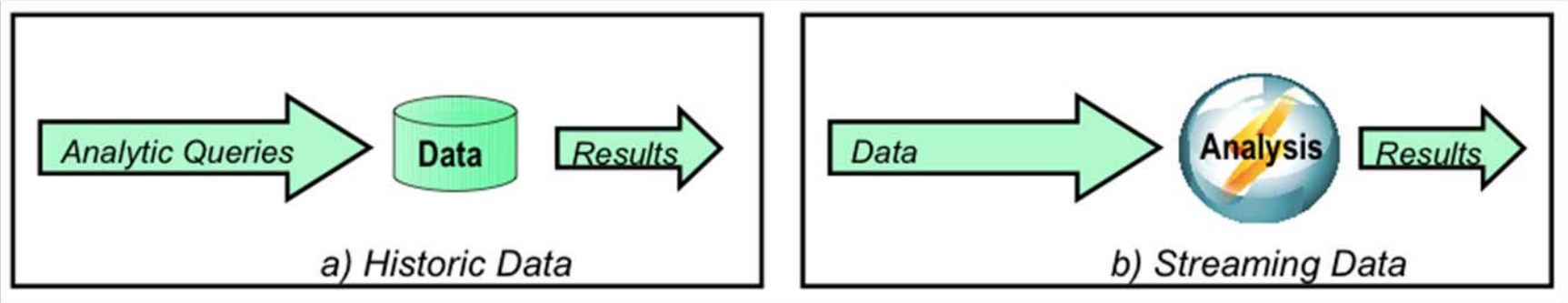
Two different technology approaches…

Realtime oriented…

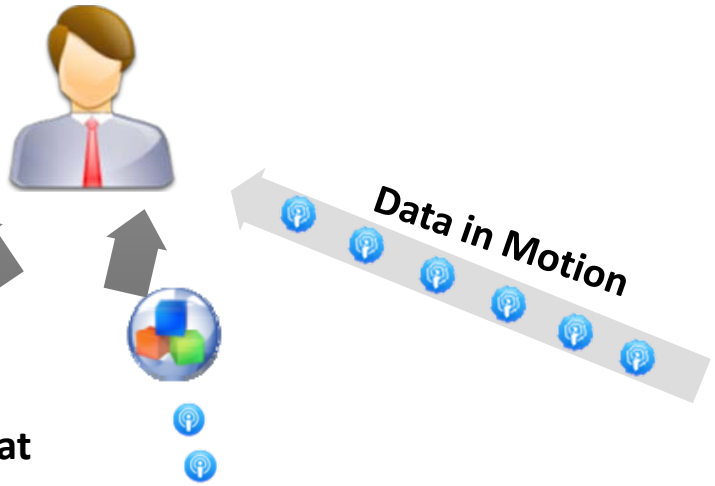
Batch oriented…

So what is Stream Computing

- A little history…
- You need to THINK a little differently about data and questions.



InfoSphere Streams – a paradigm shift



Real Time Analytic Processing (RTAP) to improve business response

Analysis of current data to improve business transactions

Reporting and human analysis on historical data

Stream Computing

Data Warehousing

Operational Databases

2003 "System S"

2009 InfoSphere Streams

1968
Hierarchical database

1970
Relational database
"System R"

1983
DB2 v1

OLTP

OLAP

RTAP

Law Enforcement and Security – US Federal Government Development Use Case

Streams of information including video surveillance, wire taps, communications, call records, etc.

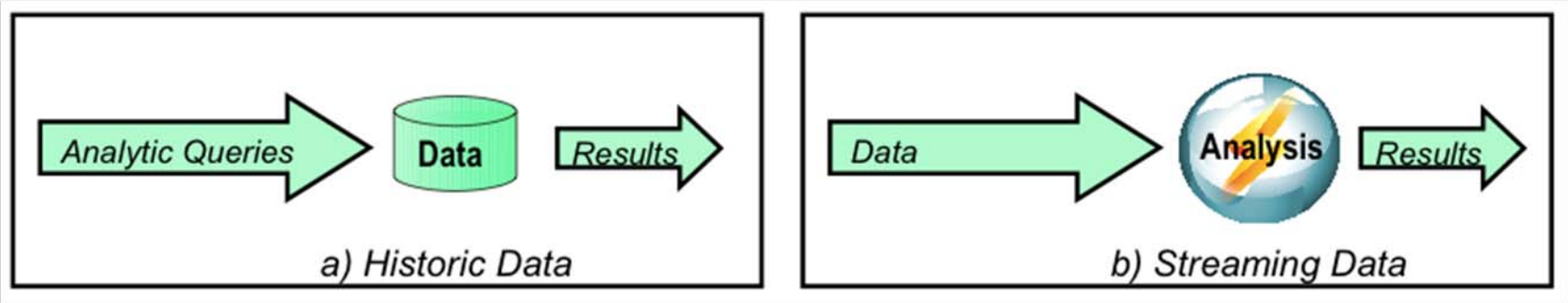
Millions of streams per second with low density of critical data

Identify patterns and relationships among vast information sources

"The US Government has been working with IBM Research since 2003 on a radical new approach to data analysis that enables high speed, scalable and complex analytics of heterogeneous data streams in motion. The project has been so successful that US Government will deploy additional installations to enable other agencies to achieve greater success in various future projects" - US Government

So what is Stream Computing

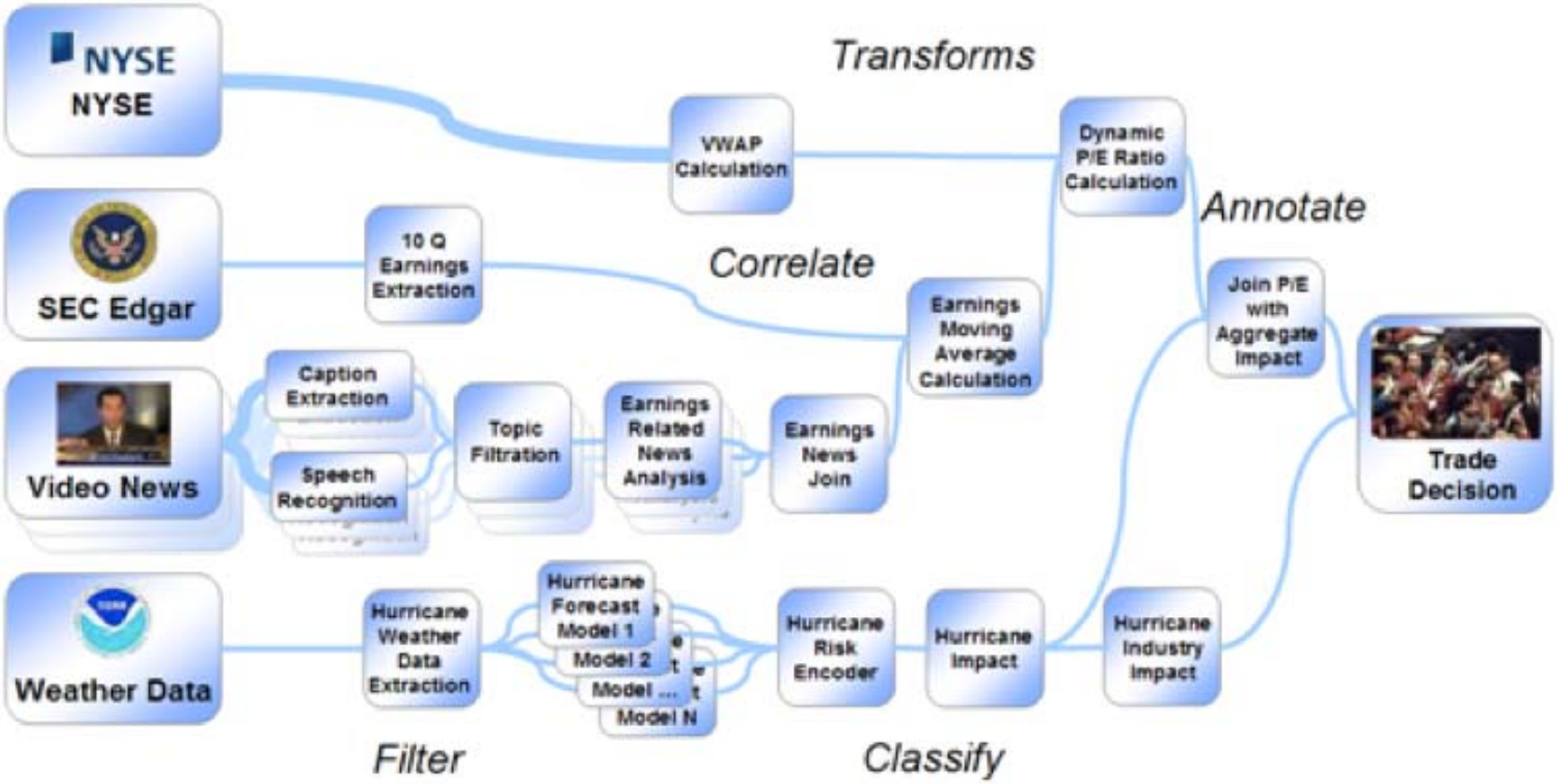
- A little history…
- You need to THINK a little differently about data and questions.



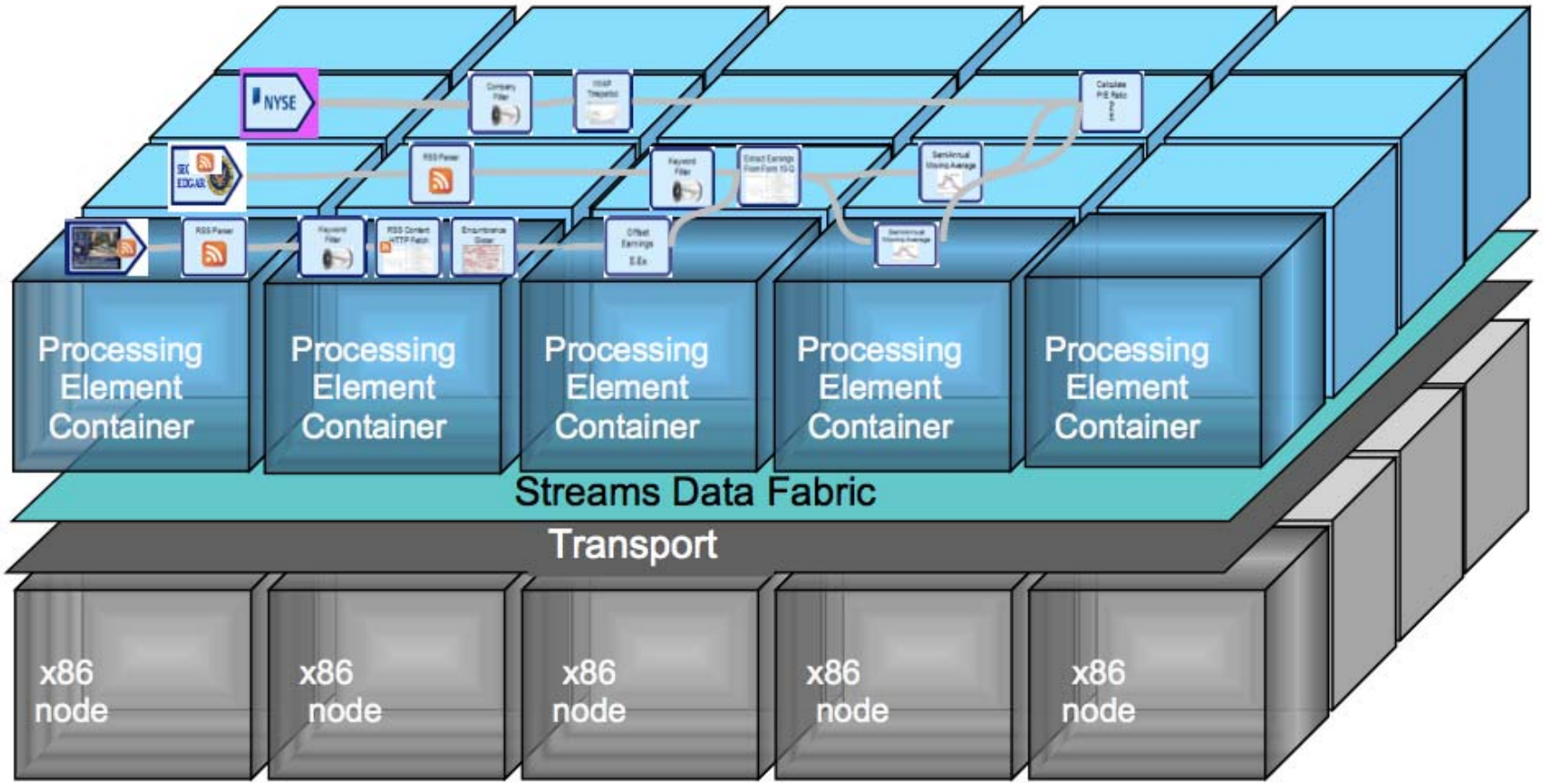
Objectives for Enterprise class Streams Processing

- Objectives for Stream Processing
  - Respond in real time to events and changing requirements
  - Continuously analyze data at rates that are orders of magnitude greater than existing systems
  - Adapt rapidly to changing data forms and types
  - Manage high availability, heterogeneity, and distribution for the new stream paradigm
  - Provide security and information confidentiality for shared information

The big idea in Streams



The system takes care of figuring out where to place the processing



Streaming Analytics in Action

Natural Systems
- Wildfire management
- Water management

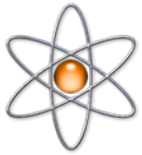
Stock Market
- Impact of weather on securities prices
- Analyze market data at ultra-low latencies

Law Enforcement, Defense & Cyber Security
- Real-time multimodal surveillance
- Situational awareness
- Cyber security detection

Transportation
- Intelligent traffic management

Fraud Prevention
- Detecting multi-party fraud
- Real time fraud prevention

Manufacturing
- Process control for microchip fabrication



e-Science
- Space weather prediction
- Detection of transient events
- Synchrotron atomic research

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

Telephony
- CDR processing
- Social analysis
- Churn prediction
- Geomapping

Other
- Smart Grid
- Text analysis
- Who's talking to whom?
- ERP for commodities
- FPGA acceleration

Use Case – Data Baby

- Now extended beyond the initial trial at the University Hospital in Ontario, Canada
- Using remote telemetry, hospitals are online from the US, Australia and China

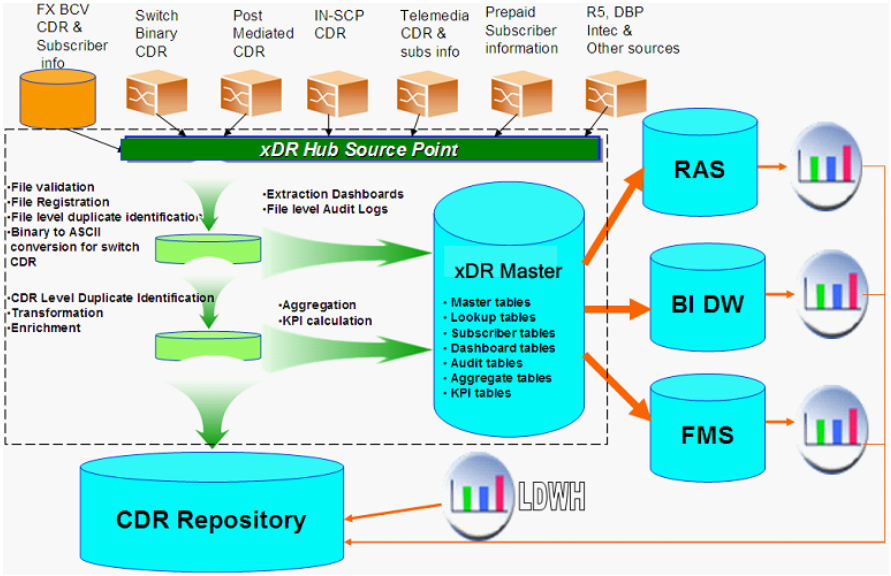
Use Case: Telco

Challenge

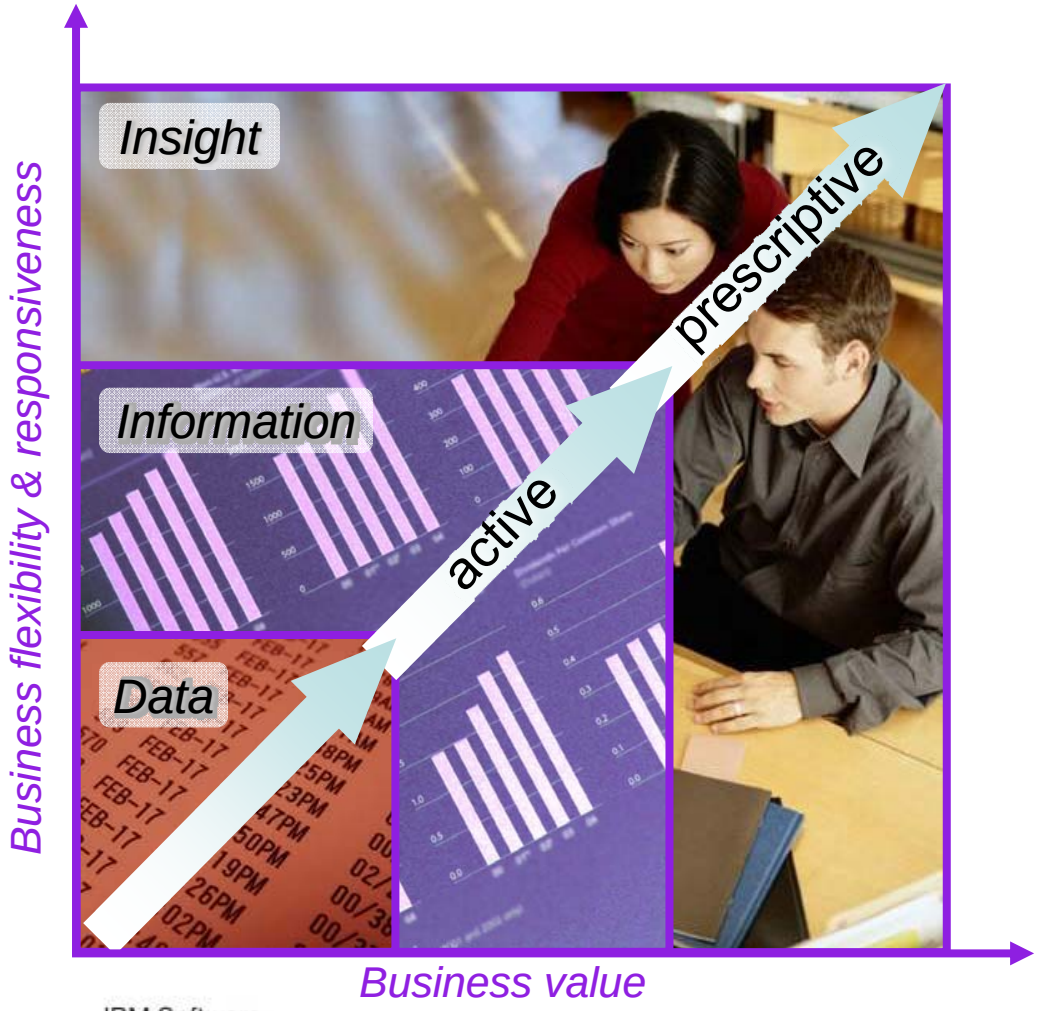
- Mediation process to support real time billing required handling billions of CDRs per day and needed de-duplication against 15 days worth of CDR data
- Simultaneous requirement was to support IT and Business with real time analytics
- CDR processing within Warehouse was sub-optimal from a loss of performance and real time requirements standpoint

Solution

- InfoSphere Streams supported real time mediation by handling 6bil CDRs each day
- Streams also provided them with a platform to run real time analytics
- Offloading CDRs processing to Streams platform increased the performance of their warehouse for other analytics
- Single platform for mediation and real time analytics reduced IT complexity



Real Time Marketing at Southeast Asian Telco



The Pain:

- 100M CDRs per day from SMS from 25M subscribers
- Used billing information to understand behavior and deliver promotions

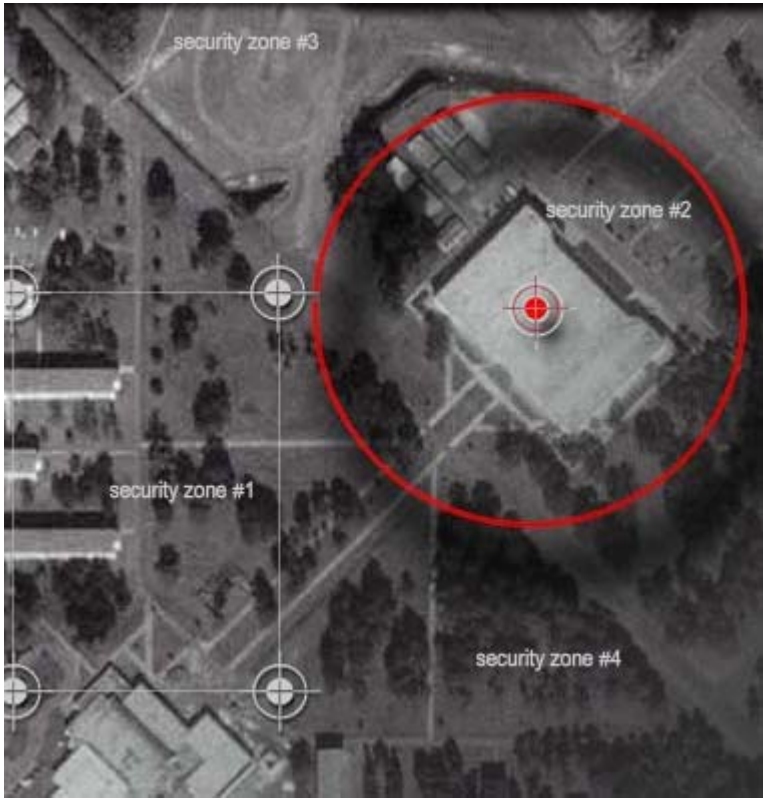
The Answer:

- InfoSphere Streams to create realtime marketing promotions

"A moment's insight is sometimes worth a life's experience."
Oliver Wendell Holmes

Surveillance and Physical Security: TerraEchos

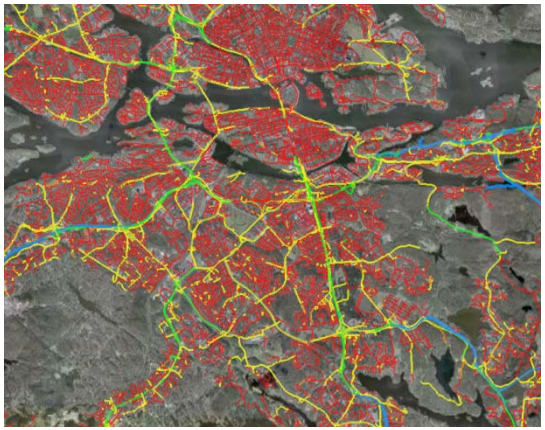
- Use scenario
  - **State-of-the-art covert surveillance system based on Streams platform**
  - Acoustic signals from buried fiber optic cables are monitored, analyzed and reported in real time for necessary action
  - Currently designed to scale up to 1600 streams of raw binary data
- Requirement
  - Real-time processing of multi-modal signals (acoustics. video, etc)
  - Easy to expand, dynamic
  - 3.5M data elements per second
- Winner 2010 IBM CTO Innovation Award

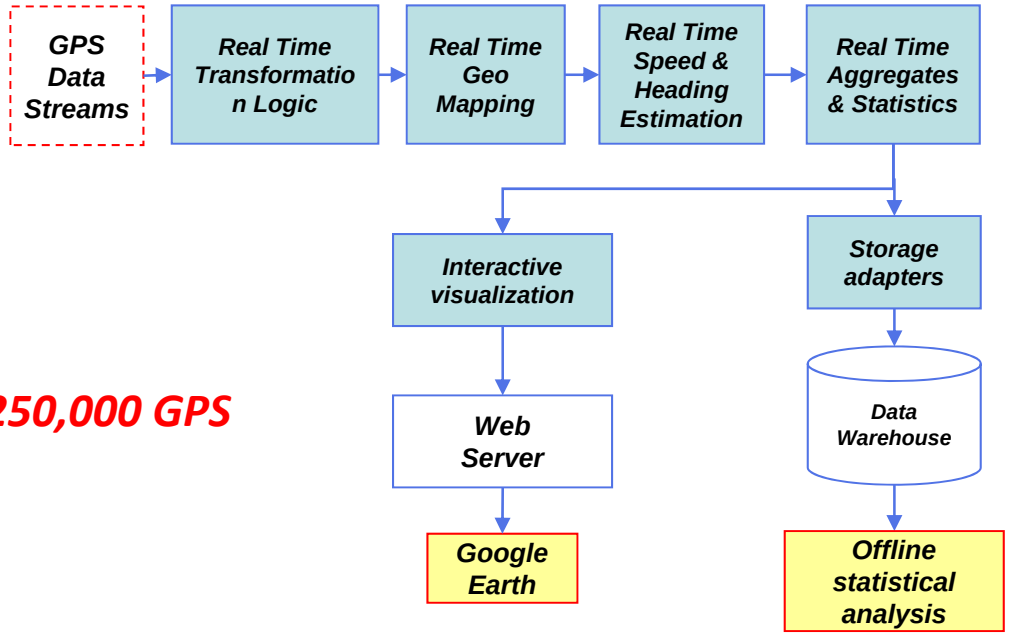


TerraEchos

Use Case – Intelligent Transport

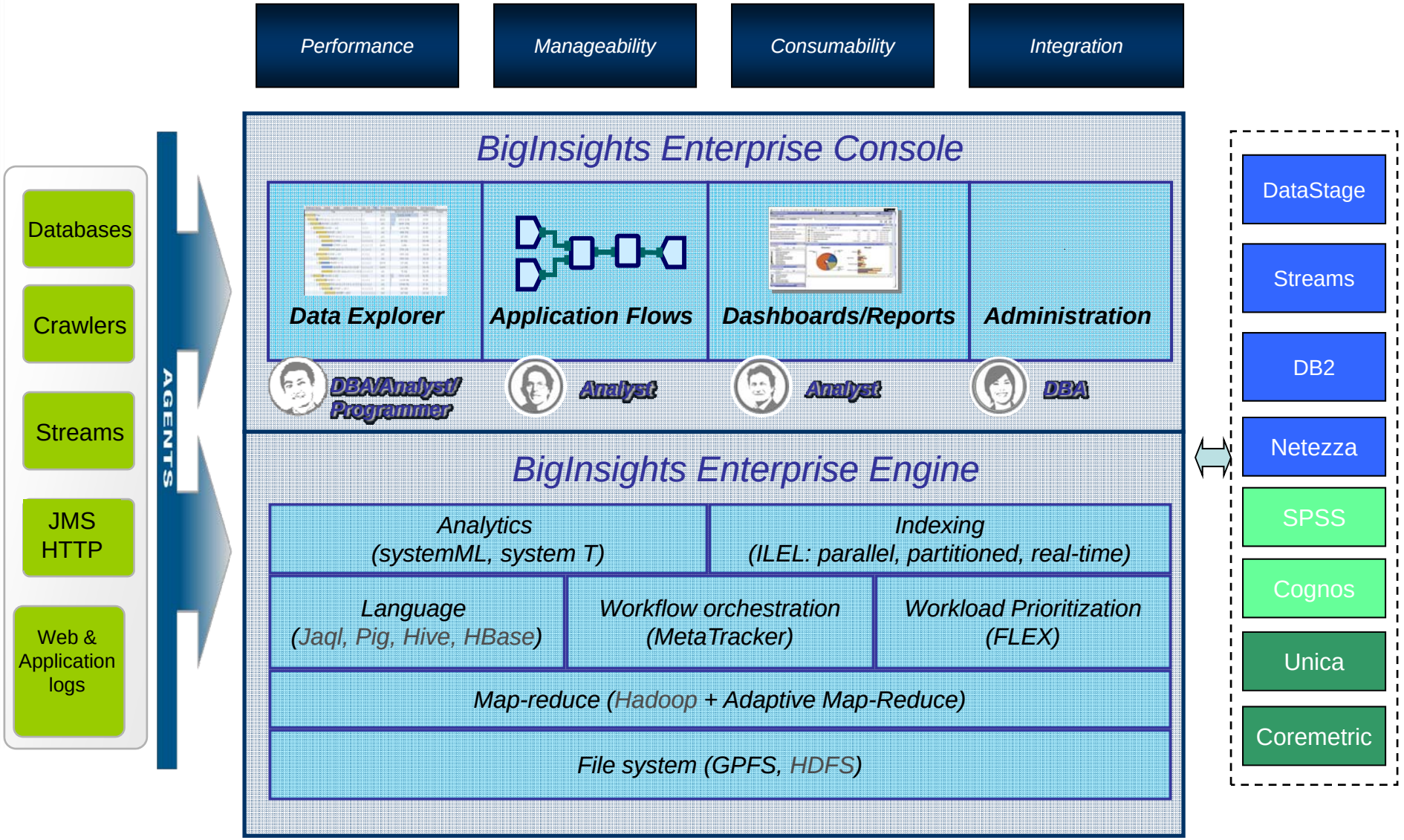
- Multimodal Data Streams
  - GPS
  - Counts, speeds, travel times
  - Public Transport
  - Pollution measurements
  - Weather Conditions
- Archiving of cleansed data
- Real Time Traffic Monitoring
- Real Time Traffic Information
- (Multimodal) Travel Planner





Only 4 x86 Blade servers to process 250,000 GPS probes per second

BigInsights Stack – Value Add on Open Software



The BigSheets Component of BigInsights

- BigSheets is the BigInsights web front end that enables non-developers to interact with BigInsights managed data and workloads.
  - Define and manage long running data collection jobs
  - Analyze content of the text on the pages that have been retrieved
  - Rich visualizations

Use Case: Credit Card Fraud

Problem

- Credit card fraud can cost up to 7 cents per 100 dollars worth of transactions – billions of dollars year
- Fraud schemes are constantly changing
  - Understanding the fraud pattern months after the fact is only partially helpful since by the time detection models have changed the crooks have moved on.

If Only Visa Could;

- Reinvent how to detect the fraud patterns
- Stop new fraud patterns before they can rack-up significant losses

Solution

- Revolutionize the speed of detection
  - Visa loaded two years of test records, or 73 billion transactions, amounting to 36 terabytes of data into hadoop. The processing time fell from one month with traditional methods to a mere 13 minutes
- Revolutionize how models are tested and promoted
  - Use hadoop to compress the processing time to test models, handle new workloads that support ad-hoc testing

Sampling Of Other Customer Use Cases

- **Use Case**: Web & system log analytics for better understanding of operational risk management
- **Use Case:** 10+ years of customer statements being converted from APF to indexed text. BigInsights is being used as the platform to power the conversion to text, create the index, and provide search results to end-users
- **Use Case**: Next generation search infrastructure based on BigInsights with Lucene++ and Unstructured analytics Modules from IBM Research
- **Use Case**: Fraud detection and prevention. Agent supplied context on claims will be combined with customer correspondence to find patterns and flag abuse
- **Use Case**: Counterparty analytics to determine credit risk exposure in commercial lending
- **Use Case**: Customer sentiment tracking and brand perception management solution
- **Use Case:** Large scale data operations and analytics capabilities for web logs, web activities, set-top info that are web enabled. Modeling has to include multiple web properties so logs will be both large in size and have widely different formats that need to be structured on the fly and then made available for machine learning operations

In my newsfeed yesterday…

JULY 11, 2011 5:36 PM PDT

Researchers mine tweets in search of health trends

by Elizabeth Armstrong Moore

Print E-mail

Recommend 6 Tweet 55 +1 3 Share 1 comment

The explosion of social media has given researchers a lot of data to mine and trends to identify, but two computer scientists at Johns Hopkins University say they've **developed sophisticated filtering software** that is attracting particular attention from public health officials.

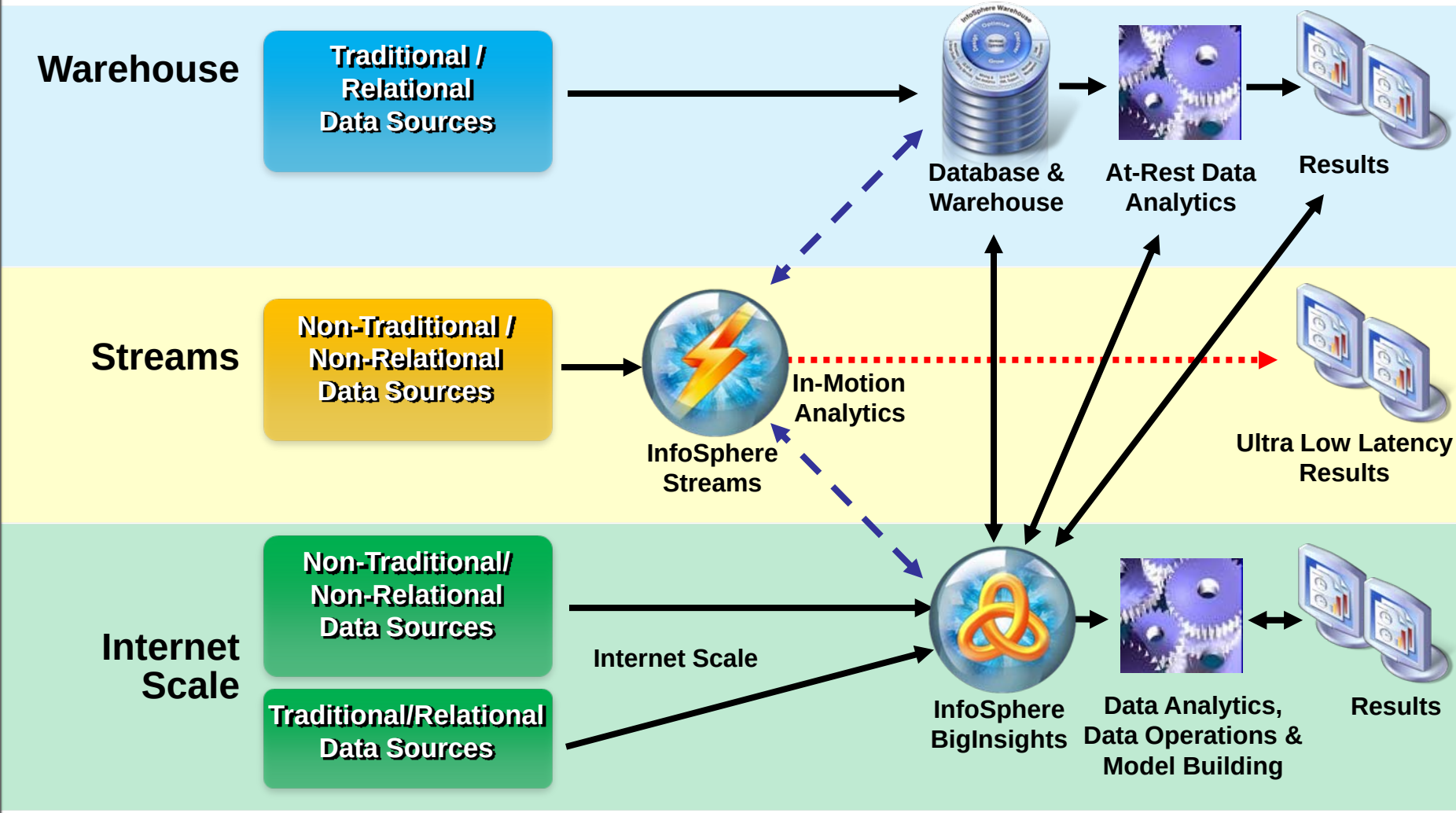
Twitter, which **launched five years ago**, has already been used by computer scientists **to try to track the flu**.

But when Johns Hopkins University computer scientists **Mark Dredze** and **Michael Paul** devised a method to filter and categorize health-related tweets, they weren't sure what they might find. So they decided to sort the tweets (they filtered 1.5 million health-related tweets from a sample of 2 billion) into electronic, ailment-specific "piles."

Johns Hopkins computer scientists Mark Dredze, left, and Michael J. Paul say that Twitter posts can provide useful public health information. (Credit: **Will Kirk**)

"There have been some narrow studies using Twitter posts, for example, to track the flu," Dredze said **in a news release**. "But to

BIG Data – Compliments what we do today



Things to consider…

- You need to think differently
  - McKinsey have published an excellent primer
    - http://www.mckinsey.com/mgi/publications/big_data/index.asp
  - Jeff Jonas Blogs
- You need good analysts
  - Ask the right question for your business
- You need flexible management systems
- Different Model
  - Traditional: Requirements -> Design -> Implementation
  - Design the platform -> Requirements evolve – rapidly -> Itereate

Evolution or Revolution

- After 150 successful implementations world wide we are experiencing rapid evolution...

Revolution is up to you…

Thankyou…

Mark McConnell
Enterprise Data Management
IBM Asia Pacific
mmcconne@au1.ibm.com

IBM Software
ExecutiveSummit2011
A Premier Leadership Exchange