Smarter Analytics Live 2013
Turning information and insight into actionable business outcomes.

Harness the power of big data for improved citizen services

Mike Kearney – Big Data Product Marketing

Canberra - 15 October 2013

twitter   #SALive2013   @ibmbaanz

SmarterAnalytics

Topics

- Some perspectives on big data

- Big data-driven case studies - extracting insights to improve service

- An enterprise model for big data analysis

- Cultivating big data adoption

Without analytics, big data is all cost and no benefit

> "*The value of big data*
> *lies in our ability to extract insights*
> *and make better decisions*"
>
> Dr. Michael Rappa
> Director of the Institute for Advanced Analytics & Distinguished University Professor
> North Carolina State University

http://analytics.ncsu.edu/?p=4770
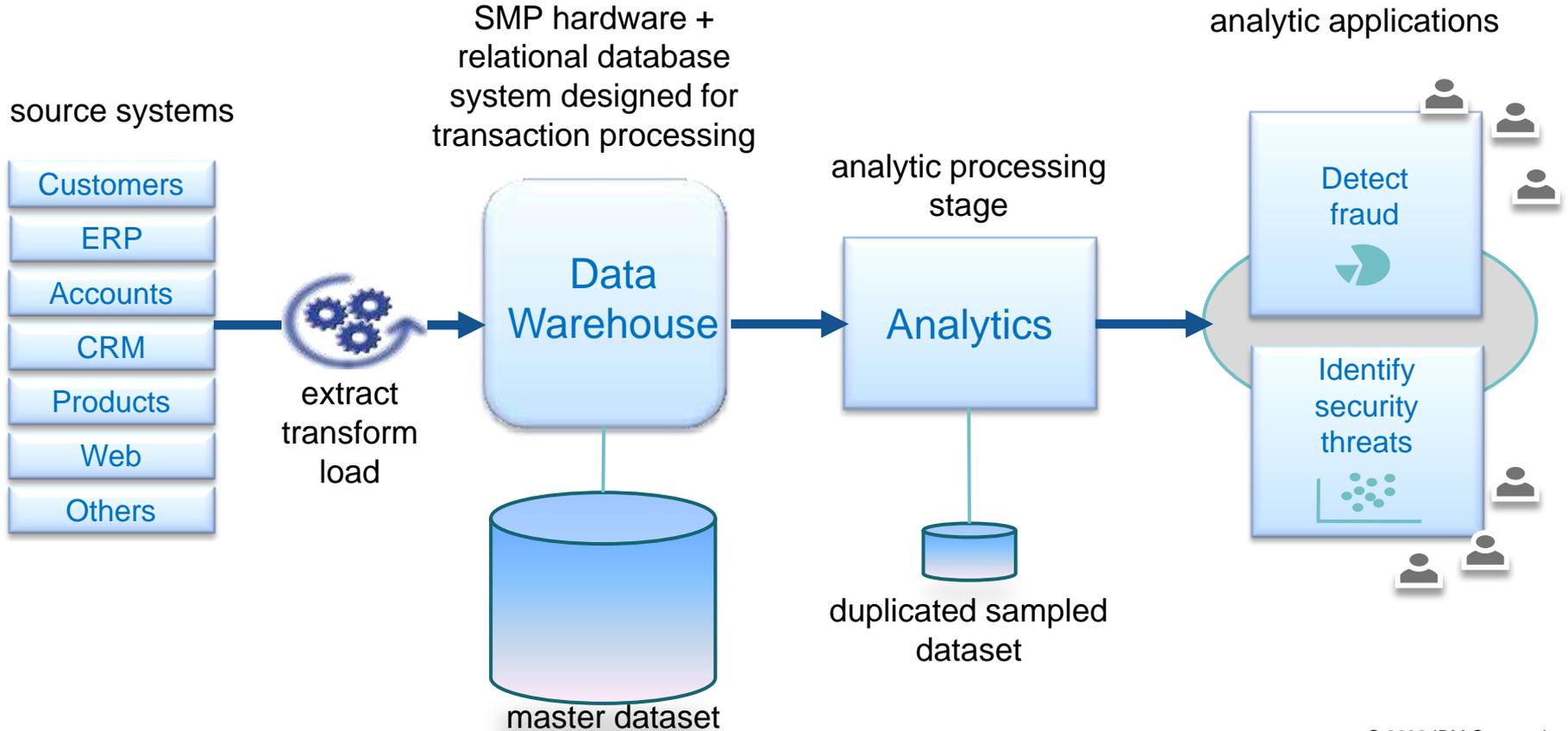
Comment on IRS Compliance data warehouse

Frequent questions that require advanced analytics must be off-loaded from the database to a separate server where numerical analysis can be performed, which creates a significant I/O bottleneck.

This is one of the most important issues that Big Data must address in the next 5---10 years.

Software vendors need to push more of their analytic APIs into the database to eliminate the need to off-load data, and hardware vendors need to provide faster disk speeds and network I/O.

www.techamericafoundation.org/content/wp-content/uploads/2012/10/Final-Big-Data-Case-Study-IRS-Compliance-Data-Warehouse.pdf

IBM

Traditional analytics: sampled, offline data

SMP hardware +
relational database
system designed for
transaction processing

analytic applications

source systems

Customers

ERP

Accounts

CRM

Products

Web

Others

extract
transform
load

Data
Warehouse

analytic processing
stage

Analytics

Detect
fraud

Identify
security
threats

master dataset

duplicated sampled
dataset

Topics

- Some perspectives on big data

- Big data-driven case studies - extracting insights to improve service

- An enterprise model for big data analysis

- Cultivating big data adoption

Study 1: Investigating multiple sclerosis at SUNY Buffalo

"*The genetics in MS are relatively complex and no single factor has been identified as singly caused. In this context we began work on gene-gene and gene-environment analysis. We have developed new metrics to identify interactions, but the problem is computationally explosive.*" Dr. Murali Ramanathan, Research Lead

public.dhe.ibm.com/common/ssi/ecm/en/.../IMC14675USEN.PDF

Single nucleotide polymorphisms
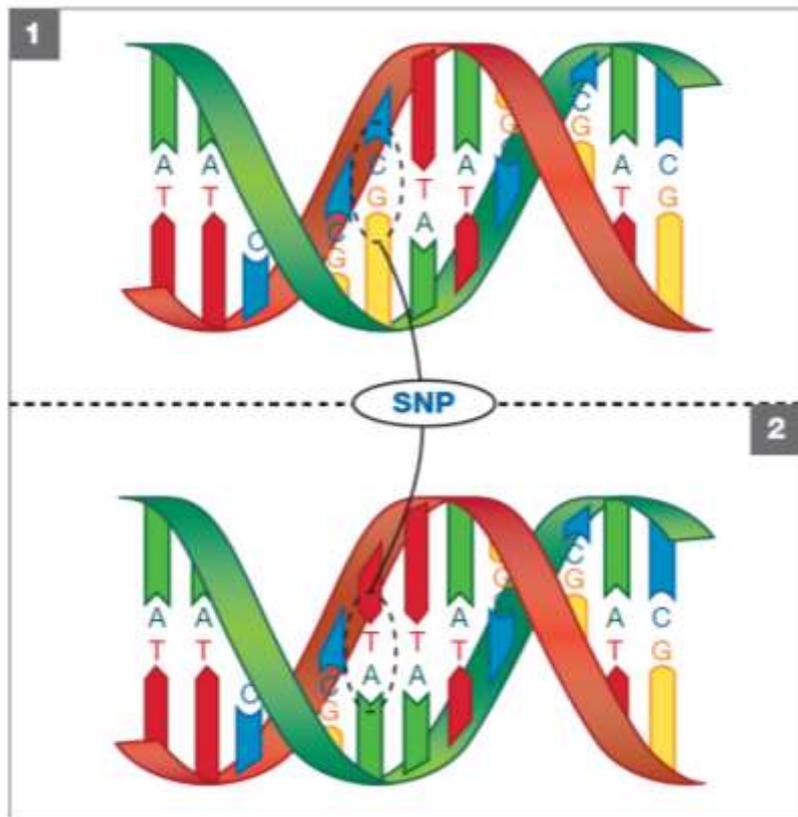

SNP is a sequence variation within a single nucleotide:

e.g., C+ G pairing replaced with T+A

Gene-to-gene interaction and MS occurrence

Individual	SNiP1	SNiP2	SNiP3	SNiPn	Phenotype
Patient 1	A	C	C	C	No Disease
Patient 2	T	G	T	A	Disease
Patient 3	A	A	T	T	Disease
Patient N	C	A	G	C	No Disease

Considering that there are thousands of single nucleotide polymorphisms , the number of combinations of SNPs that have to be assessed to uncover potential interaction becomes astronomical

Combinatorial explosive analysis
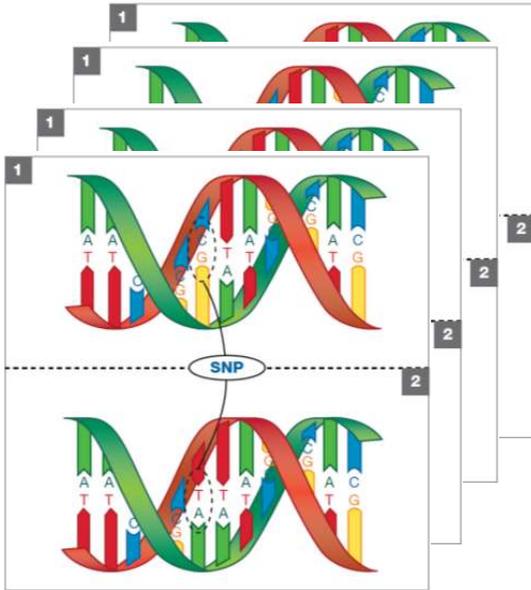
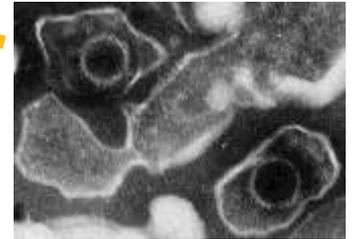interaction involving combinations of 1000s of single nucleotide polymorphisms
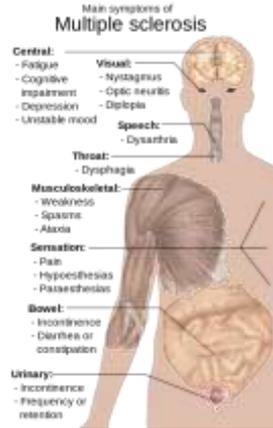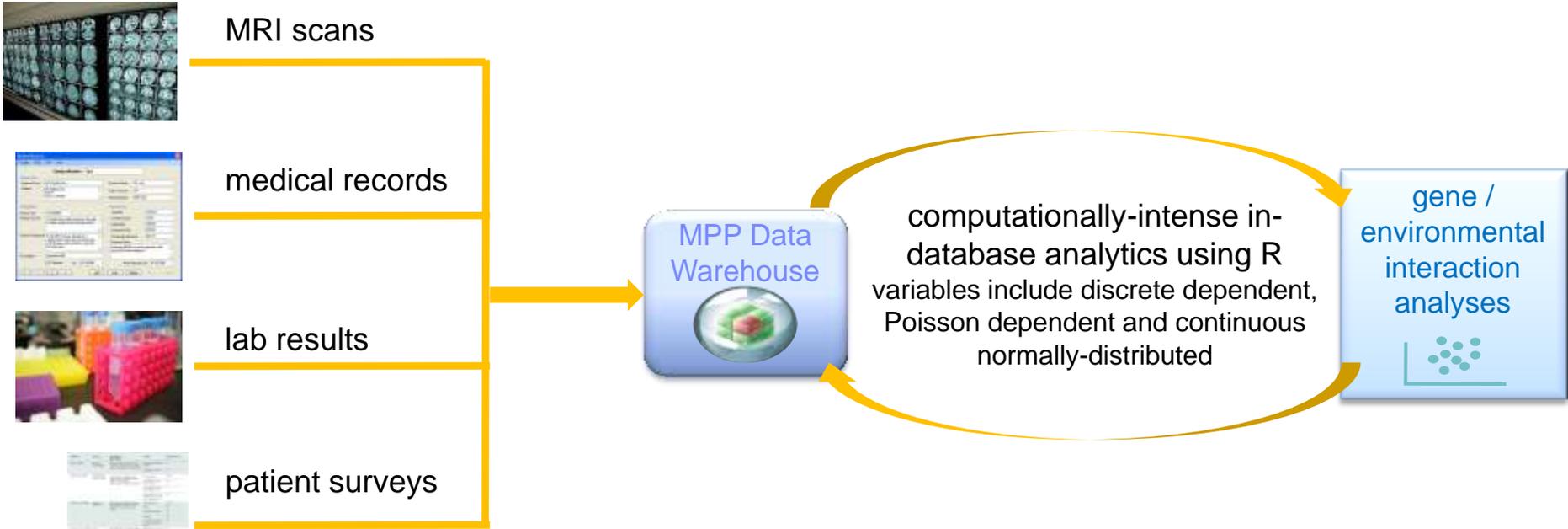
…. with data on sunlight exposure and vitamin D3 levels

…. and with data on infection with Epstein-Barr virus e.g., Mononucleosis

According to research lead, Dr. Murali Ramanathan, a critical fact in the study of MS is that "*gene products work by interacting with both other gene products and environmental factors.*"

R in-database analytics at SUNY Buffalo

MRI scans

medical records

lab results

patient surveys

MPP Data Warehouse

computationally-intense in-database analytics using R
variables include discrete dependent, Poisson dependent and continuous normally-distributed

gene / environmental interaction analyses

The algorithms were taking several days to run the data sets on standard servers. "*Once we had the data on Netezza we were able to do the same analysis and much more complex analysis in minutes*" - Dr. Murali Ramanathan

Study 2: Re-engineering neonatal healthcare

Dr. Carolyn McGregor
Canada Research Chair
in Health Informatics

University of Ontario
Institute of Technology

www.ibmbigdatahub.com/video/ibm-stream-computing-smarter-healthcare

"Being able to analyse streams in real-time is providing us with real opportunities for much earlier proactive intervention in the prediction of onset of illness"

Predicting potential health-threatening events

real-time
intervention

capacity for
38 babies
in ICU

**medical
devices**

Real-time analysis of data streaming from ICU

blood pressure

blood oxygen saturation

heart rate

**Real-time
alerts**
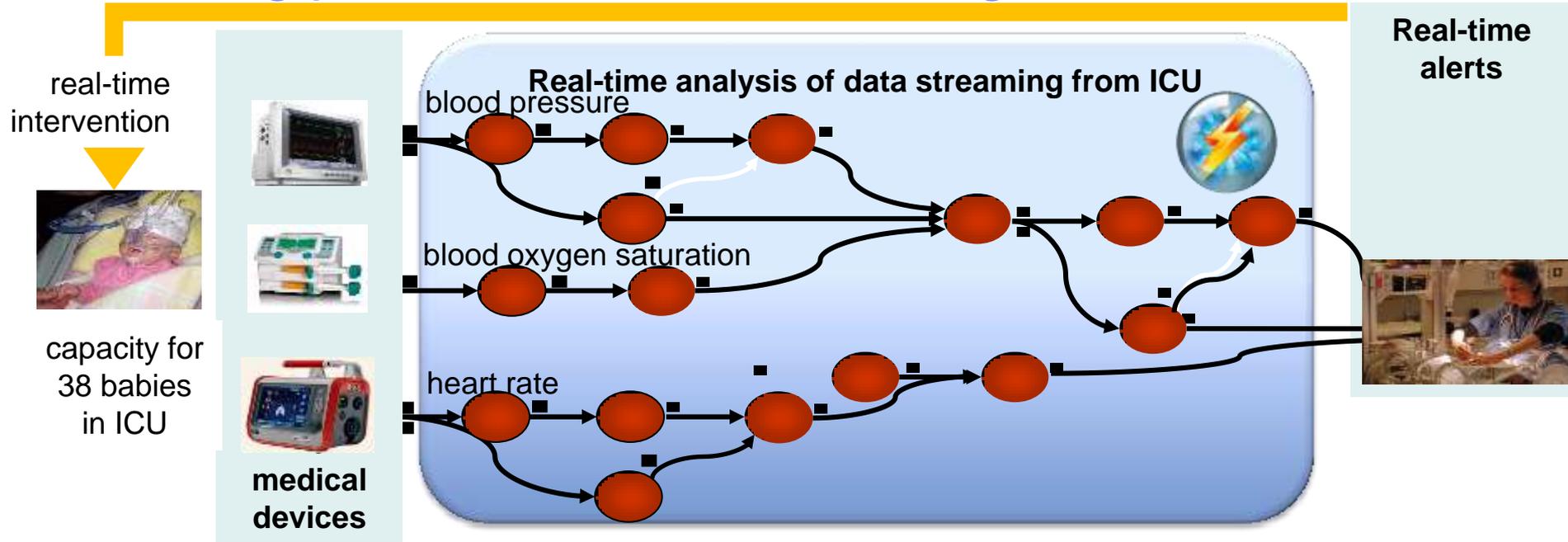
Understanding NOW in the context of history

real-time intervention

capacity for 38 babies in ICU

medical devices

Real-time analysis of data streaming from ICU

blood pressure

blood oxygen saturation

heart rate

Real-time alerts

refined / new analytic models

trickle feed updates

Event modelling and analysis

MPP data warehouse (growing and persistent store for real-time stream data)
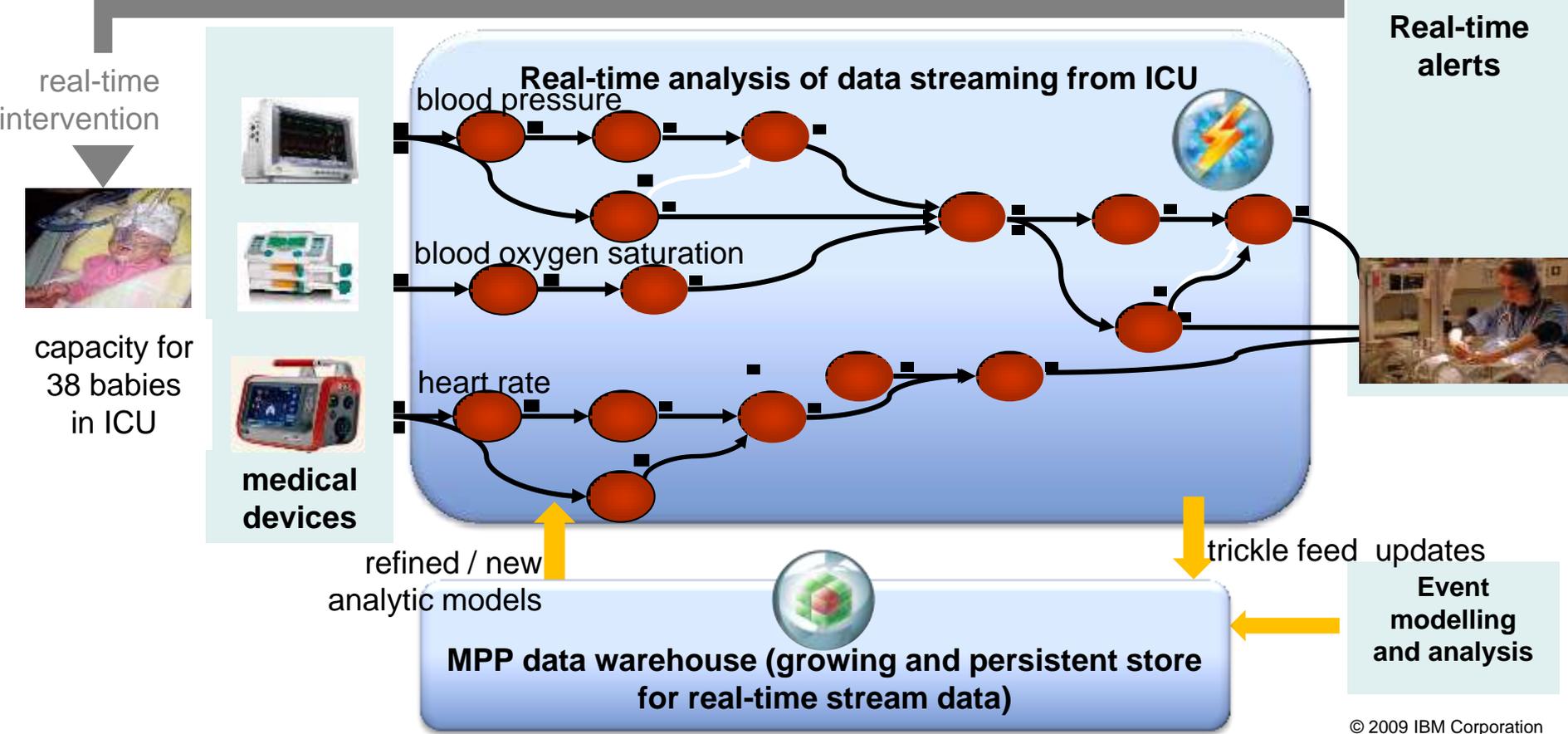
Study 3: Reducing carbon output at PNNL Smart Grid

http://www.pnnl.gov/news/release.aspx?id=776

"We looked at a bunch of direct and indirect mechanisms where the Smart Grid can help – altogether they add up to about at 12% reduction in carbon from our electric grid"

Pacific Northwest Demonstration Project

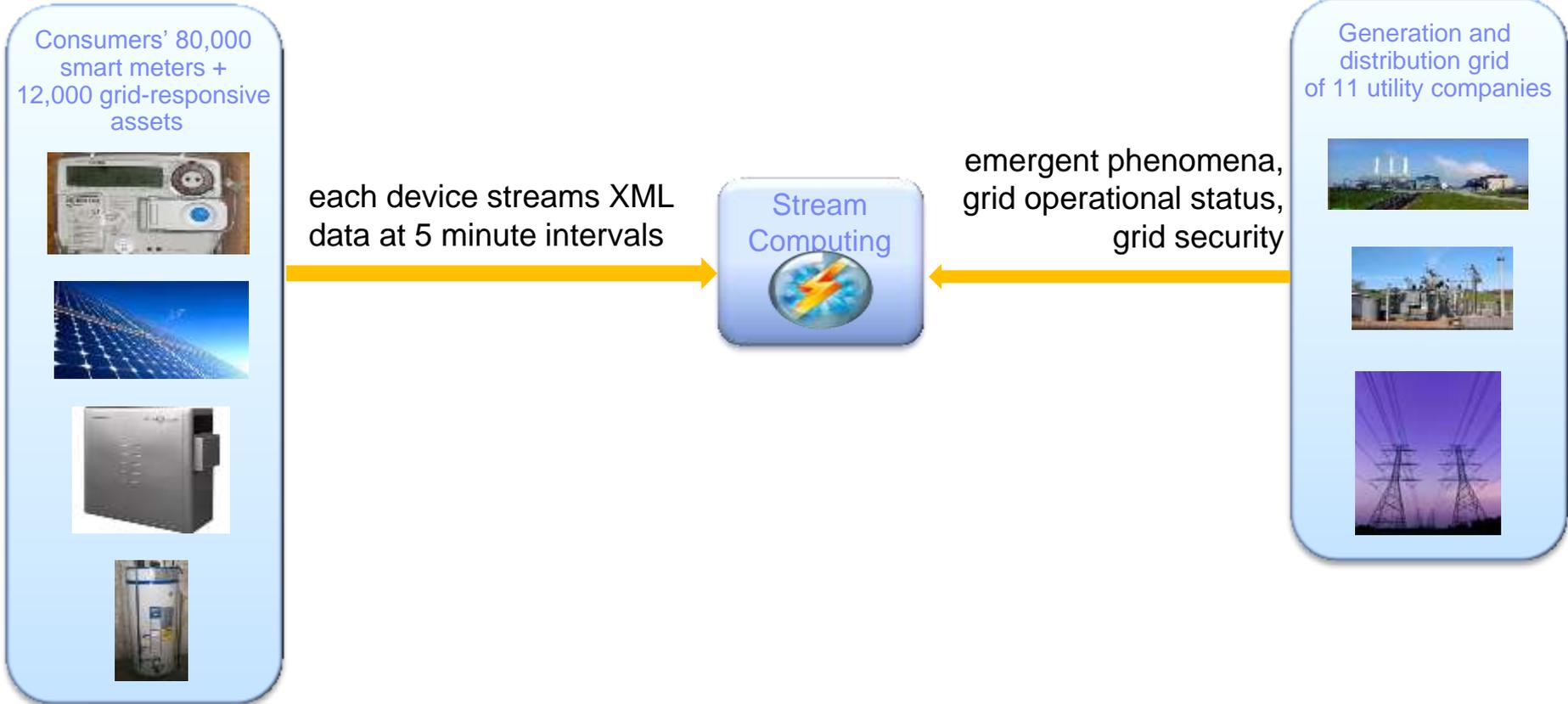www.pnwsmartgrid.org
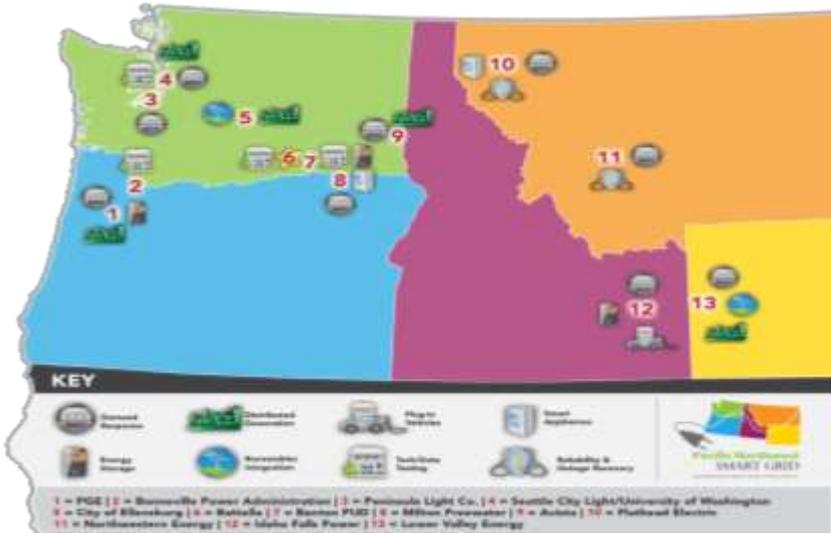
- 5 year demonstration project
- 60,000 metered customers across 5 states – largest smart grid project in USA
- Designed to let consumer participate and benefit
- Led by Battelle and partners including 11 utilities, 2 universities, and 5 vendors (IBM & Netezza)

IBM

Understanding supply and demand

Consumers' 80,000 smart meters + 12,000 grid-responsive assets

each device streams XML data at 5 minute intervals

Stream Computing

emergent phenomena, grid operational status, grid security

Generation and distribution grid of 11 utility companies

Dynamic pricing to balance supply and demand

Consumers' 80,000 smart meters + 12,000 grid-responsive assets

dynamic price for electricity

each device streams XML data at 5 minute intervals

Stream Computing

generation plan – matches supply to demand

emergent phenomena, grid operational status, grid security

Generation and distribution grid of 11 utility companies

IBM
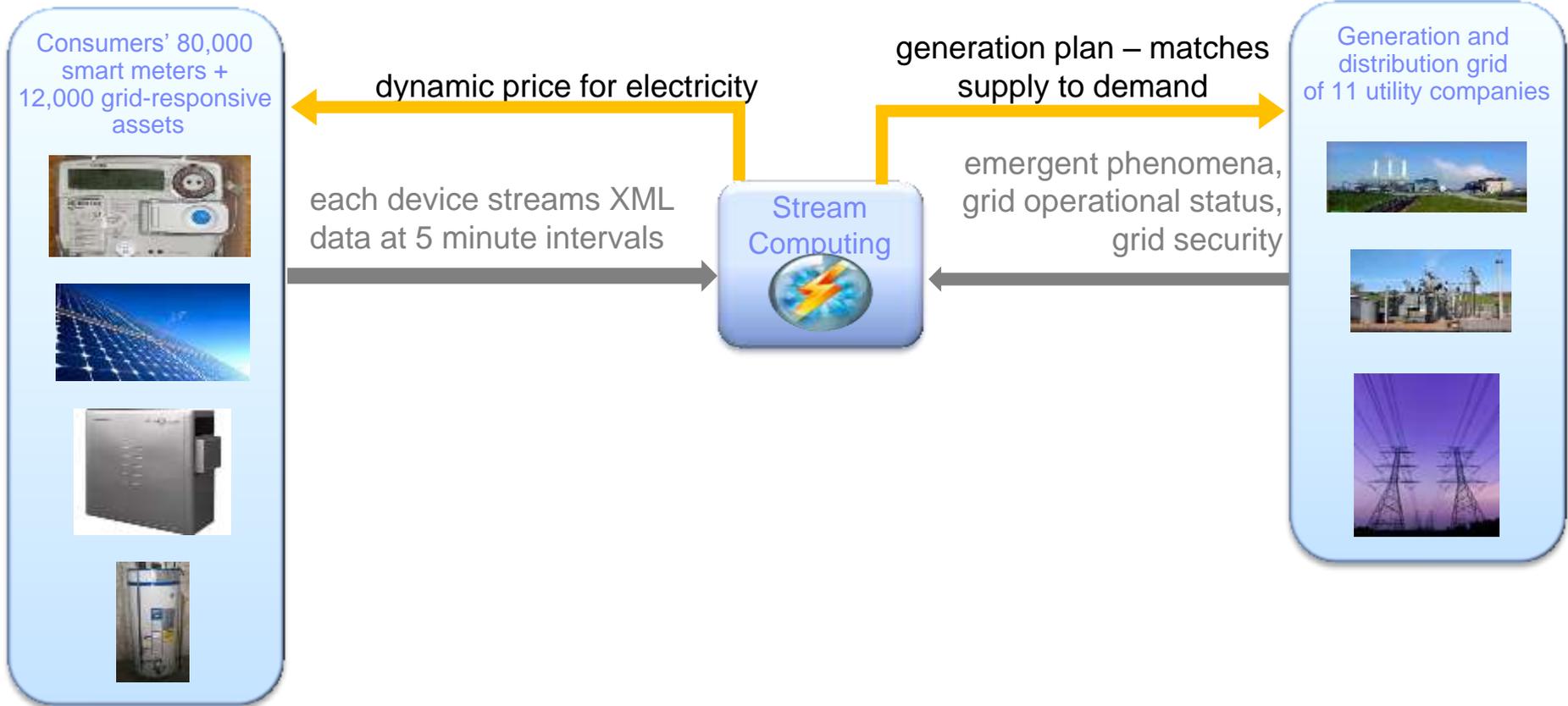
Controlling demand to maintain supply

real-time control signal, e.g. shutdown water heaters

Consumers' 80,000 smart meters + 12,000 grid-responsive assets

dynamic price for electricity

generation plan – matches supply to demand

Generation and distribution grid of 11 utility companies

each device streams XML data at 5 minute intervals

Stream Computing

emergent phenomena, grid operational status, grid security

IBM

Understanding history to optimise control

real-time control signal, e.g. shutdown water heaters

Consumers' 80,000 smart meters + 12,000 grid-responsive assets

generation plan – matches supply to demand

dynamic price for electricity

Generation and distribution grid of 11 utility companies

each device streams XML data at 5 minute intervals

Stream Computing

emergent phenomena, grid operational status, grid security

feedback loop – refining analytic models: "what set of conditions create opportunity to maintain service and reduce carbon?"

trickle-feed updates

MPP Data Warehouse

petabyte-scale analysis of 10 years' data – detailed understanding of how the grid functions

# Study 4: Informed public transportation in Dublin, Eire
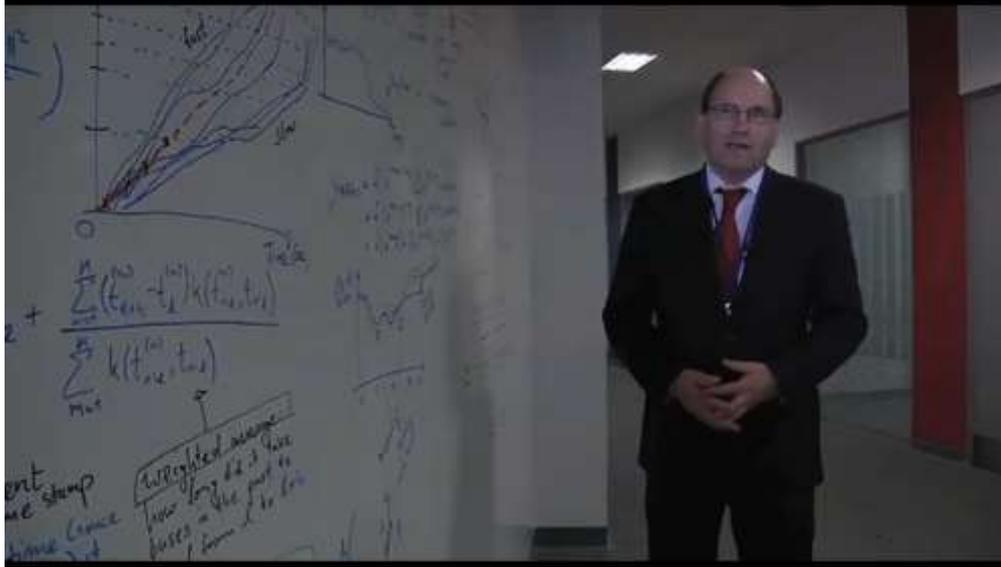
www.youtube.com/watch?v=YjKSkoiD9Yo

"*We can see how our transport network is working as a whole – and develop innovative ways to improve it for Dublin's citizens*"

Brendan O'Brien, Head of Technical Services – Roads and Traffic Department, Dublin City Council

An efficient citywide traffic awareness system for Dublin City

1000 buses
on 150 routes
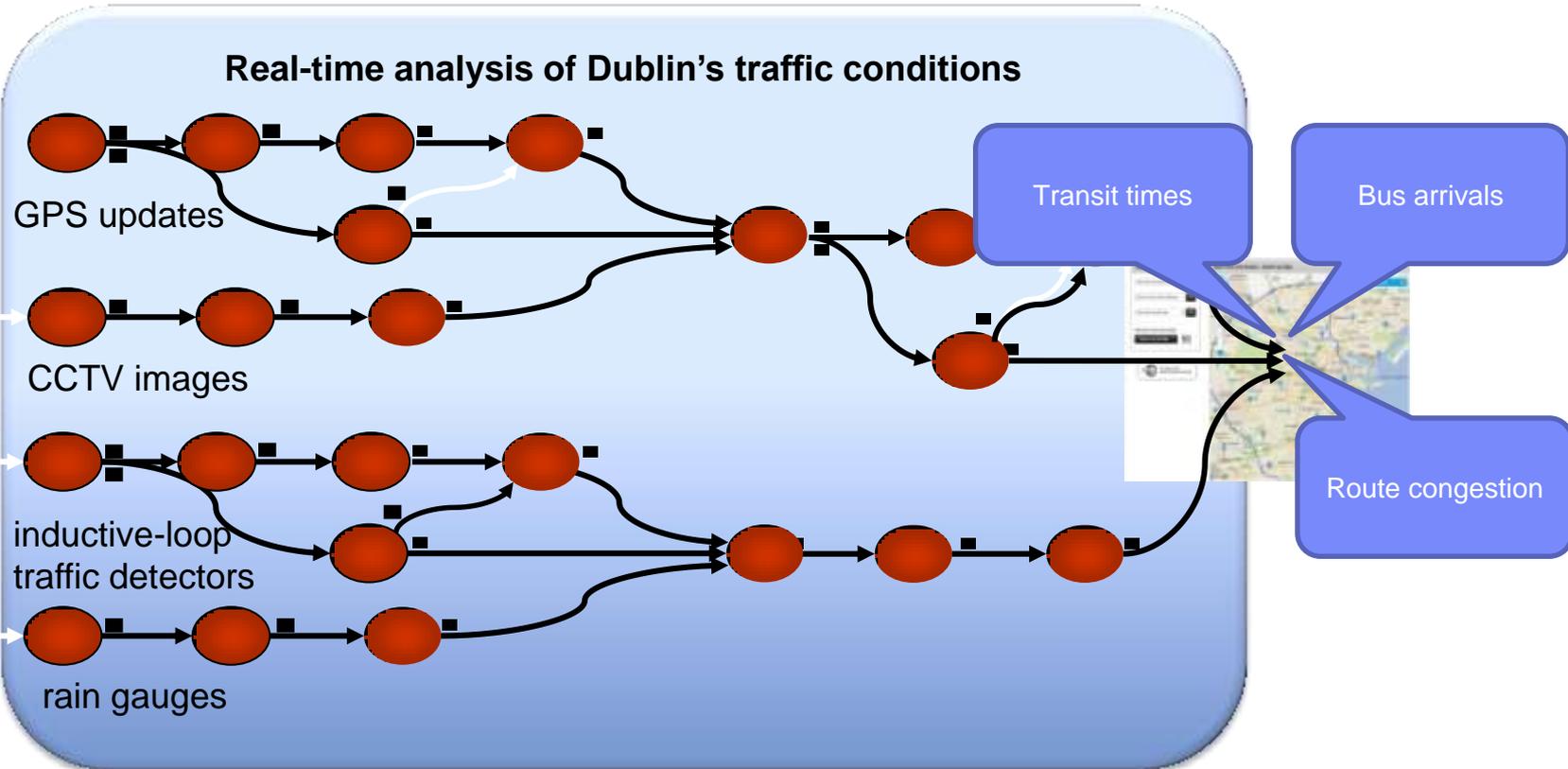
Study 5: Securing the perimeter of a U.S. Department of Energy National Lab
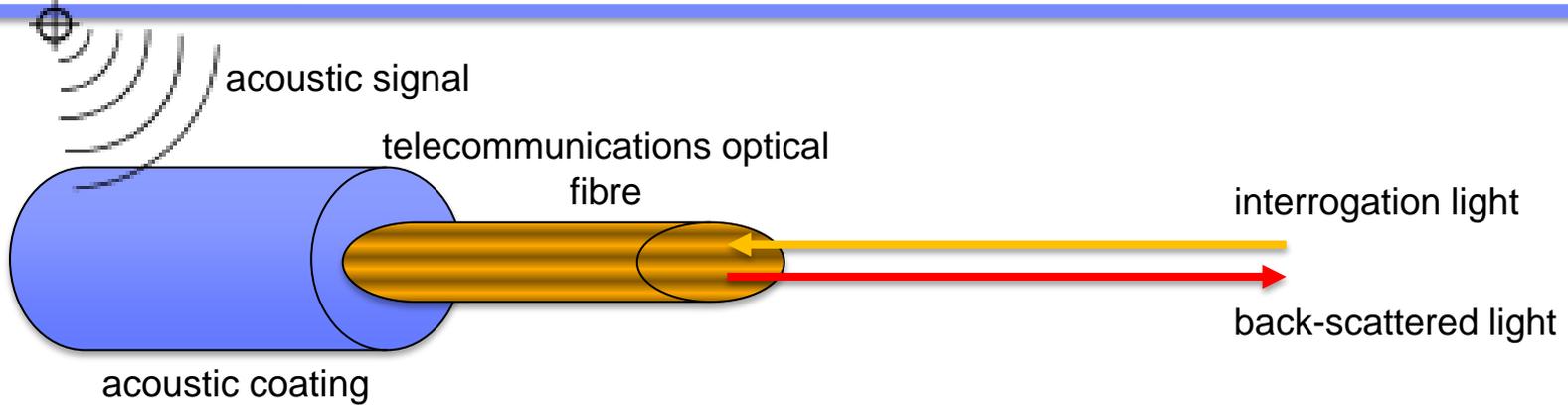

www.youtube.com/watch?v=2DQLQk_dnhM

"*We are not only capturing and reducing a tremendous amount of digital acoustic data, but we are also running intensive computational statistical analysis - and we are doing all of this in one-fourteenth of a second*"

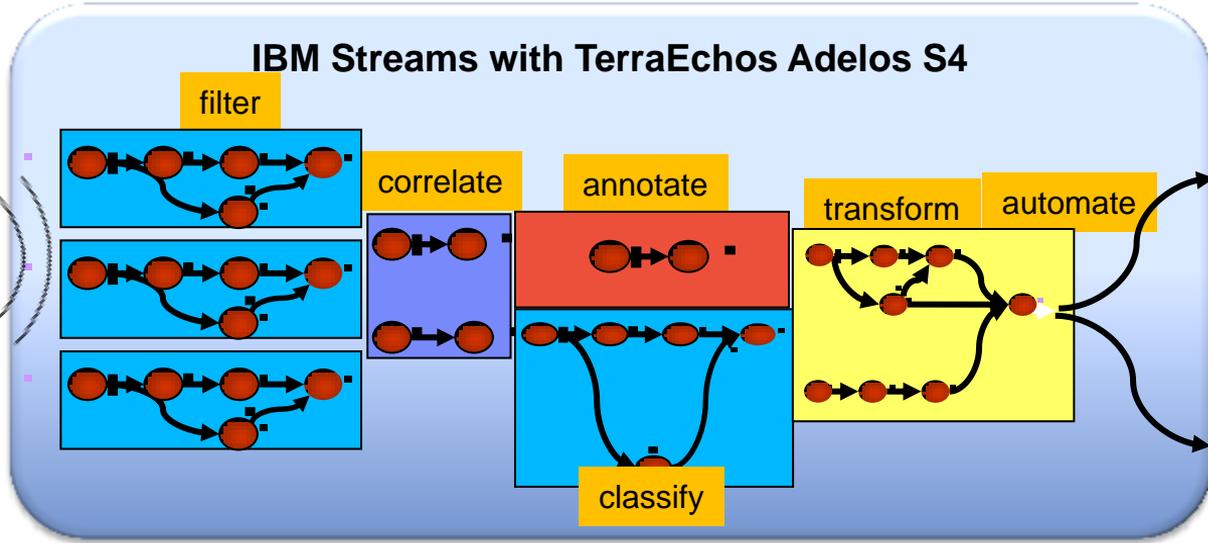- Dr. Alex Philp, Founder and CTO, TerraEchos, Inc.

Continuous and persistent sensing

underground cable multiplexed to create a continuous array of sensors

acoustic signal

telecommunications optical fibre

interrogation light

back-scattered light

acoustic coating

Detect, classify, locate and track potential threats

IBM Streams with TerraEchos Adelos S4

filter · correlate · annotate · transform · automate · classify
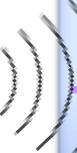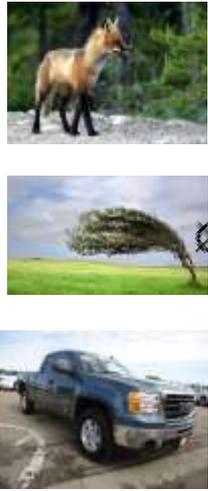
cues cameras to focus on area of interest

alerts security personnel

275 Mbit of acoustic data from 1,024 individual sensor channels

processing up to 42 terabytes of acoustic data per day

Topics

- Some perspectives on big data

- Big data-driven case studies - extracting insights to improve service

- An enterprise model for big data analysis

- Cultivating big data adoption

IBM

Managing and analyzing structured data
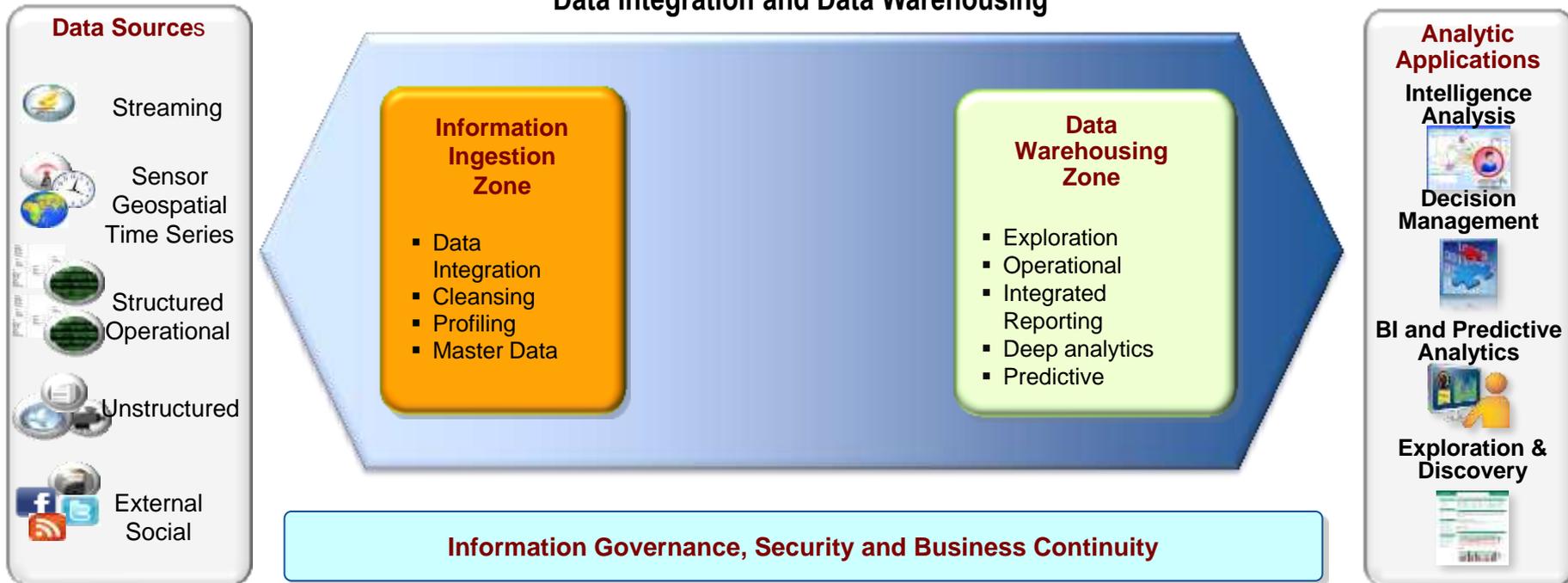
Data Integration and Data Warehousing

Data Sources

- Streaming
- Sensor Geospatial Time Series
- Structured Operational
- Unstructured
- External Social

Information Ingestion Zone

- Data Integration
- Cleansing
- Profiling
- Master Data

Data Warehousing Zone

- Exploration
- Operational
- Integrated Reporting
- Deep analytics
- Predictive

Information Governance, Security and Business Continuity

Analytic Applications

Intelligence Analysis

Decision Management

BI and Predictive Analytics

Exploration & Discovery

IBM

New architecture to analyze and manage all data

Big Data Platform

Data Sources

- Streaming
- Sensor Geospatial Time Series
- Structured Operational
- Unstructured
- External Social

Real-time Analytic Zone
- Video/Audio
- Network/Sensor
- Entity Analytics
- Predictive

Information Ingestion Zone
- Data Integration
- Cleansing
- Profiling
- Master Data

Landing Area, Analytics and Archive Zone
- Exploration
- Text Analytics
- Data Mining
- Entity Analytics
- Machine Learning

Data Warehousing Zone
- Exploration
- Operational
- Integrated Reporting
- Deep analytics
- Predictive

Analytic Applications

Intelligence Analysis

Decision Management

BI and Predictive Analytics

Exploration & Discovery

Information Governance, Security and Business Continuity

IBM

Multiple systems integrate as a comprehensive platform

Big Data Platform

Data Sources
- Streaming
- Sensor Geospatial Time Series
- Structured Operational
- Unstructured
- External Social

Information Ingestion Zone

InfoSphere Information Server

Real-time Analytic Zone

InfoSphere Streams

Landing Area, Analytics and Archive Zone

PureData
System for Hadoop

Data Warehousing Zone

PureData
System for Analytics

PureData
System for Operational Analytics

Information Governance, Security and Business Continuity

Analytic Applications
- Intelligence Analysis
- Decision Management
- BI and Predictive Analytics
- Exploration & Discovery

IBM

Traditional analytics: sampled, offline data

source systems

SMP hardware +
relational database
system designed for
transaction processing

analytic applications

analytic processing
stage

| Customers |
| ERP |
| Accounts |
| CRM |
| Products |
| Web |
| Others |

extract
transform
load

Data
Warehouse

Analytics
e.g. SAS

Fraud
detection

Reduce
security
threats

master dataset

duplicated sampled
dataset

IBM

The NOW and THEN of big data

data warehouse and/or Hadoop

real-time stream processing

C
o
n
t
i
n
u
o
u
s

m
o
n
i
t
o
r
i
n
g

data

data ingest

data ingest

text analysis

entity analysis

predictive analytics

enriched model

social media analytics

text analysis

geospatial analysis

statistical modelling

data mining

predictive analytics

analytic model development

large and growing historic data sets provide rich context for model development

automated action

insight to better decisions

forecast

Topics

- Some perspectives on big data

- Big data-driven case studies - extracting insights to improve service

- An enterprise model for big data analysis

- **Cultivating big data adoption**

DEMYSTIFYING BIG DATA
A Practical Guide To Transforming The Business of Government
TechAmerica FOUNDATION   Prepared by TechAmerica Foundation's Federal Big Data Commission

TechAmerica Foundation leadership team:

Steve Mills (Co-Chair)
Senior Vice President and Group Executive, IBM
Steve Lucas (Co-Chair)
Global Executive Vice President and General Manager, Database and Technology, SAP
Leo Irakliotis (Academic Co-Chair)
Western Governors University
Michael Rappa (Academic Co-Chair)
North Carolina State University
Teresa Carlson (Vice Chair)
Vice President Global Public Sector, Amazon Web Services
Bill Perlowitz (Vice Chair), CTO, Science Technology and Engineering Group, Wyle

http://www.techamerica.org/Docs/fileManager.cfm?f=techamerica-bigdatareport-final.pdf

Step 1: Define business requirements

Reduce the cost of tax filing by 3%

Reduce incidence of death from hospital acquired infection by 300

Enable customers to make informed decisions about their energy use to reduce carbon output by 2%

Start with a set of specific and well-defined mission requirements

versus …..

a plan to deploy a universal and unproven technical platform to support perceived future requirements
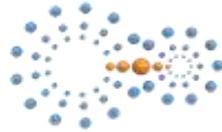
Step 2: Plan to augment and iterate
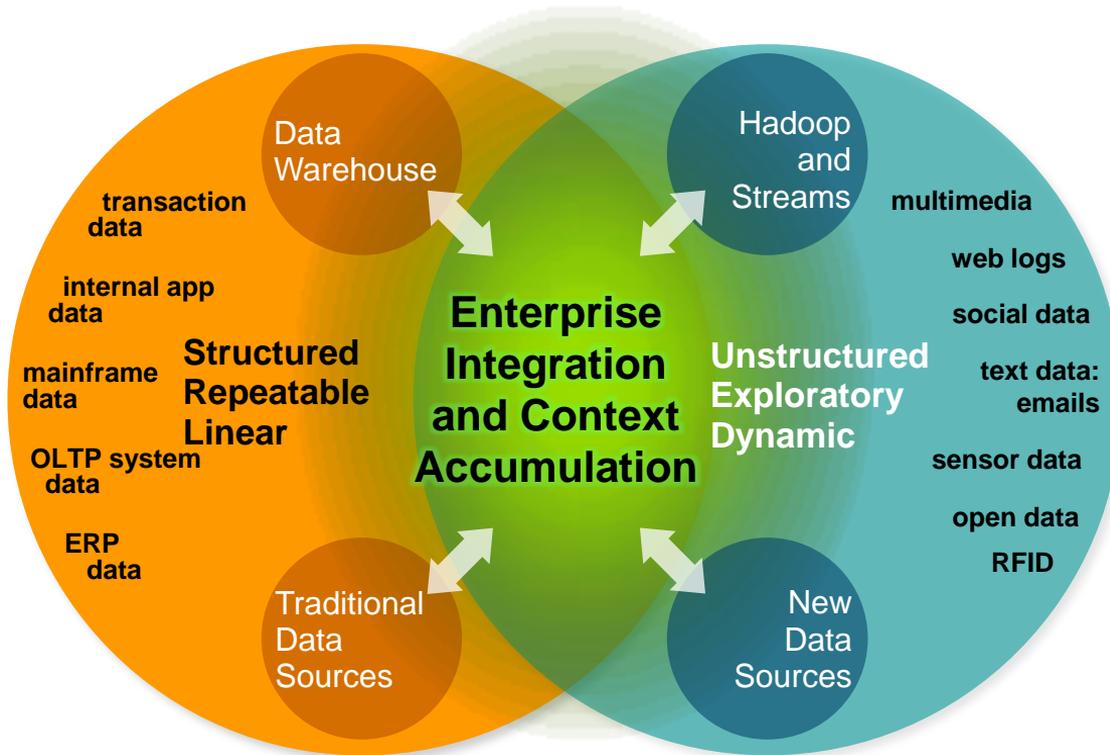
Augment current IT investments rather than building entirely new enterprise scale systems

Focus new integrated capabilities on initial requirements within a larger architectural vision that can embrace wider use cases in subsequent phases of deployment

Step 3: Select your big data entry point

Velocity: use cases requiring a high degree of velocity in data processing with real-time decision making, tend to require stream processing as an entry point

Variety: use cases requiring exploration of structured, semi-structured and unstructured data, imply Hadoop as an entry point

Volume: select as an entry point a warehouse architecture that can scale out without pre-defined bounds
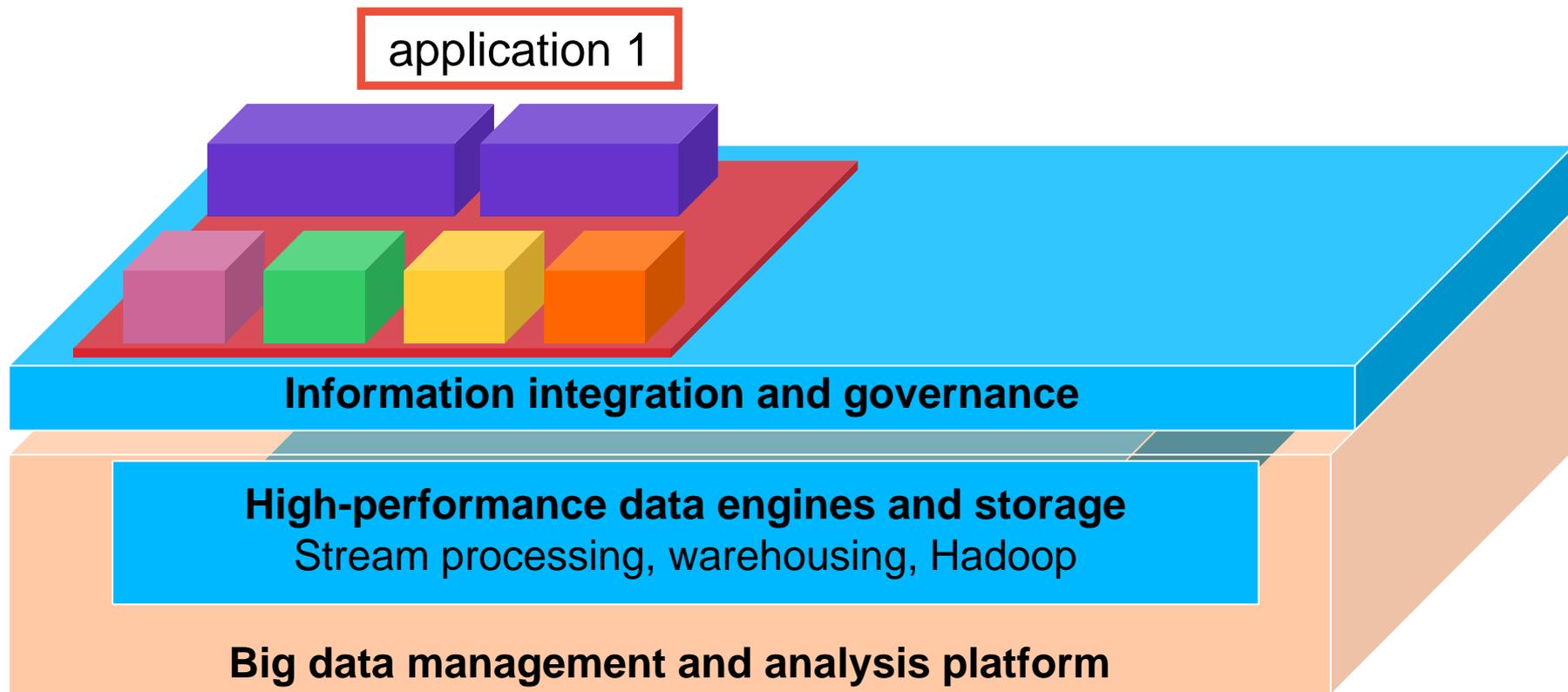
Step 4: Identify gaps and plan investments

Define an initial set of business requirements

Assess technical requirements to achieve consistency with long-term architecture
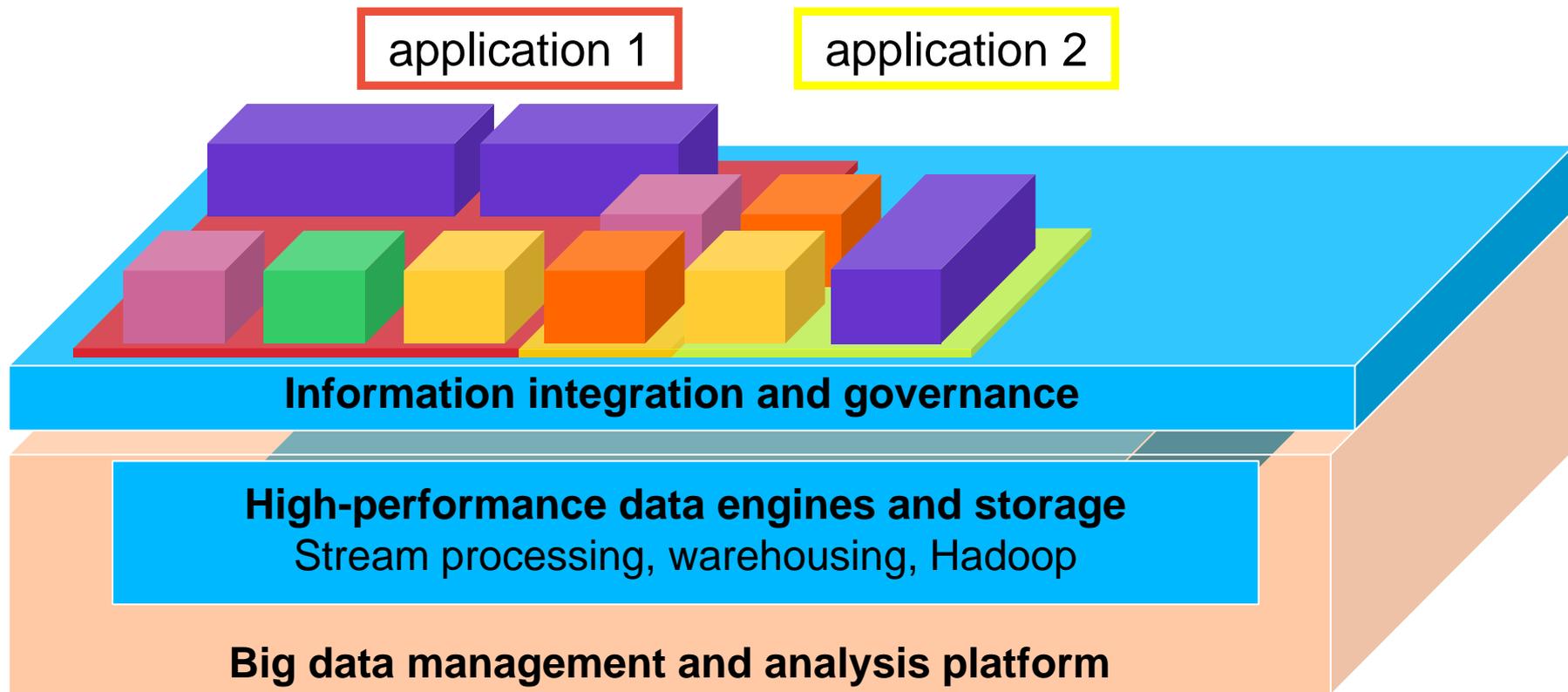
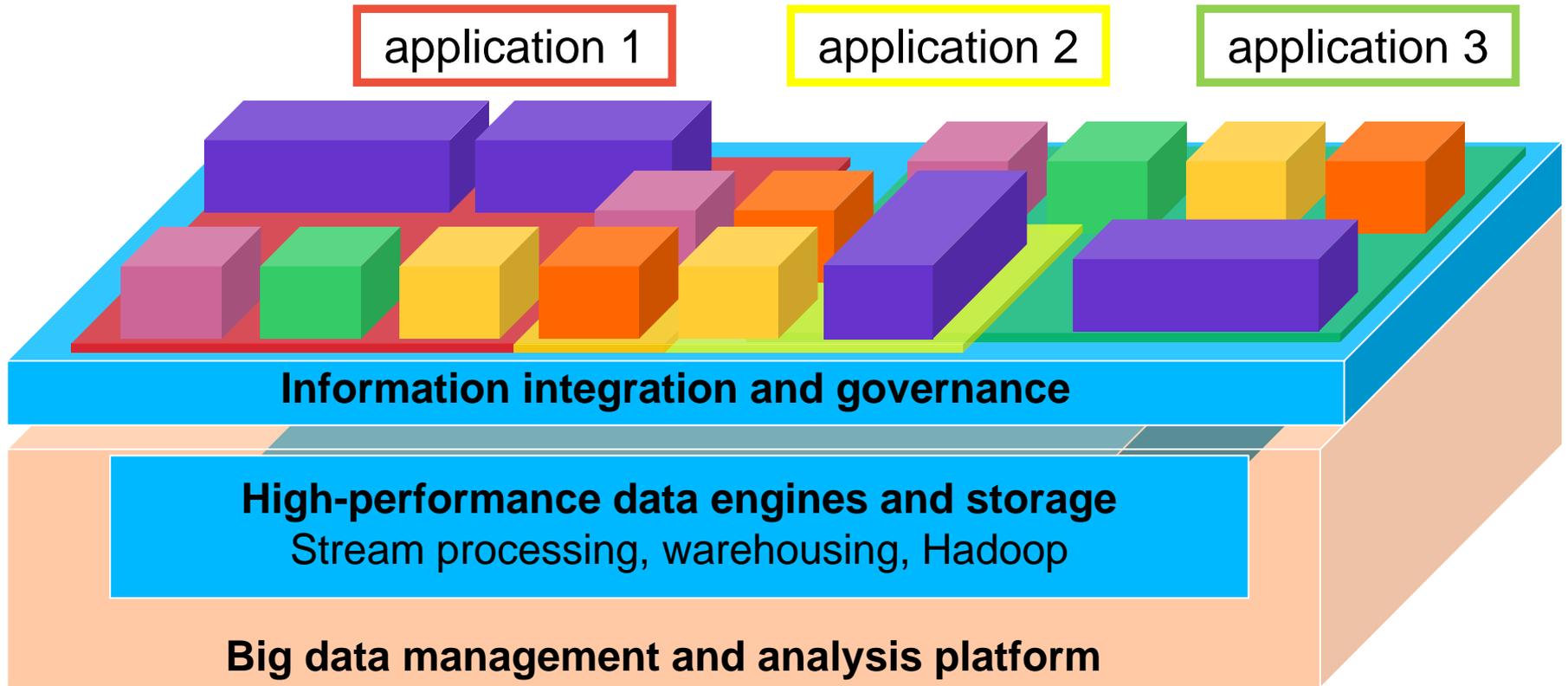**Identify** gaps in technical architecture

Plan investments to fill gaps

IBM

Build first application

application 1

Information integration and governance

High-performance data engines and storage
Stream processing, warehousing, Hadoop

Big data management and analysis platform

IBM

Build first application and iterate

application 1

application 2

Information integration and governance

High-performance data engines and storage
Stream processing, warehousing, Hadoop

Big data management and analysis platform

Build first application and iterate

Closing thoughts

- Extracting insights from big data creates opportunities for improved citizen services…..

- …. *when* big data management platforms analyse data *in-situ*

- Big data changes how organisations make decisions

- Big data changes how organisations deliver value

Thank you

Don't forget to check out the Demo Zone during the break
or talk to one of our experts at the Table Talk area.

SmarterAnalytics