Everything You Need To Know About Big Data Analytics In 20 Minutes or Less*

Tom Deutsch

Program Director, IBM

tdeutsch@us.ibm.com

* It might take more than 5 minutes, no promises…

Statement: Very Large And Diverse (Big Data) Is Faster

As Evidenced By

Your Data Flows

Reporting Handles The Edges, But What About The Middle?

Quick – Which Actually Contains More Data?

A

B

Quick - Which Of Those Would You Complete Faster?

A

B

So Why Are We Faster When We Actually Need To Process More Data?

It Is Because Our Brains Pick Up The Patterns.

The Human Brain Excels at Context And Rejecting Processing Unnecessary Information

Big Data Analytics Do The Very Same Thing

Quick – What If I Threw A Puzzle Piece From Another Puzzle In?

A

B

With Proper Analytics Big Data Is Both Faster And More Precise When While Volumes Scale Up

Unlike Human Brains However…

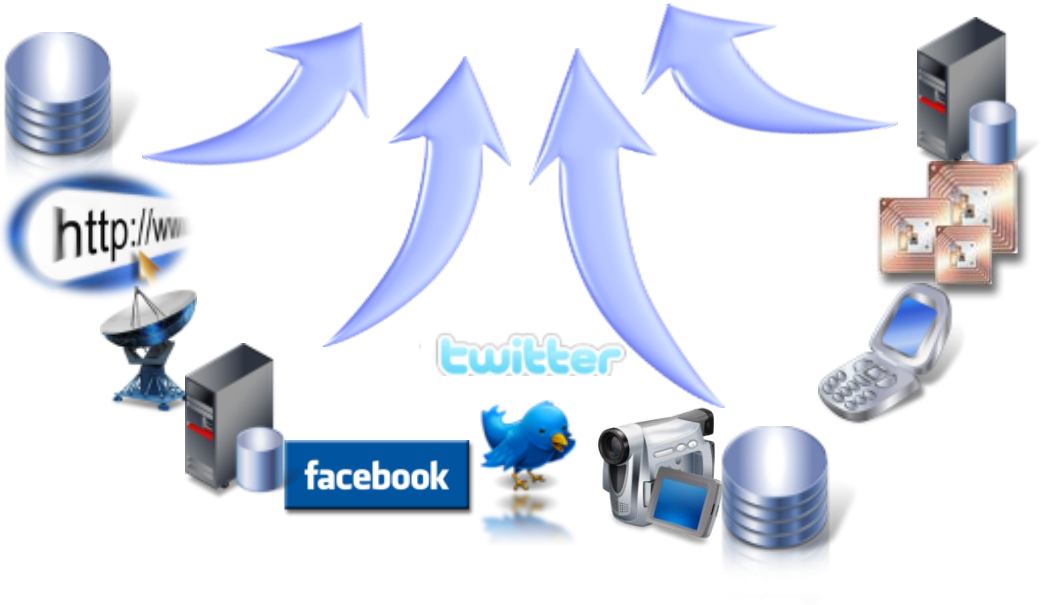
Machines Scale Better

OK – So What Is Big Data

Conventional Definition of “Big Data”

- Defined By Large Volumes
- Focused On Unstructured Data
- Valuable insight, but difficult to extract
- Basically an ETL environment

This definition is <u>wrong</u>



Functional Big Data Definition

Dealing with **information management** challenges **that don't natively fit with traditional approaches to handling the problem**

Functional Big Data Definition, Part 2

The technologies that deal with these problems are **broad and diverse, it is <u>not</u> just Hadoop**

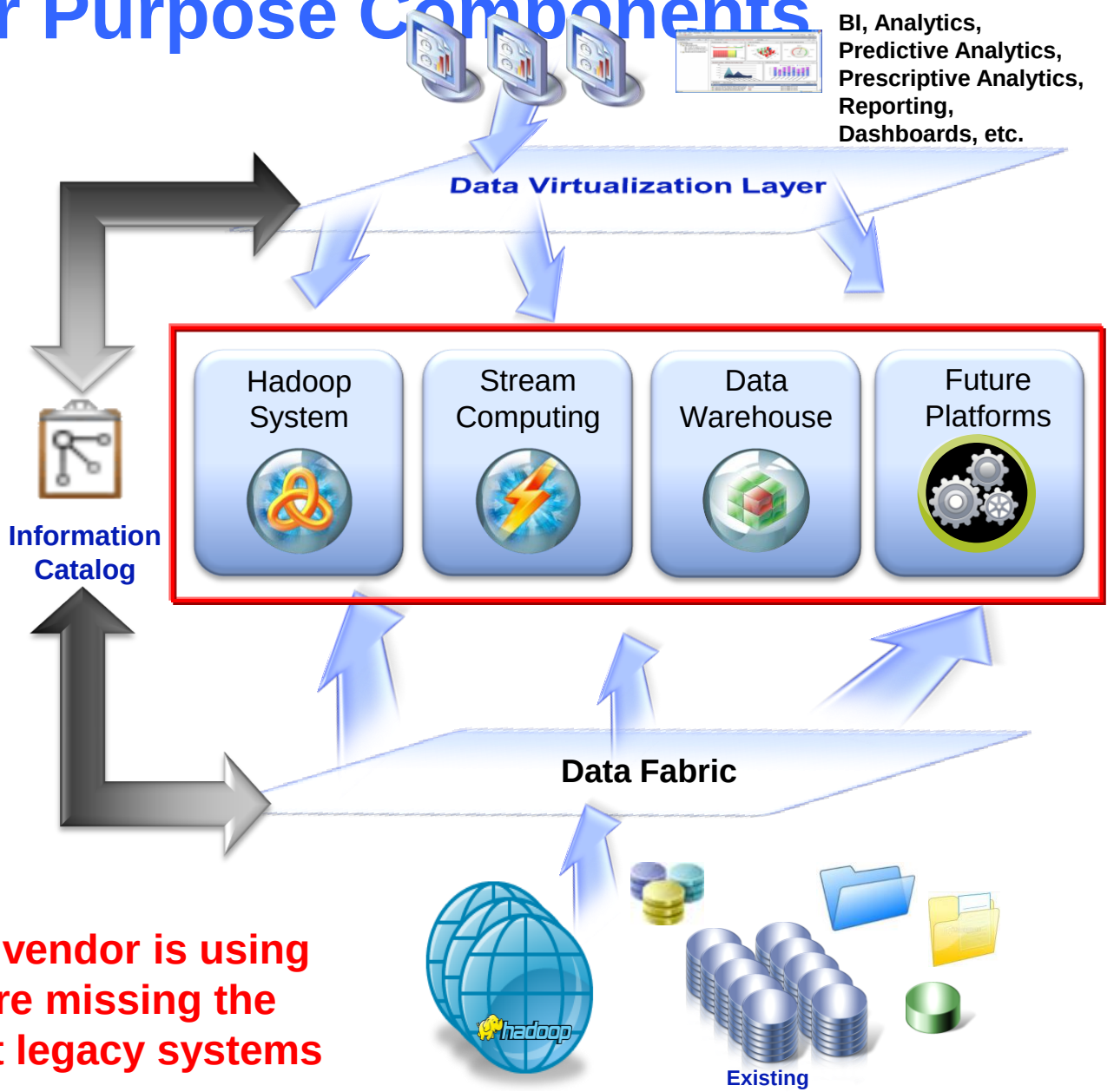
Giving Rise to “Fit For Purpose Architectures”

The introduction of purpose-built technology focused on a specific computing problem that is compelling better than existing technologies

Giving Rise to “Fit For Purpose Architectures”

Match the compute problem to the best way to handle it rather than assuming SQL by default

Fit For Purpose Components



Data Movement Platform– high speed data movement infrastructure with pre-built native connections to Data layers, Content Management, Big Data, and other types of data repositories.

Information Catalog – contains the metadata describing data domains and policies

Data Virtualization Layer– Sematic Layer that isolates applications from data location

If all you are doing as a vendor is using Hadoop for ELT you are missing the point or trying to protect legacy systems

…Which In-turn Enables Paradigm Shifts*

- The Idea Of the Super-set
- Combining Structured and Unstructured Analytics
- Move from Sampling to “Absolute Knowledge"
- Importance of Augmented Decision Making
- Streaming Computing Paradigm
- Lowering The Cost of Experimentation
- Changing Role Of Archiving

(* but the Laws of Gravity Still Apply!)

Examples Of What Using All The Data Can Do

Smart metering in **Malta** helps citizens pay only for the energy they use

Predictive analytics helped slash **Richmond's** crime rate by **40%** in one year

Data analytics helped cut crime **35%** in NYC

In **Delft**, developing enhanced flood prediction and protection systems for coastal areas and river deltas

In Taiwan, **99%** of smarter trains run on time

Miami-Dade County Public Schools have **increased academic achievement** across the board

In downtown **Stockholm** smart traffic systems helped reduce gridlock by **20%**

IBM helps **Amsterdam** Airport Schiphol move **20** million more bags every year with a smarter baggage system

Peak energy loads fell by **15%** when IBM helped homes in the Pacific Northwest talk straight to the grid

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

24

“Helps detect life threatening conditions up to 24 hours sooner”

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

Significant benefits:

- Helps detect life threatening conditions up to 24 hours sooner
- Lower morbidity and improved patient care

UOIT
CHALLENGE INNOVATE CONNECT

“Monitor and protect marine mammal life in real-time”

Marine Institute of Ireland Monitors Buoy Sensor Data to Detect Floods Sooner

Capabilities Utilized:

Stream Computing

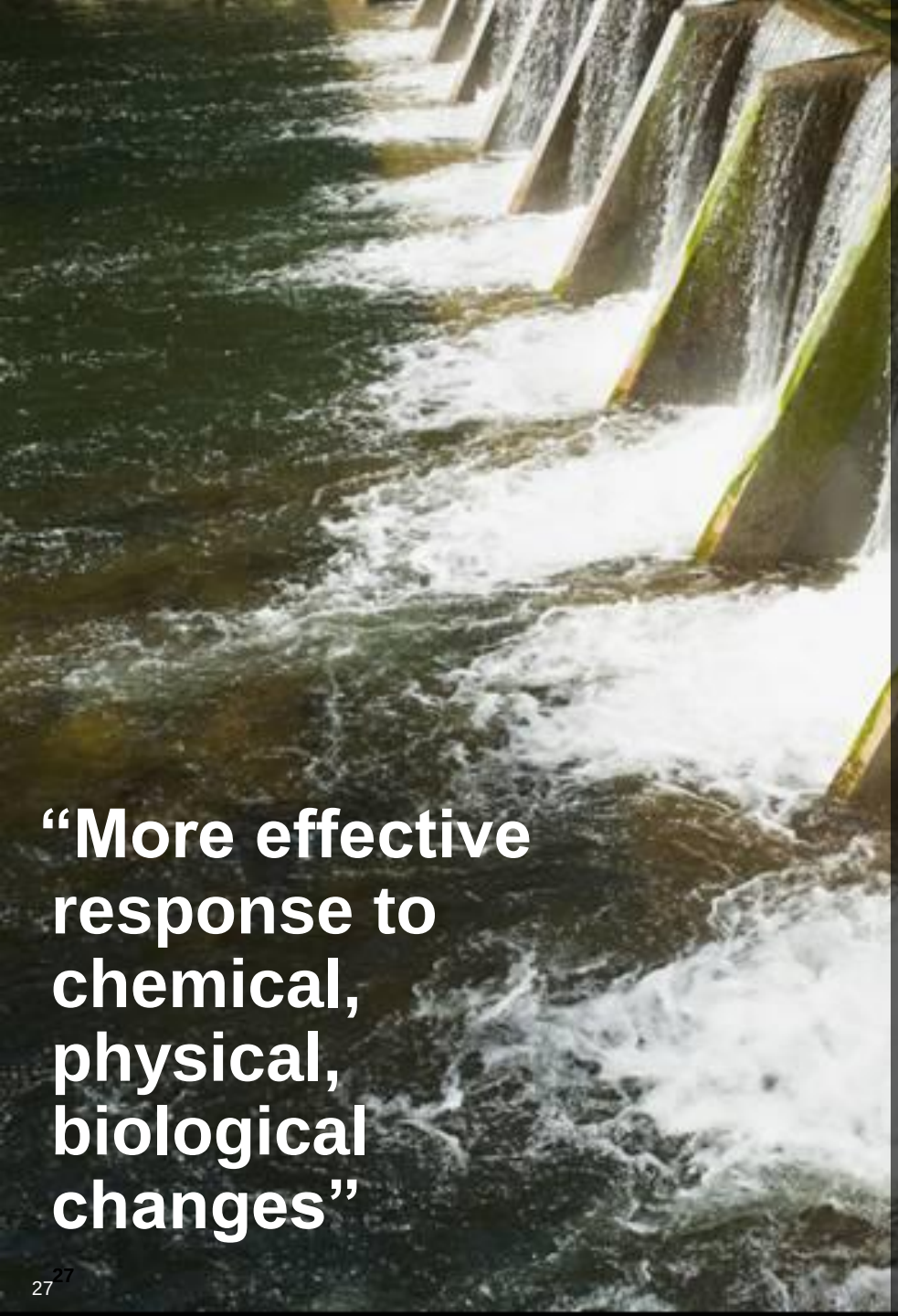
- Large amounts of sensor data is collected from buoys in local bay
- Continuously monitors environmental, pollution, and local marine life
- Information streamed to institutes’ central monitoring and analysis system
- Users can access, aggregate, analyze and set up automated alerts using web portal

Significant benefits:

- Prove adherence to regulations protecting marine mammals
- Faster and more accurate flood prediction
- Pollution and location-based debris tracking to increase public safety

Marine Institute
Foras na Mara

26



Beacon Institute Detects Environmental Changes to Water Resources in Real-time

Capabilities Utilized:

Stream Computing

- Collecting and processing multiple streams of physical, chemical, and biological data from sensors deployed in Hudson Bay
- Sensor data is analyzed against larger meteorological data and aggregated
- Real-time environmental data delivered in standard format to scientists, engineers, policy makers, and educators

Significant benefits:

- Better understanding of dynamic interactions within local river and estuaries
- Fosters increased collaboration by making real-world data available to outside systems, researchers, policy makers
- Helps resource management respond more effectively to changes to local water resources

“More effective response to chemical, physical, biological changes”

27

“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

Significant benefits:

- 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

28

KTH Swedish Royal Institute of Technology Reducing Traffic Congestion

Capabilities Utilized:

Stream Computing

- Deployed real-time Smarter Traffic system to predict and improve traffic flow.
- Analyzes streaming real-time data gathered from cameras at entry/exit to city, GPS data from taxis and trucks, and weather information.
- Predicts best time and method to travel such as when to leave to catch a flight at the airport

Significant benefits:

- Enables ability to analyze and predict traffic faster and more accurately than ever before
- Provides new insight into mechanisms that affect a complex traffic system
- Smarter, more efficient, and more environmentally friendly traffic

University of Maryland Baltimore Predicts Wildfire Dispersion to Save Lives

Capabilities Utilized:

Stream Computing

- Deployed wildfire analysis system to predict wildfire and smoke dispersion
- Analyzing large volumes of real-time streaming video data from surface, aerial, satellites sensors, and unmanned drones
- Processed video data used with air-concentration/dispersion model to predict fire/smoke location, direction, and dispersion

Significant benefits:

- Faster processing of massive wildfire data for more accurate predictive modeling
- Ability to provide real-time public safety and health alerting
- Saves lives and reduces property damage

UMBC

University of Southern California Innovation Lab Monitors Political Debates

Solution to measure public sentiment during key primary & general presidential debates

- Examines trends, volume, and content of **millions** of public Twitter messages in real-time
- Analytic accelerators to understand sentiment (positive, negative, neutral)
- Capabilities
  - Stream Computing
  - 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

Annenberg Innovation Lab

Pacific Northwest Smart Grid Demonstration Project

Capabilities:

- Stream Computing – real-time control system
- Data Warehouse 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.

Battelle
The Business of Innovation

University of Southern California Oscar Senti-meter

Solution to measure public sentiment about the Academy Awards race

- Developed by the LA Times, IBM, and the IBM Annenberg Innovation Lab
- Detects public opinion for Hollywood's awards season
- Examines trends, volume, and content of **millions** of public Twitter messages
- Analytic accelerators to understand sentiment (positive, negative, neutral)

Capabilities

- Stream Computing
- Hadoop System
- Visualization
- Consumer Sentiment Analysis application