Big Data

Ayhan Önder
Client Technical Professional, Netezza

20.09.2012

The World of DATA

THE WORLD OF DATA

| Number of Emails Sent per day | House Hold data usage per day | Video Upload per min to YouTube | Data Per Day Handled @Google | | Tweets sent per day | Total Facebook Min Per Month | Mobile Internet Data Per second | Amazon Orders every second |

2.9 MILLION — **375** MEGABYTES — **20** HOURS — **24** PETABYTES — **50** MILLION — **700** BILLION — **1.3** EXABYTES — **72.9** ITEMS

IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.

A COLLABORATION BETWEEN GOOD AND OLIVER MUNDAY

IN PARTNERSHIP WITH IBM

"Data is the new Oil"

In its raw form, oil has little value. Once processed and refined, it helps power the world.

Forbes.com

"Big Data has arrived at Seton Health Care Family, fortunately accompanied by an analytics tool that will help deal with the complexity of more than two million patient contacts a year…"

The New York Times

"At the World Economic Forum last month in Davos, Switzerland, Big Data was a marquee topic. A report by the forum, "Big Data, Big Impact," declared data a new class of economic asset, like currency or gold.

FT FINANCIAL TIMES
World business newspaper

"Increasingly, businesses are applying analytics to social media such as Facebook and Twitter, as well as to product review websites, to try to "understand where customers are, what makes them tick and what they want", says Deepak Advani, who heads IBM's predictive analytics group."

THE WALL STREET JOURNAL

"Companies are being inundated with data—from information on customer-buying habits to supply-chain efficiency. But many managers struggle to make sense of the numbers."

Forbes.com

"…now Watson is being put to work digesting millions of pages of research, incorporating the best clinical practices and monitoring the outcomes to assist physicians in treating cancer patients."

CNBC

*"Data is the new oil."*
Clive Humby

Los Angeles Times

The Oscar Senti-meter — a tool developed by the L.A. Times, IBM and the USC Annenberg Innovation Lab — analyzes opinions about the Academy Awards race shared in millions of public messages on Twitter."

The Big Data Opportunity

Extracting insight from an immense volume, variety and velocity of data, in context, beyond what was previously possible.
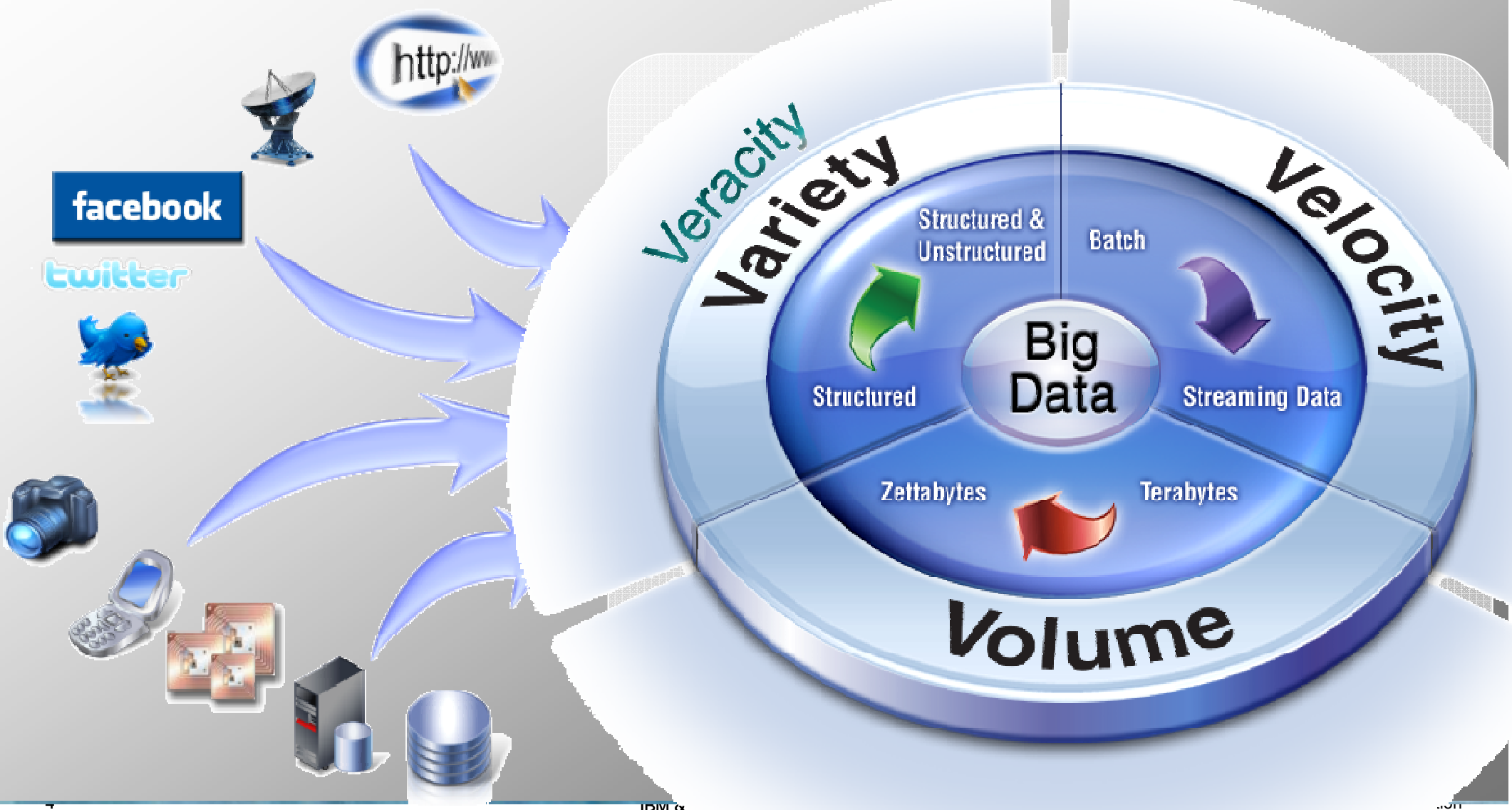
What Our Customers Tell Us . . .

- Don't know what should be analyzed

- Volumes can be extremely high
  - Potentially valuable data is dormant or discarded (size/performance)
  - Too expensive to justify integrating large volumes of unstructured data

- Much of their data is unstructured, or in widely varying structures, which are difficult to analyze

- Difficult to integrate information distributed across multiple systems and the Internet

- Some information has a short useful lifespan

- Analysis needed in the context of their existing information

Information Sources

Users

Point of Sale

Call Center

Customer

Structured Information

Traditional DW

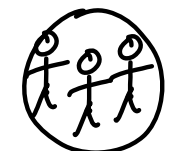Reporting, Trend Analysis, Decision Making

At Rest Analytics

Executive

Data Integration

Power User

Untapped Sources

V³

Big Data Platform

Variety
Structured & Unstructured

✓ Ability to handle V³

Ease of Use for End Users

Call Center

Velocity
Speed of data
In motion
analytics

✓ Enable Analytics
In motion & At Rest
Text, Video, Audio…

Easy Visualization Tools

✓ Ease of Use for Developers

Customers

Volume
Petabytes,
Zetabytes

✓ Enterprise Class
GPFS, Security,
Governance,
Data Integration

Better Business Outcomes

ERASER

Traditional Analytics:
Business Requirements Drive Solution Design

Business Defines Requirements – What Questions Should we ask?

IT Designs a Solution with a set structure and functionality

New requirements require redesign and rebuild

Business executes queries to answer questions over and over
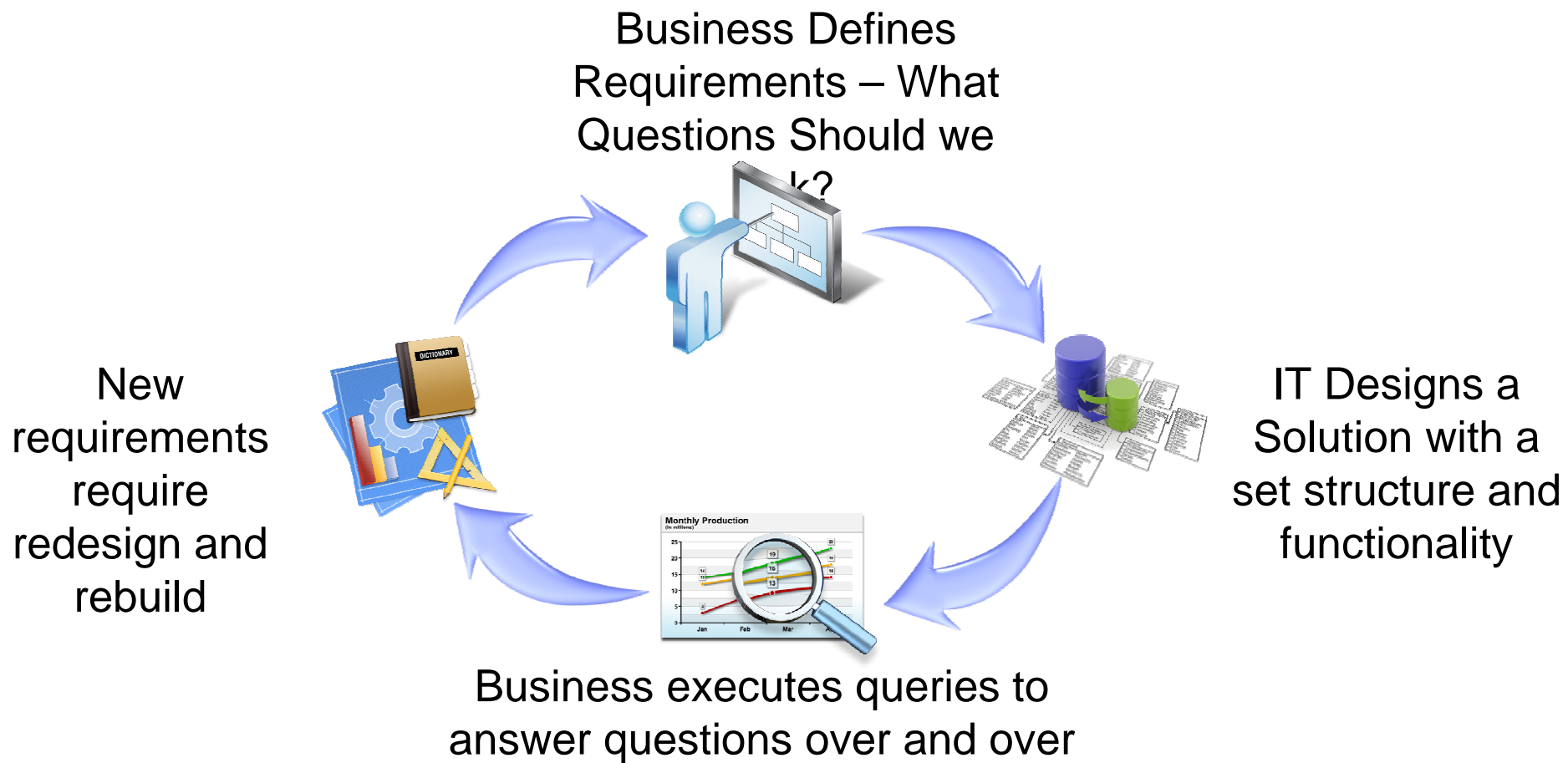
Big Data Analytics:
Information Sources Drive Creative Discovery

Business and IT Identify
Information Sources
Available

IT Delivers a
Platform that
enables creative
exploration of all
available data
and content

New insights
drive
integration to
traditional
technology

Business determines what
questions to ask by exploring
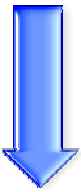the data and relationships
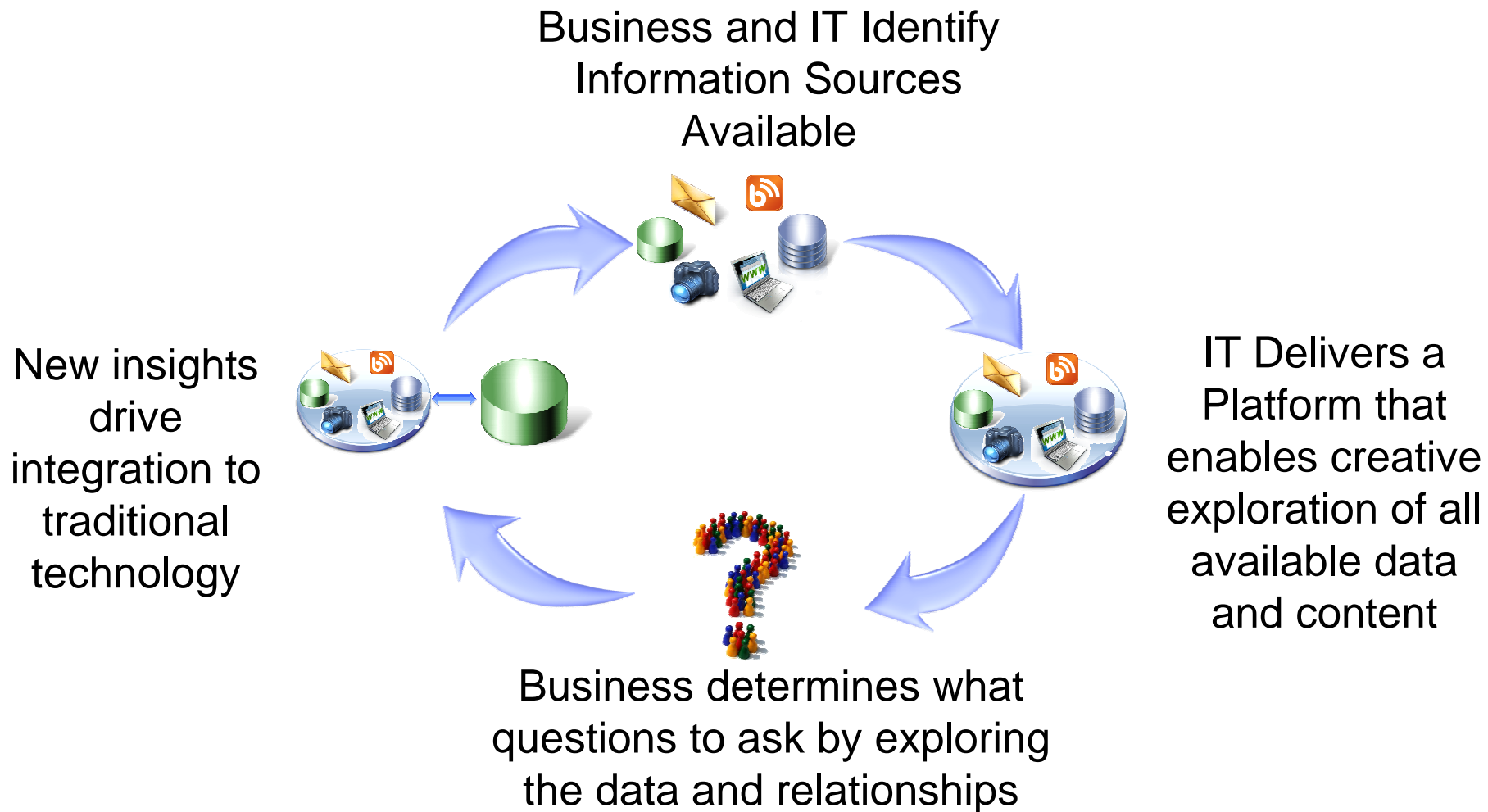
Merging the Traditional and Big Data Approaches

Traditional Approach
Structured & Repeatable Analysis

Big Data Approach
Iterative & Exploratory Analysis

Business Users
Determine what
question to ask

IT
Delivers a platform to
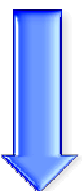enable creative
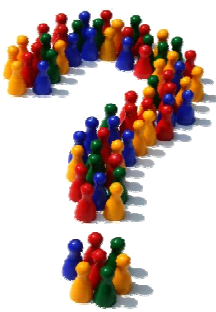discovery

IT
Structures the
data to answer
that question

Business
Explores what
questions could be
asked

Monthly sales reports
Profitability analysis
Customer surveys

Brand sentiment
Product strategy
Maximum asset utilization

Achieve Breakthrough Outcomes with Big Data capabilities

Achieve Breakthrough Outcomes	With Unique Capabilities	To Analyze Any Big Data Type
Know Everything about your Customers	Visualization & Discovery	Transactional / Application Data
Run Zero-latency Operations	Hadoop	Machine Data
Innovate new products at Speed and Scale	Data warehousing	Social Media Data
Instant Awareness of Fraud and Risk	Stream Computing	Content
Exploit Instrumented Assets	Text Analytics	
	Integration & governance	

Most Use Cases Combine Multiple Technologies
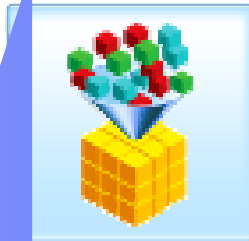
Pre-processing

Ingest and analyze unstructured data types and convert to structured data

Combine structured and unstructured analysis

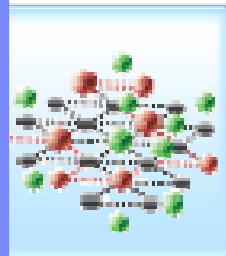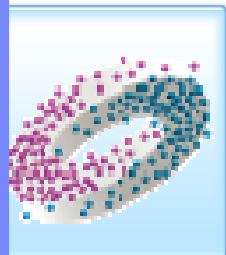Augment data warehouse with additional external sources, such as social media

Combine high velocity and historical analysis

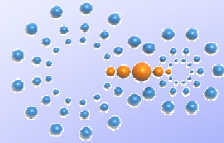Analyze and react to data in motion; adjust models with deep historical analysis

Reuse structured data for exploratory analysis

Experimentation and ad-hoc analysis with structured data

IBM & Turkcell Confidential

Opportunities To Exploit Big Data Are Everywhere

Analyze Information in Motion

- Smart Grid management
- Multimodal surveillance
- Real-time promotions
- Cyber security
- ICU monitoring
- Options trading
- Click-stream analysis
- CDR processing
- IT log analysis
- RFID tracking & analysis

Analyze Extreme Volumes of Information

- Transaction analysis to create insight-based product/service offerings
- Fraud modeling & detection
- Risk modeling & management
- Social media/sentiment analysis
- Environmental analysis

Analyze a Variety of Information

- Social media/sentiment analysis
- Geospatial analysis
- Brand strategy
- Scientific research
- Epidemic early warning system
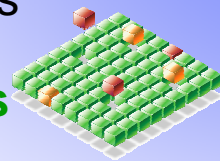- Market analysis
- Video analysis
- Audio analysis

Discovery & Experimentation

- Sentiment analysis
- Brand strategy
- Scientific research
- Ad-hoc analysis
- Model development
- Hypothesis testing
- Transaction analysis to create insight-based product/service offerings

Manage and Plan

- Operational analytics – BI reporting
- Planning and forecasting analysis
- Predictive analysis

12

Big Data: Integrated in Enterprise Information Architecture

Data Warehouse

Big Data Platform

- **Big Data is Here to Stay**

- **It Must Not Be a Silo**
  - It must be fully integrated for it to deliver value

- **It Must be Easy to Deploy and Integrate**

Enterprise Integration

Traditional Sources

New Sources
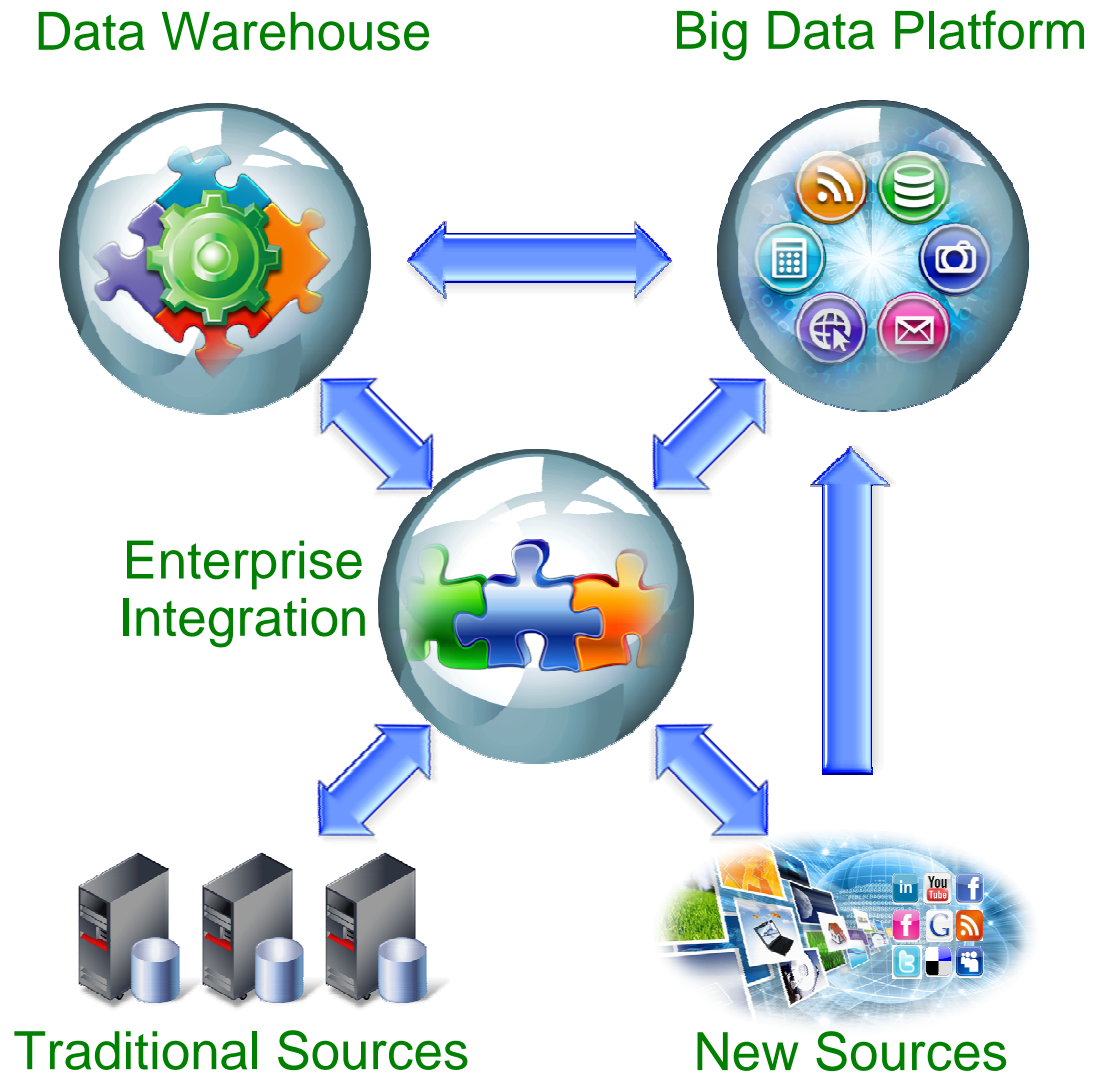
Business-centric Big Data enables you to start with a critical business pain and expand the foundation for future requirements
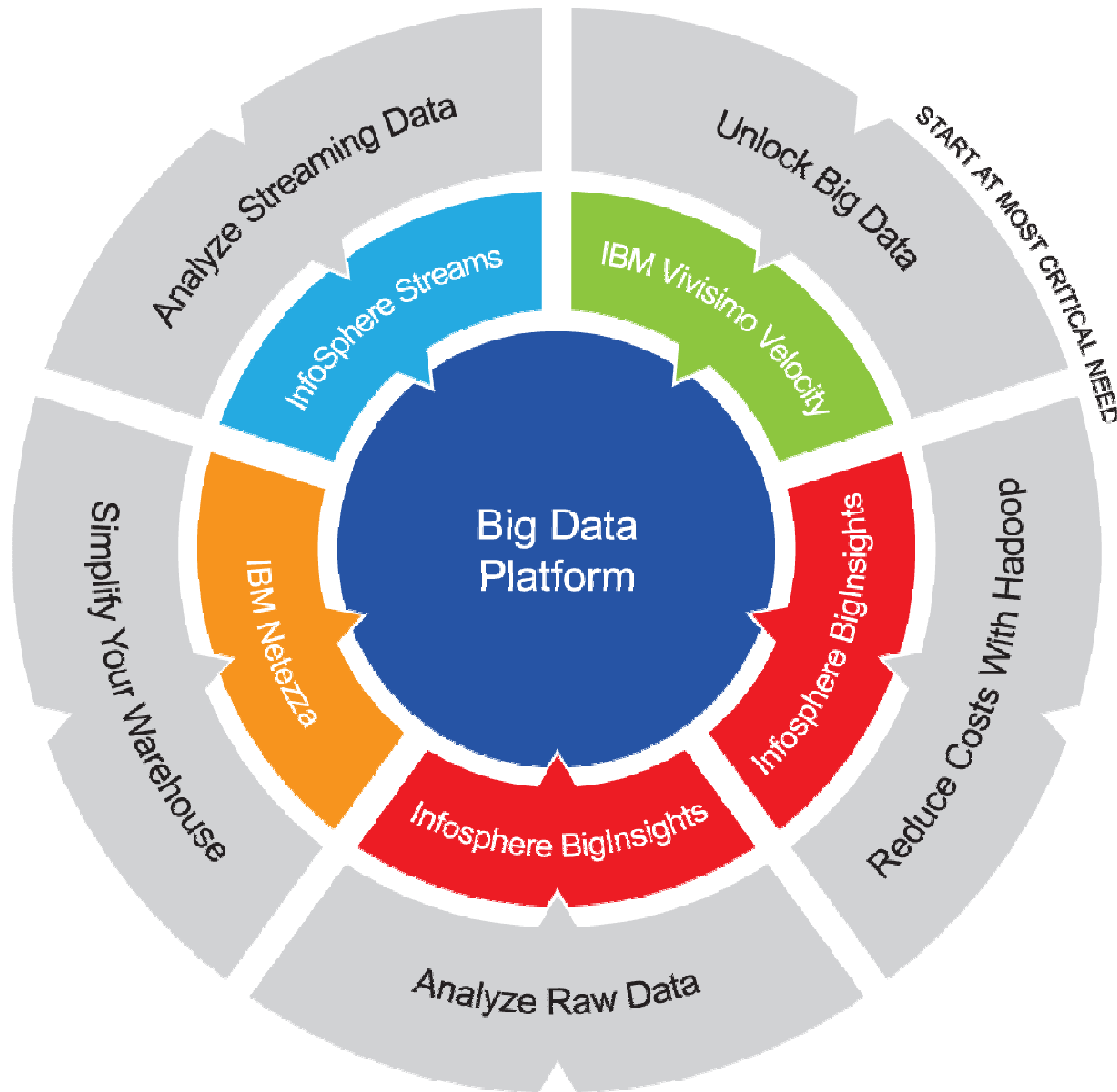
- "Big data" isn't just a technology—it's a business strategy for capitalizing on information resources

- Getting started is crucial

- Success at each entry point is accelerated by products within the big data platform

- Build the foundation for future requirements by expanding further into the big data platform

Entry Points are Accelerated by Products Within the Big Data Platform

1 – Unlock Big Data

IBM Vivisimo

2 – Analyze Raw Data

InfoSphere BigInsights

4 – Reduce costs with Hadoop

InfoSphere BigInsights

Analytic Applications

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

3 – Simplify your warehouse

IBM Warehouse Solutions

5 – Analyze Streaming Data

InfoSphere Streams

IBM

days for a single query

constant tuning

" **Nearly 70%** of data warehouses experience performance-constrained issues of various types. "

- Gartner 2010 Magic Quadrant

specialized resources required

months to deploy
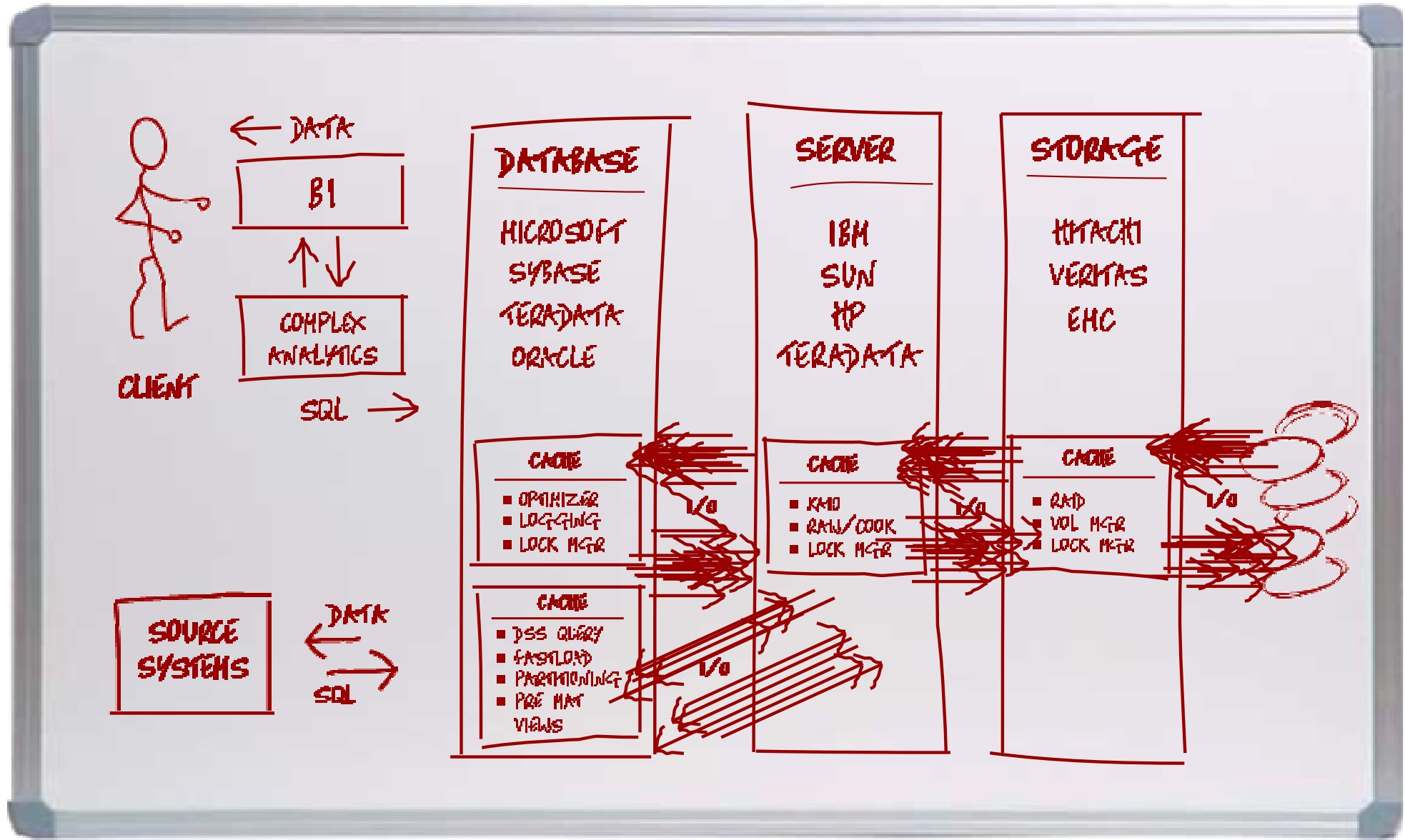
Highway or off-road? Building a machine for its workload
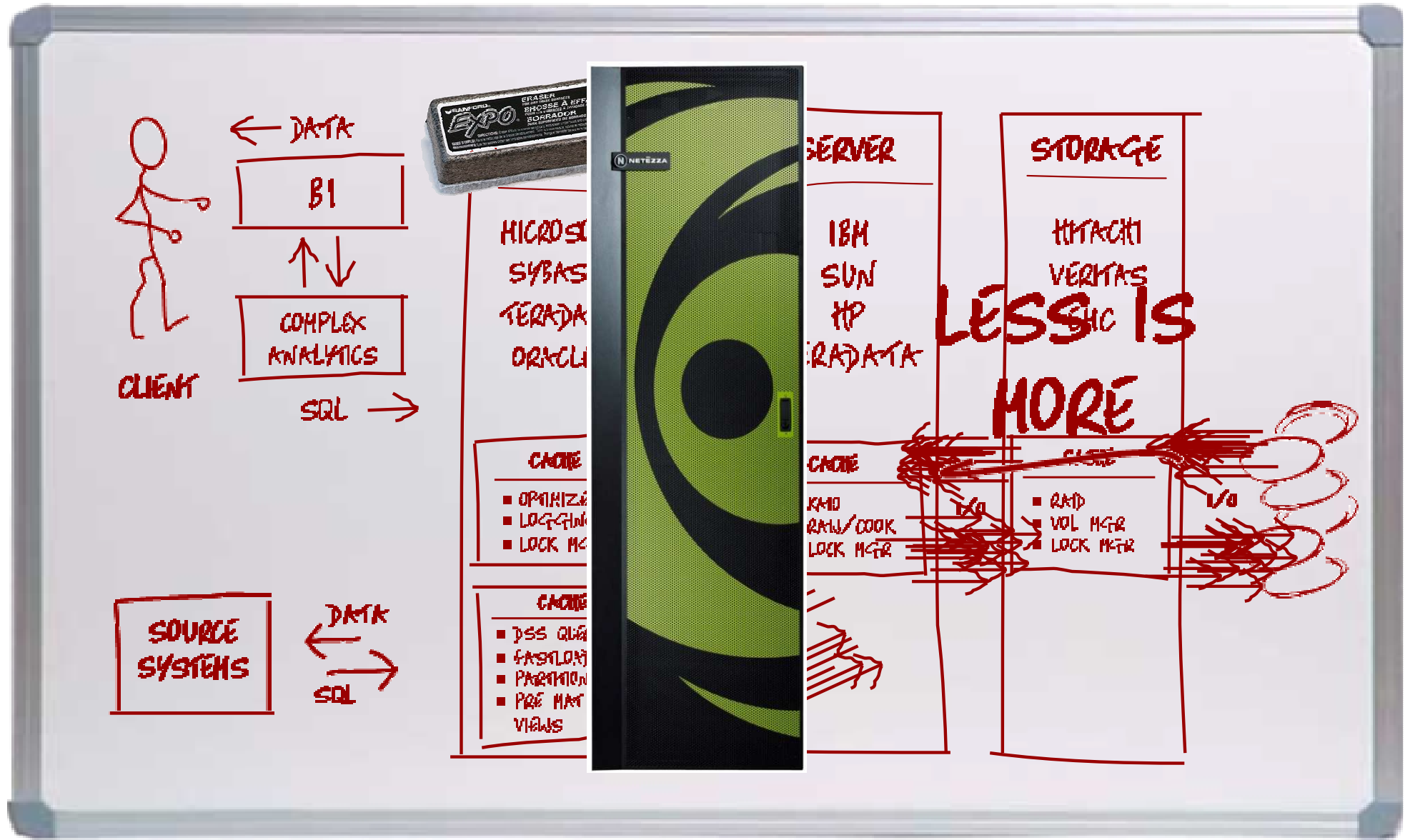
Transactional workloads vs. analytic workloads
Two VERY different requirements for storing and processing data

Customer **Business Transaction** **Transactional Database**

Transaction

Simple Query

```
Item:'Shoes'
Cost:'$34'
Cust:'James'
```

2011
Sales

Item	Cost	Cust
Shoes	$34	James

Business Analyst **Data Warehouse** **BI Reports & Dashboards**

Complex Query

```
Sales & Profit for
Shoes & Belts
Year >= 2005
```

SALES

2010
2009
2008
2007
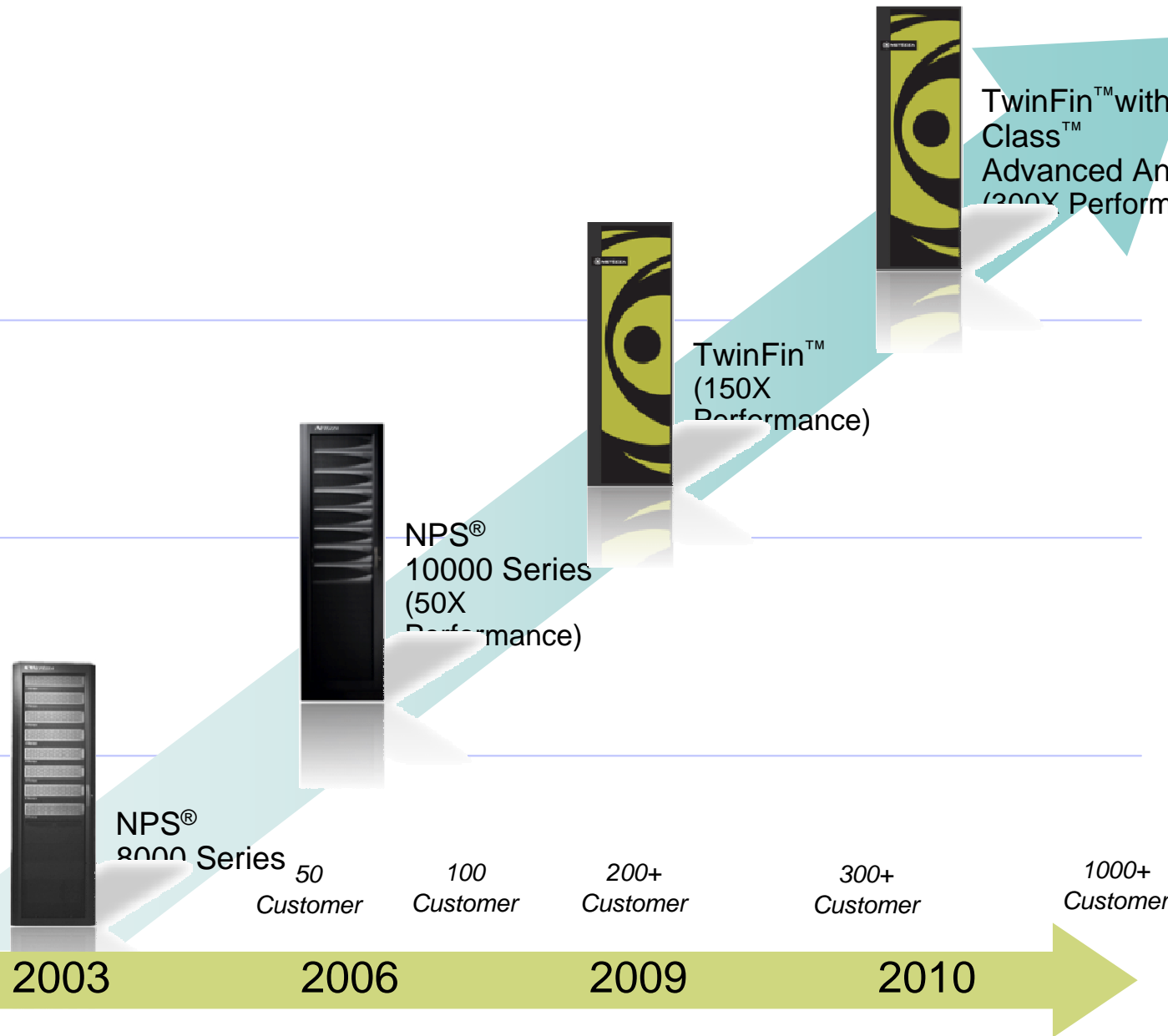2006
2005

Traditional Data Warehouse Approach

Netezza – Simplicity

Netezza

First Analytical
Appliance in the
world

First Petabyte Size
Data Warehouse

First 100 TB Size
Data Warehouse

Impact

TwinFin™ with i-
Class™
Advanced Analytics
(300X Performance)

TwinFin™
(150X
Performance)

NPS®
10000 Series
(50X
Performance)

NPS®
8000 Series

2003	2006	2009	2010	
	50 Customer	100 Customer	200+ Customer	300+ Customer	1000+ Customer

" *Netezza was part of the inspiration for Exadata …*

We'd like to thank them for forcing our hand and forcing us to go into the hardware business. "

-- Larry Ellison, CEO
Oracle Corporation
January 27 2010

http://oracle.com.edgesuite.net/ivt/4000/8104/9238/12652/lobby_external_flash_clean_480x360/default.htm

IBM Netezza
The Simple Appliance for Serious 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

Inside IBM Netezza
data warehouse appliances

Optimized Hardware + Software

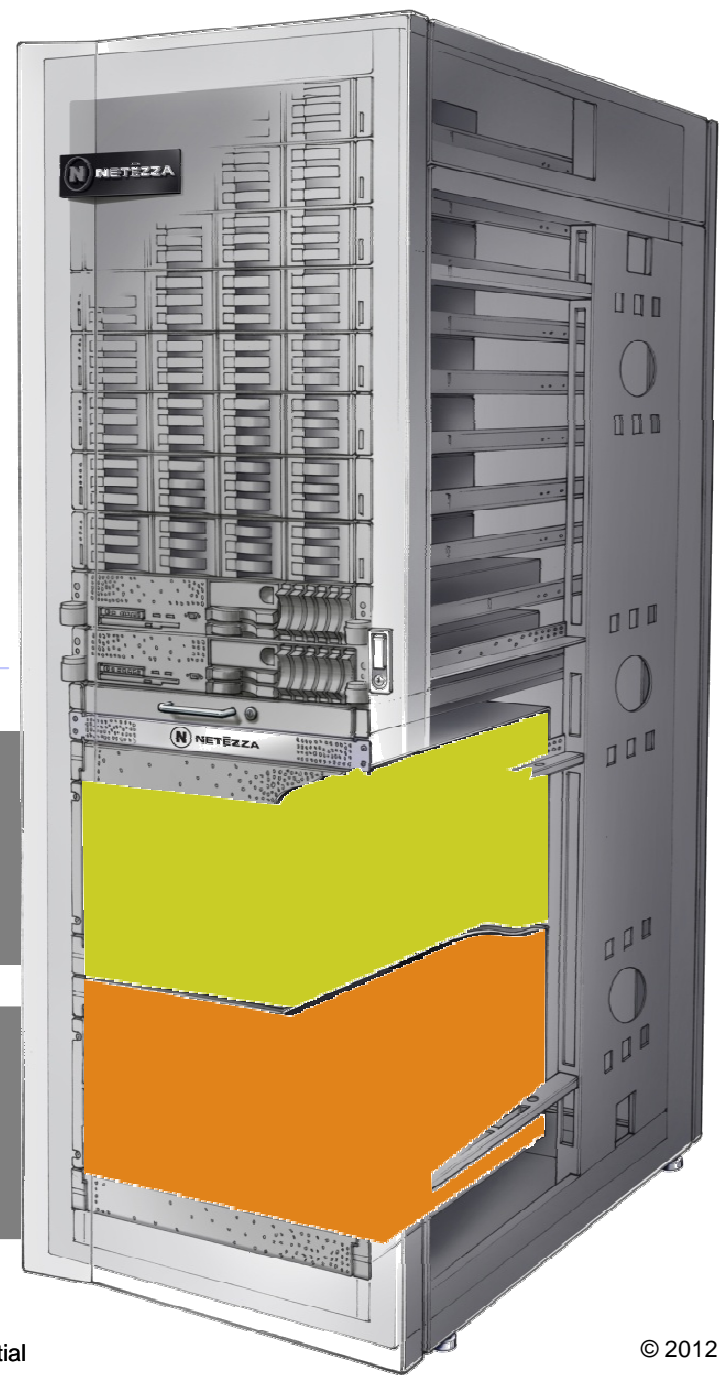Purpose-built for high performance analytics; requires no tuning

True MPP

All processors fully utilized for maximum speed and efficiency

Streaming Data

Hardware-based query acceleration   for blistering fast results

Deep Analytics

Complex analytics executed in-database for deeper insights

Netezza delivers simplicity

- Up and running 6 months before being trained

- 200X faster than Oracle system

- ROI in less than 3 months

" Allowing the business users access to the Netezza box was what sold it.

-- Steve Taff,
Executive Dir. of IT Services

XO. Communications

IBM Netezza Appliance family

IBM Netezza 100	IBM Netezza 1000	IBM Netezza High Capacity Appliance
Development & Test System	Data Warehouse High-Performance Analytics	Queryable Archiving Backup/DR
1 TB to 10 TB	1 TB to 1.5 PB	100 TB to 10 PB

IBM & Turkcell Confidential
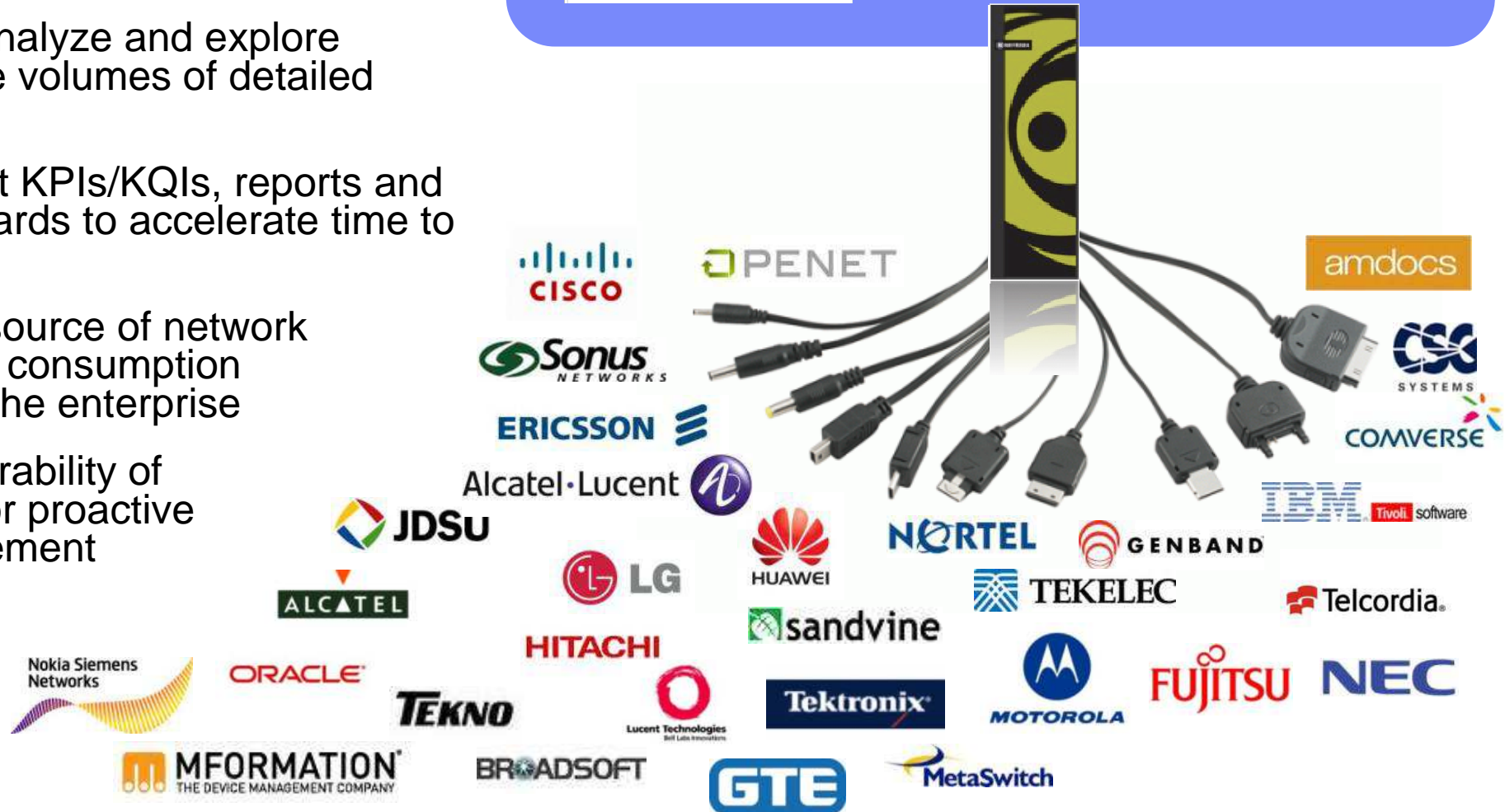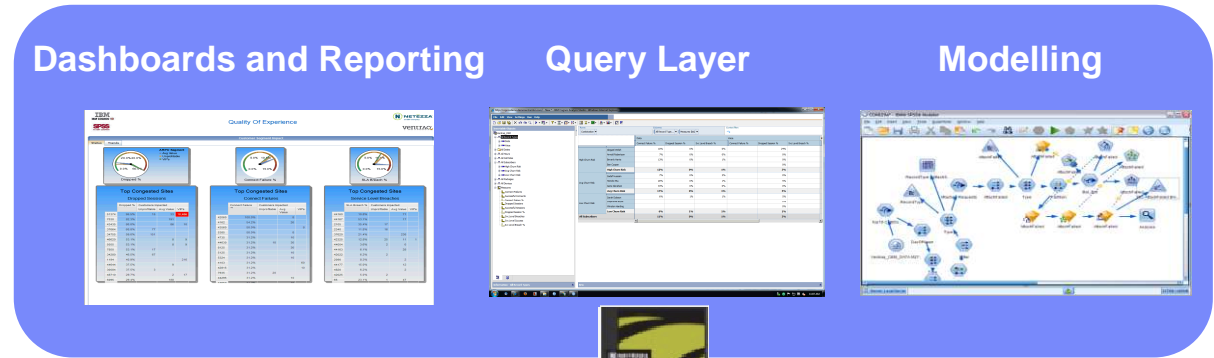
Network Analytics with Netezza
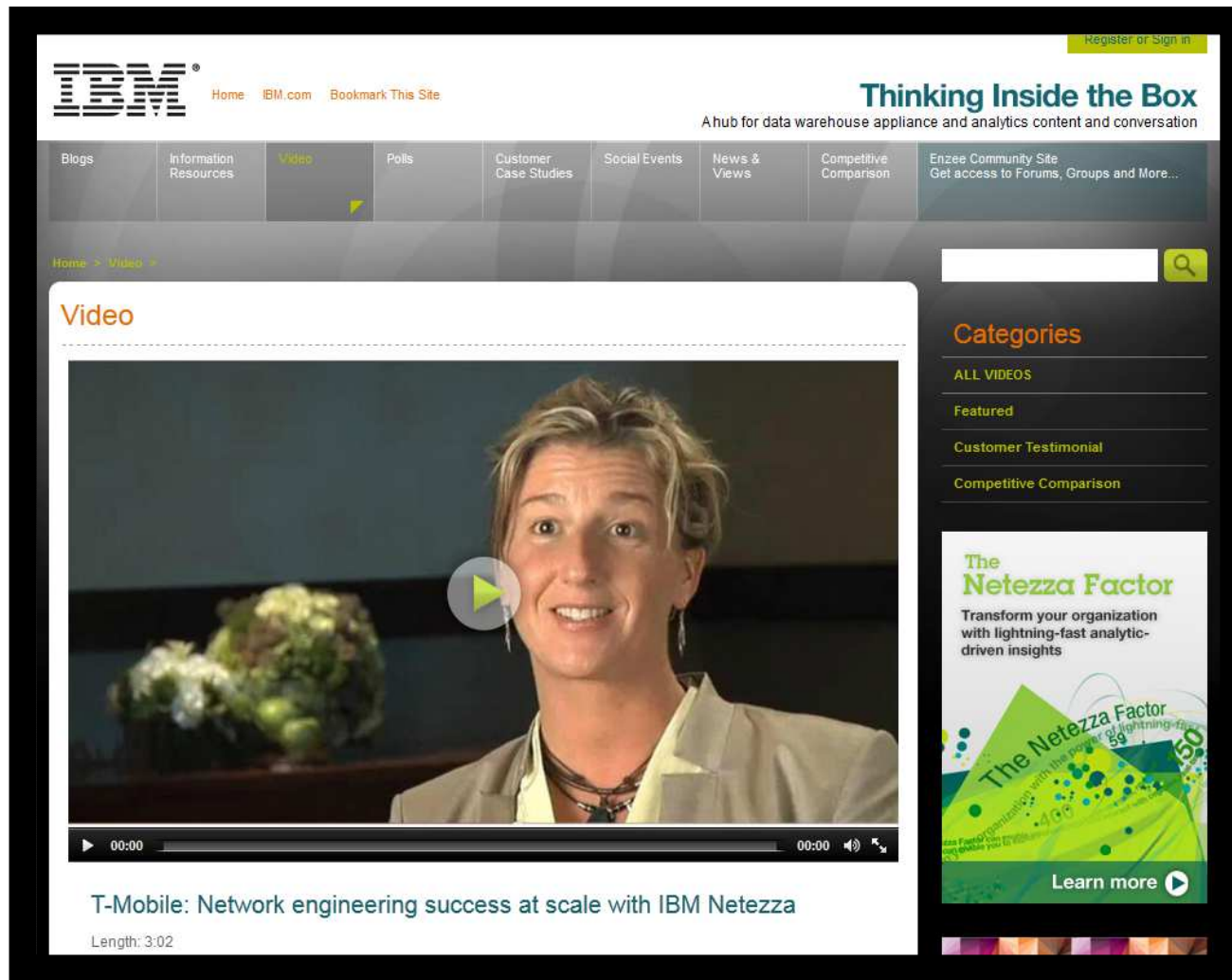## *Single view of Network, Customers, Services and Devices*
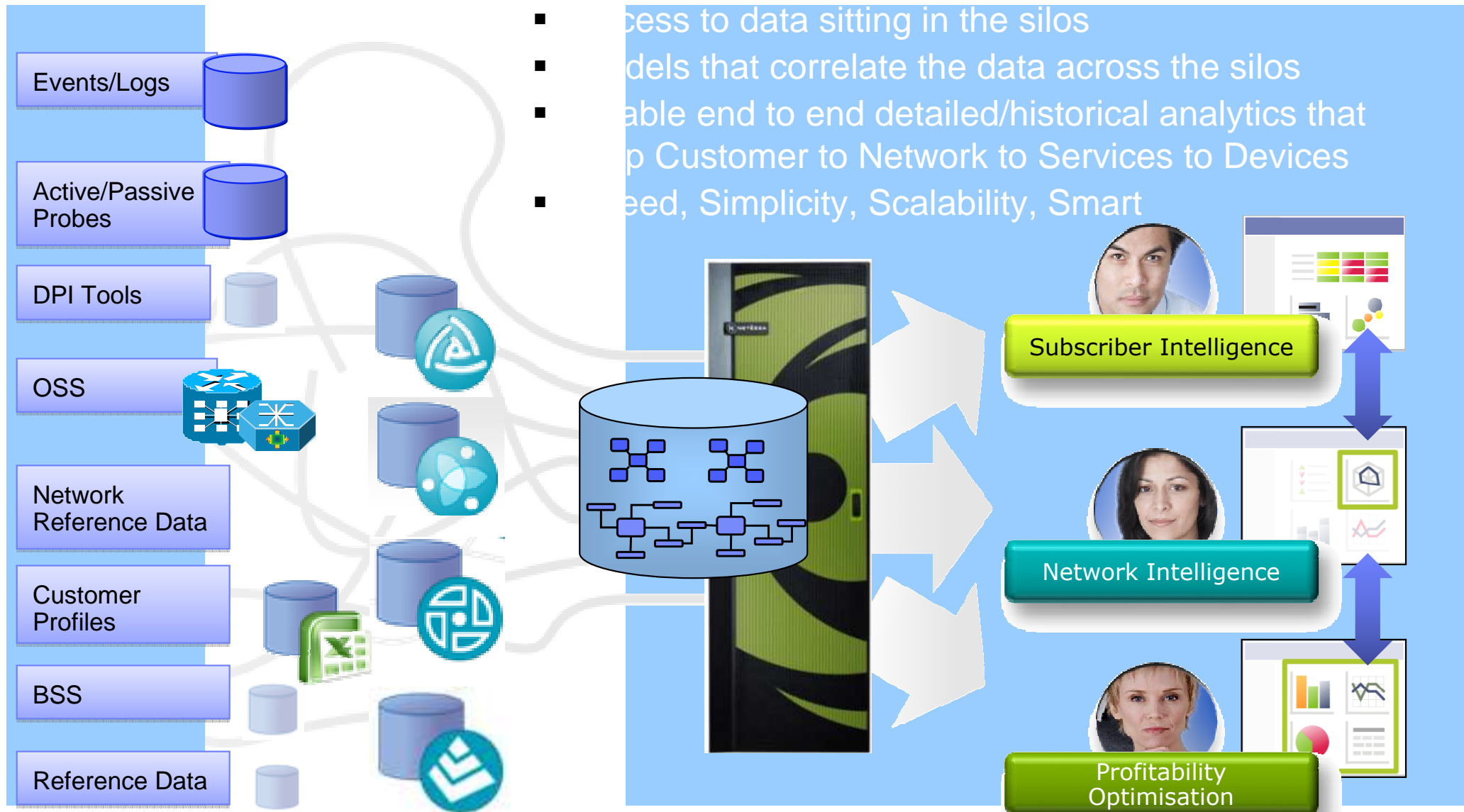
- ...cess to data sitting in the silos
- ...dels that correlate the data across the silos
- ...able end to end detailed/historical analytics that
  ...p Customer to Network to Services to Devices
- ...eed, Simplicity, Scalability, Smart

Events/Logs

Active/Passive Probes

DPI Tools

OSS

Network Reference Data

Customer Profiles

BSS

Reference Data

Subscriber Intelligence

Network Intelligence

Profitability Optimisation

Out of the box solutions for Network Analytics

- Data acquired directly from network elements or systems, OSS/BSS applications or other data stores via landing zone

- Load, analyze and explore massive volumes of detailed events

- Pre-built KPIs/KQIs, reports and dashboards to accelerate time to value

- Single source of network data for consumption across the enterprise

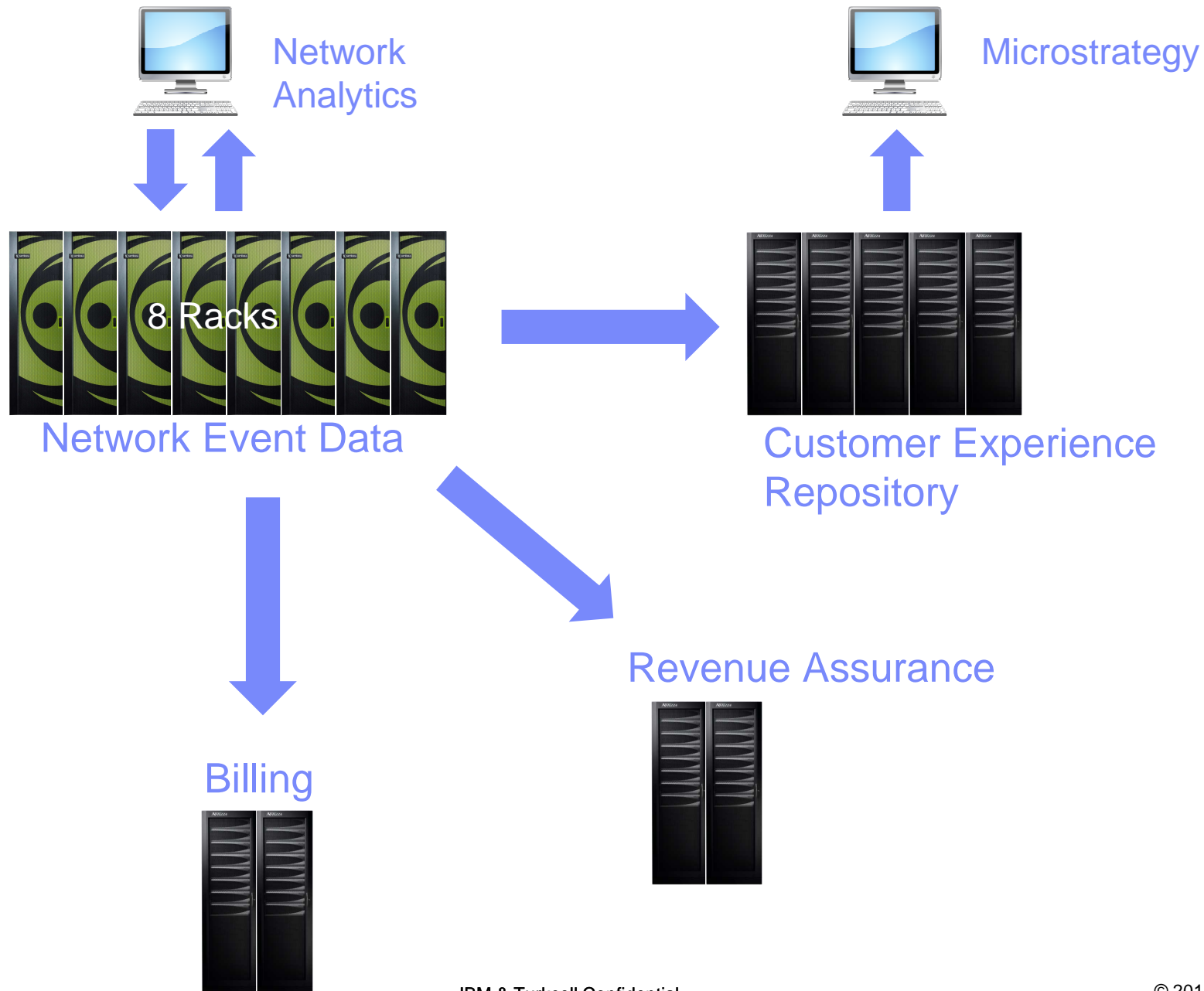- Configurability of alerts for proactive management

Dashboards and Reporting   **Query Layer**   **Modelling**

Results – In T-Mobile's (USA) own words

http://thinking.netezza.com/video/t-mobile-network-engineering-success-scale-ibm-netezza

T-Mobile USA

Network Analytics

Microstrategy

8 Racks

Network Event Data

Customer Experience Repository

Revenue Assurance

Billing

IBM & Turkcell Confidential
© 2012 IBM Corporation

Turkcell, Big Data PoC
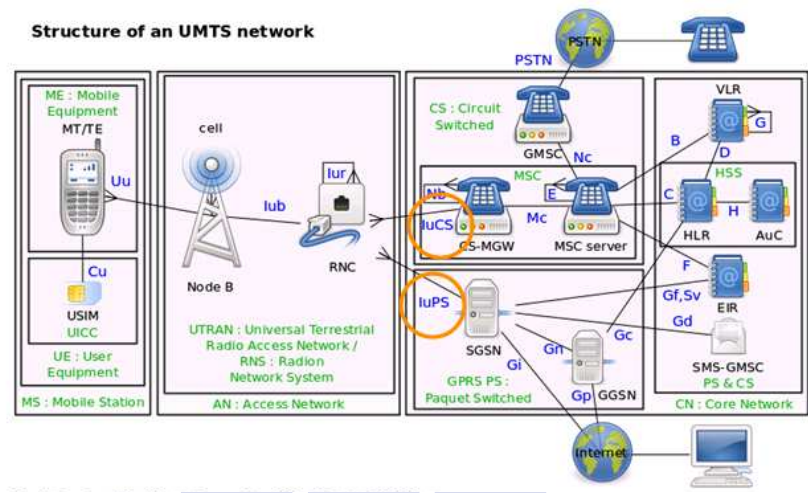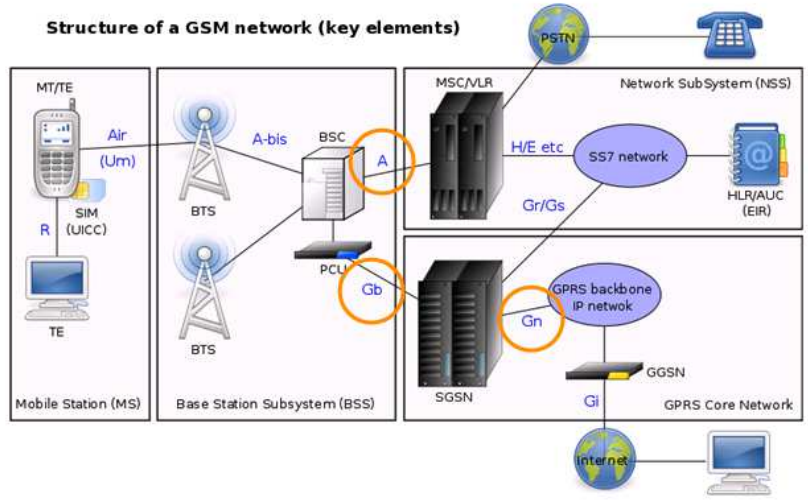
ADR — Single file in each 5 minutes Avg.1.3 billion event per day

GB — Single file in each 5 minutes Avg.440 million event per day

GN — Multiple files continuously Avg.5.2 billion event per day

IuCS — Single file in each 5 minutes Avg.440 million event per day

IuPS — Single file in each 5 minutes Avg.625 million event per day

NAT   SGSN   MAP

Cognos, Data & Voice Reports
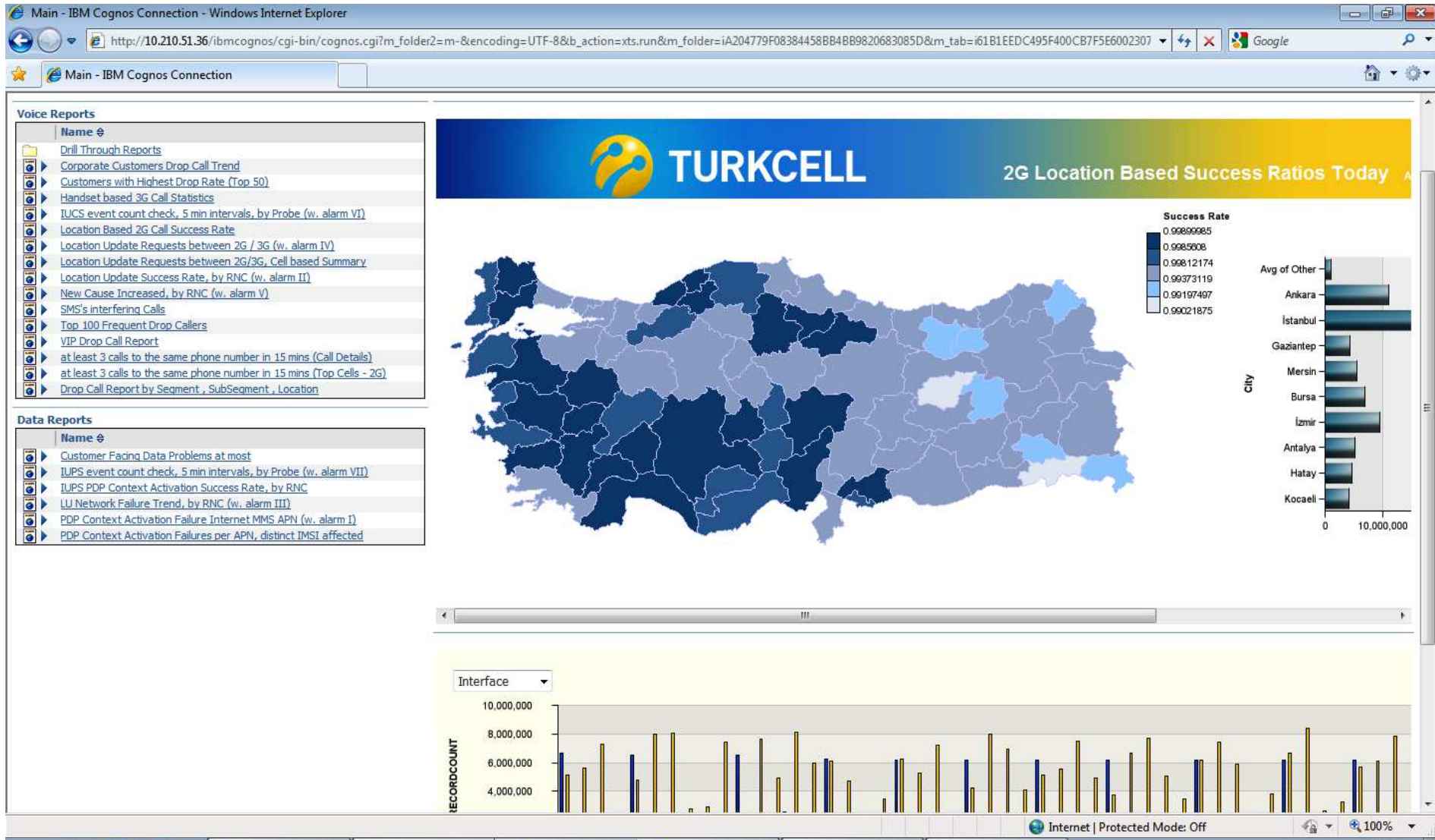
Customers with Highest Drop Rate (Top 50)

Customers facing Highest Droprate (Top50) for the given date. Produced report includes 2G / 3G filtering, as well as graphs allowing drop ratio and drop count comparison for each IMSI.

IBM & Turkcell Confidential

Query Performance

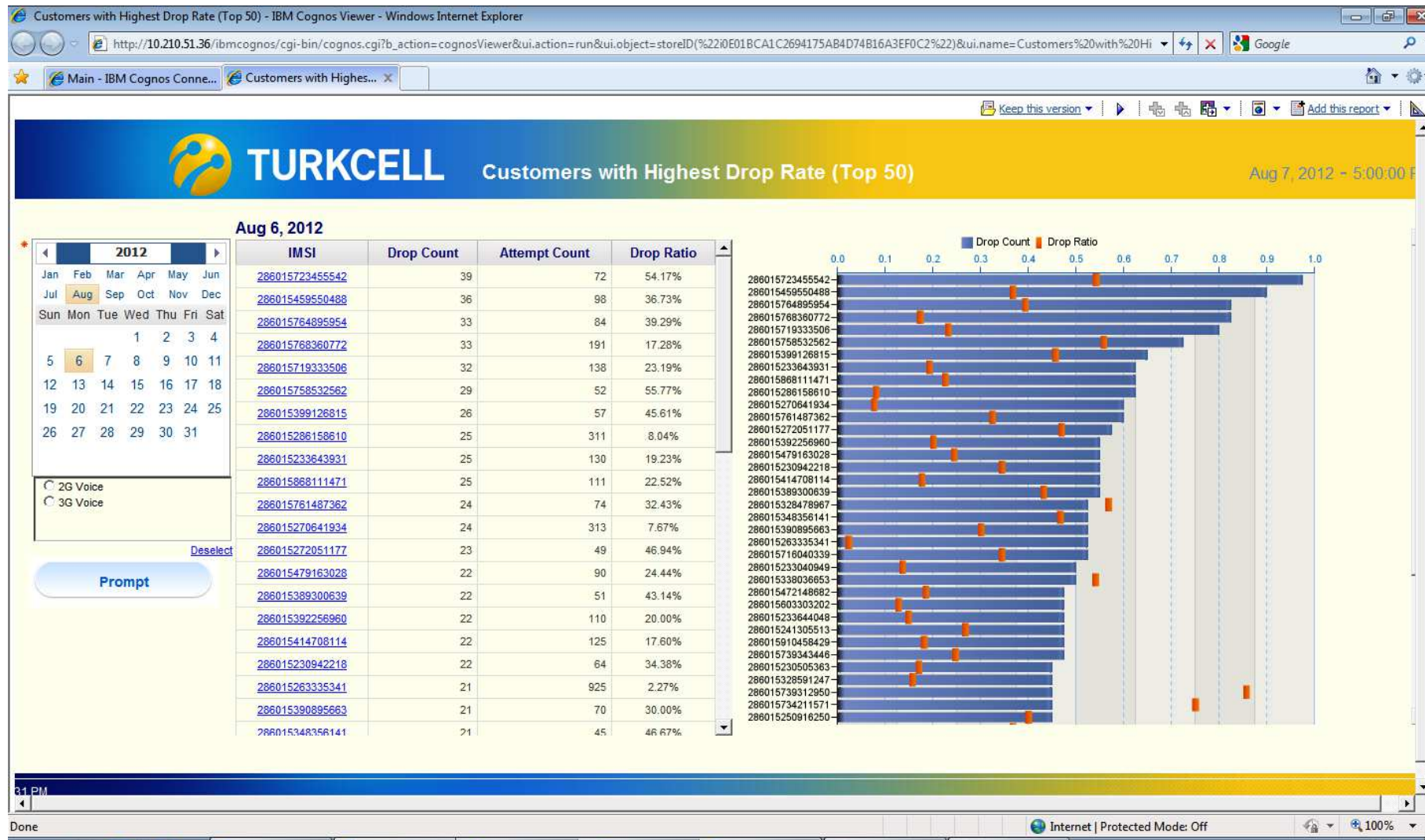|	Observed Duration in Cognos (includes report rendering / download time, while data loads are going on)
SMS' Interferrring Calls	< 15 sec.
PDP Context Activation Failures per APN, distinct IMSI affected	< 5 sec.
PDP Context Activation Failure Internet MMS APN	< 10 sec.
New Cause Increased, by RNC	< 3 sec.
LU Network Failure Trend, by RNC	< 3 sec.
Location Update Success Rate, by RNC	< 4 sec.
Location Update Requests between 2G / 3G, Cell based Summary	< 4 sec.
Location Update Requests between 2G / 3G	< 4 sec.
Location Based 2G Call Success Rate	< 4 sec.
IUPS PDP Context Activation Success Rate, by RNC	< 3 sec.
IUPS event count check, 5 min intervals, by Probe	< 3 sec.
IUCS event count check, 5 min intervals, by Probe	< 3 sec.
Handset Based 3G Call Statistics	< 6 sec. (for 3 days)
Drop Call Report by Segment , SubSegment , Location	< 10 sec.
Customers with Highest Drop Rate (Top 50)	< 10 sec.
Customer Facing Data Problems at most	< 5 sec.
Corporate Customers Drop Call Trend	< 5 sec.
at least 3 calls to the same phone number in 15 mins (Top Cells - 2G)	< 10 sec.
at least 3 calls to the same phone number in 15 mins (Call Details)	< 5 sec.

IBM & Turkcell Confidential

BigInsights Summary

- **BigInsights = analytical platform for persistent "Big Data"**
  - Based on open source & IBM technologies
  - Managed like a start-up . . . .

- **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; simplifies development and maintenance*

- **IBM advantage**
  - Combination of software, hardware, services and advanced research

Analytic Applications

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

Integration & Governance

First Ever Forrester Wave on Hadoop, 2012Q1

FORRESTER®
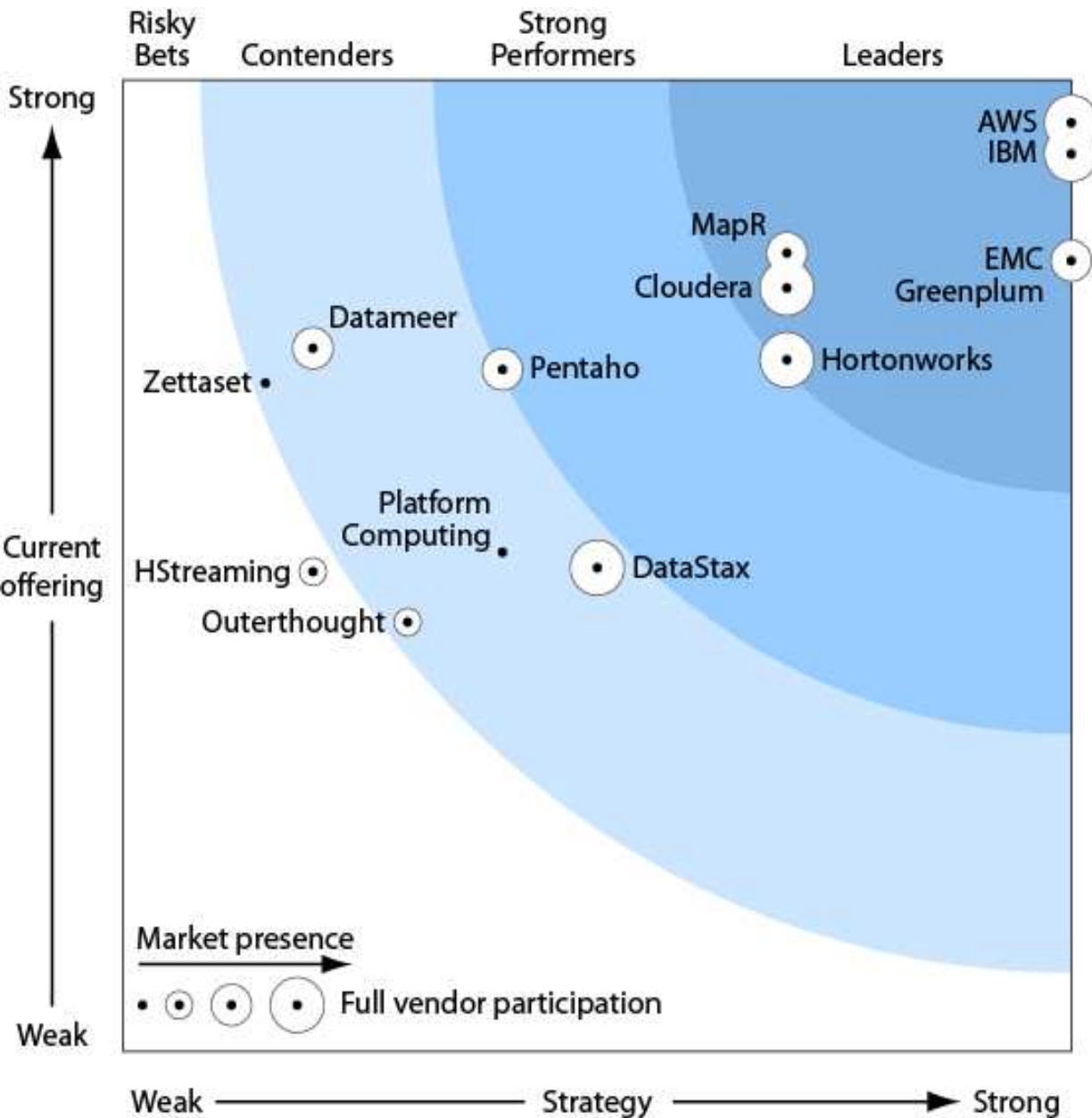
"IBM has the deepest Hadoop platform and application portfolio. IBM, an established EDW vendor, has its own Hadoop distribution; an extensive professional services force working on Hadoop projects; extensive R&D programs developing Hadoop technologies; connections to Hadoop from its EDW."
–The Forrester Wave™: Enterprise Hadoop Solutions, 1Q12

IBM & Turkcell Confidential

Big Database Vendors Adopt Hadoop

Microsoft climbs onto Hadoop bandwagon

Joint Microsoft-Hortonworks effort aims to deliver a Hadoop distribution for Windows Server and Windows Azure

By Jaikumar Vijayan
October 12, 2011 03:40 PM ET   💬 Add a comment   Like ‹ 6   +1 ‹ 0

Computerworld - Microsoft is the latest of the world's top IT vendors to climb aboard the Hadoop 'big data' bandwagon.

Oracle Jumps on Hadoop Bandwagon with Cloudera

By Jennifer LeClaire
January 10, 2012 12:10PM   Follow NewsFactor on twitter!   SHARE

Charles King, principal analyst at Pund-IT, said the Oracle-Cloudera partnership is a win-win. As he sees it, Oracle is coming late to the Hadoop-based Big Data party with Cloudera. He points to EMC's Greenplumb acquisition and IBM's Netezza acquisition, as well as HP's Vertica acquisition in this space as evidence.

Teradata taps Hortonworks to improve Hadoop story

By Derrick Harris | Feb. 21, 2012, 6:33am PT | No Comments

IBM Watson

InfoSphere Streams - Streaming Analytics for Big Data

❿ **Built to analyze data in motion**

- Multiple concurrent input streams

- Massive scalability

❿ **Process and analyze a variety of data**

- Structured, unstructured content, video, audio

- Advanced analytic operators

❿ **Enables Adaptive Real-Time Analytics**

- With Data Warehousing

- With Hadoop Systems

Analytic Applications

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

IBM Big Data Platform

Visualization & Discovery | Application Development | Systems Management

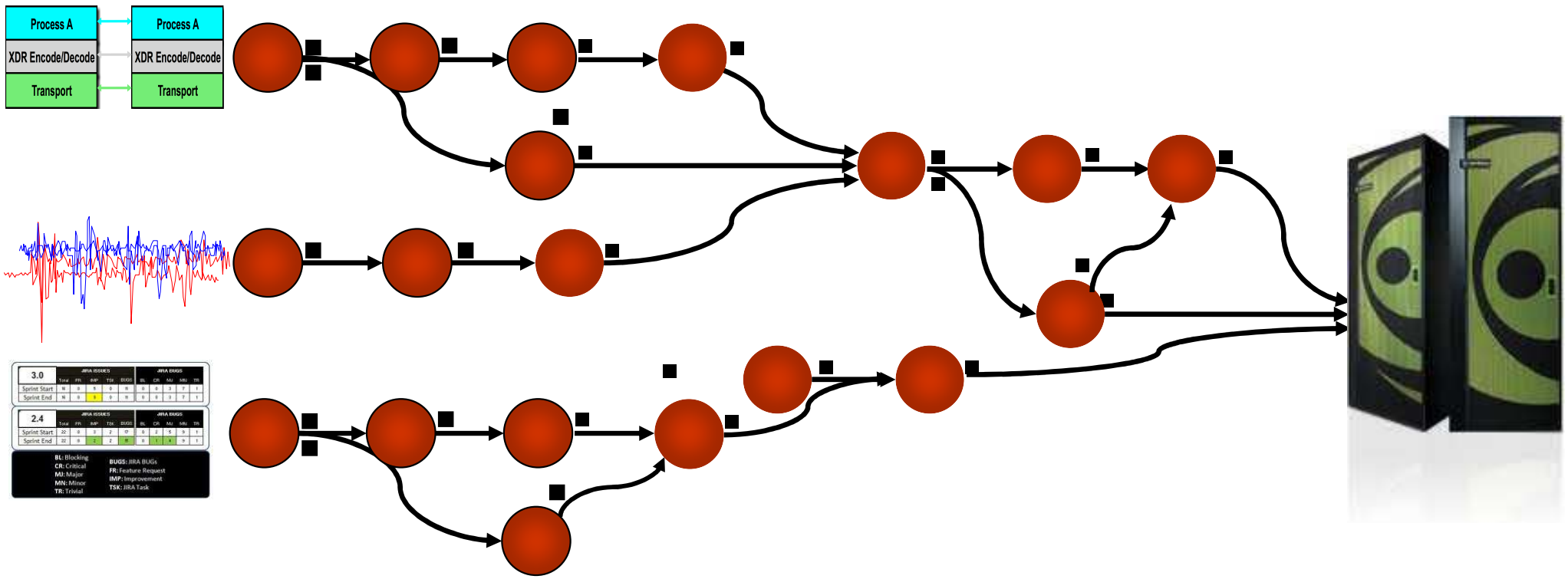Stream Computing

Information Integration & Governance

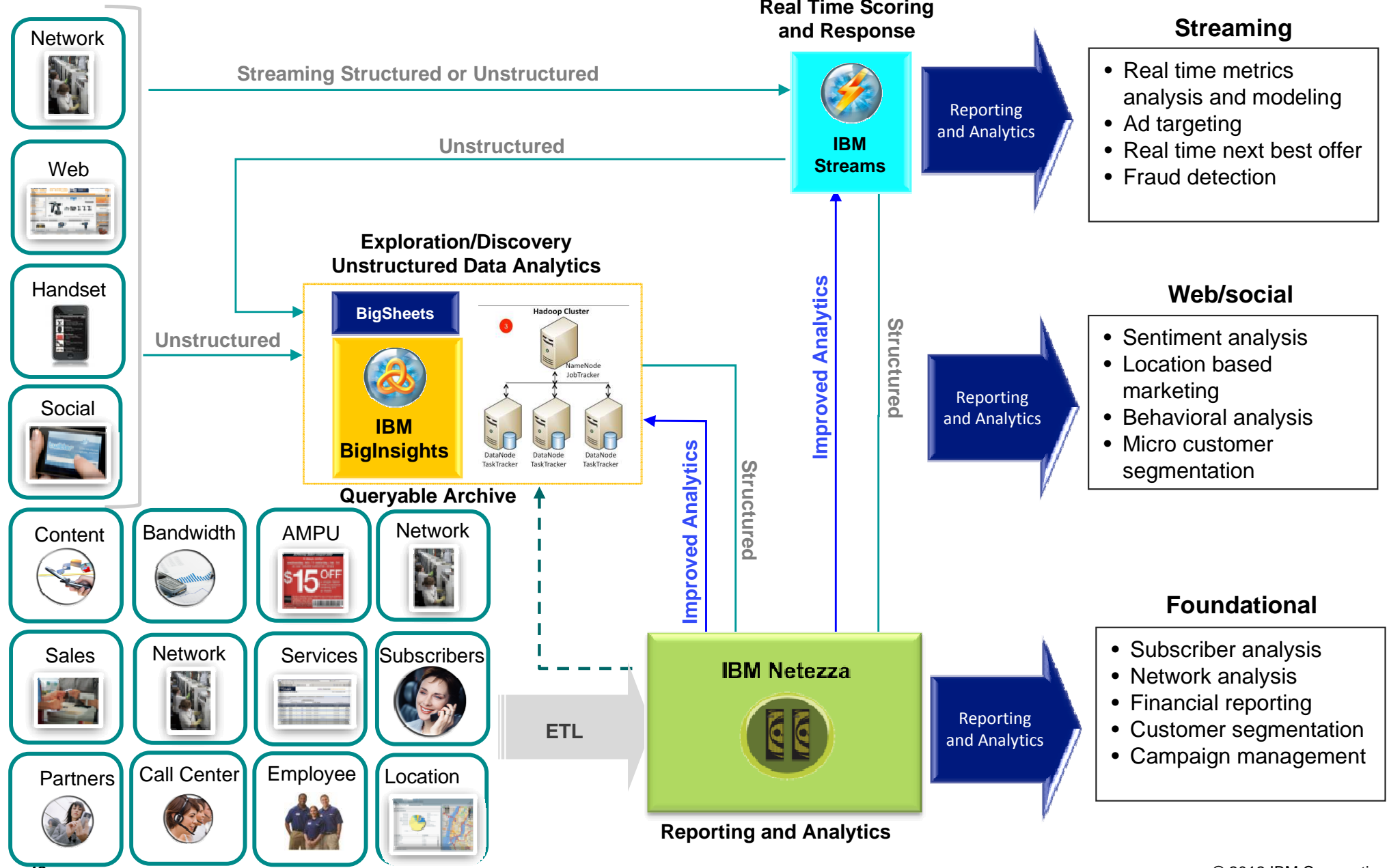# Data In Motion – Improving What They Already Have

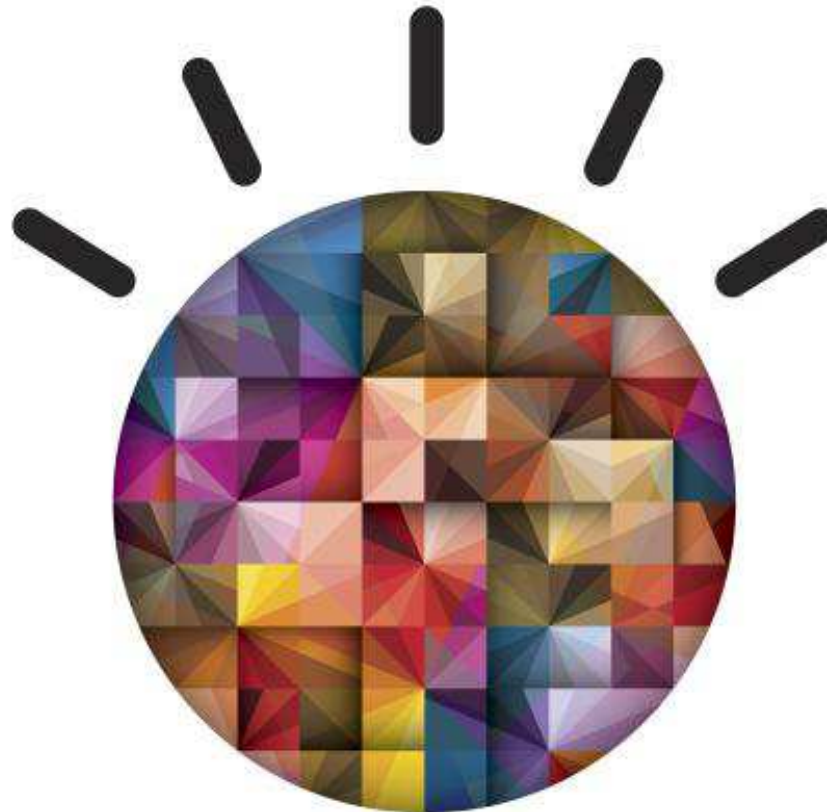Uses InfoSphere Streams and IBM Netezza

Significant Benefits:

- Over 90% reduction in time to merge/load call record data
- Over 90% reduction in storage
- Increased network quality,
- improved customer satisfaction,
- reduced churn

IBM Big Data integrates
Streams, Hadoop and Netezza in an integrated platform

Network

Web

Handset

Social

Content **Bandwidth** **AMPU** **Network**

Sales **Network** **Services** **Subscribers**

Partners **Call Center** **Employee** **Location**

Streaming Structured or Unstructured

Unstructured

Unstructured

Exploration/Discovery
Unstructured Data Analytics

BigSheets

Hadoop Cluster

NameNode
JobTracker

DataNode
TaskTracker   DataNode
TaskTracker   DataNode
TaskTracker

IBM BigInsights

Queryable Archive

ETL

IBM Netezza

Reporting and Analytics

Real Time Scoring and Response

IBM Streams

Improved Analytics

Structured

Improved Analytics

Structured

Reporting and Analytics

Reporting and Analytics

Reporting and Analytics

Reporting and Analytics

Streaming
- Real time metrics analysis and modeling
- Ad targeting
- Real time next best offer
- Fraud detection

Web/social
- Sentiment analysis
- Location based marketing
- Behavioral analysis
- Micro customer segmentation

Foundational
- Subscriber analysis
- Network analysis
- Financial reporting
- Customer segmentation
- Campaign management

For more information:

ibm.com/bigdata

#ibmbigdata