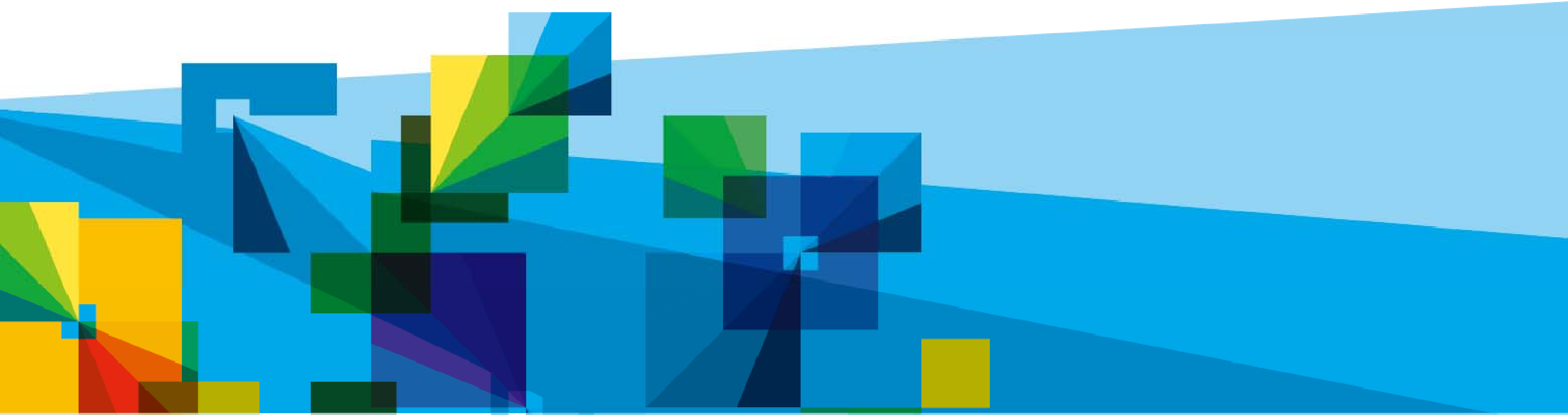
Big Data, Integration & Governance

28 - 30 August | Canberra; Melbourne; Sydney



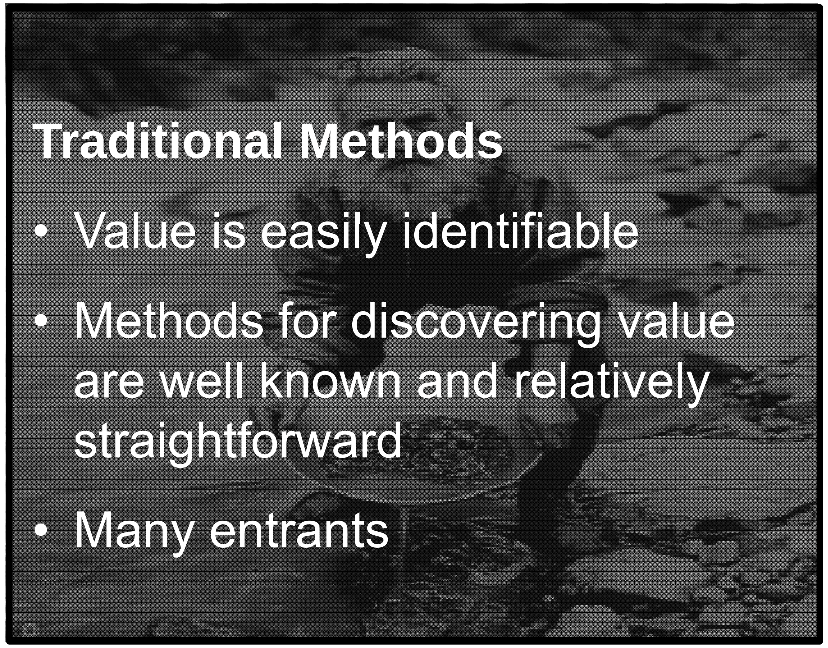
Agenda

1 What is Big Data and Why is it Compelling?

2 Big Data and the Information Supply Chain

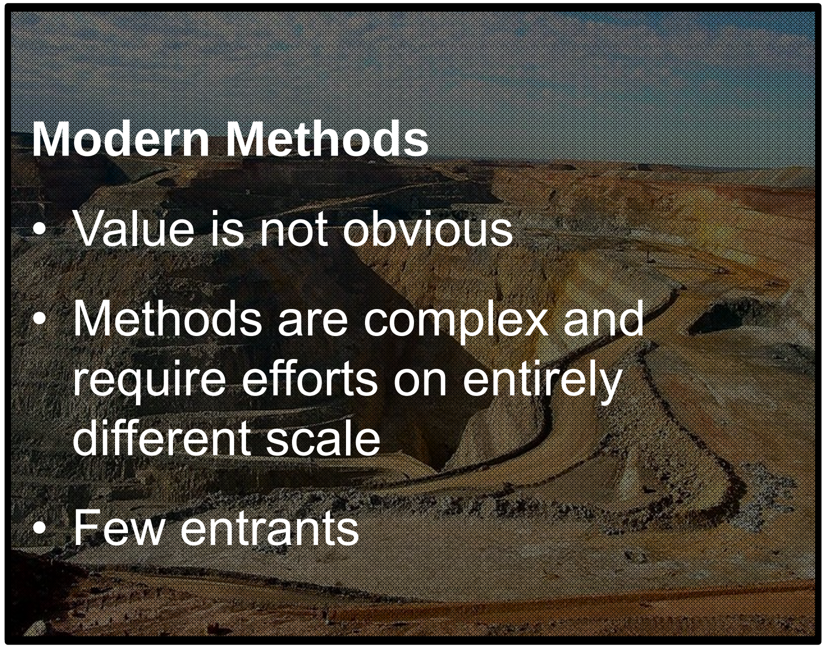
3 Data Integration, Governance and Big Data

How do organizations find value?



Traditional Methods

- Value is easily identifiable
- Methods for discovering value are well known and relatively straightforward
- Many entrants



Modern Methods

- Value is not obvious
- Methods are complex and require efforts on entirely different scale
- Few entrants

Kalgoorlie's Super Pit

What is Big Data?

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

Volume

12 terabytes
of Tweets create daily

Analyze product sentiment

350 billion
meter readings per annum

Predict power consumption

Velocity

5 million
trade events per second

Identify potential fraud

500 million
call detail records per day

Prevent customer churn

Variety

100’s video feeds
from surveillance cameras

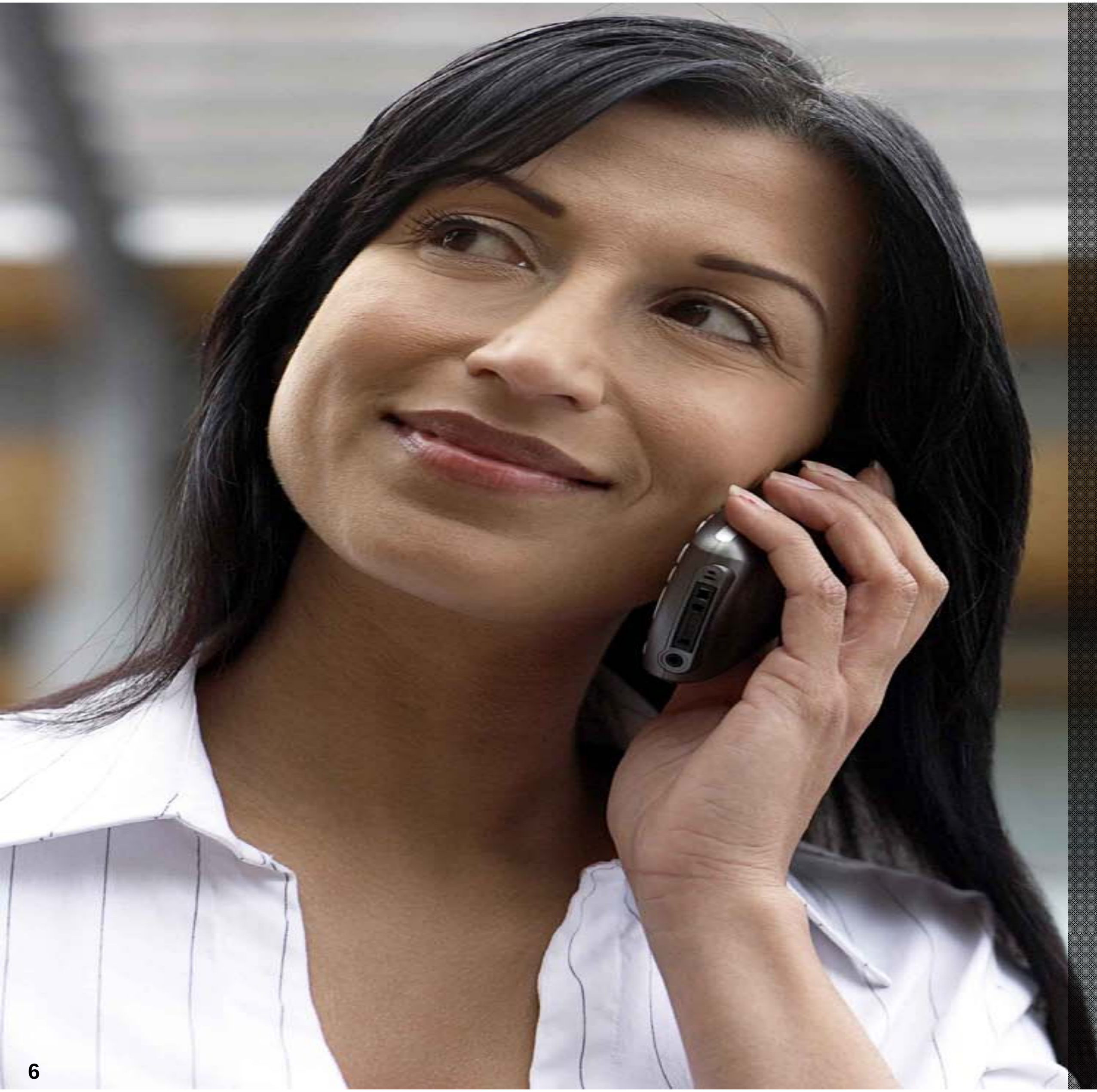
Monitor events of interest

80% data growth
are images, video, documents…

Improve customer satisfaction

Vestas optimizes wind turbine placement and operating life expectancy

- Analyze 2.8 petabytes of climate data to predict weather patterns at potential sites.
- More data means more accurate and richer models and results
    - Granularity 27km x 27km grids: driving to 9x9, 3x3 to 10m x 10m simulations
- Reduced response time for wind forecasting from weeks to hours
- Shortened time to develop a wind turbine site by nearly a month



Asian telco reduces billing costs and improves customer satisfaction

- Ensure real-time mediation and analysis of 6 billion Call Detail Records per day
- Uses stream computing for real-time data integration and analytics
  - Data processing time reduced from 12 hours to 1 second
  - Hardware cost reduced to 1/8th
- Proactively address issues (e.g. dropped calls) impacting customer satisfaction

Big Data: at-rest and in-motion

Data at-rest
Hadoop-based Analytics

- Analyze massive variety and volume of all data types
- Explore data to understand potential value to business

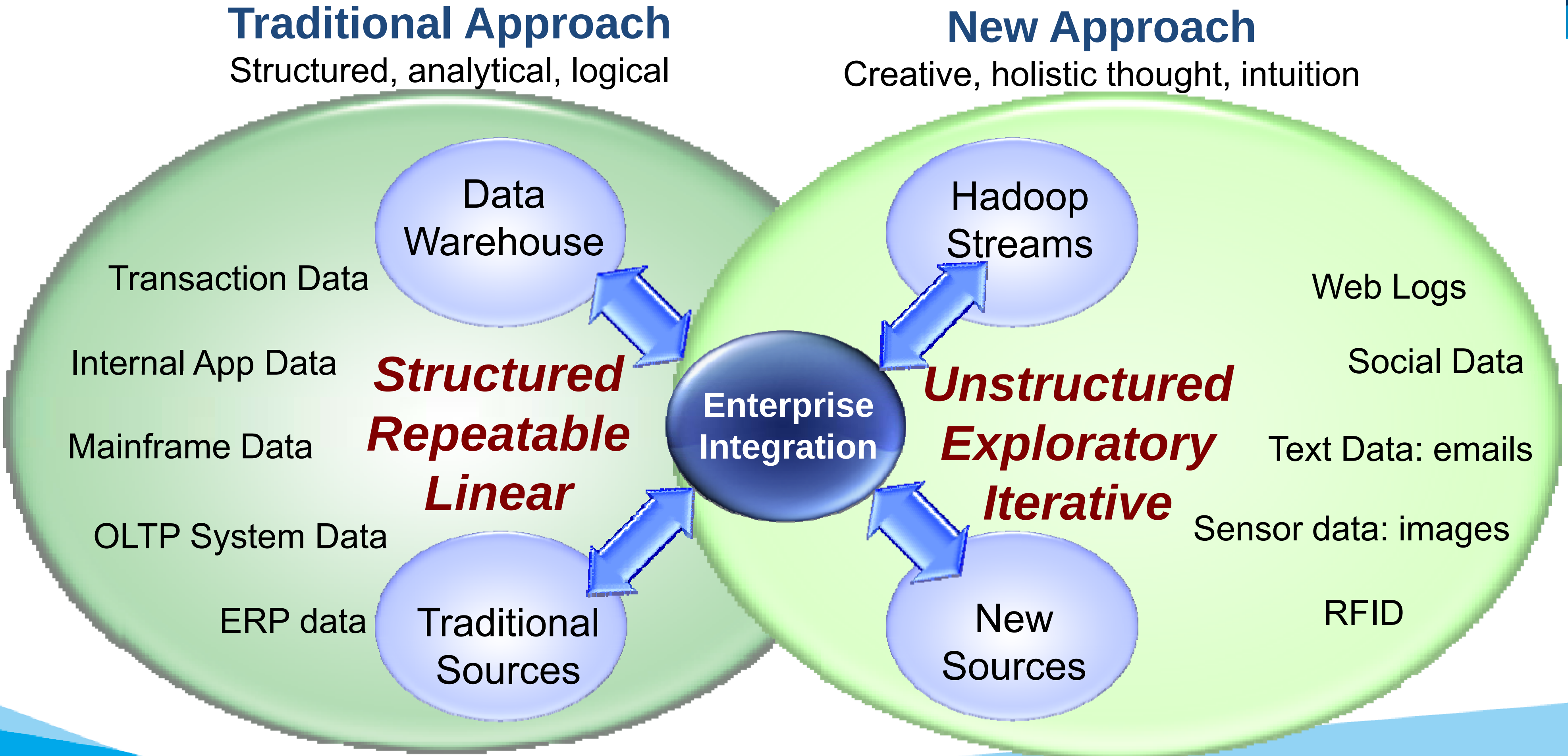
InfoSphere BigInsights

Data in-motion
Stream-based Analytics

- Analyze streaming data with multiple data types
- Respond to millions of events per second as they happen

InfoSphere Streams

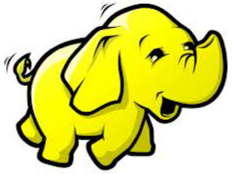
Information quantity and diversity



What is Hadoop

Description

- Apache Hadoop is a platform and a framework that supports data-intensive applications
- It enables applications to work with thousands of commodity hardware nodes in a cluster and scale out to processing petabytes of data by pushing processing to the data ("data locality")
- Two primary technologies
  1. Map/Reduce
  2. Hadoop Distributed File System (HDFS)



Observations

- Great benefits
  - ✓ Scalable
  - ✓ Fault Tolerant
  - ✓ Low cost per compute
- Some challenges
  - ✓ Relatively immature
  - ✓ Tooling is just now emerging
  - ✓ Few trained proficient resources
  - ✓ Lacks features that would be considered enterprise class

Why is Big Data so Compelling

Creates the opportunity to do something previously unachievable

- *Previously may not have had the ability to scale processing so large*

Reduces cost model and aligns opportunity to investment

- *Scaling is based on low cost commodity hardware (cost per compute)*

Removes processing burden on alternative infrastructures

- *Analytical processing across structure/unstructured lends itself to Big Data platforms over other conventional technologies*

What Big Data **is not**

A Magic Pill

“Can I run my <<insert ERP name>> on top of Hadoop?”

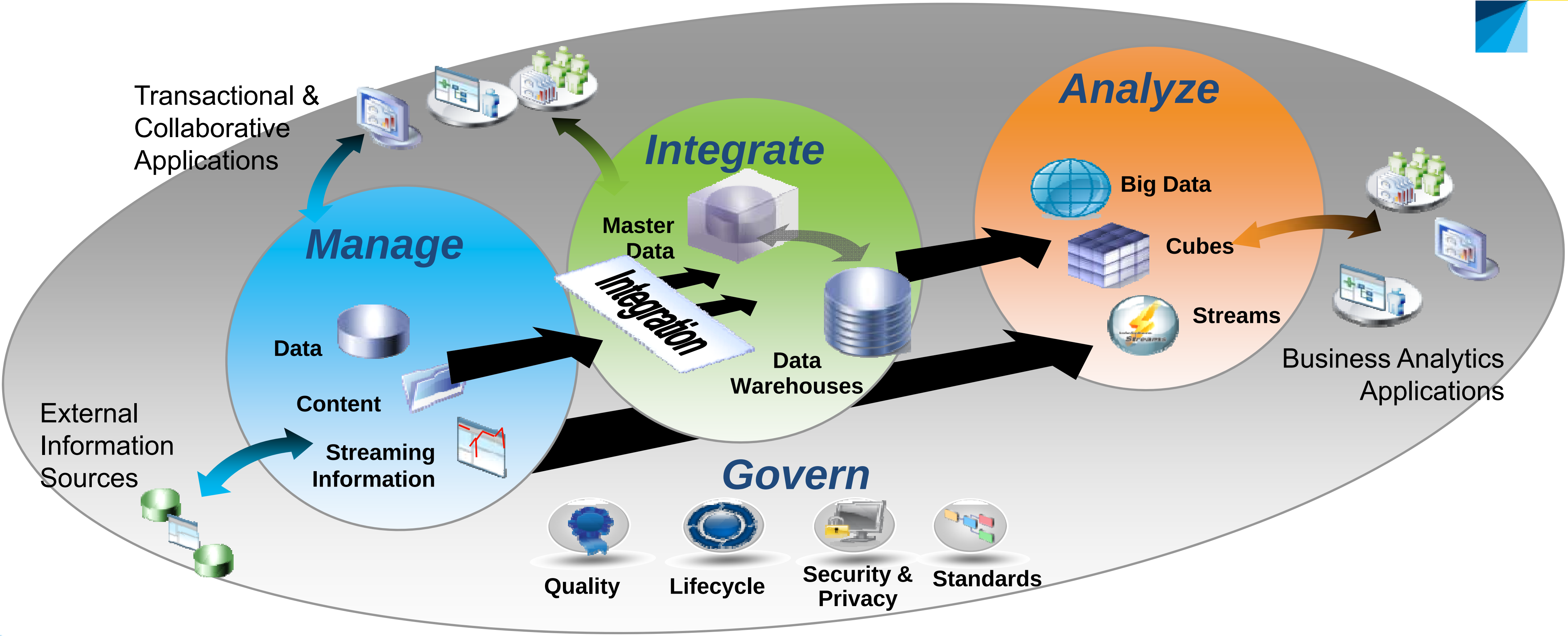
There are some things that Big Data simply doesn’t do well (updates/deletes, transactional consistency, guaranteed delivery, etc…)

People still need to consume information in a common form, using consistent values, and governed calculations.

A Big Garbage Bin

“a big garbage bin with just store all data into it without schema”

Fitting into the Information Supply Chain



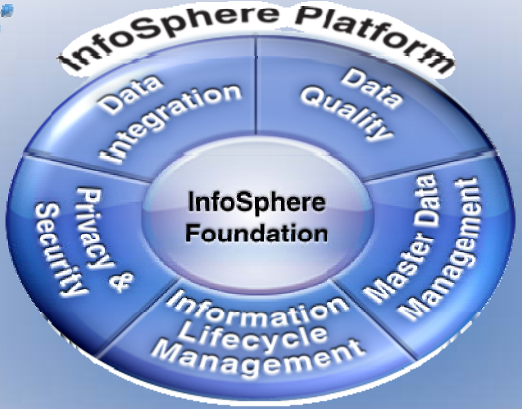
Information & Governance for Big Data

Integrate & Link Big Data

- Big Data as a Source
- Big Data as a Target
- Data Transformations
- Data Movement
- Integrate w/existing Enterprise
- Data Lineage & Impact Analysis
- Metadata Integration w/Analytics
- Realtime & Data Federation

Cleanse and Validate Big Data

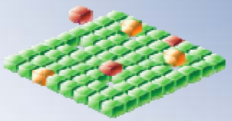
- Accuracy and Entity Matching with Social Data
- De-duplication and Standardization of Machine Data
- In-line Cleansing with Integration
- Trusted Data Dashboard and Reporting on Data Quality



Master Big Data

- Big Data as a Supplier
- Big Data as a Consumer
- Links between Big Data and Trusted Golden Records
- Leverage Master Data in Big Data Analytics
- Entity Resolution at Extreme Scale Out Levels
- Probabilistic Entity Matching

Protect Big Data



- Activity Monitoring
- Data Masking
- Data Encryption
- On-Demand / In-Place Protection
- In-Line Protection (w/ETL etc.)
- Active Detection & Alerting

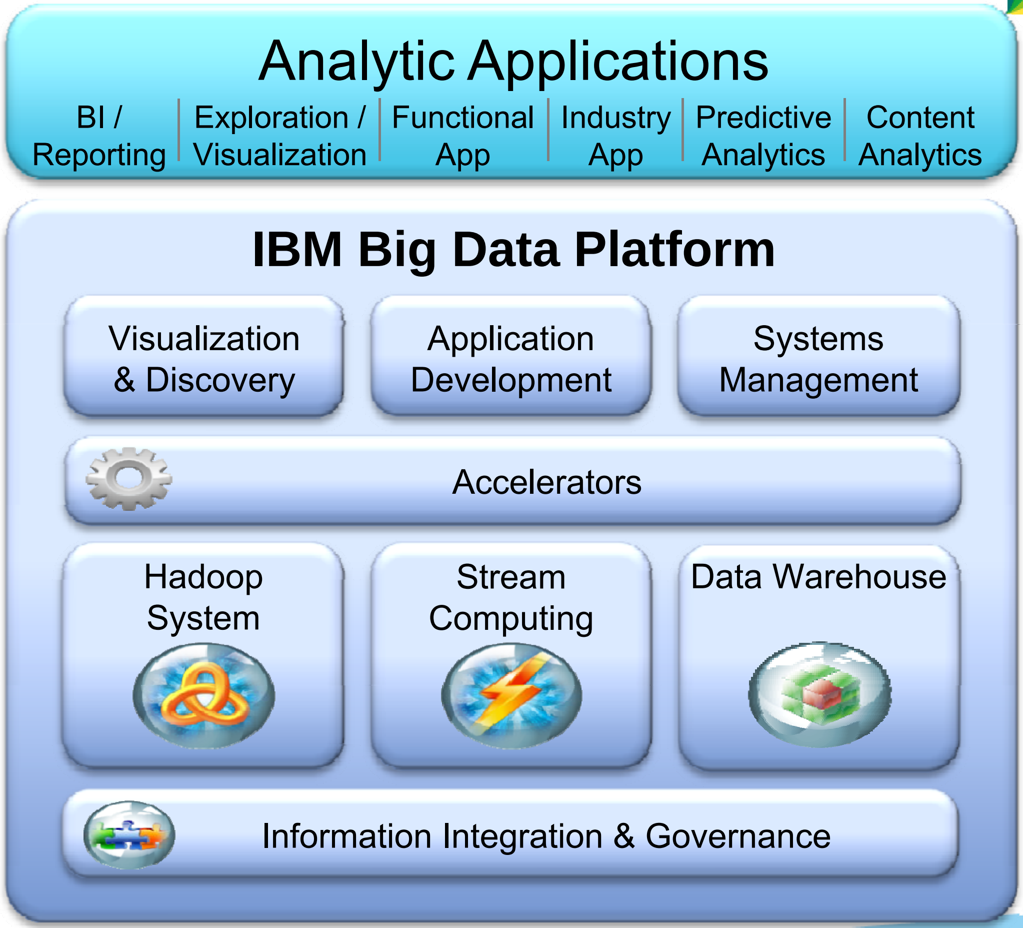
Audit & Archive Big Data

- Queryable Archive
- Structured and Semi-Structured
- Optimized Connectors to existing Apps
- Hot-Restorable On-the-Fly
- Immutable and Secure Access
- Automated Legal Hold Capability for Data Freeze

IBM’s Big Data Platform

New analytic applications drive the requirements for a big data platform

- Integrate and manage the full variety, velocity and volume of data
- Apply advanced analytics to information in its native form
- Visualize all available data for ad-hoc analysis
- Development environment for building new analytic applications
- Workload optimization and scheduling
- Security and Governance



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

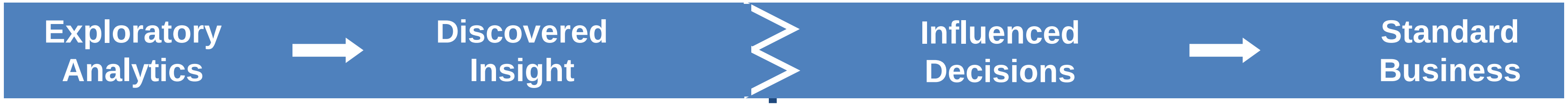
Hadoop System

Stream Computing

Data Warehouse

Information Integration & Governance

Analytic Lifecycle drives Governance Requirements



Exploratory Analytics → Discovered Insight

Influenced Decisions → Standard Business

Even in traditional Business Intelligence stack, technology in this category is typically not required to be full governed

- *Little security concerns*
- *No audit trail standards*
- *No data retention requirements*
- *Little to no regulation*
- *Data quality not strategic*
- *Sources of information are “interesting”*

Once the data becomes part of the decision making process of the organization, all of the standard controls become relevant

- *Protect, Secure, Encrypt*
- *Track access and changes*
- *Preserve data for N years*
- *Legislated requirements for industry*
- *Data quality imperatives*
- *Sources of information must be “trusted”*

No difference in data governance requirements once the data is used for making operational business decisions

Information Server and Governance

B

Business Metadata

– Business rules, Stewardship, Business Definitions, Auditing Terminology, Glossaries, Algorithms and Lineage using business language.

T

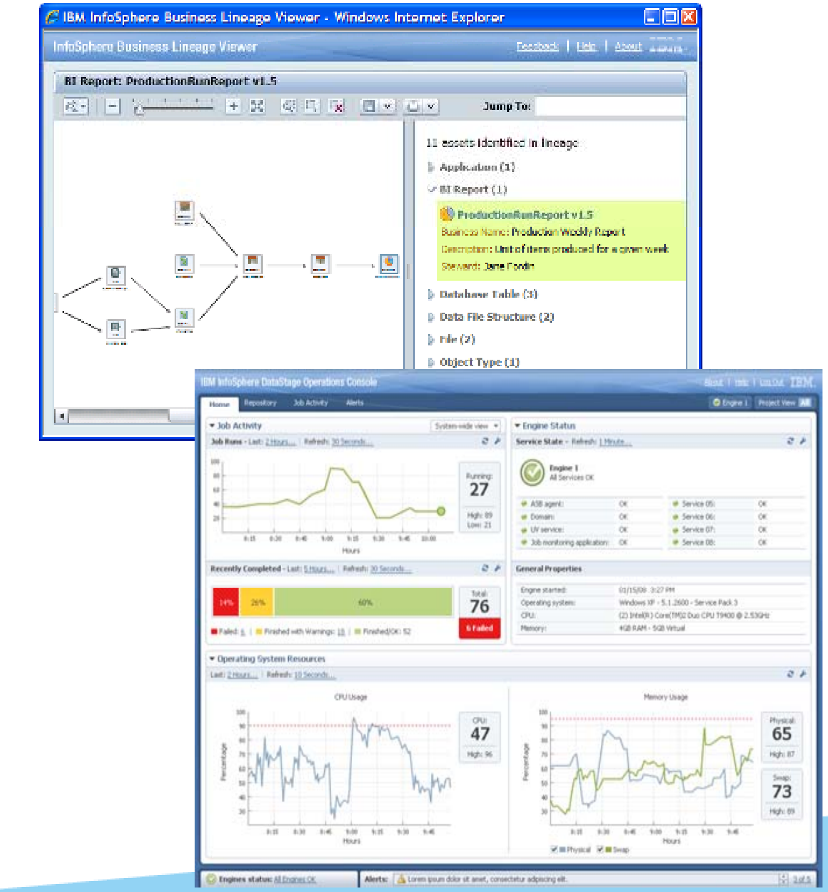
Technical Metadata

– Defines Source and Target systems, their Table and Fields structures and attributes, Documentation for Auditing Derivations and Dependencies. Audience: Specific Tool Users – BI, ETL, Profiling, Modeling.

O

Operational Metadata

– Information about application runs: their frequency, record counts, component by component analysis and other statistics for auditing purposes. Audience: Operations, Management and Business Users.



IBM InfoSphere Business Lineage Viewer - Windows Internet Explorer

InfoSphere Business Lineage Viewer

BI Report: ProductionRunReport v1.5

11 assets identified in lineage

- Application (1)
- BI Report (1)
  - ProductionRunReport v1.5
  - Business Name: Production Weekly Report
  - Description: Unit of items produced for a given week
  - Steward: Jane Fordin
- Database Table (3)
- Data File Structure (2)
- File (2)
- Object Type (1)

IBM InfoSphere DataStage Operations Console

Job Activity

Job Runs - Last: 2 Hours... | Refresh: 30 Seconds...

Running: 27

High: 89 Low: 21

Recently Completed - Last: 5 Hours... | Refresh: 30 Seconds...

Total: 76

6 Failed

🟥 Failed: 6 | 🟨 Finished with Warnings: 18 | 🟩 Finished/OK: 52

Engine Status

Service State - Refresh: 1 Minute...

Engine 1
All Services OK

ASB agent:	OK	Service 05:	OK
Domain:	OK	Service 06:	OK
UV service:	OK	Service 07:	OK
Job monitoring application:	OK	Service 08:	OK

General Properties

Engine started:	01/15/08 3:27 PM
Operating system:	Windows XP - 5.1.2600 - Service Pack 3
CPU:	(2) Intel(R) Core(TM)2 Duo CPU T9400 @ 2.53GHz
Memory:	4GB RAM - 5GB Virtual

Operating System Resources

Last: 2 Hours... | Refresh: 10 Seconds...

CPU: 47 High: 96

Physical: 65 High: 87

Swap: 73 High: 89

“Bigger” Data Integration Challenges

More sources and targets

- Big Data introduces additional data stores that need to be integrated

New data store types

- Big Data has added and will continue add new data stores (noSQL) that don’t easily lend themselves to conventional methods for data movement

New data types and formats

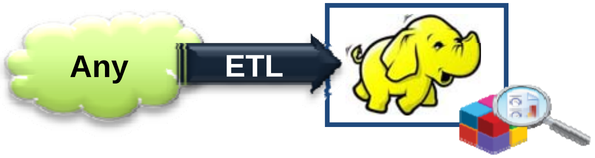
- Unstructured data; polymorphic data structures; JSON, Avro, ???

Larger volumes

- Solutions need to move, transform, cleanse and otherwise prepare huge data volumes

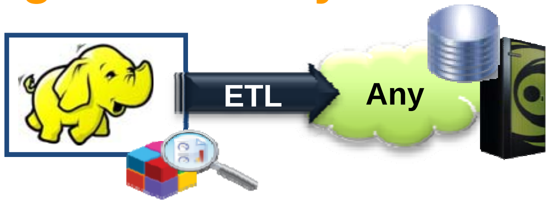
“Bigger” Data Integration Common Use Cases

Any to Big Data



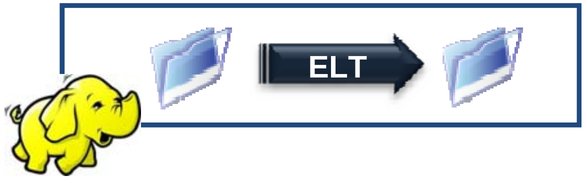
“I need to mix in traditional sources into Hadoop so that I can run the analytical models I need.”

Big Data to Any



“Now that I know something new, how do I move this back into my applications and warehouses so that it is easily consumable.”

Big Data Hub



“I need to transform and cleanse information to make it (re)usable for analytics but can’t afford to move TBs across the network frequently.”

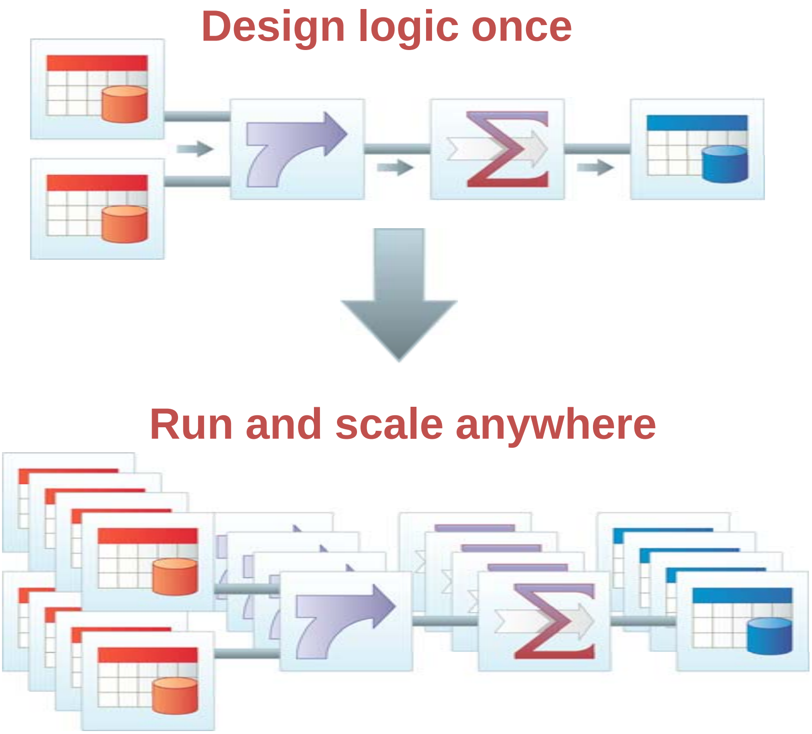
InfoSphere DataStage : “Hadoop-ish” ETL

Built upon principles of Massively Parallel Processing

- Automatic pipeline partitioning across job logic components
- Automatic data partitioning based on user-defined or dbms driven partitioning
- Ability to scale application across SMP, MPP or Grid environments as specified at job runtime to fully abstract the job logic from the processing environment.

Data Integration Specific Optimizations

- Industry unique dynamic repartitioning of data in stream to support sources & targets which are partitioned differently without having to land information to disk



A few Customer Stories

- Healthcare organization runs 200,000 programs built in Information Server on a grid/cluster of low commodity hardware.
- Financial institution desensitizes 200 TB of data one Saturday each month to populate their development environments.
- Medical research organization combines text analytics running inside Information Server to process 200 million medical documents a weekend and create indexes to support optimal retrieval by end users.

Marketing Technology and Services Company has been running a 600 node grid of Information Server for 8 years processing billions of records regularly

Comprehensive Integration Platform

Data Integration

Dozens of prebuilt transformation objects and 100s of functions

Data Quality

Data validation & cleansing applicable for multiple information domains

Connectivity

Databases, Big Data, Messages,CDC, Mainframe and more

Parallel Engine

Most scalable data integration engine supporting SMP, MPP and Grid with simple configuration file control

One Design Environment

Single design paradigm advances time to value

One Set of Design Artifacts

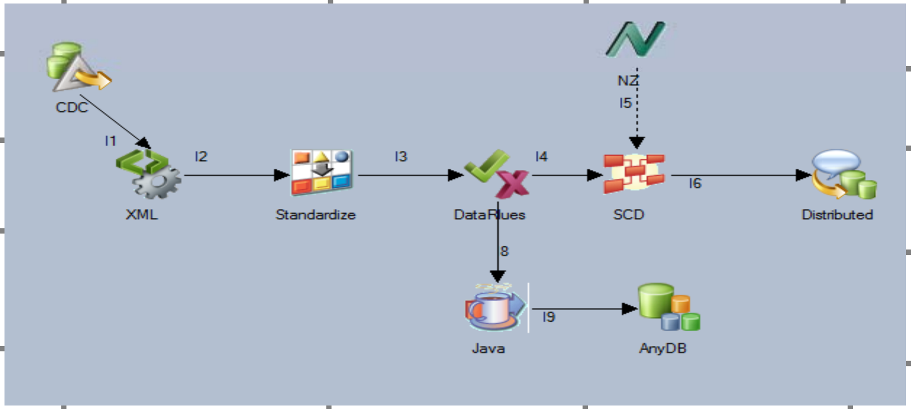
Logic represented by one set of design objects regardless of deployment styles

One Metadata Store

Maximizes business & IT collaboration and accelerates data governance efforts

One Administration Center

Integration of install, security, auditing, connectivity, logging reduces TCO



Batch and Real-time

Traditional scheduled batch or real-time through various modes of operation

Distributed Transactions

Scalable, heterogeneous information fabric with guaranteed data delivery (2PC)

Balanced Optimization

Maximize DBMS infrastructure by moving processing to it

Data Virtualization

Expose any data integration, quality, monitoring, etc… component using web services, RSS, REST, JMS...

As part of InfoSphere Information Server, directly benefits from other aspects of the suite – data profiling, mapping specifications, etc…

Enterprise Packs

Specific connectivity for leading ERP solutions

Std Industry Formats

SWIFT, EDI, HL7, etc... along with native and scalable XML and Complex File support

Business Rules

Drive critical enterprise logic from SMEs in your business

Data Masking

Protect sensitive information

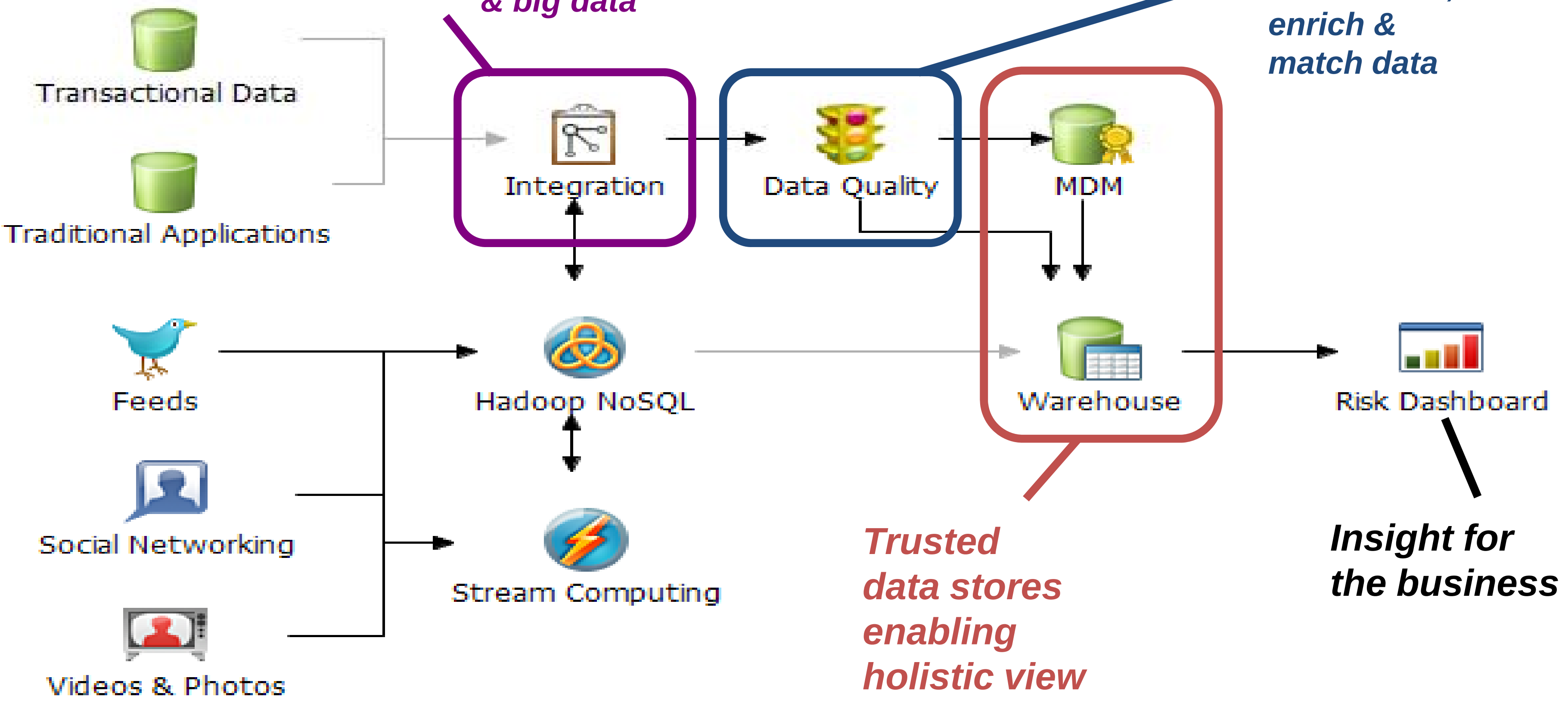
Applying Data Quality - Differences & Implications

Big Data may include invalid values, information noise, and incomplete.

Data that will be relied upon, must monitor and cleanse for such cases.

- Big data information is primarily machine-captured rather than manually entered
  - Requirement to potentially check the validity of the collection process rather than incorrect user entry
- Big data information is not owned by the enterprise
  - Information may be incomplete and degree of reliability may differ significantly
- Big data information is of a finer grain and higher volume
  - Importance of taking data volume and complexity into consideration of business value assessment / return on investment study
  - Need to "filter out the noise" before applying data quality

Making Big Data Quality Trusted



Unified & integrated access to traditional & big data

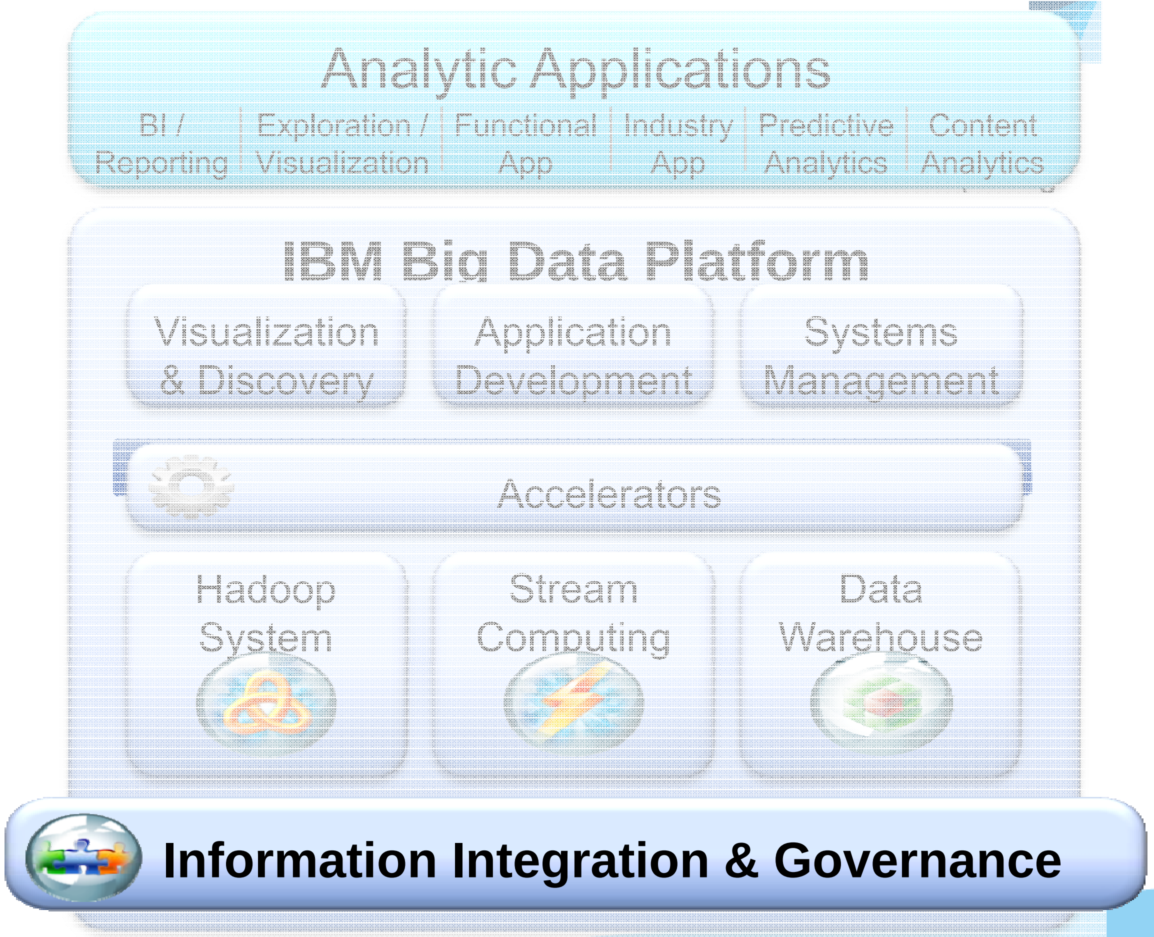
Validate, standardize, enrich & match data

Trusted data stores enabling holistic view

Insight for the business

Information Integration and Governance

- **Integrate** any type of data to the big data platform
  - Structured
  - Unstructured
  - Streaming
- **Govern** big data
  - **Secure** sensitive data
  - **Lifecycle management** to control data growth
  - Validate, cleanse & control **data quality** holistically
  - **Master data** to establish single version of the truth
  - Align business & IT based on end-to-end **metadata**

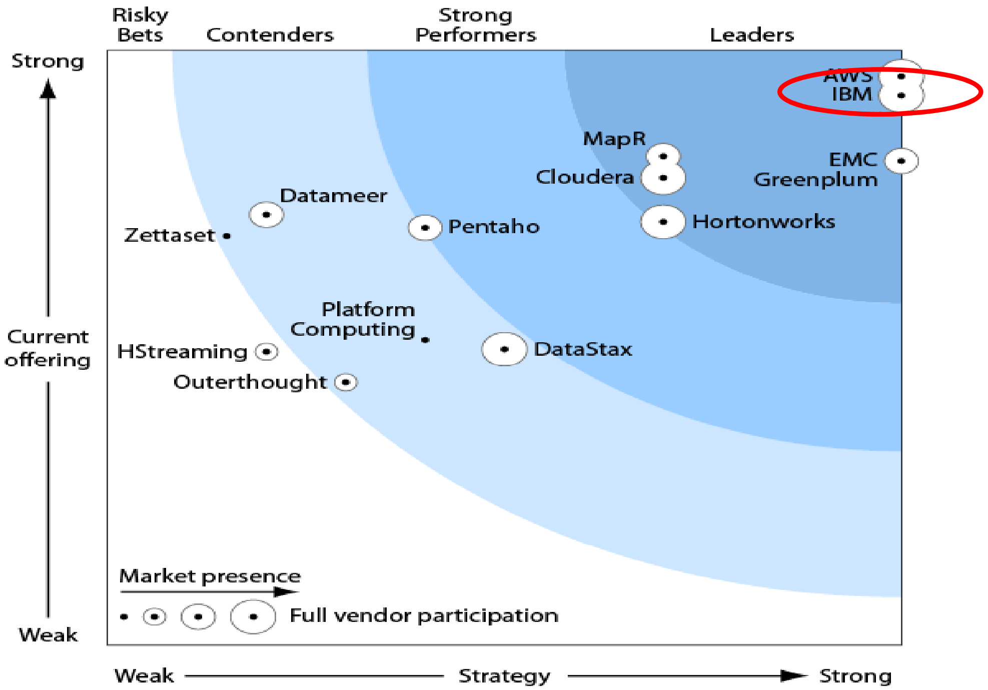


Recognized for Big Data Leadership

“IBM has the deepest Hadoop platform and application portfolio.”

FORRESTER

February 2012 **“The Forrester Wave™: Enterprise Hadoop Solutions, Q1 2012”**



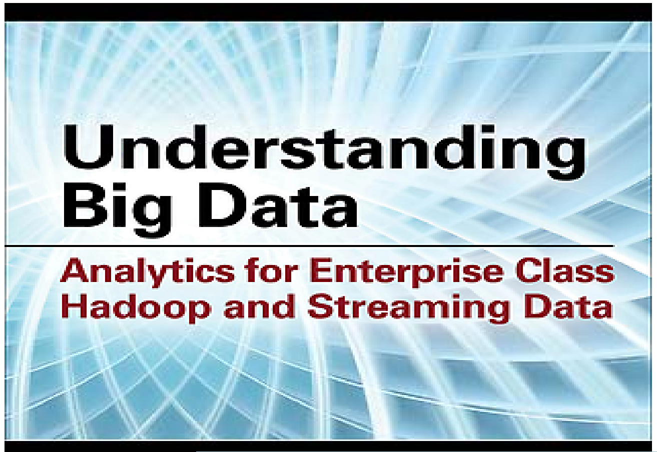
Thoughts on Getting Started

Get Educated

- Forum content
- Big Data University
- Books / Analyst papers

Schedule a Big Data Workshop

- Free of charge
- Best practices
- Industry use cases
- Business uses
- Business value assessment

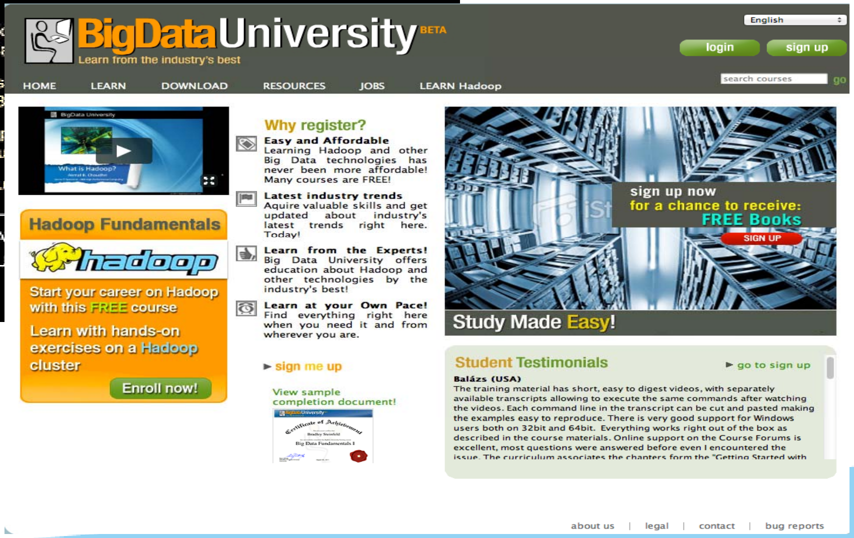


Understanding Big Data

Analytics for Enterprise Class Hadoop and Streaming Data

- Learn ho class sca
- Gain ins at-rest B
- Learn tip and solu
- Get a qu

CHRIS EA
TOM DEU



BigDataUniversity BETA
Learn from the industry's best

Hadoop Fundamentals

hadoop

Start your career on Hadoop with this FREE course

Learn with hands-on exercises on a Hadoop cluster

Why register?

Easy and Affordable
Learning Hadoop and other Big Data technologies has never been more affordable! Many courses are FREE!

Latest industry trends
Aquire valuable skills and get updated about industry's latest trends right here. Today!

Learn from the Experts!
Big Data University offers education about Hadoop and other technologies by the industry's best!

Learn at your Own Pace!
Find everything right here when you need it and from wherever you are.

► sign me up

View sample completion document!

sign up now
for a chance to receive:
FREE Books

Study Made Easy!

Student Testimonials

► go to sign up

Balázs (USA)
The training material has short, easy to digest videos, with separately available transcripts allowing to execute the same commands after watching the videos. Each command line in the transcript can be cut and pasted making the examples easy to reproduce. There is very good support for Windows users both on 32bit and 64bit. Everything works right out of the box as described in the course materials. Online support on the Course Forums is excellent, most questions were answered before even I encountered the issue. The curriculum associates the chapters form the "Getting Started with

THINK BIG

Big Data, Integration & Governance

28 - 30 August | Canberra; Melbourne; Sydney

Thank You