Transforming Data With Intelligence™

Previews of TDWI course books offer an opportunity to see the quality of our material and help you to select the courses that best fit your needs. The previews cannot be printed.

TDWI strives to provide course books that are content-rich and that serve as useful reference documents after a class has ended.

This preview shows selected pages that are representative of the entire course book; pages are not consecutive. The page numbers shown at the bottom of each page indicate their actual position in the course book. All table-of-contents pages are included to illustrate all of the topics covered by the course.

This page intentionally left blank.

BIG DATA WORKSHOP

THE LITTLE SECRETS TO KNOW

WHAT YOU WILL LEARN

Data-Driven Decisions

- Data Value Chain.
- Interrogating A Cross-Organizational Data Set.
- Migration from silos to an integrated raw data landscape.

Big Data Technological Landscape

- Ecosystem Essentials:
- Understanding technologies and vendors
- Defining the value of distributed computing
- Technology stacks.
- Selection Criteria
- Data Swamp, Data Lake and Data Hubs.

The Intersection of Data Science and Big Data

- Data Science Practice – Teams and Skills
- Defining the Data Warehouse Goals for Data Science.
- Artificial Intelligence and Machine Learning Intersects.
- Monetizing Data.
- Pitfalls of Big Data.
- Challenges and Barriers.

Big Data Resources

- Building an internal team for data science.
- Hiring Data Science Team Members.
- Data Science Teams Success Goals.

New Directions and Opportunities for Innovation

- New business opportunities.
- Leveraging Big Data.
- Leadership development.
- Innovation Lab: how to take a problem, solve it, and push that solution up within the organization.

Metrics and Measures

- What to measure
- What are the metrics
- Who will monitor
- What outcomes need governance

2

EXAMPLE

Tweet: @jdoe – very disappointed with @united @checkin lousy svc, bad mgmt, long lines #fail.
20000 retweets. 4 hours ago

IT Perspective

- Text of 140 chars will be stored as **string**.
- The data model for this will be a table with source, content, datetime.

Business Perspective

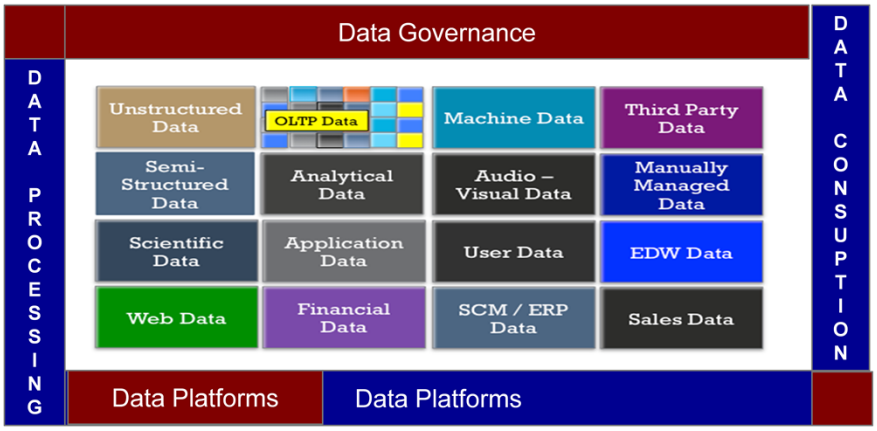
- User – JDoe
- Brand – United
- Sentiment – Negative
- Process – Check-in
- Time – Waiting in long lines
- Impact – shared 20000 times in 4 hours

16

VALUE OF DATA – WRITE IT OUT

17

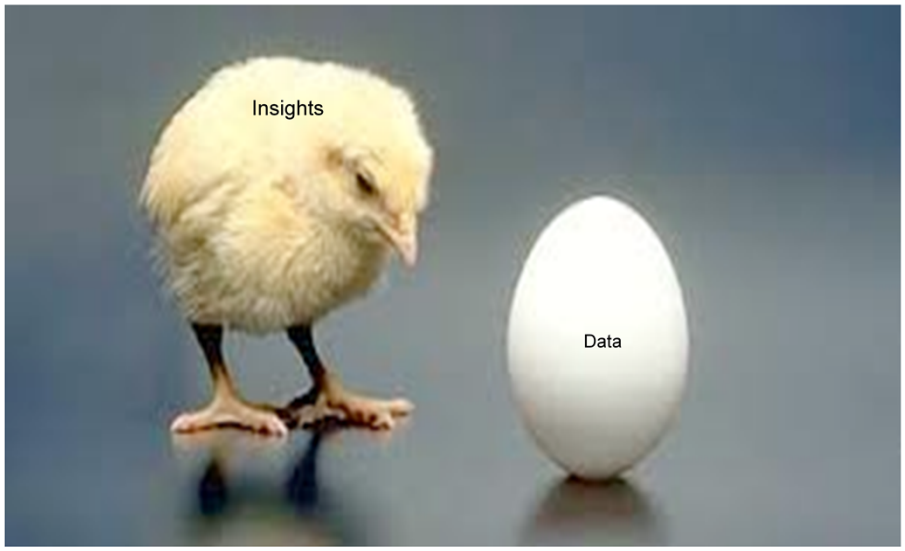
THE NEW DATA FABRIC



New Existing

18

PERSPECTIVE – FOOD FOR THOUGHT



19

ANALYSIS

	Acquire	Search	Process (Inspect / Analyze)		Metrics	Visualize
Business	3rd Party Data RSS Feeds Content Platforms Internet	Patterns Phrases Keys Tag	Categorize Filter Enrich	Contextualize Geo-Tag Meta-Tag Integrate	Score Aggregate Domain	Trends Sentiments Alerts Competitive Intel

	Acquire	Search	Process (Inspect / Analyze)		Metrics	Visualize
IT	Search Platforms Crawlers API Development			Metadata GeoSpatial Data	Platform Setup & Integration	Platform Setup & Workflow support

20

OTHER BIG DATA CHALLENGES – WRITE IT DOWN

22

INNOVATIONS

Category	New Frontiers
Infrastructure	Big Data and Data Warehouse Appliances In-Memory Technologies SSD Storage Fast Networks Cloud Mobile Technologies
Software	In-memory Databases Hadoop, Cassandra & NoSQL Ecosystems Columnar DBMS Improved ETL-Hadoop integration – Informatica, Talend
Algorithms	Mahout
Pre-Configured Architectures	IBM, Teradata, Kognitio, EMC, CloudEra, HortonWorks, Cirro, Intel, Cicso UCS, Pivotal, Oracle, MapR

BIG DATA – INFRASTRUCTURE REQUIREMENTS

- Scalable platform
- Database independent
- Fault tolerant
- Low cost of acquisition
- Scalable and Reliable Storage
- Supported by standard toolsets
- Datacenter Ready

25

BIG DATA – WORKLOAD DEMANDS

- Process dynamic data content
- Process unstructured data
- Systems that can scale up with high volume data
- Systems that can scale out with high volume of users
- Perform complex operations within reasonable response time

26

KEY VENDORS AND TECHNOLOGIES – WRITE DOWN YOUR CHOICES

29

BIG DATA TECHNOLOGIES

Apache Software Foundation

- Hadoop
- HBASE
- Zookeeper
- Oozie
- Avro
- Pig
- Sqoop
- Flume
- Cassandra
- Spark

CloudEra

HortonWorks

MongoDB

IBM BigInsights

EMC Pivotal

Teradata Aster – Big Data Appliance

Oracle Big Data Appliance

Intel Hadoop Distribution

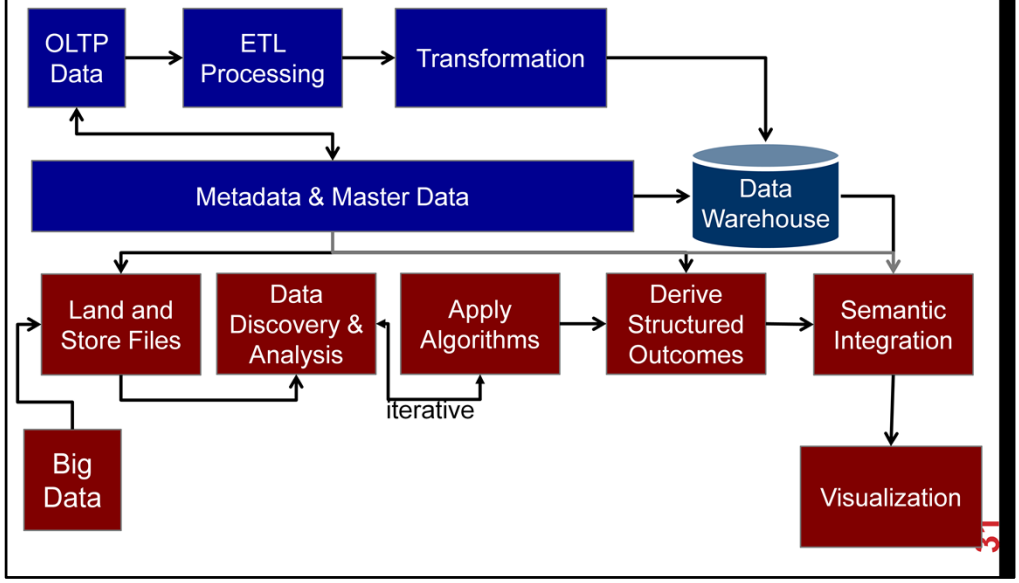
MapR

Datastax

QueryIO

30

DECISION SUPPORT ARCHITECTURE

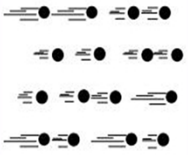
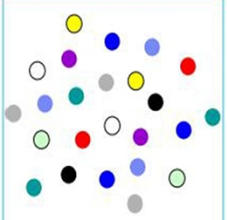


THE INTERSECTION OF DATA SCIENCE AND BIG DATA

DATA SCIENCE PRACTICE – TEAMS AND SKILLS
DEFINING THE DATA WAREHOUSE GOALS FOR DATA SCIENCE.
ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING INTERSECTS.
MONETIZING DATA.
PITFALLS OF BIG DATA.
CHALLENGES AND BARRIERS.

35

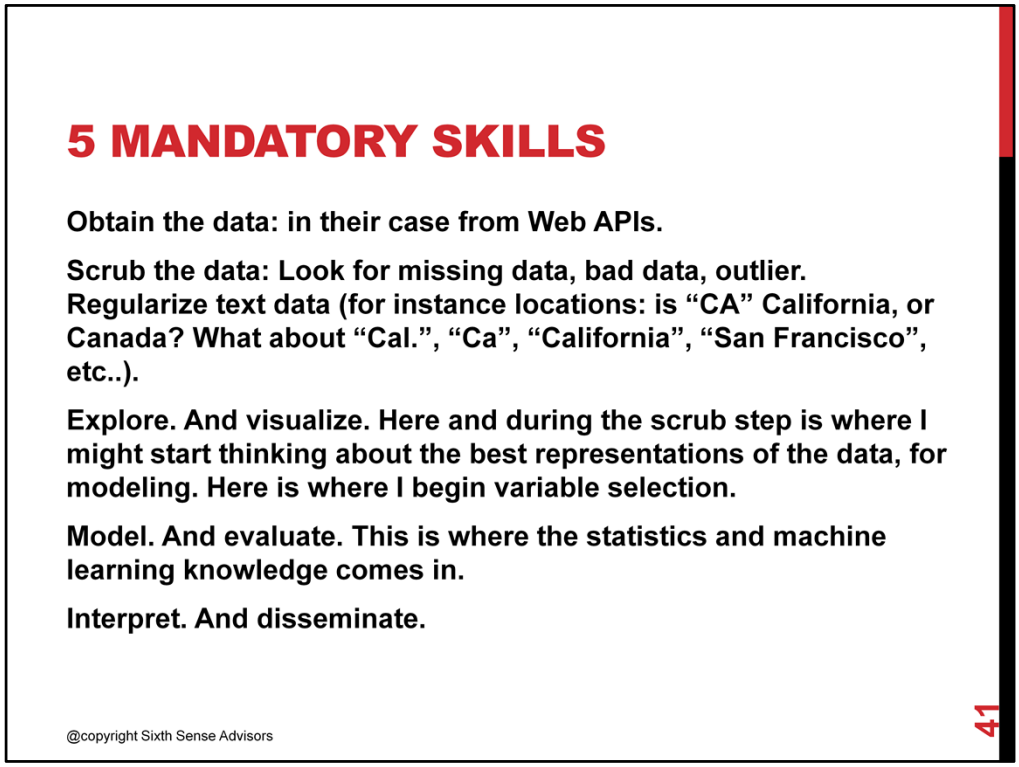
THE ISSUES…

Volume	Velocity	Variety	Veracity*
			
Data at Rest	Data in Motion	Data in Many Forms	Data in Doubt
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

36

DEFINE YOUR DATA SCIENCE TEAM

38



Source: Hillary Mason et al - http://www.dataists.com/2010/09/a-taxonomy-of-data-science/

MONETIZING DATA – HOW WILL BE BUILD A BUSINESS CASE? MONEY DRIVES

45

DISCUSSION ON HIRING DATA SCIENCE SKILLS

LEADERSHIP DEVELOPMENT - DISCUSSION

53

METRICS AND MEASURES

What to measure
What are the metrics
Who will monitor
What outcomes need governance

55

WHAT TO MEASURE - DISCUSSION

56

WHO WILL MONITOR - DISCUSSION

58