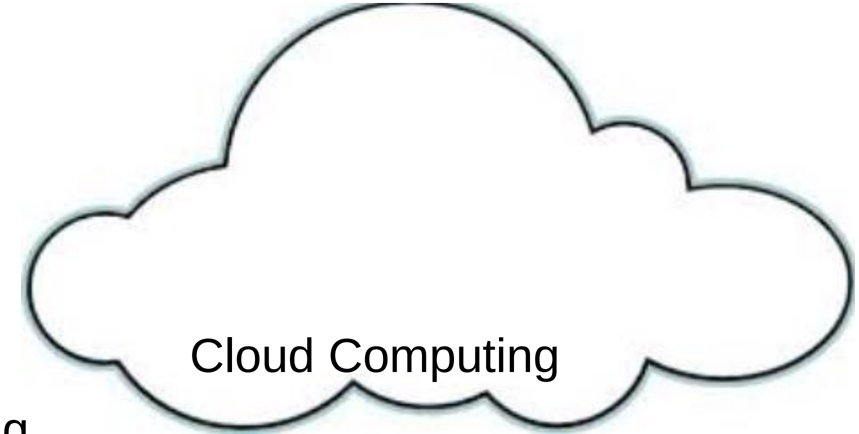
Cloud Computing and BigData



Cloud computing:

The National Institute of Standards and Technology's definition of cloud computing identifies "five essential characteristics":

1. On-demand self-service.
Demand resources anytime and anywhere (e.g. server's resources, network)

2. Broad network access. Capabilities are available over the network and accessed through thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations).

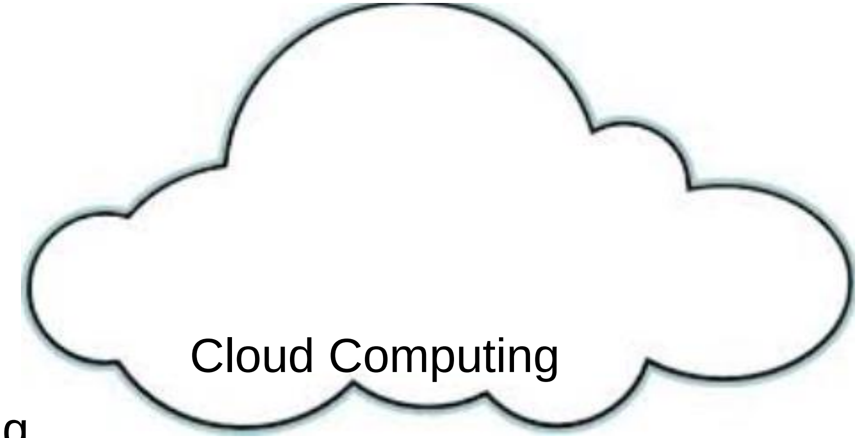
Cloud Computing and BigData

Cloud computing:

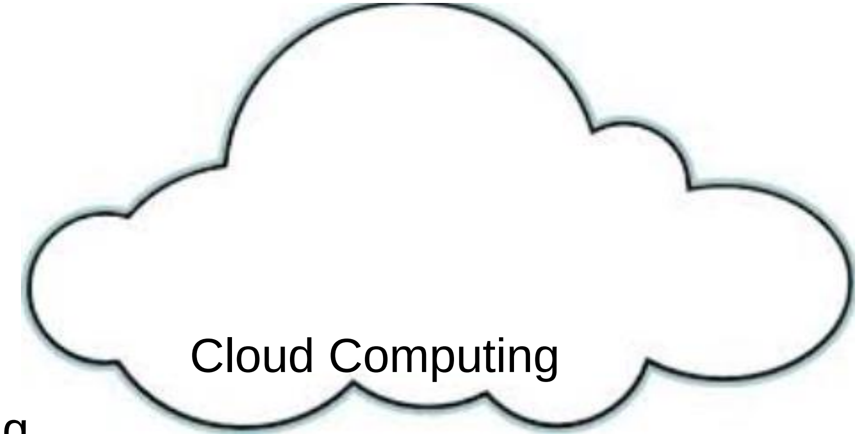
The National Institute of Standards and Technology's definition of cloud computing identifies "five essential characteristics":

3. Resource pooling. The provider's computing resources are pooled (grouped) to serve multiple consumers according to consumer demand.

4. Rapid elasticity. Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand



Cloud Computing and BigData

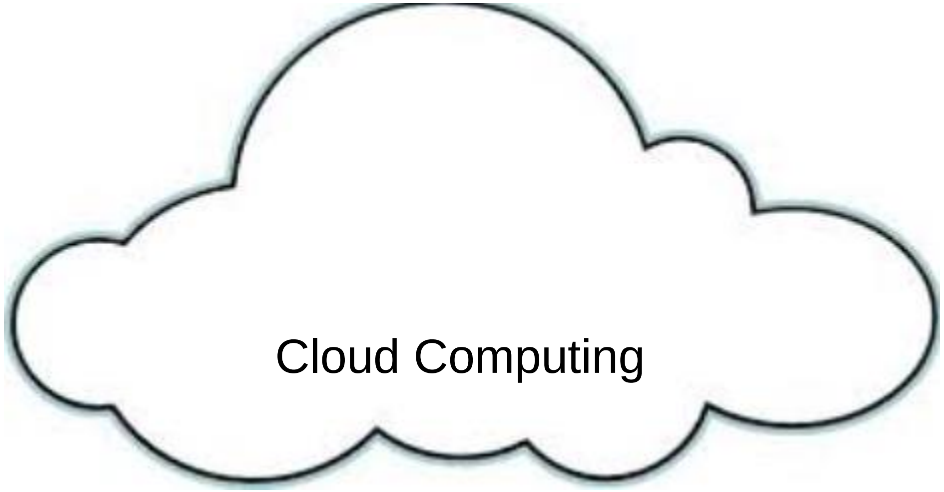


Cloud computing:

The [National Institute of Standards and Technology]'s definition of cloud computing identifies "five essential characteristics":

5. Measured service. Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts).

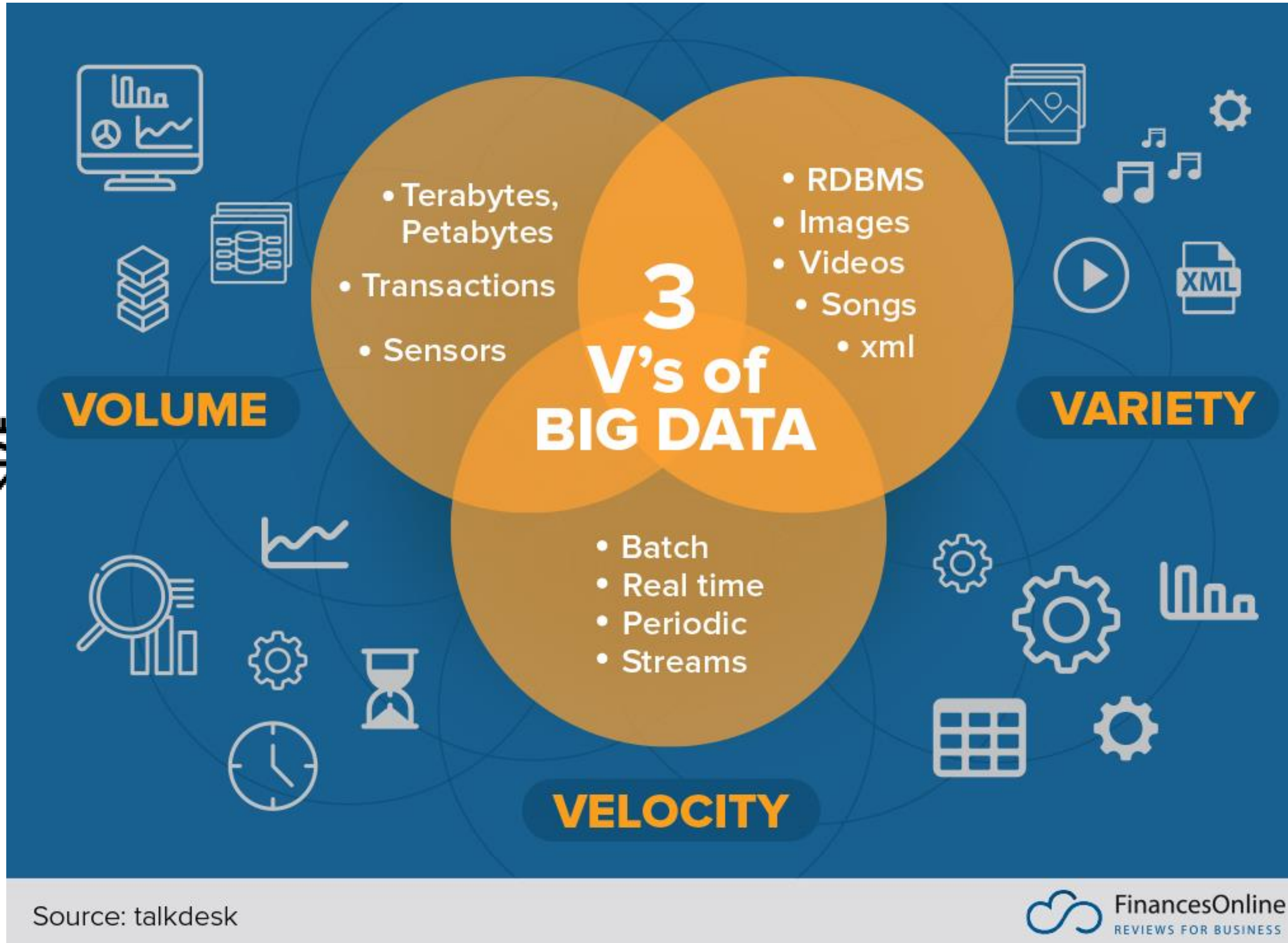
Cloud Computing and BigData



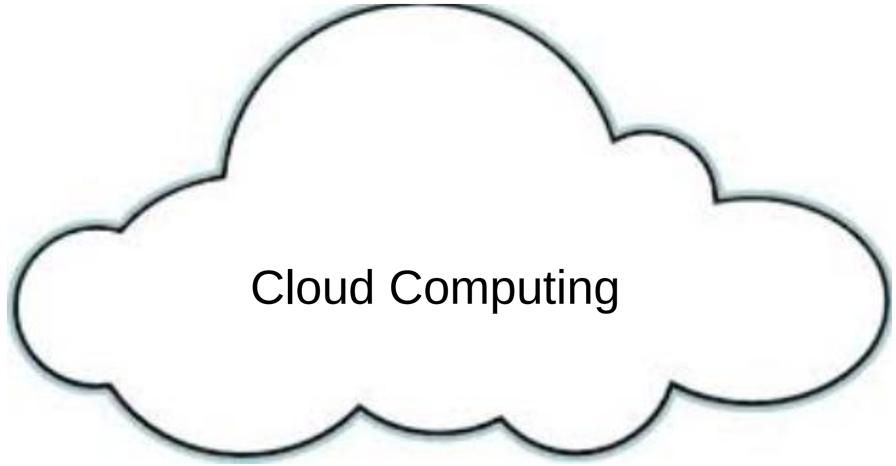
Big Data is a set of large-size data generated from different resources such as:

- Statistics
- Marketing
- Mobile applications
- IoT
- etc.


STORAGE
CONSUMER
COMPUTERS
MARKETING
SAMPLE
RESEARCH
BIG DATA
BYTES
BEHAVIOR
ANALYTICS
TECHNOLOGY
INFORMATION
INTERNET
SIZE


3 V's of BIG DATA
VOLUME
• Terabytes, Petabytes
• Transactions
• Sensors
VARIETY
• RDBMS
• Images
• Videos
• Songs
• xml
VELOCITY
• Batch
• Real time
• Periodic
• Streams
XML
Source: talkdesk
FinancesOnline
REVIEWS FOR BUSINESS

Cloud Computing and BigData



Introduction to Big Data Analytics

Objectives:

- Define big data
- Identify four business drivers for advanced analytics
- Distinguish the techniques for Business Intelligence from Data Science
- Describe the role of the Data Scientist within the new big data ecosystem
- Cite illustrative examples of big data opportunities.

What is *Big Data*?

What makes data, *“Big” Data*?

What's Big Data?

- **Big data***

  A collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

- Big Data challenges include

  capture, storage, search, sharing, transfer, analysis, and visualization.

from Wikipedia

Trend to Big Data

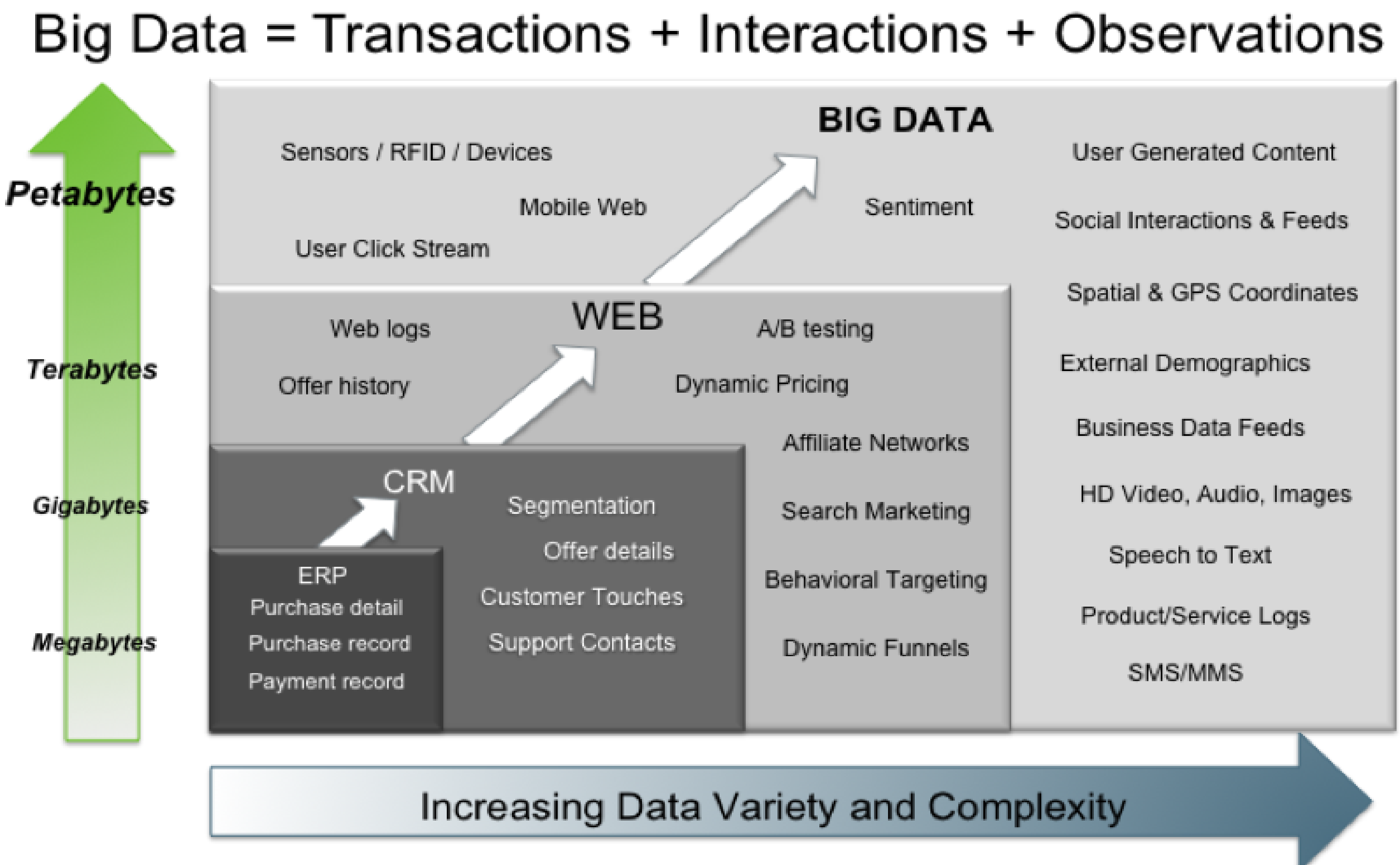
- The trend to big data is due to:

  Additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found like:

  - Prevent diseases
  - Spot business trends
  - Determine real-time roadway traffic conditions
  - Determine quality of research
  - Link legal citations

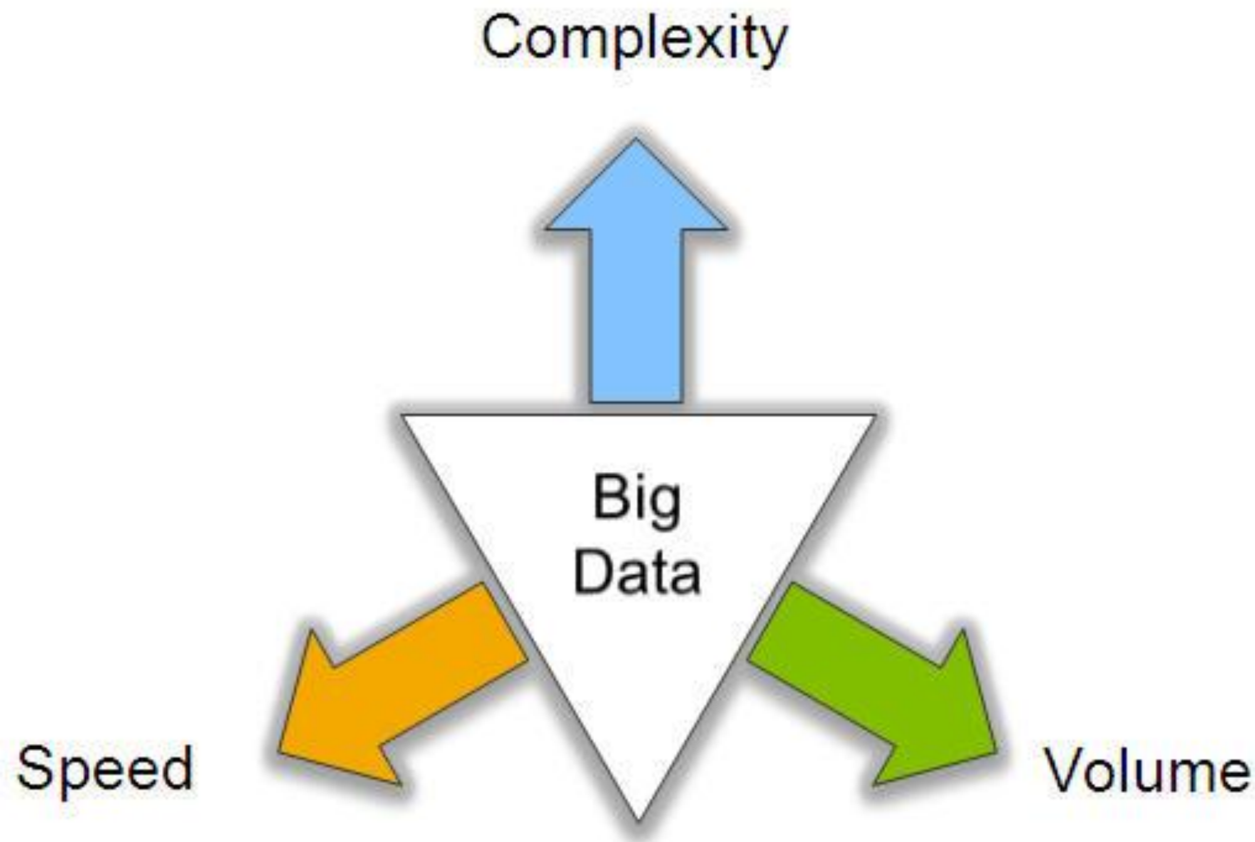
- The Big Data trend is generating an enormous amount of information that requires advanced analytics and new market players to take advantage of it.

Trend to Big Data



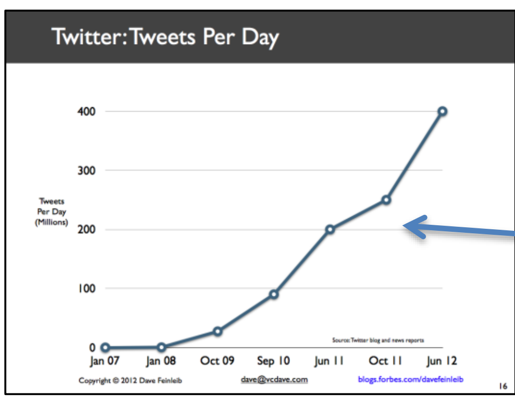
Source: *Contents of above graphic created in partnership with Teradata, Inc.*

Trend to Big Data

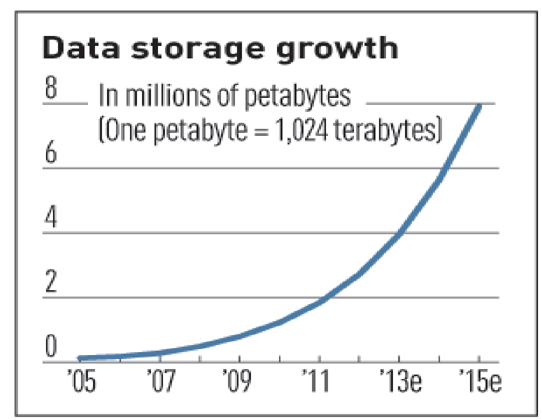


Big Data Volume

- Data volume (scale) is increasing exponentially
- Will be Increased from 0.8 zettabytes to 35zb in years 2009 to 2020. How many multiples?



Exponential increase in collected/generated data



Big Data Complexity

Knowledge can be extracted by linking together different types of data:

- Big Public Data (weather, finance, ..etc)
- Relational Data (tables, transaction, legacy Data)
- Text Data (web)
- Semi-structured Data (xml)
- Graphical Data
- Streaming Data

- Processing Complexity requirements:
  - Changing data structures
  - Use additional transformations and analytical techniques

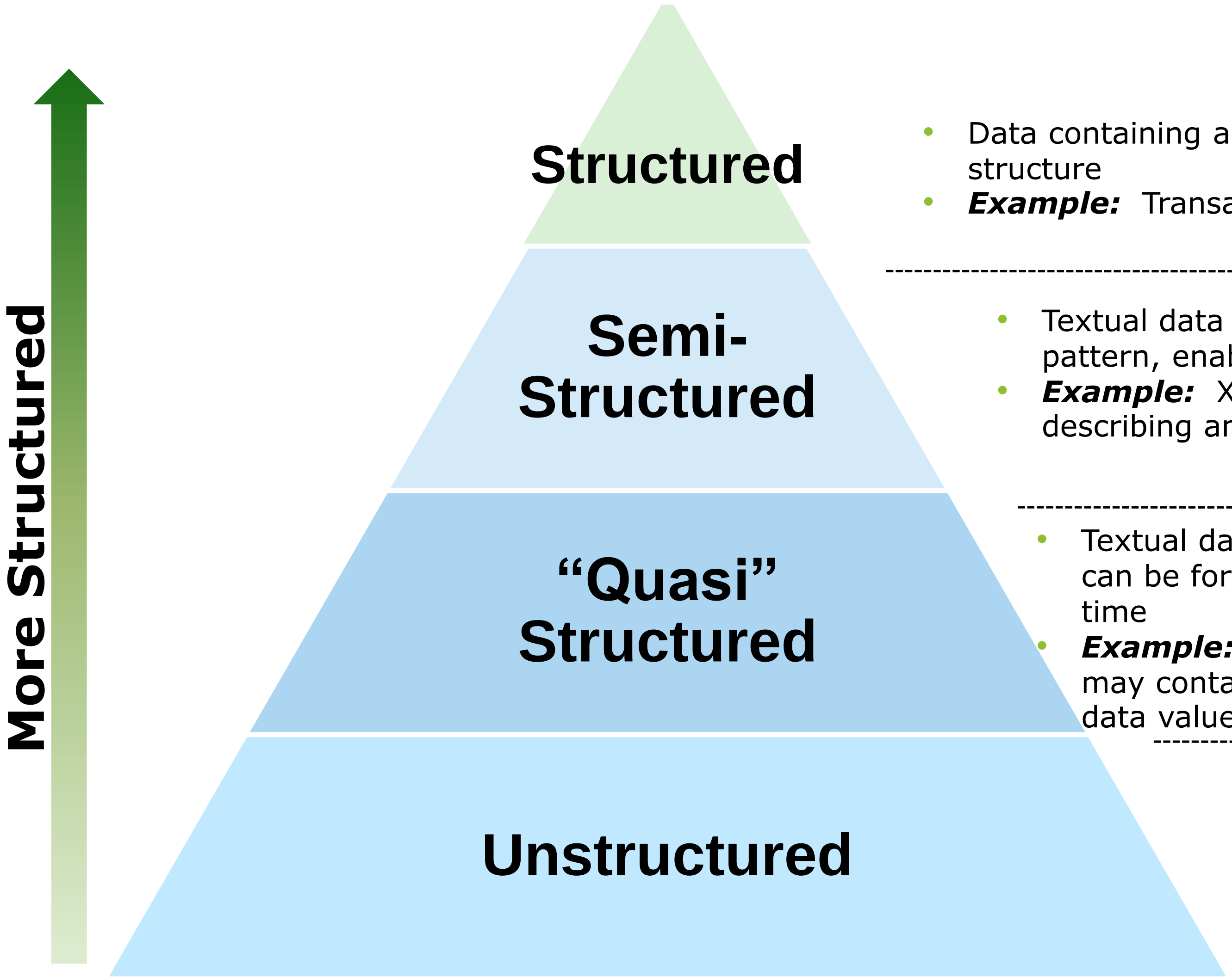
Big Data Processing Speed

- As data generating go fast, data processing need to be fast.
- Late decisions may cause missing opportunities
- Online Data Analytics needs fast processing
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like → send promotions of store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body → any abnormal measurements require immediate reaction

Big Data Type Structures

- The big data in nature is:
  - Structured
  - Semi-Structured
  - Quasi-Structured
  - Unstructured
- Greater variety of big data structures requires different techniques and tools to process and analyze.

Big Data Type Structure



- Data containing a defined data type, format, structure
- ***Example:*** Transaction data and OLAP

- Textual data files with a discernable pattern, enabling parsing
- ***Example:*** XML data files that are self describing and defined by an xml schema

---

- Textual data with erratic data formats, can be formatted with effort, tools, and time
- ***Example:*** Web clickstream data that may contain some inconsistencies in data values and formats

---

- Data that has no inherent structure and is usually stored as different types of files.
- ***Example:*** Text documents, PDFs, images and video

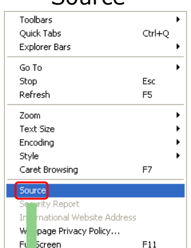
Big Data Type Structure

Structured Data

SUMMER FOOD SERVICE PROGRAM 1]				
(Data as of August 01, 2011)				
Fiscal Year	Number of Sites	Peak (July) Participation	Meals Served	Total Federal Expenditures 2]
	------------Thousands------------		--Mil.--	---Million $---
1969	1.2	99	2.2	0.3
1970	1.9	227	8.2	1.8
1971	3.2	569	29.0	8.2
1972	6.5	1,080	73.5	21.9
1973	11.2	1,437	65.4	26.6
1974	10.6	1,403	63.6	33.6
1975	12.0	1,785	84.3	50.3
1976	16.0	2,453	104.8	73.4
TQ 3]	22.4	3,455	198.0	88.9
1977	23.7	2,791	170.4	114.4
1978	22.4	2,333	120.3	100.3
1979	23.0	2,126	121.8	108.6
1980	21.6	1,922	108.2	110.1

Semi-Structured Data

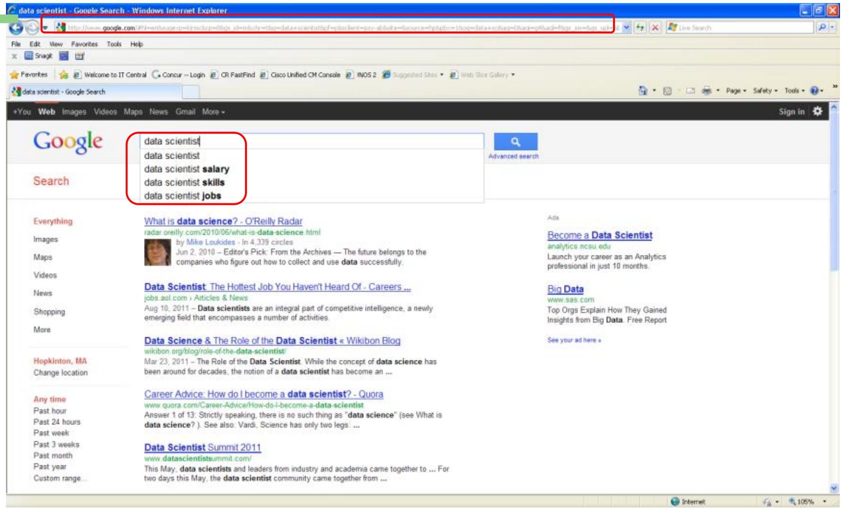
View → Source




```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-trans
<html xmlns="http://www.w3.org/1999/xhtml">

        <head>
                <meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />
                <META name="y_key" content="859b4020e1c9acec">
                        <link rel="canonical" href="http://www.emc.com/index.htm" />
                                <META NAME="verify-v1" CONTENT="yiZt9V0P4eV0jFdIPeVVIfRP32g4qtwFE0I2UvTMfSU
                <title>EMC - Data Recovery, Cloud Computing, and Storage Hardware</title>
                <META NAME="description" CONTENT="EMC is a leading provider of storage hardware solutions th
 data recovery and improve cloud computing." />
                        <META NAME="keywords" CONTENT="emc,network storage,data recovery,information manager
 software,nas storage,information protection,information management" />
                <!-- Start :stylehseet incldues -->
<link rel="stylesheet" href="/_admin/css/styles.css" />
<link rel="stylesheet" href="/_admin/css/styles_nav.css" />
<!--[if IE]>
```

Quasi-Structured Data



http://www.google.com/#hl=en&sugexp=kjrmc&cp=8&gs_id=2m&xhr=t&q=data+scientist&pq=big+data&pf=p&sclient=psyb&source=hp&pbx=1&oq=data+sci&aq=0&aqi=g4&aql=f&gs_sm=&gs_upl=&bav=on.2,or.r_gc.r_pw.,cf.osb&fp=d566e0fbd09c8604&biw=1382&bih=651

Unstructured Data

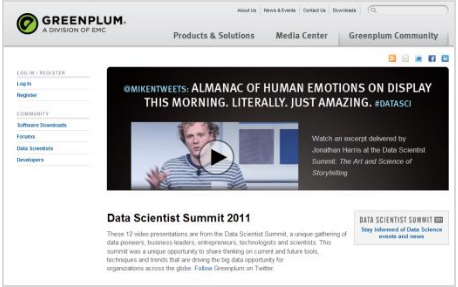
The Red Wheelbarrow, by William Carlos Williams

so much depends
upon

a red wheel
barrow

glazed with rain
water

beside the white
chickens.



Data Repositories from an Analyst Perspective

Data Islands "Spreadmarts"

Isolated data marts

- Spreadsheets and low-volume DB's for recordkeeping
- Analyst dependent on data extracts

Data Warehouses

Centralized data containers in a purpose-built space

- Supports BI and reporting, but restricts robust analyses
- Analyst dependent on IT & DBAs for data access and schema changes
- Analysts must spend significant time to get extracts from multiple sources

Analytic Sandbox

Data assets gathered from multiple sources and technologies for analysis

- Enables high performance analytics using in-db processing
- Reduces costs associated with data replication into "shadow" file systems
- *"Analyst-owned" rather than "DBA owned"*

Business Drivers for Analytics

Current Business Problems Provide Opportunities for Organizations to Become More Analytical & Data Driven

	Driver	Examples
1	Desire to optimize business operations	Sales, pricing, profit, efficiency
2	Desire to identify business risks	Customer agitate, fraud
3	Predict new business opportunities	Upsell, cross-sell, best new customer prospects
4	Comply with laws or regulatory requirements	Anti-Money Laundering, Fair Lending

Business Drivers Analytical Approaches

Business Intelligence vs. Data Science

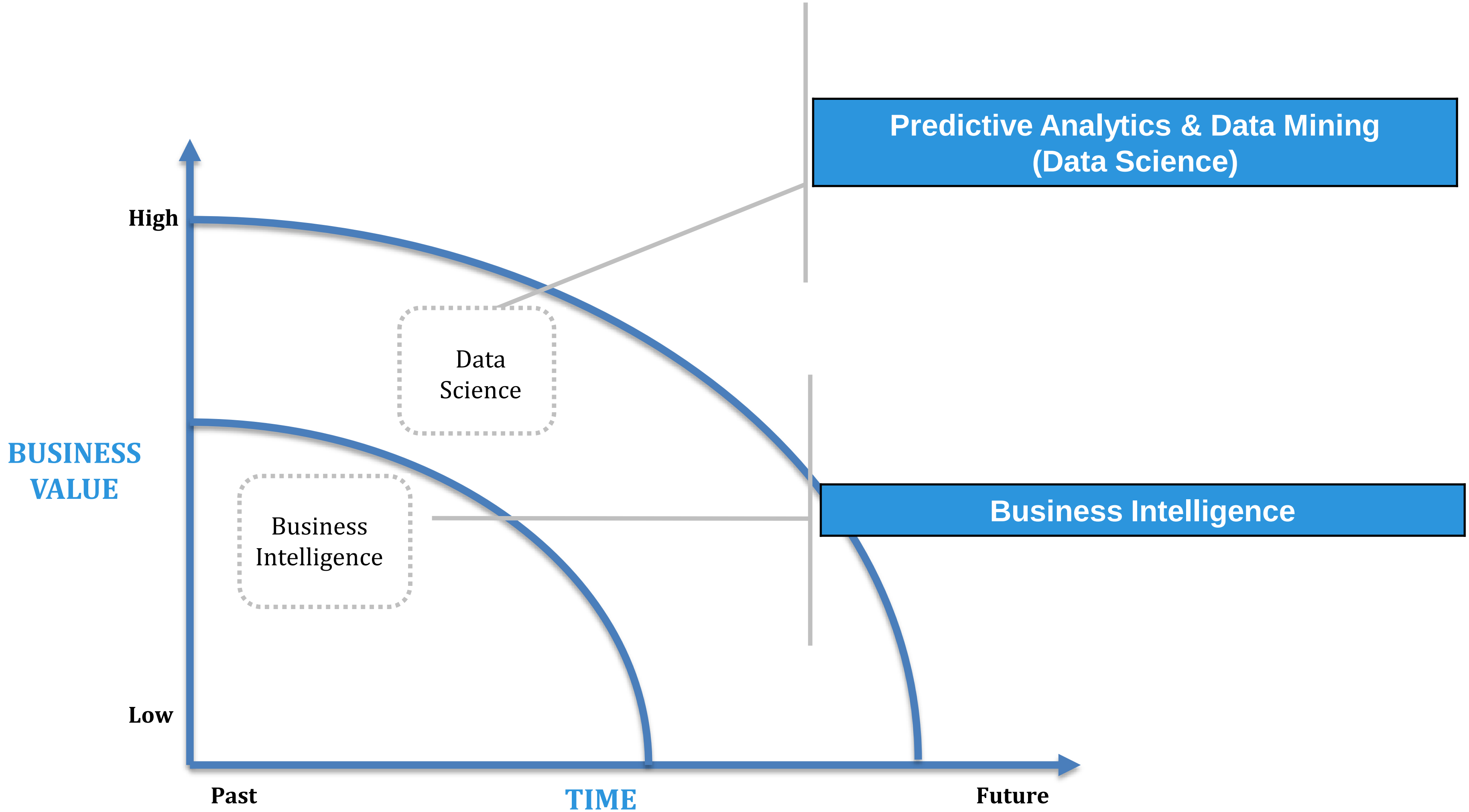
Business Intelligence	
Typical Techniques & Data Types	• Standard and immediate reporting, dashboards, alerts, queries, details on demand • Structured data, traditional sources, manageable data sets
Common Questions	• What happened last quarter? • How many did we sell? • Where is the problem? In which situations?

Predictive Analytics & Data Mining (Data Science)	
• **Typical Techniques & Data Types**	• Optimization, predictive modeling, forecasting, statistical analysis • Structured/unstructured data, many types of sources, very large data sets
• **Common Questions**	• What if…..? • What's the optimal scenario for our business ? • What will happen next? What if these trends continue? Why is this happening?

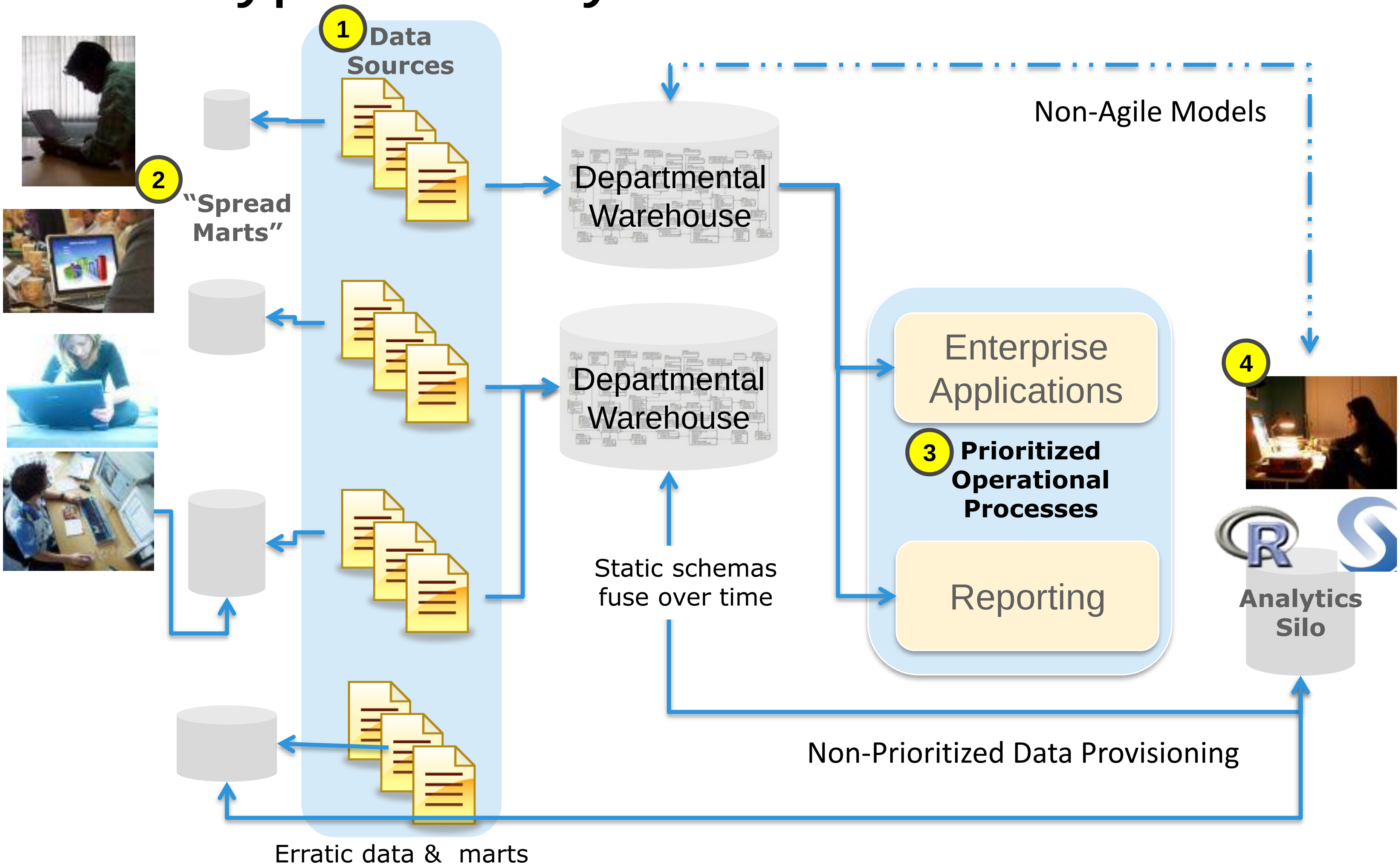
EMC² PROVEN PROFESSIONAL

Business Drivers Analytical Approaches

Business Intelligence vs. Data Science



A Typical Analytical Architecture



The graphic shows a typical data warehouse and some of the challenges that it presents.

For source data (1) to be loaded into the EDW, data needs to be well understood, structured and normalized with the appropriate data type definitions. While this kind of centralization enables organizations to enjoy the benefits of security, backup and failover of highly critical data, it also means that **data must go through significant pre-processing and checkpoints before it can enter this sort of controlled environment, which does not lend itself to data exploration and iterative analytics.**

(2) As a result of this level of control on the EDW, shadow systems emerge in the form of departmental warehouses and local data marts that business users create to accommodate their need for flexible analysis.

These local data marts do not have the same constraints for security and structure as the EDW does, and allow users across the enterprise to do some level of analysis.

However, these one-off systems reside in isolation, often are not networked or connected to other data stores, and are generally not backed up.

(3) Once in the data warehouse, data is fed to enterprise applications for business intelligence and reporting purposes.

These are high priority operational processes getting critical data feeds from the EDW.

(4) At the end of this work flow, analysts get data provisioned for their downstream analytics.

Since users cannot run custom or intensive analytics on production databases, analysts create data extracts from the EDW to analyze offline in R or other local analytical tools.

Many times these tools are limited to in-memory analytics with desktops analyzing samples of data, rather than the entire population of a data set.

Because these analyses are based on data extracts, they live in a separate location and the results of the analysis – and any insights on the quality of the data or anomalies, rarely are fed back into the main EDW repository.

Lastly, because data slowly accumulates in the EDW due to the rigorous validation and data structuring process, **data is slow to move into the EDW and the schema is slow to change**.

EDWs may have been originally designed for a specific purpose and set of business needs, but over time evolves to house more and more data and enables business intelligence and the creation of OLAP (Online Analytical Processing) cubes for analysis and reporting.

The EDWs provide limited means to accomplish these goals, achieving the objective of reporting, and sometimes the creation of dashboards, but generally limiting the ability of analysts to iterate on the data in an separate environment from the production environment where they can conduct in-depth analytics, or perform analysis on unstructured data.

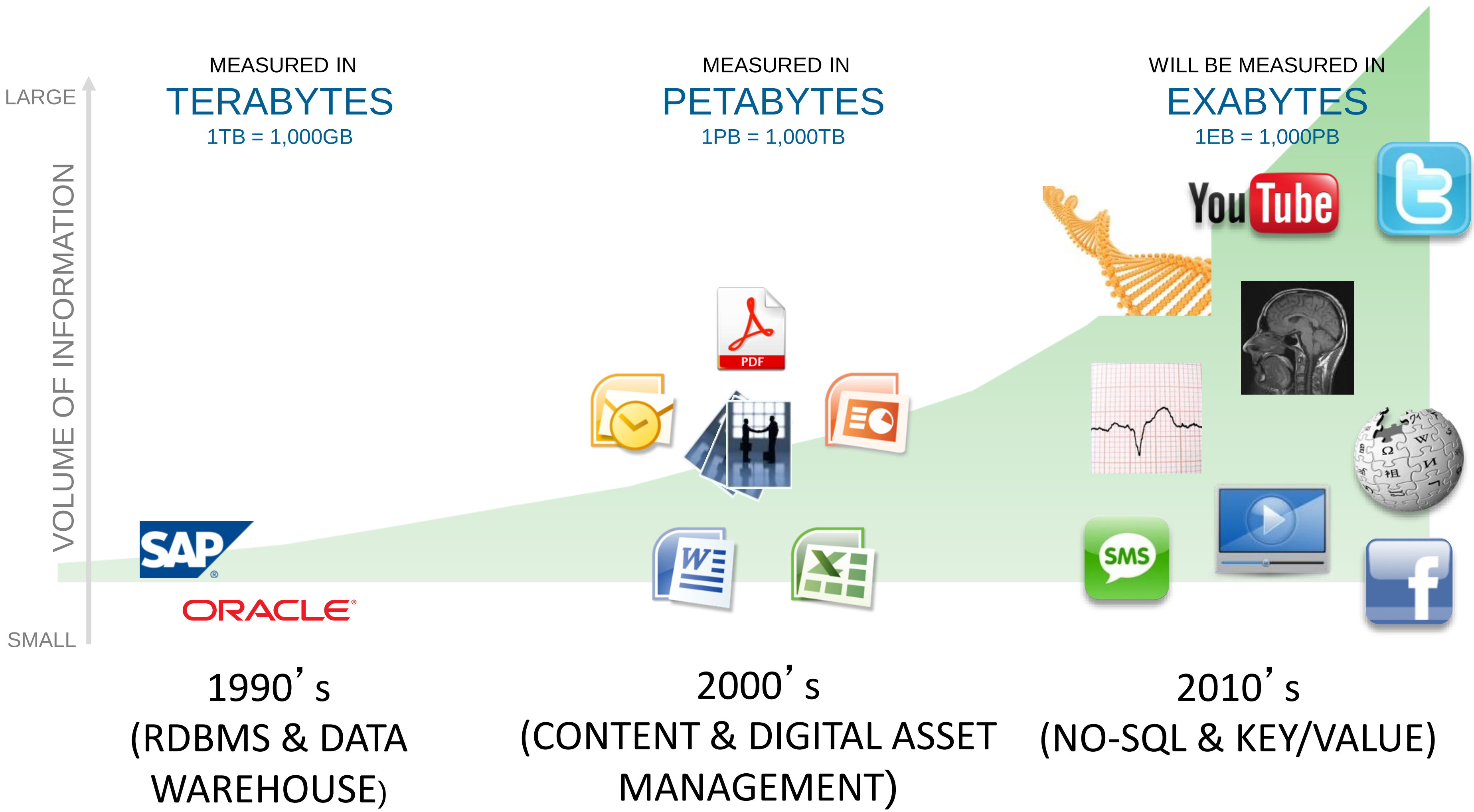
Implications of Typical Data Architecture

- High-value data is hard to reach and leverage
- Predictive analytics & data mining activities are last in line for data
  - Queued after prioritized operational processes
- Data is moving in batches from EDW to local analytical tools
  - In-memory analytics (such as R, SAS, SPSS, Excel)
  - Sampling can skew model accuracy
- Isolated, *ad hoc* analytic projects, rather than centrally-managed harnessing of analytics
  - Non-standardized initiatives
  - Frequently, not aligned with corporate business goals

Slow "time-to-insight" & reduced business impact

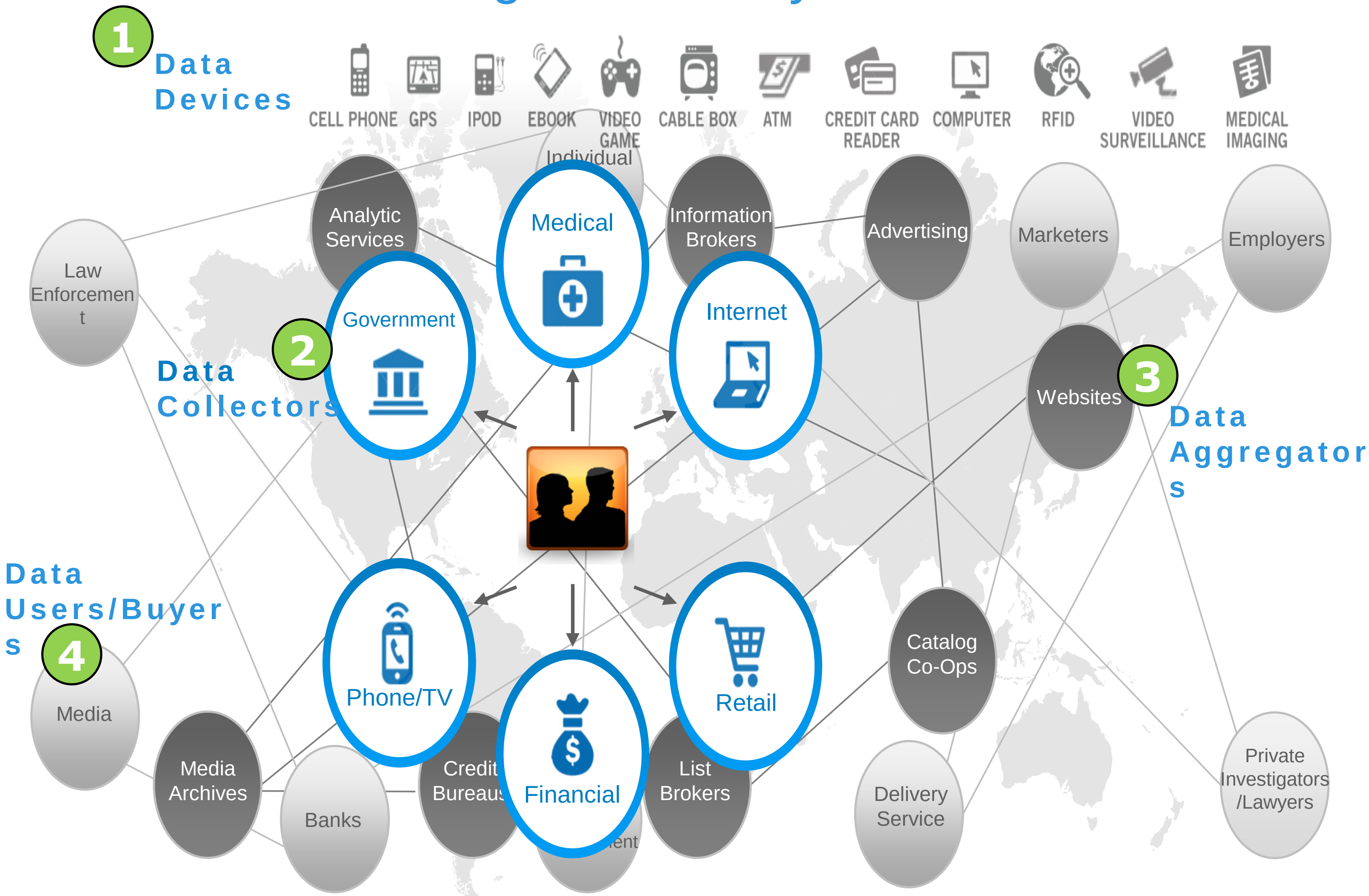
Opportunities for a New Approach to Analytics

New Applications Driving Data Volume



Opportunities for a New Approach to Analytics

Big Data Ecosystem



Considerations for Big Data Analytics

Criteria for Big Data Projects

1. Speed of decision making
2. Throughput
3. Analysis flexibility

New Analytic Architecture

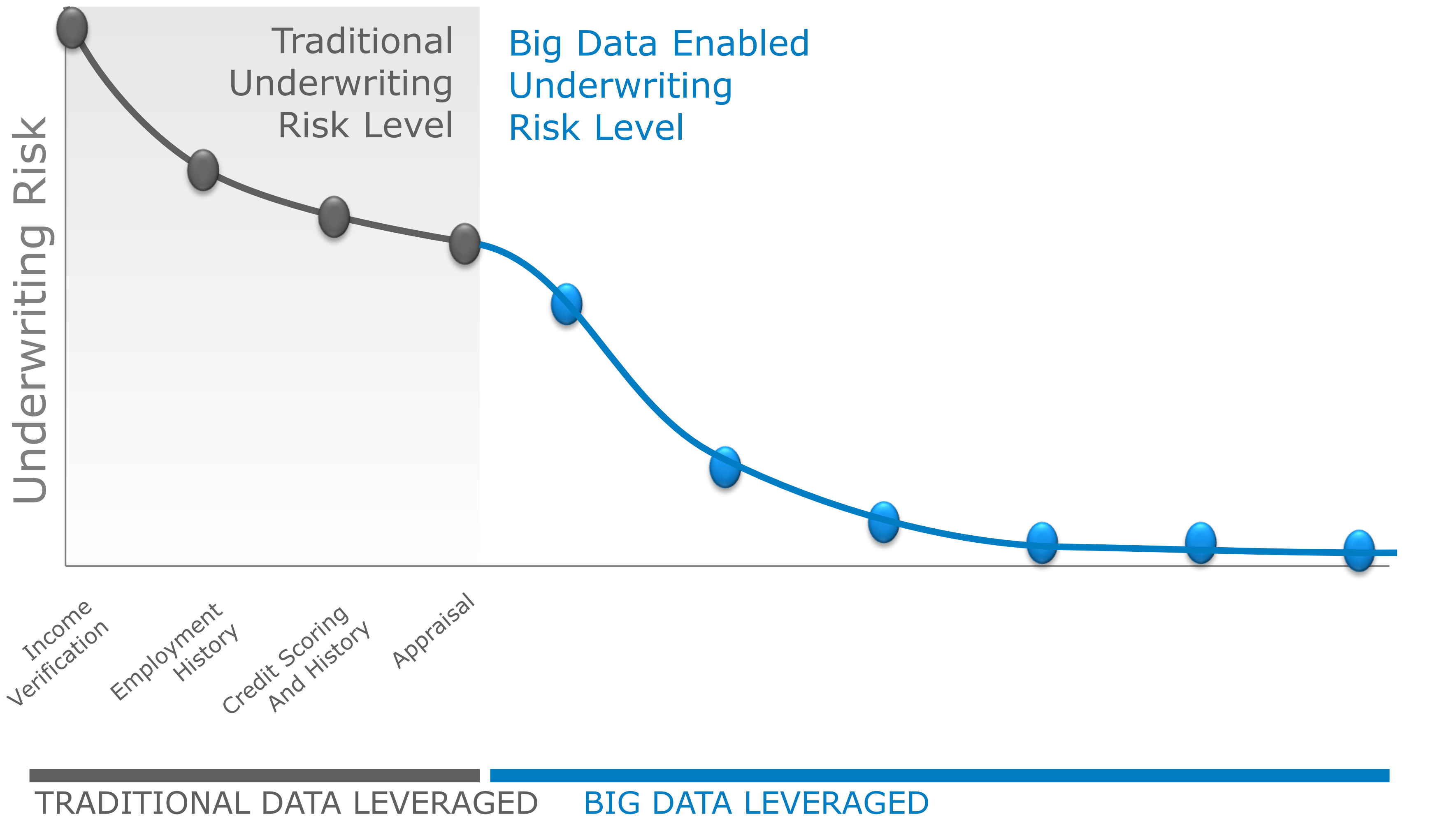
Analytic Sandbox

Data assets gathered from multiple sources and technologies for analysis

- Enables high performance analytics using in-db processing
- Reduces costs associated with data replication into "shadow" file systems
- “Analyst-owned” rather than “DBA owned”

Practice in Analytics - Case Study

Exercise – Big Data sets in Loan Processing Improvement



EMC² PROVEN PROFESSIONAL

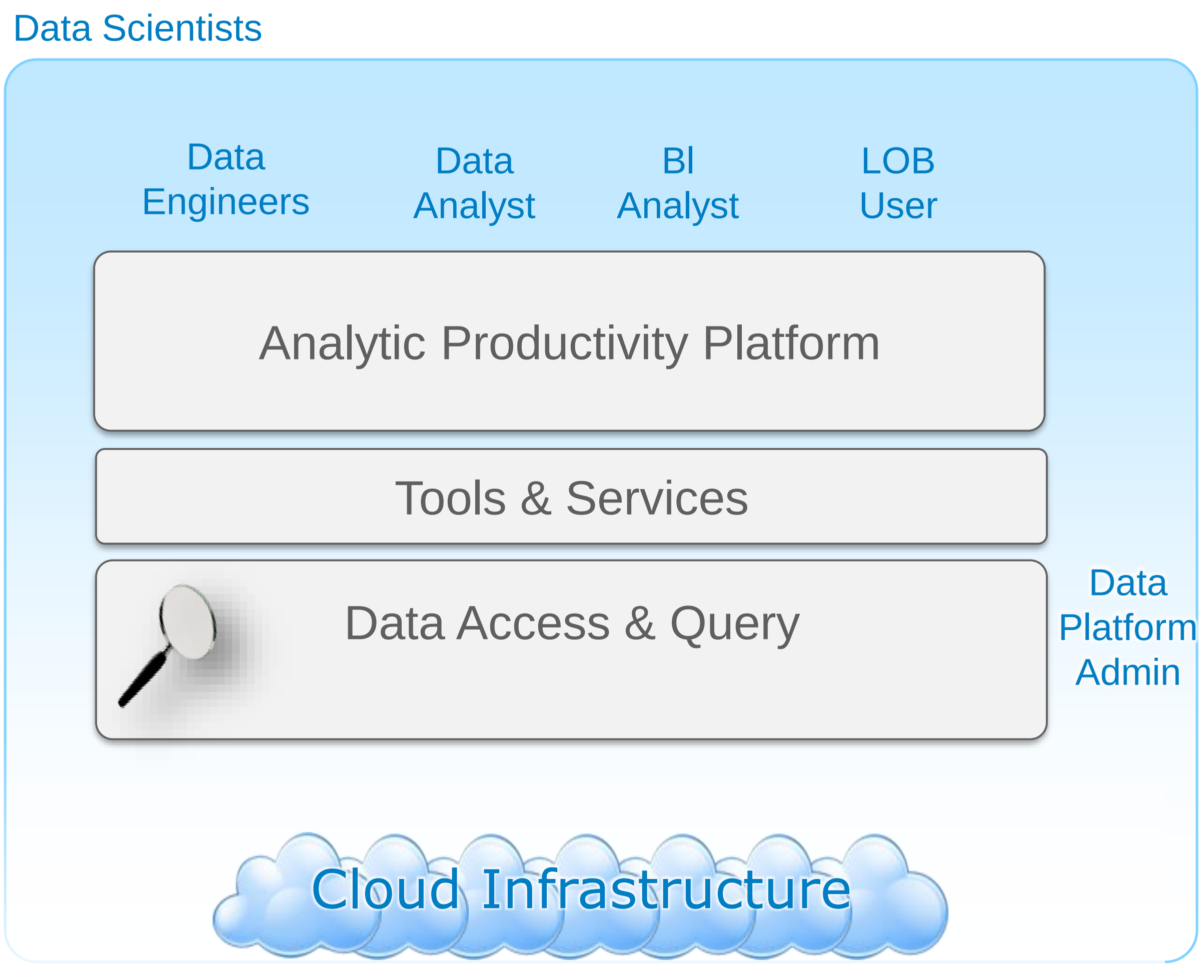
Key Roles of the New Data Ecosystem

Role	Role Description
Deep Analytical Talent (Data Scientist**)**	People with advanced training in quantitative disciplines, such as mathematics, statistics, and machine learning.
Data Savvy Professionals	People with a basic knowledge of statistics and/or machine learning, who can define key questions that can be answered using advanced analytics
Technology & Data Enablers	People providing technical expertise to support analytical projects. Skills sets including computer programming and database administration

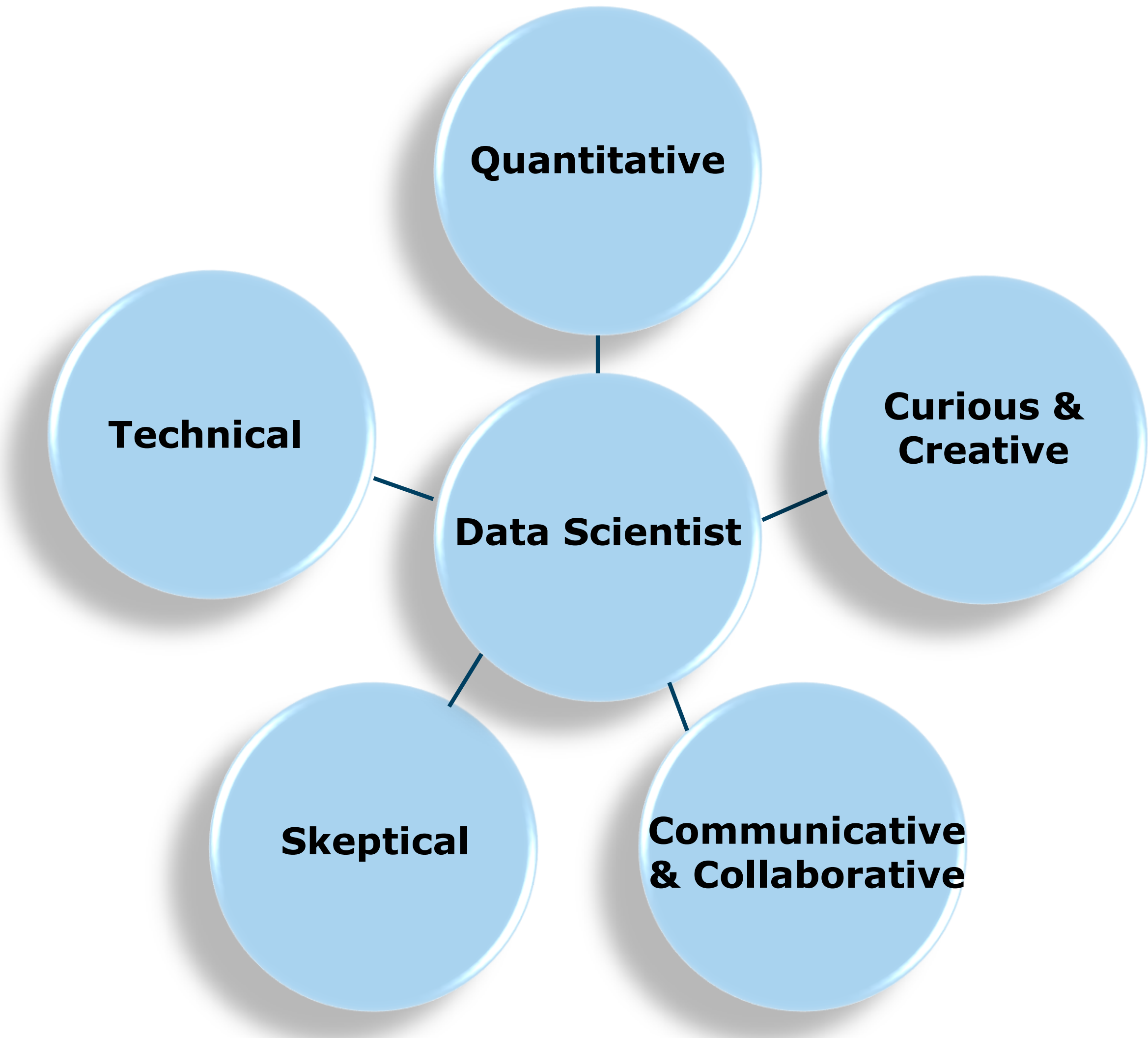
Data Scientist *Key Activities* for Analytical Projects

Key Activities

- Reframe business challenges as analytics challenges
- Design, implement and deploy statistical models and data mining techniques on big data
- Create insights that lead to actionable recommendations



Profile of a Data Scientist



Big Data Analytics - Industry Examples

1 Health Care
- Reducing Cost of Care

2 Public Services
- Preventing Pandemics

3 Life Sciences
- Genomic Mapping

4 IT Infrastructure
- Unstructured Data Analysis

5 Online Services
- Social Media for Professionals

1

Big Data Analytics: *Healthcare*

Situation

- Poor police response and problems with medical care, triggered by shooting student
- The event drove local doctor to map crime data and examine local health care

Use of Big Data

- The doctor generated his own crime maps from medical billing records of 3 hospitals

Key Outcomes

- City hospitals & ER's provided expensive, low quality care
- Reduced hospital costs by 56% by realizing that 80% of city's medical costs came from 13% of its residents, mainly low-income or elderly
- Now offers preventative care over the phone or through home visits

2 Big Data Analytics: *Public Services*

Situation

- Threat of global pandemics has increased exponentially
- Pandemics spreads at faster rates, more resistant to antibiotics

Use of Big Data

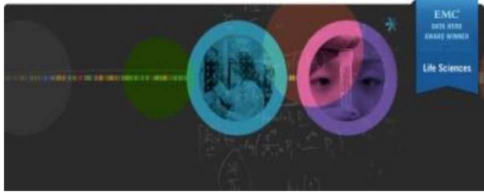
- Created a network of viral listening posts
- Combines data from viral discovery in the field, research in disease hotspots, and social media trends
- Using Big Data to make accurate predications on spread of new pandemics

Key Outcomes

- Identified a fifth form of human malaria, including its origin
- Identified why efforts failed to control swine flu
- Proposing more proactive approaches to preventing outbreaks

3

Big Data Analytics: *Life Sciences*



Situation

- Broad Institute (MIT & Harvard) mapping the Human Genome

Use of Big Data

- In 13 yrs, mapped 3 billion genetic base pairs; 8 petabytes
- Developed 30+ software packages, now shared publicly, along with the genomic data

Key Outcomes

- Using genetic mappings to identify cellular mutations causing cancer and other serious diseases
- Innovating how genomic research informs new pharmaceutical drugs

4 Big Data Analytics: *IT Infrastructure*

Situation

- Explosion of unstructured data required new technology to analyze quickly, and efficiently

Use of Big Data

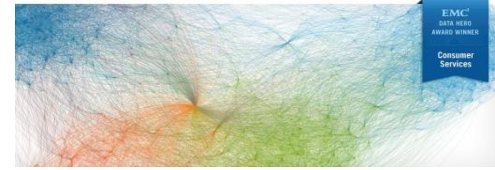
- Doug Cutting created Hadoop to divide large processing tasks into smaller tasks across many computers
- Analyzes social media data generated by hundreds of thousands of users

Key Outcomes

- New York Times used Hadoop to transform its entire public archive, from 1851 to 1922, into 11 million PDF files in 24 hrs
- Applications range from social media, sentiment analysis, wartime chatter, natural language processing

5

Big Data Analytics: *Online Services*



Situation

- Opportunity to create social media space for professionals

Use of Big Data

- Collects and analyzes data from over 100 million users
- Adding 1 million new users per week

Key Outcomes

- LinkedIn Skills, InMaps, Job Recommendations, Recruiting
- Established a diverse data scientist group, as founder believes this is the start of Big Data revolution