CSCI446/946 Big Data Analytics

Week 1 Introduction to Big Data Analytics

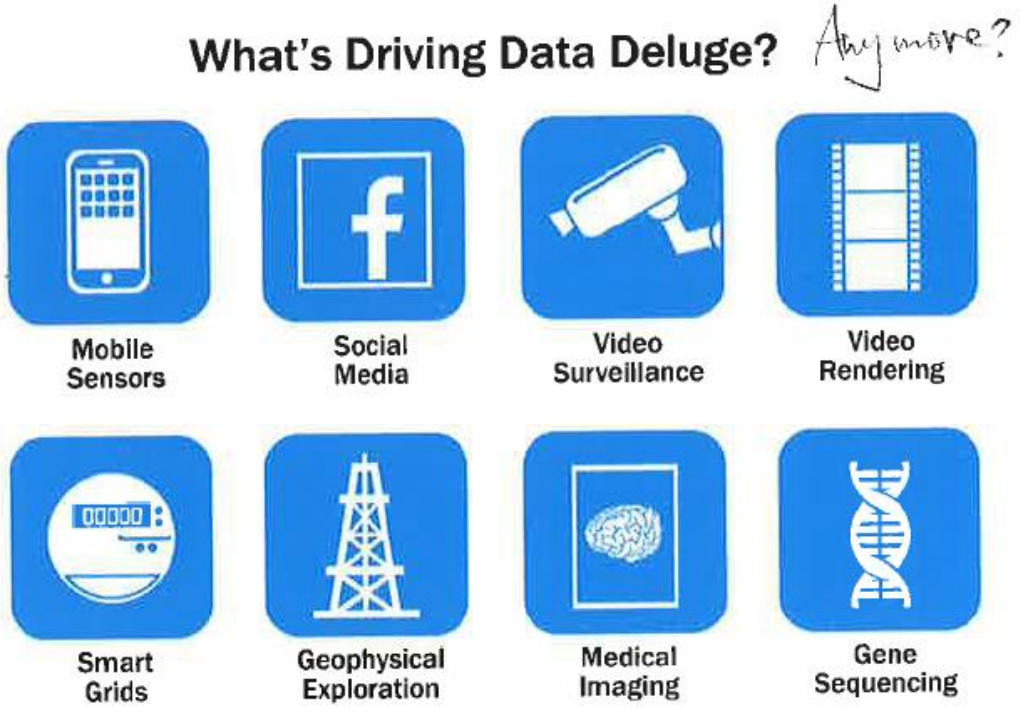
School of Computing and Information Technology
University of Wollongong Australia

Introduction to Big Data Analytics

- Big Data Overview
- State of the practice in Analytics
- Key Roles for the New Big Data Ecosystem
- Examples of Big Data Analytics
  - See more details in Chapter 1 of *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*, EMC Education Services (Editor)

Big Data Overview

- What's your idea on Big Data?
- What's driving data deluge?
  - Can you name a source of big data?



Big Data Overview

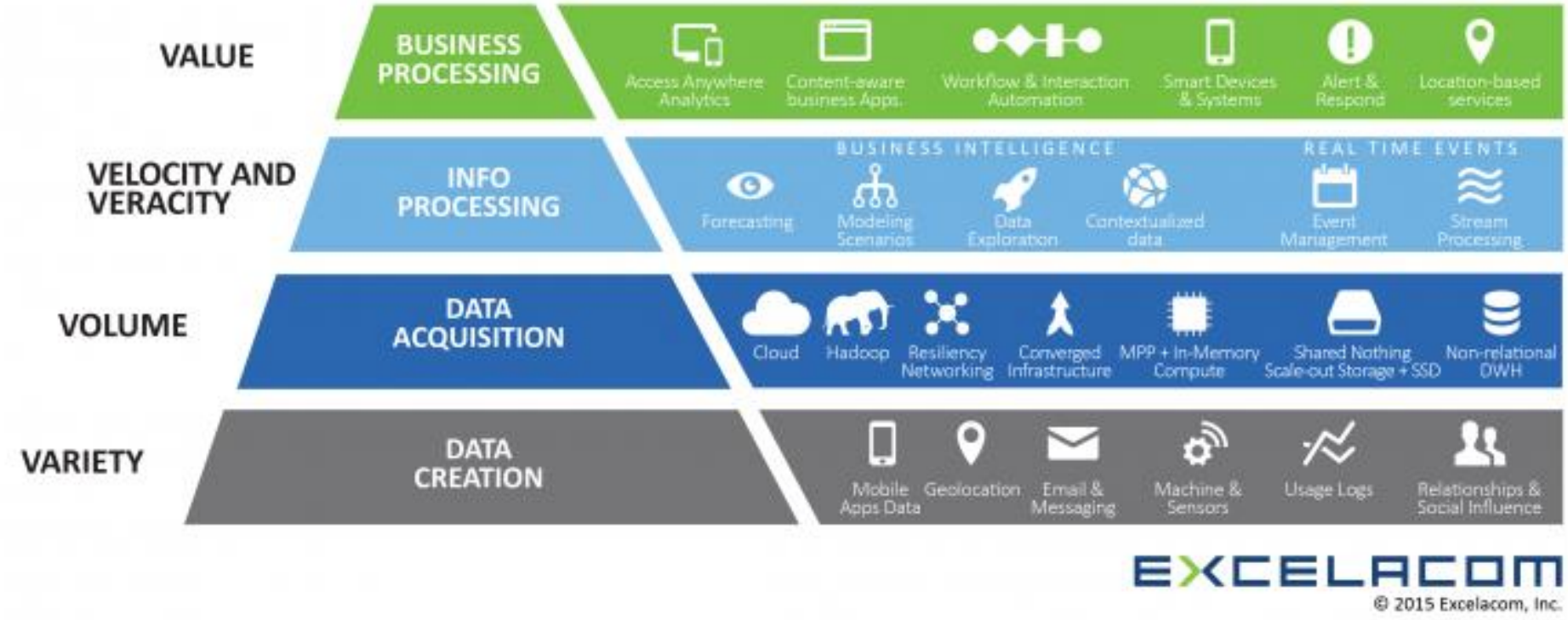
- Keeping up with this high influx of data is difficult
- Analysing vast amounts of data is more challenging, especially when the data does not conform to traditional structure
- Can you name any real applications of Big Data Analytics you have been aware of?

Big Data Overview

- Three attributes defining Big Data
  - Huge volume of data (billions x millions)
  - Complexity of data types and structures
  - Speed of new data creation and growth
- 3V: Volume, Variety, and Velocity
- 5V: 3V + Value and Veracity (accuracy, truthfulness)

Big Data Overview

- 5V: Volume, Variety, Velocity, Value and Veracity



From http://www.excelacom.com/resources/blog/the-5-vs-of-big-data-predictions-for-2016

Big Data Overview

- So, Big data analysis needs new tools and technologies

- *Big Data is data whose scale, distribution, diversity, and/or timeliness require the use of **new** technical architectures and analytics to enable insights that unlock **new** source of business value*

  – McKinsey & Co.; Big Data: The Next Frontier for Innovation, Competition, and Productivity, 2011

Big Data Overview

- This implies the need of
  - New data architectures
  - New analytic sandboxes
  - New tools
  - New analytical methods
  - An integration of multiple skills
  - New role of data scientist

Big Data Overview

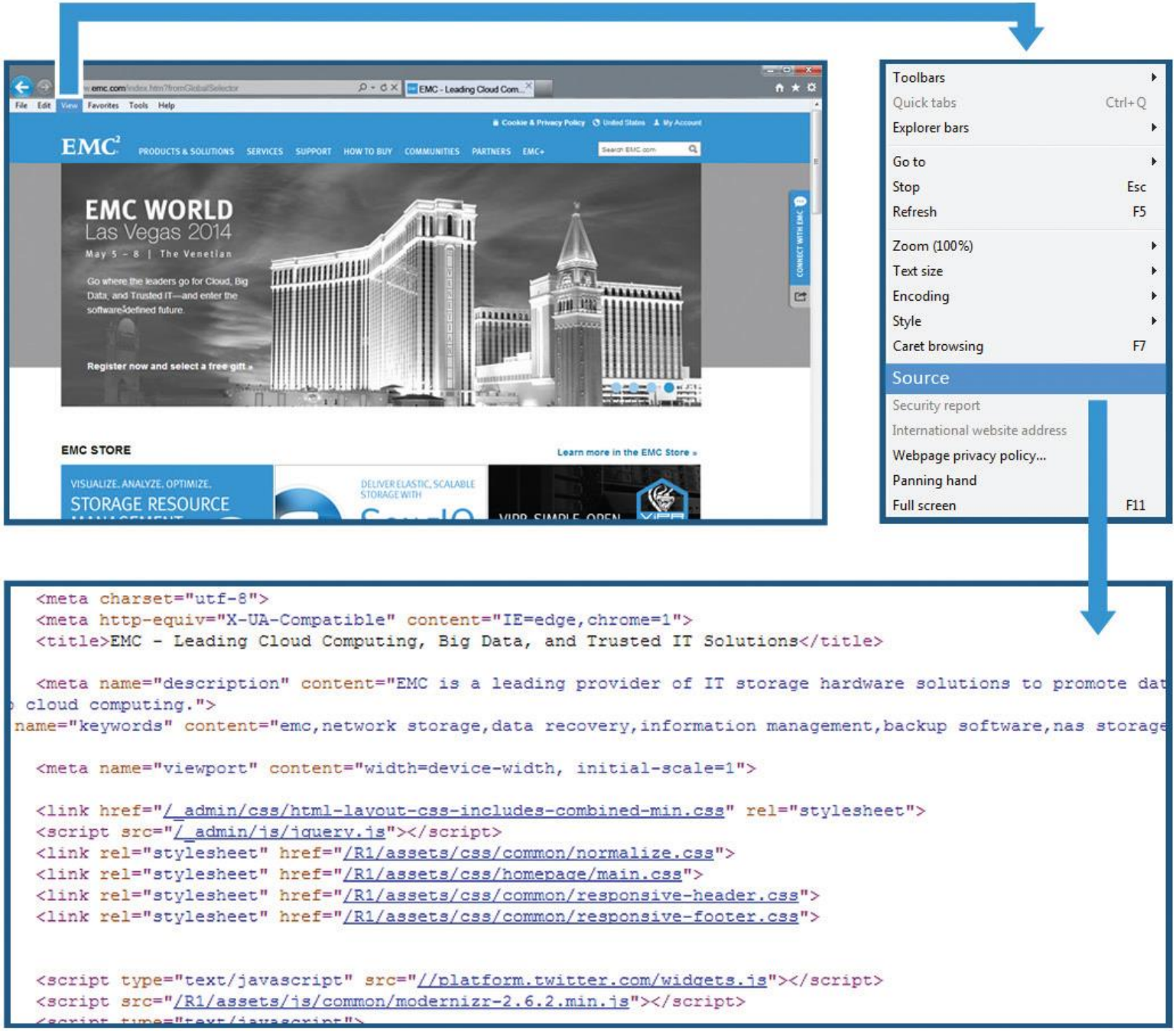
- Data Structures
- Structured data
  - Can you name some examples?

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
1981	20.6	1,726	90.3	105.9
1982	14.4	1,397	68.2	87.1
1983	14.9	1,401	71.3	93.4
1984	15.1	1,422	73.8	96.2
1985	16.0	1,462	77.2	111.5
1986	16.1	1,509	77.1	114.7
1987	16.9	1,560	79.9	129.3
1988	17.2	1,577	80.3	133.3
1989	18.5	1,652	86.0	143.8
1990	19.2	1,692	91.2	163.3

Structured

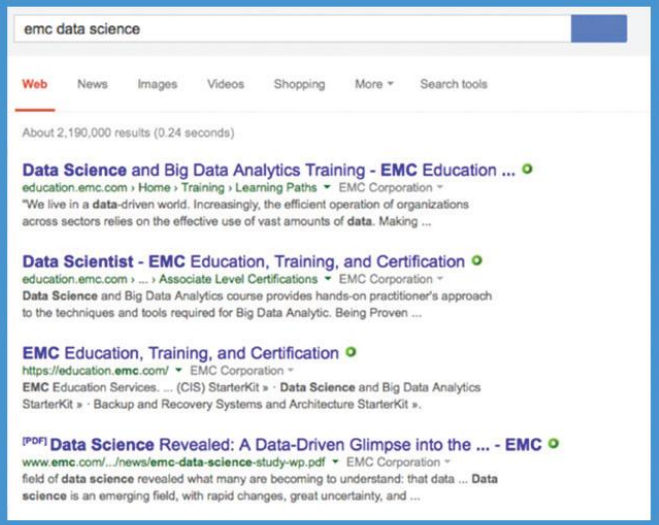
Big Data Overview

- Data Structures
- Structured data
  - Can you name some examples?
- Non-structured data (80-90% of data growth)
  - Semi-structured (XML data file)
  - Quasi-structured (Web clickstream data)
  - Unstructured (text documents, images, videos)
  - Can you name some examples?



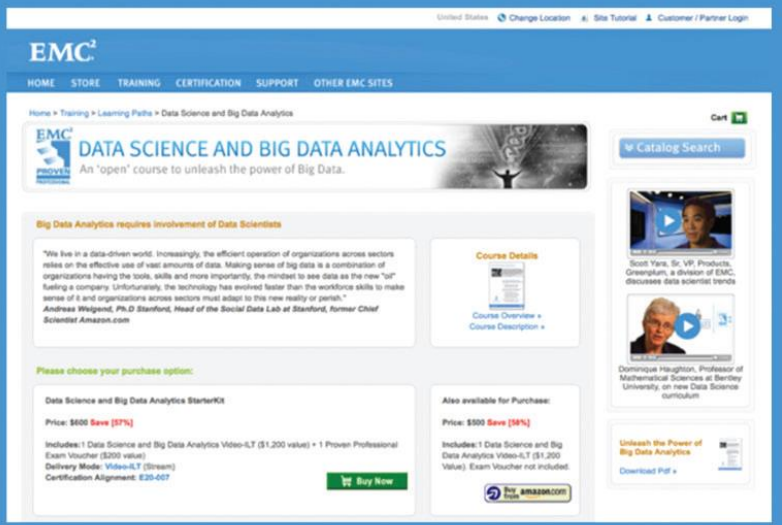
Semi-structured

1



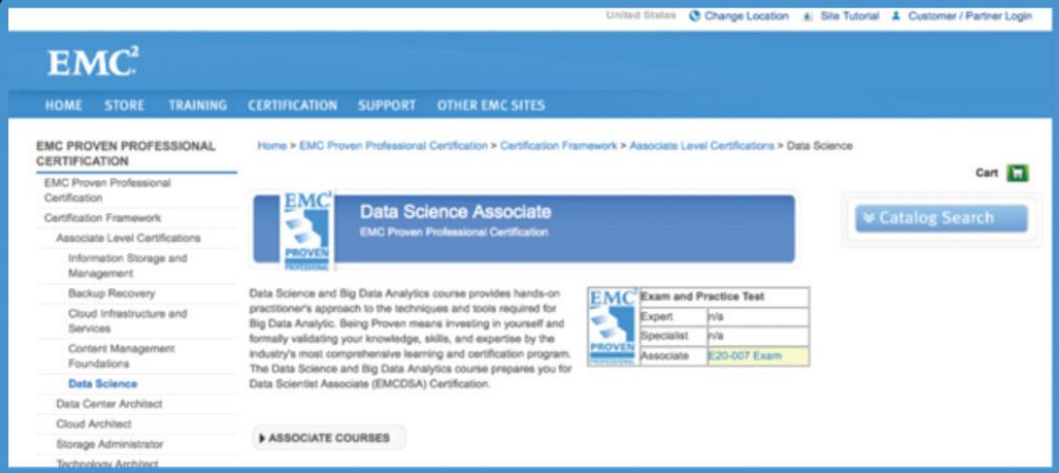
https://www.google.com/#q=EMC+data+science

2



https://education.emc.com/guest/campaign/data_science.aspx

3

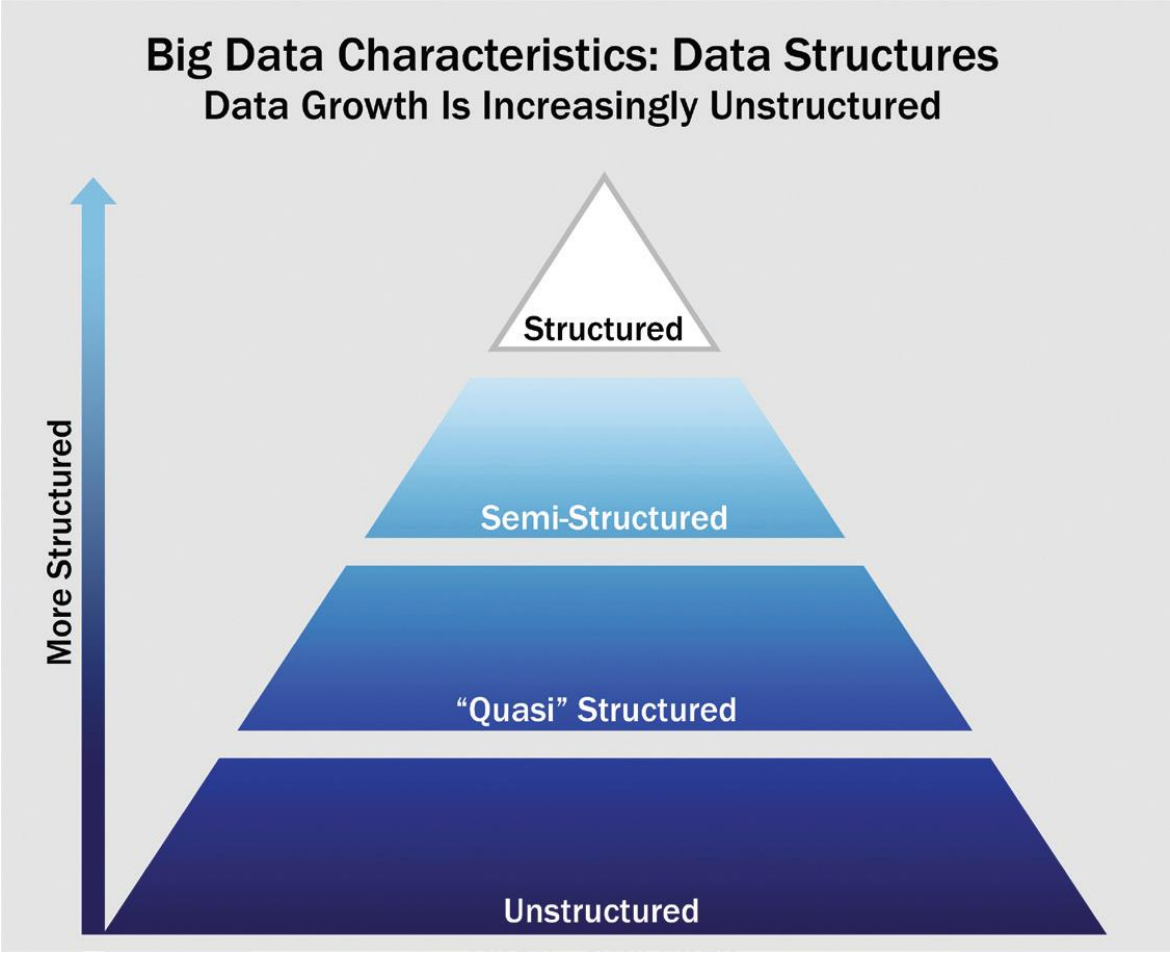


https://education.emc.com/guest/certification/framework/stf/data_science.aspx

Quasi-structured

Big Data Overview

- Data Structures



Big Data Overview

- Analyst Perspective on Data Repositories
  - Data accuracy and availability
  - Flexibility and agility of analysis

Big Data Overview

- Types of data repositories

Data Repository	Characteristics
Spreadsheets and data marts ("spreadmarts")	Spreadsheets and low-volume databases for recordkeeping Analyst depends 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 and DBAs for data access and schema changes Analysts must spend significant time to get aggregated and disaggregated data extracts from multiple sources.
Analytic Sandbox (workspaces)	Data assets gathered from multiple sources and technologies for analysis Enables flexible, high-performance analysis in a nonproduction environment; can leverage in-database processing Reduces costs and risks associated with data replication into "shadow" file systems "Analyst owned" rather than "DBA owned"

Big Data Overview

- Analyst Perspective on Data Repositories
  - Data accuracy and availability
  - Flexibility and agility of analysis
- Types of data repositories
  - Spreadsheets and data marts
  - Data Warehouses
  - Analytics Sandbox (workspaces)
- Approach shall fit with the desired goals

State of the Practice in Analytics

- Business drivers for Advanced Analytics

Business Driver	Examples
Optimize business operations	Sales, pricing, profitability, efficiency
Identify business risk	Customer churn, fraud, default
Predict new business opportunities	Upsell, cross-sell, best new customer prospects
Comply with laws or regulatory requirements	Anti-Money Laundering, Fair Lending, Basel II-III, Sarbanes-Oxley (SOX)

https://au.linkedin.com/jobs/view/head-of-ai-machine-learning-at-woolworths-supermarkets-746700893

State of the Practice in Analytics

- Business drivers for Advanced Analytics
  - Optimise business operations
  - Identify business risk
  - Predict new business opportunities
  - Comply with laws or regulatory requirements
- Leverage advanced analytics to create competitive advantage
- Advanced analytical techniques + Big Data

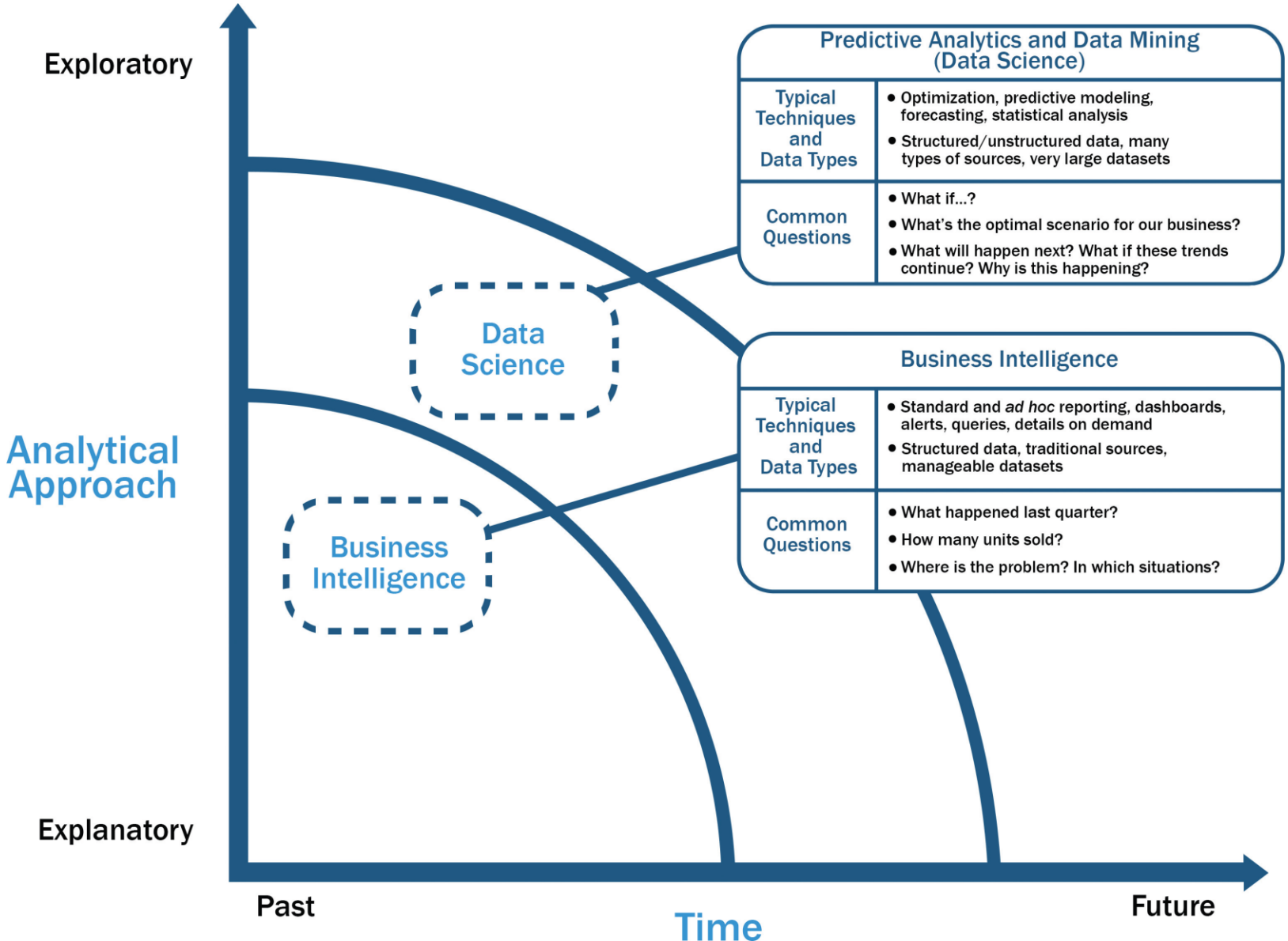
  ➔ More impactful analyses

https://au.linkedin.com/jobs/view/head-of-ai-machine-learning-at-woolworths-supermarkets-746700893

State of the Practice in Analytics

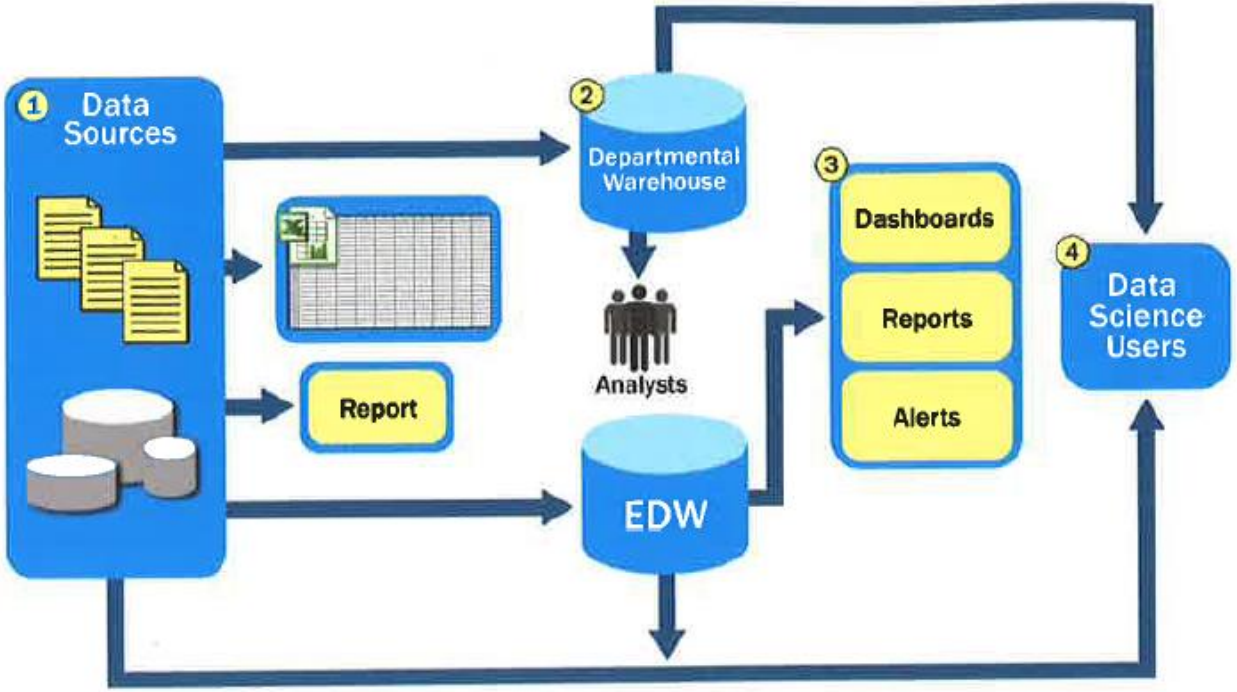
- Business Intelligence vs. Data Science
  - Time,
  - Analytical Approach,
  - Data type
  - Can you name more?
- What & How have we done in the past?
- What & How can we do in the future?

State of the Practice in Analytics



State of the Practice in Analytics

- Current Analytical Architecture



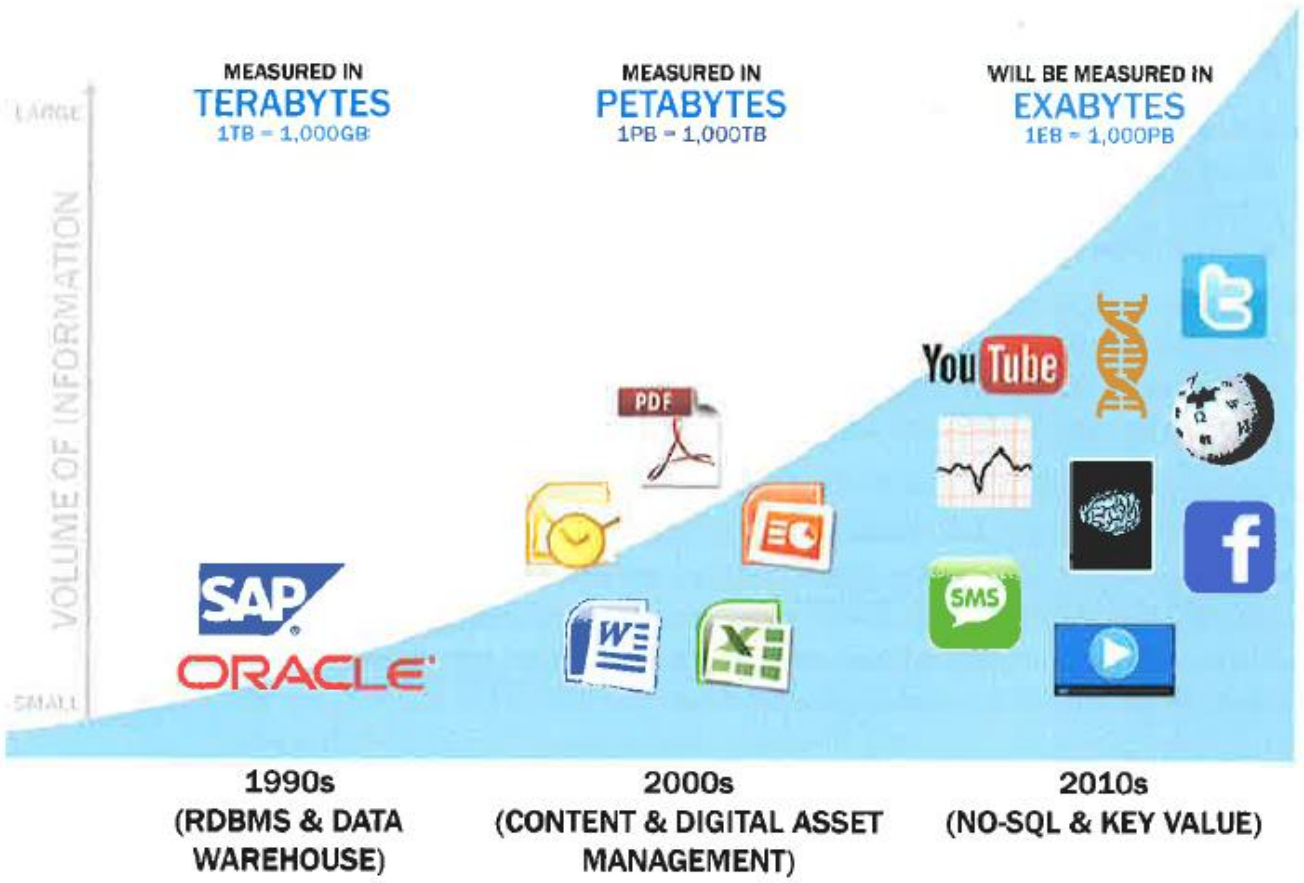
  - Traditional data architectures inhibit data exploration and more sophisticated analysis

State of the Practice in Analytics

- Traditional data architectures have several additional implications for data scientists
  - Predictive analytics and data mining activities are last in the line for data (i.e., low priority)
  - Limited to perform in-memory analytics, restricting the size of the datasets they can use
  - Projects remain isolated and ad hoc, rather than centrally managed. Exist as nonstandard initiatives
- One solution: analytic sandboxes

State of the Practice in Analytics

- Drivers of Big Data



State of the Practice in Analytics

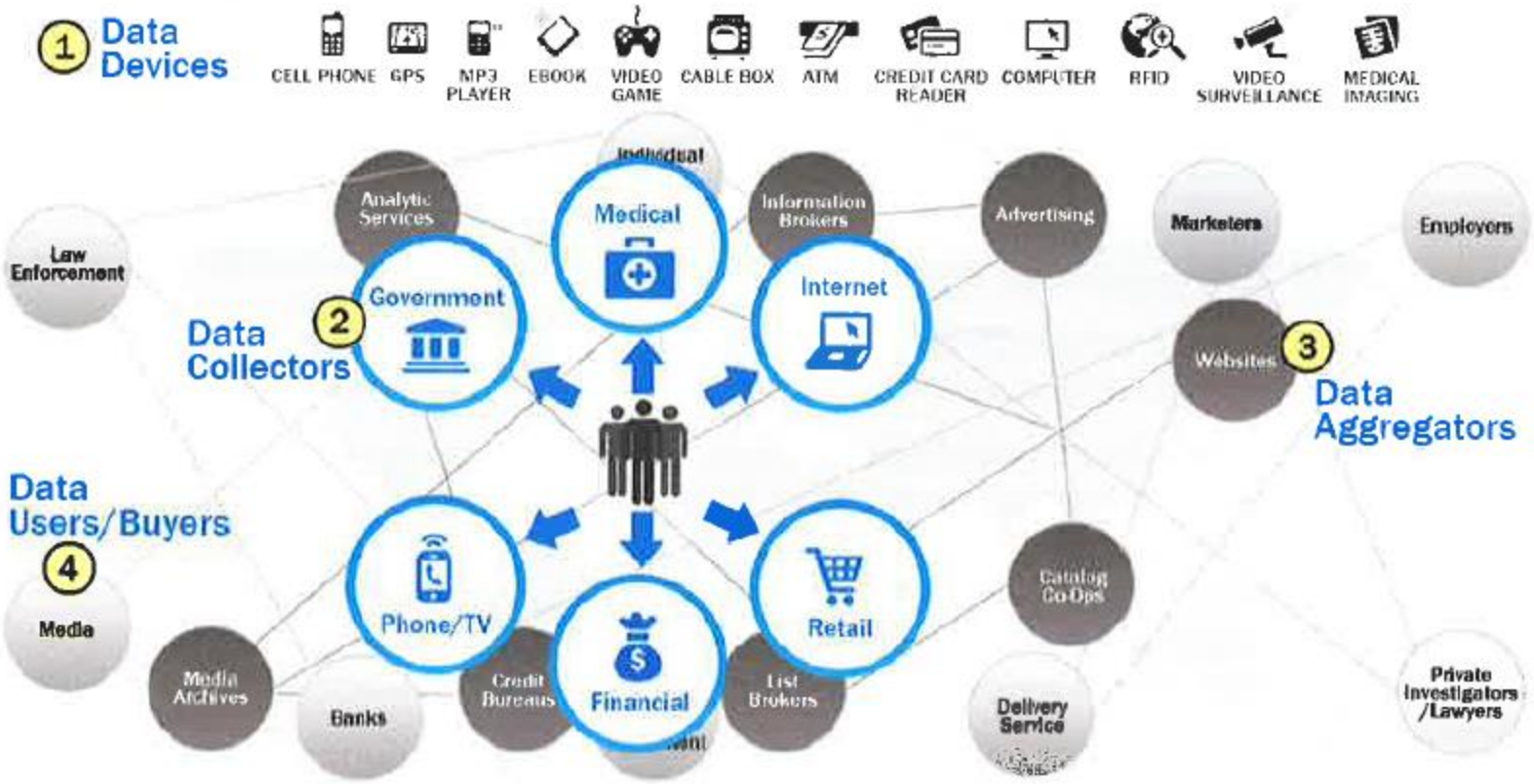
- Emerging Big Data Ecosystem & a New Approach to Analytics
  - Data → intrinsic value → a new economy
  - Data vendors, data cleaners
  - Repackaging and simplifying open source tools
  - Data is the king!

https://www.linkedin.com/pulse/data-king-one-who-masters-rule-world-so-you-ready-vinay-rao-mle-

State of the Practice in Analytics

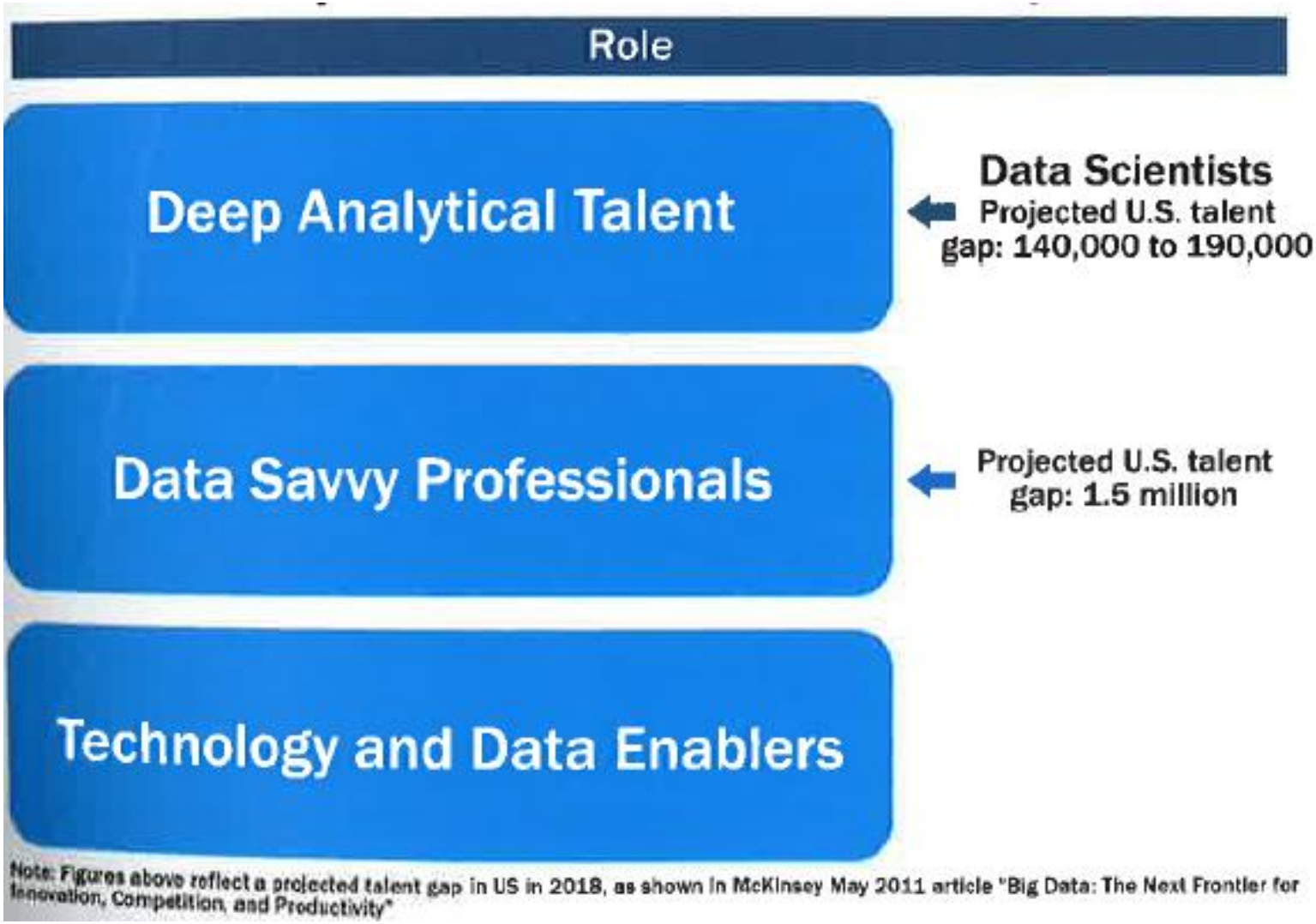
- Four main groups of players here
  - Data devices
    - Video game, Smartphone, Retail shopping card
  - Data collectors
    - Cable TV provider, shopping cart with RFID chips
  - Data aggregators
    - Compile, transform and package data to sell
  - Data users and buyers
    - Retail banks, common people

State of the Practice in Analytics



- So, Big Data problems and projects require new approach to succeed

Key roles for the New Ecosystem



Key roles for the New Ecosystem

- Data Analytical Talent (Data Scientist)
  - Advanced training in mathematics, statistics, and machine learning
  - Newest role, least understood
- Data Savvy Professionals
  - Less technical depth but can define key questions
- Technology and Data Enablers
  - Support data analytical projects
- These three groups must work together

Key roles for the New Ecosystem

- What do data scientists do?
  - Reframe business challenges to analytical challenges
  - Design, implement, and deploy statistical models and data mining techniques on Big Data
    - This is mainly what people think about them
  - Develop insights that lead to actionable recommendations to derive new business value

Examples of Big Data Analytics

- Three examples
  - US retailer Target
    - Infer Marriage, Divorce, and Pregnancy
    - Manage its inventory correspondingly
  - IT Infrastructure
    - Apache Hadoop
    - Process vast amount of information parallelly
  - Social media
    - Leverage social interactions to derive new insights

Summary

- Big Data comes from myriad sources
- Big Data addresses business needs and solves complex problems
- Companies and organisations move toward Data Science
- Require new architectures, new ways of working, new skill sets, new roles, etc.
- A growing talent gap

Questions for you

- What are the three (or five) characteristics of Big Data?
- What is an analytic sandbox, and why is it important?
- Explain the difference between BI and Data Science
- Describe the challenges of the current analytical architecture for data scientists
- What are the key skills and characteristics of a data scientist?

Q&A

Images Courtesy of Google Image