

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?

What's Driving Data Deluge?

Any more?



Mobile
Sensors



Social
Media



Video
Surveillance



Video
Rendering



Smart
Grids



Geophysical
Exploration



Medical
Imaging



Gene
Sequencing

Big Data Overview

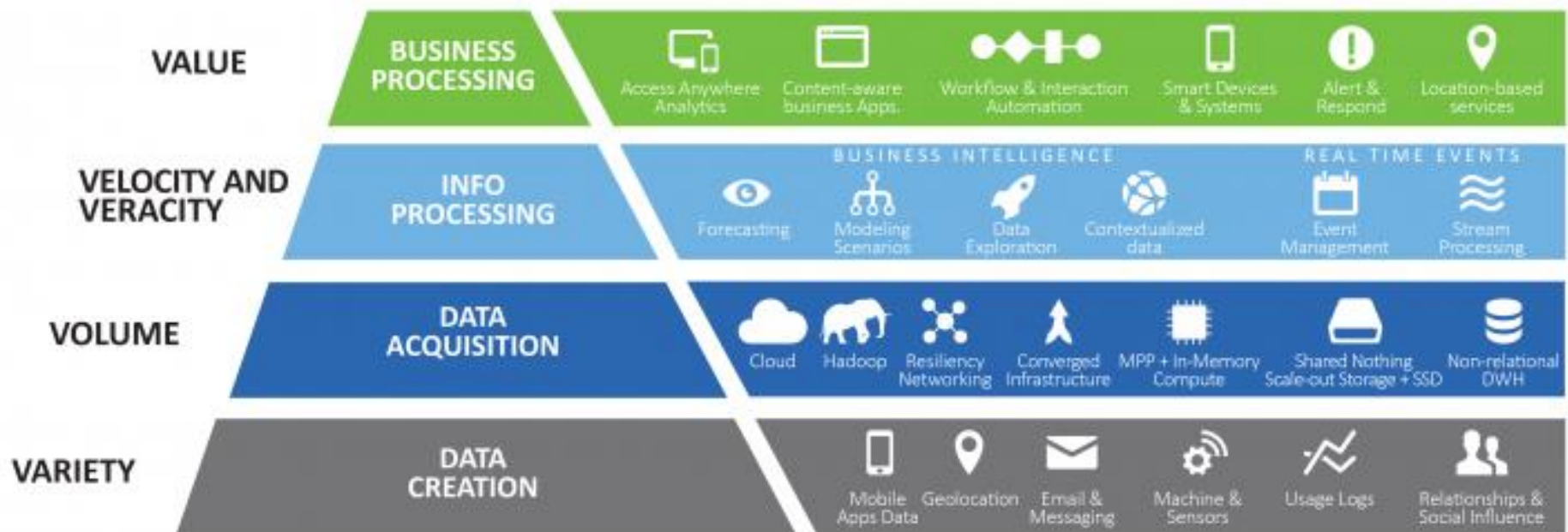
- Keeping up with this **high influx of data** is difficult
- Analysing **vast amounts of data** is more challenging, especially when the data does **not** conform to **traditional** structure
- Can you name any real applications of Big Data Analytics you have been aware of?

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



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

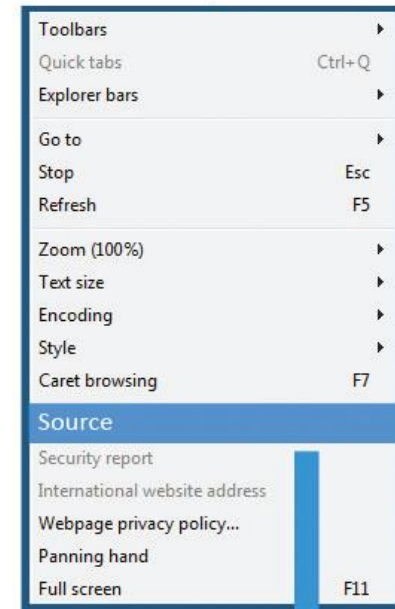
- Data Structures
- Structured data
 - Can you name some examples?

| 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?



```
<meta charset="utf-8">
<meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1">
<title>EMC - Leading Cloud Computing, Big Data, and Trusted IT Solutions</title>

<meta name="description" content="EMC is a leading provider of IT storage hardware solutions to promote data
cloud computing.">
name="keywords" content="emc,network storage,data recovery,information management,backup software,nas storage

<meta name="viewport" content="width=device-width, initial-scale=1">

<link href="/_admin/css/html-layout-css-includes-combined-min.css" rel="stylesheet">
<script src="/_admin/js/igquery.js"></script>
<link rel="stylesheet" href="/R1/assets/css/common/normalize.css">
<link rel="stylesheet" href="/R1/assets/css/homepage/main.css">
<link rel="stylesheet" href="/R1/assets/css/common/responsive-header.css">
<link rel="stylesheet" href="/R1/assets/css/common/responsive-footer.css">

<script type="text/javascript" src="//platform.twitter.com/widgets.js"></script>
<script src="/R1/assets/js/common/modernizr-2.6.2.min.js"></script>
<script type="text/javascript">
```

Semi-structured

1

emc data science

Web News Images Videos Shopping More Search tools

About 2,190,000 results (0.24 seconds)

Data Science and Big Data Analytics Training - EMC Education ...
 education.emc.com > Home > Training > Learning Paths > EMC Corporation >
 "We live in a data-driven world. Increasingly, the efficient operation of organizations across sectors relies on the effective use of vast amounts of data. Making ...

Data Scientist - EMC Education, Training, and Certification
 education.emc.com > ... > Associate Level Certifications > EMC Corporation >
 Data Science and Big Data Analytics course provides hands-on practitioner's approach to the techniques and tools required for Big Data Analytics. Being Proven ...

EMC Education, Training, and Certification
 https://education.emc.com/ > EMC Corporation >
 EMC Education Services. ... (CIS) StarterKit > Data Science and Big Data Analytics StarterKit > Backup and Recovery Systems and Architecture StarterKit >.

[PDF] Data Science Revealed: A Data-Driven Glimpse into the ... - EMC
 www.emc.com/.../news/emc-data-science-study-wp.pdf > EMC Corporation >
 field of data science revealed what many are becoming to understand: that data ... Data science is an emerging field, with rapid changes, great uncertainty, and ...

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

2

United States Change Location Site Tutorial Customer / Partner Login

EMC²

HOME STORE TRAINING CERTIFICATION SUPPORT OTHER EMC SITES

Home > Training > Learning Paths > Data Science and Big Data Analytics

DATA SCIENCE AND BIG DATA ANALYTICS
 An 'open' course to unleash the power of Big Data.

Big Data Analytics requires involvement of Data Scientists

"We live in a data-driven world. Increasingly, the efficient operation of organizations across sectors relies on the effective use of vast amounts of data. Making sense of big data is a combination of organizations having the tools, skills and more importantly, the mindset to see data as the new 'oil' fueling a company. Unfortunately, the technology has evolved faster than the workforce skills to make sense of it and organizations across sectors must adapt to this new reality or perish."
 Andreas Weigend, Ph.D. Stanford, Head of the Social Data Lab at Stanford, former Chief Scientist Amazon.com

Course Details
 Course Overview > Course Description >

Please choose your purchase option:

Data Science and Big Data Analytics StarterKit
 Price: \$600 **Save [37%]**
 Includes: 1 Data Science and Big Data Analytics Video-ILT (\$1,200 value) + 1 Proven Professional Exam Voucher (\$200 value)
 Delivery Mode: Video-ILT (Stream)
 Certification Alignment: E20-007

Also available for Purchase:
 Price: \$500 **Save [38%]**
 Includes: 1 Data Science and Big Data Analytics Video-ILT (\$1,200 Value). Exam Voucher not included.

Unleash the Power of Big Data Analytics
 Download PDF >

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

3

United States Change Location Site Tutorial Customer / Partner Login

EMC²

HOME STORE TRAINING CERTIFICATION SUPPORT OTHER EMC SITES

Home > EMC Proven Professional Certification > Certification Framework > Associate Level Certifications > Data Science

EMC PROVEN PROFESSIONAL CERTIFICATION

EMC Proven Professional Certification

Certification Framework

- Associate Level Certifications
 - Information Storage and Management
 - Backup Recovery
 - Cloud Infrastructure and Services
 - Content Management Foundations
 - Data Science**
 - Data Center Architect
 - Cloud Architect
 - Storage Administrator
 - Technology Architect

Data Science Associate
 EMC Proven Professional Certification

Data Science and Big Data Analytics course provides hands-on practitioner's approach to the techniques and tools required for Big Data Analytics. Being Proven means investing in yourself and formally validating your knowledge, skills, and expertise by the industry's most comprehensive learning and certification program. The Data Science and Big Data Analytics course prepares you for Data Scientist Associate (EMCDSA) Certification.

Exam and Practice Test

| | |
|------------|--------------|
| Expert | n/a |
| Specialist | n/a |
| Associate | E20-007 Exam |

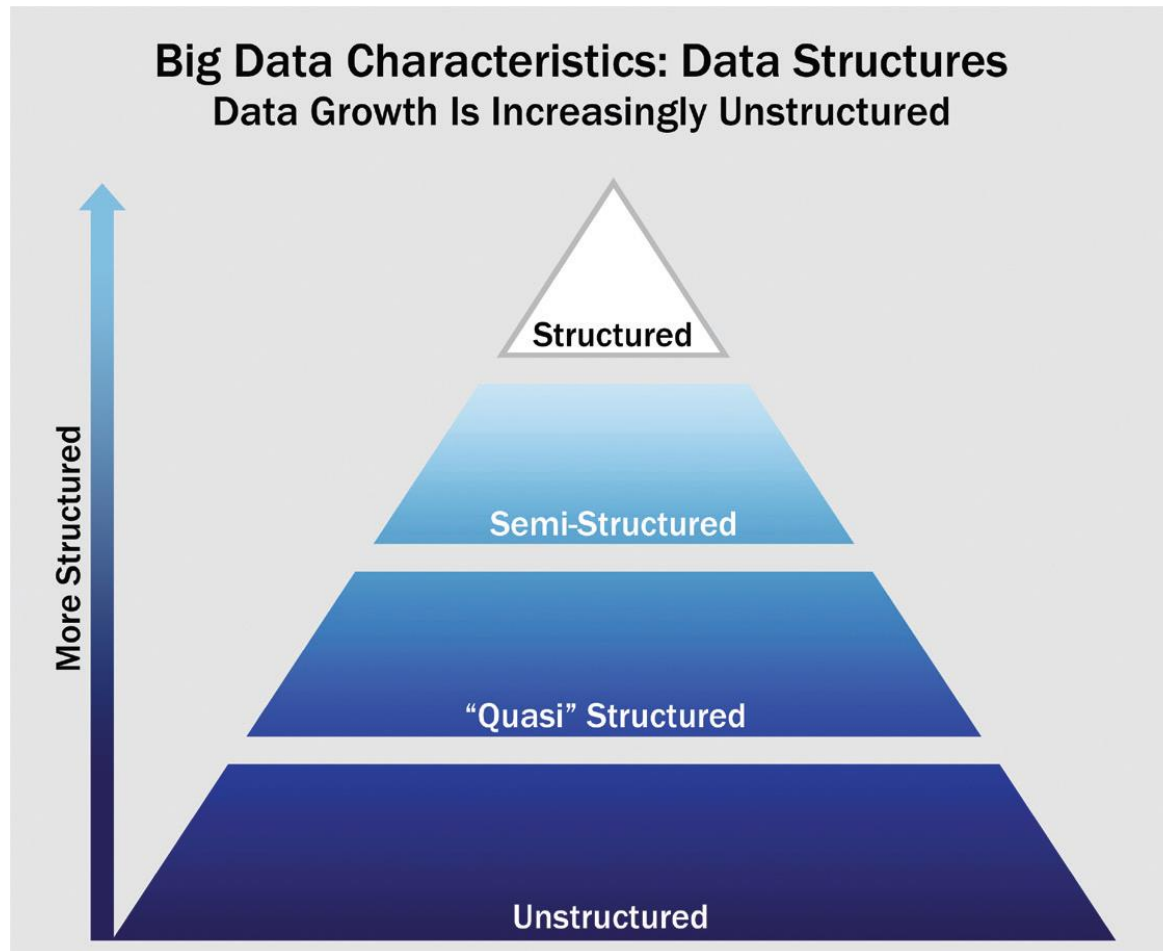
ASSOCIATE COURSES

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) |

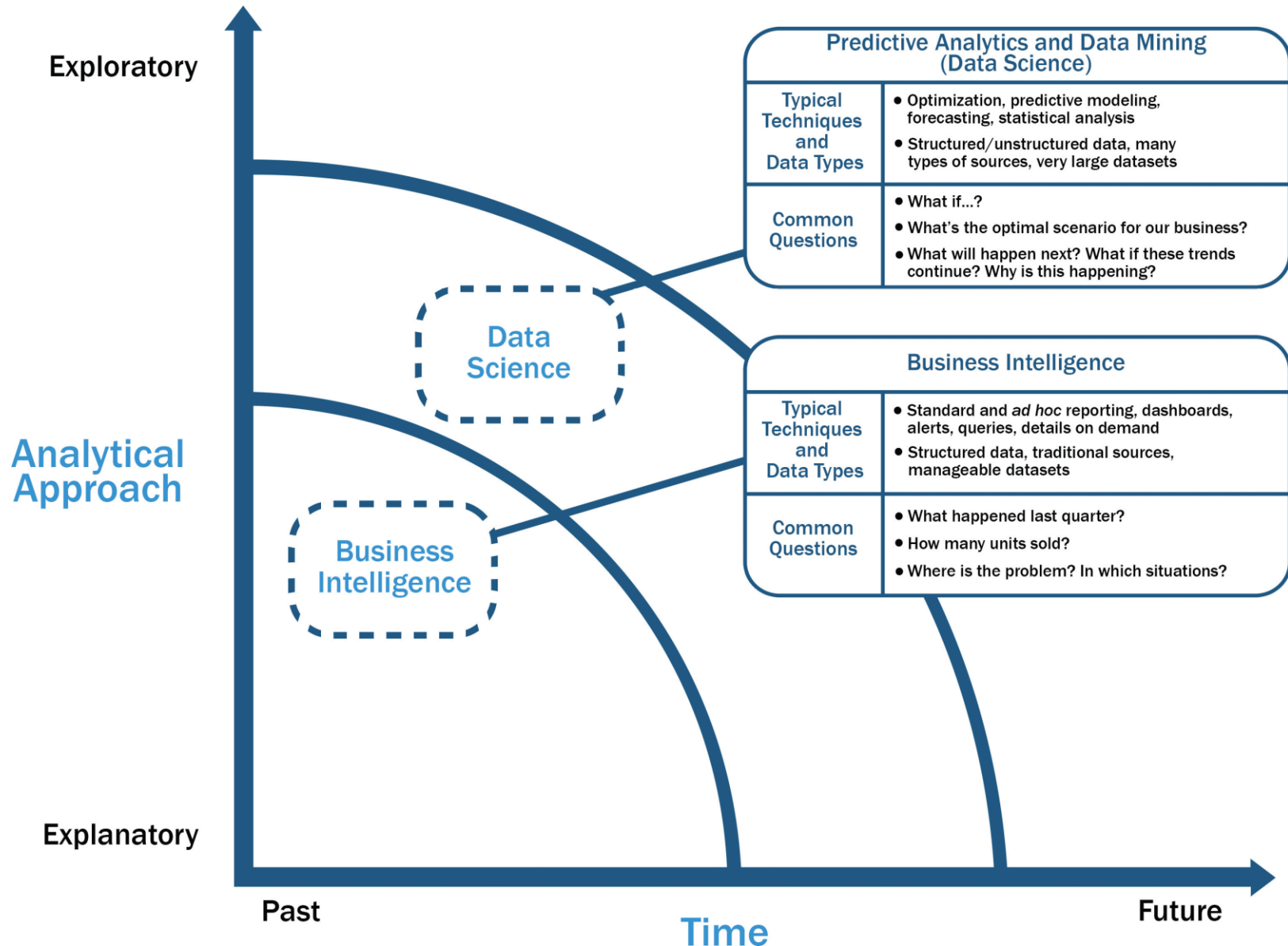
State of the Practice in Analytics

- **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**

State of the Practice in Analytics

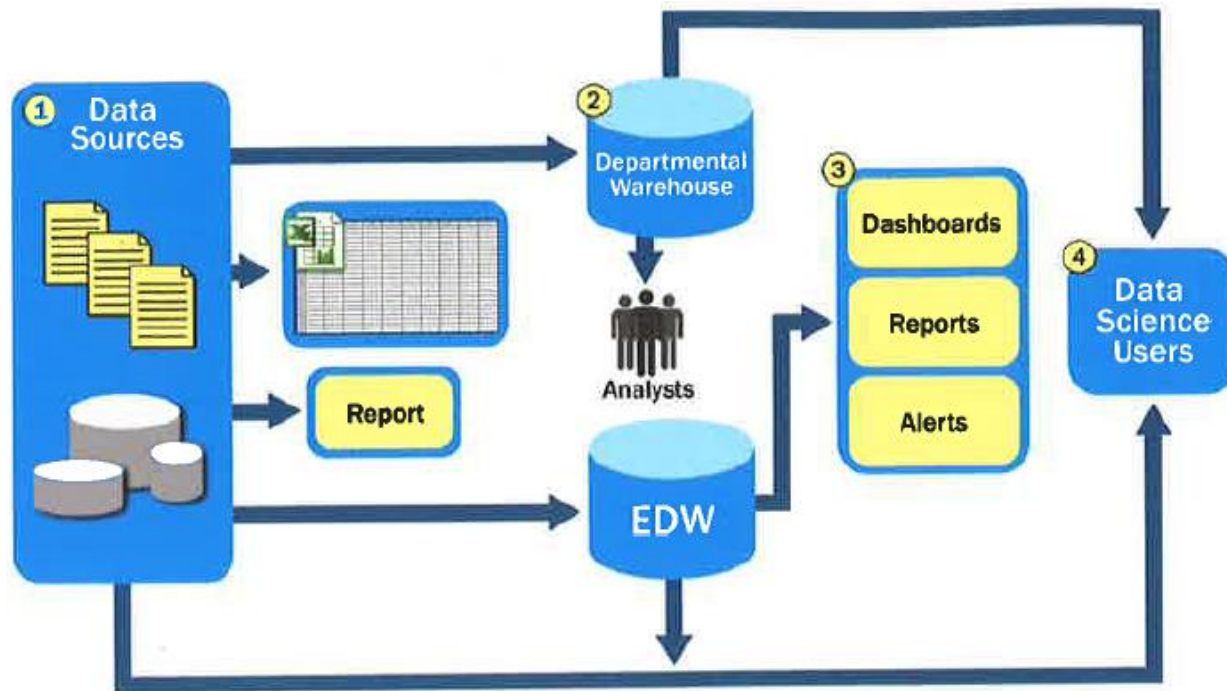
- Business Intelligence vs. Data Science
 - Time,
 - Analytical Approach,
 - Data type
 - Can you name more?
- What & How have we done in the past?
- What & How can we do in the future?

State of the Practice in Analytics



State of the Practice in Analytics

- Current Analytical Architecture



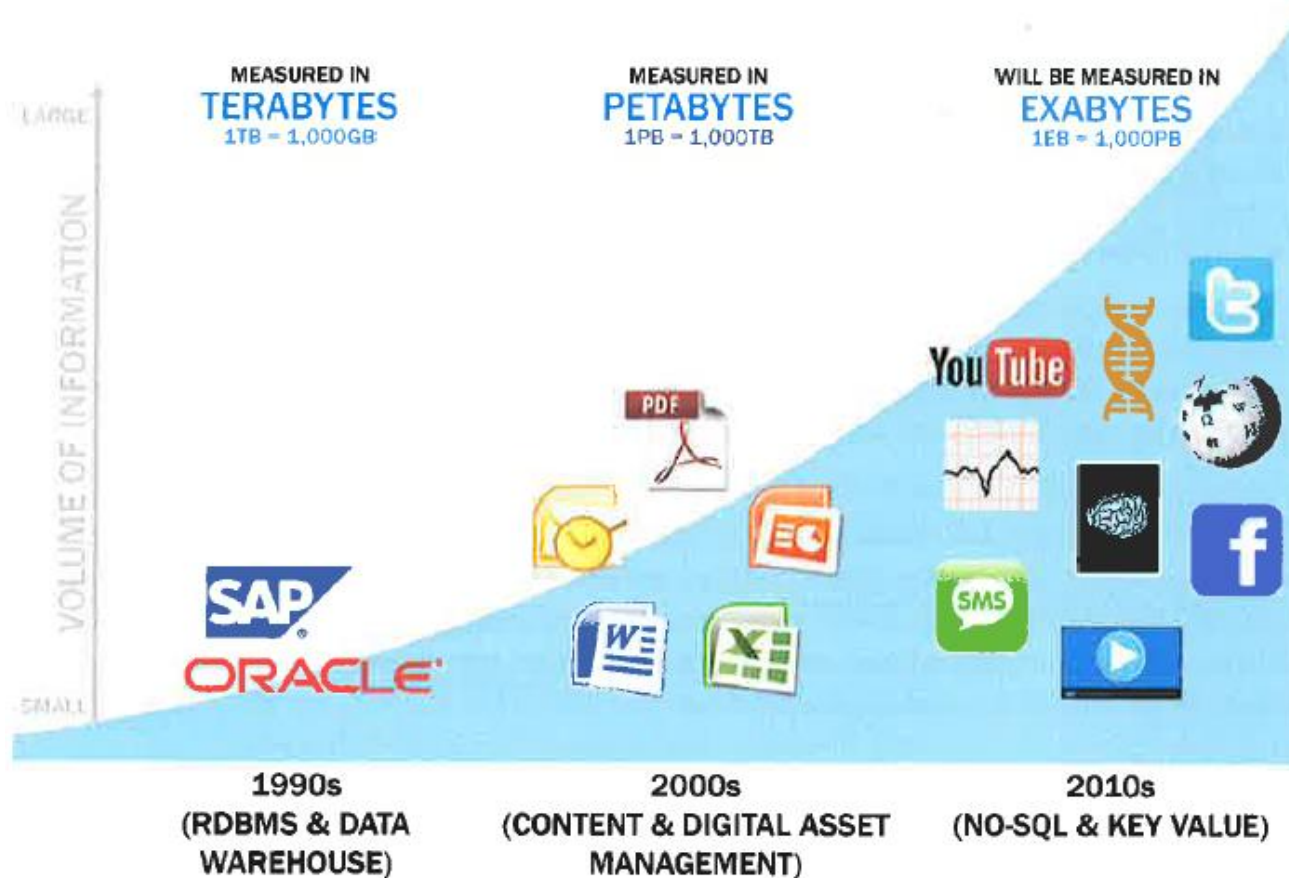
— Traditional data architectures **inhibit** data exploration and more sophisticated analysis

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

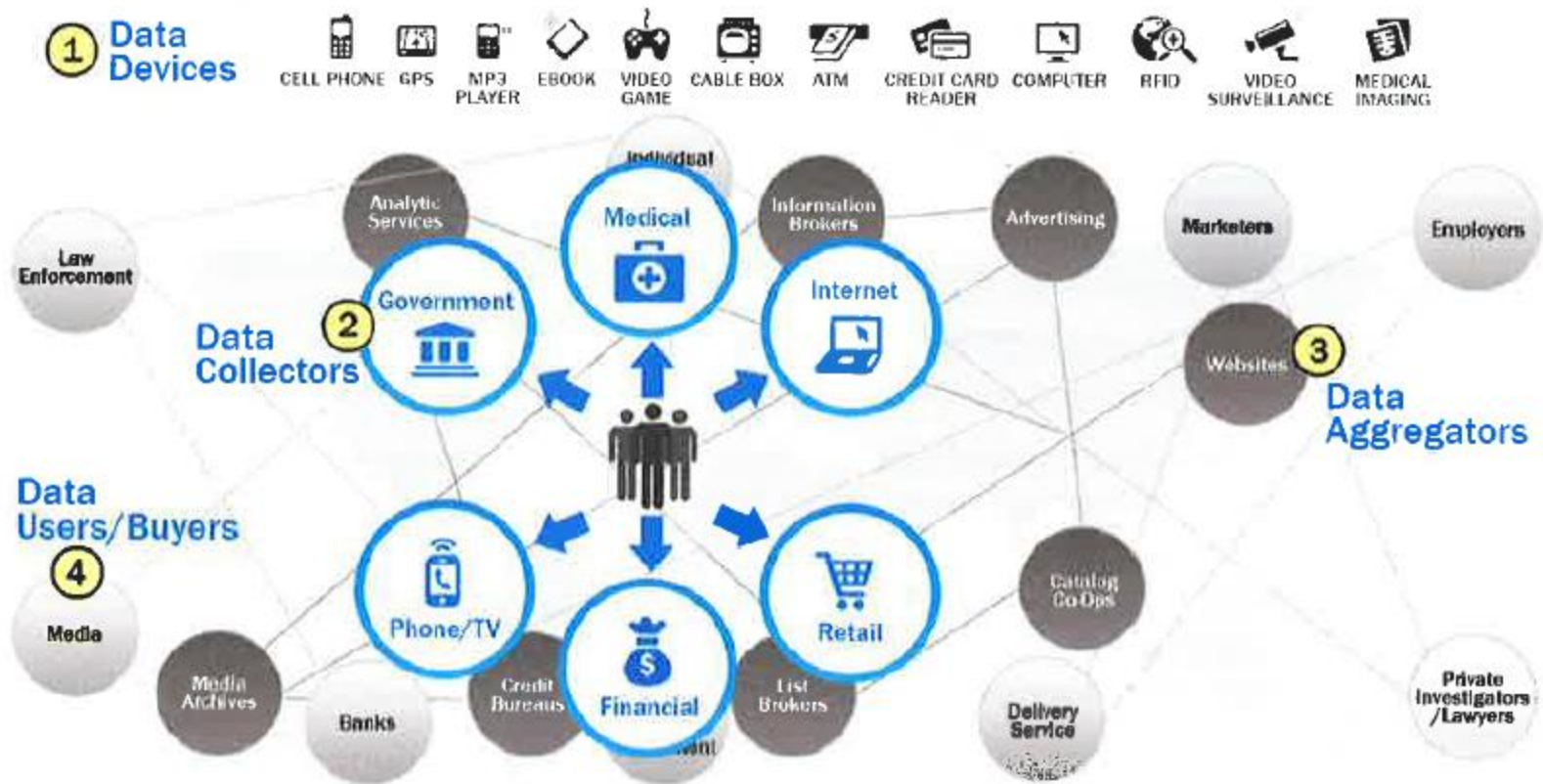
- Emerging Big Data Ecosystem & a New Approach to Analytics
 - Data → intrinsic value → a new economy
 - Data vendors, data cleaners
 - Repackaging and simplifying open source tools
 - Data is the king!



State of the Practice in Analytics

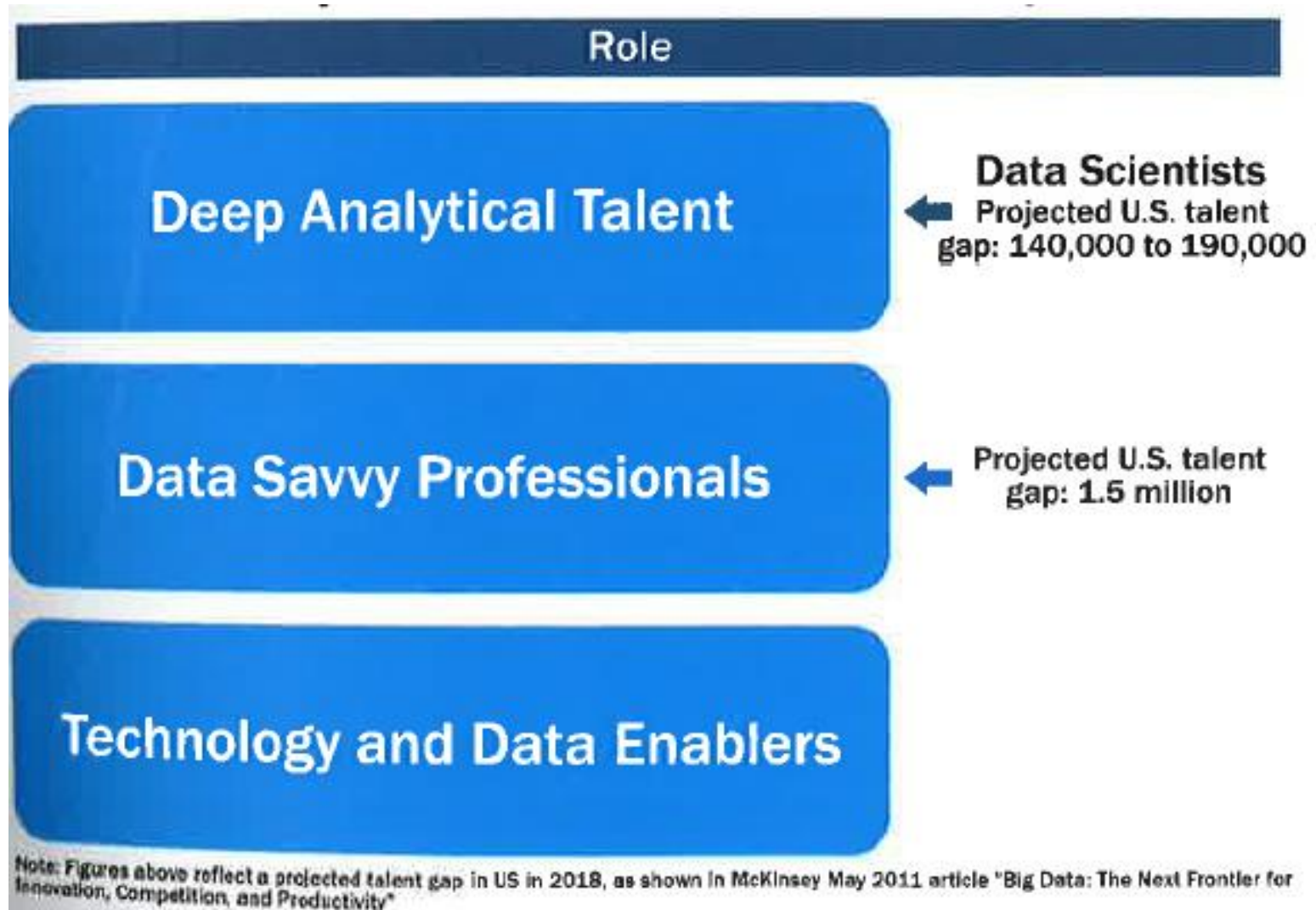
- **Four** main groups of players here
 - Data devices
 - Video game, Smartphone, Retail shopping card
 - Data collectors
 - Cable TV provider, shopping cart with RFID chips
 - Data aggregators
 - Compile, transform and package data to sell
 - Data users and buyers
 - Retail banks, common people

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?

