CBD2204: Week 1

Takis Zourntos

overview

- data is created constantly, at an ever-increasing rate
- mobile phones, social media, medical imaging technologies, all of these create data, which must be stored and processed
- devices and sensors (e.g., IoT technologies) generate information that needs real-time (deadline-driven) processing
- two major challenges:
  - keeping up with the rate of data generation
  - analyzing the vast amount data, which may not be structured

leading the way in big data are industries such as:

- credit card companies (monitoring every purchase made by each customer)
- mobile phone companies (analyzing, for example, subscriber calling patterns)
- social media companies (where the data about users has inherent value)

But what makes something “Big Data”?

1. huge volume of data
2. complexity/diversity of data types and structures
3. speed of new data creation and growth (“high velocity data”)

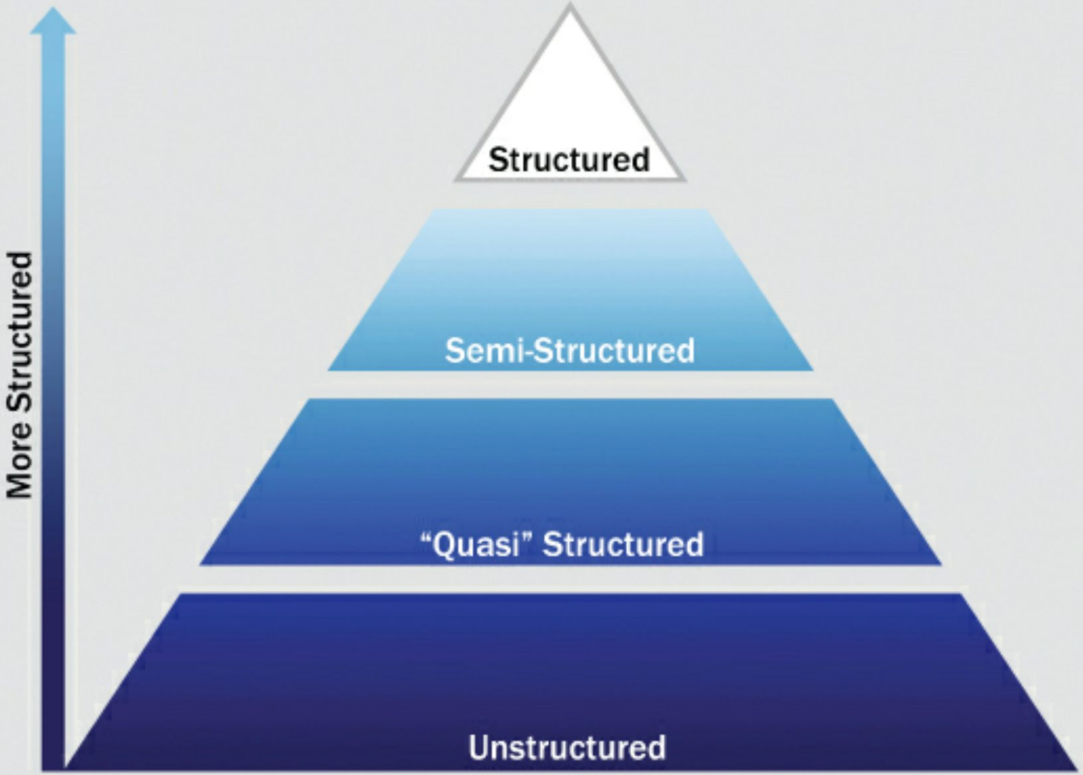
Big Data is sometimes described as having the three Vs: volume, variety and velocity

What's Driving Data Deluge?

- Mobile Sensors
- Social Media
- Video Surveillance
- Video Rendering
- Smart Grids
- Geophysical Exploration
- Medical Imaging
- Gene Sequencing

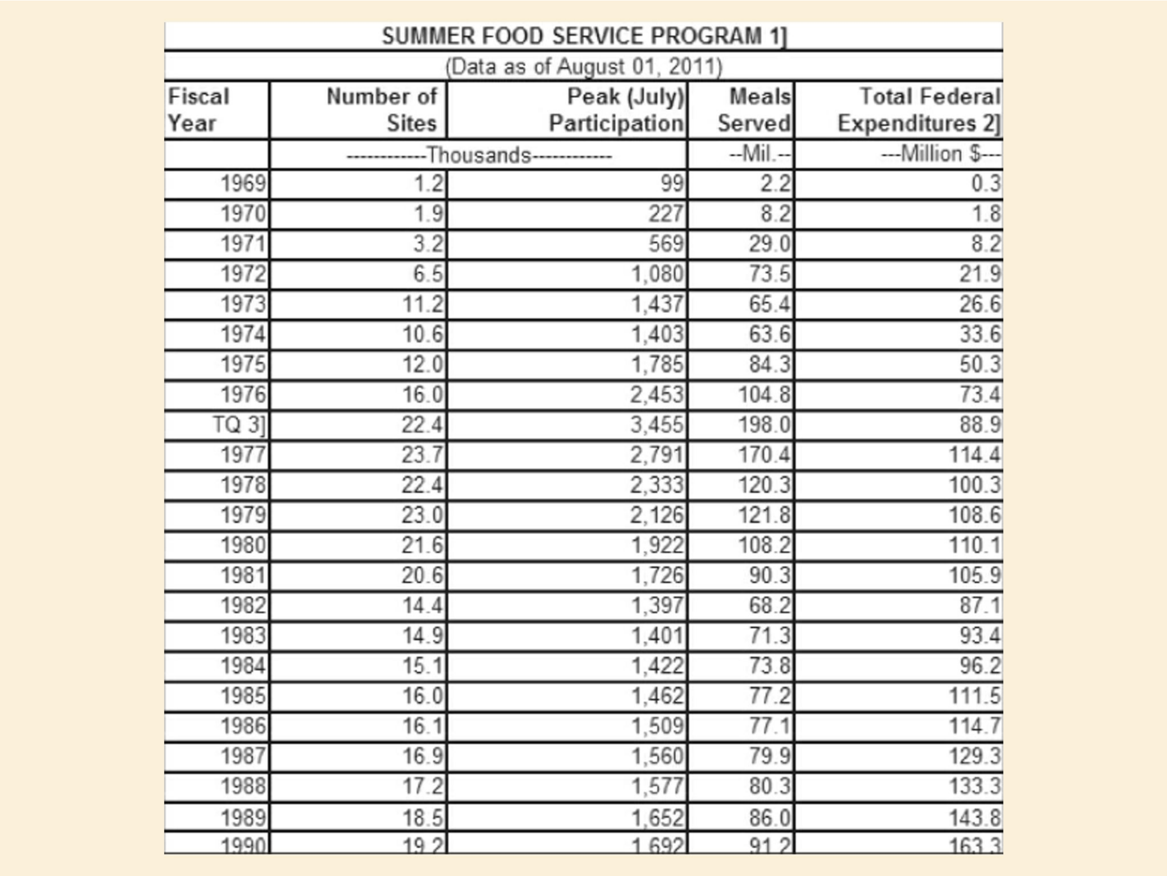
Big Data Characteristics: Data Structures

Data Growth Is Increasingly Unstructured



data structures

- **structured data**: data containing defined type and format (think: spreadsheet, CSV files, relational databases)
- **semi-structured data**: textual data files with a discernable pattern, enables parsing (e.g., XML files)
- **quasi-structured data**: textual data with erratic data formats that needs effort to format (e.g., clickstream data containing inconsistencies in data values and formats)
- **unstructured data**: data with no inherent structure such as text documents, images and video

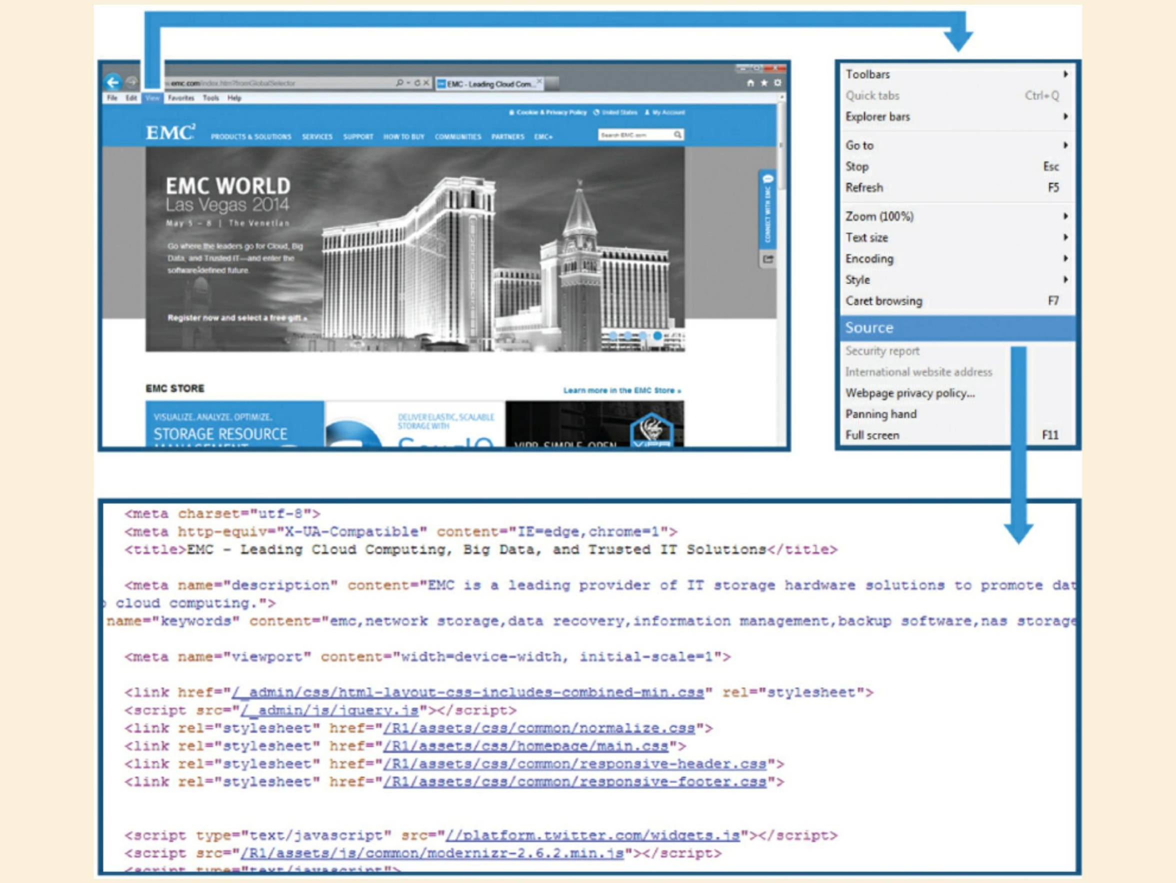


SUMMER FOOD SERVICE PROGRAM 1]

(Data as of August 01, 2011)

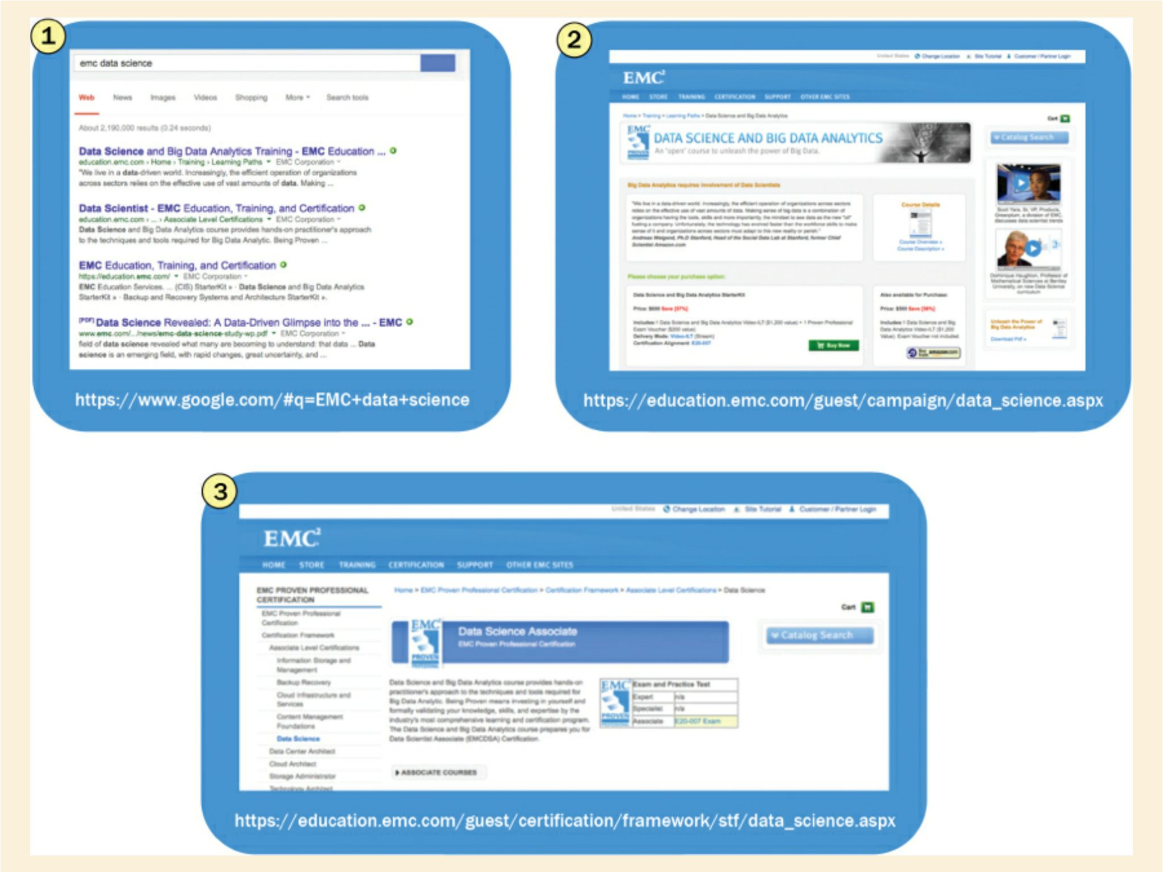
Fiscal Year	Number of Sites	Peak (July) Participation	Meals Served	Total Federal Expenditures 2]
	------------Thousands------------		--Mil.--	---Million $---
1969	1.2	99	2.2	0.3
1970	1.9	227	8.2	1.8
1971	3.2	569	29.0	8.2
1972	6.5	1,080	73.5	21.9
1973	11.2	1,437	65.4	26.6
1974	10.6	1,403	63.6	33.6
1975	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

example of structured data



```
  <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 dat
 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/jquery.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">
```

example of semi-structured data



1

emc data science

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 Analytic. 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

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 »

Scott Yara, Sr. VP, Products, Greenplum, a division of EMC, discusses data scientist trends

Dominique Haughton, Professor of Mathematical Sciences at Bentley University, on new Data Science curriculum

Please choose your purchase option:

Data Science and Big Data Analytics StarterKit

Price: $600 Save (57%)

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 (58%)

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

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

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

Data Science Associate

EMC Proven Professional Certification

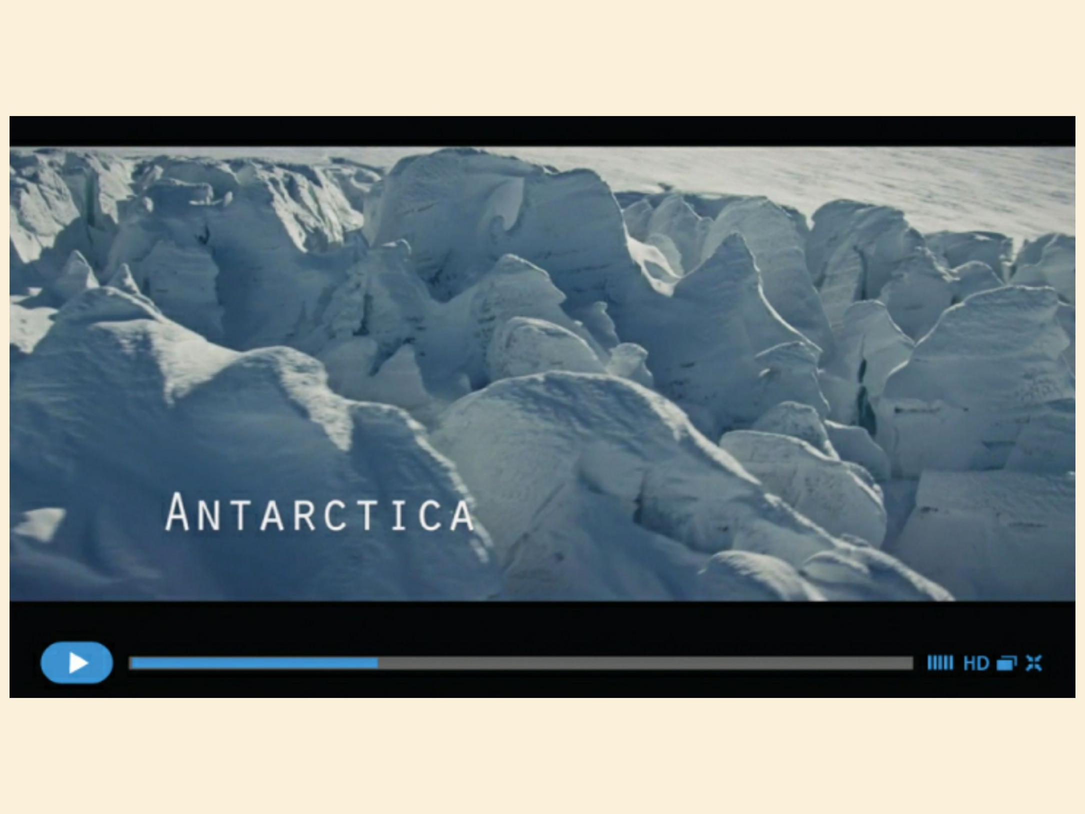
Data Science and Big Data Analytics course provides hands-on practitioner's approach to the techniques and tools required for Big Data Analytic. Being Proven means investing in yourself and formally validating your knowledge, skills, and expertise by the industry's most comprehensive learning and certification 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

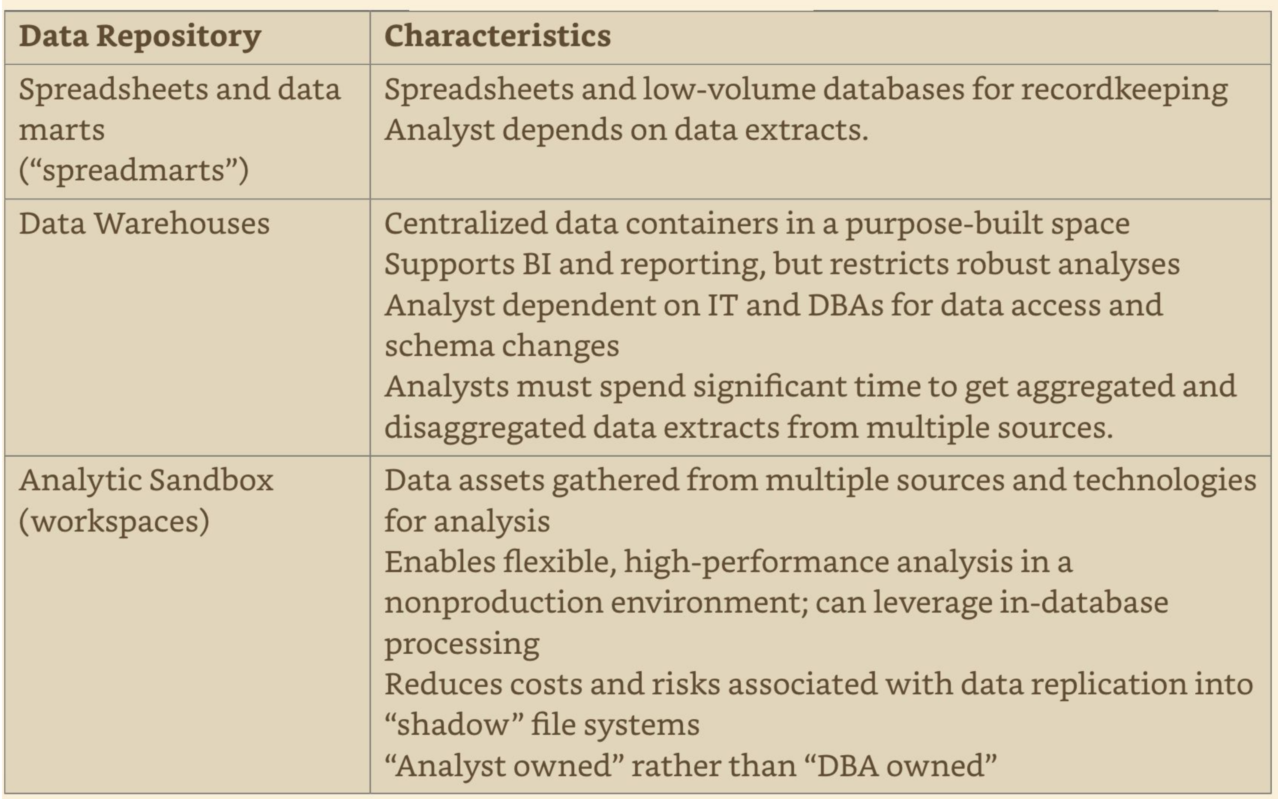
https://education.emc.com/guest/certification/framework/stf/data_science.aspx

example of quasi-structured data

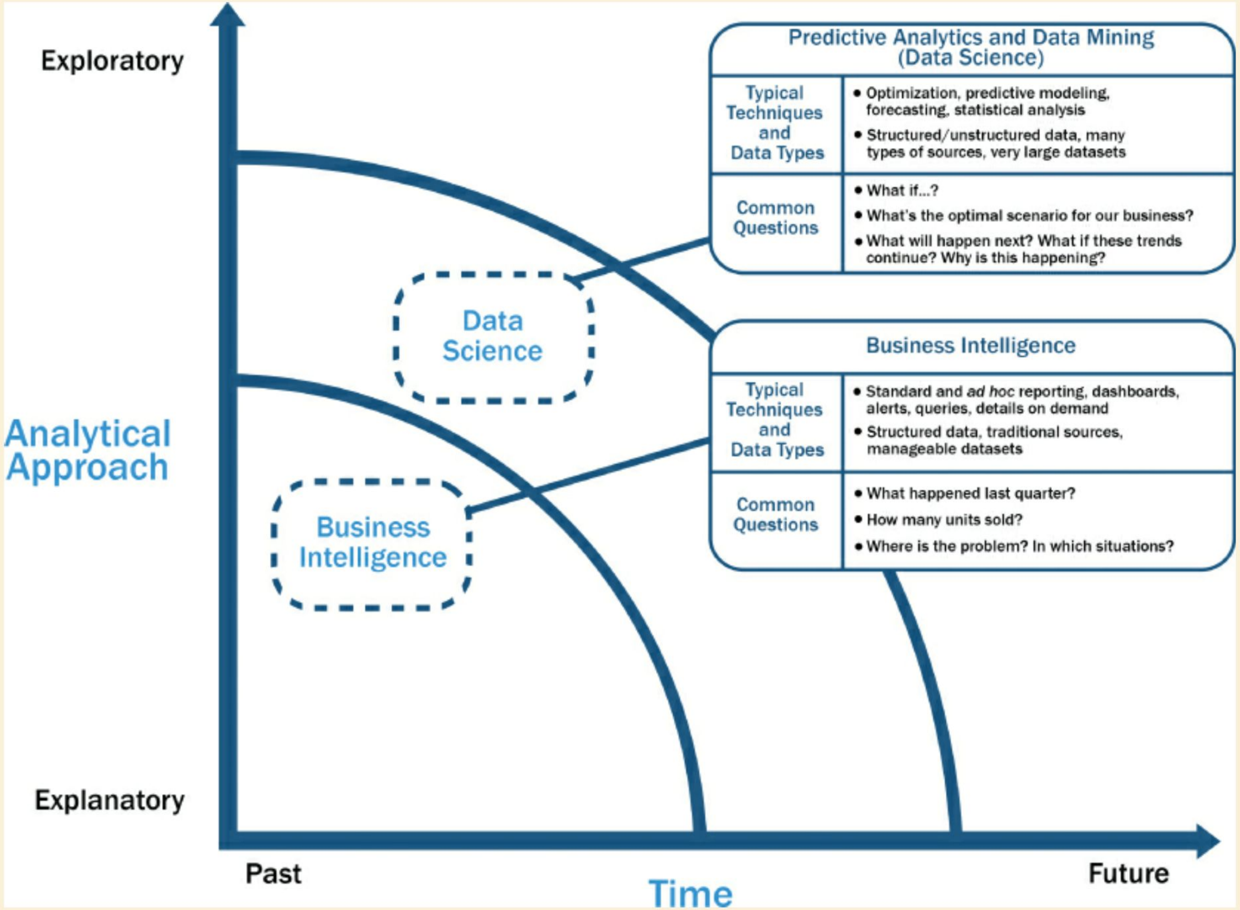


example of unstructured data

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”



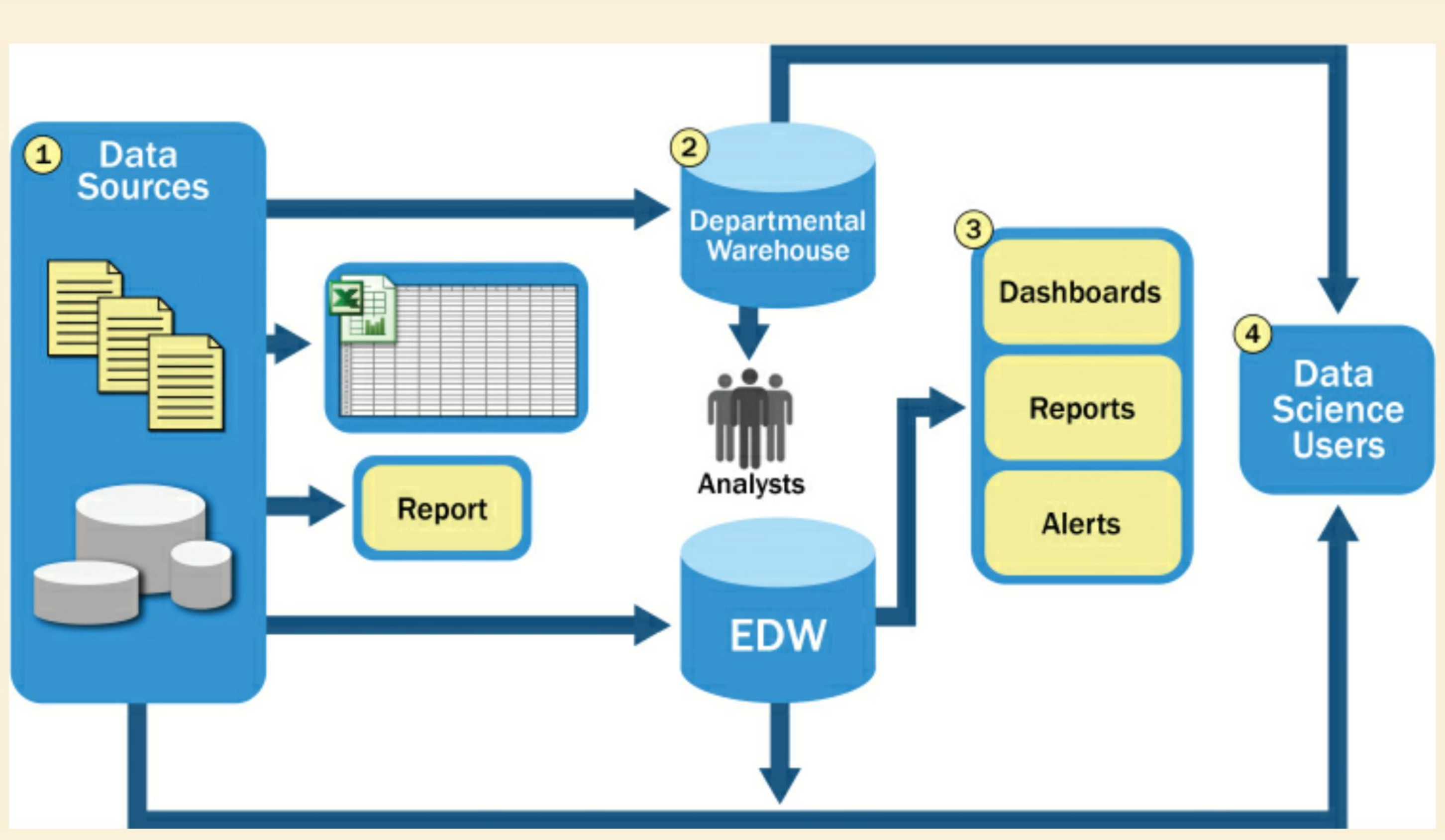
Predictive Analytics and Data Mining (Data Science)

Typical Techniques and Data Types	• Optimization, predictive modeling, forecasting, statistical analysis • Structured/unstructured data, many types of sources, very large datasets
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?

Business Intelligence

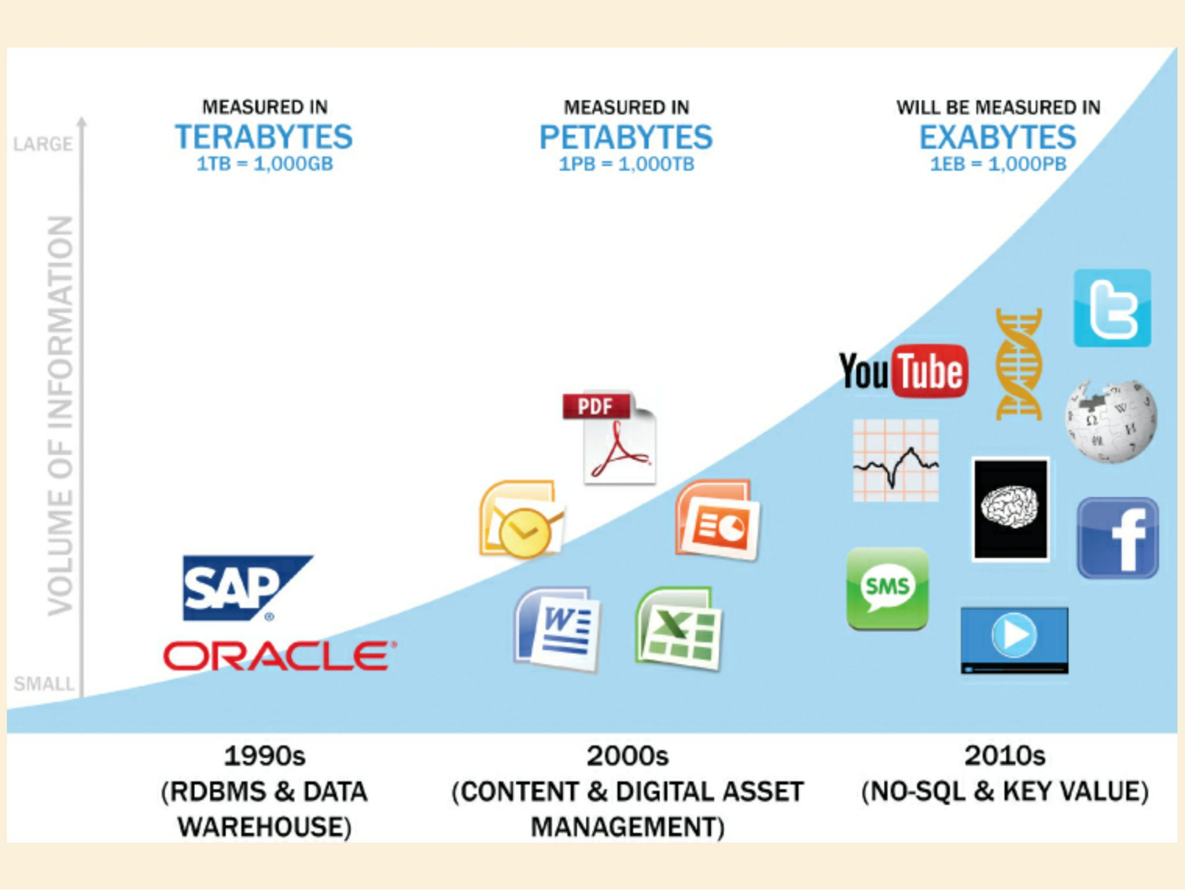
Typical Techniques and Data Types	• Standard and *ad hoc* reporting, dashboards, alerts, queries, details on demand • Structured data, traditional sources, manageable datasets
Common Questions	• What happened last quarter? • How many units sold? • Where is the problem? In which situations?

typical analytical architecture

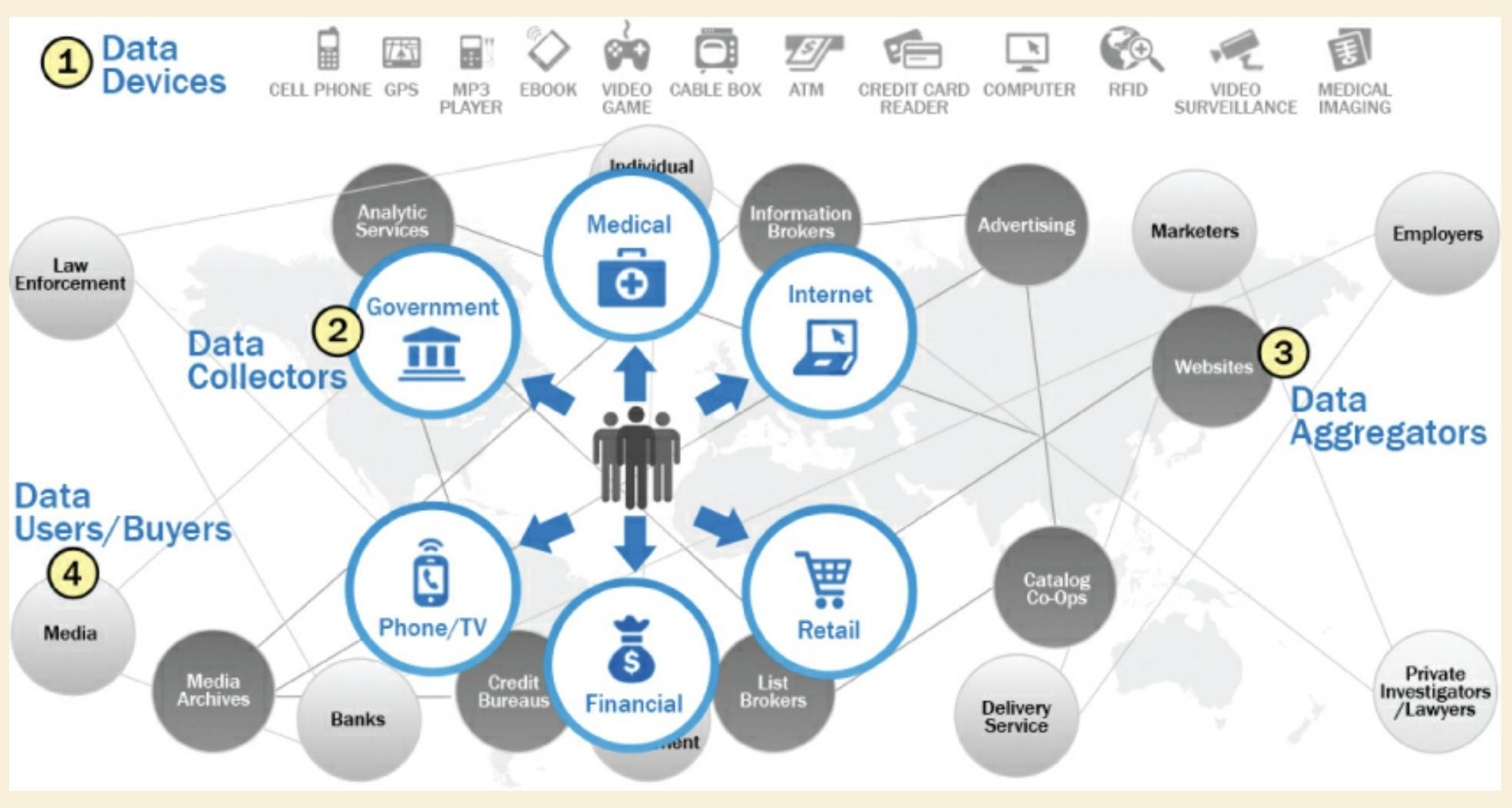


EDW: Enterprise Data Warehouse

rise of big data sources



emerging Big Data ecosystems



Main Reference:

Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, EMC Education Services

Amazon.ca link:
https://www.amazon.ca/Data-Science-Big-Analytics-Discovering-ebook/dp/B00RXHVQF6/ref=tmm_kin_swatch_0?_encoding=UTF8&qid=&sr=