Big Data Defined

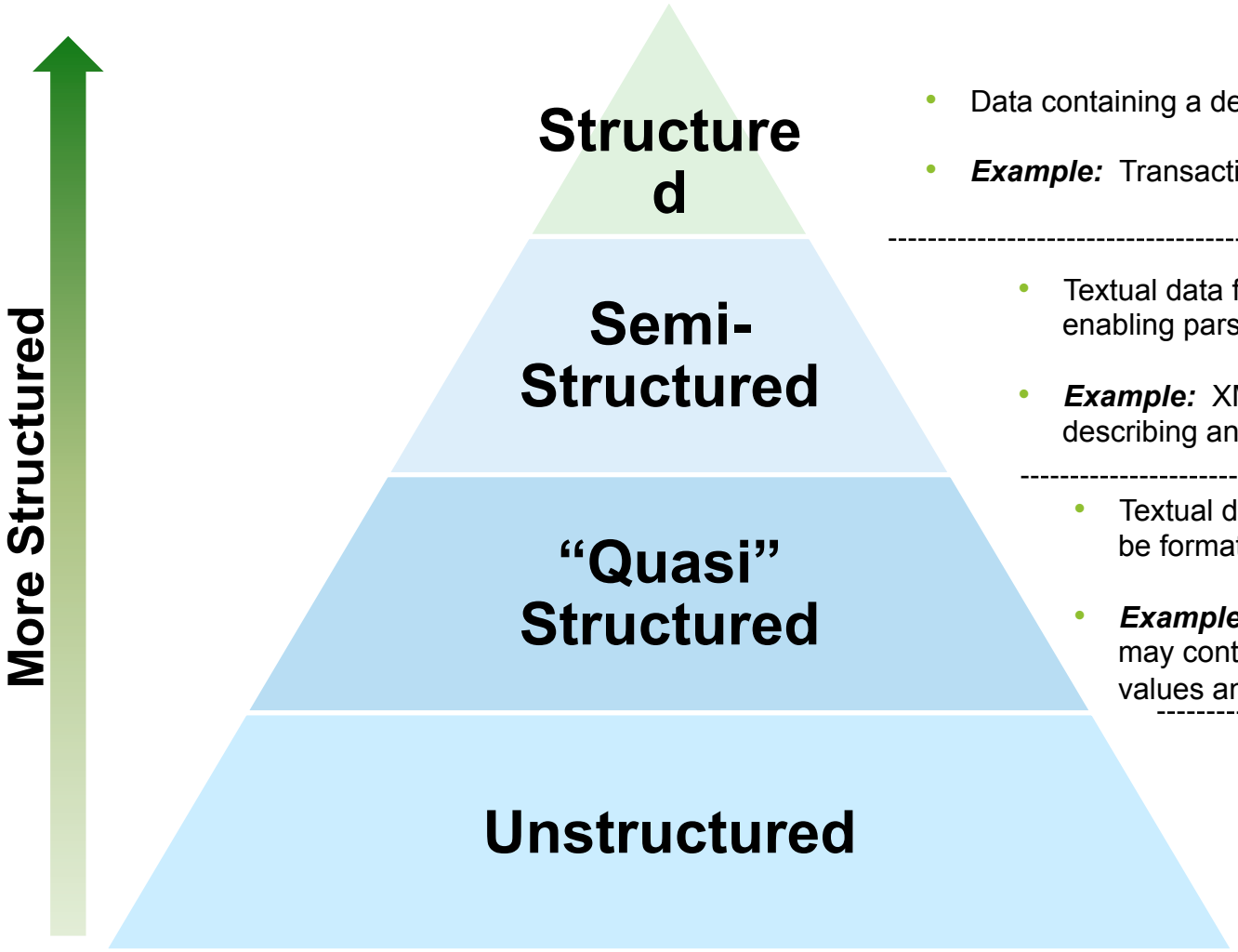
- ***"Big Data" is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value.***
  - Requires new data architectures, analytic sandboxes
  - New tools
  - New analytical methods
  - Integrating multiple skills into new role of data scientist

- Organizations are deriving business benefit from analyzing ever larger and more complex data sets that increasingly require real-time or near-real time capabilities

Source: McKinsey May 2011 article *Big Data: The next frontier for innovation, competition, and productivity*

Big Data Characteristics: Data Structures

Data Growth is Increasingly Unstructured



- 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

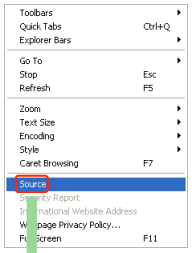
Four Main Types of Data Structures

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

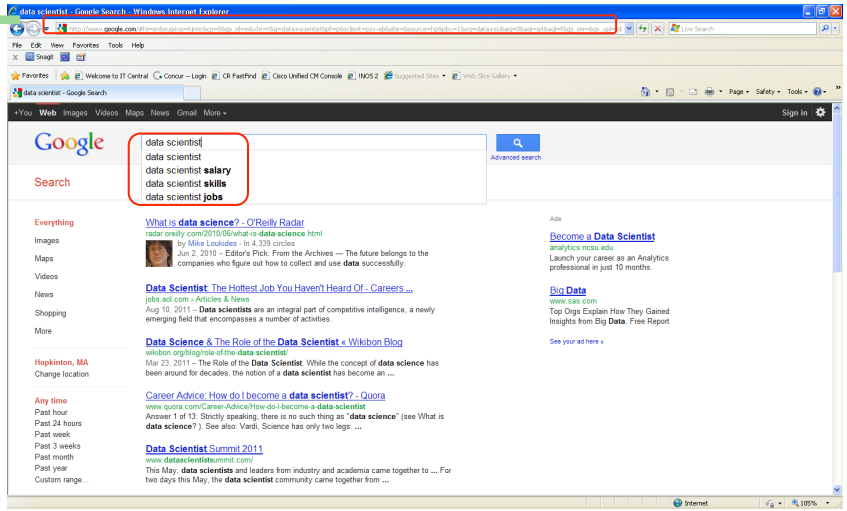
View → Source




```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transi
<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

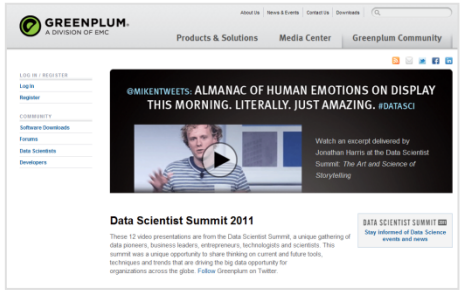
The Red Wheelbarrow, by William Carlos Williams

so much depends
upon

a red wheel
barrow

glazed with rain
water

beside the white
chickens.



Data Repositories, 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

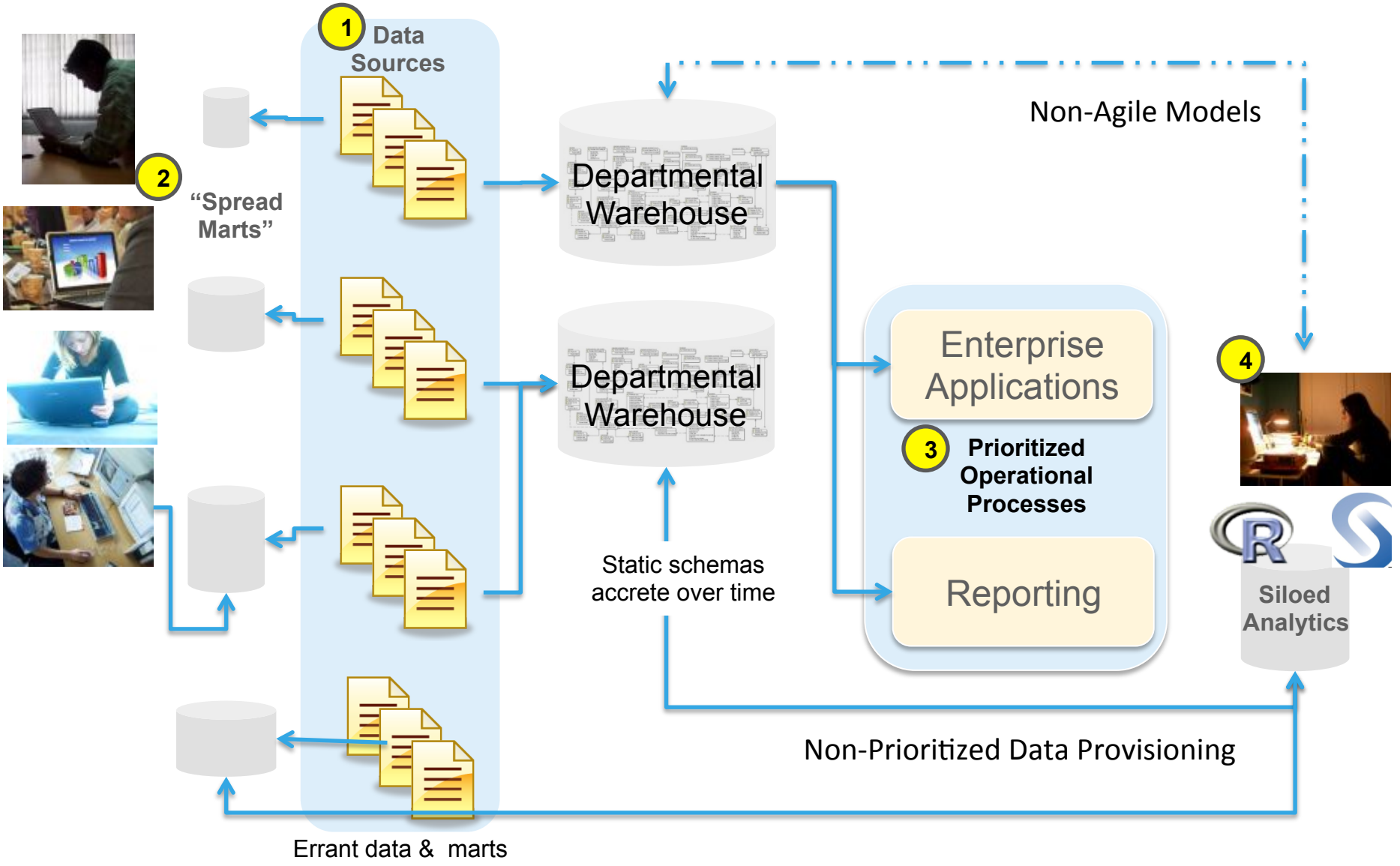
- Supports BI and reporting, but restricts robust analyses
- Analyst dependent on IT & DBAs for data access and schema changes
- Analysts must spend significant time to get extracts from multiple sources

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"*

A Typical Analytical Architecture



Implications of Typical Architecture for Data Science

- 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

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”