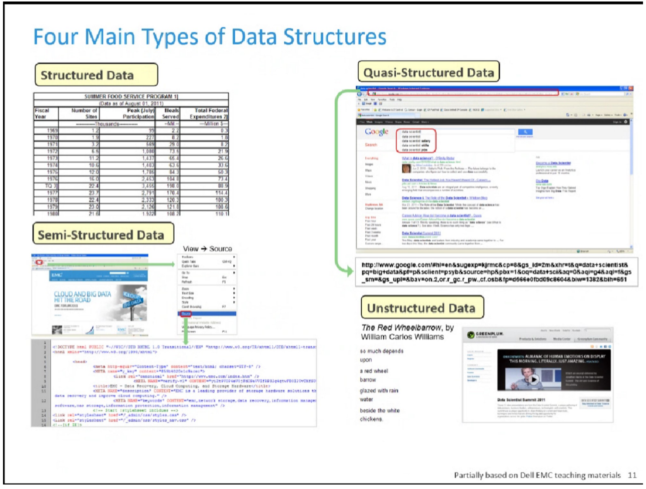
Lec 1

5 сентября 2024 г. 18:24

Notes:

1. Big data not new. It's just raw data of different scale and variety
2. 4V: volume(storage problems), velocity(real time analys challenge), variety(very different data), varacity(quality data challenge), variability, value
3. Analys models can be hacked by being given wrong data in learning process in order to get specific output on specific dataset.(For example: hacking banking credit analys model to get credit on big money aprooved)
4. Data can be structured and not structured.
5. Structured - database-like. Transactions and etc
6. Unstructured - videos, images, text. Data that should be analysed as a whole.
7. Semi structured - xml, json.
8. Quasi structured data - very exotic data formats. Not full or none documentation at all. Can be encrypted. Exp: clickstream data from web browser, that can be used to optimize site structure.
9. 

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

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.

GREENPLUM
ALMANAC OF HUMAN EMOTIONS ON DISPLAY THIS MORNING. LITERALLY. JUST AMAZING.
Data Scientist Summit 2011

Partially based on Dell EMC teaching materials 11

10. Data repositories problem. How to gather data? If every analyst have its own data extract, than developed model can be inaccurate on whole data. It's called data islands
11. Data islands - decentralized data marts.
12. Data warehouses - centralized data centers that store gathered and structured data. Can store BIG amount of data, but hard to maintain and expensive in resources. Not everything should be stored despite scale of storage. Resources are limited.
13. Data analysis and storage technologies are developing together
14. Many areas like banking prefer to use easy-interpreted models. Like regressive. Neural networks are hard to predict and rely on in some ways
15. Data lake - exact copy or near-exact copy of data. Unstructured. Organized and classified only when accessed. Convenient for analysis. Allows to use more advanced analysis technics.
16. Server oriented solutions are not good anymore due to petabytes of data that need to be send to data centre.
17. Decentralized storage is more convenient now. Files are stored in parts distributed between several machines.
18. Every machine analyze only its own part. Then all analysis results combined. Batch processing.
19. Query - user-friendly interface and language for non-programmers to access analytics system and get results. Enhance productivity among personnel.