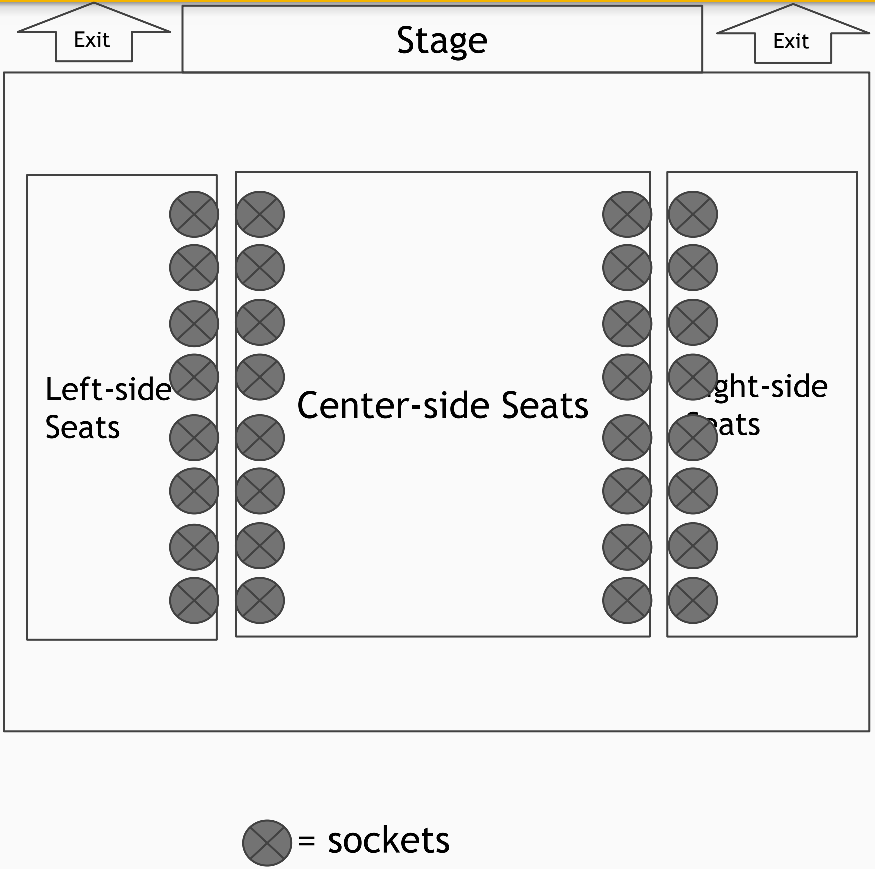
Yasumoto International Academic Park - YIA LT6

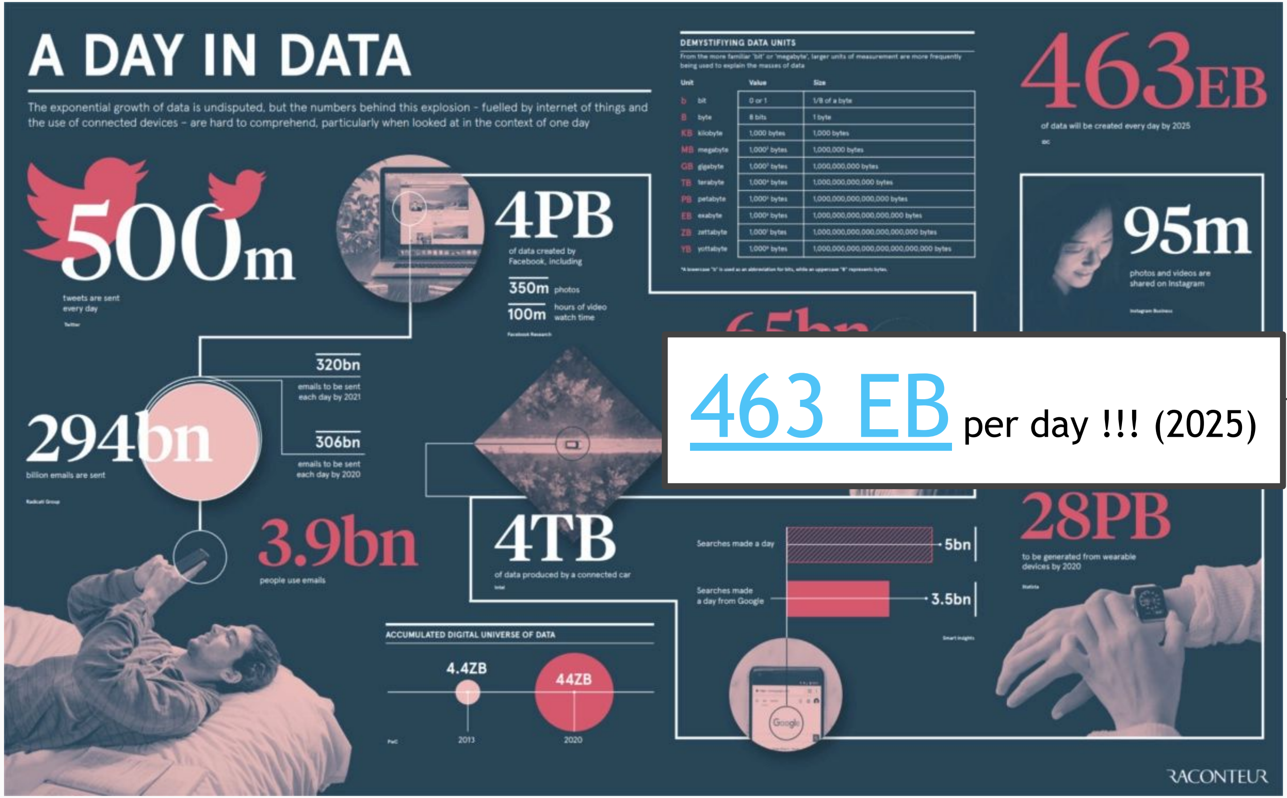
Limited Mobile Signal. Please use on-campus wifi.
32 Sockets. But please bring your own charger.



Python Data Structures
Regular Expressions

CUHK MSc Data Science & Biz Stat. Program
STAT5106 - Programming Techniques for Data Science
Week 3 @ 26 Sept 2024

What is Data ?



DEMYSTIFIYING DATA UNITS

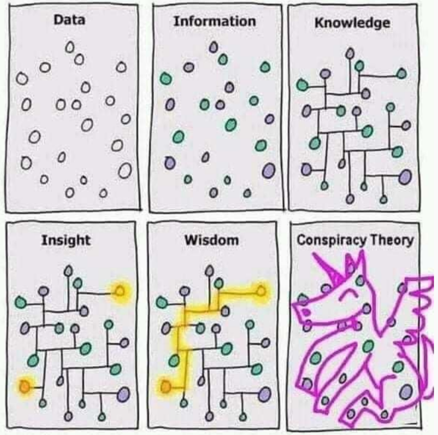
From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

Unit		Value	Size
b	bit	0 or 1	1/8 of a byte
B	byte	8 bits	1 byte
KB	kilobyte	1,000 bytes	1,000 bytes
MB	megabyte	$1,000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1,000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1,000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1,000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1,000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1,000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1,000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

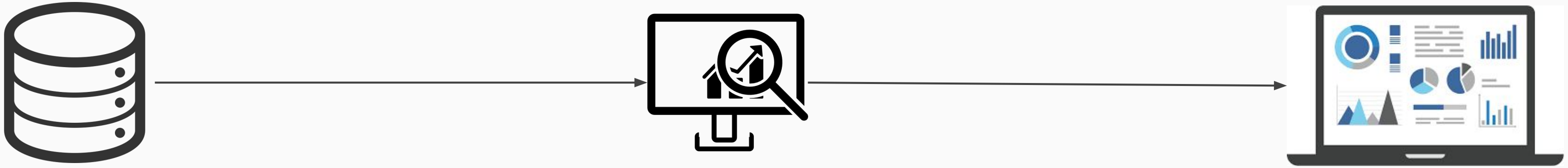
*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

463 EB per day !!! (2025)

See whether we can obtain...



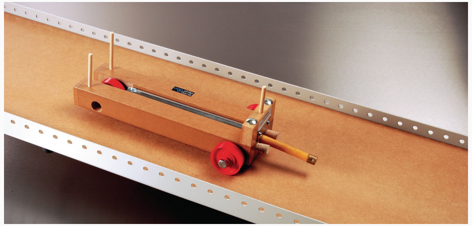
Data Analysis flow

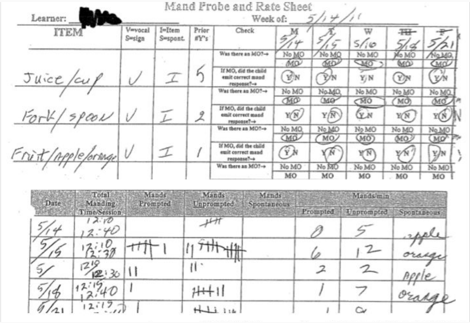


Data

Analysis

Visuals





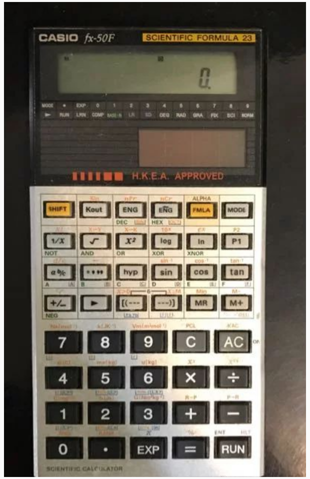
Mand Probe and Rate Sheet

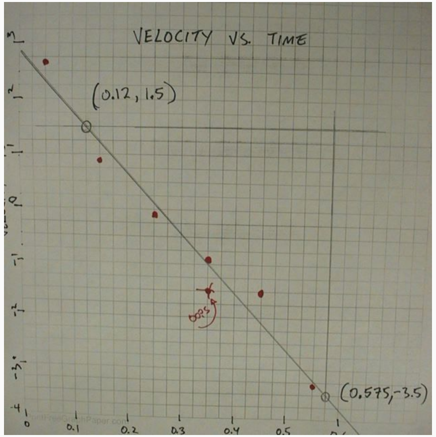
Learner: [redacted]

Week of: 5/14/11

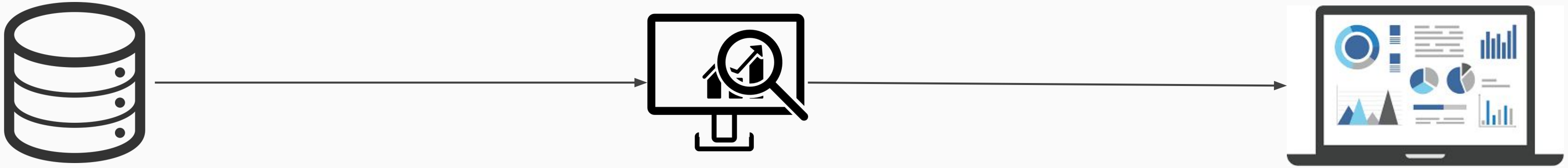
ITEM	V=vocal S=sign	I=Item S=spont.	Prior #Y's	Check	M 5/14	~~T~~ 5/15	W 5/16	~~TH~~ 5/18	~~F~~ 5/21
Juice/cup	V	I	5	Was there an MO?→	No MO [✓ MO]	No MO [✓ MO]	No MO [✓ MO]	No MO [✓ MO]	No MO [✓ MO]
				If MO, did the child emit correct mand response?→	[✓ Y] N	[✓ Y] N	Y N	[✓ Y] N	[✓ Y] N
Fork/spoon	V	I	2	Was there an MO?→	No MO [✓ MO]	No MO [✓ MO]	No MO [✓ MO]	No MO [✓ MO]	No MO [✓ MO]
				If MO, did the child emit correct mand response?→	Y [✓ N]	Y [✓ N]	[✓ Y] N	Y [✓ N]	Y [✓ N]
Fruit/apple/orange	V	I	1	Was there an MO?→	No MO [✓ MO]	No MO [✓ MO]	No MO [✓ MO]	No MO [✓ MO]	No MO [✓ MO]
				If MO, did the child emit correct mand response?→	[✓ Y] N	Y [✓ N]	Y [✓ N]	Y [✓ N]	[✓ Y] N
				Was there an MO?→	No MO MO	No MO MO	No MO MO	No MO MO	No MO MO

Date	Total Manding Time/Session	Mands Prompted	Mands Unprompted	Mands Spontaneous	Mands/min		
					Prompted	Unprompted	Spontaneous
5/14	12:10 12:40		卌		0	5	apple
5/15	12:10 12:30	卌 I	II 卌 卌		6	12	orange
5/	12:19 ~~12~~:30	II	II		2	2	Apple
5/18	12:19 12:40	I	卌 II		1	7	Orange
5/21	12:19	I	卌 II		1	[illegible]	





Data Analysis flow



Data

Analysis

Visuals

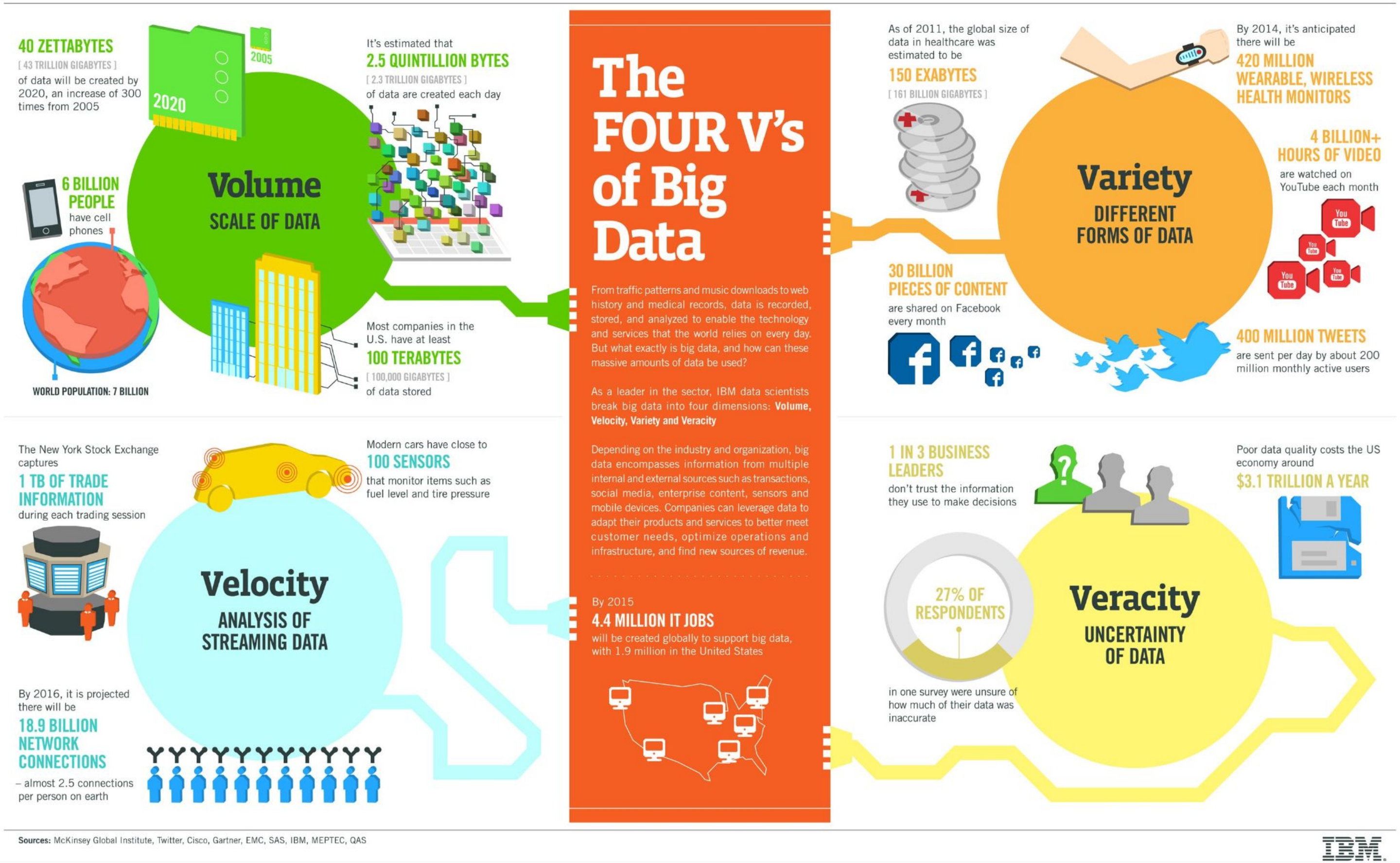
- xlsx
- csv
- sas7bdat
- SQL db
- Datalake
- HBase
- jpg
- mp4
- ...

- Excel chart
- R ggplot
- SAS VA
- Python matplotlib
- Tableau
- PowerBI
- Google DataStudio
- ...

What is BIG DATA ?

大

多



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate

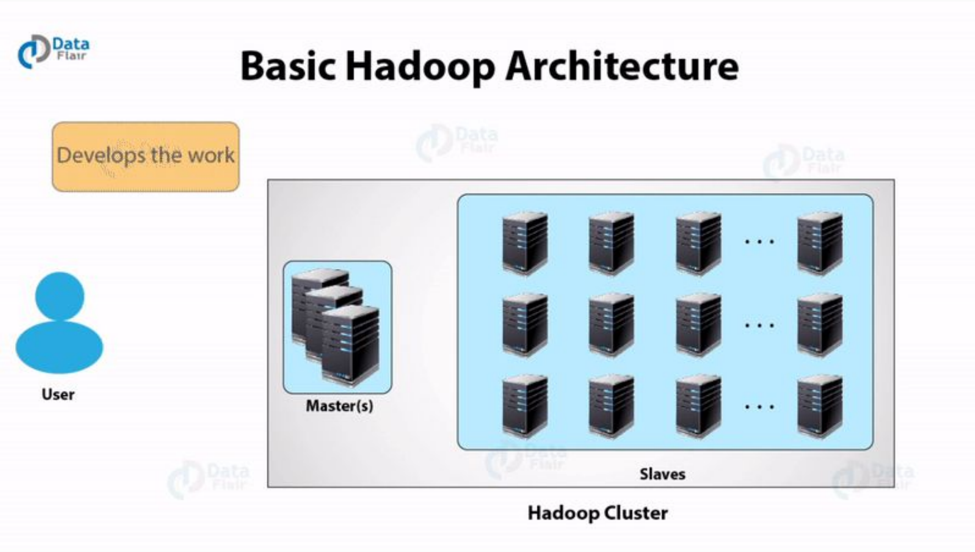
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

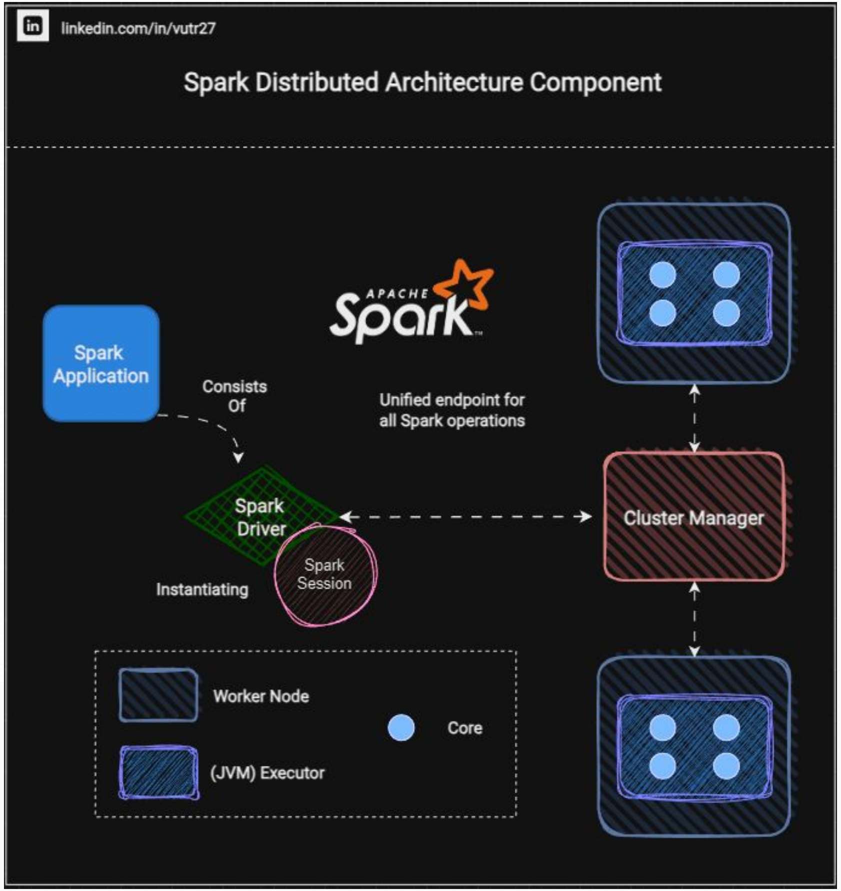
IBM.

快

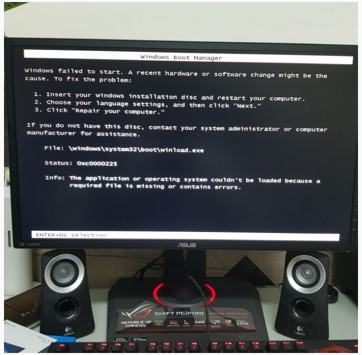
真

Big Data Computing Architecture - Hadoop & Spark



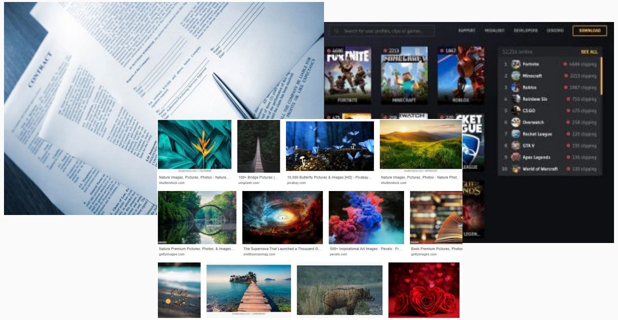


Data Structures, Algorithm in Programming



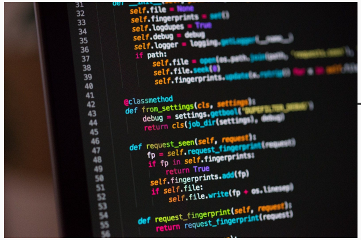
```
Windows Boot Manager

Windows failed to start. A recent hardware or software change might be the
cause. To fix the problem:

  1. Insert your Windows installation disc and restart your computer.
  2. Choose your language settings, and then click "Next."
  3. Click "Repair your computer."

If you do not have this disc, contact your system administrator or computer
manufacturer for assistance.

    File: \windows\system32\boot\winload.exe

    Status: 0xc0000225

    Info: The application or operating system couldn't be loaded because a
          required file is missing or contains errors.

ENTER=OS Selection
```

(Big) Data



+

Algorithm



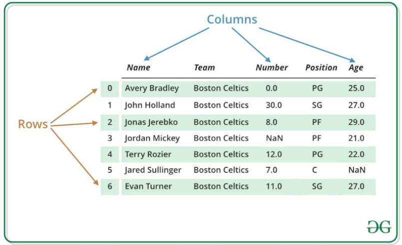
```python
    def __init__(self,
        self.file = None
        self.fingerprints = set()
        self.logdupes = True
        self.debug = debug
        self.logger = logging.getLogger(__name__)
        if path:
            self.file = open(os.path.join(path, 'requests.seen'),
            self.file.seek(0)
            self.fingerprints.update(x.rstrip() for x in self.file

    @classmethod
    def from_settings(cls, settings):
        debug = settings.getbool('DUPEFILTER_DEBUG')
        return cls(job_dir(settings), debug)

    def request_seen(self, request):
        fp = self.request_fingerprint(request)
        if fp in self.fingerprints:
            return True
        self.fingerprints.add(fp)
        if self.file:
            self.file.write(fp + os.linesep)

    def request_fingerprint(self, request):
        return request_fingerprint(request)
```

- Algorithm: (coding in previous lesson)
  A set of rules or steps used to solve a problem
- Data Structure: (variables)
  A particular way of organizing data in a computer

- NOW more Data Structure

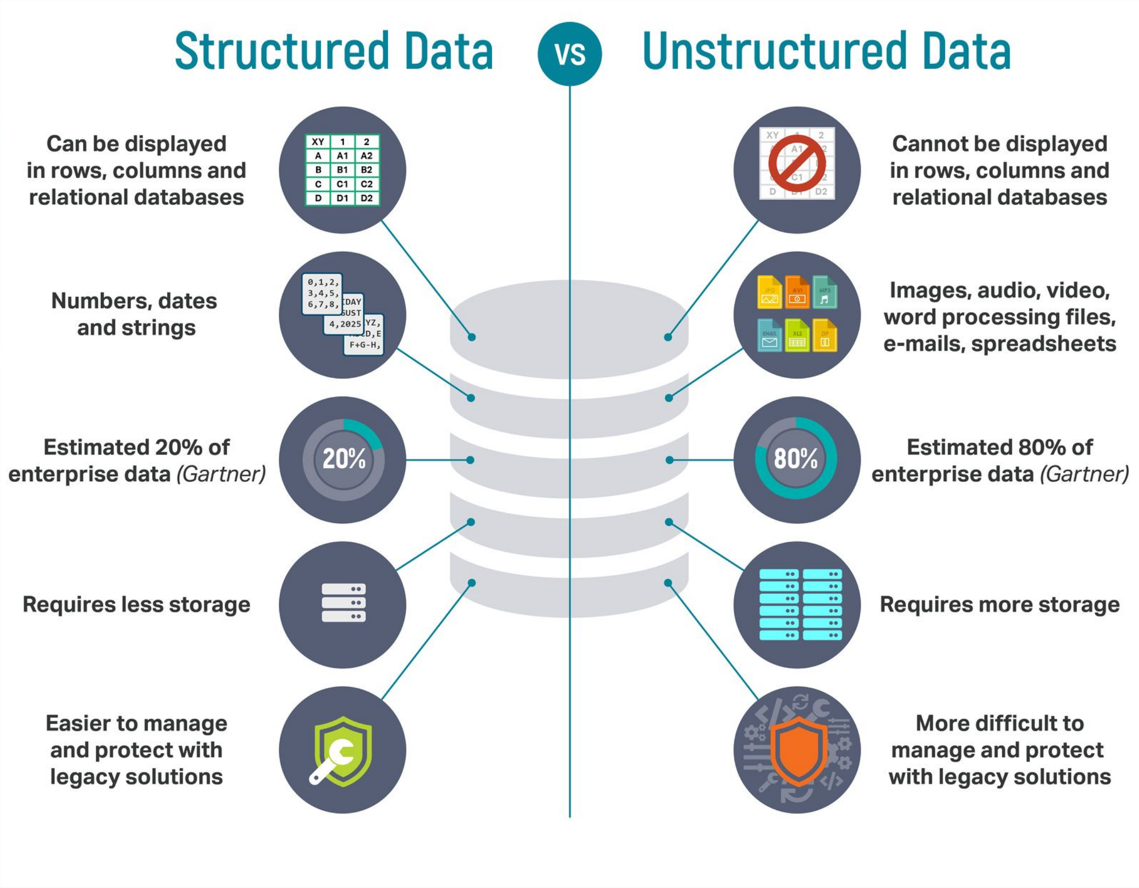
In Python	In R
Tuple / List / Dictionary	Vector / List
Series / DataFrame in Pandas ✓ & Dictionary	data.frame / tibble / data.table
Array in Numpy	Matrix

Structured Data vs Unstructured Data



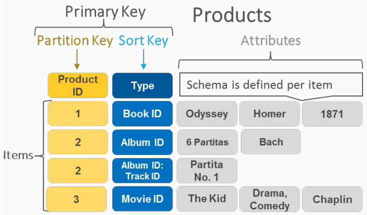
Columns
Rows

	Name	Team	Number	Position	Age
0	Avery Bradley	Boston Celtics	0.0	PG	25.0
1	John Holland	Boston Celtics	30.0	SG	27.0
2	Jonas Jerebko	Boston Celtics	8.0	PF	29.0
3	Jordan Mickey	Boston Celtics	NaN	PF	21.0
4	Terry Rozier	Boston Celtics	12.0	PG	22.0
5	Jared Sullinger	Boston Celtics	7.0	C	NaN
6	Evan Turner	Boston Celtics	11.0	SG	27.0



Structured Data vs Unstructured Data

Structured Data	Unstructured Data
Can be displayed in rows, columns and relational databases	Cannot be displayed in rows, columns and relational databases
Numbers, dates and strings	Images, audio, video, word processing files, e-mails, spreadsheets
Estimated 20% of enterprise data *(Gartner)*	Estimated 80% of enterprise data *(Gartner)*
Requires less storage	Requires more storage
Easier to manage and protect with legacy solutions	More difficult to manage and protect with legacy solutions



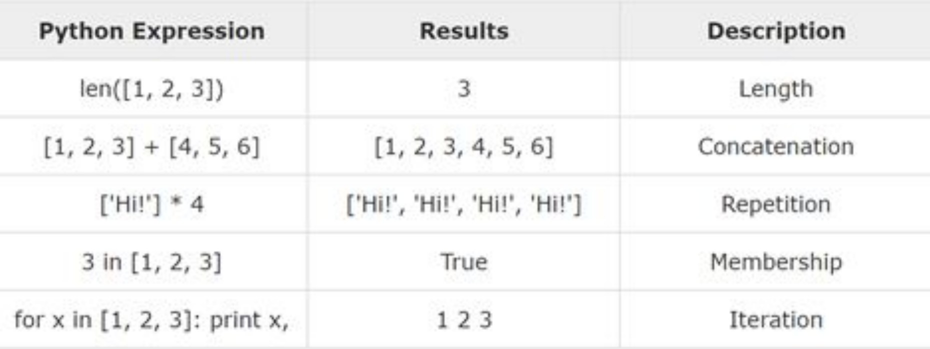
Product ID	Type	Schema is defined per item		
1	Book ID	Odyssey	Homer	1871
2	Album ID	6 Partitas	Bach	
2	Album ID: Track ID	Partita No. 1		
3	Movie ID	The Kid	Drama, Comedy	Chaplin

Start Coding...

Please access...Week_3_Tuple_List_Dict.ipynb

Lists, Tuples, Dictionaries

- Tuple - `x = (1, 'a', 2, 'b’)`
  - cannot be altered
- List - `x = [1, 'a', 2, ‘b’]`
  - can be altered
- Dictionary - `x = {‘a’: 1; ‘b’: 2}`
  - Collection of key-value pairs
  - `x.keys() = [‘a’, ‘b’] ;`
    `x.values() = [1, 2]`
- Note
  - string = tuple of characters
  - Tuple & List can use slicing properties (`x[0:5]`)
  - List vs Dictionary
    - List: A linear collection of values that stay in order
    - Dictionary: A “bag” of values, each with its own label
  - Dictionary input = JSON



Python Expression	Results	Description
len([1, 2, 3])	3	Length
[1, 2, 3] + [4, 5, 6]	[1, 2, 3, 4, 5, 6]	Concatenation
['Hi!'] * 4	['Hi!', 'Hi!', 'Hi!', 'Hi!']	Repetition
3 in [1, 2, 3]	True	Membership
for x in [1, 2, 3]: print x,	1 2 3	Iteration

Shorten For loops

NOTE: This is not R - NO VECTORLIZED CALCULATION ! until you have pandas and numpy.

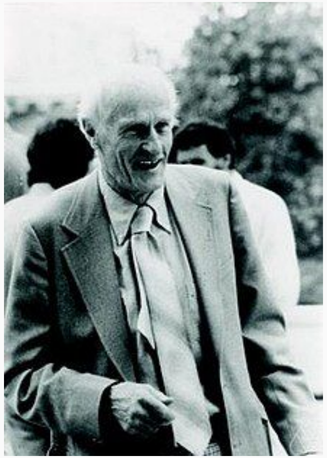
- Map
  - `Map(function f, list)` = output each f(x) in list
- Lambda = one line function
  - `my_function = lambda a, b, c : a + b`
- List Comprehensions
  - `my_list = [number for number in range(0,1000) if number % 2 == 0]`

Start Coding...again

Please access...Week_3_Regular_Expressions.ipynb

Regular Expressions

- The concept arose in the 1950s when the American mathematician Stephen Cole Kleene formalized the description of a regular language - An area of Math Logic.
- Very powerful in string searching and quite cryptic
- Fun once you understand them
- Again, further reference from Dr. Chuck



Basics

.	Matches any character
*	Repeats a character zero or more times
+	Repeats a character one or more times
?	Appears a character zero or one time only

Regular Expression Cheat Sheet

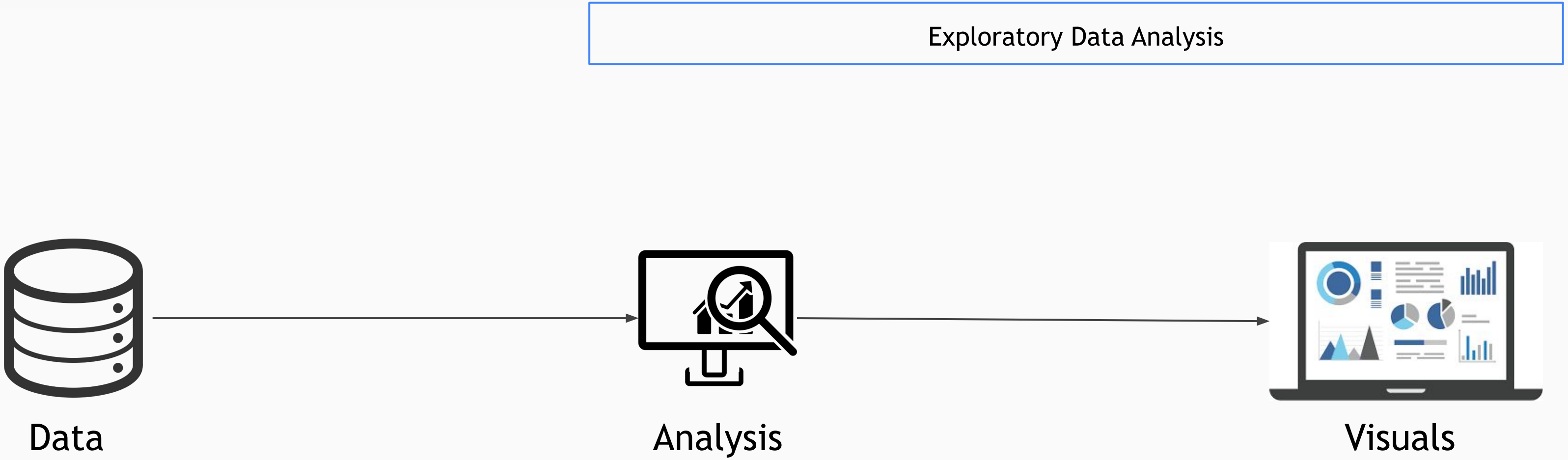
^	Matches the beginning of a line
$	Matches the end of the line
\s	Matches whitespace
\S	Matches any non-whitespace character
[aeiou]	Matches a single character in the listed set
[^XYZ]	Matches a single character not in the listed set
[a-z0-9]	The set of characters can include a range
(	Indicates where string extraction is to start
)	Indicates where string extraction is to end
*?	Repeats a character zero or more times (non-greedy)
+?	Repeats a character one or more times (non-greedy)

Use in python

```
import re

re.search(pattern, text)
re.findall(pattern, text)
```

More Example in the jupyter notebook...

Data Analysis Process



Exploratory Data Analysis

Data

Analysis

Visuals

pandas

NumPy

Week 4, 7

StatsModels
Statistics in Python

Week 8

matplotlib

Seaborn

Week 9

re / urllib /
beautifulsoup

Week 5, 6

Assignment 2

Please check the link of Assignment 2 here

To be continue…