Advanced Topics in Data Management

Dr. Neamat el Tazi

Pre-Masters Course (2015-2016)

Information Systems Department
Cairo University, Egypt

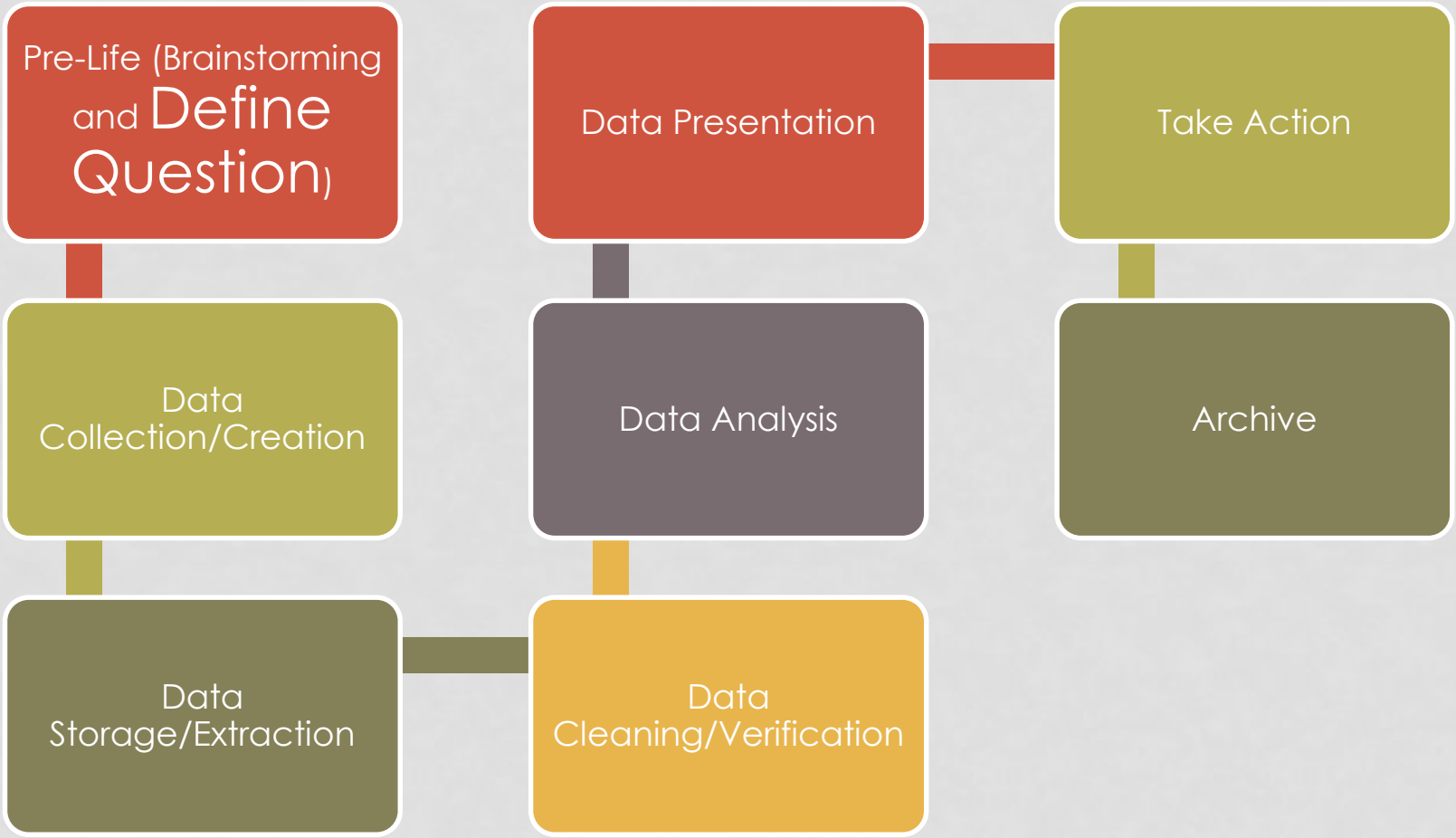
DATA MANAGEMENT

"Data management is the **development**, **execution** and **supervision** of plans, policies, programs and practices that **control**, **protect**, **deliver** and **enhance** the value of data and information assets"

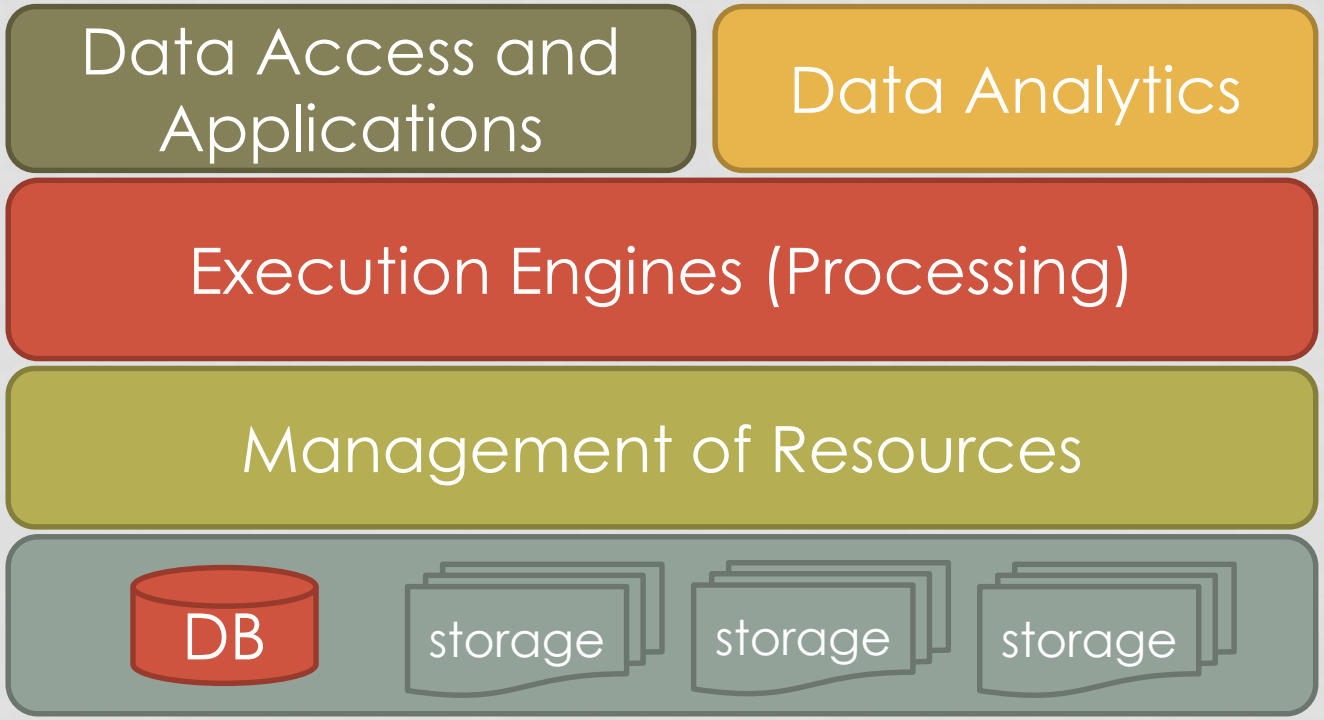
~Wikipedia

(https://en.wikipedia.org/wiki/Data_management)

DATA LIFE CYCLE



TECHNOLOGY STACK



WHAT IS DATA???

Value assigned to a thing

DATA

TYPES OF DATA

- **Qualitative**: describes the quality of an object. Example: color, texture, description…etc
- **Quantitative**: refers to a countable piece of data. Example: Number of umbrellas in the photo.
- **Categorical**: puts the data into categories. Example: small, medium and large umbrellas
- **Discrete**: Numerical data that has gaps. Example: Number of umbrellas will be a whole number, it can't be a float one. Number of chairs in the café…etc.
- **Continuous**: Numerical data with no gaps. Data can take any value. Example: Dimensions of the café.

DATA -> INFORMATION

Data Item	Value
Object	Umbrella
Context	Café
Color	Yellow
Size	1 m2
Condition	New
…	

INFORMATION-> KNOWLEDGE

- Knowledge is created when information is analyzed and understood.

- **Example**: One can say that such umbrellas are not useful for personal use because of their size and weight.

STRUCTURED VS. UNSTRUCTURED

There are 6 yellow umbrellas in Solo café.

Unstructured Data

Structured Data

Count	6
Object	Umbrella
Café Name	Solo
Color	Yellow

Structured Data

Count:6, object: umbrella, café: solo, color: yellow

DEFINE QUESTION

- What do you want to find in the data?
- Are you searching for an irregular pattern?

"You never know what you might find in a dataset, so just have a look" ~ Caelainn Barr, Citywire

http://datajournalismhandbook.org/1.0/en/understanding_data_4.html

DATA COLLECTION

- Where to find data?
- Is the data created and stored and you're only searching inside it?
- Are you going to create the data?

3 Sources of Data:

1. Searching for data that have been already released
2. Asking for release of data from a data source.
3. Collecting data

DATA COLLECTION

- Finding Data:
  - **Government**
    - data.gov.uk (UK)
    - www.data.gov (US)
    - Dados.gov.br (Brazil)
    - Opendata.go.ke (Kenya)
    - Datacatalogs.org (Check other countries)
  - **Organizations**
    - World Bank (data.worldbank.org)
    - World Health Organization (http://www.who.int/research/en/)
  - **Science**
    - NASA (http://data.nasa.gov/), Dryad (http://datadryad.org/)
  - **Data Repository List** (http://oad.simmons.edu/oadwiki/Data_repositories)
  - **Open Knowledge Foundation** (http://datahub.io/)

Source: http://schoolofdata.org/handbook/courses/finding-data/

DATA FORMAT

- TIP:

You can search for data in Google with a certain format by adding "**filetype:**csv" to your search keywords.

BASICS READINGS

- Basic Math To Start:

http://schoolofdata.org/handbook/courses/the-math-you-need-to-start/

- Data Representation:

http://schoolofdata.org/handbook/courses/data-to-diagrams/

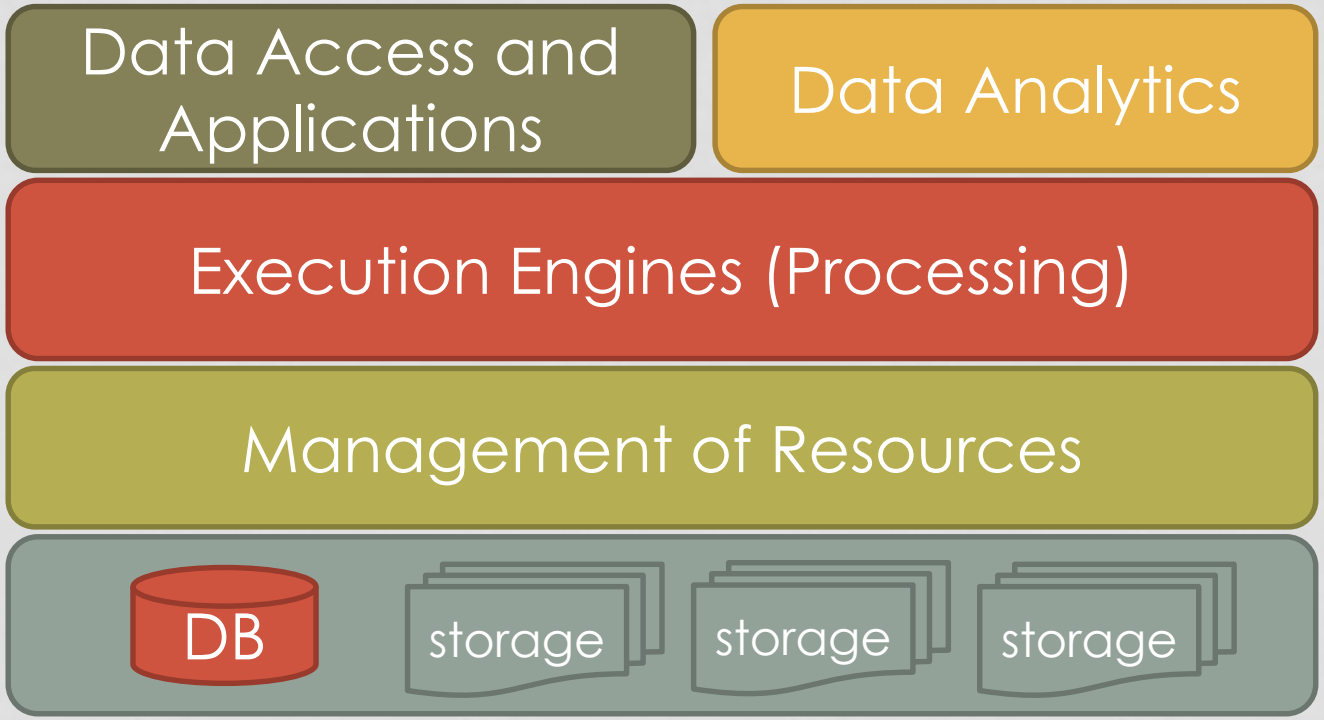
- Common Misconceptions:

http://schoolofdata.org/handbook/courses/common-misconceptions/

- Data Provenance:

http://schoolofdata.org/handbook/courses/data-provenance/

TOPICS


Data Access and Applications
Data Analytics
Execution Engines (Processing)
Management of Resources
DB
storage
storage
storage

TOPICS

- **Data Storage/Databases**(HDFS, S3 on AWS, HBase, NFS with MapR, CloudStore,..)
- **Data Processing** (MapReduce Framework)
- **Data Access** (MLlib, Hive(SQL), Pig (data flow),Shark, Avro (JSON), Mahout (Machine Learning), Sqoop (Data Connector)
- **Management** (OOzie (workflow), EMR (AWS worldflow), Chukwa (Monitoring), Flume(Monitoring), Zookeepter(Mgmt)
- **Data Analytics** (Query, Reporting, Data Mining, Predective Analysis)

TOPICS AND PAPERS

1. Data Storage/Distributed File Systems
   - **Google File System (2003) –Reading Assignment**: -- http://static.googleusercontent.com/media/research.google.com/en//archive/gfs-sosp2003.pdf
   - **HDFS (2010) – Reading Assignment:**

     http://zoo.cs.yale.edu/classes/cs422/2014fa/readings/papers/shvachko10hdfs.pdf
   - **Google Spanner (2012)- Distributed Database – Reading Assignment:**

     http://static.googleusercontent.com/media/research.google.com/en//archive/spanner-osdi2012.pdf
   - **BigTable (2006) – Reading Assignment:**

     http://static.googleusercontent.com/media/research.google.com/en//archive/bigtable-osdi06.pdf

TOPICS AND PAPERS

2. Data Processing

- **Map Reduce (**2004) –**Reading Assignment**

http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf

3. Data Access

- **Survey**: A Survey of Large-Scale Analytical Query Processing in MapReduce (VLDB 2013) –**Recommended** http://www.chinacloud.cn/upload/2013-06/13061810335030.pdf

TOPICS AND PAPERS

3. Data Access
   - **MLlb (CIDR 2013) --Recommended**

   (http://www.cidrdb.org/cidr2013/Papers/CIDR13_Paper118.pdf)
   - **Hive (ICDE 2010) – Reading Assignment**

   (http://infolab.stanford.edu/~ragho/hive-icde2010.pdf)
   - **Pig (SIGMOD 2008) – Recommended**

   (http://infolab.stanford.edu/~olston/publications/sigmod08.pdf)
   - **Shark (SIGMOD 2013) – Reading Assignment**

   (https://people.csail.mit.edu/matei/papers/2013/sigmod_shark.pdf)
   - **Spark (HOTCloud 2010) --Recommended**

   (http://www.cs.berkeley.edu/~matei/papers/2010/hotcloud_spark.pdf)

READING ASSIGNMENTS

1. Google File System (2003):
http://static.googleusercontent.com/media/research.google.com/en//archive/gfs-sosp2003.pdf

2. HDFS (2010):
http://zoo.cs.yale.edu/classes/cs422/2014fa/readings/papers/shvachko10hdfs.pdf

3. Map Reduce (2004)
http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf

4. Google Spanner (2012):
http://static.googleusercontent.com/media/research.google.com/en//archive/spanner-osdi2012.pdf

5. Hive (2010):
http://infolab.stanford.edu/~ragho/hive-icde2010.pdf

6. Shark(2013):
https://people.csail.mit.edu/matei/papers/2013/sigmod_shark.pdf

7. BigTable (2006):
http://static.googleusercontent.com/media/research.google.com/en//archive/bigtable-osdi06.pdf

PAPER READING AND REVIEWING

- **How to Read a paper:**

http://ccr.sigcomm.org/online/files/p83-keshavA.pdf

- **How to Review a paper:**

https://people.inf.ethz.ch/troscoe/pubs/review-writing.pdf

READING ASSIGNMENTS

1. Read the paper
2. Prepare a review document of **at most 2 pages** before the deadline that will be listed in the schedule. Failure to send it by deadline, will result in no marks for that reading assignment.
   1. **A review template document** is available as a guide for you to write your reviews.
   2. **Negative 5 grades** will be deducted from any student(s) who attempt to copy (internally or externally) their assignment.

IBM BLUEMIX

- **IBM Bluemix** is a cloud platform as a service (PaaS) developed by IBM. It supports several programming languages and services as well as integrated DevOps to build, run, deploy and manage applications on the cloud. Bluemix is based on Cloud Foundry open technology and runs on SoftLayer infrastructure.

- Bluemix supports several programming languages including **Java**, **Node.js**, **Go**, **PHP**, **Python**, **Ruby Sinatra**, **Ruby on Rails** and can be extended to support other languages such as **Scala** through the use of buildpacks.

Source: https://en.wikipedia.org/wiki/Bluemix

IBM BLUEMIX

- Availability of Promo Codes for you to use Bluemix
- Extension for 6 months.
- Build your project using IBM bluemix

PROJECT

- Data Analysis Project.
- Research Area Topic that involves data management or data analysis.
- Project is a **group of 3 students max.**

- You can use **IBM Bluemix Cloud account**.. Ask for the promo code to give it to you.
- You can use **IBM Watson for data analytics** .. Ask for the promo code to give it to you.

GRADING SCHEME

- **Reading Assignments**: 21%
  - 3% on each paper review.
- **Paper Presentation**: 9%
- **Project Presentation**: 20%
- **Project Paper and Implementation**: 20%

- **Final**: 30% -- Project grade will be given as the rest of the 30% of the legal grades of the final project and presentation.