CS286 AI for Science and Engineering

Lecture 12: Systems Perspectives of AI: Big Data Techniques

Shu Yin (殷树)

PhD, Assistant Professor

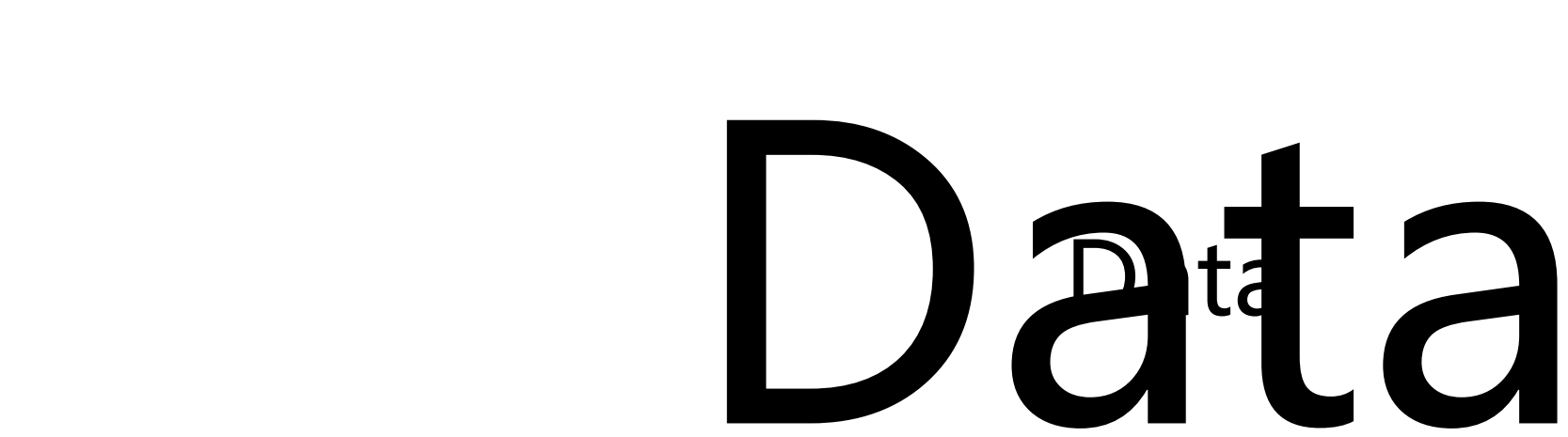
School of Information Science and Technology (SIST), ShanghaiTech University

Fall Semester, 2020

Outline

- What’s Big Data
- Thinking Reform
- Summary

What is Big Data?



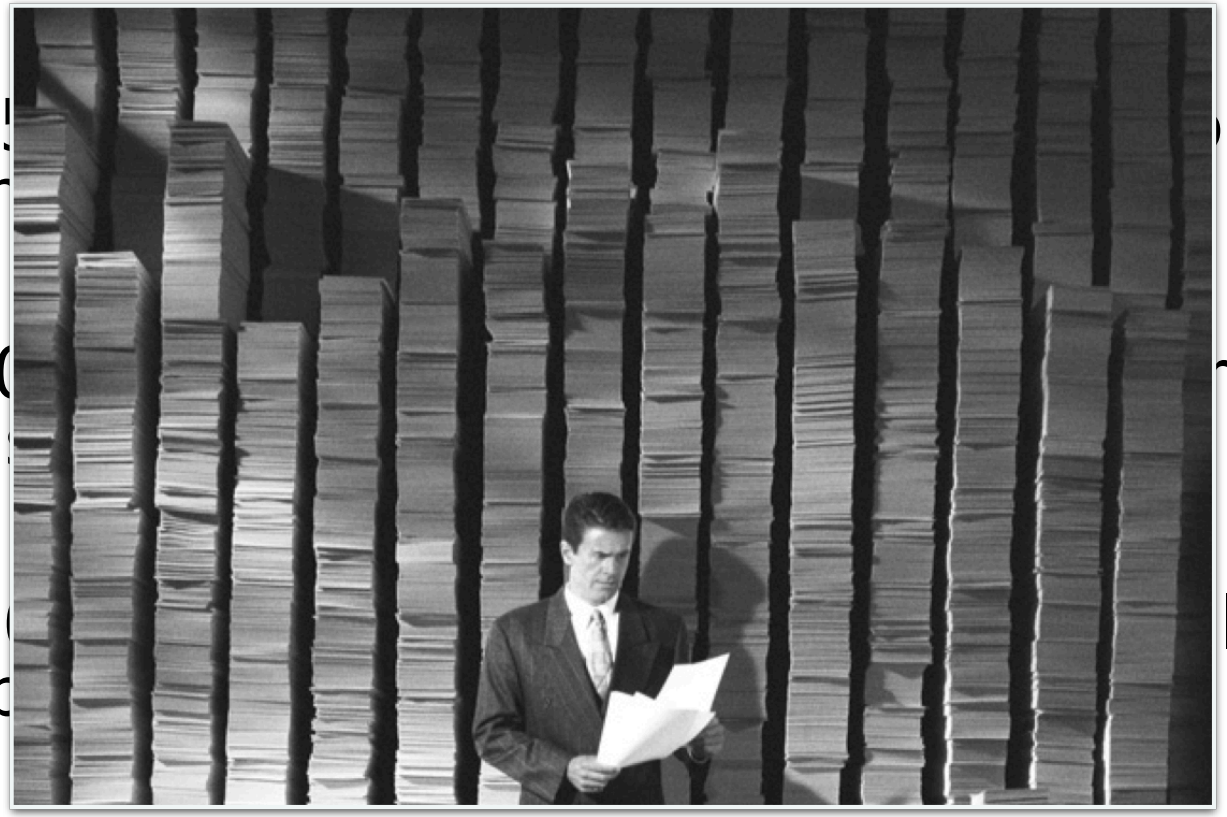
What is Big Data?

"Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." -- Gartner

Which was derived from:

"While enterprises struggle to consolidate systems and collapse redundant databases to enable greater operational, analytical, and collaborative consistencies, changing economic conditions have made this job more difficult. E-commerce, in particular, has exploded data management challenges along three dimensions: ***volumes, velocity and variety***. In 2001/02, IT organizations much compile a variety of approaches to have at their disposal for dealing each." – Doug Laney

- Volume
  - Convert 3 ...etter predict power cor...
- Velocity
  - Analyze 5... ...real-time to predict cu...
- Variety
  - Monitor 1... ...lance cameras to target p...
- Veracity



What Can We Do with Big Data

- Example: H1N1 in 2009 (A Very Old example)
  - CDC : a 2-week lag
  - Google: a few weeks before

How Does Google Do?

- 50 million most common search terms
- Compare with CDC data 2003-2008
- Look for correlations
- 450 million different mathematical models
- Struck gold: strong correlation

So …

- Google could tell where the flu has spread
- Could tell it in near real time
- Not weeks after the fact, like CDC
- Ironically, COVID-19 …

Big Data Refers to

- Things can do at a large scale
- To extract new insights
- To create new forms of value
- Change markets, organizations, and more…

Some Numbers

- Google: 1.2 Trillion searches in 2012 (we don’t have updated data afterwards)
- Youtube: 800 million 1-hr video every second
- Tweet: 500 million tweets per day in 2013

The Core of Big Data

- Predictions
- Not to “teach” a computer to “think”
- Apply math to infer probabilities
  - Email spams
  - Auto-correction
  - Self-driving cars
- The key is that system perform well because of huge amount of data

The Core of Big Data

- Big Data is all about seeing and understanding the relations
- Let the data “speak to you”
- Three major shifts of mindset:
  - Vast vs. Small amounts of data
  - Messiness vs. Exactitude
  - Correlation vs. Causality

Thinking Reform

- More
- Messy
- Correlation

Think Reform

More:
Using all the data at hand
instead of just a small portion of it

More:
not just the random sample, but the entire data

- From some to all
  - What we did: sampling
  - Aim: to confirm the richest finding using the smallest amount of data
  - But: how to choose a sample:
    - Randomness matters more than sample size
  - So: bottleneck is Randomness

Think Reform： More （example）

- Light Field Camera

Think Reform： More

- With MORE data, the information is more re-useable
- Big data allow us to look at details or explore new analysis w/o risk of blurriness

Think Reform

Messy:
Increasing the volume opens the door to inexactitude

Messy:
not just accuracy, but mix

- Embrace the in-accuracy
  - What we did: small data
  - Aim: reduce errors & ensure high data quality
  - But: errors may get amplified
    - A limited number of data
  - So: let the probability speak

Think Reform： Messy （example）

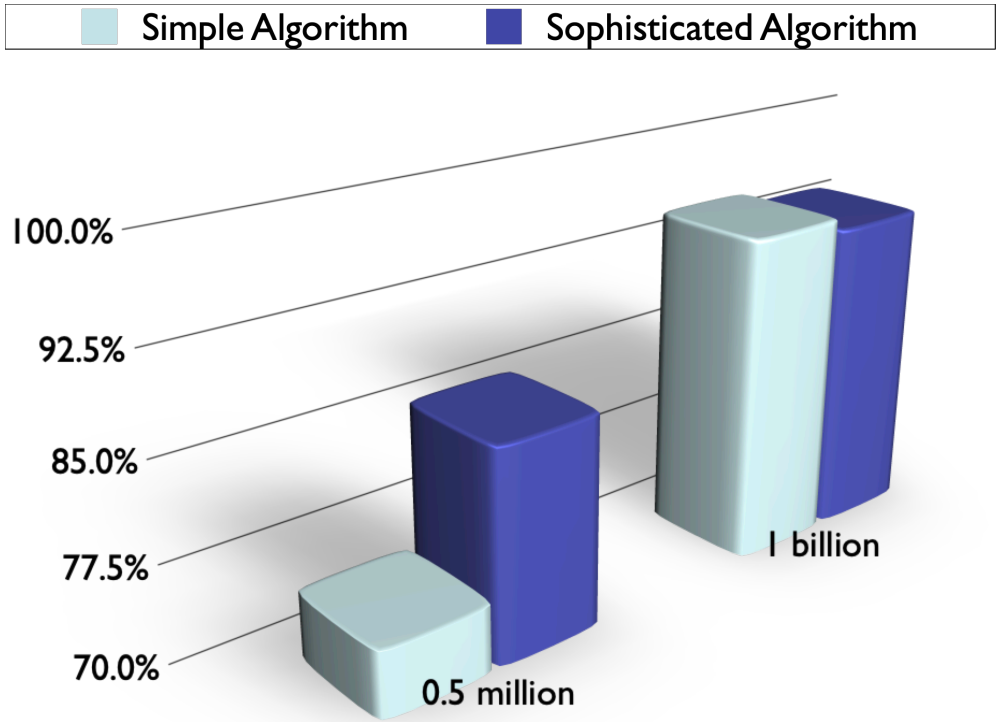
- Google Translate
  - Works the best
  - Not because of a smarter algorithm
  - But because of more data
    - Billions of pages
    - 95 billion sentences
    - Trillion words

simple vs. sophisticated algorithm

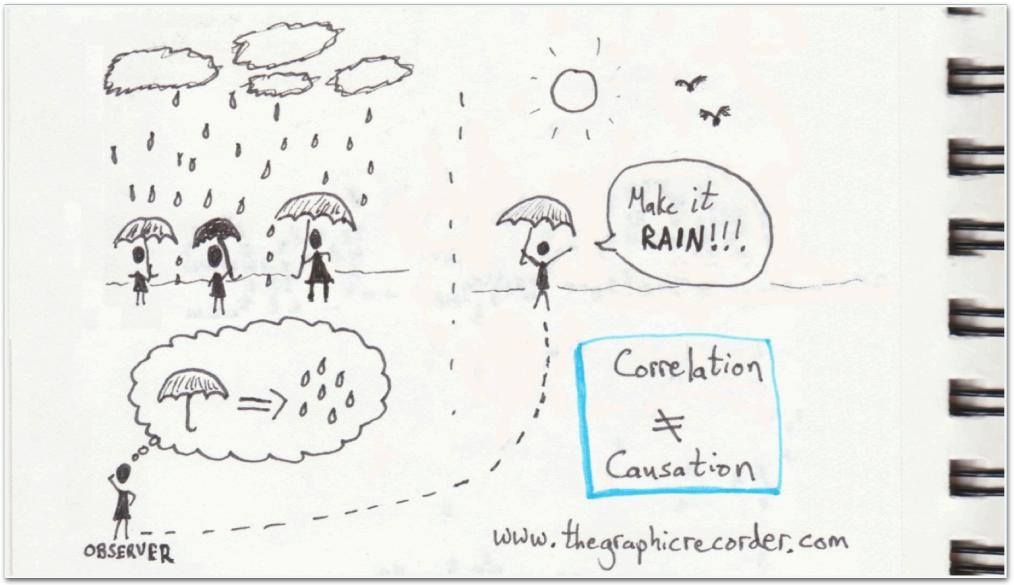
- Simple algorithm + big data
- Sophisticated algorithm + small data

"These results suggest that we may want to reconsider the tradeoff between spending time and money on algorithm development versus spending it on corpus development"

--Michele Banko, Microsoft R&D



Think Reform



Correlation:
The idea of understanding the reasons behind all that happens

Correlation:
“what” is enough, no need for “why”

- Prediction
  - What we did: choose proxy manually
  - Aim: forecast
  - But: some correlation maybe neglect
    - A limited number of data
  - So: let the machine do the job

Think Reform: Correlation (example)

- Notice: prior to a hurricane
  - Sales of flashlights increase
  - Also: sales of Pop-Tarts increase
  - So, as storm apporaches
    - Stocked boxes of Por-Tarts at the front of stores next to the hurricane supplies

Heart of Big Data

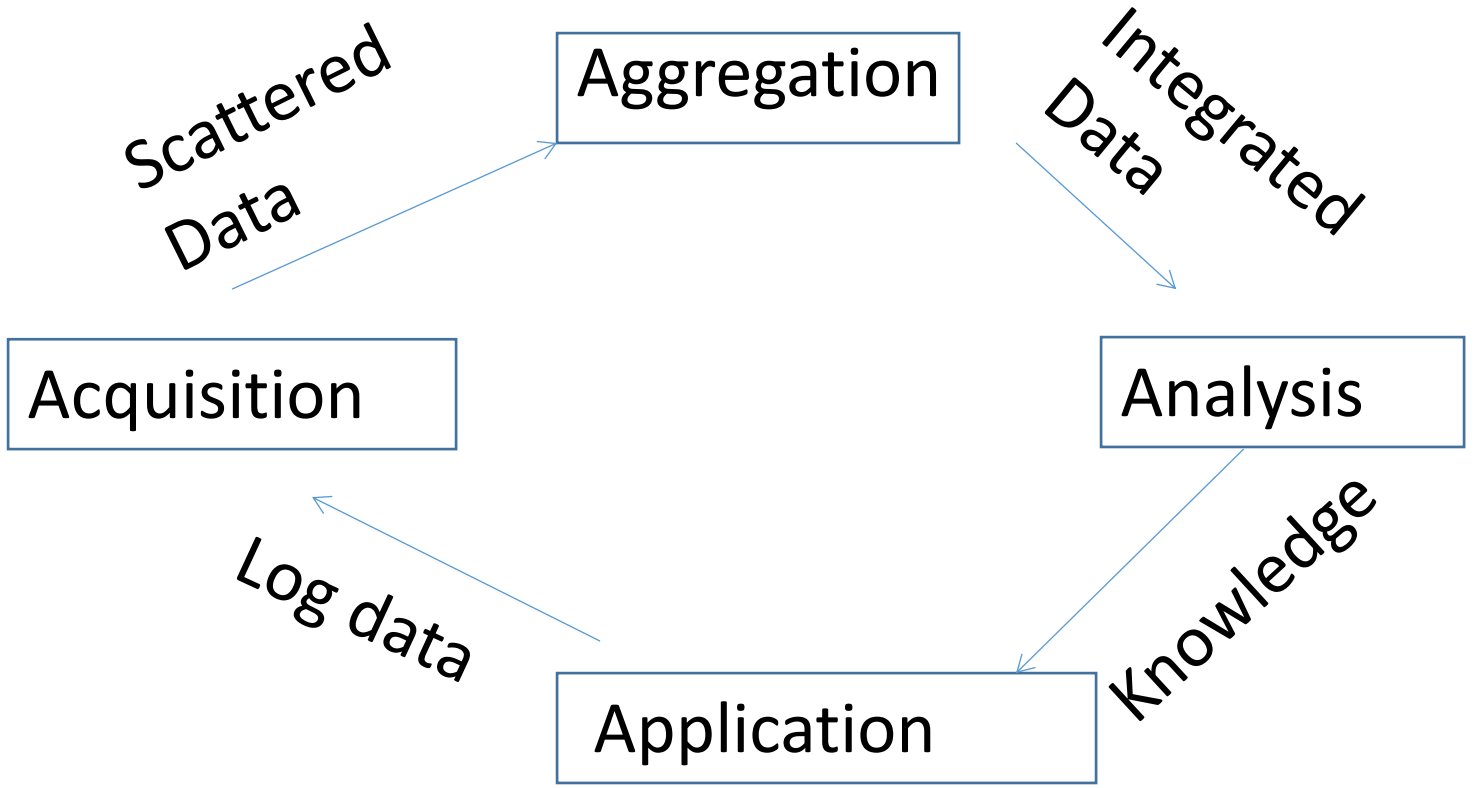
- Predictions based on correlations
- Aim:
  - Identify a good proxy
  - Watch it
  - Predict future events
- E.g., UPS trucks
  - What they did: replace certain parts after 2-3 years
  - Aim: avoid breakdown on the road
  - But: inefficient – some of them were fine
  - So: switch predictive analysis
    - Measure and monitor individual parts
    - Replace them when necessary
    - One case: new vehicles had a defective parts

Heart of Big Data

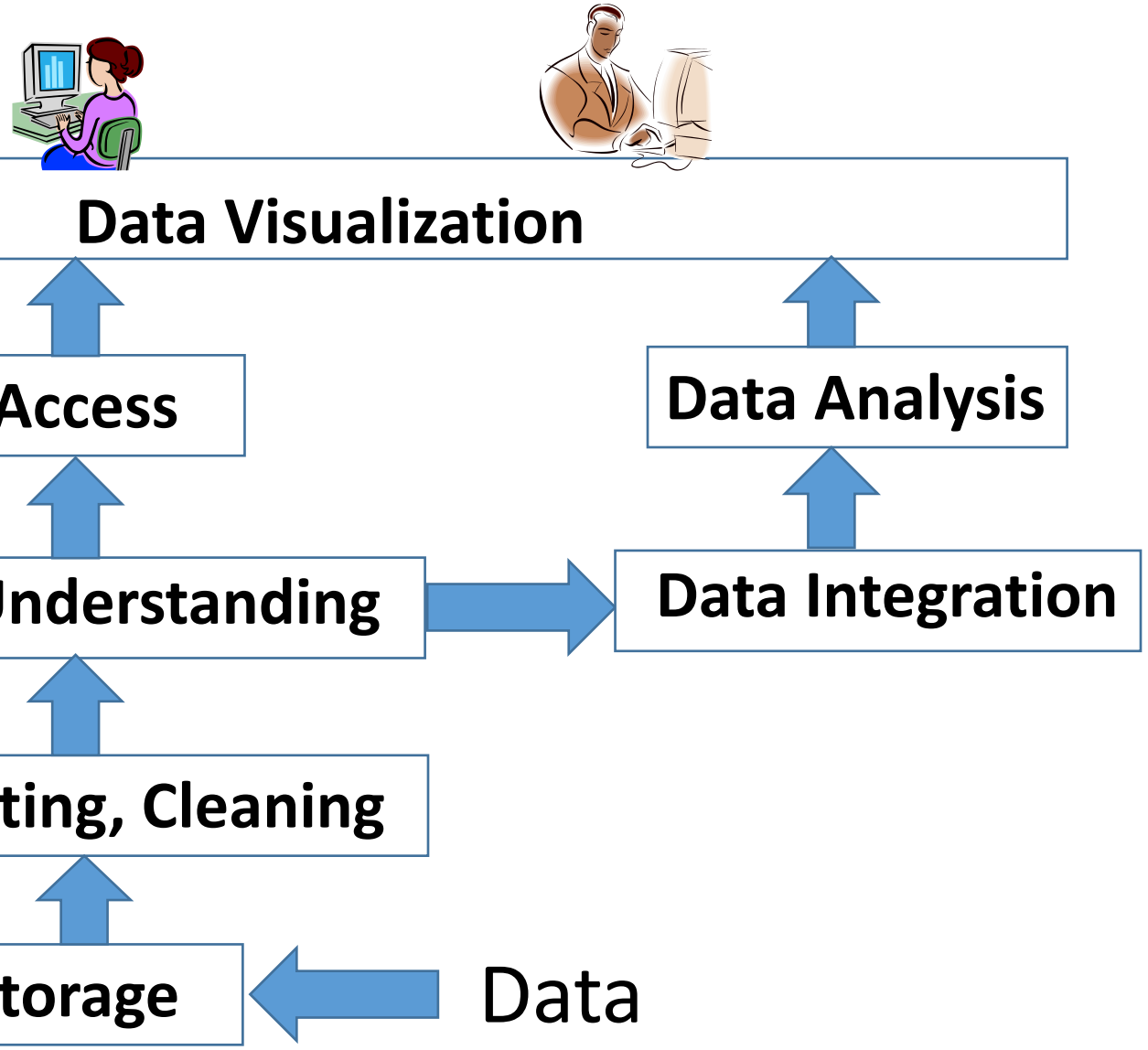
- Predictions based on correlations
- Aim:
  - Identify a good proxy
  - Watch it
  - Predict future events
- E.g., UPS trucks
  - What they did: replace certain parts after 2-3 years
  - Aim: avoid breakdown on the road
  - But: inefficient – some of them were fine
  - So: switch predictive analysis
    - Measure and monitor individual parts
    - Replace them when necessary
    - One case: new vehicles had a defective parts

Lifecycle of Data

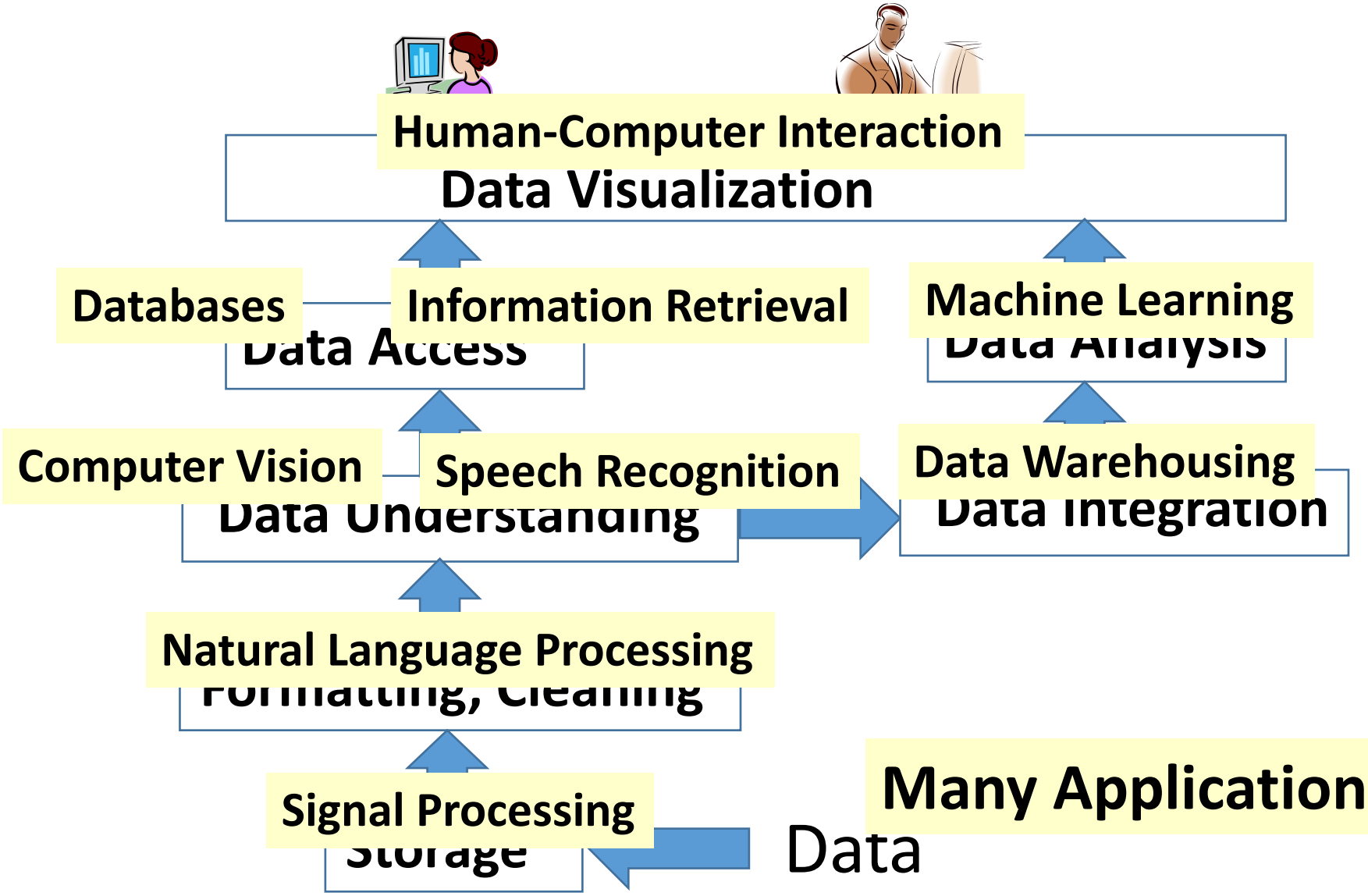
- Acquisition, Aggregation, Analysis, Application



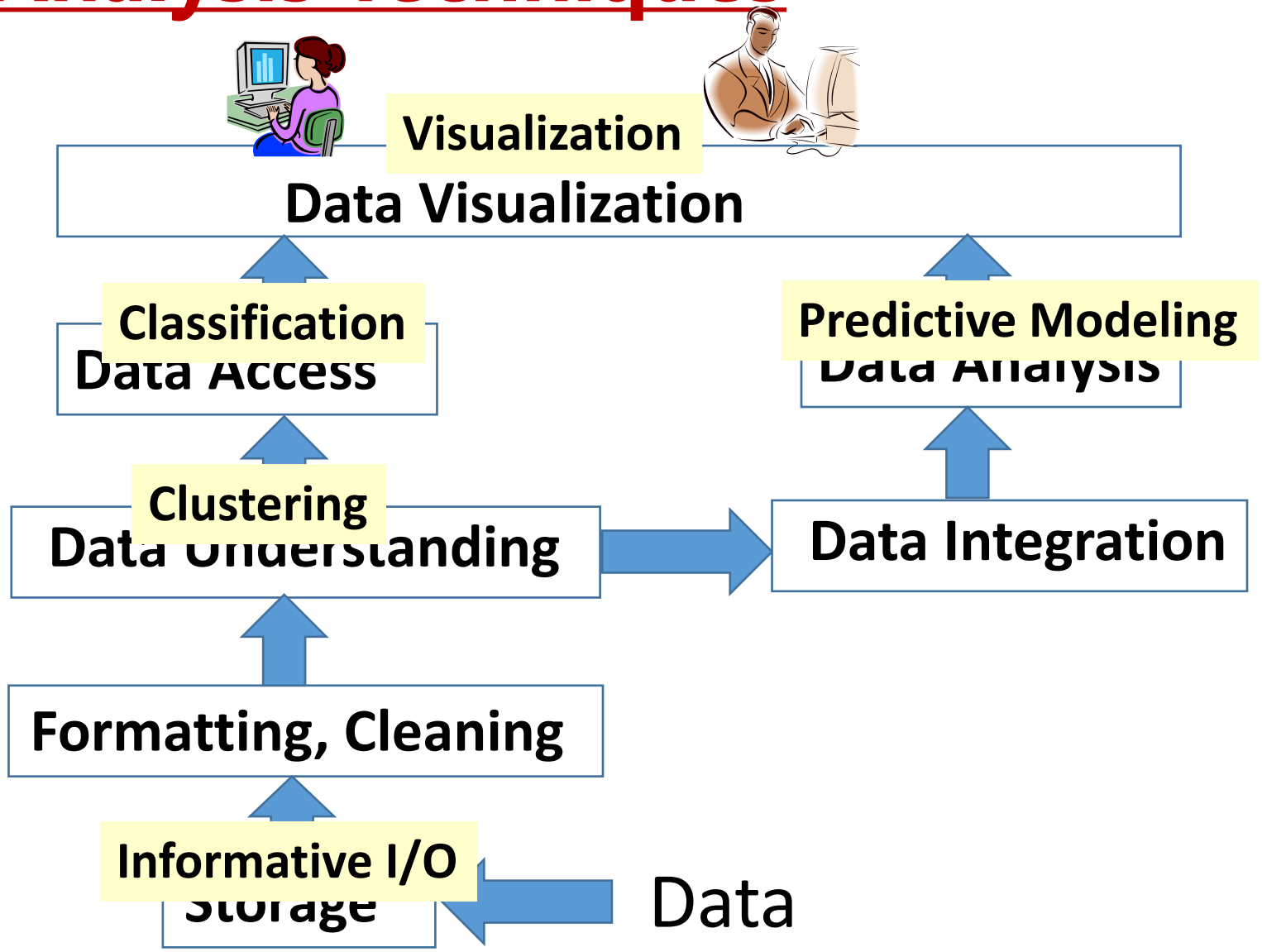
Computational View of Big Data



Big Data & Related Topics



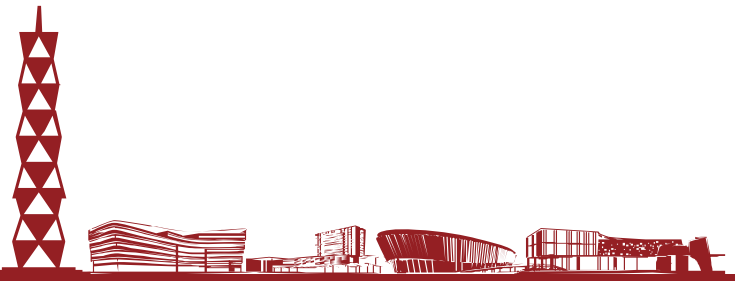
Big Data Analysis Techniques



Key Open Source Big Data Foundations

Key Questions

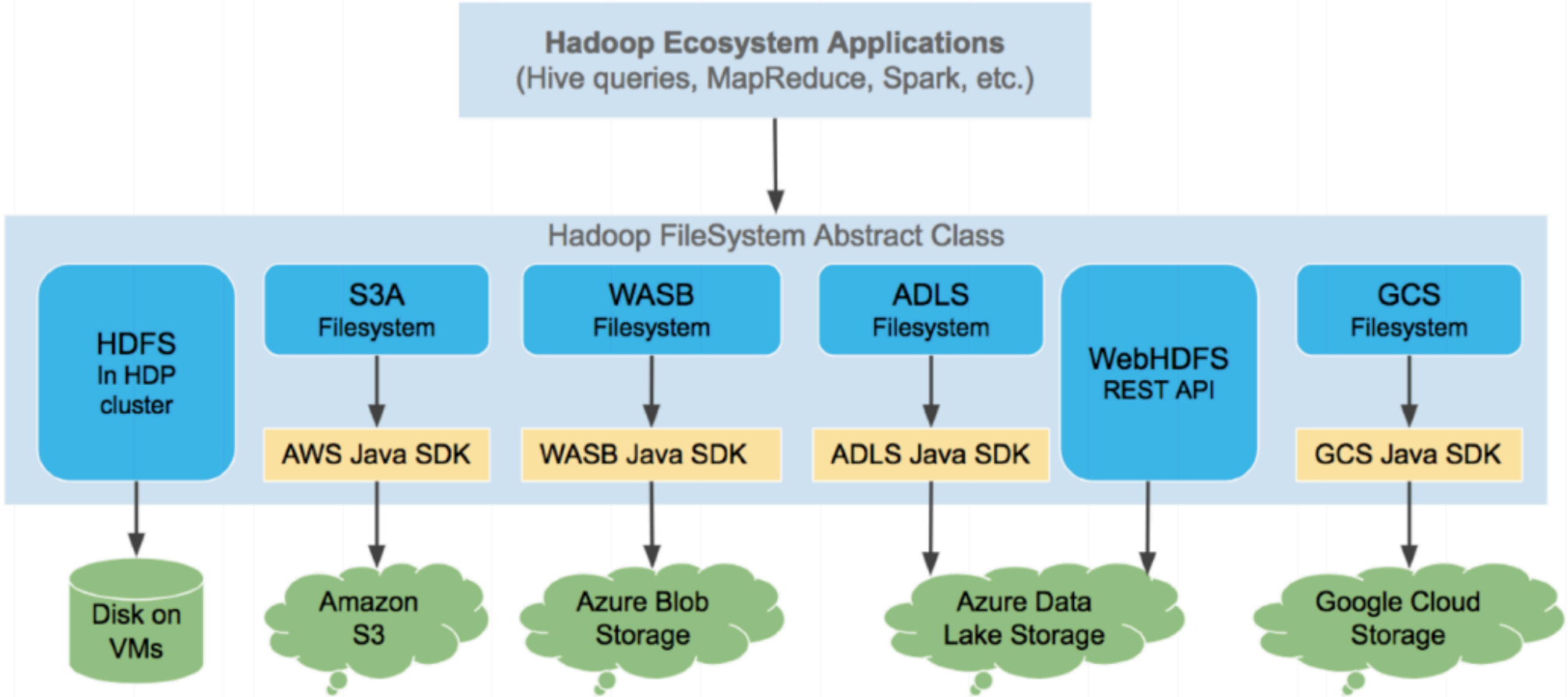
Where to store data?
How to get data in and out?
How to manage access of data?



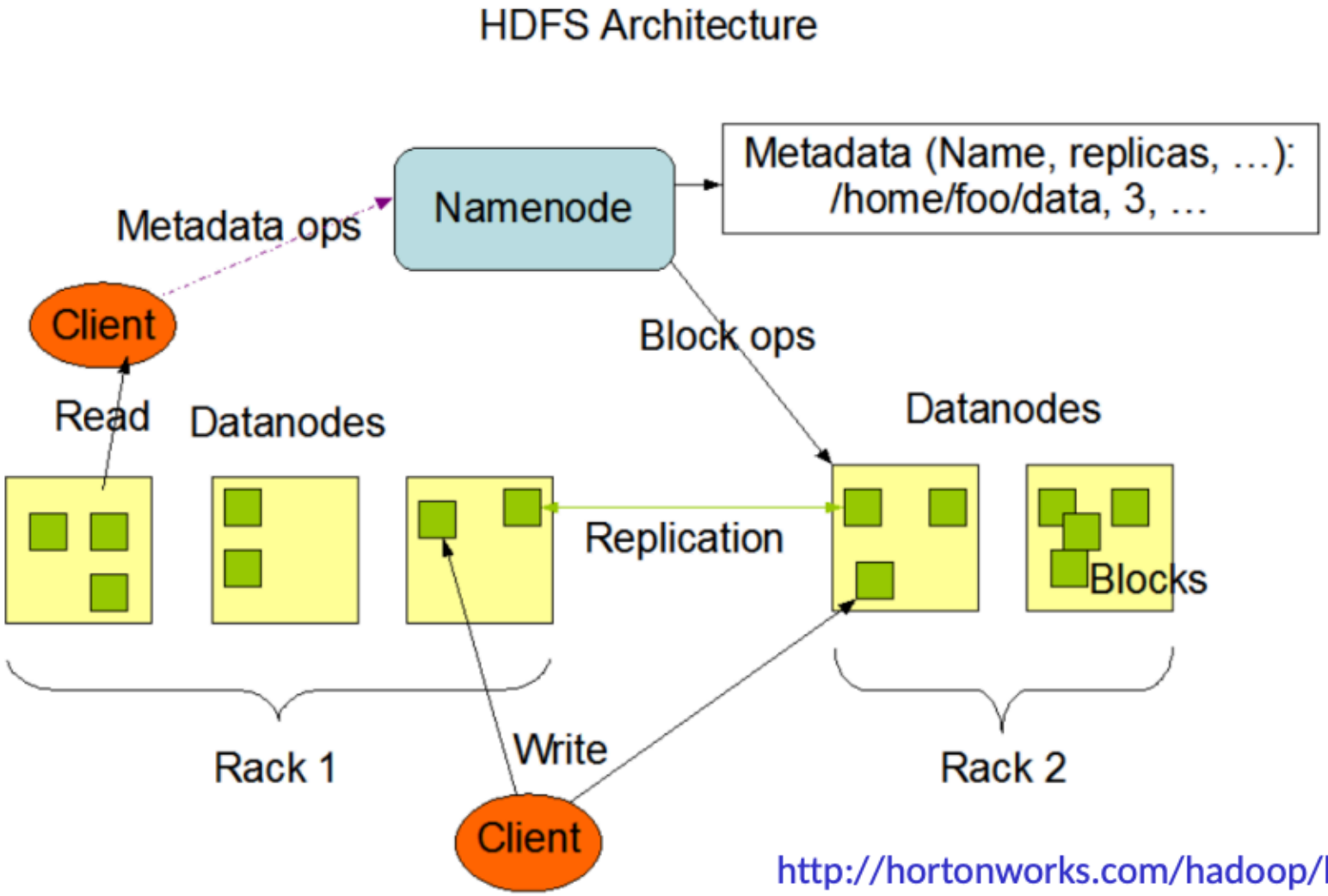
Hadoop

- Key Layers
  - Storage
  - Compute Platform
  - Shared Services
  - Compute Framworks

Hadoop Storage is widely used



Hadoop Distributed File System (HDFS)



Basic Data Storage Operations in HDFS

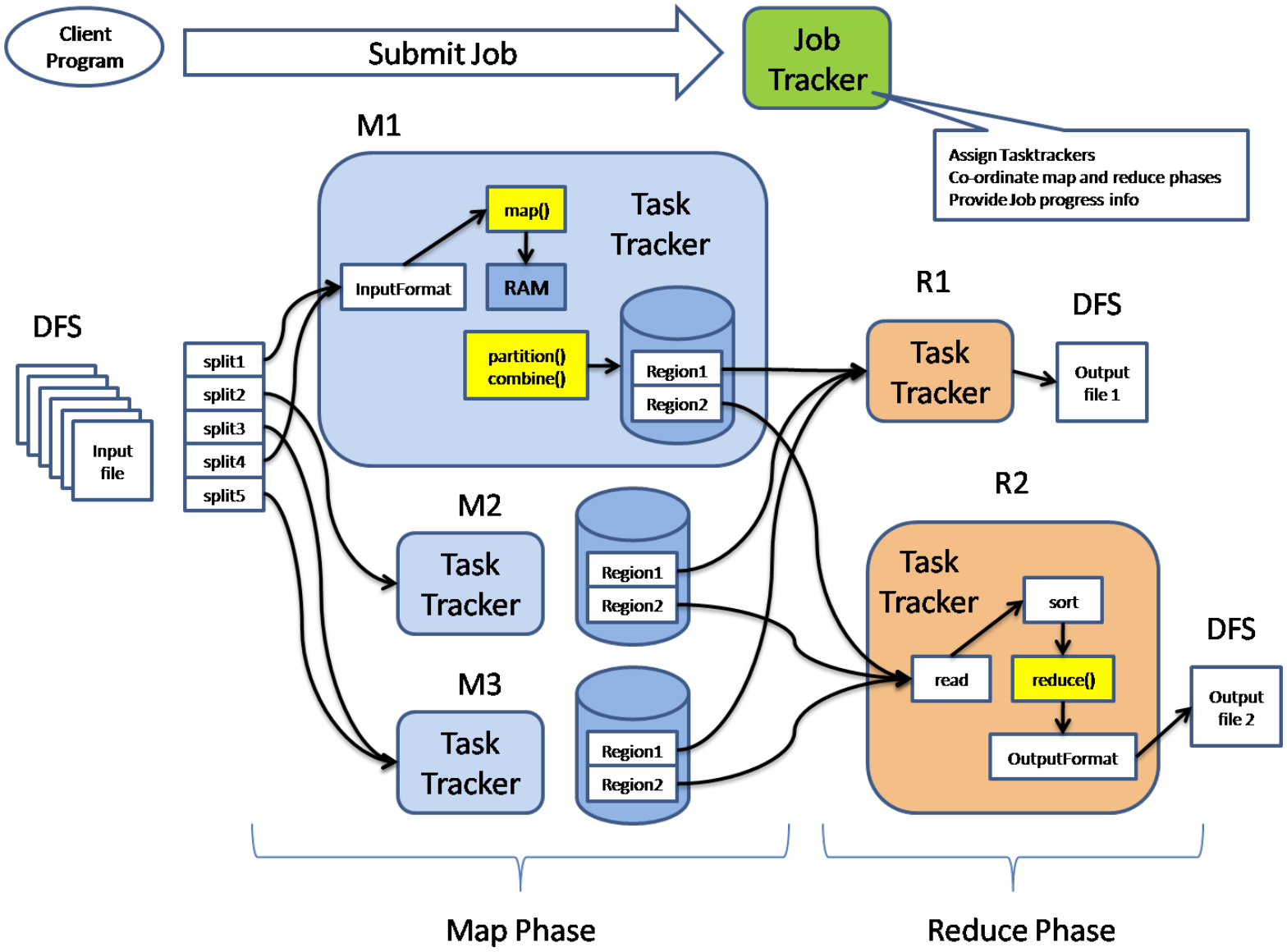
- Hadoop is designed to work best with a modest number of extremely large files
- Average file sizes: > 500MB
- Write Once, Read Often model
- Content of individual files cannot be modified, other than appending new data at the end of the file.
- General operations:
  - Create, append, delete, rename, modify file attr.
- High Availability

Hive: A data warehouse for Hadoop

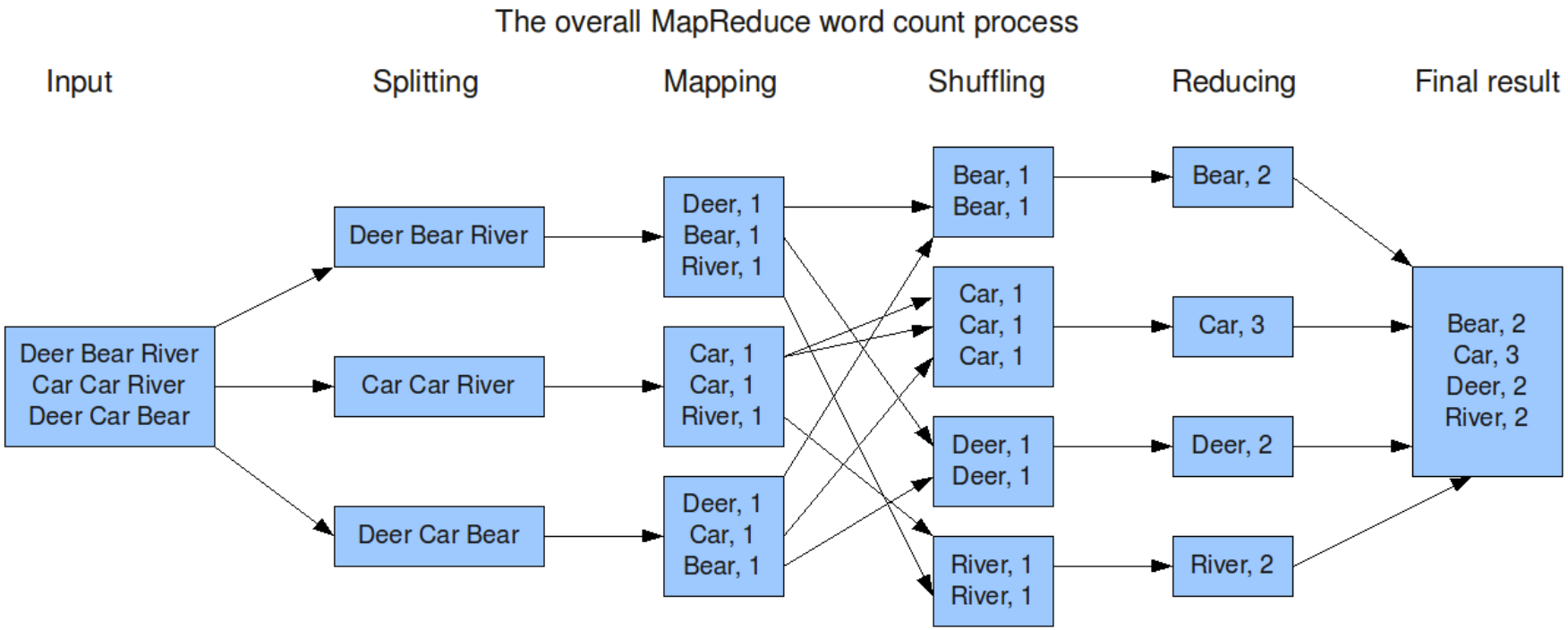
- Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data query and analysis.
- SQL-like interface to query data
- HiveQL example: word count

```
1 DROP TABLE IF EXISTS docs;
2 CREATE TABLE docs (line STRING);
3 LOAD DATA INPATH 'input_file' OVERWRITE INTO TABLE docs;
4 CREATE TABLE word_counts AS
5 SELECT word, count(1) AS count FROM
6  (SELECT explode(split(line, '\s')) AS word FROM docs) temp
7 GROUP BY word
8 ORDER BY word;
```

MapReduce Architecture

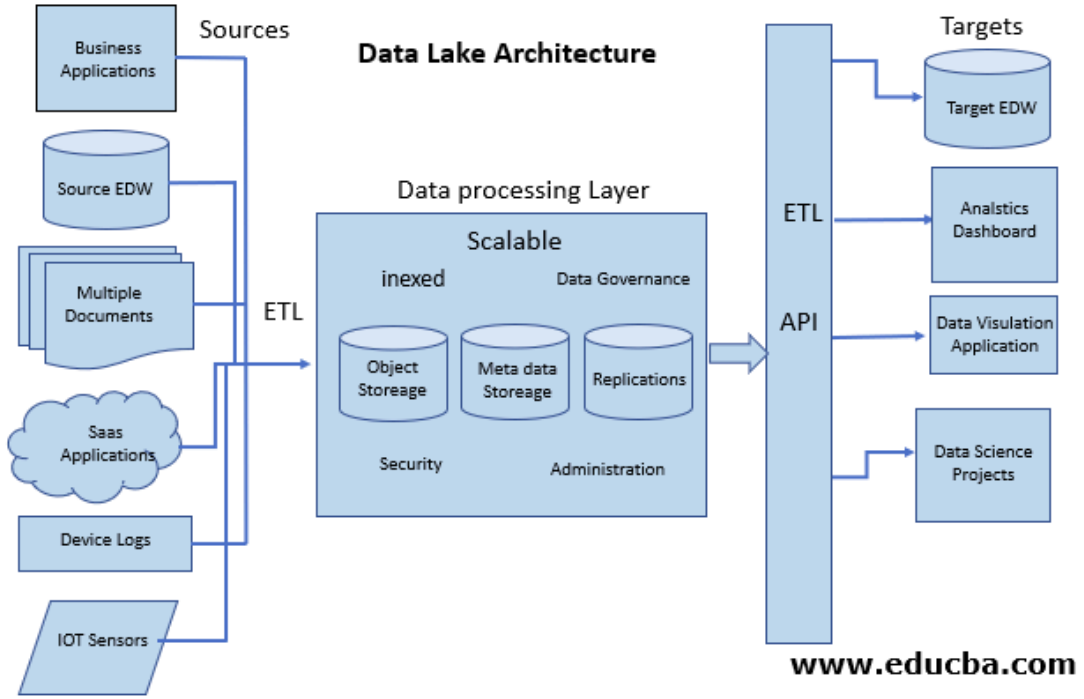


MapReduce Example



Data Lake

- A data lake provides a way to centrally apply and enforce authentication, authorization, and audit policies across multiple ephemeral workload clusters.

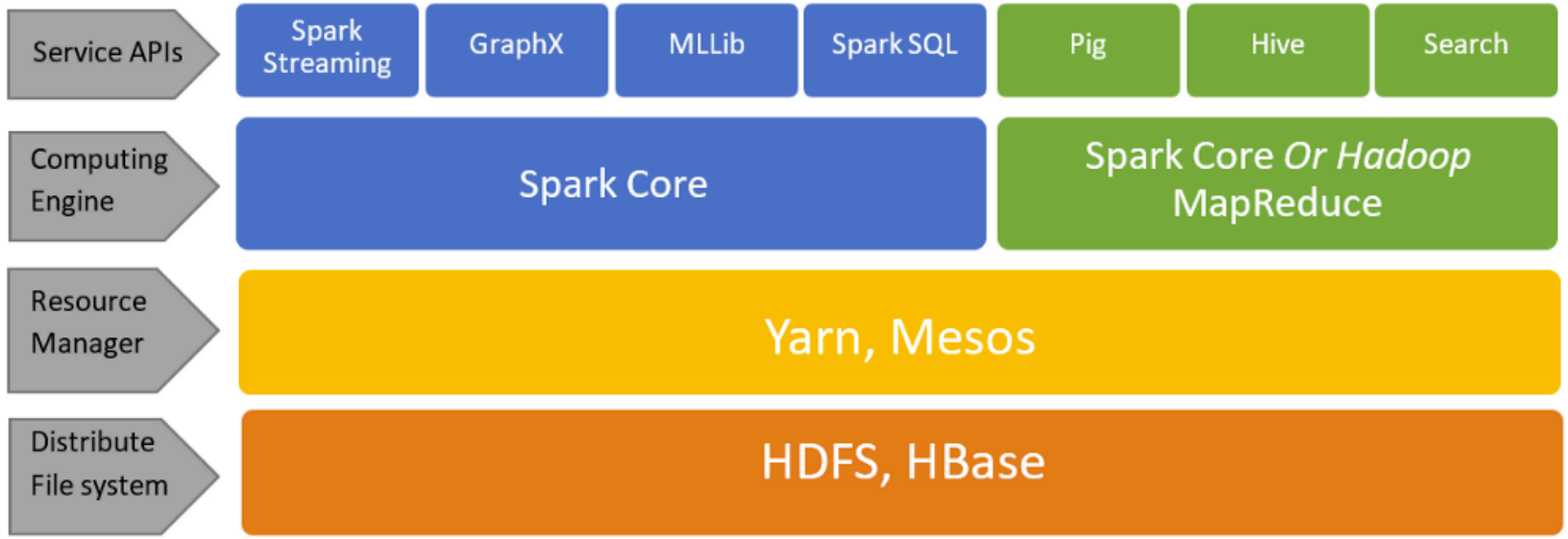


Key Questions 2

How do I process the data?
How to execute ML from the data?
How do I tell people my analytics results?

Spark

- Apache Spark is an open source distributed general-purpose cluster-computing framework. Spark provides an interface for programming entire clusters with implicit data parallelism and fault tolerance.



Spark MLlib

- ML algorithm: common learning algorithms such as classification, regression, clustering, and collaborative filtering
- Featurization: feature extraction, transformation, dimensionality reduction, and selection
- Pipelines: tools for constructing, evaluating, and tuning ML Pipelines
- Persistence: saving and load algorithms, models, and Pipelines
- Utilities: linear algebra, statistics, data handling, etc

Spark MLlib Basic Statistics

- Includes: correlation, hypothesis testing, summarizer
- Example of calculating correlation of time sequence

```
from pyspark.ml.linalg import Vectors
from pyspark.ml.stat import Correlation

data = [(Vectors.sparse(4, [(0, 1.0), (3, -2.0)]),),
        (Vectors.dense([4.0, 5.0, 0.0, 3.0]),),
        (Vectors.dense([6.0, 7.0, 0.0, 8.0]),),
        (Vectors.sparse(4, [(0, 9.0), (3, 1.0)]),)]
df = spark.createDataFrame(data, ["features"])

r1 = Correlation.corr(df, "features").head()
print("Pearson correlation matrix:\n" + str(r1[0]))

r2 = Correlation.corr(df, "features", "spearman").head()
print("Spearman correlation matrix:\n" + str(r2[0]))
```

Spark MLlib Features

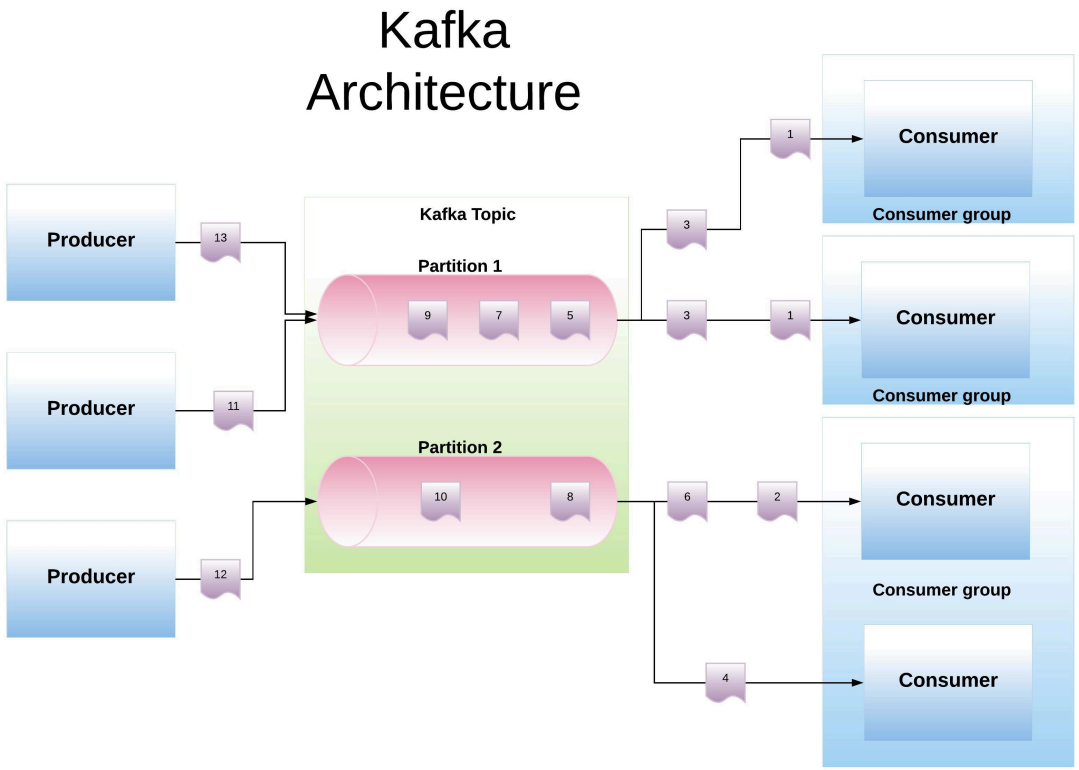
- Includes:
  - Extraction: extracting features from “raw” data
  - Transformation: scaling, converting or modifying features
  - Selection: selecting a subset from a larger set of features
  - LSH (locality sensitive hashing): combines aspects of feature transformation with other algorithms

Spark MLlib ML Algorithms

- Supervised:
  - Classification (logistic regression, decision tree classifier …)
  - Regression (linear regression, decision tree regression)
- Un-Supervised:
  - Clustering (K-means, Latent Dirichlet allocation…)
  - Collaborative Filtering (explicit vs. implicit feedback)

Kafka

- Kafka is a distributed streaming platform that is used publish and subscribe to streams of records.



More Advanced Topics

Parallel Computing with GPU
CUDA
OpenGL
Will be discussed next Tue. by Dr. Rui Fan

End of Lecture 12