Fundamentals of Big Data Analytics
(Fundamentals of Data Science)

Rudolf Mathar, TI

www.ti.rwth-aachen.de
usr: datascience passwd: 613410

Table of contents:

1. Introduction

What is data analytics / data science?
The discovery of "models" for data to extract information, draw conclusions and make decisions.
"Model" can be one of several things.

- Statistical model, an underlying distribution from which the data is drawn
  Example: Given a set of numbers, assume Gaussian, estimate the mean and variance
  Model: $N(\mu, \sigma^2)$, independent samples.

- Use the data as a training set for algorithms of machine learning, e.g., Bayes nets, support vector mach., decision tree, etc.
  Example: "Netflix challenge" devise an algorithm which predicts the rating of movies.

- Extract the most prominent features of the data and ignore the rest.
  Example: Feature extraction, similarity, PCA

- Summarization of features

  Examples: - Page rank (Google's web mining) = probability that a random walker on the web graph meets that page at any given time.

  - Clustering: points that are closed are assigned to clusters, clusters are summarized, e.g., by their centers.

Don't overstress the effect of data analytics.

Example:

Find evil-doers by looking for people who both were in the same hotel on two different days.

Assumptions: 1. 10^5 hotels

2. Everyone goes to a hotel one day in 100.
3. 10^9 people
4. People pick days and hotels at random independently
5. Examine hotel records for 1000 days

Prob. that any two people visit a hotel on any given day:

$$\frac{1}{100} \cdot \frac{1}{100} = 10^{-4}$$

Prob. that they pick the same hotel: $\frac{1}{10^4} \cdot \frac{1}{10^5} = 10^{-9}$

Prob. that 2 people visit the same hotel on 2 diff. days:

$$(10^{-9})^2 = 10^{-18}$$

Cardinality of the event space:

Pairs of people: $\binom{10^9}{2}$

Pairs of days: $\binom{10^3}{2}$

Exp. no. of evil-doing events $\left(\text{use } \binom{n}{2} \approx \frac{n^2}{2}\right)$

$$\binom{10^9}{2} \cdot \binom{10^3}{2} \cdot 10^{-18} \approx 5 \cdot 10^{18} \cdot 5 \cdot 10^5 \cdot 10^{-18}$$

$$= 25 \cdot 10^4 = 250 \cdot 000$$

Implementations to deal with huge data sets.
Note: For huge data sets hardware errors will occur almost surely.

MapReduce and Hadoop

Key idea: use parallelism from computing clusters (not a super-comp) built of commodity hardware, connected by Ethernet or inexpensive switches.

Software stack:

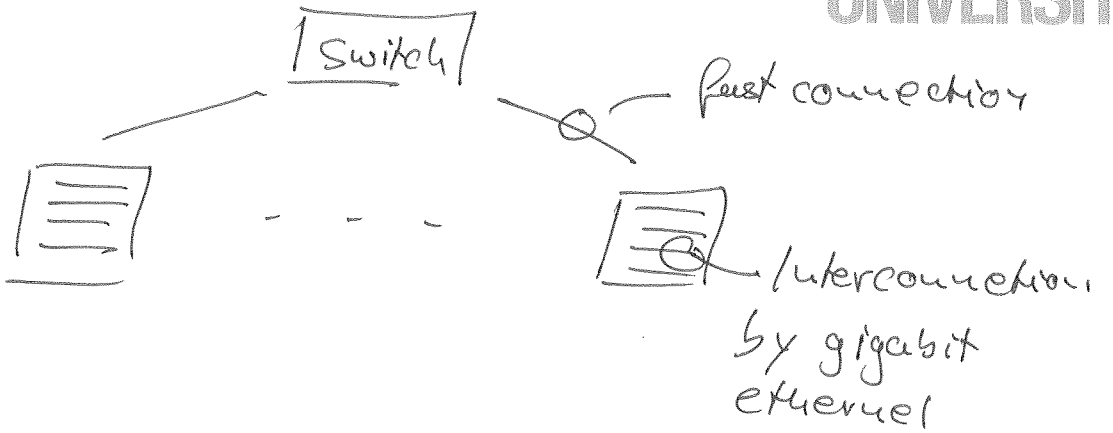
(i) Distributed file system (DFS)
- large blocks
- redundancy by replication

(ii) Programming system: MapReduce
- tolerant to hardware failure
- able to handle large data efficiently

Architecture:

(i) Compute nodes stored in a rack, each with its own processor and storage device

(ii) Racks are connected by switches



Principles:

(i') Files are stored redundantly to protect against failures

(ii') Comp. are downloaded into independent tasks. If one fails it can be restarted without affecting others.

add(): Distr. File system (DFS)

- Files are devided into chunks (typ. 64 MB)
- chunks are replicated (typ. 3 times on diff. racks)
- there is a File-master node or name node with information where to find copies of files.

Implementations

- GFS (Google File system)
- HDFS (Hadoop distr. File system, Apache)
- Cloud Store (open source DFS)

ad (ii): Map Reduce (computing paradigm)

- System manages parallel execution, coordination of tasks
- 2 Functions are written by the user
  Map and Reduce

Implementations:

- MapReduce (Google, internal)
- Hadoop (Open Source, Apache)

Detailled information LRU p. 21-70