# Implementations to deal with huge data sets.

Note: For huge data sets hardware errors will occur almost certainly.

## MapReduce and Hadoop

(Not the main focus of this lecture, only briefly summarized)

Key idea:
Use parallelism from 'computing clusters' (not a super-computer), built of commodity hardware, connected by Ethernet and inexpensive switches.

### Software stack:

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

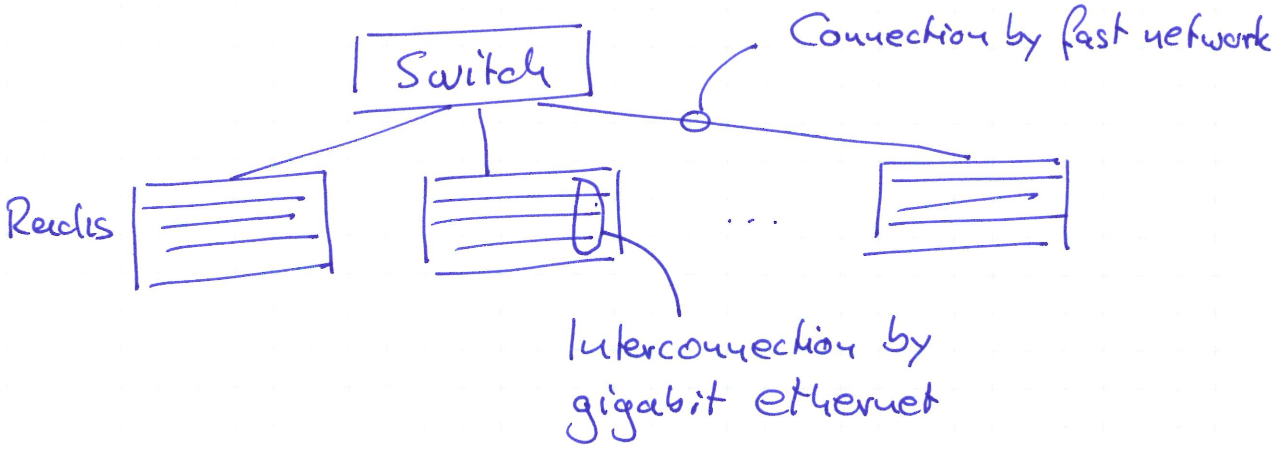
(ii) Programming system: MapReduce
- tolerant to hardware failure
- able to handle large data sets efficiently

RWTH AACHEN UNIVERSITY

## Architecture:

(i) Compute nodes stored on racks, each with its own processor and storage device.

(ii) Racks are connected by switches




## Principles:

(i) Files are stored redundantly to protect against failure of nodes.

(ii) Computations are devided into independent tasks. If one fails it can be restarted without affecting others.

Thinking the Future
Zukunft denken

**Remarks (i): Distributed File system (DFS)**

- Files are devided into chunks (typically 64 MB)
- Chunks are replicated (typically 3 times on different racks)
- A file master node or name node has information where to find copies of files

Implementations:

- GFS (Google file system)
- HDFS (Hadoop distributed file system, Apache)
- Cloud Store (open source DFS)

**Remarks (ii) Map Reduce (computing paradigm)**

- System manages parallel execution and coordination of tasks.
- 2 functions are written by the user: Map and Reduce

Implementations:

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

RWTH AACHEN UNIVERSITY

http://hadoop.apache.org/

## What Is Apache Hadoop?

The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

The project includes these modules:

- **Hadoop Common**: The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.

Co-founders: Doug Cutting and Mike Cafarella, January 2006

(Doug Cutting named the system after his son’s toy elephant.)

RWTH AACHEN UNIVERSITY

# 2. Prerequisites from Matrix Algebra

Real $(m \times n)$ matrices will be written as

$$M = (m_{ij})_{\substack{1 \le i \le m \\ 1 \le j \le n}} \in \mathbb{R}^{m \times n} \quad (\text{or } \mathbb{C}^{m \times n})$$

Diagonal matrices as $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n) \in \mathbb{R}^{n \times n}$

Matrix $U \in \mathbb{R}^{n \times n}$ is called orthogonal if

$$U U^T = U^T U = I_n \quad (\text{identity matrix})$$

$O(n)$ set of orthogonal $(n \times n)$-matrices.

Th. 2.1. (Singular value decomposition, SVD)
Given $M \in \mathbb{R}^{m \times n}$. There exist $U \in O(m)$ and $V \in O(n)$ and some $\Sigma \in \mathbb{R}^{m \times n}$ with non-negative entries in its diagonal and zeros otherwise such that

$$M = U \Sigma V^T.$$

The diagonal elements of $\Sigma$ are called singular values. The columns of $U$ and $V$ are called left and right singular vectors. ⊥

Thinking the Future
Zukunft denken

<u>Remark</u> If $m<n$

$$\Updownarrow \quad \boxed{\;M\;} = \boxed{U}\ \boxed{\begin{matrix}\circ & & \\ & \ddots & \\ & & \circ\end{matrix} \;\Big|\; 0}\ \boxed{\;V^T\;}$$

then the SVD may be written as

$\exists U \in \mathbb{R}^{m\times n}$, $UU^T = I_m$, $\exists V \in O(n)$, $\exists \Sigma \in \mathbb{R}^{n\times n}$ diagonal with $\geq 0$ entries

such that $\quad M = U\Sigma V$. $\lrcorner$

<u>Th 2.2.</u> (Spectral decomposition)

Given $M \in \mathbb{R}^{n\times n}$ symmetric. There exist $V \in O(n)$, $V = (v_1, \dots, v_n)$ and $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$ such that

$$M = V\Lambda V^T = \sum_{i=1}^{n} \lambda_i v_i v_i^T .$$

$v_i$ are the eigenvectors of $M$ with eigenvalues $\lambda_i$. $\lrcorner$

- If $\lambda_i > 0$, $i = 1, \dots, n$, $M$ is called positive definite ($M > 0$) (p.d.)

  If $\lambda_i \geq 0$, $i = 1, \dots, n$, $M$ is called non-negative definite ($M \geq 0$) (n.n.d.)

- If $M \geq 0$, then it has a Cholesky decomposition

$$M = V\Lambda^{\frac{1}{2}} (V\Lambda^{\frac{1}{2}})^T$$

where $\Lambda^{\frac{1}{2}} = \mathrm{diag}(\lambda_1^{\frac{1}{2}}, \dots, \lambda_n^{\frac{1}{2}})$

- $M \succeq 0 \iff x^T M x \geq 0 \quad \forall x \in \mathbb{R}^n$

  [ "$\Rightarrow$" $M \succeq 0 \Rightarrow M = V \Lambda V^T, \ \Lambda \succeq 0$
  $\Rightarrow x^T M x = x^T V \Lambda V^T x \geq 0 \quad \forall x \in \mathbb{R}^n$. ]

- $M \succ 0 \iff x^T M x > 0 \quad \forall x \in \mathbb{R}^n, \ x \neq 0$.

<u>Def. 2.3.</u> a) $M = (m_{ij}) \in \mathbb{R}^{n \times n}$. $tr(M) = \sum_{i=1}^{n} m_{ii}$ is called <u>trace</u> of $M$.

b) Given $M \in \mathbb{R}^{n \times n}$. $\|M\|_F = \left( \sum_{i,j} m_{ij}^2 \right)^{1/2} = \left( tr(M^T M) \right)^{1/2}$

is called the <u>Frobenius norm</u> of $M$.

c) $M \in \mathbb{R}^{n \times n}$, $M$ symmetric. $\|M\|_S = \max_{1 \leq i \leq n} |\lambda_i|$

is called <u>spectral norm</u>. $\lrcorner$

- It holds that $tr(A \cdot B) = tr(B \cdot A)$, $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times m}$.
- $tr(M) = \sum_{i=1}^{n} \lambda_i(M)$, $\det(M) = \prod_{i=1}^{n} \lambda_i(M)$

  for any symm. matrix $M$.

Th. 2.4. (Ky Fan, 1950)

Given $M \in \mathbb{R}^{n\times n}$ symm., $k \leq n$, $\lambda_1(M) \geq \dots \geq \lambda_n(M)$ eigenvalues.

$$\max_{\substack{V\in\mathbb{R}^{n\times k}\\ V^TV = I_k}} \operatorname{tr}(V^TMV) = \sum_{i=1}^{k} \lambda_i(M)$$

$$\min_{\substack{V\in\mathbb{R}^{n\times k}\\ V^TV = I_k}} \operatorname{tr}(V^TMV) = \sum_{i=1}^{k} \lambda_{n-i+1}(M)$$

$\lrcorner$

Special case: $k=1$

$$\max_{\|v\|=1} v^TMv = \lambda_{max}(M)$$

$$\min_{\|v\|=1} v^TMv = \lambda_{min}(M)$$

Also note:

$$\max_{\|v\|=1} v^TMv = \max_{v\neq 0} \frac{v^TMv}{v^Tv}$$

Th. 2.5. Given $A, B \in \mathbb{R}^{n\times n}$, symmetric with eigenvalues $\lambda_1 \geq \dots \geq \lambda_n$ and $\mu_1 \geq \dots \geq \mu_n$, respectively. Then

$$\sum_{i=1}^{n} \lambda_i \mu_{n-i+1} \leq \operatorname{tr}(A\cdot B) \leq \sum_{i=1}^{n} \lambda_i \mu_i$$

$\lrcorner$

RWTH AACHEN UNIVERSITY

Proof. First show for any $x_1, \dots, x_n, y_1, \dots, y_n \in \mathbb{R}$ with ordered values $x_{[1]} \geq \dots \geq x_{[n]}$ and $y_{[1]} \geq \dots \geq y_{[n]}$.

$$\sum_{i=1}^{n} x_{[i]} y_{[n-i+1]} \leq \sum_{i=1}^{n} x_i y_i \leq \sum_{i=1}^{n} x_{[i]} y_{[i]} \qquad (*)$$

Left as an exercise. Now:

It holds

$$\begin{aligned}
\operatorname{tr}(A \cdot B) &= \operatorname{tr}\left(V \Lambda V^T U M U^T\right) \\
&= \operatorname{tr}\Big(\underbrace{U^T V}_{Q} \Lambda \underbrace{V^T U}_{Q^T} M\Big), \quad \underset{\text{orthogonal}}{Q = (q_1, \dots, q_n) = U^T V} \\
&= \operatorname{tr}\left((\lambda_1 q_1, \dots, \lambda_n q_n) \begin{pmatrix} \mu_1 q_1^T \\ \vdots \\ \mu_n q_n^T \end{pmatrix}\right) \\
&= \operatorname{tr}\left(\sum_{i=1}^{n} \lambda_i \mu_i q_i q_i^T\right) \\
&= \sum_{i=1}^{n} \lambda_i \mu_i \operatorname{tr}(q_i q_i^T) \\
&= \sum_{i=1}^{n} \lambda_i \mu_i \operatorname{tr}\big(\underbrace{q_i^T q_i}_{=1}\big) = \sum_{i=1}^{n} \lambda_i \mu_i .
\end{aligned}$$

Assertion follows by $(*)$. $\square$

Thinking the Future
Zukunft denken

Let $\lambda^+ = \max\{\lambda, 0\}$ denote the positive part of $\lambda \in \mathbb{R}$.

Th. 2.6. Given $M \in \mathbb{R}^{n \times n}$ symmetric with spectral decomposition $M = V \operatorname{diag}(\lambda_1, \dots, \lambda_n) V^T$, $\lambda_1 \geq \dots \geq \lambda_n$. Then for $k \leq n$

$$\min_{A \succeq 0,\ \mathrm{rk}(A) \leq k} \|M - A\|_F^2$$

is attained at $A^* = V \operatorname{diag}(\lambda_1^+, \dots, \lambda_k^+, 0, \dots, 0) V^T$ with optimum value $\sum_{i=1}^{k} (\lambda_i - \lambda_i^+)^2 + \sum_{i=k+1}^{n} \lambda_i^2$. $\lrcorner$

Proof.

$$\|M - A\|^2 = \|M\|^2 - 2\,\mathrm{tr}(MA) + \|A\|^2$$

$$\geq \sum_{i=1}^{n} \lambda_i^2 - 2 \sum_{i=1}^{n} \lambda_i \mu_i + \sum_{i=1}^{n} \mu_i^2$$

$\mu_1 \geq \dots \geq \mu_n \geq 0$ eigenv. of $A$

$$= \sum_{i=1}^{n} (\lambda_i - \mu_i)^2$$

$$= \sum_{i=1}^{k} (\lambda_i - \mu_i)^2 + \sum_{i=k+1}^{n} (\lambda_i - 0)^2, \quad \text{if } \mathrm{rk}(A) \leq k$$

$$\geq \sum_{i=1}^{k} (\lambda_i - \lambda_i^+)^2 + \sum_{i=k+1}^{n} \lambda_i^2$$

Lower bound is attained if $A = V \operatorname{diag}(\lambda_1^+, \dots, \lambda_k^+, 0, \dots, 0) V^T$! $\blacksquare$