Implementations to deal with huge data sets.

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

Map Reduce 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:

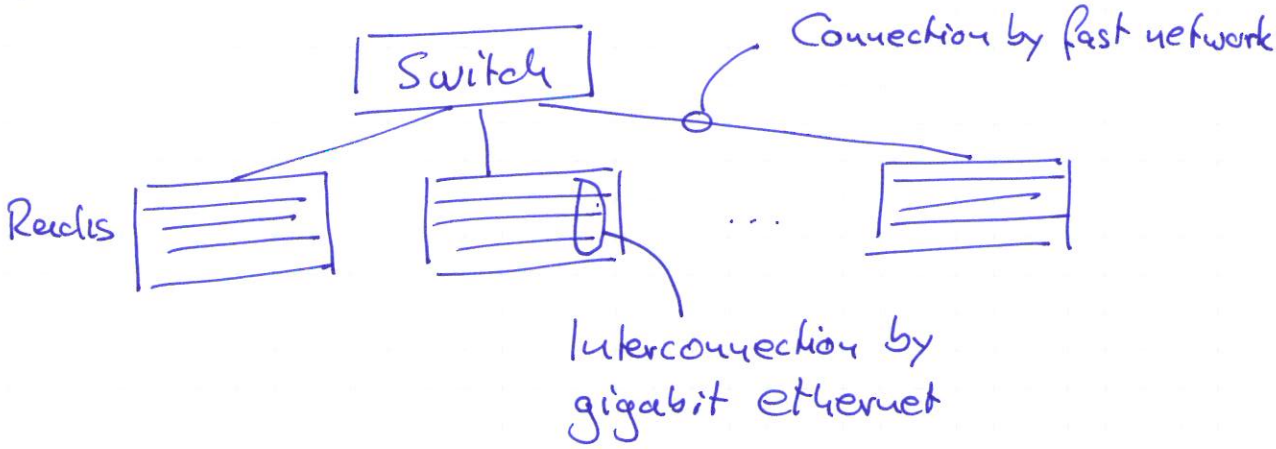
(i) Distributed file system (DFS)
- large blocks
- redundancy by replication

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

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.

Remarks: 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: MapReduce (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)

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.)

2. Prerequisites from Matrix Algebra

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

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

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

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

$$UU^T = U^T U = I_n \quad (I_n = \operatorname{diag}(1, \dots, 1) \in \mathbb{R}^{n \times n})$$

$O(n)$ denotes the set of orthogonal matrices.

Th. 2.1. (Singular value decomposition, SVD)
Given $M \in \mathbb{R}^{m \times n}$. There exists $U \in O(m)$ and $V \in O(n)$ and some $\Sigma \in \mathbb{R}^{m \times n}$ with non-negative entries in its diagonal and 0's otherwise such that

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

The diagonal elements of Σ are called singular values. The columns of U and V are called left and right singular vectors of M. ∎

Remark: If $m < n$, say, 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 such that $M = U \Sigma V$.

Th 2.2. (Spectral decomposition)

Given $M \in \mathbb{R}^{n \times n}$ symmetric. There exists $V \in \mathcal{O}(n)$, $V = (v_1, \dots, v_n)$ and $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ 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 λ_i.

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

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

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

$$M = \left(V \Lambda^{1/2}\right)\left(V \Lambda^{1/2}\right)^T$$

where $\Lambda^{1/2} = \operatorname{diag}(\lambda_1^{1/2}, \dots, \lambda_n^{1/2})$.

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

Def. 2.3.

a) Given $M = (m_{ij}) \in \mathbb{R}^{n \times n}$.

$tr(M) = \sum_{i=1}^{n} m_{ii}$ is called trace of M.

b) Given $M = (m_{ij}) \in \mathbb{R}^{m \times n}$.

$$\|M\|_F = \sqrt{\sum_{i,j} m_{ij}^2} = \sqrt{tr(M^T M)}$$

is called the Frobenius norm of M.

c) $M \in \mathbb{R}^{n \times n}$, M symmetric.

$$\|M\|_S = \max_{1 \le i \le n} |\lambda_i|$$

is called the spectral norm.

[A "norm" only for symm. matrices, otherwise not] ⌟

- It holds that $tr(AB) = tr(BA)$, $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times m}$
- $tr(M) = \sum_{i=1}^{n} \lambda_i(M)$ (*), $\det(M) = \prod_{i=1}^{n} \lambda_i(M)$, if M is symmetric

(*) $tr(M) = tr(V \Lambda V^T) = tr(\Lambda V^T V)$

$= tr(\Lambda) = \sum_{i=1}^{n} \lambda_i(M)$

Th 2.4. (Ky Fan, 1950)

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

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

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

⊣

- Special case of Th. 2.4., $k = 1$

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

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

Also note $\max_{\|v\| = 1} = \max_{v \neq 0} \frac{v^T M v}{v^T v}$

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

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

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

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

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

Proof.

$$\begin{aligned}
\|M - A\|_F^2 &= \|M\|^2 - 2\operatorname{tr}(MA) + \|A\|^2 \\
&\overset{\text{(Th. 2.5)}}{\geq} \sum_{i=1}^{n} \lambda_i^2 - 2 \sum_{i=1}^{n} \lambda_i \mu_i + \sum_{i=1}^{n} \mu_i^2 \quad , \ \mu_1 \geq \dots \geq \mu_n \geq 0 \\
&= \sum_{i=1}^{n} (\lambda_i - \mu_i)^2 \\
&= \sum_{i=1}^{k} (\lambda_i - \mu_i)^2 + \sum_{i=k+1}^{n} (\lambda_i - 0)^2 \quad , \text{ since } \operatorname{rk}(A) \leq k \\
&\geq \sum_{i=1}^{k} (\lambda_i - \lambda_i^+)^2 + \sum_{i=k+1}^{n} \lambda_i^2 .
\end{aligned}$$

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