# Big Data
# NoSQL

Dr. Wenceslao PALMA
wenceslao.palma@pucv.cl

PONTIFICIA UNIVERSIDAD CATOLICA DE VALPARAISO

Escuela de Ingeniería Informática

# What is a database?

- Collection of records.
- Stored as a table on disk (persistency).
- Records are accessed via queries.

# What is a database?

- Collection of records.
- Stored as a table on disk (persistency).
- Records are accessed via queries.

Databases store the most valuable data of organizations

# Relational databases

- The most popular database approach for data storage, retrieval and management.
- ACID properties (Atomicity, Consistency, Isolation, Durability) place Relational databases as the solution for almost all data management systems.

# Relational databases

- The most popular database approach for data storage, retrieval and management.
- ACID properties (Atomicity, Consistency, Isolation, Durability) place Relational databases as the solution for almost all data management systems.

How to tackle scalability in web-scale systems?
How to query data colletions where the query and the storage model of a relational database does not fit?

# NoSQL

- The term NoSQL was first coined in 1988.
- It has been restated as "Not Only SQL".
- NoSQL is a very broad term.
- NoSQL databases are usually divided in four categories:
    - Key-value stores: DynamoDB.
    - Document stores: MongoDB, CouchBase, CouchDB.
    - Column stores: HBase, Cassandra.
    - Graph databases: Neo4J, Infinite Graph.

# NoSQL

- The term NoSQL was first coined in 1988.
- It has been restated as "Not Only SQL".
- NoSQL is a very broad term.
- NoSQL databases are usually divided in four categories:
  - Key-value stores: DynamoDB.
  - Document stores: MongoDB, CouchBase, CouchDB.
  - Column stores: HBase, Cassandra.
  - Graph databases: Neo4J, Infinite Graph.

The "one size fits all" approach of relational databases no longer applies.

# NoSQL

- Performance and flexibility are the main reason to move to NoSQL databases.
  - Performance: availability, horizontal scalability (Big Data).
  - Flexibility: semi-structured or unstructured data.

# The CAP theorem

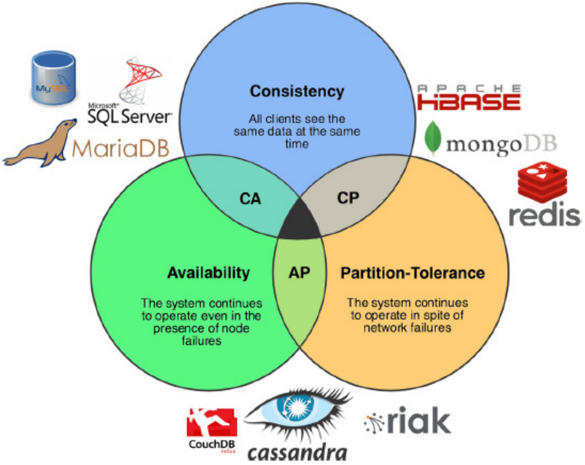
Proposed by Brewer, the CAP theorem states that no distributed system can simultaneously guarantee Consistency, Availability and Partition-Tolerance.

# The CAP theorem

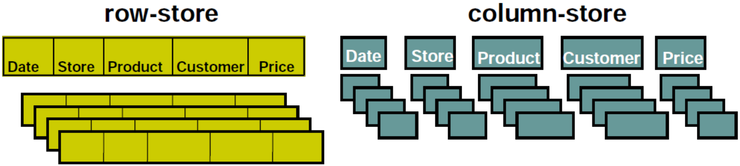
Proposed by Brewer, the CAP theorem states that no distributed system can simultaneously guarantee Consistency, Availability and Partition-Tolerance.

NoSQL databases lose the support for ACID transactions as a trade-off for increased availability and scalability.

# NoSQL databases w.r.t. the CAP theorem


MySQL
Microsoft SQL Server
MariaDB
Consistency
All clients see the same data at the same time
APACHE HBASE
mongoDB
redis
CA
CP
Availability
The system continues to operate even in the presence of node failures
AP
Partition-Tolerance
The system continues to operate in spite of network failures
CouchDB relax
cassandra
riak

# Column store



- Each column is stored in a separate file.
- Only retrieves relevant data. :)
- Add/modify a record require multiple accesses. :(

# Column store

- Suppose a database of one table which has 5 columns (each columns is same size). 1 Billion of records, 100 Bytes each → 100GB.
- A query that retrieves 3 columns from all the records.
- A retrieval time of 100MB/sec.
- What is the retrieval time in a:
    - row-oriented database: 1000sec
    - column-oriented database: 100GB×3/5 → 600sec

# Column store

- Suppose a database of one table which has 5 columns (each columns is same size). 1 Billion of records, 100 Bytes each → 100GB.
- A query that retrieves 3 columns from all the records.
- A retrieval time of 100MB/sec.
- What is the retrieval time in a:
  - row-oriented database: 1000sec
  - column-oriented database: 100GB$\times$3/5 → 600sec

A column store database is more suitable for read-intensive of large data repositories.

# References

- *Column-Oriented Database Systems.* VLDB 2009 Tutorial.
- *Choosing the rigth NoSQL database for the job: a quality attribute evaluation.* Journal of Big Data, (2015) 2:18