EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# Advanced Database Systems

## Principles

Strategic project of TBU in Zlín, reg. no. CZ.02.2.69/0.0/0.0/16_015/0002204

Zdenka Prokopova
TBU in Zlín

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# Content

- Basic principles of NoSQL databases
  - Scalability
- Data distribution
  - Sharding
  - Replication: master-slave, peer-to-peer
  - Combination: Sharding + Replication
- Consistency
  - Write-Write vs. read-write
  - Strategies and techniques

Zdenka Prokopova
TBU in Zlín

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# Objective and assumptions

- **The main goal**: to process increasing amounts of data and queries without losing performance
- **Basic solution**
  - Scalability
  - Data distribution
- **Compromise solution**
  - It is necessary to give up some ACID properties of RDBS (strict high consistency requirement)
  - CAP theorem

Zdenka Prokopova
TBU in Zlín

Fowler, A. (2015). *NoSQL for dummies*. Hoboken, NJ: John Wiley & Sons, Inc.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# Basic principles

- **Scalability** – vertical or horizontal
  - The property of the system to respond flexibly to changing requirements
  - Achievable by data distribution
- **Distribution models**
  - A method of processing data using a distributed approach
  - Specific ways to do sharding, replication, or a combination of both
- **Consistency**
  - CAP theorem for distributed systems
  - Strong consistency → eventual consistency

Zdenka Prokopova
TBU in Zlín

Brewer, E. A. (2000). *Towards robust distributed systems.* Paper presented at the PODC.
Holubová, I., Kosek, J., Minařík, K., & Novák, D. (2015). *Big Data a NoSQL databáze*: Grada.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# Scalability - vertical (scaling up)

- **Increasing the computing power** of the respective **machine** (server)
  - Large disk storage using disk arrays
  - Massively parallel architectures
- A traditional choice
  - Easy to implement (no need to distribute data)
  - In favor of strong consistency

Zdenka Prokopova
TBU in Zlín

Brewer, E. A. (2000). *Towards robust distributed systems.* Paper presented at the PODC.
Holubová, I., Kosek, J., Minařík, K., & Novák, D. (2015). *Big Data a NoSQL databáze*: Grada.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# Disadvantages of vertical scalability

- Higher costs
  - Powerful machines cost more than regular commercial HW
- Data growth limit
  - Each (even the most powerful) machine has its upper limit
- Proactive attitude
  - At the beginning of the implementation of the application, it is necessary to plan the maximum data size in advance
- Vendor lock-in
  - Powerful machines are produced by only a few manufacturers
  - The customer is dependent on a single supplier (proprietary HW)

Zdenka Prokopova
TBU in Zlín

Brewer, E. A. (2000). *Towards robust distributed systems.* Paper presented at the PODC.
Holubová, I., Kosek, J., Minařík, K., & Novák, D. (2015). *Big Data a NoSQL databáze*: Grada.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# Scalability - horizontal (scaling out)

**The system is distributed over multiple machines** (computers) / nodes (cluster)

- The possibility of using common commercial machines
  - Cost effective
  - It provides higher scalability than the vertical approach
  - Data is distributed across many machines
  - The application can use the main memory of all machines
- It introduces new problems
  - Synchronization, consistency, partial failure processing, etc.

Zdenka Prokopova
TBU in Zlín

Brewer, E. A. (2000). *Towards robust distributed systems.* Paper presented at the PODC.
Holubová, I., Kosek, J., Minařík, K., & Novák, D. (2015). *Big Data a NoSQL databáze*: Grada.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# Disadvantages of horizontal scalability

- Several conditions are necessary for distributed computing to work, which are **not always** guaranteed:
  - A reliable network
  - Secure network
  - Zero delay (latency)
  - Unlimited bandwidth
  - Immutable network topology
  - Homogeneous network
  - Zero data transfer costs
  - The network has only one administrator

Zdenka Prokopova
TBU in Zlín

Brewer, E. A. (2000). *Towards robust distributed systems.* Paper presented at the PODC.
Holubová, I., Kosek, J., Minařík, K., & Novák, D. (2015). *Big Data a NoSQL databáze*: Grada.

# Data distribution - distribution models

**Data distribution** – necessary if the scope of the data does not allow processing on one DB server

- Two orthogonal data distribution techniques:
  - **Sharding** – distribution of different parts of data (shards) on different nodes in the cluster
    - Capacity increase
  - **Replication** - creating copies of data on multiple nodes in a cluster
    - Increase availability and throughput
    - Number of required data replicas on different nodes - replication factor
- Sharding and replication can be used **separately** or **in combination**

Zdenka Prokopova
TBU in Zlín

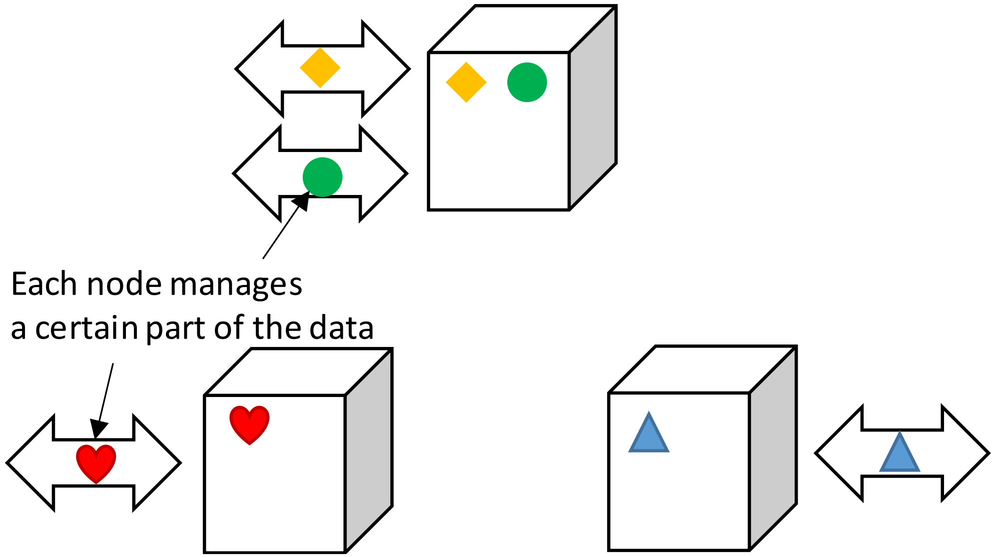
Brewer, E. A. (2000). *Towards robust distributed systems.* Paper presented at the PODC.
Holubová, I., Kosek, J., Minařík, K., & Novák, D. (2015). *Big Data a NoSQL databáze*: Grada.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# Data division - sharding

- Due to horizontal scaling, it is necessary to divide the data into subsets and place them on different nodes in the cluster
- Users access different parts of the dataset, i.e. different nodes, servers...
- Each node maintains a certain part of the data
  - It only reads and writes its data




Zdenka Prokopova
TBU in Zlín

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# Data division - sharding

**We try to ensure:**

- Even distribution of data on nodes
  - Maintain a balanced load (may change over time)
- Minimizing the number of nodes the user must access
  - E.g. the user gets all the data from one server
- Optimizing the physical distribution of data according to their geographical affiliation
  - E.g. companies, cities, countries...

Many NoSQL databases offer automatic sharding

A node **failure** causes data unavailability

- Sharding is often combined with replication

Zdenka Prokopova
TBU in Zlín

Sadalage, P. J., & Fowler, M. (2013). *NoSQL distilled: a brief guide to the emerging world of polyglot persistence*: Pearson Education.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# Replication

**Replication** is copying data to multiple nodes.

Replication Types:

1. master-slave - based on the principle of master and slave nodes
2. peer-to-peer - based on the principle of node equality - "equal to equal"

Zdenka Prokopova
TBU in Zlín

Sadalage, P. J., & Fowler, M. (2013). *NoSQL distilled: a brief guide to the emerging world of polyglot persistence*: Pearson Education.

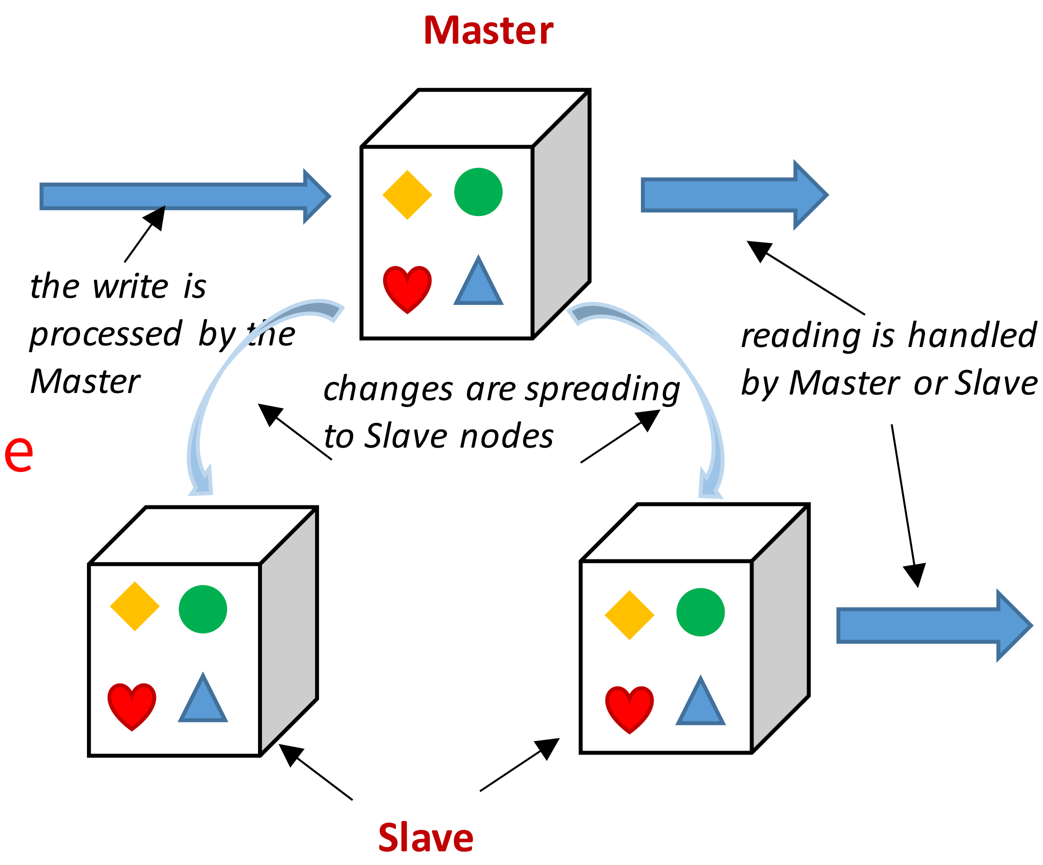
EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# Replication: master-slave

One node is designated as **primary** - **master**, the others as **secondary** - **slave**

- **Writing** (updating) data is processed only by the master
- **Reads** can be processed by any node




**Suitable primarily for reading data!**

Zdenka Prokopova
TBU in Zlín

Sadalage, P. J., & Fowler, M. (2013). *NoSQL distilled: a brief guide to the emerging world of polyglot persistence*: Pearson Education.

# Replication: master-slave

- In case of multiple read requests
  - Multiple slave nodes
- In case the master fails
  - Slaves can still process read requests
  - Slave can quickly become the new master (it is a replica)
- Limited by the master node's ability to process updates
- Node masters are selected by the administrator or automatically

Zdenka Prokopova
TBU in Zlín

Sadalage, P. J., & Fowler, M. (2013). *NoSQL distilled: a brief guide to the emerging world of polyglot persistence*: Pearson Education.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

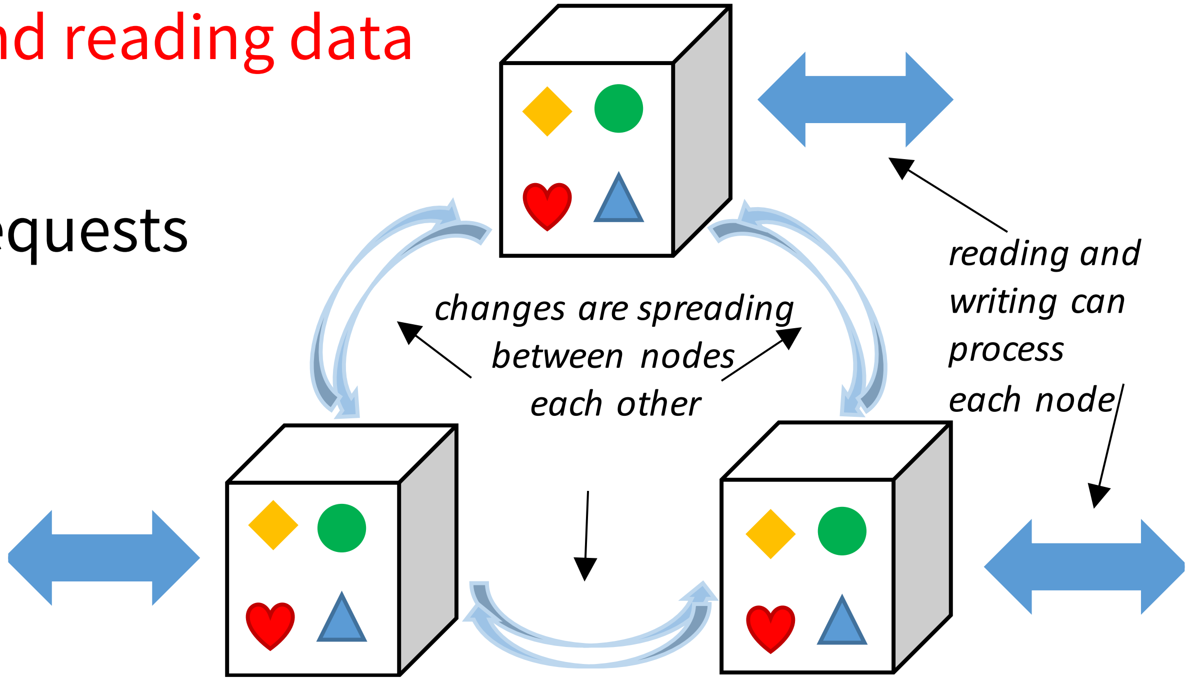
MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# Replication: peer-to-peer

- All replicas are equal (nodes are equal)
- Each node can handle both writing and reading data

Centralized management of data write requests

**Suitable primarily for data recording!**




Zdenka Prokopova
TBU in Zlín

Sadalage, P. J., & Fowler, M. (2013). *NoSQL distilled: a brief guide to the emerging world of polyglot persistence*: Pearson Education.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# Replication: peer-to-peer

- **A problem** with maintaining data consistency (timeliness of data).
  - Users can write simultaneously on two different nodes
  - Threat of write-write conflict
- **Solution**:
  1. When writing, the replicas (nodes) coordinate
     - At the cost of network traffic – it slows down the system
  2. Application of the electoral principle (majority decides)
     - It is implemented through a quorum (the minimum number of nodes on which the operation must take place in order to be successful is determined)



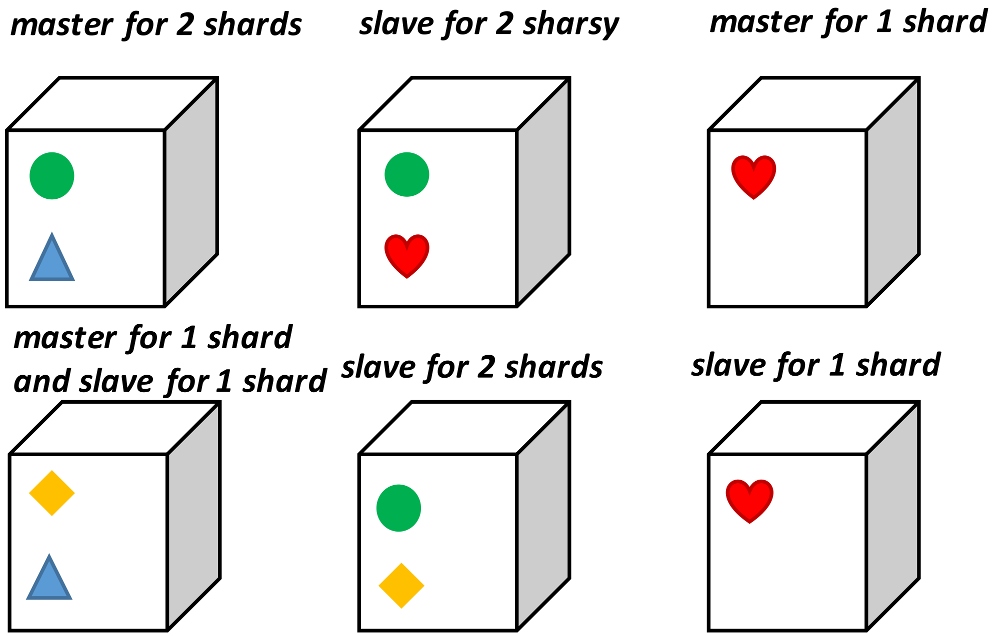
Brewer, E. A. (2000). *Towards robust distributed systems.* Paper presented at the PODC.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# Combination: sharding and M-S replication

- Sharding and master-slave replication:
  - **Each data part** (after sharding) is replicated through a **single master node**
  - **A node** can be **a master** for some data **and a slave** for other data




Zdenka Prokopova
TBU in Zlín

Sadalage, P. J., & Fowler, M. (2013). *NoSQL distilled: a brief guide to the emerging world of polyglot persistence*: Pearson Education.

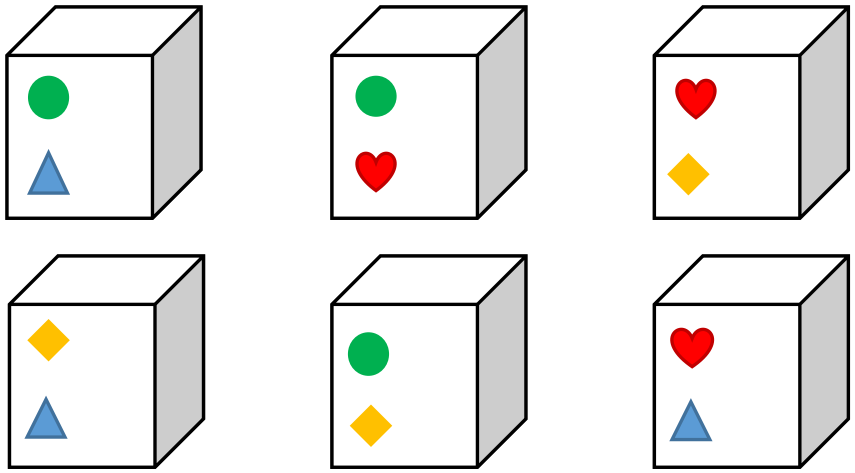
EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# Combination: sharding and P-P replication

- Sharding and peer-to-peer replication
  - Each piece of data is stored on N nodes (N-replication factor)
  - A typical default setting is a **replication factor** of 3
  - Individual nodes contain different pieces of data
  - Each shard is present on N nodes

=> The need to address **data consistency** issues



- each shard is present on three nodes

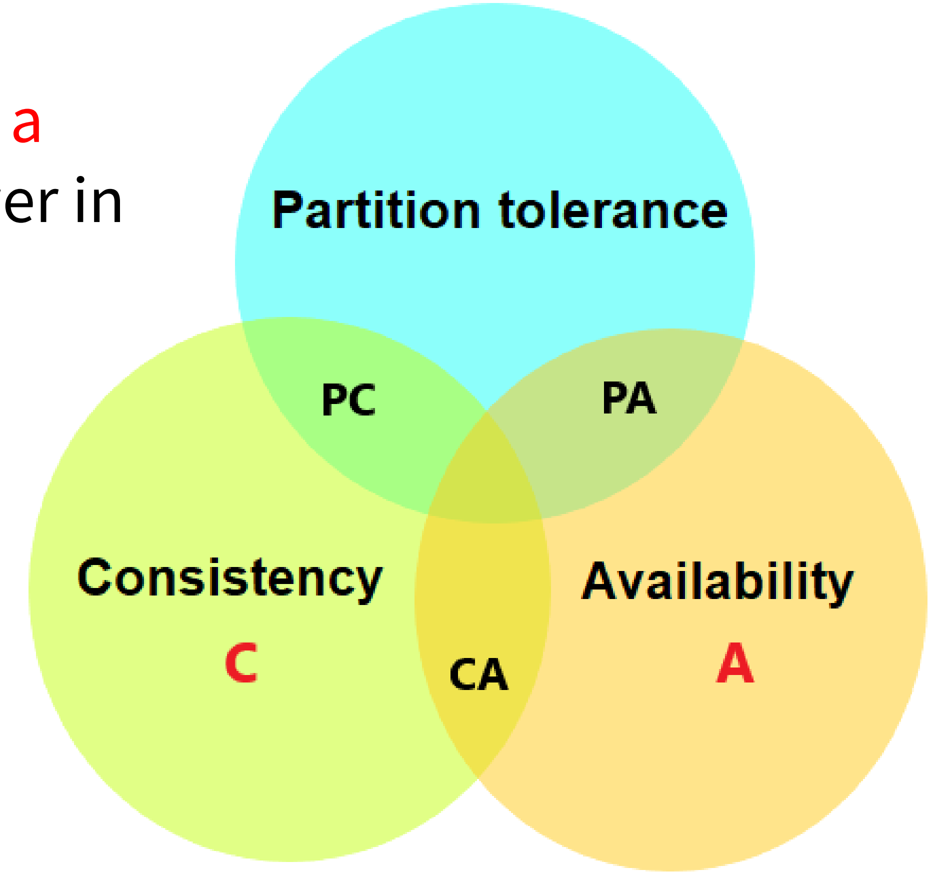
Zdenka Prokopova
TBU in Zlín

Sadalage, P. J., & Fowler, M. (2013). *NoSQL distilled: a brief guide to the emerging world of polyglot persistence*: Pearson Education.

# CAP theorem

The **CAP theorem** characterizes the properties of a distributed system (formulated by Prof. Eric Brewer in 2000):

- **C**onsistency
  - Ability to always return the same/correct data
- **A**vailability
  - Ability to respond to every request
- **P**artition tolerance
  - Resistance to network breakdown




Zdenka Prokopova
TBU in Zlín

Brewer, E. A. (2000). *Towards robust distributed systems.* Paper presented at the PODC.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# CAP theorem

Consistency

- After the update, all readers in the distributed system see the same data (replication)
  - All nodes should always contain the same data

Availability

- If the node (server) is running, it can read and/or write data
  - Every inquiry must result in an answer

Partition Tolerance

- The system continues to operate even if the network goes down and the servers are isolated
  - A network connection failure should not shut down the system

Zdenka Prokopova
TBU in Zlín

Brewer, E. A. (2000). *Towards robust distributed systems.* Paper presented at the PODC.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# CAP theorem

- It is not possible to provide all three CAP properties in a distributed system
  - Only two guarantees are possible - PC, PA or CA

- In real operation, various NoSQL databases offer network partition tolerance - P and a suitable trade-off between consistency - C and availability - A
  - We can wish for availability and thereby gain partial consistency
  - We can wish for strong consistency and thus get partial availability

Zdenka Prokopova
TBU in Zlín

Brewer, E. A. (2000). *Towards robust distributed systems.* Paper presented at the PODC.

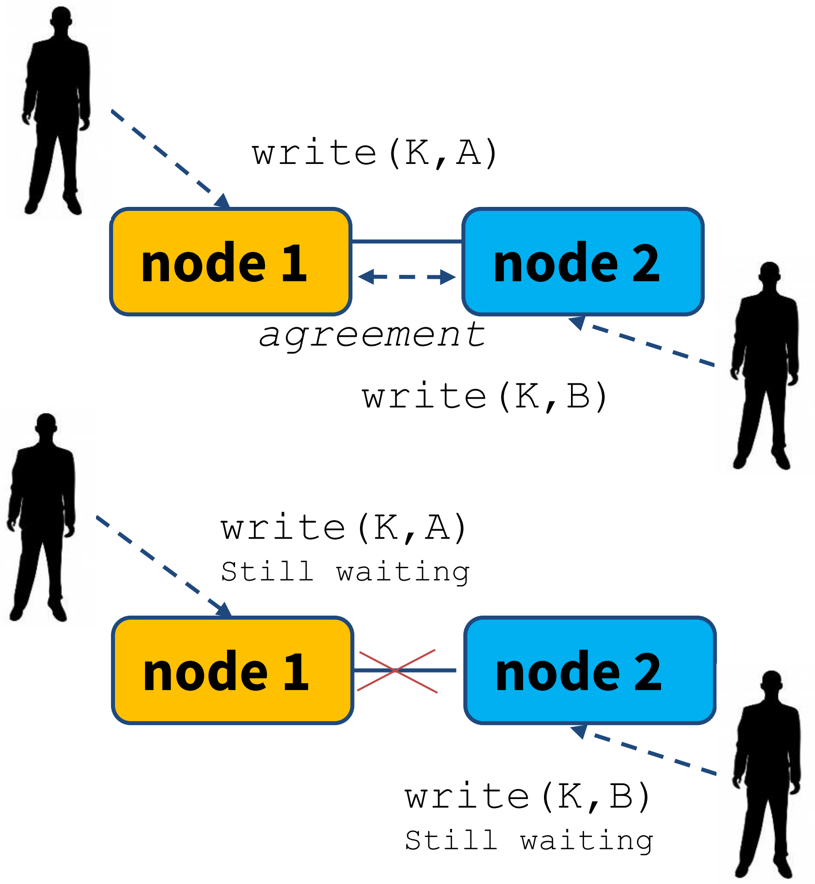
# PC: Partition Tolerance and Consistency

**Example**:

Two users, two nodes, two write attempts

**Strong consistency**:

- Before writing, both nodes must agree on the order of the writes
- If the nodes are split, we lose availability - writes are not possible,
- - reads are still available




Zdenka Prokopova
TBU in Zlín

Brewer, E. A. (2000). *Towards robust distributed systems.* Paper presented at the PODC.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# Consistency

- Timeliness / correctness of data
- **Centralized RDBSs** ensure strong consistency
  - Required fulfillment of ACID properties (Atomicity, Consistency, Isolation and Durability)
- Distributed NoSQL databases must choose compromise solutions
  - **CAP** – compromise between consistency and availability
  - **BASE** (possible consistency, essentially available, soft state…)

Zdenka Prokopova
TBU in Zlín

Hills, T. (2016). *NoSQL and SQL data modeling: bringing together data, semantics, and software*: Technics Publications.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# Occasional consistency – BASE model

**BASE model** - alternative to ACID properties of RDBS

- **B**asically **A**vailable
  - The system works basically all the time
  - Partial failures may occur, but without total system failure
- **S**oft state
  - The system is unstable - a non-deterministic state
  - Changes are happening all the time
- **E**ventual consistency
  - The system will eventually be in a consistent state sometime in the future

Zdenka Prokopova
TBU in Zlín

Brewer, E. A. (2000). *Towards robust distributed systems.* Paper presented at the PODC.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# Transaction processing in NoSQL

- **Centralized databases - RDBS**
  - Hassle-free transaction processing
  - ACID properties may be implemented
- **Distributed systems**
  - Concurrency of transactions
  - The emergence of conflict situations
    - Write-read conflict
    - Read-write conflict
    - Write-write conflict

Zdenka Prokopova
TBU in Zlín

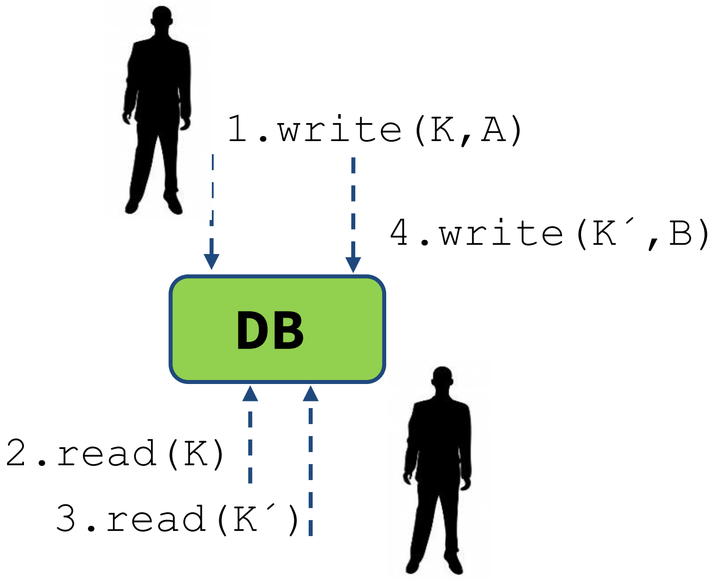
Brewer, E. A. (2000). *Towards robust distributed systems.* Paper presented at the PODC.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# Write-read / Read-write conflict

- **Problem**: one user reads in the middle of another user's writes (inconsistent read)
  - Logical inconsistency
- **Ideal solution**: transaction (ACID)
  - Ensuring strong consistency



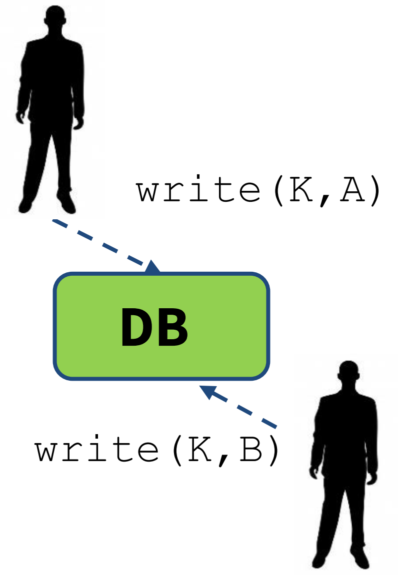

Zdenka Prokopova
TBU in Zlín

Brewer, E. A. (2000). *Towards robust distributed systems.* Paper presented at the PODC.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# Write-write conflict




- **Problem**: two users want to change the same record
  - Lost update, second update is based on old data

**Two general solutions**:

- Pessimistic approach: avoiding conflicts
  - Acquiring lock write before update
- Optimistic approach: enables conflicts - detects them and takes steps to resolve them
  - Conditional updates, save both updates and record the conflict
  - Implementation of e.g. time stamp

Zdenka Prokopova
TBU in Zlín

Brewer, E. A. (2000). *Towards robust distributed systems.* Paper presented at the PODC.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# Transaction processing in NoSQL

**Distributed transactions**

- X/Open Distributed Trans. Processing Model (X/Open XA)
- Two-Phase Confirmation Protocol (2PC)
- Strong (Strict) Two-Phase Locking (SS2PL)

Zdenka Prokopova
TBU in Zlín

Pivert, O. (2018). *Nosql data models*. Hoboken, NJ: ISTE Ltd / John Wiley and Sons Inc.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

# References


- Brewer, E. A. (2000). *Towards robust distributed systems.* Paper presented at the PODC.
- RNDr. Irena Holubova, Ph.D. MMF UK course PA195: NoSQL Databases.
- Fowler, A. (2015). *NoSQL for dummies*. Hoboken, NJ: John Wiley & Sons, Inc.
- Hills, T. (2016). *NoSQL and SQL data modeling: bringing together data, semantics, and software*: Technics Publications.
- Pivert, O. (2018). *Nosql data models*. Hoboken, NJ: ISTE Ltd / John Wiley and Sons Inc.

# References


- Lemahieu, W., Broucke, S. v., & Baesens, B. (2018). *Principles of database management : the practical guide to storing, managing and analyzing big and small data* (First edition. ed.). New York, NY: Cambridge University Press.
- Sadalage, P. J., & Fowler, M. (2013). *NoSQL distilled: a brief guide to the emerging world of polyglot persistence*: Pearson Education.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research, Development and Education

MINISTRY OF EDUCATION, YOUTH AND SPORTS

# Questions?

Strategic project of TBU in Zlín, reg. no. CZ.02.2.69/0.0/0.0/16_015/0002204

Zdenka Prokopova
TBU in Zlín