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

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Advanced Database Systems

Key-value stores

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

- Principles of key-value databases
- Main techniques
  - Data distribution
  - Data Consistency
  - Availability of data
- Applicability and use
  - Basic and advanced functions

Zdenka Prokopova
TBU in Zlín

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

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Basic principles

- A key-value store we can imagine as:
  - Associative array (map)
  - A simple hash table storing arbitrary values according to a unique key
  - That is the principle is **key-value mapping**
- Key - unique data identifier - access to data via the primary key
- Values - stored data - can be divided by type into separate storage (buckets)
  - Repositories can be understood as **namespaces** of keys
- The repository does not examine the contents of the values (any type) - the client application does

Zdenka Prokopova
TBU in Zlín

Holubová, I., Kosek, J., Minařík, K., & Novák, D. (2015). *Big Data a NoSQL databáze*: Grada.

Basic principles

- Advantages of this type of storage:
  - Simple structure
  - Simple operations
  - Allows data to be distributed efficiently (sharding and replication)
  - It works very fast
  - It is used by both **write and reads-intensive** applications (e.g. write every millisecond)

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

Basic principles

Basic operations:

- **PUT** – entering / updating the key value
- **GET** - getting the key value
- **DELETE** - delete key and value from storage

- Key - identifier of stored data objects - key role for work
  - Native - The natural primary key
  - Artificial (derived, created) primary key
    - E.g. hash function and timestamp (generated by the system)

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

Basic principles

Unsuitable use:

- Relationships between data
  - Relationships between different data sets
- Data queries
  - Search based on data from the value part
- Operations by sets
  - Operations that are limited to one key at a time
- Multi-operational transaction
  - Storing multiple keys

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

Basic principles

Typical use:

- Web Application Caching
- Saving information from the so-called session
- Management of shopping cart items (e-shop)
- User Profiles

Comparison with RDBS:

  - A table with two columns:
  - ID column (primary key)
  - DATA column to store the value

A database instance	Riak cluster
table	bucket (namespace)
row	key-value
row id	key

Zdenka Prokopova
TBU in Zlín

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

Basic principles - example

E-shop: storage of user information

- User Profiles
- Web sessions
- Contents of shopping carts

1. Storing all data in one namespace
   - Use: if the application will need all the information at once

namespace User

Key	UserID
Value:	UserProfile Sessions ShoppingList .item1 .item2

Zdenka Prokopova
TBU in Zlín

Holubová, I., Kosek, J., Minařík, K., & Novák, D. (2015). *Big Data a NoSQL databáze*: Grada.

Basic principles - example

2. Storing data in multiple namespaces
   - Use: when part of the application only needs some information
   - All data will require multiple accesses (3)

namespace Sessions

Key	UserID
Value:	Sessions

namespace ShoppingList

Key	UserID
Value:	ShoppingList .item1 .item2

namespace UserProfiles

Key	UserID
Value:	UserProfile

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

Representatives

- First - Amazon Dynamo (2007)
- Each representative has a slightly different purpose
- Various techniques were used to complete the tasks
- Other functions may vary - advanced system-specific functions
- The most popular representatives:
  - Persistent distributed systems (Riak, Redis, Infinispan...)
  - Libraries for creating embedded disk storage (BerkeleyDB, LevelDB, RocksDB, MapDB...)
  - Memory caches (Memcached, Ehcache, Hazelcast...)

Zdenka Prokopova
TBU in Zlín

DB-Engines Ranking. (2019). Retrieved from https://db-engines.com/en/ranking

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

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Representatives

redis
MEMCACHED
hazelcast
AEROSPIKE
EHCACHE
riak
ORACLE NOSQL DATABASE
levelDB
ORACLE BERKELEY DB 12c
Infinispan
RocksDB
amazon DynamoDB
MapDB
Project Voldemort

Zdenka Prokopova
TBU in Zlín

DB-Engines Ranking. (2019). Retrieved from https://db-engines.com/en/ranking

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

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Selected challenges and solutions

Challenge	Technical solution
Data distribution (sharding)	Consistent hashing
Read scalability & reliability	Data replication
Node detection connection/abandonment/failure	Gossip protocol (no centralized registry of node membership and activity)
Managing Replicas	Version stamps, vector clock
Competition, transactions	Two Phase Confirmation Protocol, MVCC

Zdenka Prokopova
TBU in Zlín

Erl, T., Khattak, W., & Buhler, P. (2016). *Big Data Fundamentals: Concepts, Drivers \& Techniques*: Prentice Hall Press

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

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Data distribution

- Data can be distributed among multiple nodes of a distributed system
- On which node the key-value pair will be stored determines:
  - Direct key
  - Hash function - hash(key)

Standard hashing - based on the modulo operation

– uniform distribution of values under the condition of a constant number of nodes

node number = hash(key) **mod** (number of nodes)

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

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Consistent hashing

- It is used for **automatic data sharing** between nodes (automatic sharding) in case of a **change in the number of nodes**
- It is based on standard hashing

It applies: **each node is responsible for a continuous interval of hash keys**

- Each node in the system is assigned a hash key
- We use the same hash function for data and nodes

Zdenka Prokopova
TBU in Zlín

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

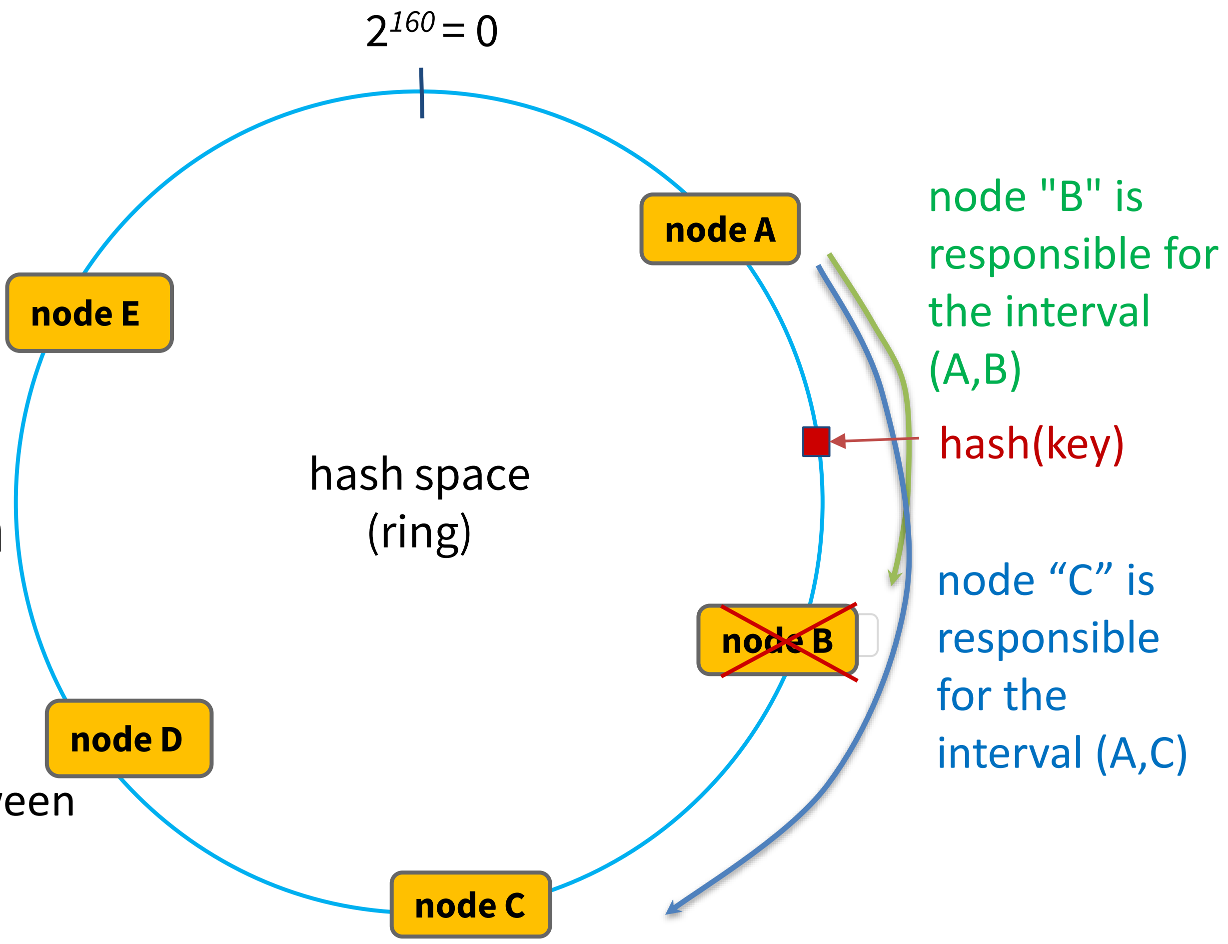
Consistent hashing

- Each node is assigned a hash key from the same domain, e.g.

  hash: Keys->$[0,2^{160}-1]$

- Each subsequent node in a clockwise direction is "responsible" for all keys from the interval of the previous one and its own key

- the new node is assigned a hash (key) – 1 key interval between two nodes is split – the data is moved to the new node
- I only need to move the data between nodes A and B



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

Consistent Hashing - Node detection

- Information about the change in the number of nodes is "spread" by so-called **gossip protocols**
- Principle:
  - At regular time intervals, each node randomly selects one of its known nodes, contacts it and gives it "news" (up-to-date information about the nodes in the system)

- Disadvantages of consistent hashing:
  - Uneven distribution of data between cooperating nodes
  - The concept of virtual nodes is used for load balancing

Zdenka Prokopova
TBU in Zlín

Holubová, I., Kosek, J., Minařík, K., & Novák, D. (2015). *Big Data a NoSQL databáze*: Grada.

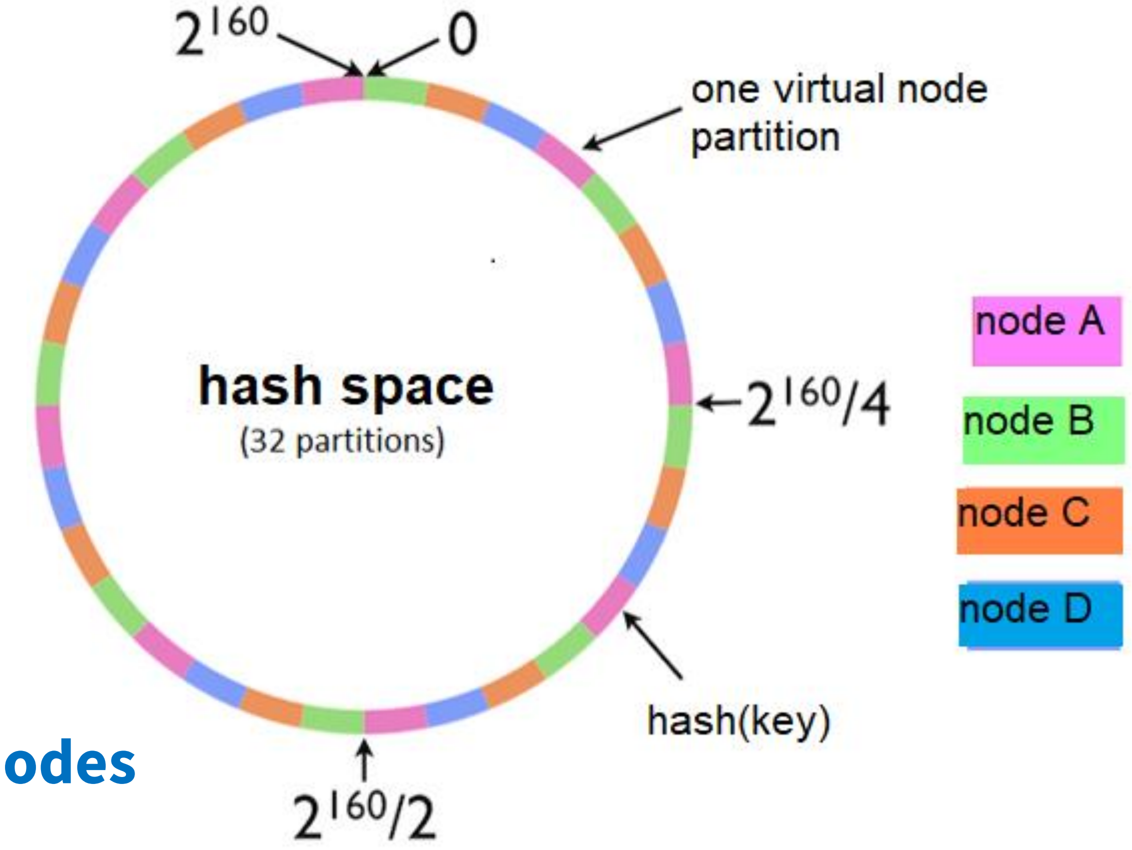
EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Consistent hashing

Virtual nodes

- The domain of the hashing function is divided into a *fixed number of equally sized intervals* – virtual nodes (partitions) – Q
- These are gradually assigned to physical nodes (servers) – S
- Q/S – partitions per 1 node
- It is assumed: Q >> S



Balanced distribution of data to physical nodes

Zdenka Prokopova
TBU in Zlín

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

Consistent Hashing - Replication

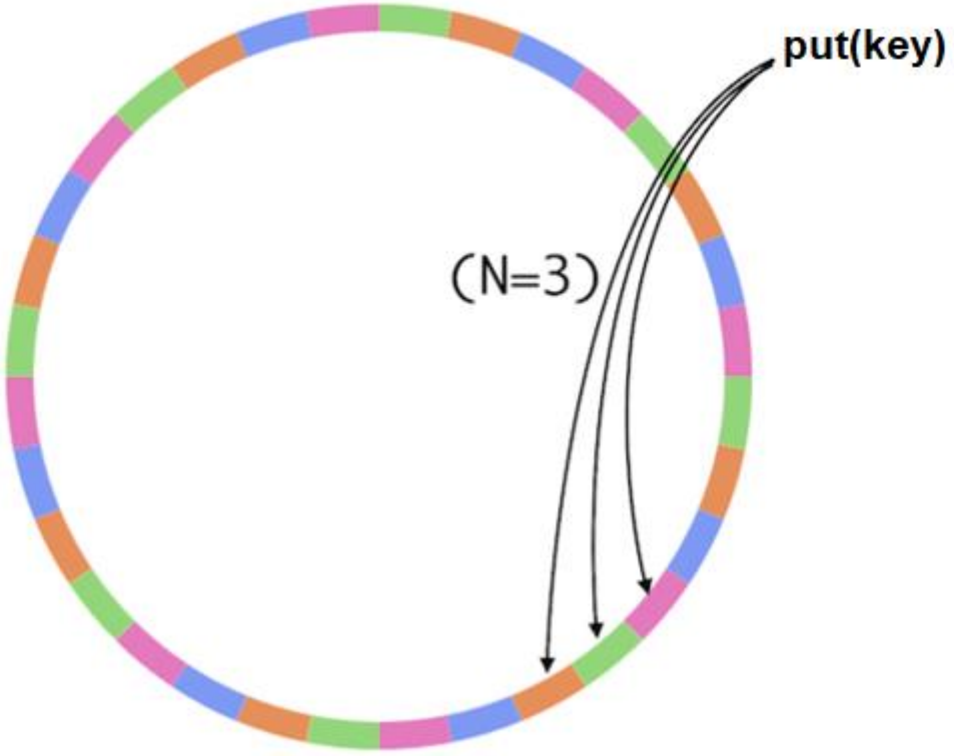
Data replication – ensures reliability

Performance boost - read/write operations

Increased availability - tolerance of the system to outages

Each object stored on N (replication factor) consecutive nodes

Both master-slave and peer-to-peer replication are possible



Zdenka Prokopova
TBU in Zlín

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

Consistency and availability

Manage replicas

- Using the concept of quora
  - N = replication factor (each pair (key, value) is replicated on 3 nodes by default, i.e. N = 3)
  - W = data must be written to at least W nodes
  - R = data must be found on at least R nodes

$$\boxed{W > N/2}$$

$$\boxed{R + W > N}$$

- R/W values:
  - Can be set by the user for each individual operation
  - all / one / quorum / default / integer value

Zdenka Prokopova
TBU in Zlín

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

Version stamps

- Mechanisms for: avoiding / detecting conflicts
  - On one data record
  - Write-write conflicts
- Version stamp in general:
  - A helper attribute stored with a key that changes each time the value for that key changes
- Basic use (also in centralized system):
  - The client reads the stamp along with the record
  - When the record is updated later, the stamp is sent back along with the new value and checked
  - If the stamp is different from the actual stamp => conflict

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

Version stamps

Stamps can be created in several ways:

- Counter – changes after each record update
  - Advantages - it is clear which version is newer
  - Disadvantages - Duplications must be avoided
- GUID(globally unique identifier) – generated large unique random number
  - Advantages - anyone (client) can generate
  - Disadvantages - you can't check up-to-date
- Hash from the data
  - Advantages - anyone can generate it, it is deterministic
  - Disadvantages - you can't check up-to-date

Zdenka Prokopova
TBU in Zlín

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

Version stamps

Time stamps

- Advantages - Timeliness is clear
- Disadvantages - clock synchronization required, sufficient granularity required

Combination:

- Counter + Hash:
  - Counter - comparison of actuality
  - Hash - if two updates appear simultaneously on two servers (with the same counter), the hash identifies the conflict

Zdenka Prokopova
TBU in Zlín

Holubová, I., Kosek, J., Minařík, K., & Novák, D. (2015). *Big Data a NoSQL databáze*: Grada.

Version stamps on multiple nodes

- Master-slave replication: one "master", everything works fine
- Peer-to-peer replication:
  - Any node can handle the update (multiple masters)
  - When contacted for an update, the node must respond to the client immediately after saving the new value
    - It cannot wait for all partners to confirm the update (two-phase confirmation protocol)
- Goal: A distributed algorithm that would
  - Reliably detect write-write conflict
  - Algorithm struck a balance between doing the write-up and avoiding conflicts
    - Allows user settings

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

Version stamps on multiple nodes

Possible solutions:

1. Using counters - remembering the history of counters
    - If the counters on 2 nodes do not match and one of the counters is not in the history of the other=>conflict
    - Space intensive (not used in NoSQL databases)

2. Time stamps
    - Clock synchronization problems
    - A write-write conflict is not detectable for multiple masters

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

Version stamps on multiple nodes

3. Vector stamps

- Algorithms for generating partial ordering of events in a distributed system and detecting "conflicts"
- Each node has its own counter
    - For each data item
- A node's counter is incremented when a node updates the value for a given key
- Each node maintains information about the counter values of all nodes
- Nodes exchange information about the state of counters and values, in order to find out
    - Which value is new
    - Whether it is a conflict

Zdenka Prokopova
TBU in Zlín

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

Example: Vector stamps

We have three nodes A, B, C

- At the beginning, the nodes share the same value - the counters on the nodes [A: 1, B: 1, C: 1]
- Node C updates its value - the value of the counter increases[A: 1, B: 1, C: 2]

A (key, value1) [A:1, B:1, C:1]

B (key, value1) [A:1, B:1, C:1]

C (key, value2) [A:1, B:1, **C:2**]

If synchronization occurs
-
The C:2 value is propagated to all nodes

A (key, value2) [A:1, B:1, C:2]

B (key, value2) [A:1, B:1, C:2]

C (key, value2) [A:1, B:1, C:2]

Zdenka Prokopova
TBU in Zlín

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

Example: Vector stamps

- Node A updates its value to value3 and at the same time node B updates its value to value4 – values of vectors [A: 2, B: 1, C: 2] and [A: 1, B: 2, C: 2] are incomparable – write conflict

A (key, value3)
[**A:2**, B:1, C:2]

B (key, value4)
[A:1, **B:2**, C:2]

C (key, value2)
[A:1, B:1, C:2]

A conflict occurs
during
synchronization
-
Which value is
newer?
A or B?

Zdenka Prokopova
TBU in Zlín

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

Vector stamps - conflict resolution

Specific techniques - differ in the way of communication

The most frequently used techniques (and their variants):

- Vector clock (used by dynamo, etc.)
- Lamport's Time Stamps (1987)
  - Counters updated whenever nodes communicate
  - The last value can only be retrieved during a read
- Matrix Clock
- …

Zdenka Prokopova
TBU in Zlín

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

Conflict solving

There are three general ways to resolve write conflicts

- Differences between copies of distributed data need to be reconciled

1. Read correction - (Optimistic Strategy)
   - A correction is made when a read detects a mismatch
   - The read operation is slow
2. Write correction
   - The correction takes place during the write operation - deceleration
3. Asynchronous correction
   - Correction is performed independently of reading and writing - separate operations

Zdenka Prokopova
TBU in Zlín

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

Gossip protocols

It is a set of distributed protocols for:

- Each node periodically sends its current information
  - To a randomly selected partner
- Dissemination of information about the current status
  - Input / output / failed nodes
  - Asynchronous conflict reconciliation (anti-entropy)
  - Other properties

Zdenka Prokopova
TBU in Zlín

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

Transactions

- Some key-value stores allow full transactional processing
- The most used techniques:
  - Two-phase commit protocol (2PC – two-phase commit)
  - Multi-version concurrency control (MVCC -Multi-version concurrency control)
  - Transaction isolation levels

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

References

- Sadalage, P. J., & Fowler, M. (2013). *NoSQL distilled: a brief guide to the emerging world of polyglot persistence*: Pearson Education.
- Erl, T., Khattak, W., & Buhler, P. (2016). *Big Data Fundamentals: Concepts, Drivers \& Techniques*: Prentice Hall Press.
- Holubová, I., Kosek, J., Minařík, K., & Novák, D. (2015). *Big Data a NoSQL databáze*: Grada.
- DB-Engines Ranking. (2019). Retrieved from https://db-engines.com/en/ranking

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