Big Data

Tuur Vanhoutte

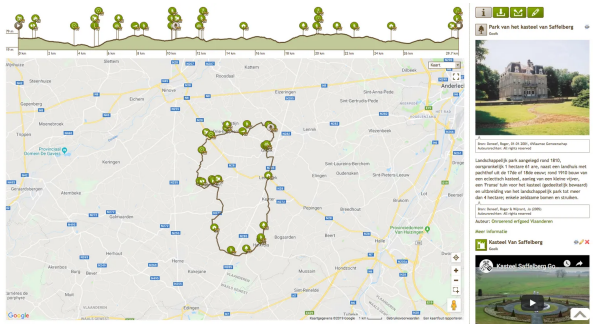
8 maart 2021

Inhoudsopgave

1 Understanding Data Intensive Applications

1.1 Why Big Data?

1.1.1 Use case: data intensive application RouteYou



Figuur 1: RouteYou

- Routes - user preferences & interests
- Searcheable Text data
- Geospatial data
- Community driven
  - Exponential user growth is necessary to make the application posssible
  - Server power/bills should grow linearly

1.2 Data Intensive Application: RAMS!

- **Reliable**
  - tolerating human mistakes
- **Available**
- **Maintainable**
  - Easy to adapt (evolvability)
  - Easy to deploy & operate (operations/sys admins)
- **Scalable**
  - User growth while maintaining low response times

1.2.1 Common similar abbreviations

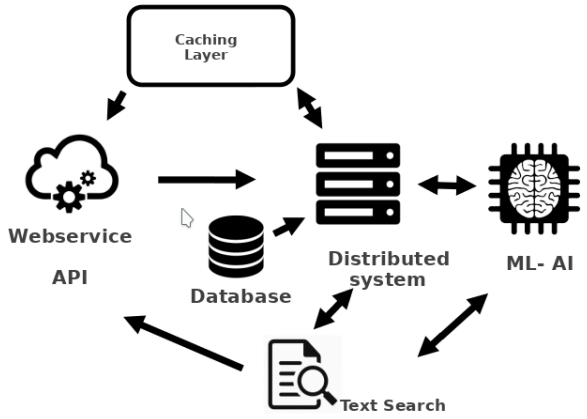
- Infrastructure: RAS (Reliable, Available, Serviceable)
- Developer: RMS (Reliable, Maintainable, Scalable)

1.2.2 Methods to improve Maintainability

- Github
- Error handling
- Relative paths (not absolute)
- Abstraction (REST API, . . .)
- Documentation

1.2.3 RAMS applied to RouteYou application

- Geospatial data (longitude, latitude)
- Available & scalable
- Scalable & low response time
- Community driven - unstructured text
- Maintainable: automatic classification of community input (ML)



Figuur 2: To support many users, you need a caching layer

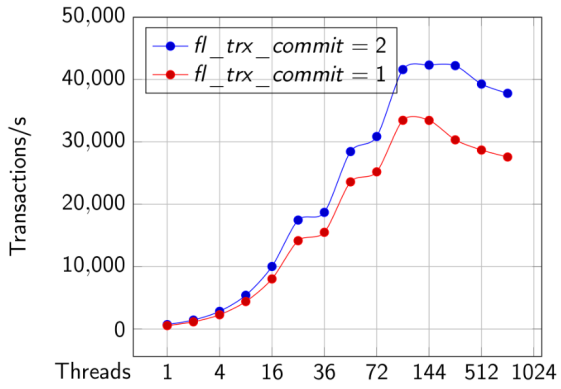
1.3 Learning outcome for this module

Being able to make infrastructure & software choices to build a Reliable, Available, Maintainable & Scalable (RAMS) data intensive application.

- Deep insights into database technology & cloud services
- Connecting with Machine Learning & AI
- Configuring a data back-end (in the cloud or locally)

1.4 Scaling

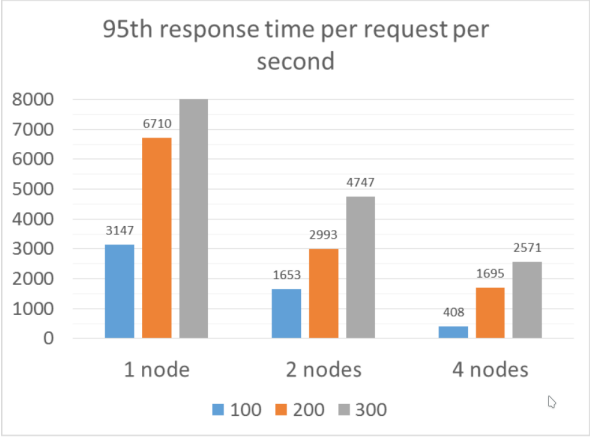
1.4.1 MySQL scaling



Figuur 3: Transactions/sec

- Processing power of 16-64 = slightly less then 4x
- Real performance: 2.3x
- = scaling up: add more processing power to the system

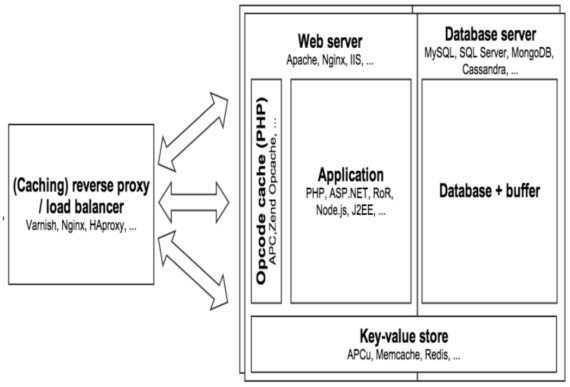
1.4.2 ElasticSearch Scaling: distributed system



Figuur 4: Response time per request

- Scaling out: add more servers to your data system

1.4.3 Professional architecture (Dev oriented)



Figuur 5: Professional architecture diagram

- **Reverse proxy** / **Load balancer:** improves scalability
- **Opcode/app/Webserver:** webservice + API
- **Key-value store:** 'caching layer'
- **Database server:** distributed storage system + relational database

1.4.4 Time series Distributed database (OpenTSDB, InfluxDB)

Location	DateTime	maxTemp	minTemp	prcp	dewpoint	gust	seaLevelPres	windSpeed
3927	2011-09-16T00:00:00	82	80.1	0	75.9		1015.9	5.9
3864	2011-09-16T00:00:00	82.4	71.6	0	70.2			4.2
3860	2011-09-16T00:00:00	86	75	0	68.7	22	1019	12
3858	2011-09-16T00:00:00	84	79	0	71.4		1018.5	4.9
3855	2011-09-16T00:00:00	84.9	78.1	0	70		1019.1	9.8
3853	2011-09-16T00:00:00	82	62.1	0	69.8	25.1	1014.9	12.5
3847	2011-09-16T00:00:00	87.1	78.1	0	56.6		1013	5.3
3846	2011-09-16T00:00:00	93.9	81	0	55.2		1006.4	2.9
3845	2011-09-16T00:00:00	91	78.1	0	62.4			3.3
3843	2011-09-16T00:00:00	59	52	0.05	39.2		1024.1	5.9
3837	2011-09-16T00:00:00	54	42.1	0.07	40.3	20	1017.7	5.3
3835	2011-09-16T00:00:00	57.9	53.1	0	40.9	28	1017.9	13.5
3833	2011-09-16T00:00:00	55.9	52	0.01	38.8	24.1	1019.5	14
3830	2011-09-16T00:00:00	57.2	48.9	0.04	42	20	1016.7	10.1
3827	2011-09-16T00:00:00	63	46.9	0	41.8	22	1010.8	13

Figuur 6: Data from windmill sensors. Most sensors log about every second

- Losing data is not that big a problem
- Massive amount of data to write

1.5 Scalability & application performance management

Response times and percentiles rule the web

1.5.1 The need for speed: some insights from Google

- Speed is a ranking factor
- When your site has high response times, less URLs will be crawled from your site
- 53% of visits are abandoned if a site takes longer than 3 seconds to load

- Slow websites will be labeled by Google Chrome

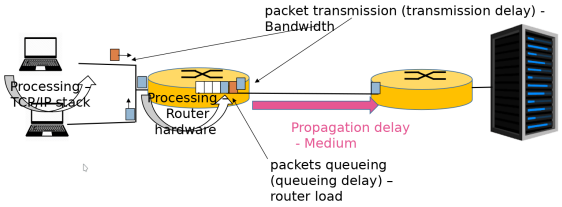
1.5.2 Response times for websites

- **Ideal:** "blink of an eyeïs 300-400 ms
- **Excellent:** 500ms to 1.5 seconds at most
- **Barely acceptable:** 3 seconds

Response time = Network latency + processing

- 2.9 seconds is faster than 50% of the web
- 1.7 seconds is faster than 75% of the web
- 0.8 seconds is faster than 94% of the web

1.5.3 4 components of network latency

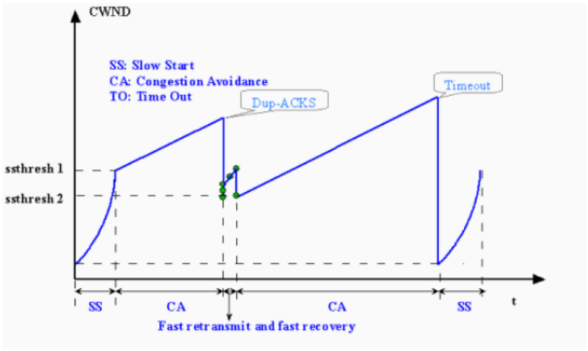


Figuur 7: Network latency diagram

- Processing delay
  - Processing network software stack (TCP/IP layers)
  - Routing decisions
- Transmission delay
  - Bits on physical link (Bandwidth plays a big role, ex: 1Gbit/s)
- Propagation delay
  - Speed of EM signals in fiber: 200.000 km/s (67% of lightspeed)
  - Changes with distance and medium (Copper: 64% of lightspeed)
- Queing delay
  - Time spent in router & NIC buffers

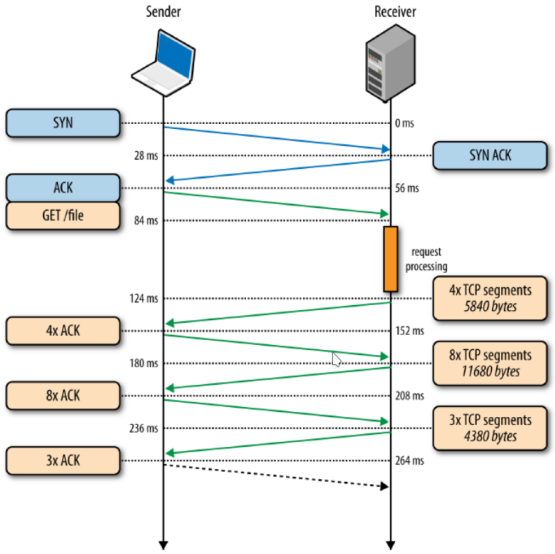
1.5.4 TCP Congestion Window - slow start

- Network congestion = a network node or link is carrying more data than it can handle
- The internet is built around dropped packages



Figuur 8: TCP Congestion window

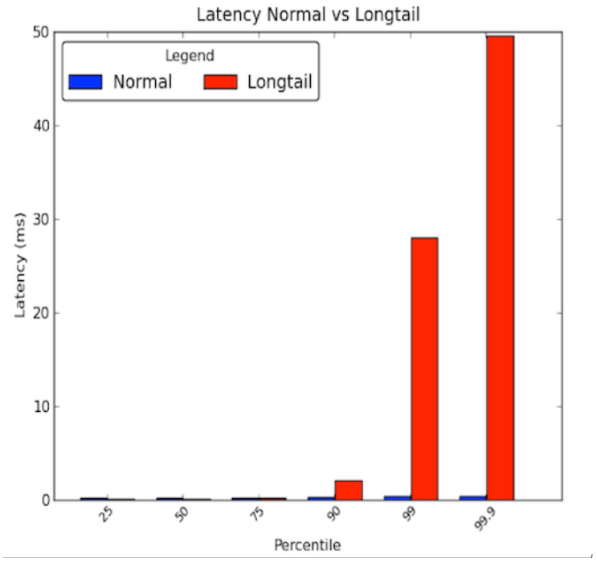
- 4-8-16-32 TCP segments (Win 2008, Win7)
- 10-20-40 (Linux 2.6+, Windows Server 2016 / Windows 10)



Figuur 9: Because of many handshakes, there is a lot of latency

- Solution: KeepAlive of a HTTP Persistent Connection
  - Only one 3-way handshake for many requests
  - Lower network & CPU load
  - Lower response times
  - **Downside**: more connections open ⇒ more memory, more connection failures, app crashing, . . .
- Measure parallel requests of a website using `https://www.webpagetest.org/`
- Get a waterfall view of a webpage

1.5.5 Long tail latency



Figuur 10: Long tail latency vs Normal latency

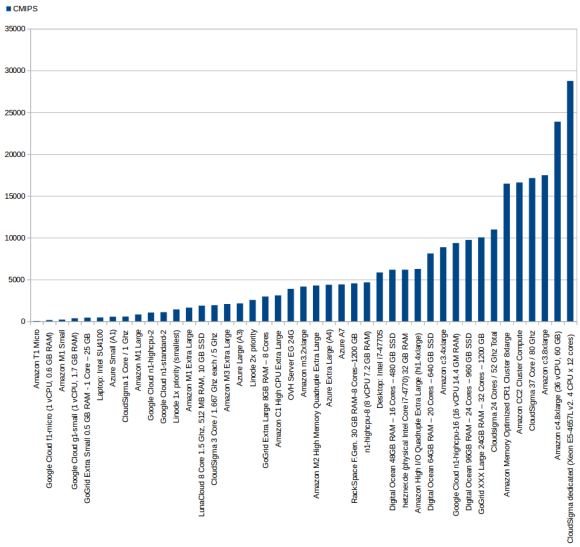
- Average = useless
- Long tail latency = 99th percentile
  - To be experienced by a lot more than 1% of users!
- Best customers encounter highest percentiles
- URL consists of many requests

1.6 Conclusion

- Our goal is RAMS (or RASS)
- Many data models & stores: transactional, timeseries, text search
- Website 99th percentile + DNS + TCP $\Rightarrow$ < 2s response time
  - Efficient caching
  - Think about your architecture (infrastructure + software) before coding

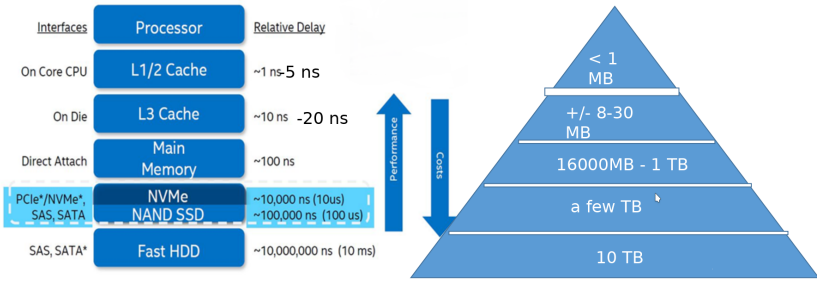
2 Professional storage

2.1 Cloud MIPS



Figuur 11: MIPS = Million Instructions Per Second

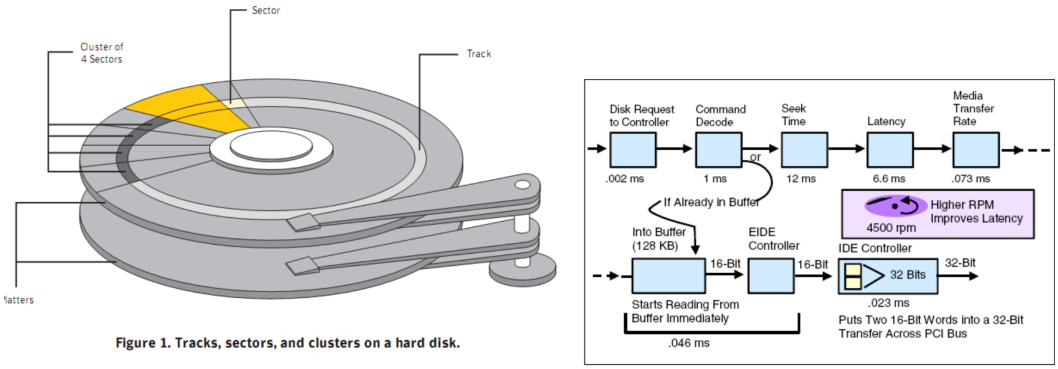
2.2 Latency vs storage space pyramid



Figuur 12: The higher the performance, the higher the cost per byte of storage

2.3 Storage media

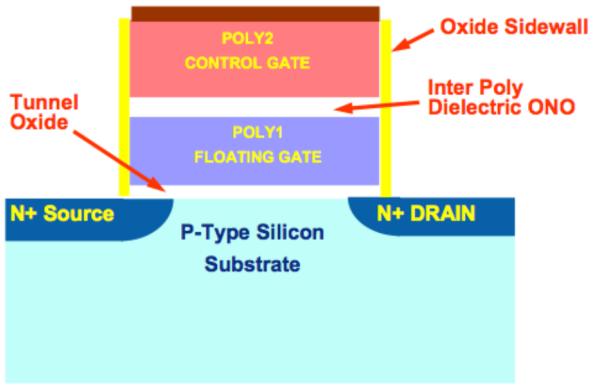
2.3.1 Magnetic disks



Figuur 13: Massive capacity but mechanical latency

- Seek time and latency are the key bottlenecks
- Need large quantity of disks for good server performance

2.3.2 Flash (NAND) / SSDs



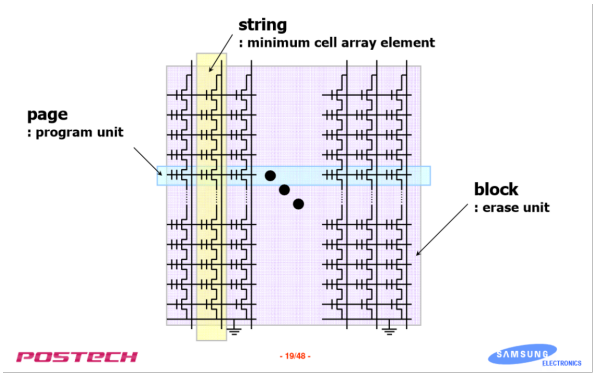
Figuur 14: Flash storage

- SSD = Solid State Drive
- NAND = MOSFET + floating gate
- Voltage between control gate and N+ : electrons in floating gate
- This works very quickly

Architecture

- Page = 4 KB, pages are in block
- Block = 128 pages (4KB * 128 = 512 KB)
- You can read or write page per page

- Erasing has to erase the entire block



Figuur 15: Diagram of a flash Block

2.3.3 Big difference between read and writing

	MLC NAND flash
Random Read (page)	50-100 µs
Erase (block)	1000-2000 µs per block
Programming (page)	40-250 µs

Figuur 16

- Limited number of writes
- Slow block write
- Limited "normal"write (programming)

2.3.4 IOPS vs Bandwidth

- Transactions & virtualized workloads: lots of random access
- Timeseries fileserving: mostly sequential
- HDD: random performance can be extremely low to medium
- IOPS = Input/Output Operations Per Second

Storage device	Seagate Enterprise HDD	Intel SSD NVMe
	ST8000NE0001	DC3700
Capacity	8 TB	800 GB
Spindle speed (rpm)	7200	N/A
Max. BW (MB/s)	**230**	**600**
Latency (ms)	4,16	N/A
Seek time	8	N/A
Total Random read time ms	12	0,08
Random Performance	1000 Random 4 KB blocks	1000 Random 4 KB blocks
Total Random read time (ms)	12000	80
Transfer time (ms)	17,4	6,7
Sustained Transfer rate (MB/s)	**0,33**	**46,15**
IOPS	83	11538
Sequential Performance	1x 4 MB block	1x 4 MB block
Total Random read time (ms)	12	0,08
Transfer time (ms)	17	7
Sustained Transfer rate (MB/s)	**136**	**593**

Figuur 17: An enterprise HDD vs an NVME SSD

2.3.5 Storage options

	Media Type	Interface	Read Latency (µs)	Write latency (µs)	Random IOPS	BW (MB/s)
HDD	Magnetic	SATA	10.000	10.000	100	1-200
Low-end SSD	NAND Flash	SATA	100-300+	40-2000+	5k-20k	100-550
High-end SSD	NAND Flash	NVMe	100-200+	20-1000+	50-200k	100-1800
3D-Xpoint	Electric resistance	NVMe	10-40	10-60	500+k	200-2000

Figuur 18: Storage options

2.3.6 Performance Conditions

Type	Queue depth	Random?	Write vs Read	Perf consistency
HDD	As low as possible (1-2)	Sequential ! Random as low as 50 IOPS	Write slightly slower	Terrible (1 -200 MB/s)
Low-end SSD	8-16	Random	Write can be a lot slower	IOPS writes can vary 2-4x
High-end SSD	16+	Both	Write can be a lot slower	IOPS writes can vary 10-30 percent
3D-Xpoint	2+	Both	Does not matter	Very good

Figuur 19: Performance Conditions

2.4 RAID

2.4.1 Definition

Redundant Array of Inexpensive Disks is a storage technology that combines multiple physical drives into one logical unit.

Purpose:

- Data redundancy
- Performance improvement
- Both

2.4.2 Hardware <> chip

2.4.3 Raid levels

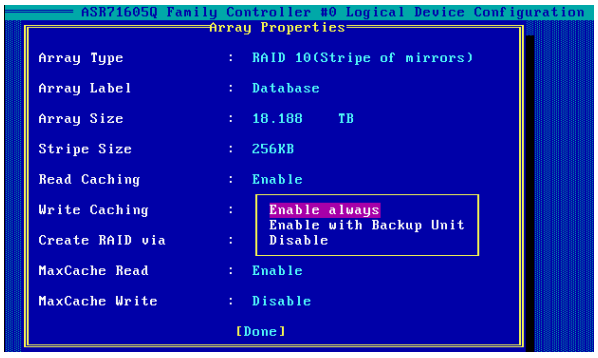
- RAID 0
- RAID 1
- RAID 5
- Combinations are possible (RAID 10, 01, 51, 15)

Level	Benaming	Schijven	Capaciteits verlies	Availability	Lezen sequentieel	Schrijven sequentieel	Lezen random	Schrijven random
RAID 0	Striping	Min. 2	geen	Slechter!	Sneller	Sneller	gelijk	Gelijk
RAID 1	Mirror	Min. 2	50%	Beter	iets Sneller	Gelijk	Sneller	Iets trager
RAID 10	Stripe + Mirror	Min. 4	50%	Beter	Sneller	Sneller	iets Sneller	Gelijk
RAID 01	Mirror + Stripe	Min. 4	50%	Beter	iets Sneller	iets Sneller	Gelijk	Gelijk
RAID 5	Stripe+ Parity	Min. 3	33%	Beter, slechter tijdens rebuild	Sneller	iets Sneller	iets Sneller	Gelijk

Figuur 20: RAID level choices

2.4.4 Caching & BBU

- RAM caching: to allow more users to access your data at a time
- RAID = lower latency by caching
- Not always durable: backup solutions needed like Battery Backup Unit (BBU)
- RAID = more bandwidth, +- same latency
  - Latency does not increase as fast when load increases (vs single disk)
  - More bandwidth & capacity available



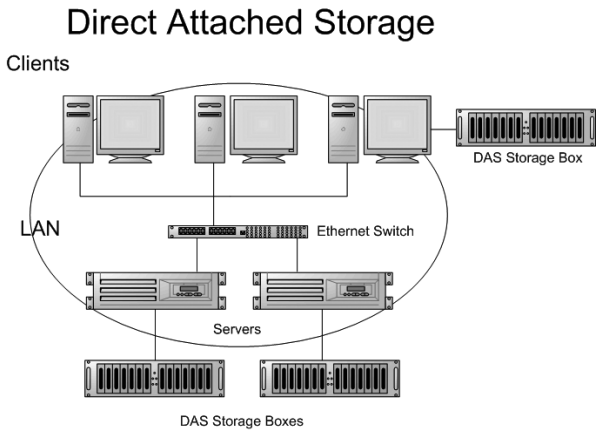
Figuur 21: RAID configuration

2.5 Professional Storage Topology

2.5.1 Components

- Enclosure
- Controller
- Disk Array
- HotSpare (=backup disk if a disk fails)
- LUN (logical unit number) / Volumes (= logical storage areas)

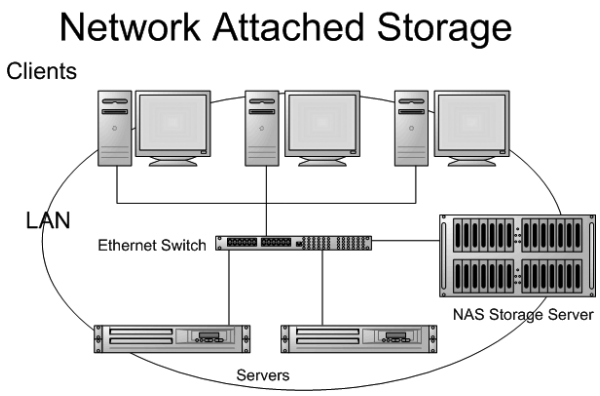
2.5.2 DAS - Block storage



Figuur 22

- Up to 122 disks per SAS controller
- Similar to disks inside the server
- No centralized back-up

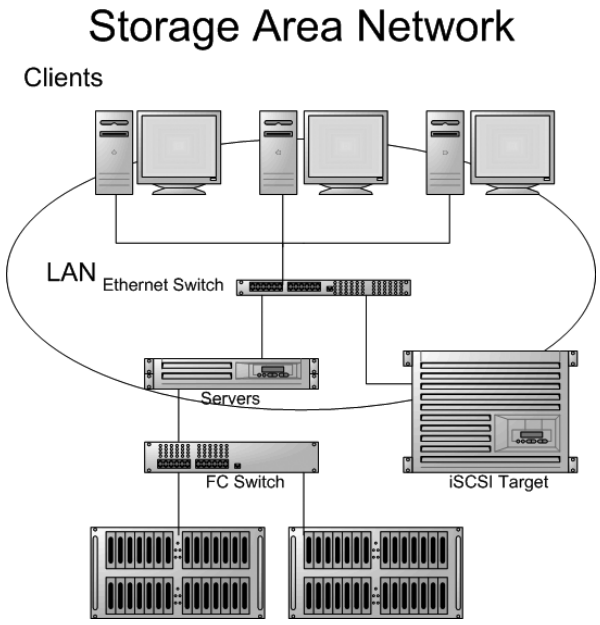
2.5.3 NAS - File storage



Figuur 23

- Common Internet File System (CIFS) for Windows
- $\rightarrow$ SMB protocol
- Network File System (NFS) for UNIX $\Rightarrow$ mounting via network
- SMB also available in Linux

2.5.4 SAN - Block storage on a network



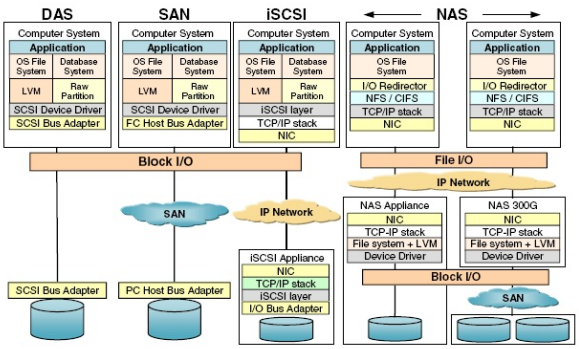
Figuur 24

- Seperate Block storage network

- Centralized backup & management
- Good scaling, no load on LAN
- But:
  - No standards - proprietary
  - Expensive

2.5.5 iSCSI terminology

- iSCSI Target = the iSCSI 'server'
  - IP + port = Portal
  - Portal: LUNs / Volumes
  - Volume = IQN
- iSCSI Initiator = the iSCSI 'client'
  - Connects targets
  - Find LUNs/Volumes



Block I/O compared to file I/O

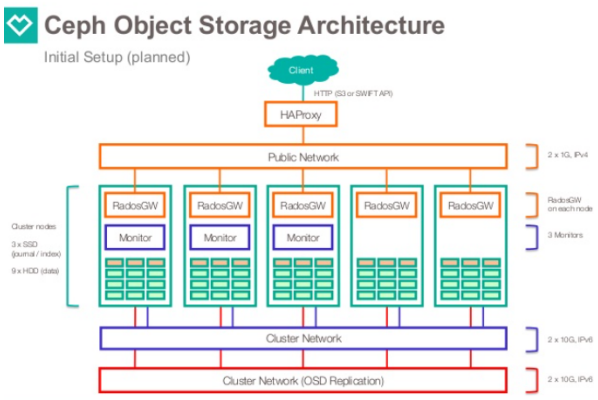
Storage characteristic	iSCSI SAN	Fibre Channel SAN	NAS
Protocol	Serial SCSI	Fibre Channel Protocol	NFS, CIFS
Network	Ethernet, TCP/IP	Fibre Channel	Ethernet, TCP/IP
Source / target	Server / Device	Server / Device	Client / Server or Server / Server
Transfer	Blocks	Blocks	Files
Storage device connection	Direct on network	Direct on network	I/O bus
Embedded file system	No	No	Yes

Figuur 25

2.5.6 Object storage

- NAS hardware
  - Distributed over multiple datacenters
- Object Data
  - Metadata
- Globally Unique Identifier

  - URL
  - RESTful API
- Examples:
  - AWS S3
  - Ceph - Lustre
  - Google Cloud storage



Figuur 26

2.5.7 Link with Databases & other data storage

- Transactional database: needs block storage
  - Performance
  - Durability
  - Consistency
- Block storage best for ‘raw data’ (no meta data involved)
- NAS = ‘file based’ services like sharepoint
- static objects on Object Cloud storage
  - good match for OOP & ‘unstructured data’
  - highly available
  - ‘Eventually’ consistent

3 Relational databases

Data intensive application: needs RAMS!

- **Reliable**
- Available
- Maintainable

- Scalable

3.1 Components of a relational database

- **Tables** = Relations are saved in the format of tables
- **Relationships** = a logical connection between different tables
  - Join, key, foreign key
  - Relation schema
- **Tuple** = A single row (record) of a table, which contains a single unordered record for that relation
  - A dataset representing an object, an item ('person')
  - Columns represent the attributes
  - Tuples are unique
  - Tuples are similar to Python dictionaries or JavaScript objects

SNAME	AGE	MAJOR	ID	SEX	ADDRESS	CITY	STATE
Anderson B.	19	CS	55555501	M	101 Rocket Way	Atlantis	CA
Barnes D.	17	MATH	55555502	M	1402 Elf Lane	Ruston	LA
Bronson P.	26	MATH	55555503	M	1 Web Master	Ruston	LA
Brooks D.	18	CS	55555504	F	900 Baird Street	Dallas	TX
Garrett D.	20	PSY	55555505	M	BGB Consulting	Dallas	TX
Howard M.	21	CS	55555506	M	5 Scarborough	Dallas	TX
Huey B.	20	CS	55555507	F	1 Historic Place	Jackson	MS
Kleinpeter J.	24	CS	55555508	M	69 Watson Lane	Ruston	LA
Kyzar D.	18	CS	55555509	M	49 Animae Way	Hammond	LA
Moore D.	19	MATH	55555510	M	No. 7 Seagram	Ruston	LA
Moore L.	20	MATH	55555511	F	2 Pot Place	New York	NY
Morton M.	30	ACCT	55555512	M	2010 Skid Row	Compton	CA
Pittard S.	22	ACCT	55555513	M	111 Easy Street	Ruston	LA
Plock C.	22	MGT	55555514	M	13 NSF Road	Ruston	LA
Slack J.	28	PSY	55555515	M	1 Pirate's Cove	Ruston	LA
Talton J.	19	PSY	55555516	M	666 Microsoft	Redmond	WA
Teague L.	18	PSY	55555517	F	Fern Gully Farm	Terry	LA
Tucker T.	45	MGT	55555518	F	Prop Wash Way	Eldorado	AR
Walker J.	23	CS	55555519	M	42 Ocean Drive	Venice	CA
Walker R.	21	CS	55555520	M	9 Iron Drive	Monroe	LA

Figuur 27: 1 relation 'student': 20 tuples, 8 attributes

3.2 Reliability problems

- Applications crash
- Client (website) - network - database
  - $\Rightarrow$ network is very unreliable
- Multi-threaded code: race conditions $\Rightarrow$ who gets access to 1 piece of data
- Disks can fail

3.3 Example

1 database: bank

- Checking account = table 1
- Savings account = table 2

3.3.1 The problem

```
SELECT saldo FROM checking WHERE customer_id = 10233276;
UPDATE balance SET balance = balance - 200.00 WHERE customer_id = 10233276;

# CRASH: -200 but not on savings account!

UPDATE Savings SET balance = balance + 200.00 WHERE customer_id = 10233276;

# Crash: +200, and application might try again: +400
```

3.3.2 The solution: Transactions

= multiple operations are executed on multiple objects as one unit

```
START TRANSACTION;
SELECT balance FROM checking WHERE customer_id = 10233276;
UPDATE checking SET balance = balance - 200.00 WHERE customer_id = 10233276;
UPDATE savings SET balance = balance + 200.00 WHERE customer_id = 10233276;
COMMIT;
```

VERY IMPORTANT! Every transaction is ACID

- **Atomic**
  - Each transaction is treated as a single 'unit', which either succeeds completely, or fails completely.
  - If all succeed ⇒ Commit transaction
  - If at least one fails ⇒ Rollback transaction
- **Constistent**
  - Data cannot get 'magically' deleted or added
  -
  - Example: when sending money to another bank account, the money cannot exist on both accounts after a transaction
- **Isolated**
  - Transactions cannot interfere with each other
- **Durable**
  - Data is written in a reliable way
  - Storage medium must be reliable

Commit / Rollback does not protect agains threads that overwrite each other! It only protects agains crashes from one thread.

3.4 Single object entry

Situation:

- Input = 1 record - row - object

  - What if the network fails while sending the input
- Single Object Atomicity & isolation:
  - Create log entry (WAL = Write Ahead Log)
  - Write lock when writing
  - Create log entry if successful
  - Restart if fail
- (Almost) all database - storage engines support this
- This is not a transaction!

3.5 Concurrency Control

3.5.1 Dirty Reads

Definitie 3.1 (Dirty Reads) *Dirty reads (aka uncommitted dependency) occur when a transaction is allowed to read data that has been modified by another running transaction, and not yet committed.*

Example:

```
1  -- 1: start the transaction
2  START TRANSACTION;
3  -- 2: check the current balance
4  SELECT balance FROM checking WHERE customer_id = 10233276;
5  -- 3: money is taken from the balance account
6  UPDATE checking SET balance = balance - 200.00 WHERE customer_id = 10233276;
7  -- 4: money is put on the savings account
8  UPDATE savings SET balance = balance + 200.00 WHERE customer_id = 10233276;
9  -- 5: Commit the transaction
10 COMMIT;
```

If someone reads the data after command #3 happens, the savings and total values will be wrong

Tx	balance	savings	Total
1	1000	1000	2000
2	1000	1000	2000
3	800	1000	1800
4	800	1200	2000
5	800	1200	2000

Dirty read

Figuur 28: Dirty read example: every command Tx is a row

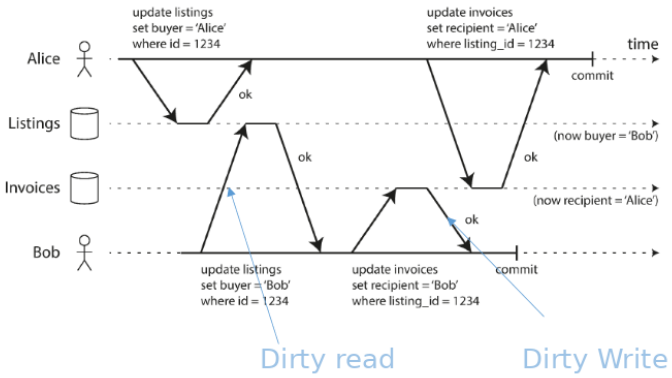
Solutions:

- Read locks (=very bad performance)

- Remember the old value until commit
  - Until the commit happens, every value will be what it was at Tx = 1

3.5.2 Dirty Writes

Definitie 3.2 (Dirty Write) *A dirty write happens when a transaction writes data that has been changed on disk by another transaction. The last transaction will overwrite what the first transaction wrote.*



Figuur 29: Dirty write example

1. Alice buys a car from a dealership
2. Bob buys the same car from the dealership
3. Bob gets an invoice before Alice because his internet is faster
4. Alice gets an invoice after Bob. Two people now own the same car?

Solution: Write lock:

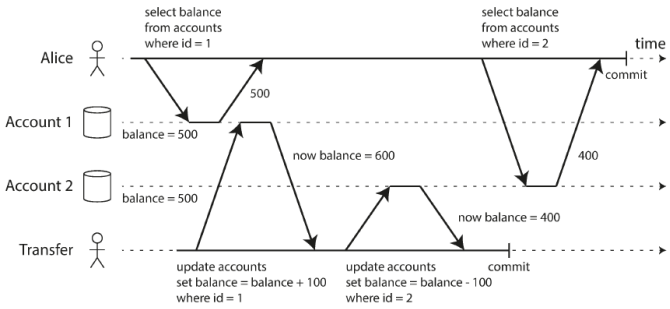
- If a row is claimed by a transaction, that row should be locked until commit
- Bob cannot write to the invoices, because it has been locked by Alice.

3.5.3 Read skew

Definitie 3.3 (Read skew) *Read skew happens when a commit reads the same data twice, with different results because another transaction updated the data.*

1. Alice checks the balance of the first account
2. Bob updates the balance of the first account
3. Bob updates the balance of the second account
4. Alice checks the balance of the second account

Result: Alice 'loses' $100 in one commit, because another transaction changed data.

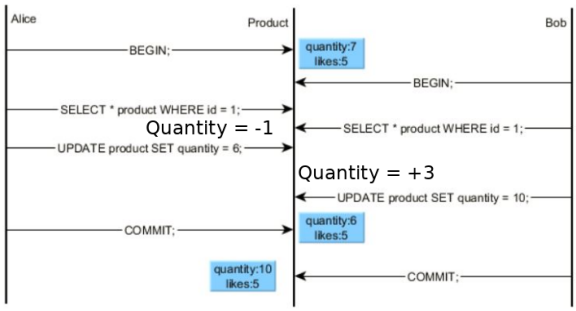


Figuur 30: Read skew example

Solution:

- Reading the values again solves the problem
- Except for backups: If a backup saves data while another transaction changes it, you will come across problems.

3.5.4 Lost updates & Atomic updates



Figuur 31: Lost updates: example

1. Alice checks the quantity of the product (quantity = 7)
2. Bob checks the quantity of the product (quantity = 7)
3. Alice buys the product (quantity = 6)
4. Bob thinks the quantity is 7 and he wants to add 3: he sets the quantity to 10 (7+3)
5. Alice commits her changes. According to her, the quantity should be 6
6. Bob commits his changes. The quantity is 10, overwriting Alice's changes. The actual quantity should be 9 (7-1+3)

Solution: Atomic updates

- Problem: Two read - modify - write transactions with different outcomes
- Repeatable read does not fix this
- Solutions: 'atomic updates' or manual lock

  - = Exclusive read lock on the data
  - = No reads or update object until commit
  - Update ‘X’ SET value = "X2"⇒ (Read - modify - write in one operation)

3.5.5 Write Skew

Atomic Updates don’t protect against everything:

- Multi object updates & lost updates

Pattern:

1. Read something
2. Make decision
3. Write new data
4. By the time the write is committed, the premise of the decision (step 2) is no longer true.

3.5.6 2-phase lock - Serial execution

With weak isolation levels:

- Readers never block writers
- Writers never block readers (you can read the old value while it is being overwritten)

With 2-phase lock, there are two fases (duh):

1. Exclusive read-lock on data
2. Exclusive write-lock on data

Problem: Deadlocks

- Transactions keep waiting on other transactions’ locks
- Result: the whole database can crash because of this

Examples that support 2-phase locking:

- MySQL InnoDB
- SQL server
- DB2 (but DB2 mistakenly calls this "Repeatable read")

3.6 Isolation levels

= Choose between strong isolation or strong performance

- Modern processing 8 - 100+ threads
- Choose an isolation level:
  - Read Uncommitted (weakest isolation, most performance)
  - Read Committed
  - Repeatable Read (=snapshot isolation)

  - Serial Execution (strongest isolation, least performance)
- Isolation problems are hard to debug:
  - It's a timing problem
  - Very hard to reproduce
  - No errors are logged

	Default	Max	Source
SQL Server	Read Committed	Serializable	https://docs.microsoft.com/en-us/sql/t-sql/statements/set-transaction-isolation-level-transact-sql?redirectedfrom=MSDN&view=sql-server-ver15
MySQL InnoDB	Repeatable read	Serializable	https://dev.mysql.com/doc/refman/5.7/en/innodb-transaction-isolation-levels.html
MySQL MyISAM	No Transactions!	No Transactions!	
Oracle	Read Committed	Snapshot Isolation ("Serializable" (*))	https://docs.oracle.com/cd/B14117_01/server.101/b10743/consist.htm#i17856
MongoDB/Cassandra	No Transactions!	No Transactions!	

Figuur 32: (default) isolation levels in current databases. (*) Wrong, Oracle does not comply with ANSI

3.6.1 Isolation level 1: Read Uncommited

- Read Uncommitted offers no protection against concurrency threats
- Fastest performance, lowest isolation
- One transaction may see not-yet-committed changes made by other transactions

3.6.2 Isolation level 2: Read Committed

Offers protection against:

- Dirty reads
- Dirty writes

Solutions

1. Read locks (bad performance)
2. Remember the old value until 'commit' (better performance)

3.6.3 Isolation level 3: Repeatable read or Snapshot Isolation

- Also called "Multi Version Concurrency Control (MVCC)"
- Solves dirty reads, dirty writes and read skew
- If a commit happens before everything is fully backed up, every commit started after the start of the backup will be ignored.
- To accomplish this, every transaction gets a number
- 'Readers do not block writes, writers do not block reads'

3.6.4 Isolation level 4: Serial execution

- One single fast thread (in RAM) for writing
- Multiple threads for reading
- Not very fast, definitely not very scalable
- Only use if your database is not too complex
  - Redis
  - VoltDB
  - Other databases that can be kept in memory...
- No write locks necessary, no overhead from thread synchronisation, ...
- Limited by a single thread on your CPU
- How to use multiple threads?
  - Partition data
  - Multiple threads, one thread per partition
  - Speed will be much slower if a transaction accesses multiple partitions
- Complete transaction in one serial stored procedure (=piece of code, already compiled and ready to execute)

3.6.5 Conclusion

- Isolation levels are a complex trade-off between...:
  - Consistency
  - Scalability
- Check your application: which level is the best for your usecase?
-

3.7 ACID: Durable

= A database should be durable: every write transaction has to be written to disk, and should be stored safely and reliably.

But durability is also a trade-off:

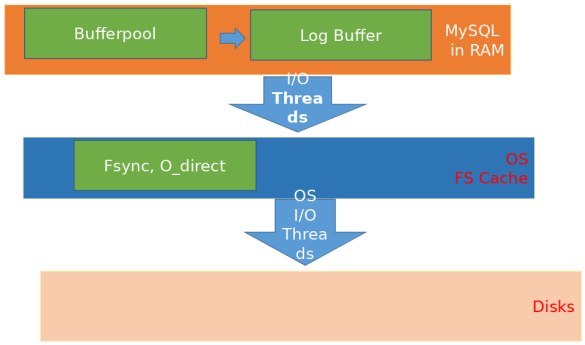
- Higher durability $\Leftrightarrow$ lower performance
- Higher performance $\Rightarrow$ more risks

3.7.1 Caching & BBU

- If we choose the highest isolation on sofware, the hardware can still fail.
- For a RAID configuration:
  - Most RAID configurations use a RAM cache before writing changes to disks
  - RAM cache should have a Battery Backup Unit (BBU) in case the power goes out

  - This is because RAM is by definition not durable, but volatile
- Disks also have RAM caches
  - This is mostly to sort the data before it gets written
  - Use professional storage: disks with capacitors!
  - RAM caches in SSDs have to be Non-Volatile (NV)!

3.7.2 The Transaction chain



Figuur 33: The transaction chain when writing to disk

Steps a transaction takes to write to disk:

1. The transaction gets buffered in the buffer pool
2. The data gets written to the log buffer
3. Using multiple I/O threads, the log buffer gets flushed to the OS
4. The OS chooses how the data is written (cache first, or write immediately)
5. The OS writes the data to the disks

3.7.3 The transaction chain: innodb_flush_log_at_trx_commit

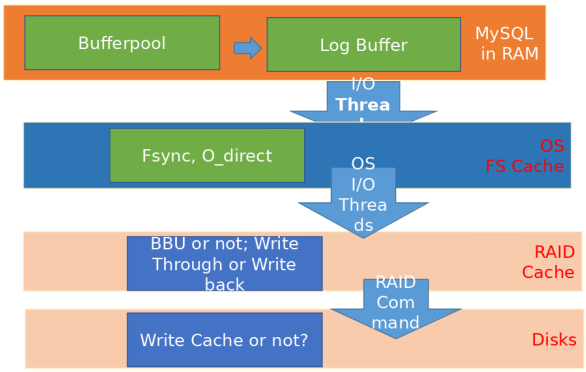
= a setting in MySQL InnoDB with three options:

- 0: Write the log buffer to the log file and flush the log file **every second**, but do nothing at transaction commits (fastest)
  - Fastest
- 1: Write the log buffer to the log file and flush it to durable storage **at transaction commits**
  - This is the only option that is fully ACID compliant
  - It is also the slowest
- 2: Write the log buffer to the log file **at every commit**, but flush it every second

3.7.4 innodb_flush_method

= a setting that tells the OS how the data has to be written

- fdatasync
  - InnoDB uses fsync() to flush both data and log files (unix)
- O_DIRECT
  - This setting still uses fsync() to flush the files to disk, but it instructs the operating system not to cache the data and not to use read-ahead. Avoids double buffering
- async_unbuffered
  - Default value on Windows
  - Causes InnoDB to use unbuffered I/O for most writes
  - Exception: it uses buffered I/O to the log files when innodb_flush_log_at_try_commit = 2



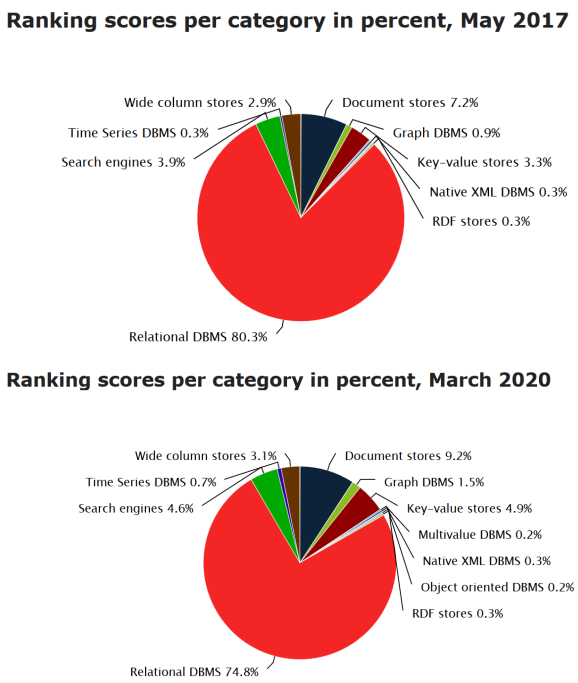
Figuur 34: The full transaction chain, with RAID

4 NoSQL

4.1 SQL

4.1.1 Possibilities:

- Relationeel
- Column store
- Document store
- Graph
- Key-value
- Specialisten:
  - Time series
  - Text search



Figuur 35: Popularity: Relational DBs are the most popular

4.1.2 Imperative languages vs Declarative languages

```
Public class TokenizerMapper extends Mapper<Object, Text, Tekst>{
    private final static IntWriteable one = newIntWriteable(1);

    public void map(Object key, Text alue, Context context) throws
            IOException, InterruptedException {
        StringTokenizer itr = new StringTokenizer(value.toString());
        while(itr.hasMoreTokens()) {
            word.set(itr.nextToken()); context.write(word, one); }}
    ....Much more code
```

```
SELECT word, count(*) FROM lines
    LATERAL VIEW explode(split(text, ' ') Table as
words
    GROUP BY word;
```

Figuur 36: Imperative (left) vs Declarative languages (right)

- Imperative: tell the system how to retrieve/handle/mutate the data, in what order
  - C#, python, . . .
- Tell the system the structure of the data you're looking for. Don't tell the system how it has to happen, the query optimizer.
  - SQL, HTML + CSS

4.2 Index

Index of a book:

- Summary/copy that allows to search faster in the main structure (book/database)
- Redundant (copy), needs disk space (needs "pages")

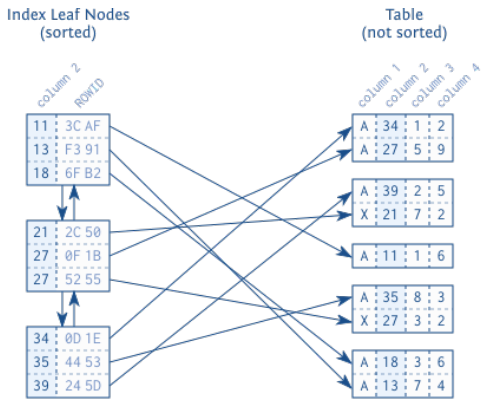
Index of a database:

Definitie 4.1 (Index) *A copy of some columns from a table, sorted, that improves the speed of data retrievel operations at the cost of additional writes and storage space to*

TODO

`https://use-the-index-luke.com/sql/anatomy`

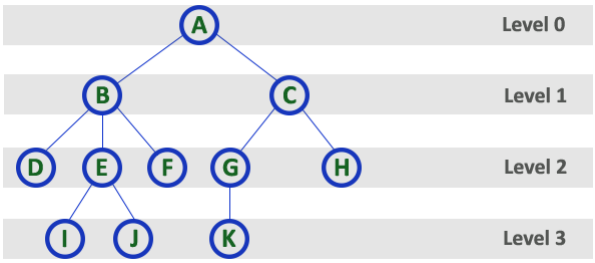
4.2.1 B-tree index



Figuur 37

- Leaf nodes = a double linked list
  - Index + Row ID
  - or index + value (in 'key value' DBs)
  - Every block refers to other blocks: the next and previous block
  - insert = add new links to the list
- This index is stored in RAM:
  - every block refers to another block
  - If you have to jump to another block, sequential disks are too slow for random access
  - random read/writes are much faster with RAM
  - If you have 10 million records: serial search is way too slow!

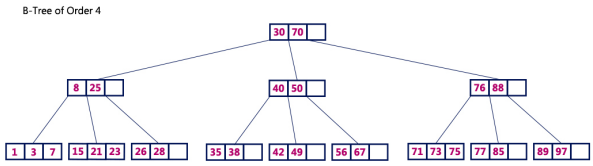
4.2.2 Tree architecture



Figuur 38: A balanced tree

- A = root
- B & C = child (pages)
- AC = edge
- Depth A to K = 3 (=amount of edges)
  - From root to leaf
- I, J, K = leaf nodes (point to data)

4.2.3 Searching for an index



Figuur 39: Balanced tree example

Find 38:

1. Search from node 30
2. Find subnode 40
3. Search from node 35 (depth 3), now read serially
4. Find row id in index 38

4.2.4 Size

- Block size = 16KB
- Branching factor = 100 (=amount of nodes in a page)
- Depth = 3
- $100 \cdot 100 \cdot 100 \cdot 16\text{KB} = 16\text{GB}$

  - Typical depth: 4
  - Typical branches = 100s

4.2.5 B-trees: getting faster & more reliable

- WAL (Write Ahead Log) or REDO log = sequential database
  - append only
  - update are critical moments
- 1 update = two writes
  - WAL
  - Page update
- Less levels vs more branches:
  - Each level can be a disk seek
  - Using means less disk seeks

TODO

4.2.6 Python + Postgres

Python + Postgres can handle almost any analytics challenge

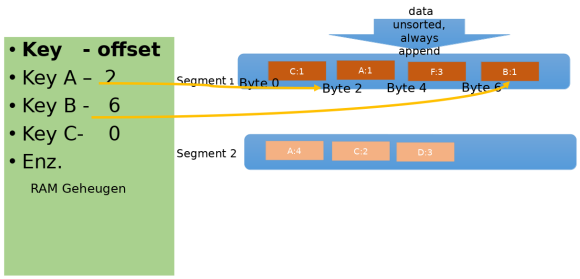
Figuur 40

4.2.7 When is SQL not the answer

- Volume = When you have petabytes of data (rare)
- Velocity = Too many writes per second
- Scalability
  - Want to avoid expensive servers
  - Want to avoid expensive SANs (Storage Area Network)
- Variety = When you don't want to turn an object (with unstructured data) into a relational row
  - = Object - relational database mismatch
  - `https://en.wikipedia.org/wiki/Object%E2%80%93relational_impedance_mismatch`

4.3 Key-Value

TODO: betere image

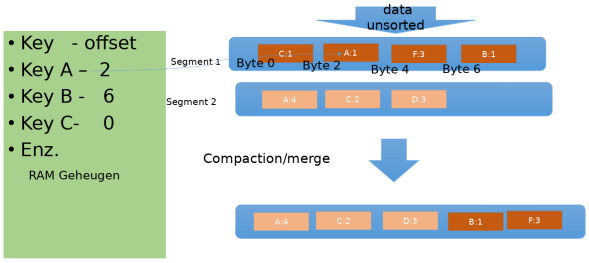


Figuur 41: Log structure + hash index

- TODO
-

4.3.1 Compacting

TODO: betere image



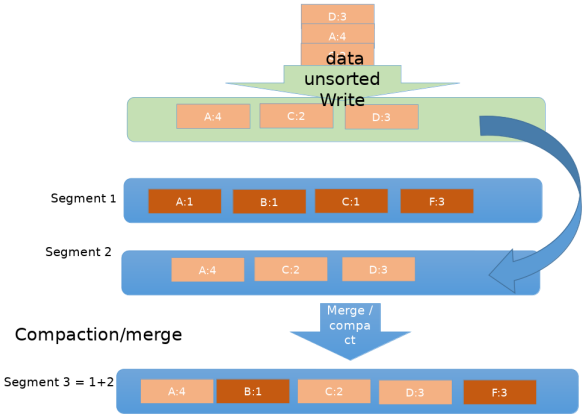
Figuur 42: Compaction/merge of 2 segments

- TODO
- Segment 2 is newer: we ignore the first segment's A and B

4.3.2 Principles

- Very fast writes (append only: you can write sequentially because disks don't need to change tracks)
- When crash: no corruption because of wrong update
- very fast reads if hash index is in RAM
  - if not: very slow
  - SELECT * FROM A TO ZZZ (ordered full scan == slooow)
- Example Riak Bitcask (`https://en.wikipedia.org/wiki/Riak`)

4.4 LSM: Log Structured Merge Tree

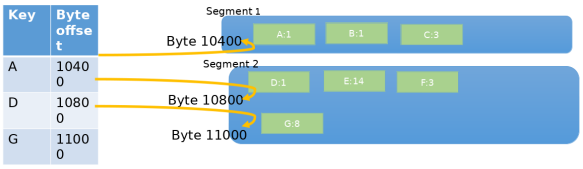


Figuur 43

TODO: betere image

- Sort in RAM (memtable)
  - Log segment file as backup
- Write to disk after several MB
  - To sorted string table file
- Merge, compact and sort
- TODO

4.4.1 Log Structure Merge + Sparse tree index



Figuur 44: LSM + Sparse tree index

- We can simplify the LSM
- Sparse tree index = remember where some milestones are
  - If you need F, and you know where D and G are
  - Start reading from D (byte-offset 10800)
  - Read sequentially until G (byte-offset 11000)
  - This sequential read is very quick, because
- Merge, compact & sort every time = string sorted table

- Every delete: create new segment and merge. ‘Tombstone the old segment’ = marking key/value pairs for deletion

4.4.2 Applications of Sorted String & LSM-tree

- First application: Google Big Table (`https://en.wikipedia.org/wiki/Bigtable`)
- LevelDB, Rocks-DB
- Hbase, Cassandra (Facebook)
- ElasticSearch: Lucene text search (key = text, value = document) or inverted index

4.4.3 Advantages LSM

- Very fast writes
- Quite fast reads (sparse index)
- Easily scaled over nodes (segments)

4.4.4 Disadvantages LSM

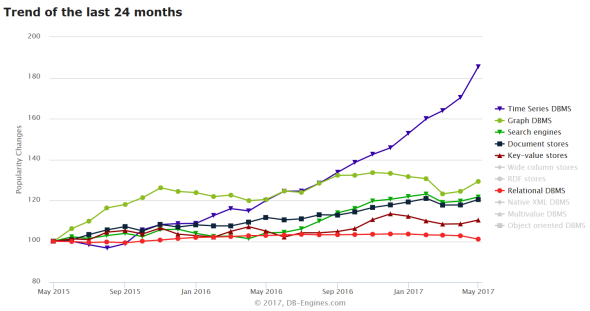
- Write (merge & delete) in background can influence speeds
- Deletes are costly!
- You have to specify the rate between compaction & write/read (ongoing action) yourself

4.4.5 Summary: B-Trees vs LSM trees

- TODO

4.5 Time Series

= LSM with a twist



Figuur 45: Popularity

4.5.1 Properties

- Lots of individual data points: ‘a row is not important’
- High write throughput

- High read throughput (aggregation per hour/day)
- Large deletes (data expiration)
- Mostly an insert/append workload, very few updates

4.5.2 Use case: windmill sensors

Situation: a windmill has many sensors that produce data that needs to be logged

- Turbine sensor data needs to be stored every second
  - 30 sensor readings per second
  - > 300GB per windmill per year
- Both aggreation as realtime
- Issue: MySQL database read locks during queries, INSERT failed, causing data loss

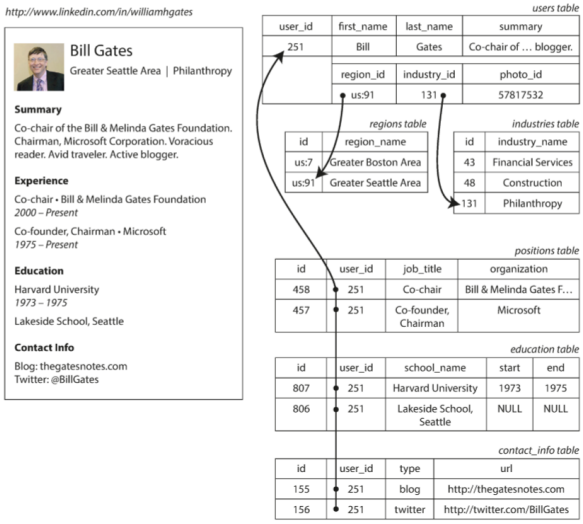
4.5.3 Case study: influx DB

`https://docs.influxdata.com/influxdb/v1.4/concepts/storage_engine/`

TODO: 'read the entire page and understand everything, and answer these questions:'

- Why important for MCT?
- How can you scale?
- TSM?
- WAL? How durable? What is it?
- What is a compactor? Compaction planning?
- Give one example of unique functionality typical for a time series environment
  - Tip: say you need to aggregate every hour

4.6 Object - relational mismatch



Figuur 46: Representing an object in relational rows

- TODO
- Better match for objects with unstructured data: JSON documents

4.6.1 PostgreSQL

- Mid 2014: PostgreSQL 9.4 natively supports JSON
- The speed to ingest documents as quickly as MongoDB, but ACID!
- Fully indexable

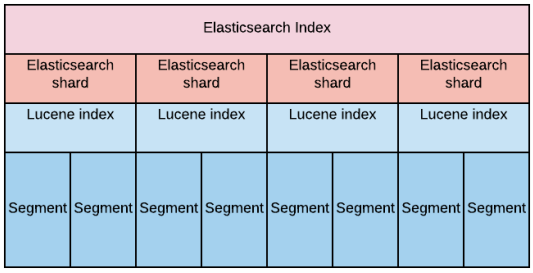
4.7 ElasticSearch

- Document Store
  - So naturally suited for describing objects
  - JSON serialized
- Easy access to an advanced fulltext search-engine library
- Lucene is very complex but very advanced
- Automatized sharding (and thus scalable) in containers
- RESTful API
- Slower data ingest (‘index’)

4.7.1 Elastic Search architecture

- Document = JSON data
- Index = a collection of documents

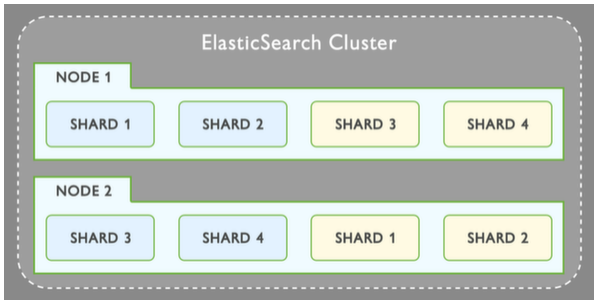
- Shards = scalable pieces of index
- Segments = sequential pieces of a shard



Figuur 47

4.7.2 Elastic Search Cluster

- Index is split over shards - nodes: scalability
- Shards can be replicated over nodes: availability



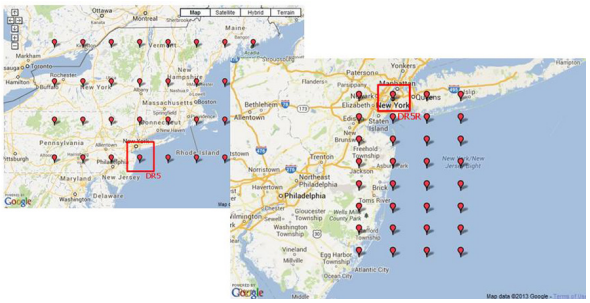
Figuur 48

4.7.3 Inverted index

- TODO
- = 'Lucene Index'

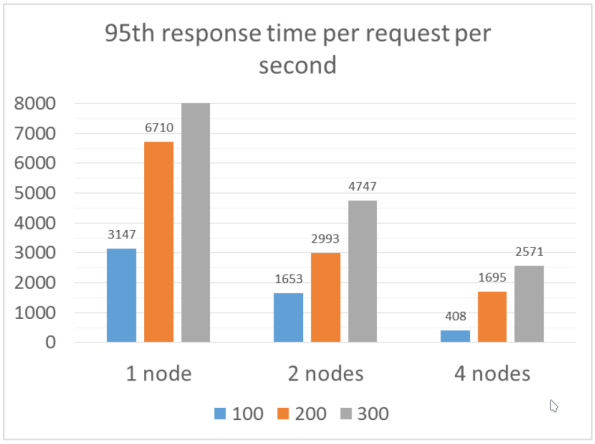
4.7.4 GeoHashes: Representing Geospatial data in ElasticSearch

- Since ElasticSearch 0.90
- Base32 encoded strings, interleaving the latitude and longitude
- Max resolution: 40mm * 20mm
- Each extra symbol divides the grid in 26 cells
- Use ElasticSearch's text search capabilities



Figuur 49

4.7.5 ElasticSearch scaling



Figuur 50

4.8 Which storage engine is the best and the worst

Which storage engine is the best/worst for the following situations:

- High amount of writes every second (sensor data)
- Hoog aantal updates iedere seconde
  - Web analyse data (Marketing campagne)
- Constante updates en reads van persoonlijke data?
- Full scans op gestructureerde data?
- ACID compliant OLTP?

TODO: overzicht van slides 5 en 6 van H4

5 Distributed Stores

= perfect match for big data (Volume, Velocity, Variety)

5.1 Terminology

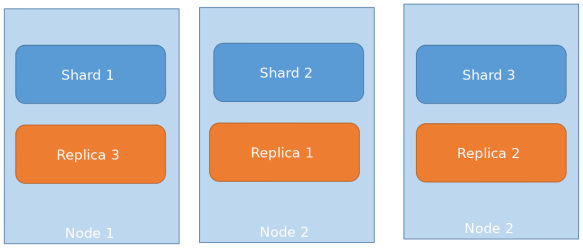
5.1.1 Shard/partition

= A shard or partition is a small subset of the database that can be assigned to a node

- ElasticSearch, MongoDB, MySQL: "Shard"
- Hbase: "Region"
- Cassandra, Riak: "vnode"

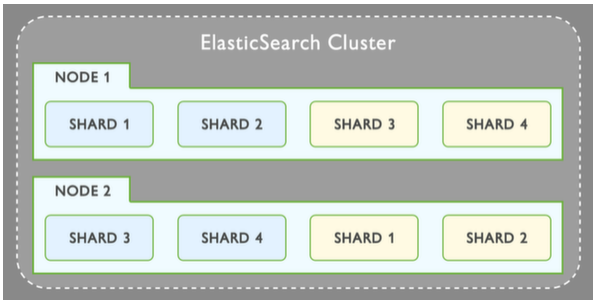
5.1.2 Replica

= A copy of a shard/database that is kept on a different machine



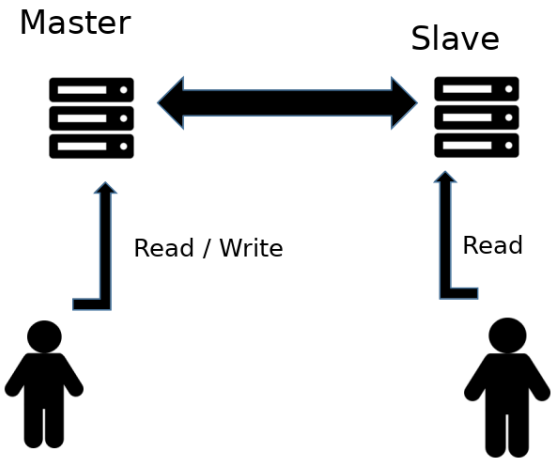
Figuur 51: Partitions/Replicas/Nodes

5.2 Elastic Search Cluster



- Index is split over shards - nodes: **Scalability**
- Shards can be replicated over nodes: **Availability**

5.3 Replication: master/slave or leader/follower



Figuur 52

Synchronously

- Acknowledgment of slave
- Consistent data
- Slow & unreliable with many slaves

Asynchronously

- No acknowledgment of slave
- Inconsistent data
- Fast, even with many slaves

5.3.1 Consistency choices

Strong Consistency

- At a certain point in time after a write, all replicas return the same update record/document (almost realtime)
- consistent with order in which write operations are submitted by clients

Eventual Consistency

- Faster, easier to be ävailable"
- Hard for devs: update a row - not sure when it will be updated in other nodes

High availability

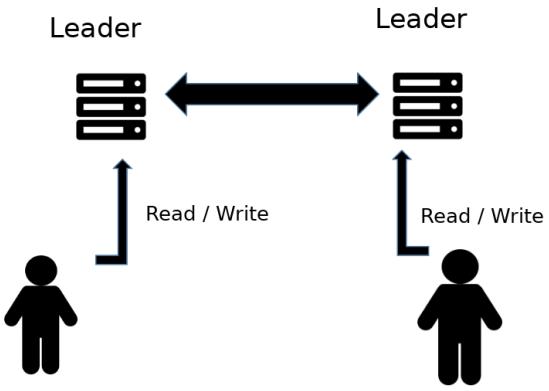
- Every request received by a non-failing node must result in a "non-error"response
- Nodes can read/write - even if that means that not all replicas have the same content

5.3.2 Amount of leaders

Single leader

- Only one node can write
- Strong consistency is possible

Multi leader



- More than one node can write
- $\Rightarrow$ write conflicts are possible
- Not consistent = replica may be out of sync

5.3.3 Split brain or Network partition

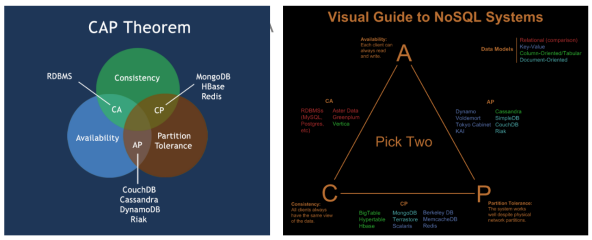
What if the network connection drops?

- Do we allow the follower to become the leader?
- Do we block from writing to follower?
- But leader might still be alive and serving...

5.4 CAP Theoreum

3 mogelijke eigenschappen:

- **Consistency**: every (later) read operation will always return the last version (*) of the data, or an error message (single object)
  - (*): last written by a write operation X older than this read operation Y
- Availability: there is always a server available, if necessary: with older data
- PartitionTolerance: the system must keep working, even if de nodes can't communicate with each other ("network partitioning")



Figuur 53: Oversimplification of CAP theorem

5.4.1 CAP: either consistent or available when partitioned

- When there are problems (partitiontolerance) with the network, we will have to choose:
  - Consistency
  - Availability
- Both at the same time is NOT possible with Partition Tolerance
  - Network issues: (group of) nodes can't communicate
- When there is no network partitioning, then both Consistency and Availability are satisfied

5.4.2 Relational DB: CA

- Only consistent if perfectly synchronized
- In reality, only allows read to slave (so no "A"for writes)

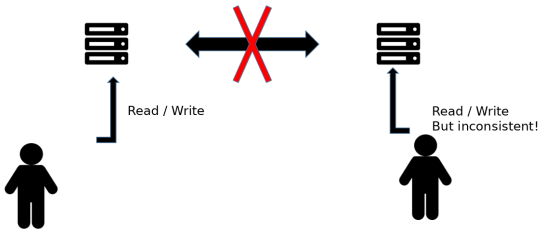
5.4.3 Relational DB: CA: single node (no P possible)

TODO

5.4.4 AP

Availability and PartitionTolerance (AP):

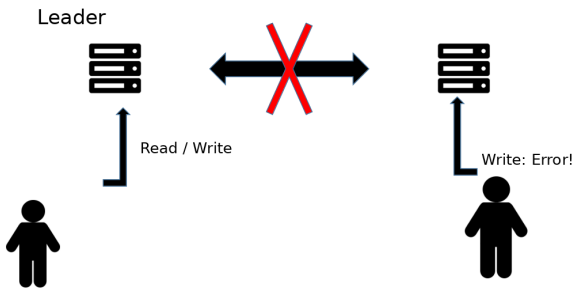
- The system will always return data, but not always consistent (with network issues between nodes)
- If we read or write data and this fails, another node will take control, no error messages
- We choose AP in situations where reading data is more important than writing



Figuur 54: AP: write once, read many times

5.4.5 CP

- Consistency and PartitionTolerance: The system will always return the correct data
- We will not be able to read the data as long as these were not written to all nodes
- If we can't write data, we will see an error message ⇒ no availability
- We can use CP if data consistency is very important



Figuur 55

5.4.6 Conclusion 1

CAP theorem is not a good architectural consideration

- CA doesn't exist: > 1 node & partitioning: no real availabilty anymore (only for read)
- Consistency is single object! (so less transaction isolation)
- No CP: with many replicas, you choose for asynchronous replication
  - So no strong consistency
  - No availability (no writes)

5.4.7 Conclusion 2

CAP theorem only serves to clarify the difference between AP and Single node consistency

- When you use multiple nodes, there is only ëventual consistencyänd "Read availability"⇒ CP is impossible
- Multi node systems are always a little inconsistent
- AP is possible but only for partition problems
  - Is being available in 1 situation (network partition, quite rare) even that useful

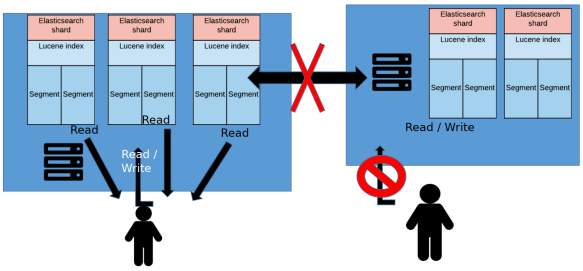
5.5 Distributed system: Elastic Search

5.5.1 Consistency and Network partitioning

`https://www.elastic.co/guide/en/elasticsearch/reference/2.4/docs-index_.html#index-consistency`

- To prevent writes from taking place on the "wrong"side of a network partition
- By default, index operations only succeed if a quorum(=voting procedure) (>replicas/2+1) of active shards are available.

- This default can be overridden on a node-by-node basis using the action.write_consistency setting.
- To alter this behavior per-operation, the consistency request parameter can be used.
- Valid write consistency values are one, quorum, and all.



Figuur 56: TODO

5.5.2 Architectural considerations

Do you need transactions? (Multi-object transactions?)

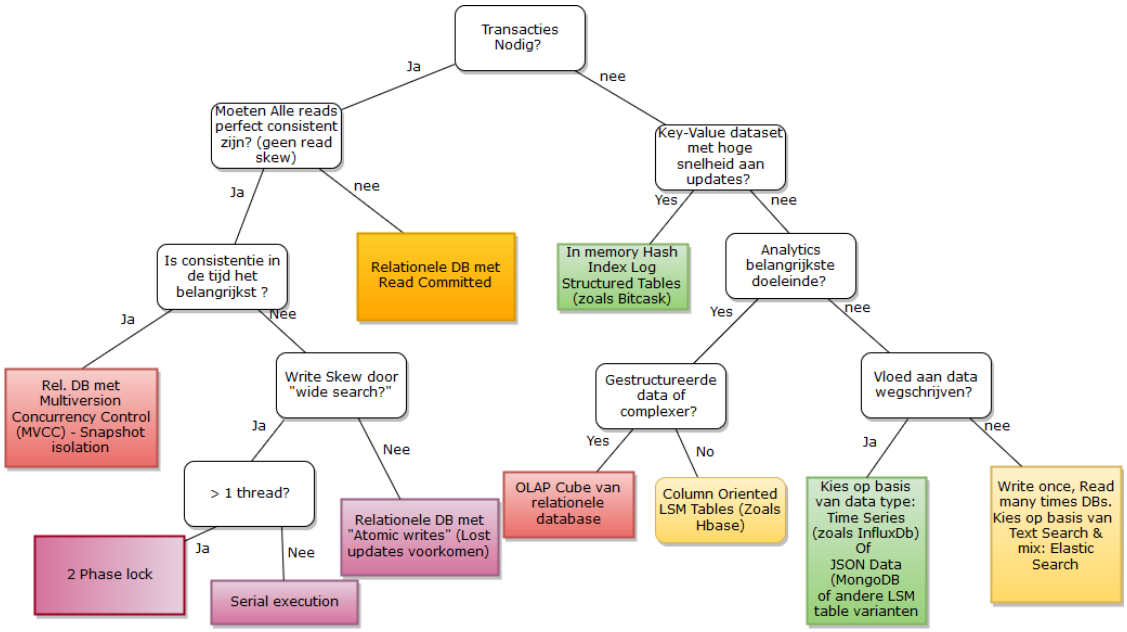
- If very important: pursue high isolation (atomic)
- How important is availability: choose reliable (expensive) hardware
- How important is performance: very expensive hardware for high load (limited scalability)
- If not very important: Read committed + hardware dependant on load
- If availability very important: fast network (synchronized!) and expensive reliable hardware + few (1) replicas

How important is scalability and performance:

- Very high load = low consistency & limited availability
- Rather choose for databases that are easily sharded (so no transactional DBs)
- Availability: high amount of replicas + "normal hardware"⇒ lower consistency

	Garanties	Hoe?	Nadeel
Read committed	Geen dirty writes	Row-level locks (or table locks)	Inconsistente data: Read Skew
	Geen dirty reads	Oude waarde weergeven zolang transactie bezig is	Lost updates : 2 updates tegelijkertijd
snapshot isolation	+Geen read skew	+multiple versions of a row (or object)	Lost updates : 2 updates tegelijkertijd
		Resultaten van latere transacties zijn niet zichtbaar voor vroegere	
Atomic updates	Geen lost updates	Exclusive lock - geen Read tijdens transactie	"write skew"
Serial execution	"perfect"	Single threaded execution	Traag naarmate DB groter wordt - Single Thread is vooral interessant voor "perfect" gepartitioneerde DBs. Beperkte complexiteit van data (bvb. eenvoudige Key-value)
Two Phase Lock	"perfect"	Multi threaded, goed voor ruime "checks"	Werkt beter voor "Complexe data", nog altijd zeer traag. Blokkeert potentieel hele veel data

Figuur 57: Isolation overview



Figuur 58: Simplified overview data store choices