

Datacaptatie & visualisatie

Xander Vandooren

December 2, 2024

Contents

1 Sessie 1:	1
1.1 CSV files:	1
1.1.1 Key characteristics:	1
1.1.1.1 Advantages and disadvantages:	1
2 Sessie 2:	2
3 sessie 3:	4
3.1 Timeseries database:	4
3.1.1 Specific concepts:	5
3.1.2 Query data with Flux (version >= 1.8)	5
3.1.3 Influxdb elements:	5
3.1.4 InfluxDB 3.0 vs open source	5
3.1.5 Line protocol	5

1 Sessie 1:

1.1 CSV files:

Comma separated values (CSV) is een text fike om tabellen data op te slaan, zoals spreadsheets of databases. alle data is afgezonderd via een comma.

1.1.1 Key characteristics:

- Structuur:
 - Each row of the file corresponds to a record.
 - Each column within the row is separated by a comma (,) by default.
 - The first row often contains headers that describe the columns.
- File extension: The typical file extension is .csv
- Delimiter: Although commas are the default delimiter, other characters like tabs (␣), semicolons (;), or pipes (|) can be used in some variations of CSV files.
- Plain text: CSV files are stored in plain text format, making them easy to create and read using basic text editors like Notepad or TextEdit.

1.1.1.1 Advantages and disadvantages:

- Advantages:
 - Simplicity: Easy to create, read, and parse.
 - Compatibility: Supported by most data processing tools.
 - Lightweight: Smaller in size compared to binary formats.
- disadvantages:
 - No Formatting: Does not support rich formatting like bold text or colors.

- Limited to Simple Data: Cannot handle complex data structures like nested or hierarchical data.

CSV files are commonly used for data exchange, importing/exporting data from databases, and performing data analysis.

2 Sessie 2:

Wat is docker?

- Docker is a container runtime
- Containers zijn een technologie die bepaalde processen isoleert. Het is alsof er maar 1 bepaald proces op een cleane installatie draait.
- Het verschil met een Virtuele Machine is dat een container de kernel van een operating systeem deelt.
- Door deze Lightweight opzet kan je makkelijk verschillende containers draaien op 1 enkel OS.
- Containers zijn sneller en efficient
- Docker is de default container runtime
- Docker is open source

Docker:

- Was eerst ontwikkeld op Linux
- Vanaf 2016 op andere OS
- Op windows via Hyper-V
- in 2019 WSL2 (Windows Subsystem Linux)
- In 2020 is docker wereldwijde keuze, omdat het alle implementaties verenigt in een CLI (Command line interface) en een Daemon. (background process)

Docker begrippen:

- Docker engine → applicatie die je installeert op de host machine to build, run and manage de Docker containers.
- Docker Daemon → Het werpaard van de Docker systeem, dit component luistert naar alle processen, verstuurt vragen en regelt alles tussen de images, containers en volumes.
- Docker client → De user interface om met de Docker Daemon te communiceren via de REST API.
- Docker image → een template, die bestaat uit een all-in-one package die nodig is om een container omgeving te kunnen maken.
- Docker container → Een draaiende instantie van een Docker image.
- Docker registry → catalog systeem voor het beheer van Docker images (lokaal of van een externe partij of Docker Hub)
- Container orchestration → tools om de containers te beheren, b.v. Kubernetes, Docker Swarm,...

Docker architecture:

- Advantages of using Containers:
 - Isolation
 - Portability
 - Multi-cloud platform support
 - Version control
 - Scalability and management
 - Easier Development
 - Easy and rapid development
 - Consistency → It always Works, everywhere

Docker build,Docker run,Docker volume

3 sessie 3:

3.1 Timeseries database:

- handle time-related data:
 - Such as IoT
 - Financial applications
 - Industrial applications
- Time is primary axis
- These datasets primarily have 3 things in common:
 - The data that arrives is almost always recorded as a new entry
→ almost append only
 - The data typically arrives in time order
 - Time is a primary axis (time intervals can be either regular or irregular)
- Data at scale
 - One flight → half a terabyte of data
 - A connected car → 300 TB of data a year
- Throw data away?
 - Cars are subject to recalls, safety investigations, lawsuits, and much more
- Huge amount of timeseries data
- fast ingestion rate
- RD → reindex
- TSDB are built for client plurality

3.1.1 Specific concepts:

- Data retention → dropping data after certain time
- Bucket = database + retention policy
- Organization → workspace for users

3.1.2 Query data with Flux (version >= 1.8)

Flux = functional datascripting. Not very similar to SQL → use graphical interface querybuilder → script editor

3.1.3 Influxdb elements:

- Tags are indexed
- Tags are additional information
- Fields are not indexed
- Fields are the value of the sensors

3.1.4 InfluxDB 3.0 vs open source

3.0 is betaalde versie maar kan wel meer rows wegschrijven/ophalen.

3.1.5 Line protocol

Whitespace (spatie) in line protocol determines how influxDB interprets the data point. The first unescaped space delimits the measurement and the tag set from the field set. The second unescaped space delimits the field set from the timestamp.

- Timestamp optional in nanoseconds
- Upload file to db