DuckDB …

Minimalist Data Engineering

Daniar Achakeev

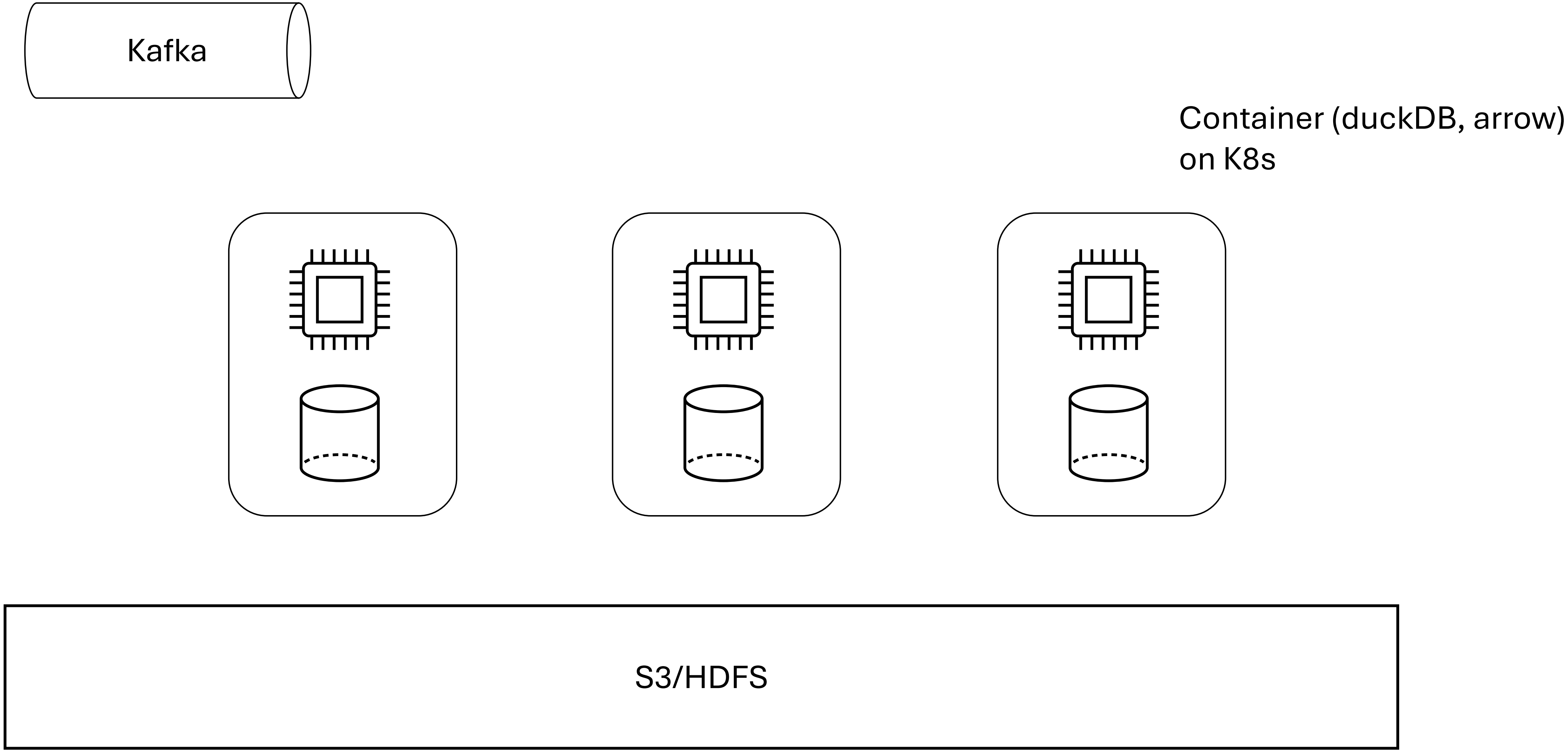
HMS Analytical Software GmbH

Software Engineer with focus on data and backend engineering

Introduction

- Client Story

- Migrating from Apache Spark to DuckDB and Apache Arrow



Kafka

Container (duckDB, arrow)
on K8s

S3/HDFS

DuckDB

- Pipeline
    - Pure functions
    - Fetch, Validate (Quarantine), Clean, Map, …
- SQL
- Unit tests

DuckDB

- CWI Amsterdam (Mark Raasveldt and Hannes Mühleisen)
- C++ (blazing fast)
- Vectorized OLAP Engine
- Single Process, single file database
- SQL
  - HOFs, Map/Array/Structs, CTEs (recursive), …
- CSV, Parquet, Json, …
- Arrow

DuckDB

- Out-of-core operators
- Excellent docs
- CLI
- WASM
- Extensions (FTS, Geo, …)
- Various filesystems
- ODBC, JDBC, python DB, Go, …
- UDFs
- …

Lakehouse formats

- Delta (without delta-rs, but with delta kernel)
  - https://duckdb.org/2024/06/10/delta.html

- Iceberg
  - https://duckdb.org/docs/extensions/iceberg.html

- Other use cases talk by Hannes Mühleisen :
- https://www.youtube.com/watch?v=NarcDUhHwQw