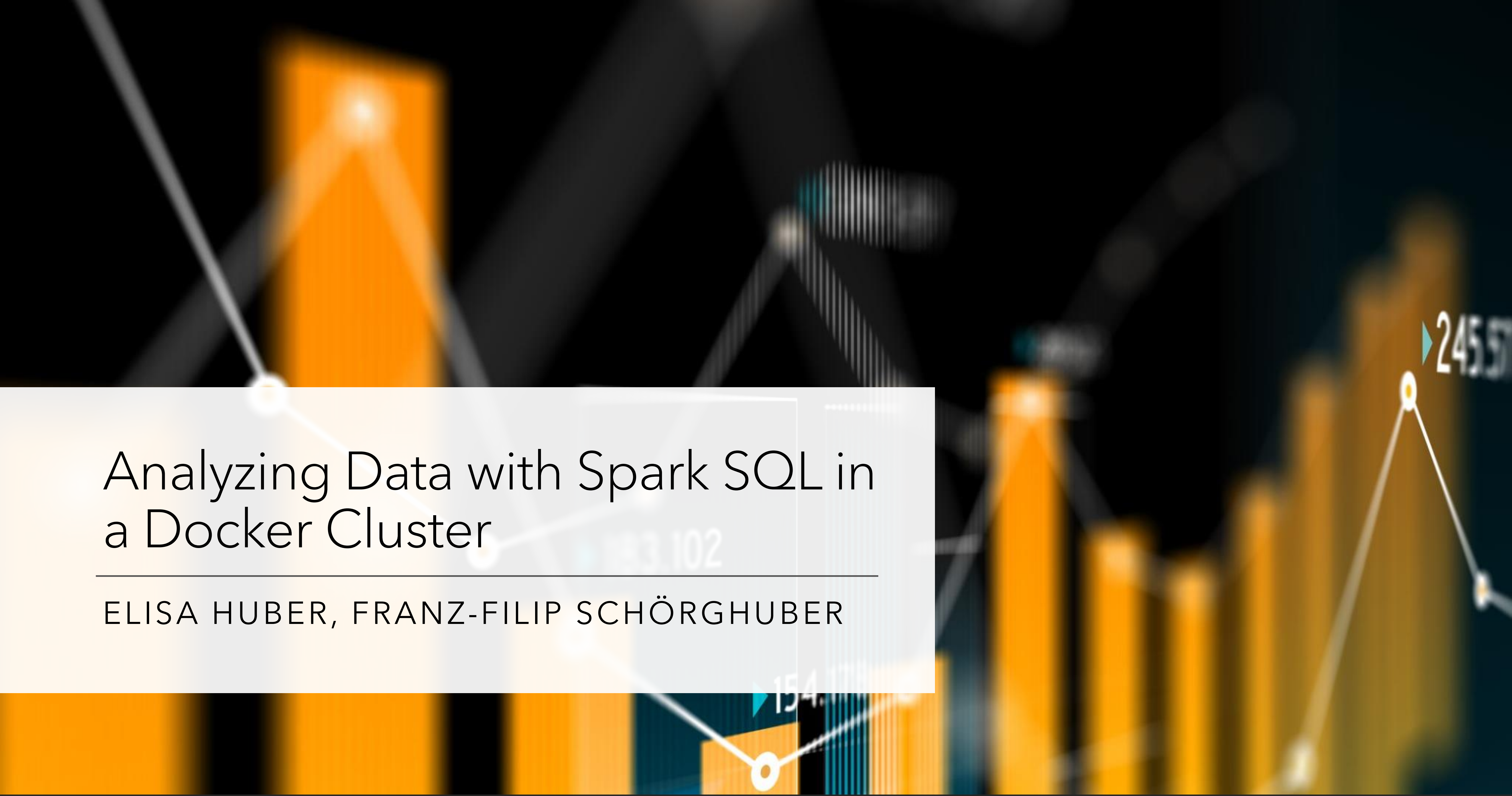


Analyzing Data with Spark SQL in a Docker Cluster

ELISA HUBER, FRANZ-FILIP SCHÖRGHUBER

Überblick

- Intro Spark
- Projektidee
- Bestandteile
- Vorgehensweise
- Demo

Apache Spark

Nachfolger von Hadoop

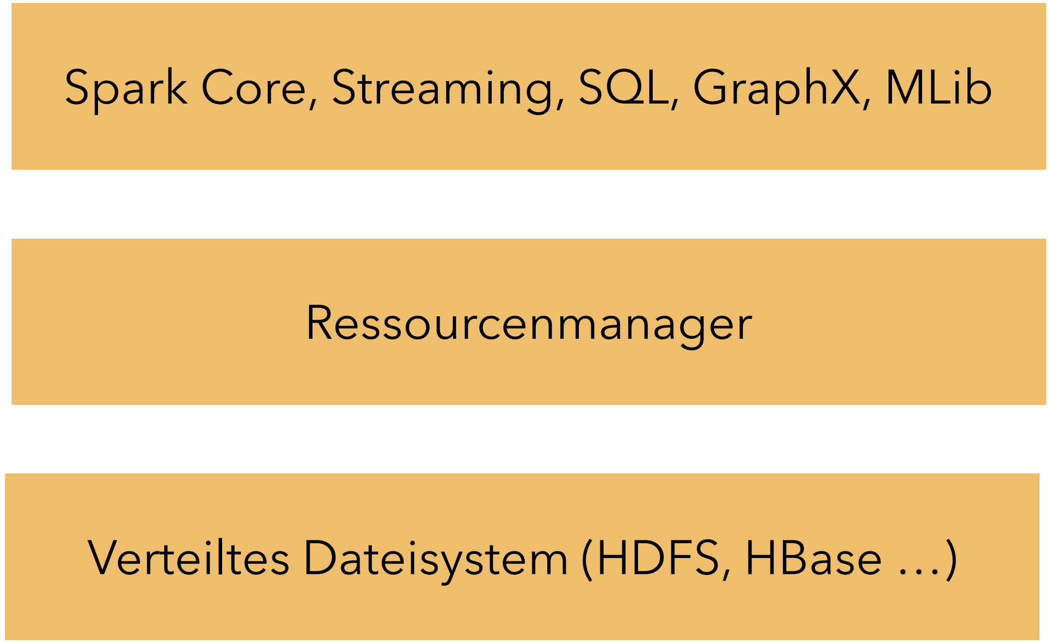
Schnelle Datenverarbeitung

Hohe Skalierbarkeit

Stapelverarbeitung

Nahe-Echtzeitverarbeitung (microbatches)

Architektur Spark



Spark Core, Streaming, SQL, GraphX, MLib

Ressourcenmanager

Verteiltes Dateisystem (HDFS, HBase …)

Datenstruktur Spark

RDD

- Verteilte Struktur mit beliebigen Objekten
- Kein Schema
- Transformationen und Aktionen

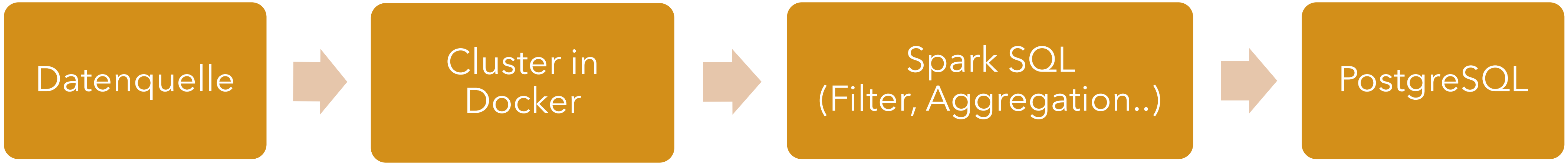
[401, 16.2, „OK“]
[412, 31.2, „OK“]
[1101, -123.2, „ERROR“]

DataFrame

- Teil von Spark SQL
- Baut auf RDD auf
- Mit Schema

ID: Integer	Temperature: Double	Status: String
401	16.2	OK
412	31.2	OK
1101	-123.2	ERROR

Projektidee



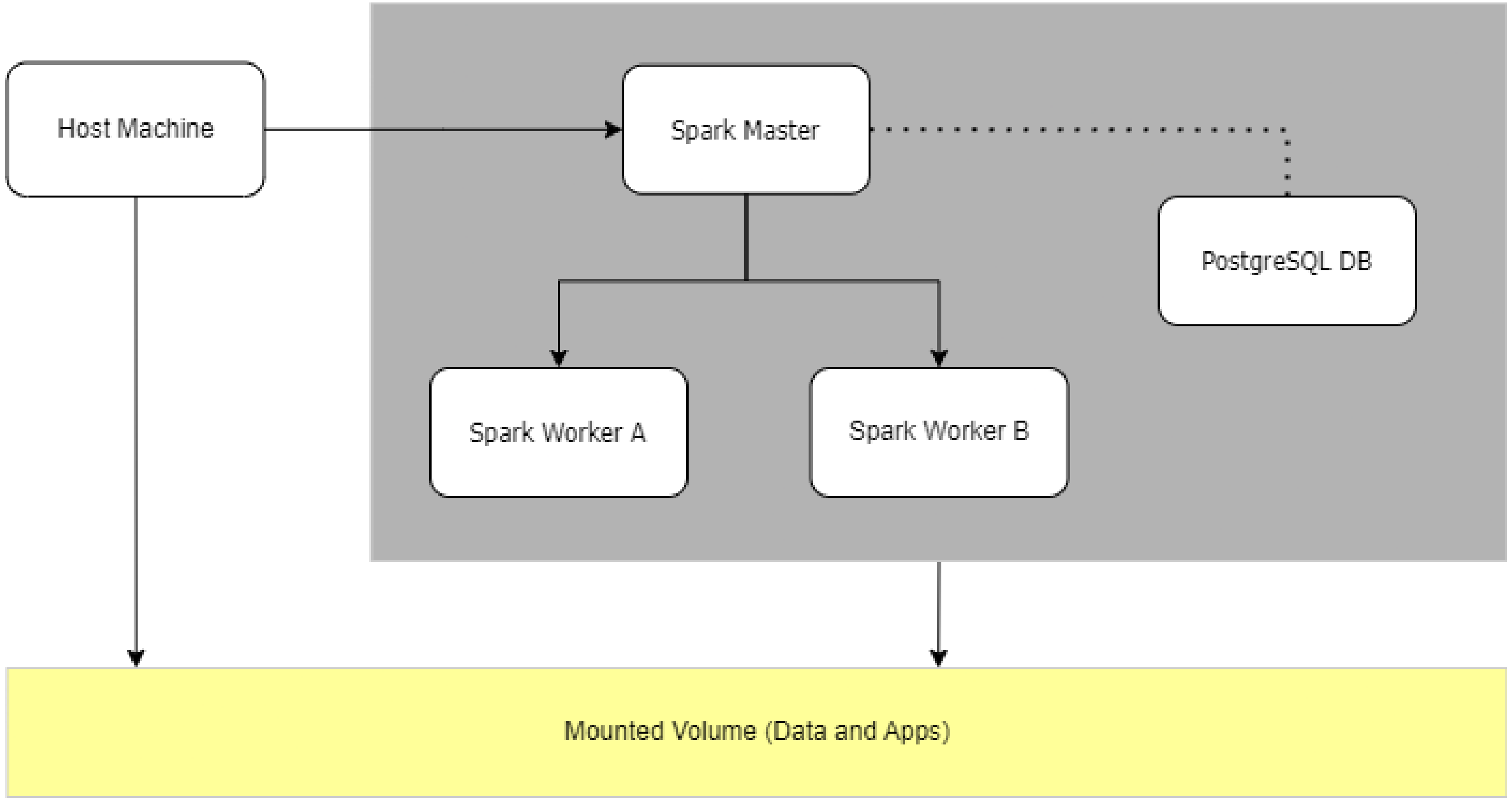
Datenquelle

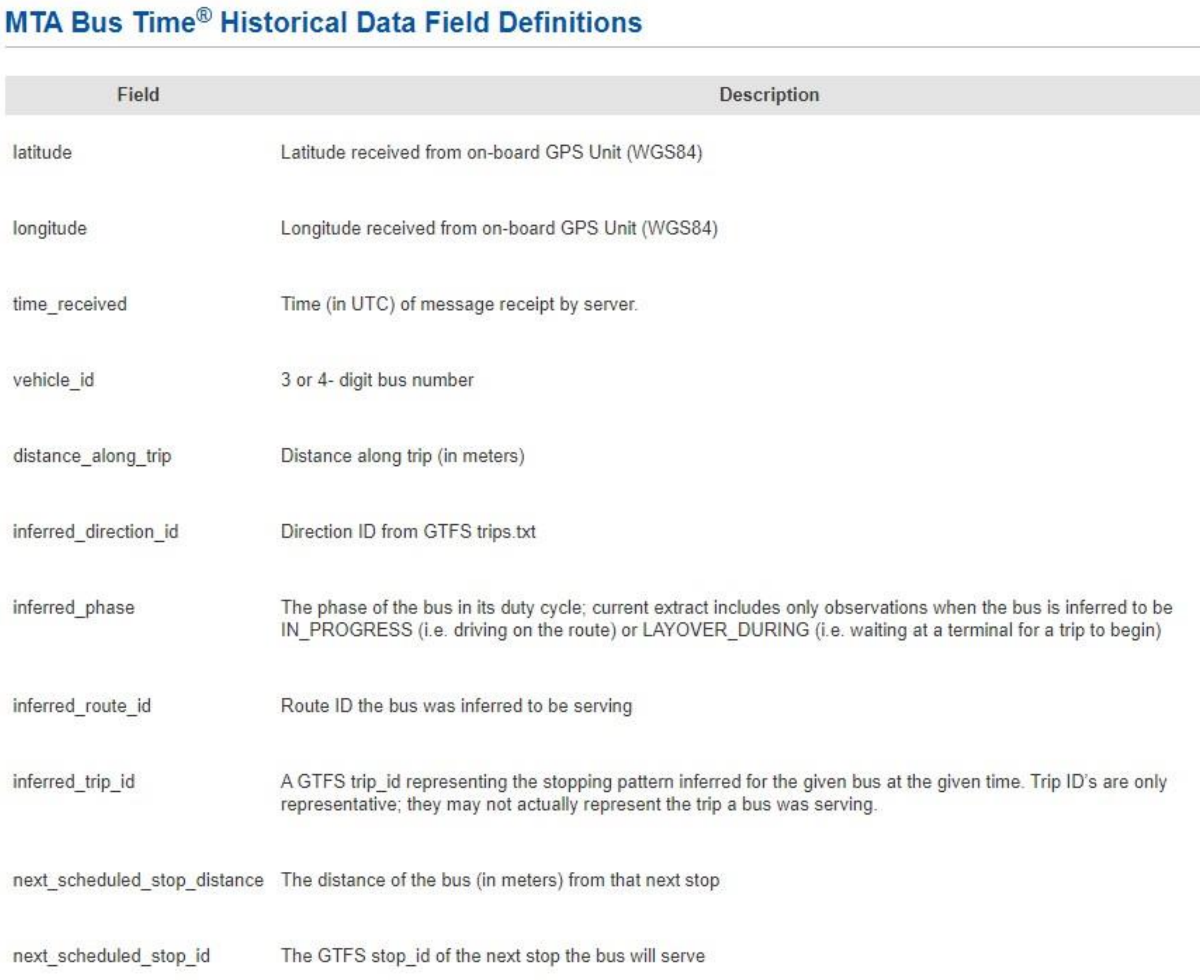
Cluster in Docker

Spark SQL
(Filter, Aggregation..)

PostgreSQL

Cluster


Host Machine
Spark Master
PostgreSQL DB
Spark Worker A
Spark Worker B
Mounted Volume (Data and Apps)

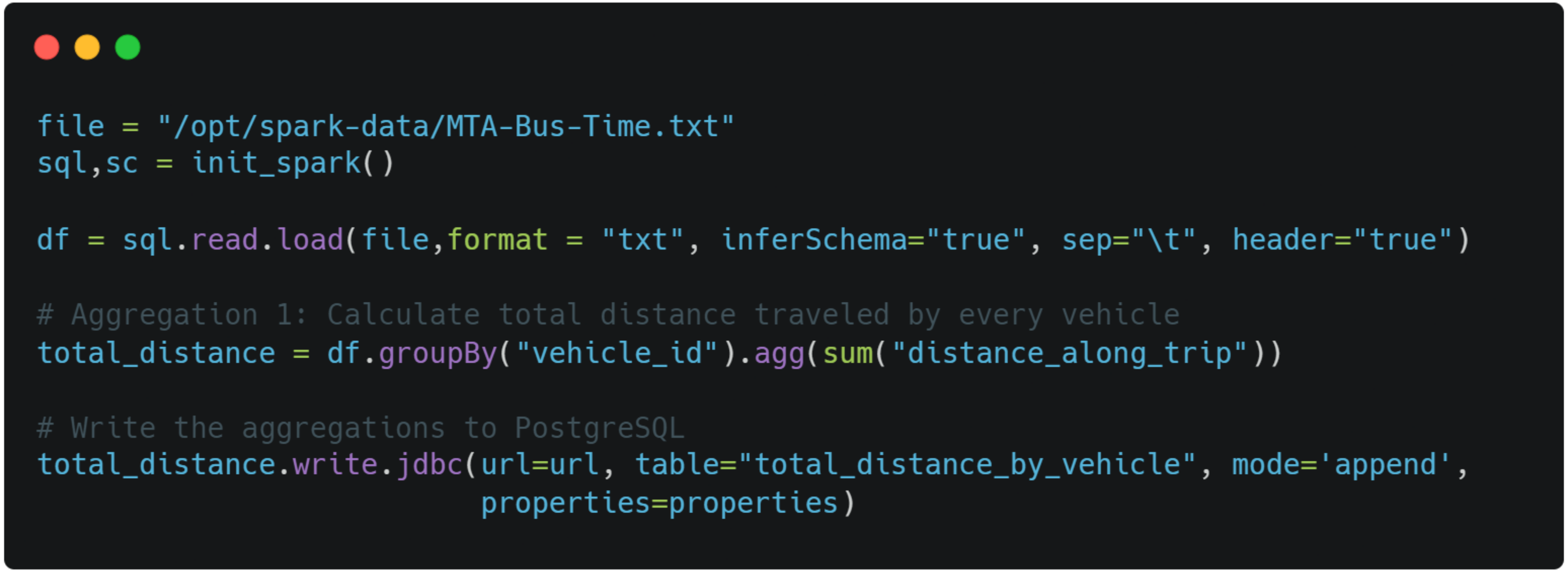


MTA Bus Time® Historical Data Field Definitions

Field	Description
latitude	Latitude received from on-board GPS Unit (WGS84)
longitude	Longitude received from on-board GPS Unit (WGS84)
time_received	Time (in UTC) of message receipt by server.
vehicle_id	3 or 4- digit bus number
distance_along_trip	Distance along trip (in meters)
inferred_direction_id	Direction ID from GTFS trips.txt
inferred_phase	The phase of the bus in its duty cycle; current extract includes only observations when the bus is inferred to be IN_PROGRESS (i.e. driving on the route) or LAYOVER_DURING (i.e. waiting at a terminal for a trip to begin)
inferred_route_id	Route ID the bus was inferred to be serving
inferred_trip_id	A GTFS trip_id representing the stopping pattern inferred for the given bus at the given time. Trip ID's are only representative; they may not actually represent the trip a bus was serving.
next_scheduled_stop_distance	The distance of the bus (in meters) from that next stop
next_scheduled_stop_id	The GTFS stop_id of the next stop the bus will serve

Spark SQL

unterstützt relationale Daten, komplexe Datenstrukturen wie JSON und Parquet



```
file = "/opt/spark-data/MTA-Bus-Time.txt"
sql,sc = init_spark()

df = sql.read.load(file,format = "txt", inferSchema="true", sep="\t", header="true")

# Aggregation 1: Calculate total distance traveled by every vehicle
total_distance = df.groupBy("vehicle_id").agg(sum("distance_along_trip"))

# Write the aggregations to PostgreSQL
total_distance.write.jdbc(url=url, table="total_distance_by_vehicle", mode='append',
                          properties=properties)
```

Demo