FOSDEM 2024: Ingesting and analyzing millions of events per second in real-time using open source tools

README.md

FOSDEM 2024: Ingesting and analyzing millions of events per second in real-time using open source tools

JAVIER RAMIREZ

@supercoco9 / @j@chaos.social

Data Advocate at

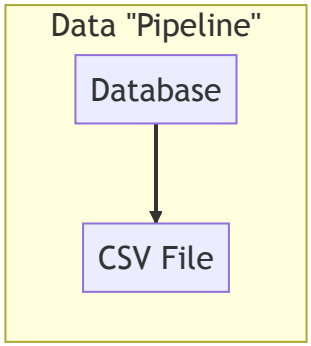
QuestDB

https://github.com/questdb/time-series-streaming-analytics-template

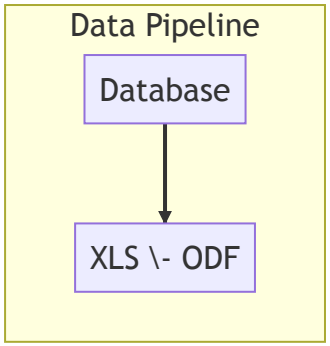

How you
can benefit
from using
Redis
FOSDEM '14
Brussels 1 & 2 February 2014
www.fosdem.org »
teowaki
javier ramirez
@supercoco9

- Apache Cassandra 2008
- Redis 2009
- Apache Kafka 2011
- Storm 2011 (released as Apache Storm in 2014)
- Apache Spark Streaming 2013
- InfluxDB 2013
- Amazon Kinesis November 2013
- Apache Flink 1st stable release 2014
- Grafana 2014
- NFSdb 2014 (rebranded as QuestDB in 2016)
- Google Cloud Dataflow 2015
- Clickhouse 2016

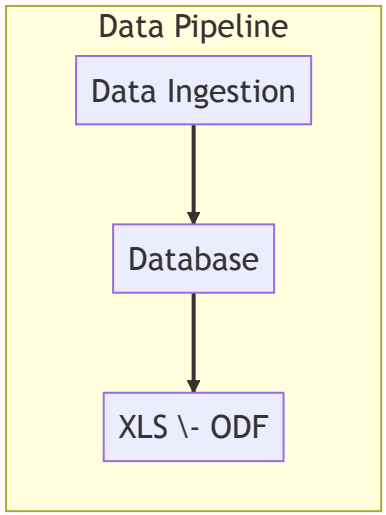
Loading a CSV every few seconds (Micro-batching)



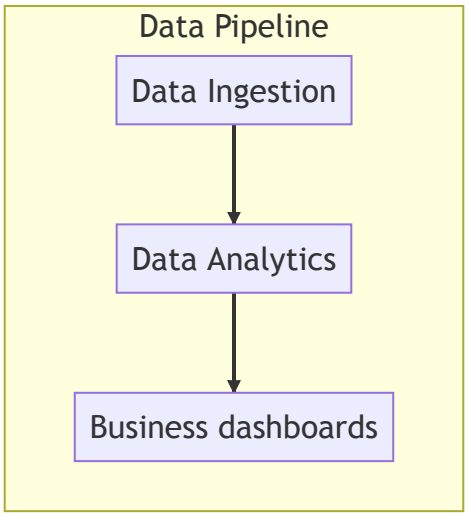
CSV Micro-batching. Look Ma! A Piechart!



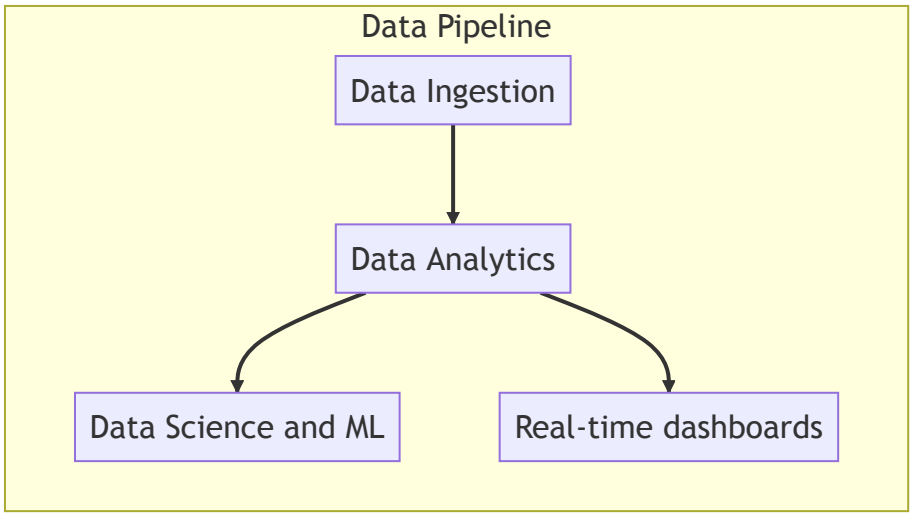
Micro-batching, but with a decoupled ingestion buffer



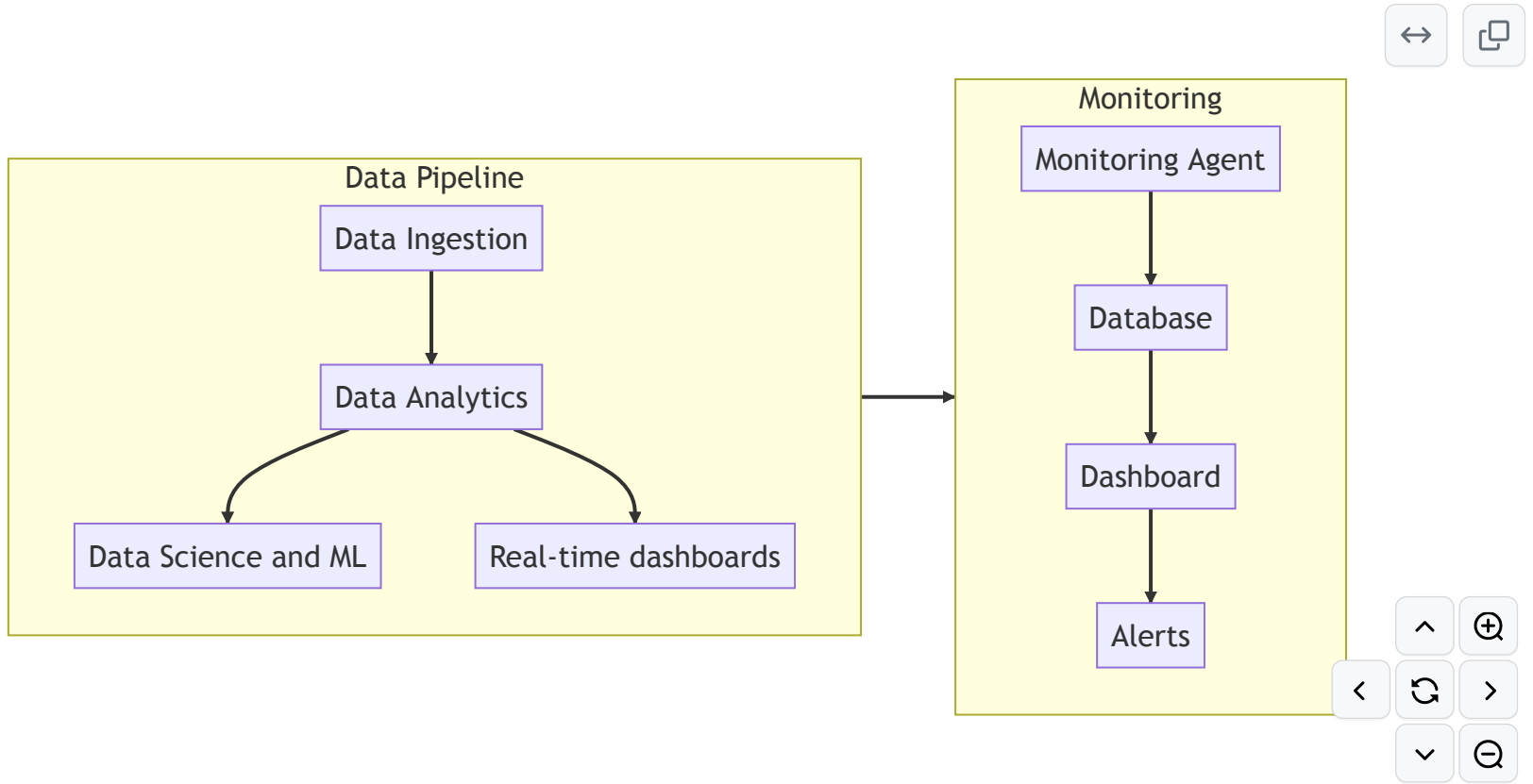
Bye micro-batches. Hello richer dashboards



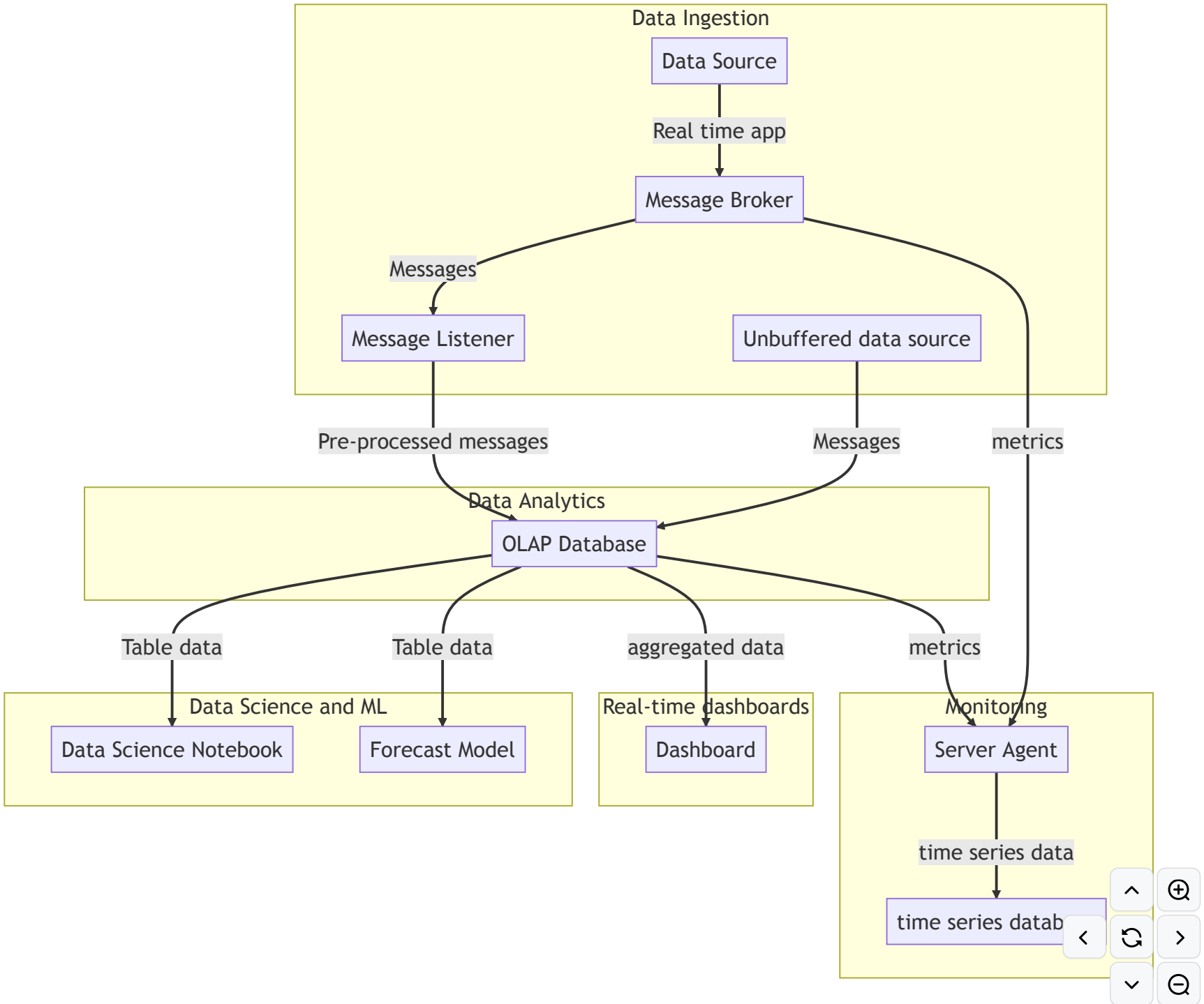
Auto-refresh, no more F5. Predict the future!



It worked on my machine. End-to-end monitoring

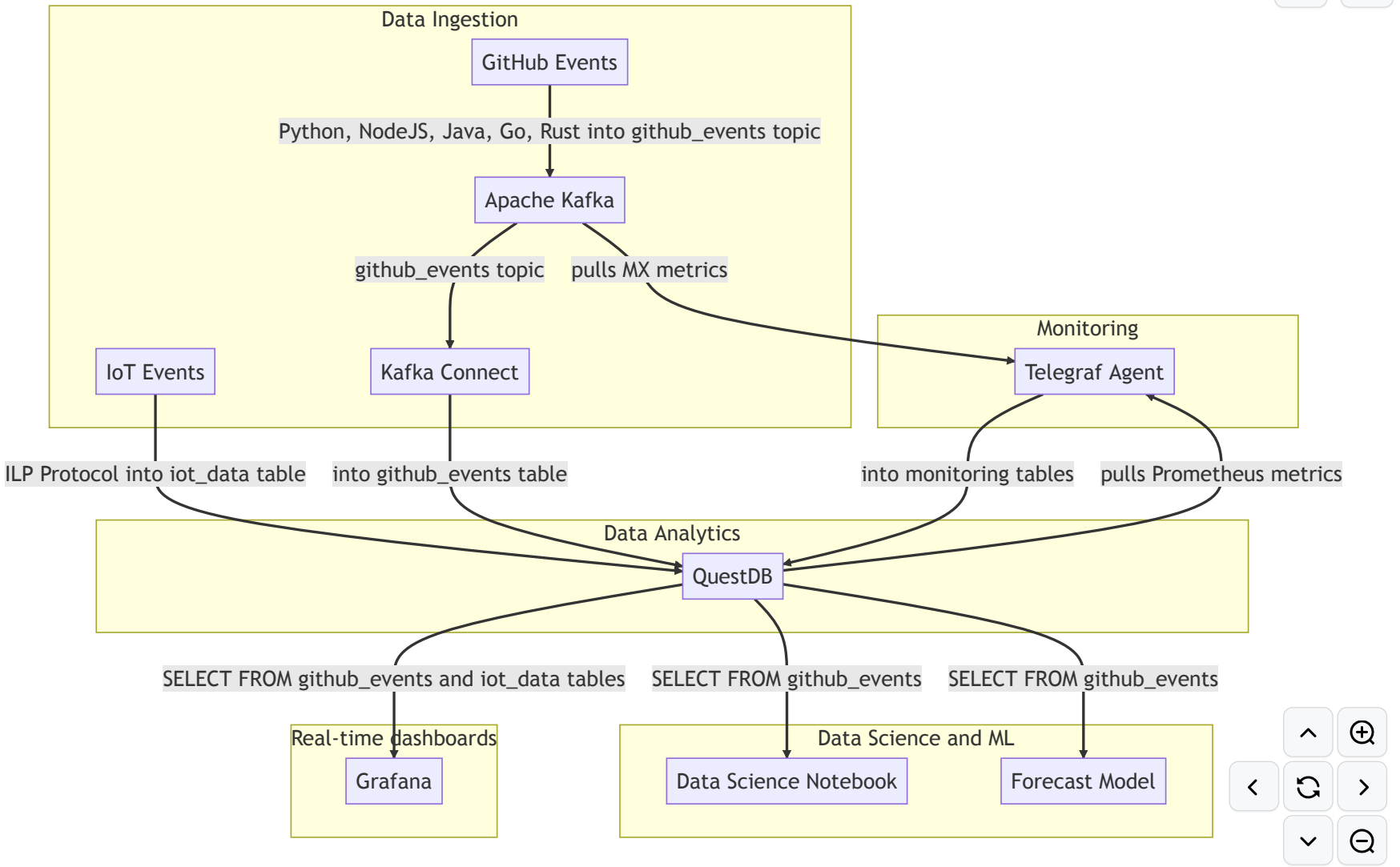


Everything, Everywhere, All At Once


Data Ingestion
Data Source
Real time app
Message Broker
Messages
Message Listener
Unbuffered data source
Pre-processed messages
Messages
metrics
Data Analytics
OLAP Database
Table data
Table data
aggregated data
metrics
Data Science and ML
Data Science Notebook
Forecast Model
Real-time dashboards
Dashboard
Monitoring
Server Agent
time series data
time series datab

https://github.com/questdb/time-series-streaming-analytics-template

What the time-series streaming analytics template contains


Data Ingestion
GitHub Events
Python, NodeJS, Java, Go, Rust into github_events topic
Apache Kafka
github_events topic
pulls MX metrics
IoT Events
Kafka Connect
Monitoring
Telegraf Agent
ILP Protocol into iot_data table
into github_events table
into monitoring tables
pulls Prometheus metrics
Data Analytics
QuestDB
SELECT FROM github_events and iot_data tables
SELECT FROM github_events
SELECT FROM github_events
Real-time dashboards
Grafana
Data Science and ML
Data Science Notebook
Forecast Model

How I made my first billion (of data points)

- a factory floor with 500 machines, or
- a fleet with 500 vehicles, or
- 50 trains, with 10 cars each, or
- 500 users with a mobile phone

Sending data every second

86,400

seconds in one day

604,800

seconds in any given week

2,628,288

seconds on your average month of 30.437 days

The data billionaire

- 43,200,000 rows a day.....
- 302,400,000 rows a week....
- 1,314,144,000 rows a month

Some not too nice things about streaming data

- It can get very big
- It never stops; always incomplete
- It is bursty
- It will lag at times
- It will come out of order
- It might get updated after you already emited results
- Individual data points lose value over time, but long-term aggregations are priceless
- Analysts prefer low latency and data freshness

The template is a good starting point but

- More data formats
- Data lifecycle policies
- Data quality
- Data Governance
- Configuring replication
- Integration with your data lake
- Monitoring dashboard and alerts
- (...)

https://github.com/questdb/time-series-streaming-analytics-template

https://questdb.io

https://kafka.apache.org/

https://jupyter.org/

https://github.com/influxdata/telegraf

https://grafana.com/

THANKS

JAVIER RAMIREZ

@supercoco9 / @j@chaos.social

Developer relations at