Cloud Chats
by aiven

Aiven for M3, Flink, Clickhouse

The whys and the whats explained

Aiven brings the best Open Source data technologies to all public clouds

Kafka

PostgreSQL

Elasticsearch

Cassandra

Grafana

Redis

InfluxDB

MySQL

Coming soon...

Clickhouse

M3

Flink

amazon web services™

Google Cloud Platform

Microsoft Azure

DigitalOcean

UpCloud

How we pick new services

- Project has to be Open Source
- Technically excellent and provides strong new capabilities
- Strong community with future outlook and ability to accept changes from other participants
- Customer demand for the solution
- ..and we also like to be able to use these ourselves (=dogfooding)

The Need for Analytics Database

As data volumes grow to make sense of it all
Analytics databases ride to the rescue

The need for an analytics database

- Aiven provides multiple relational databases (MySQL, PostgreSQL)
    - Work well for OLTP use cases
    - Can scale vertically to bigger hardware
        - ..but writes cannot be easily scaled horizontally
    - Once they have enough data management gets harder. (recovery time objective)
    - Because of need to be able to update data, they are row-oriented by nature which leads to poor compression

The need for an analytics database (Aiven view)

- Historically many of our customers have used services like Redshift, Snowflake or Google Bigquery
- All come with their own limitations and cost models
- The new service would ideally integrate well with our other services like Kafka
  - A lot of companies use Apache Kafka as their firehose so ingestion support needed
- Needs to have good support for contemporary hardware

Aiven for Clickhouse

SQL queries allowing you to analyze vast quantities of data in real-time

Aiven for Clickhouse

- Clickhouse is column oriented database horizontally scalable database
- Originally created by Yandex
  - Today the largest deployments of Clickhouse are probably at ByteDance and CloudFlare
- Claim to fame comes from it being incredibly fast
- Scale both horizontally and vertically
- Extensive support for compression
  - Huge read performance and storage space usage advantages
  - Multiple use cases from time-series data to text search

Aiven for Clickhouse

- Supports data ingestion from Apache Kafka
- Ability to scale horizontally to enormous data volumes
- Proven at enormous scale
- Varied use cases from text searching to time series to analytical queries
- Not merely quick, but consistently optimized for speed at all levels
- Data warehouse use for structured data

Aiven for Clickhouse

- Initially offered in configurations starting from one node up to n-node clusters
  - Easy scaling according to customer's data needs
- Automatic backups
- Integrates well with Aiven for Apache Kafka and other Aiven services like PostgreSQL/MySQL
- Ready-made metrics integrations from DataDog to Prometheus
- Availability Beta: late Q3

The Need for Time Series

Need for a time-series database (market view)

- Data volumes are growing very rapidly with the advent of IOT and containerization
- Data store needs to be horizontally scalable, Highly Available time-series database
- Needs to support good compression ratios for data
- Support for multiple ingestion data formats
- Ability to scale globally a nice plus

Need for time-series database (Aiven view)

- Our customer node counts grew 5x last year
- We operate globally so support for a globally distributed system a plus
- For us to offer richer, more detailed views for our customers services data volumes increasing rapidly
- Existing time-series options not that great

Aiven for M3

Globally scalable, highly available Open
Source distributed timeseries database

M3

- Horizontally scalable, HA time-series database
- Originated from Uber where proven at enormous scale
- Good compression ratio for data
- Geographically distributed operation supported
- Support for data resolution (aggregation) changes
- We also take care of the needed etcd clusters
- Provides the backbone for “unlimited” metrics use for us also internally

Aiven for M3

- Initially offered in configurations starting from three nodes up to n-node clusters
    - Allows no-downtime scaling according to customer needs
- Automatic backups
- Ingestion of time series data in multiple formats
    - Graphite
    - InfluxDB line protocol
    - Prometheus formats
- Integrates with Grafana for ready-made dashboards
- Availability: Beta Q2, GA Q3

The Need for Stream Processing System Tech

Trends (data processing)

- Underlying need is faster computation of data
  - 24 hour batch ETL cycles won’t cut it in the future
- More and more firms looking into this with ~Apache Kafka as the transport
- SQL is starting to look like the next language of choice for stream processing
  - Apache Beam, Apache Flink, Apache Spark, KSQL
- Unification of batch and real-time streaming use cases

Aiven for Flink

Framework and distributed processing engine for stateful computations over data streams.

Apache Flink

- Stateful Computations over Data Streams
  - Flink Allows you to run stream and batch processing computations over any supported data source
- Processing can take place as SQL or as custom code running against the data
- Custom code written in Java or Python

Apache Flink

- Supports an incredible variety of data sources and destinations
  - Object storage (S3, GCS, Azure Storage)
  - Streaming data platforms (Apache Kafka, Kinesis, Pub/Sub, Pulsar)
  - Relational databases (PostgreSQL, MySQL etc)
  - NoSQL services (Cassandra, Elasticsearch, InfluxDB)
- Ties many different data services into one coherent data pipeline
- Ready-made metrics integrations from DataDog to Prometheus

Aiven for Apache Flink

- Initially offered in configurations starting from three nodes up to n-node clusters
    - Allows easy scaling according to customer needs
- Initial support for running SQL queries over different data sources/destinations
- Will form the basis of many data pipeline solutions
- Flink + object storage supports data lake kind of use cases
- Availability: Beta Q3, GA Q4

Q & A

Questions?