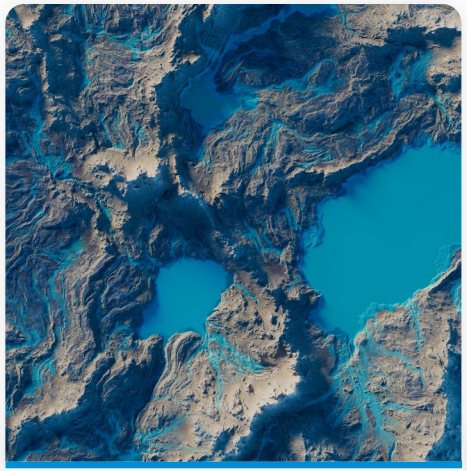
Real Time Data Lake

Client Success Story



Industry

Pharmaceutical
(Human Pharma)

Client

One of the leading international pharmaceutical company

Challenges

Largest Data Lake in Top Pharma Corporation runs 30k batch jobs daily storing data in 2PTB AWS S3 bucket and Snowflake database.

- Users need to have data faster, preferably in near real time
- There are tens of source systems – some of them can publish data as streams some can not
- SQL transformations must be applied on data before it is published to the users

Solution

- Event streaming platform (Kafka) was deployed on AWS Cloud. Data is streamed to it:
  - directly by source systems
  - with custom connectors running on Kafka Connect cluster
  - with commercial/external connectors
- Spark application was deployed on Kubernetes cluster. It can blend data from different topics and from AWS S3 using provided SQL transformation
- Starburst Trino SQL Access was deployed in Kubernetes cluster allowing to access data stored on S3 with SQL tools
- Snowflake connector allowing to streams directly from Kafka to Snowflake
- All the actions are orchestrated with simple declarative configuration provided in git

Benefits

- Better **quality of data** thanks to ability to apply data quality checks in real time – that allows business to find issues at early manufacturing stage
- **Data is provided** to Snowflake and SQL Access tool users in near **real time**
- Simple **transformations** can be applied in **real time**
- Solution is **scalable** new sources/streams can be added easily as the git configuration