

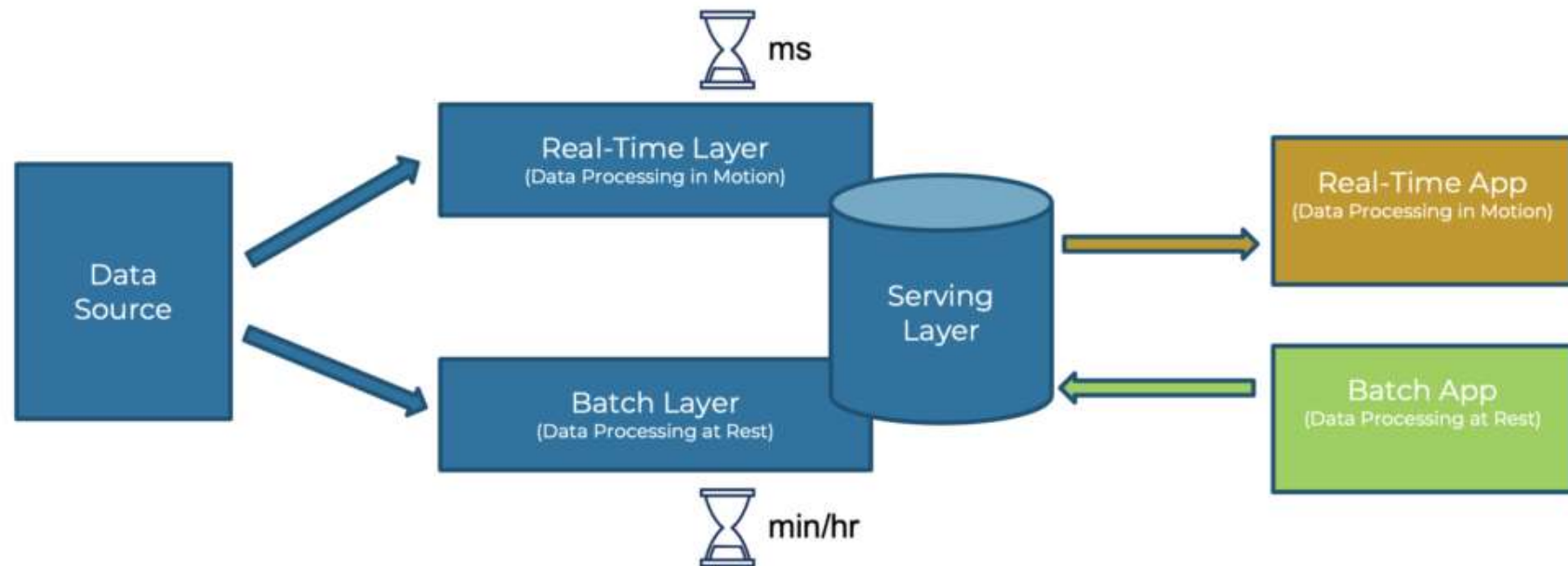
# Other Data Architectures

- ▶ Lambda Architecture
- ▶ Kappa Architecture
- ▶ Delta Architecture
  - Lakehouse Architecture
    - Data Lakes
    - Data Warehouse

# Lambda Architecture I

## Lambda Architecture

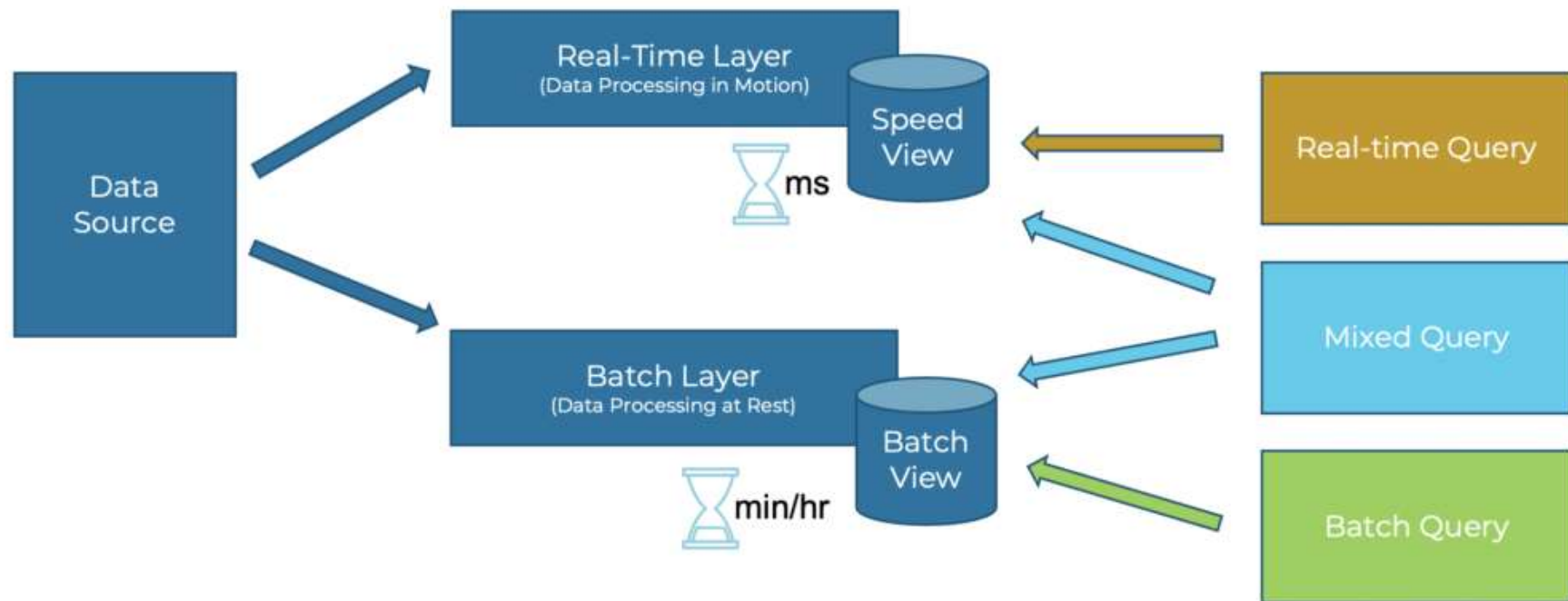
Option 1: Unified serving layer



# Lambda Architecture II

## Lambda Architecture

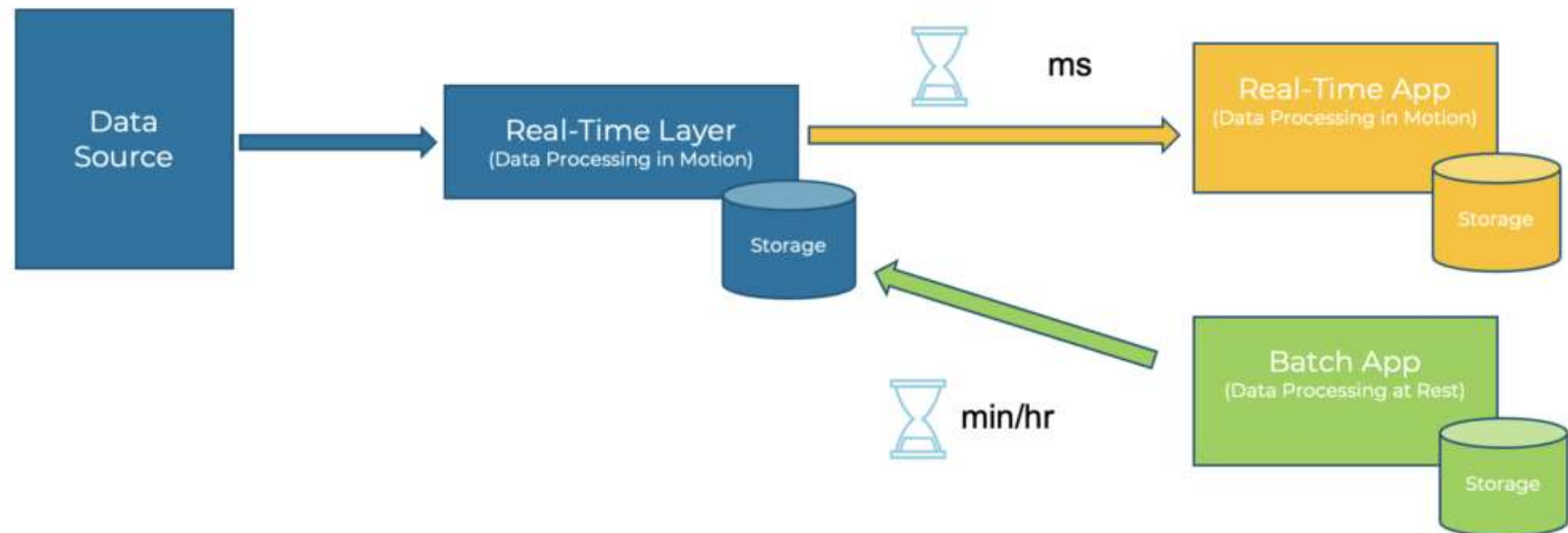
Option 2: Separate serving layers



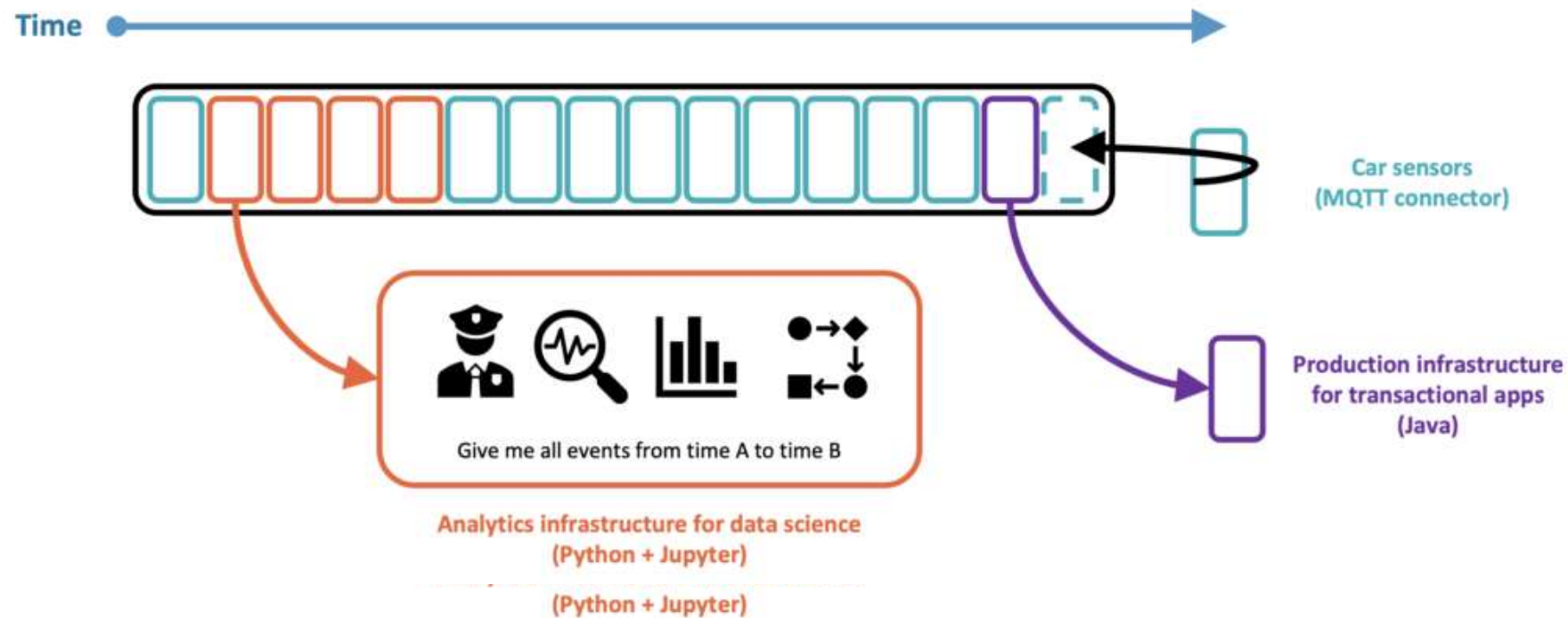
# Kappa Architecture

## Kappa Architecture

One pipeline for real-time and batch consumers



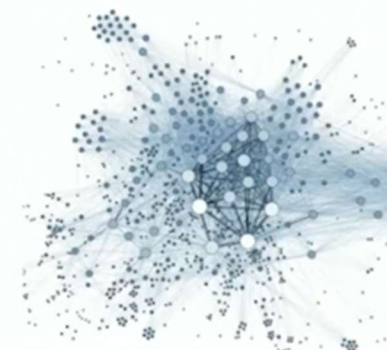
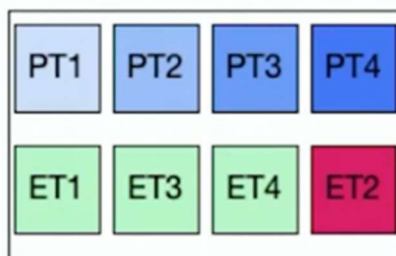
# Kappa for Transactional and Analytical Workloads



# Kappa is not free for lunch!

## Kappa Architecture - Challenges & Limitations

Disney  
STREAMING



### Re-Processing Data

What happens when you need to add a field?  
Or fix your algorithm?

### Out of order data

Event time vs. processing time - what do you do with records that arrive late?

### Added Cost?

Paying for compute vs. cold storage. Trade - soft costs (developers) vs. increased hardware costs.

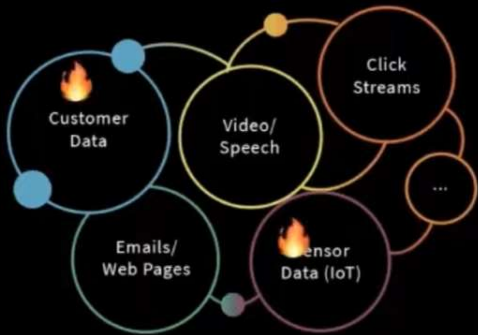
### Complex Joins

A few joins are ok - but what happens when you want to join together 25 "tables" from a relational data store?



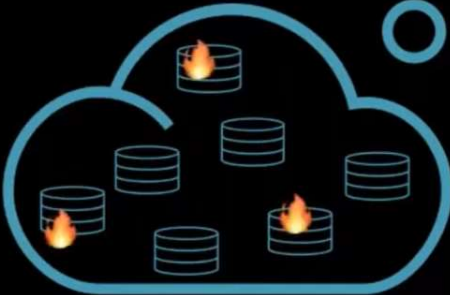
# The Promise of the Data Lake

## 1. Collect Everything



Garbage In

## 2. Store it all in the Data Lake



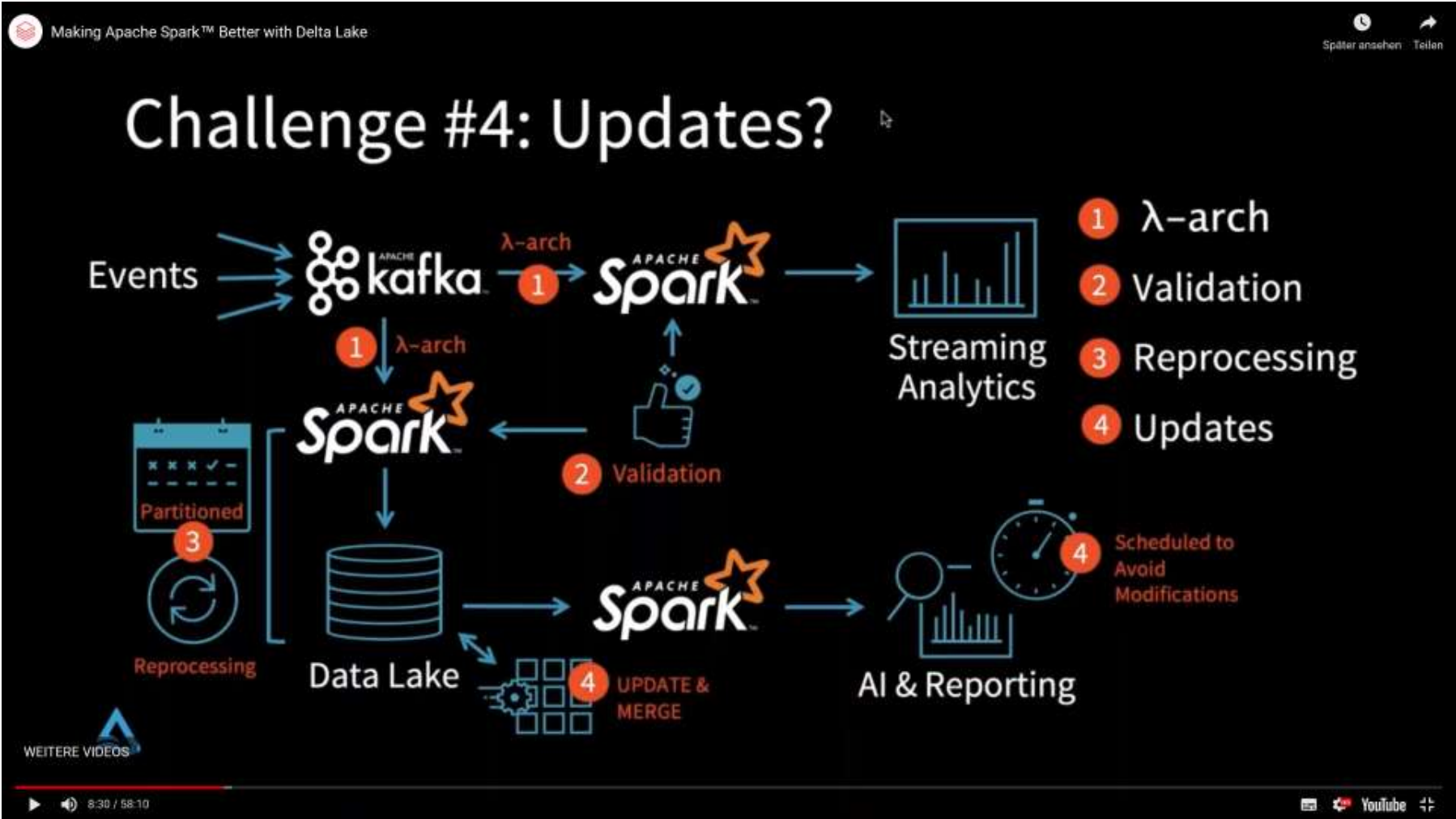
Garbage Stored

## 3. Data Science & Machine Learning



- Recommendation Engines
- Risk, Fraud Detection
- IoT & Predictive Maintenance
- Genomics & DNA Sequencing





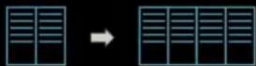




# Data Lake Distractions



**No atomicity** means failed production jobs leave data in corrupt state requiring tedious recovery




**No quality enforcement** creates inconsistent and unusable data



**No consistency / isolation** makes it almost impossible to mix appends and reads, batch and streaming




 Making Apache Spark™ Better with Delta Lake

Später ansehen

Teilen


# The DELTA LAKE Architecture




The diagram illustrates the Delta Lake architecture. On the left, four data sources are shown: two Kafka instances (represented by the Kafka logo), Kinesis (represented by the Kinesis logo), and a Data Lake (labeled 'Data Lake' with 'CSV, JSON, TXT...' below it). Arrows from these sources point to a central column of four database cylinder icons, representing the Delta Lake storage layer. From this central layer, arrows point to two output destinations on the right: 'Streaming Analytics' (represented by a bar chart icon) and 'AI & Reporting' (represented by a magnifying glass over a bar chart icon).


Open Standards, Open Source (Apache License)

Store petabytes of data without worries of lock-in. Growing community including Presto, Spark and more.

 WEITERE VIDEOS

▶ 🔊 11:51 / 58:10





 YouTube

 Making Apache Spark™ Better with Delta Lake


youtube.com befindet sich jetzt im Vollbildmodus. [Vollbild beenden \(Esc\)](#)

Später ansehen Teilen

# The DELTA LAKE


 kafka  
 Kinesis  
 CSV, JSON, TXT...  
Data Lake  
 **SPARK**

Bronze




Raw Ingestion

Silver



Filtered, Cleaned Augmented



Gold




Business-level Aggregates



\*Data Quality Levels\*


Quality


 Streaming Analytics  
 AI & Reporting

Delta Lake allows you to *incrementally* improve the quality of your data until it is **ready for consumption**.

 WEITERE VIDEOS

  13:02 / 58:10


 YouTube


 Making Apache Spark™ Better with Delta Lake

Später ansehen

Teilen


# The DELTA LAKE

 kafka


 Kinesis

CSV,  
JSON, TXT...


Data Lake

 **spark**


Bronze

  
Raw Ingestion

Silver

  
Filtered, Cleaned  
Augmented


Gold



  
Business-level  
Aggregates




Streaming Analytics


AI & Reporting

- Dumping ground for raw data
- Often with long retention (years)
- Avoid error-prone parsing

 WEITERE VIDEOS

  13:37 / 58:10


  YouTube 


 Making Apache Spark™ Better with Delta Lake

Später ansehen


Teilen


# The DELTA LAKE

 kafka


 Kinesis

CSV,  
JSON, TXT...

 Data Lake


 **SPARK**

Bronze




Raw Ingestion

Silver




Filtered, Cleaned  
Augmented

Gold




Business-level  
Aggregates


Streaming Analytics









AI & Reporting




Intermediate data with some cleanup applied.  
Queryable for easy debugging!

 WEITERE VIDEOS

  14:38 / 58:10


   YouTube 


 Making Apache Spark™ Better with Delta Lake

Später ansehen

Teilen


# The DELTA LAKE

 kafka


 Kinesis

CSV,  
JSON, TXT...

Data Lake


 **spark**

Bronze




Raw Ingestion

Silver




Filtered, Cleaned  
Augmented

Gold




Business-level  
Aggregates

Streaming Analytics






AI & Reporting






Clean data, ready for consumption.  
Read with Spark or Presto\*


\*Coming Soon

 WEITERE VIDEOS

  15:12 / 58:10

  YouTube 







 Making Apache Spark™ Better with Delta Lake


Später ansehen

Teilen

# The DELTA LAKE


 kafka  
 Kinesis  
 CSV, JSON, TXT...  
Data Lake  
 spark

Bronze




Raw Ingestion

Silver



Filtered, Cleaned Augmented

Gold




Business-level Aggregates



Streaming Analytics




AI & Reporting

Streams move data through the Delta Lake

- Low-latency or manually triggered
- Eliminates management of schedules and jobs

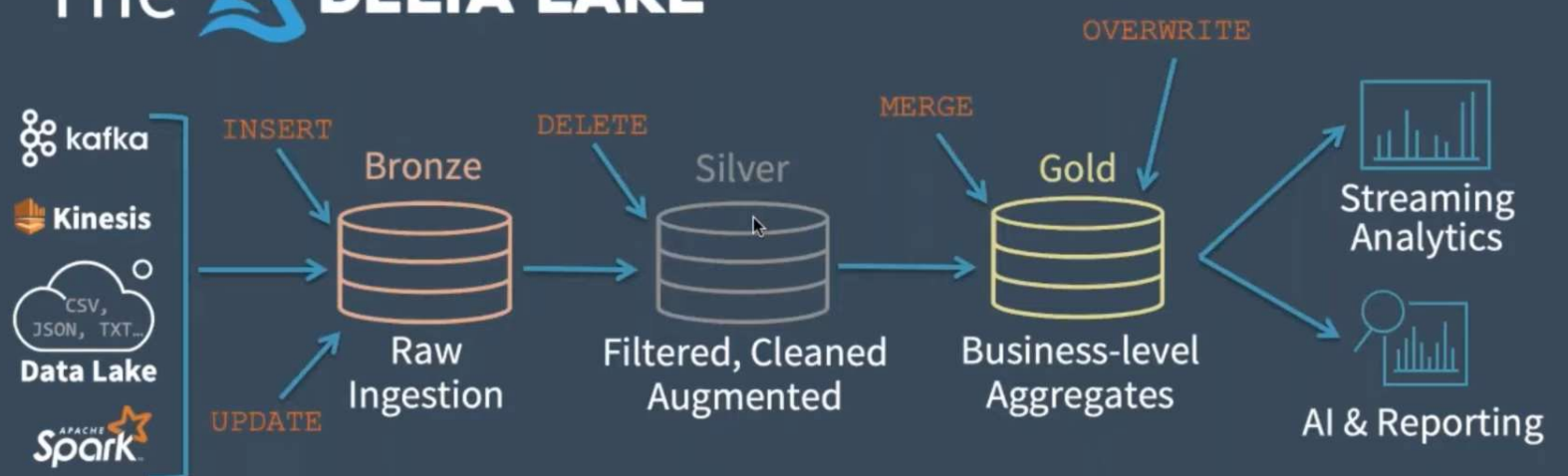
 WEITERE VIDEOS

  15:24 / 58:10

  YouTube 

Making Apache Spark™ Better with Delta Lake

# The DELTA LAKE



The diagram illustrates the Delta Lake data pipeline. On the left, data sources (Kafka, Kinesis, Data Lake with CSV, JSON, TXT, and Apache Spark) feed into the 'Bronze' stage, labeled 'Raw Ingestion'. An 'INSERT' arrow points to Bronze, and an 'UPDATE' arrow points from the sources to Bronze. An arrow from Bronze to the 'Silver' stage, labeled 'Filtered, Cleaned Augmented', has a 'DELETE' arrow pointing to it. An arrow from Silver to the 'Gold' stage, labeled 'Business-level Aggregates', has a 'MERGE' arrow pointing to it. An 'OVERWRITE' arrow points to the Gold stage. From the Gold stage, arrows point to 'Streaming Analytics' and 'AI & Reporting'.

Delta Lake also supports batch jobs and standard DML

- Retention
- Corrections
- GDPR
- UPSERTS

\*DML Coming in 0.3.0

18:58 / 58:09

Für Details scrollen



Making Apache Spark™ Better with Delta Lake

# The DELTA LAKE



Easy to recompute when business logic changes:

- Clear tables
- Restart streams



19:57 / 58:09

Für Details scrollen



Making Apache Spark™ Better with Delta Lake

# Delta On Disk

Transaction Log

Table Versions

(Optional) Partition Directories

my\_table/

\_delta\_log/

00000.json

00001.json

date=2019-01-01/

file-1.parquet

29:43 / 58:09

Für Details scrollen

Making Apache Spark™ Better with Delta Lake

# Solving Conflicts Optimistically

1. Record start version

2. Record reads/writes

3. Attempt commit

4. If someone else wins, check if anything you read has changed.

5. Try again.

Read: Schema

Write: Append

User 1

Read: Schema

Write: Append

User 2

000000.json

000001.json

000002.json

←

→

↗

↘

▶

⏮

⏪

⏩

⏭

35:07 / 58:09

Für Details scrollen

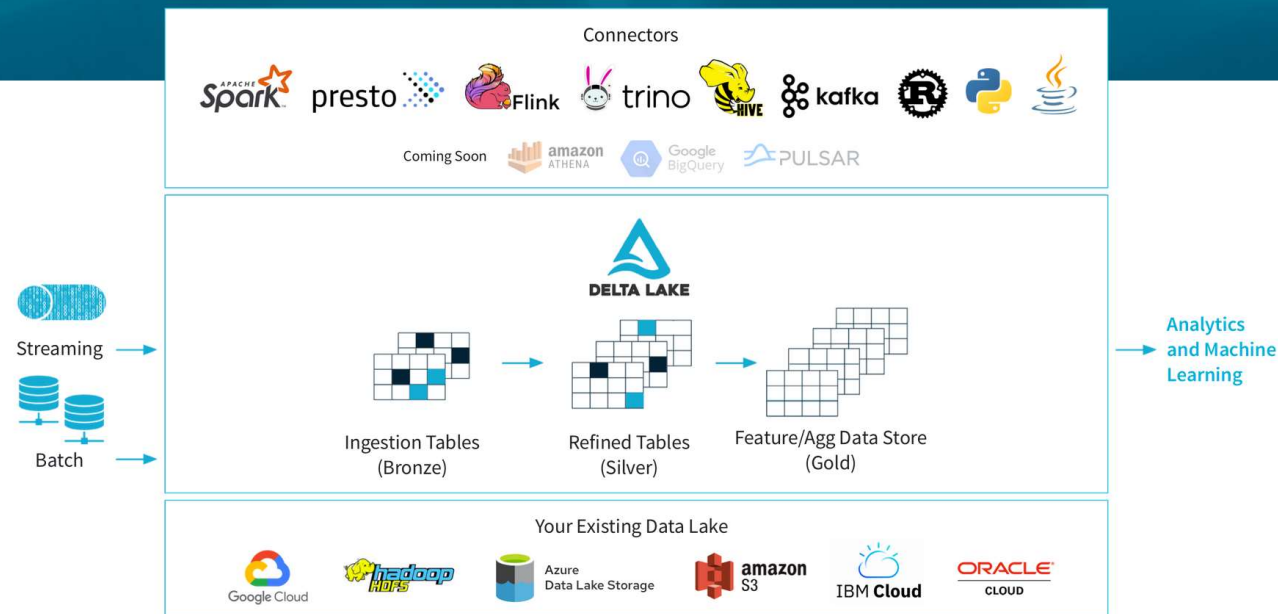
🔍

📄

🔧

🔗

# Delta Lake → delta.io



<https://delta.io/>

## Organizations using and contributing to Delta Lake

Thousands of companies are processing exabytes of data per month with Delta Lake. See more [here](#).

databricks

Tencent 腾讯

COMCAST

Alibaba Group

VIACOM

ciena