SQL KONFERENZ 2026



Data quality in the Databricks Lakehouse

Vitalija Bartuseviciute & Alexander Klein

Who are we?



Vitalija Bartuseviciute

Databricks MVP

Senior Consultant - Capgemini



Alexander Klein

Senior BI consultant

Data quality not just nowadays problem!

SQL KONFERENZ 2026

Data quality topic began in Mesopotamia c. 3000 BCE

- Grain storage
- Livestock counts
- Labor obligations
- Tax payments



Records kept on cuneiform – clay tablets

Accuracy and completeness was important, but noone called it data quality.

Incorrect records could cause famine mismanagement.

Ancient Egypt

Strong management systems made by pharaohs!

- Harvest yields
- Nile flood levels
- Labor assignments for pyramid building



Early quality controls had standardized formats!

Accuracy affected taxation and state stability.

Well, actually with data-driven decisions!



SQL KONFERENZ 2026

SQL KONFERENZ 2026

Why is Data Quality important?

- Builds trust
- Enables good data - Driven decisions
- Allows meaningful Machine learning & AI solutions



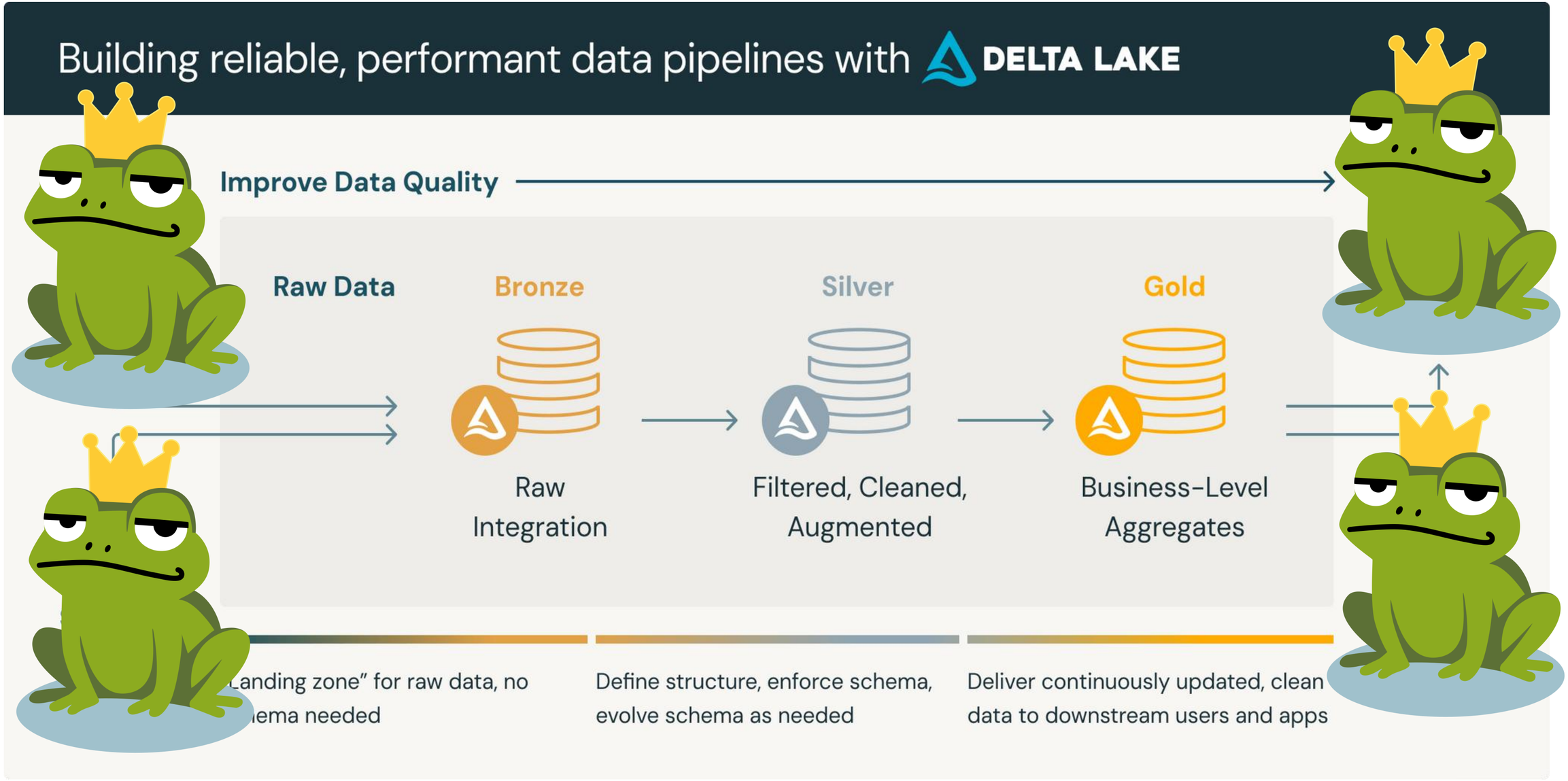
SQL KONFERENZ 2026

How did we get to the topic of data quality?



Client in Retail

Medallion architecture – conceptual foundation



Why should one care about data quality in 2026?

All starts with good data



AI Hallucinations

Wrong insights

Reduced trust in data

Wrong decisions

SQL KONFERENZ 2026

Data quality solutions from Databricks

DQX	Lakehouse Monitoring	LDP Expectations
Proactive monitoring Quarantine bad data; Custom quality rules; Works with all types of clusters.	Reactive monitoring Unity Catalog.	Use with LDP pipelines. Suitable for DLT only.



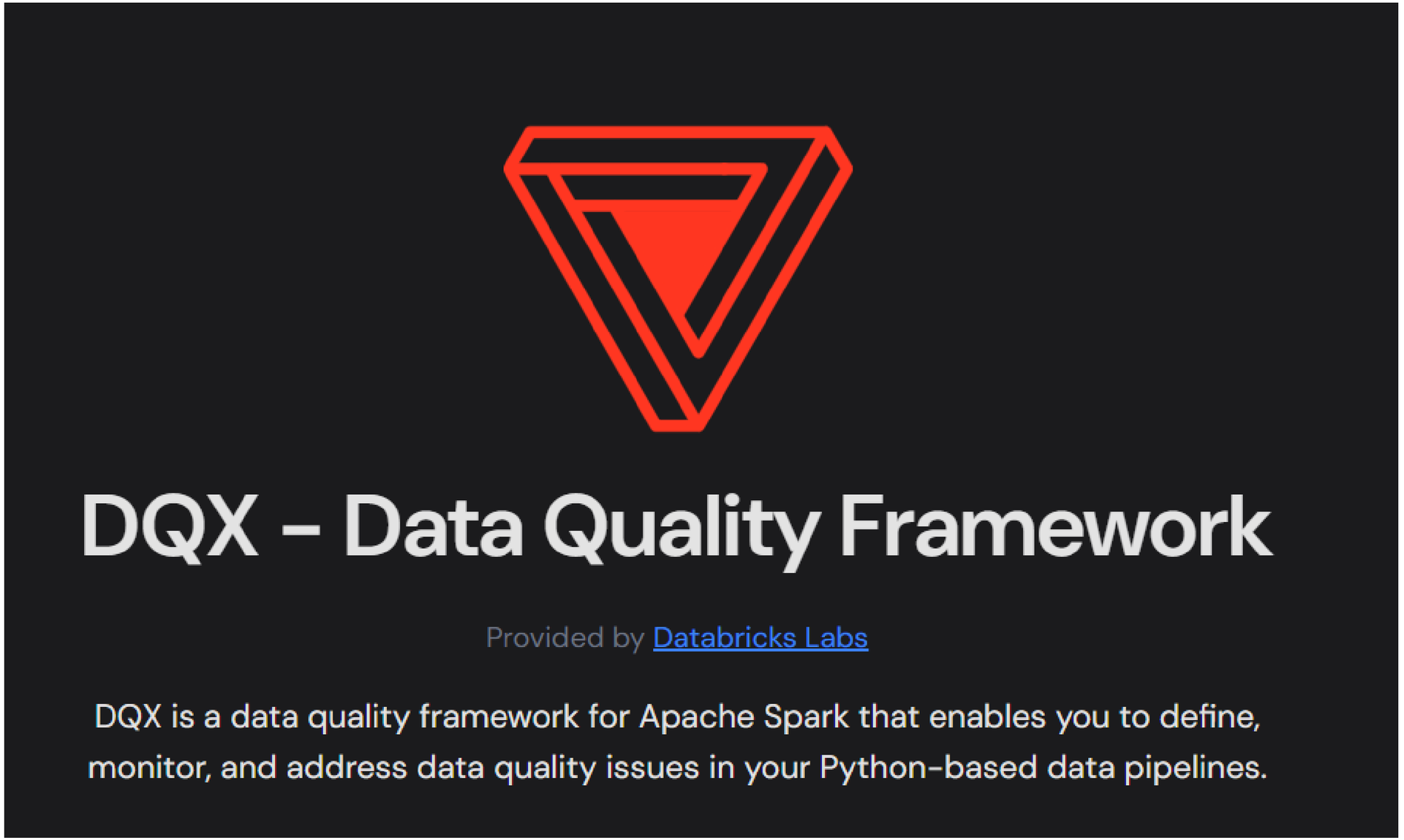
SQL
KONFERENZ
2026



Chapter 1:

DQX

DQX - Data Quality Framework



SQL KONFERENZ 2026

What makes DQX Unique?



- **Easy to use**

Designed to be simple and intuitive. Different data personas can define data quality checks without extensive training

SQL KONFERENZ 2026

What makes DQX Unique?



- **Detailed information on data quality issues**

Granular insights into data quality problems – users can understand specific reasons behind data quality failures.

Reduced troubleshooting time – reduced cost.

SQL KONFERENZ 2026

What makes DQX Unique?

- **Supports streaming**

Built to work both with data at-rest and in-transit, enabling real-time data quality validation



SQL KONFERENZ 2026

What makes DQX Unique?



- **Proactive monitoring**

Can monitor your data quality before writing to the target table. Reactive is also supported.

In practice – less things to clean up!

SQL KONFERENZ 2026

What makes DQX Unique?

- **Custom data quality rules**

Users can define their own data quality rules.

But also can dqx can suggest some checks.



SQL KONFERENZ 2026

What makes DQX Unique?



- **Easy integration with Databricks**

Designed to work with Databricks accross all engines – Spark Core, Spark Structured Streaming and Lakeflow pipelines (ex.DLT).

Cluster types (Standard and serverless)

SQL KONFERENZ 2026

What makes DQX Unique?



- **Row, Column and Dataset – level rules**

Supports row-level, column- level and dataset-level quality checks.

Validation for data at different granularities.

SQL KONFERENZ 2026

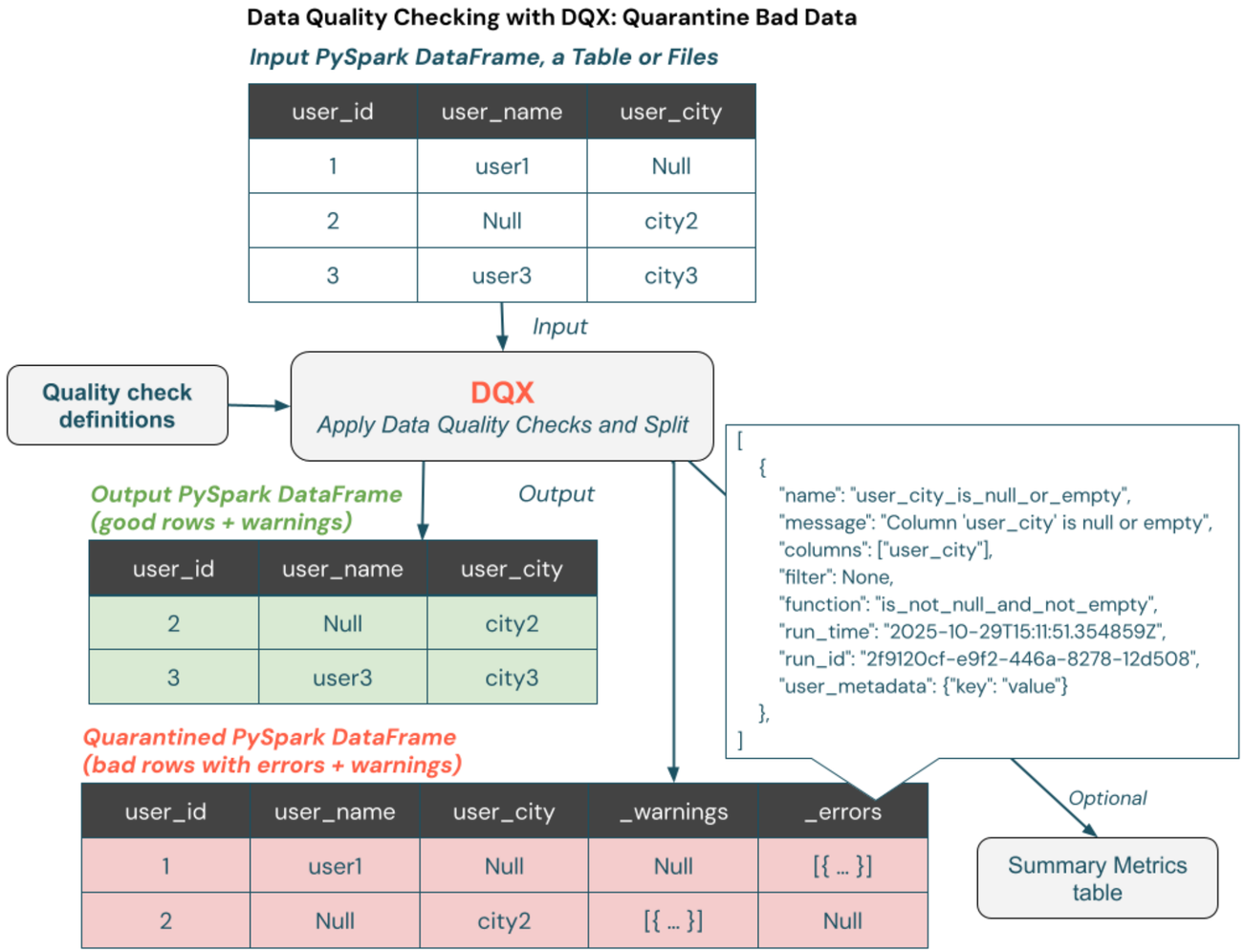
Core functionalities



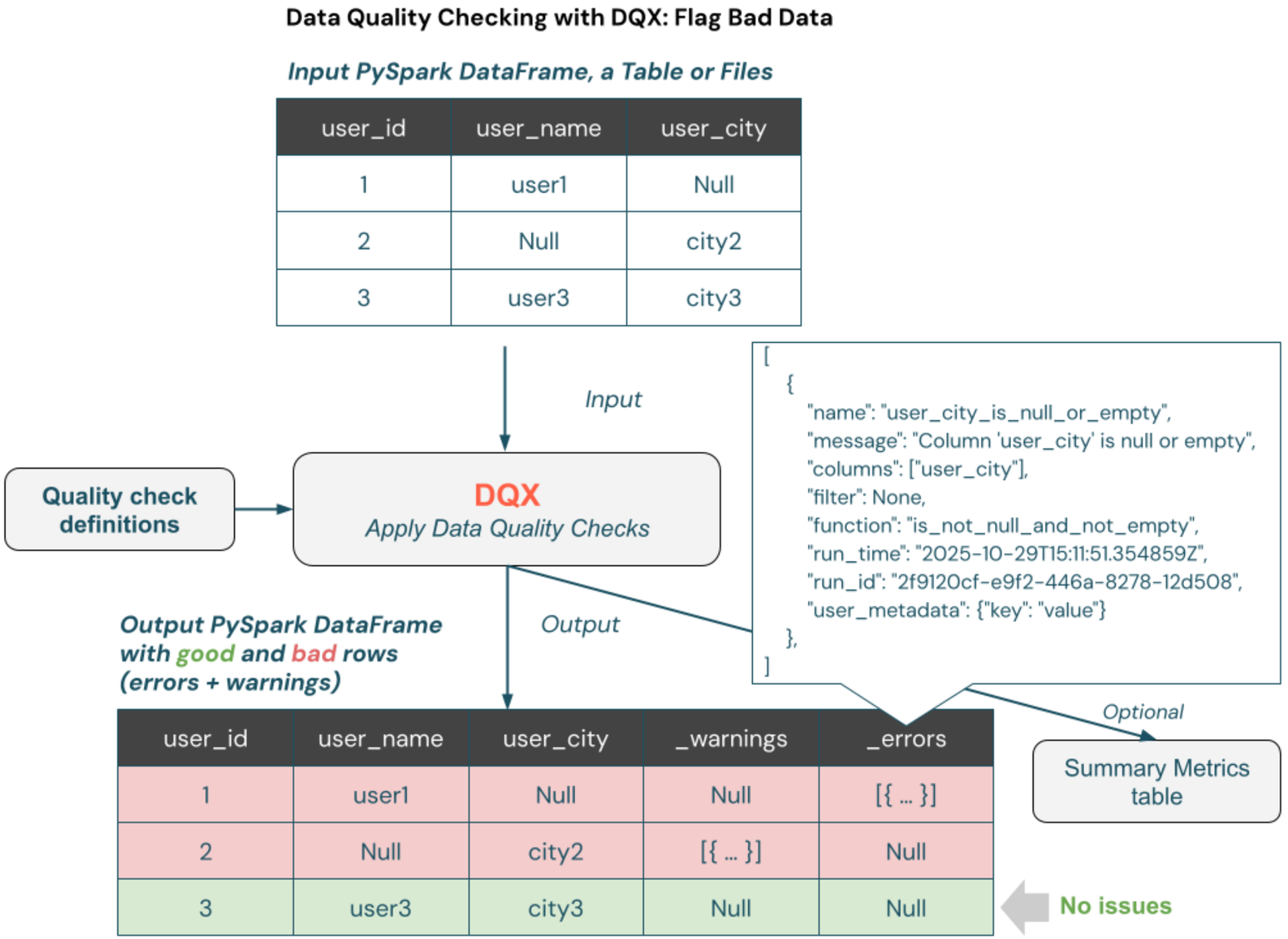
- **Option 1: Apply checks and quarantine "bad" data.**
- Apply checks on the DataFrame and quarantine invalid records to ensure "bad" data is never written to the output.

- **Option 2: Apply checks and annotate "bad" data.**
- Apply checks on the DataFrame and annotate invalid records as additional columns.

Option 1: Apply checks and quarantine "bad" data.



Option 2: Apply checks and annotate "bad" data.



SQL KONFERENZ 2026

Why DQX?

- Provides granular insights into data quality problems.

- Allows users to understand specific reasons behind data quality failures
- Reduces troubleshooting time



SQL KONFERENZ 2026

Main difference between Data Quality monitoring?

Proactive monitoring



- Allows to monitor data quality before writing to the target table
- Reactive monitoring also supported

SQL KONFERENZ 2026

Custom quality rules

- Users can define own data quality roles

- Makes DQX flexible and adaptable to various data quality requirements and domains



SQL KONFERENZ 2026

DQX profiler

- Profile and generate candidate data quality

The profiler analyzes datasets to generate summary statistics and data quality rule candidates automatically.



SQL KONFERENZ 2026

DQX usage in the Lakehouse Architecture



- Ongoing data quality monitoring between layers is crucial to detect degradation
- Monitoring in gold ensures high-quality outputs for end-users and business decisions
- Dead-letter pattern to quarantine invalid data and re-ingest after curation to ensure data quality constraints are met.

Standard DQX Quality Rules/Checks

Built-in Checks Available Today

Row-level checks

- Not-null, not-empty, not-null and not-empty
- Value within a list
- Value not less (greater) than
- Value not in the (near) future
- Regex match
- SQL expression
- PII detection (custom check)
- Support for complex types: structs, maps, and arrays
- ...

*Data Set-level checks

- Row count not less than, not greater than
- Row count equal / mean / median / avg less/greater than
- Column in another DataFrame
- Primary Key (PK) / Uniqueness check
- Foreign Key check
- Outliers detection
- Anomaly detection

**coming soon*

SQL KONFERENZ 2026

When to use DQX?

- Granular validation – row, column, dataset level data quality checks
- Custom quality rules – define and enforce data quality rules specific for your needs
- Data in-transit checking – monitor & validate before data is saved
- Data quarantining – isolate and manage bad data before it affects downstream processes



SQL
KONFERENZ
2026



Chapter 2:

LDP Expectations

SQL KONFERENZ 2026

Lakeflow Spark Declarative Pipelines (LDP)



Databricks Lakeflow is a unified data engineering solution designed to simplify the entire data lifecycle - from ingestion and transformation to orchestration.

It aims to replace fragmented tools with a single, integrated experience within the Databricks Data Intelligence Platform.

SQL KONFERENZ 2026

Lakeflow Spark Declarative Pipelines (LDP)

The Core: Spark Declarative Pipelines



The "secret sauce" of Lakeflow is the **Declarative Pipeline** (built on the evolution of Delta Live Tables).

SQL KONFERENZ 2026

Declarative vs. Imperative:



Imperative - you tell the system *how* to do every step (e.g., "open a connection, read this file, handle this error, save to that path").

Declarative - you define the *desired state* of your data (e.g., "This table should always contain the cleaned version of that source").

SQL KONFERENZ 2026

Automatic Infrastructure:



Spark automatically handles the underlying complexities like scaling clusters, managing checkpoints for streaming, and recovering from failures.

SQL KONFERENZ 2026

Three components of a lakeflow

1. Lakeflow connect



- Provides native connectors to easily ingest data from apps (Salesforce, Workday) and databases (SQL Server, MySQL) without needing third-party ETL tools.

SQL KONFERENZ 2026

Three components of a lakeflow
2. Lakeflow Pipelines



- Uses **Spark Declarative Pipelines** to transform data. It treats batch and streaming data the same way, ensuring high data quality through built-in testing ("Expectations").

SQL KONFERENZ 2026

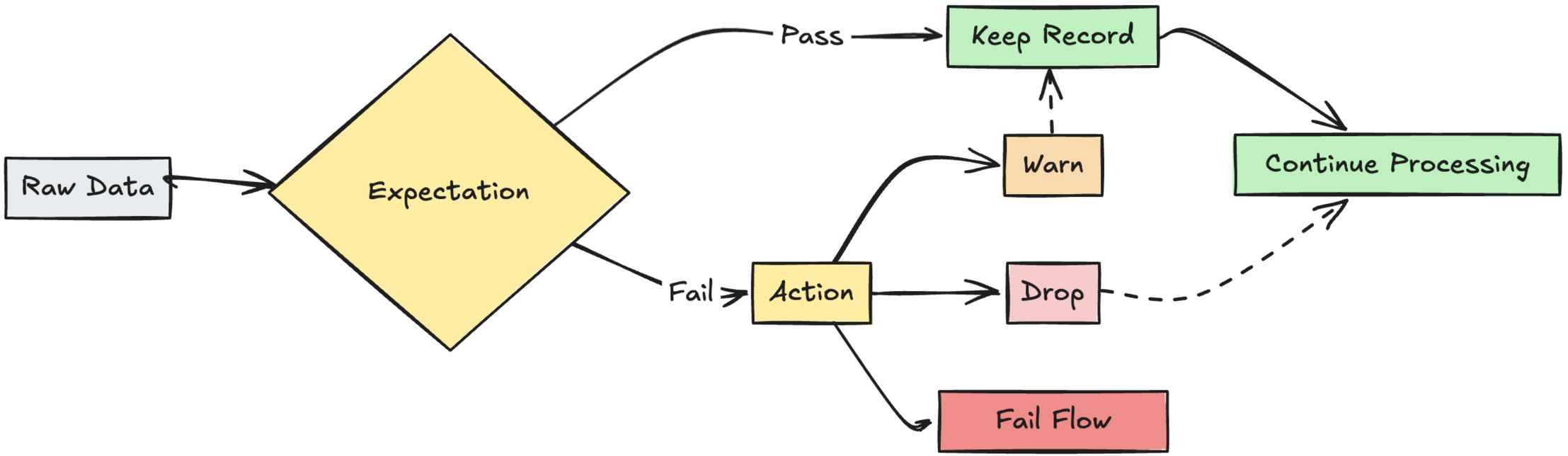
Three components of a lakeflow

3. Lakeflow Jobs

- An advanced orchestrator that schedules tasks, manages dependencies, and provides health monitoring across the entire workflow.



Manage data quality with pipeline expectations



SQL
KONFERENZ
2026

What are expectations?



Expectations are optional clauses in pipeline materialized view, streaming table, or view creation statements that **apply data quality checks** on each record passing through a query.

SQL KONFERENZ 2026

What are expectations?



Expectations use standard SQL Boolean statements to specify constraints.

You can combine multiple expectations for a single dataset and set expectations across all dataset declarations in a pipeline.

What are expectations?



Python

```python
@dp.table
@dp.expect("valid_customer_age", "age BETWEEN 0 AND 120")
def customers():
  return spark.readStream.table("datasets.samples.raw_customers")
```


SQL

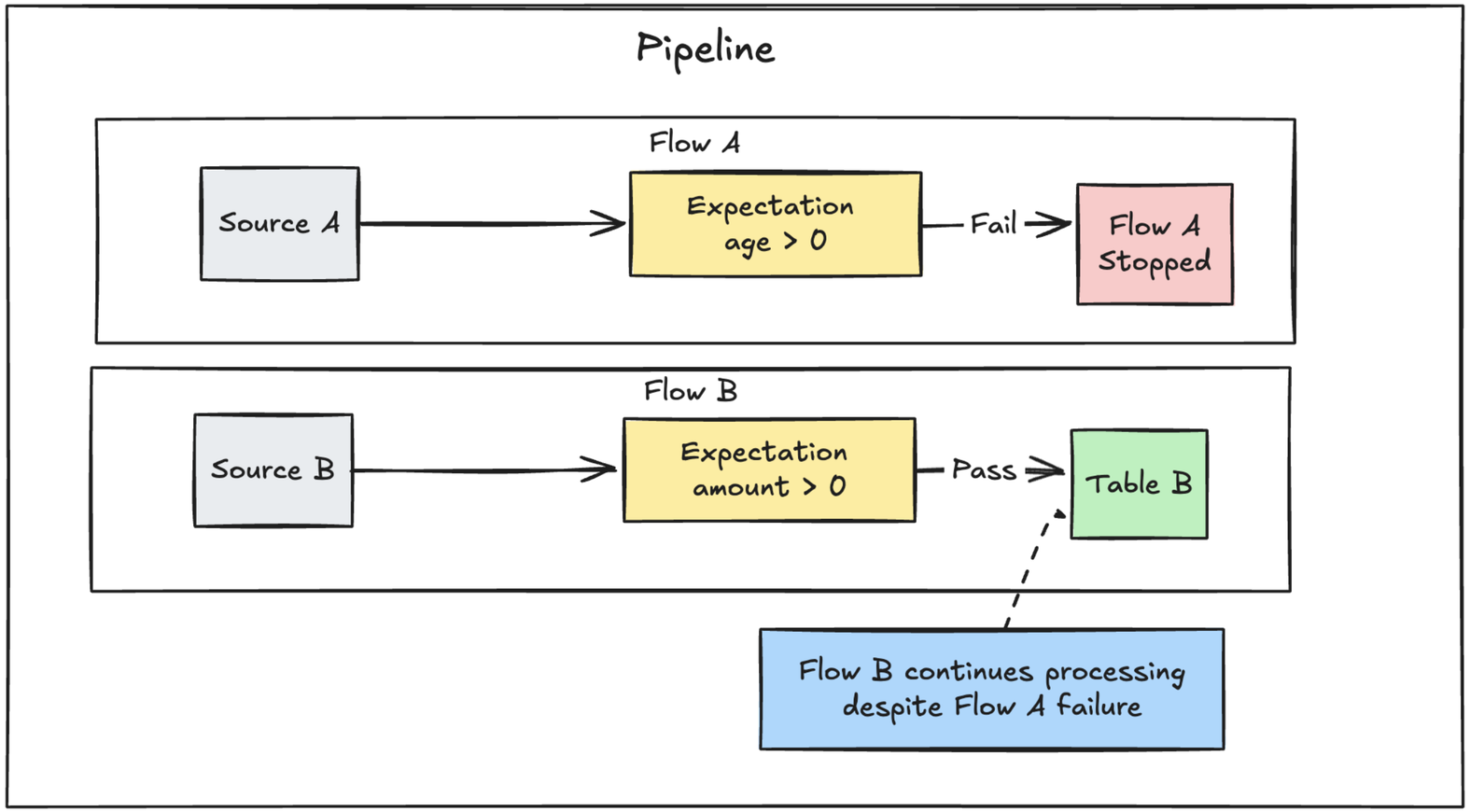
```sql
CREATE OR REFRESH STREAMING TABLE customers(
  CONSTRAINT valid_customer_age EXPECT (age BETWEEN 0 AND 120)
) AS SELECT * FROM STREAM(datasets.samples.raw_customers);
```

Action on invalid record



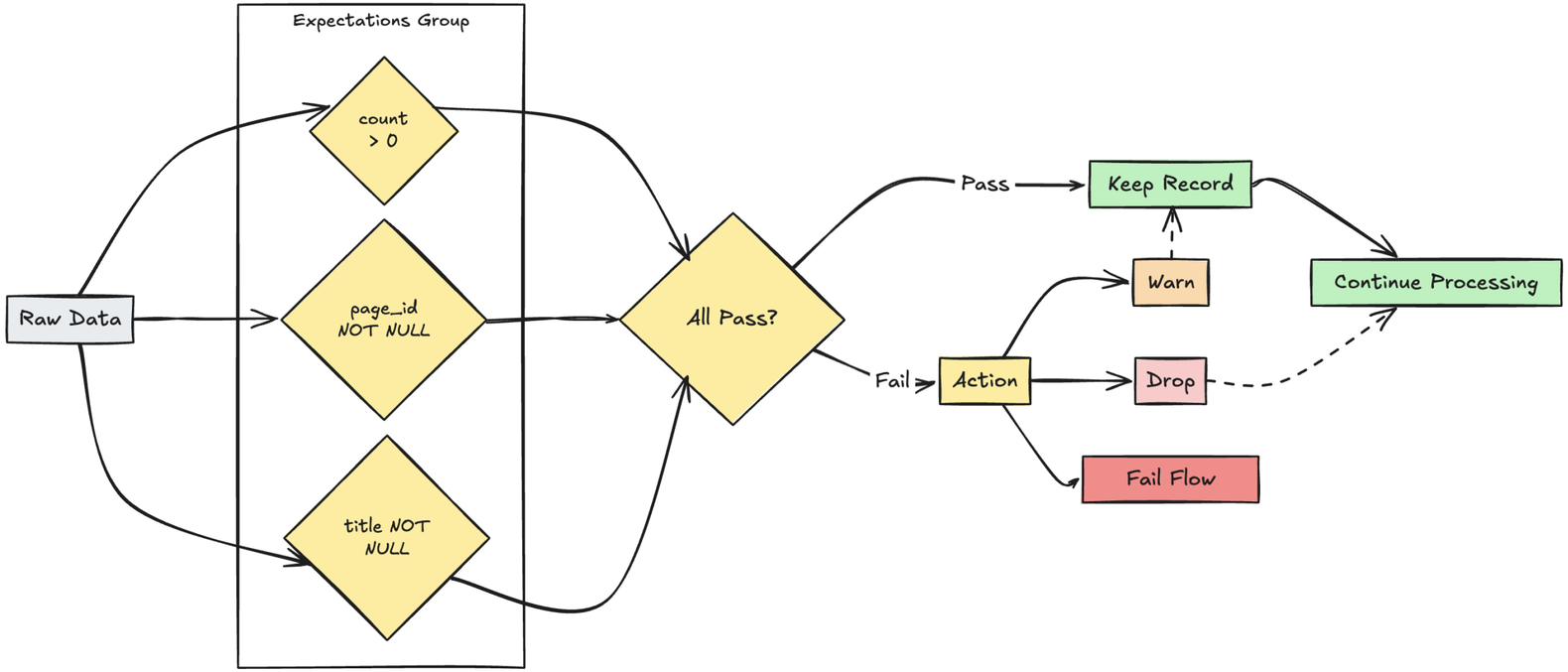
Action	SQL syntax	Python syntax	Result
warn (default)	`EXPECT`	`dp.expect`	Invalid records are written to the target.
drop	`EXPECT ... ON VIOLATION DROP ROW`	`dp.expect_or_drop`	Invalid records are dropped before data is written to the target. The count of dropped records is logged alongside other dataset metrics.
fail	`EXPECT ... ON VIOLATION FAIL UPDATE`	`dp.expect_or_fail`	Invalid records prevent the update from succeeding. Manual intervention is required before reprocessing. This expectation causes a failure of a single flow and does not cause other flows in your pipeline to fail.

If you have multiple parallel flows defined in a pipeline, failure of a single flow does not cause other flows to fail



Multiple expectations management

While both SQL and Python support multiple expectations in a single dataset, only Python allows you to group multiple expectations and specify collective actions.



SQL KONFERENZ 2026

Why does it matter?

- **Reduced Complexity:**
  You write significantly less "boilerplate" code. You focus on the business logic, not the plumbing.



SQL KONFERENZ 2026

Why does it matter?



- **Unified Batch & Streaming:**
  You can use the same code for real-time data and daily batches. The engine decides the most efficient way to process it.

SQL KONFERENZ 2026

Why does it matter?



- **Data Intelligence:**
  Because it's integrated with the Unity Catalog, Lakeflow automatically tracks data lineage
  (where data came from and where it goes) and applies security policies.

SQL
KONFERENZ
2026



Chapter 3:
Databricks Data Quality monitoring

SQL KONFERENZ 2026

What is Data quality monitoring?

Previously known as lakehouse monitoring

- Allows teams to monitor the health of their dataset without added tools or complexity.

- Powered by Unity Catalog



SQL KONFERENZ 2026

What does Data quality monitoring do?



- Automatically monitors data quality metrics, statistical trends and anomalies over time.

SQL KONFERENZ 2026

Enable anomaly detection for your schema



Advanced

Data Quality Monitoring — Disabled — Enable

Monitor data quality at scale. Learn more

Predictive Optimization — Enabled (Inherited)

Advanced

Data Quality Monitoring — Enabled — View results

Monitor data quality at scale. Learn more

Predictive Optimization — Enabled (Inherited)

Automate maintainance operations for this table. Learn more

Anomaly detection



SQL KONFERENZ 2026

- Enables scalable data quality monitoring with one click
- Automatically analyzes historical data patterns to detect anomalies in table freshness and completeness
- Tables monitored as they update

Anomaly detection



SQL KONFERENZ 2026

- **Freshness** - how recently a table has been updated.

  Analyzes the history of commits to a table and builds a per-table model to predict the time of the next commit.

If a commit is unusually late, the table is marked as stale.

SQL KONFERENZ 2026

Anomaly Detection



- **Completeness** number of rows expected to be written to the table in the last 24 hours.
- Data quality monitoring analyzes the historical row count, and based on this data, predicts a range of expected number of rows.
- If the number of rows committed over the last 24 hours is less than the lower bound of this range, a table is marked as incomplete.

SQL
KONFERENZ
2026

Catalog Explorer ›

Data Quality Monitoring

Send feedback

Vitalija "_"

Catalog: techgirls

Schema: medallion

Data Quality Monitoring in progress

Analyzing historical data to detect anomalies and intelligently monitoring tables as they update.

Estimated time: ~30 minutes



SQL
KONFERENZ
2026

Data Quality Monitoring

Send feedback

Vitalija "_"

Catalog: techgirls | Schema: medallion

Data Quality

100% healthy

14 tables healthy | 0 tables unhealthy

Tables Monitored

14

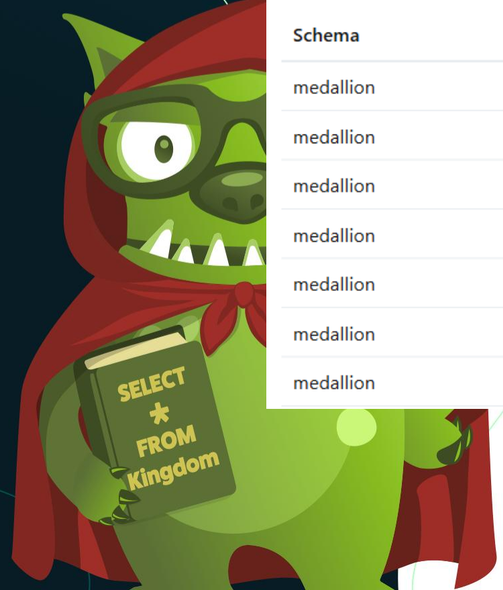
100% of all tables

All Monitored Tables

The most critical tables appear first, based on downstream usage.

Unhealthy 0 | Healthy 14 | Error 0

Schema	Table	Status	Last Scanned	Impact	Scan Frequency	Results
medallion	rock_silver	Healthy	1 hour ago	High	Every 1 week	Review
medallion	rock_silver_clone	Healthy	1 hour ago	High	Every 1 week	Review
medallion	rock_bronze	Healthy	1 hour ago	Medium	Every 1 week	Review
medallion	rock_checks	Training	1 hour ago	Medium	Every 1 week	Review
medallion	rock_raw	Healthy	1 hour ago	Medium	Every 1 week	Review
medallion	rock_checked	Training	1 hour ago	Low	Every 1 week	Review
medallion	rock_checks_checked	Training	1 hour ago	Low	Every 1 week	Review



Data profiling

Provides quantitative measures that help you track and confirm the quality and consistency of your data over time.



SQL KONFERENZ 2026

- Summary statistics for tables
- Tracking historical trends
- Unexpected changes accross your data and ML
- Helps better understand data and reduce downstream operational toll

How does data profiling actually work?



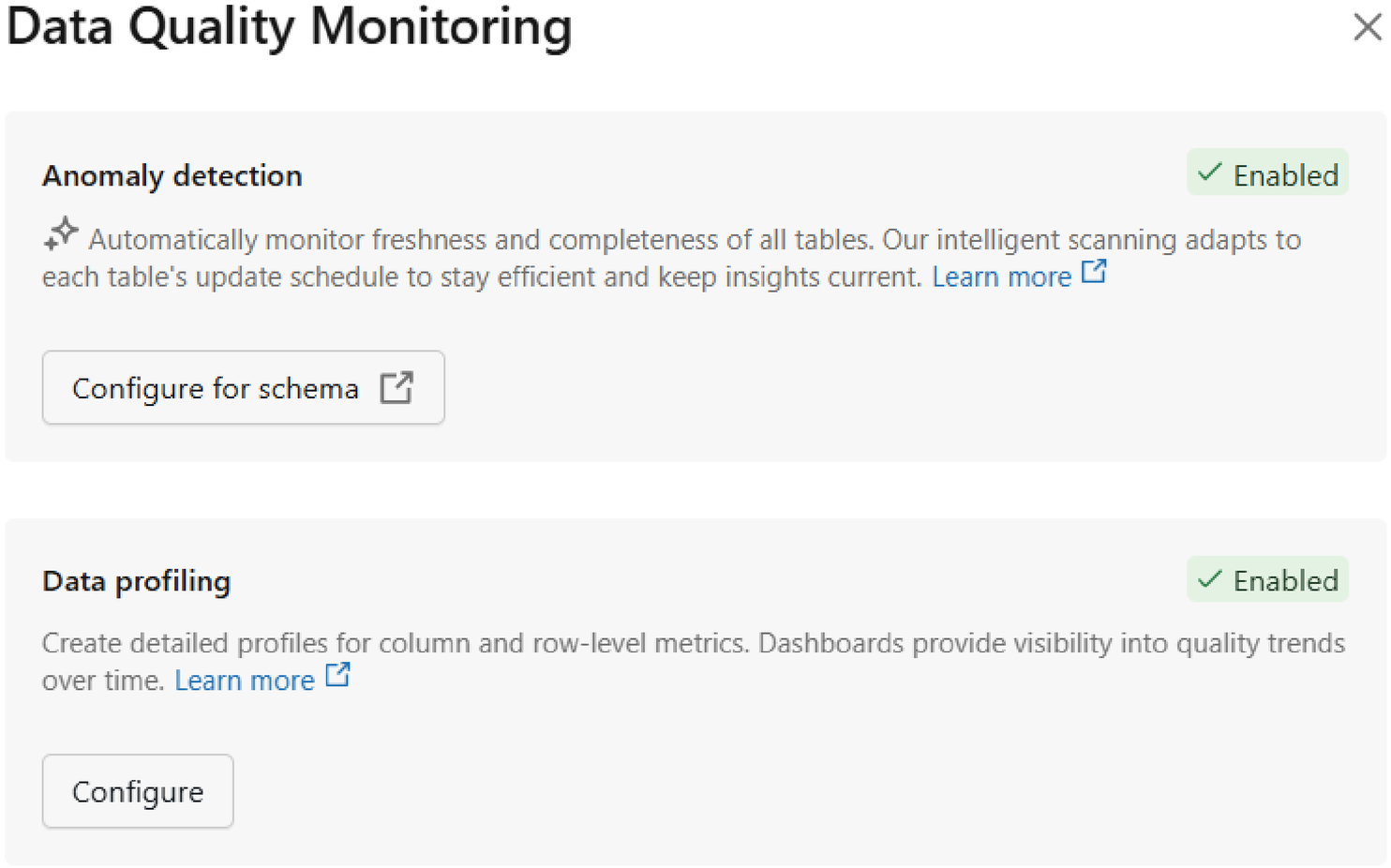
SQL KONFERENZ 2026

- The metric tables are Delta tables and are stored in a Unity Catalog schema that you specify.
- You can view these tables using the Databricks UI, query them using Databricks SQL, and create dashboards and alerts based on them.

Enable data profiling on your table

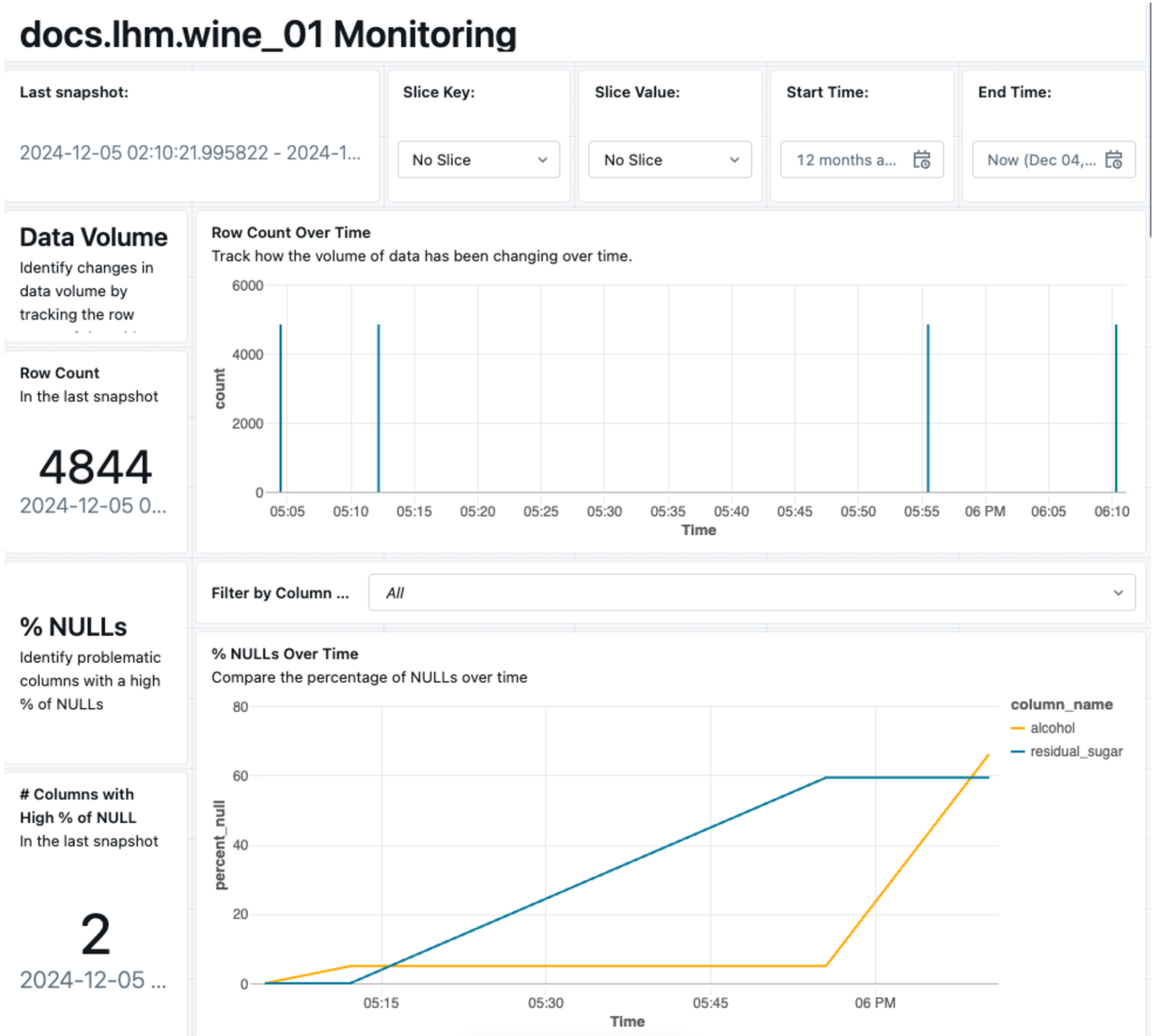


SQL KONFERENZ 2026



Data Profiling Dashboard

SQL KONFERENZ 2026



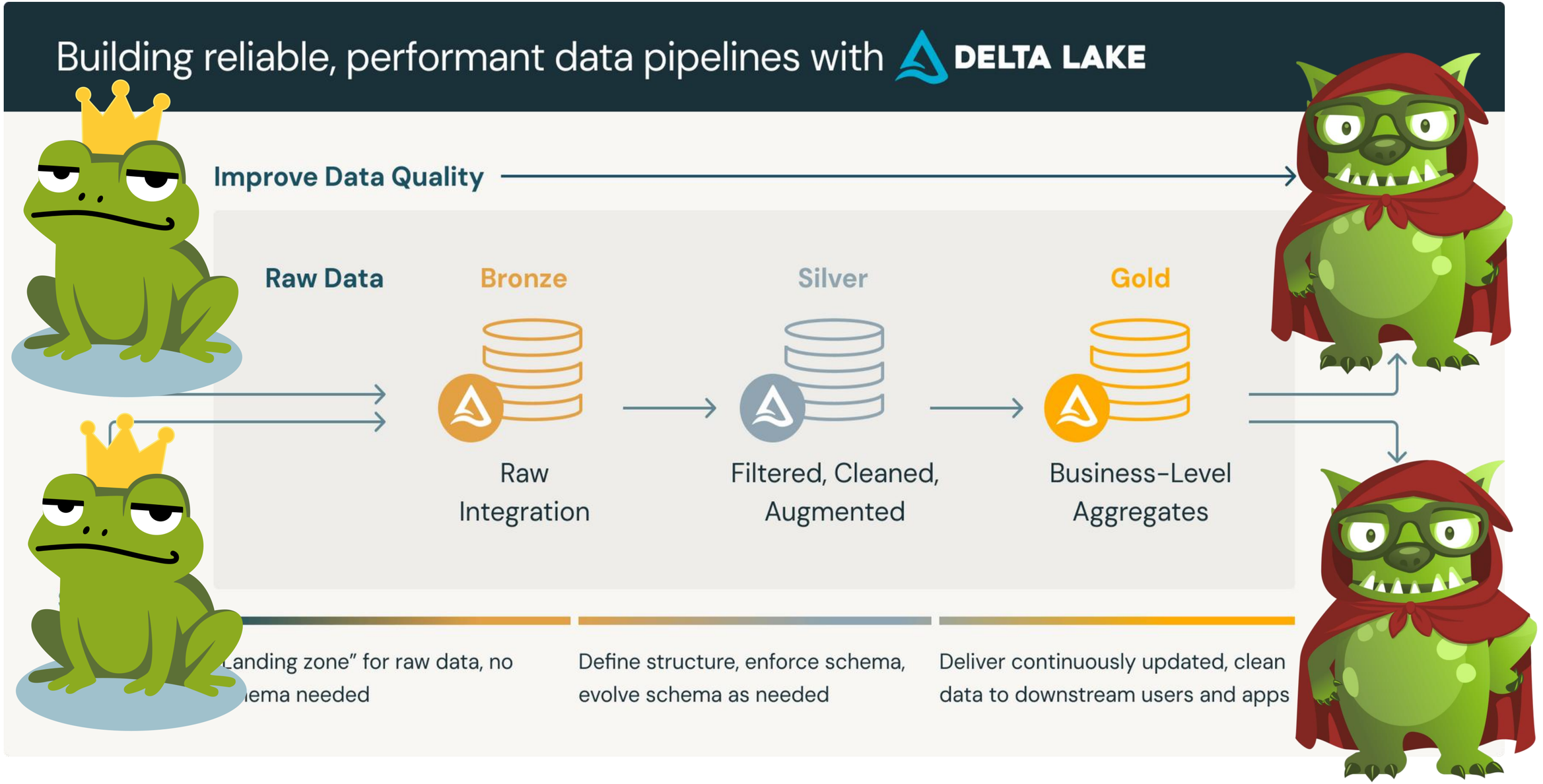
To sum up



SQL KONFERENZ 2026

- Better data = Better decisions
- Many tools, not needing to do it manually!
- So one can focus from «fixing data» to driving innovation

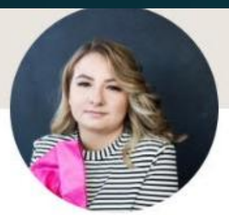
Hope your frogs turn into princes! ☺



Please provide us feedback!



Thank you!



Vitalija Bartusevičiūtė

Databricks MVP | Senior Consultant - Data Engineer @ Capgemini



Alexander Klein

- bei Alexander Klein

Requirements for Data profiling



SQL KONFERENZ 2026

- Your workspace must be enabled for Unity Catalog and you must have access to Databricks SQL.
- To enable data profiling, you must have the following privileges:
  - USE CATALOG on the catalog and USE SCHEMA on the schema containing the table.
  - SELECT on the table.
  - MANAGE on the catalog, schema, or table.

SQL KONFERENZ 2026

Accellerated root cause analysis



- Debug data and model quality issues faster by using data profiles, historical trends and anomaly signals to trace back issues to their source.

- Helps teams reduce time to resolution and improve reliability of production pipelines.