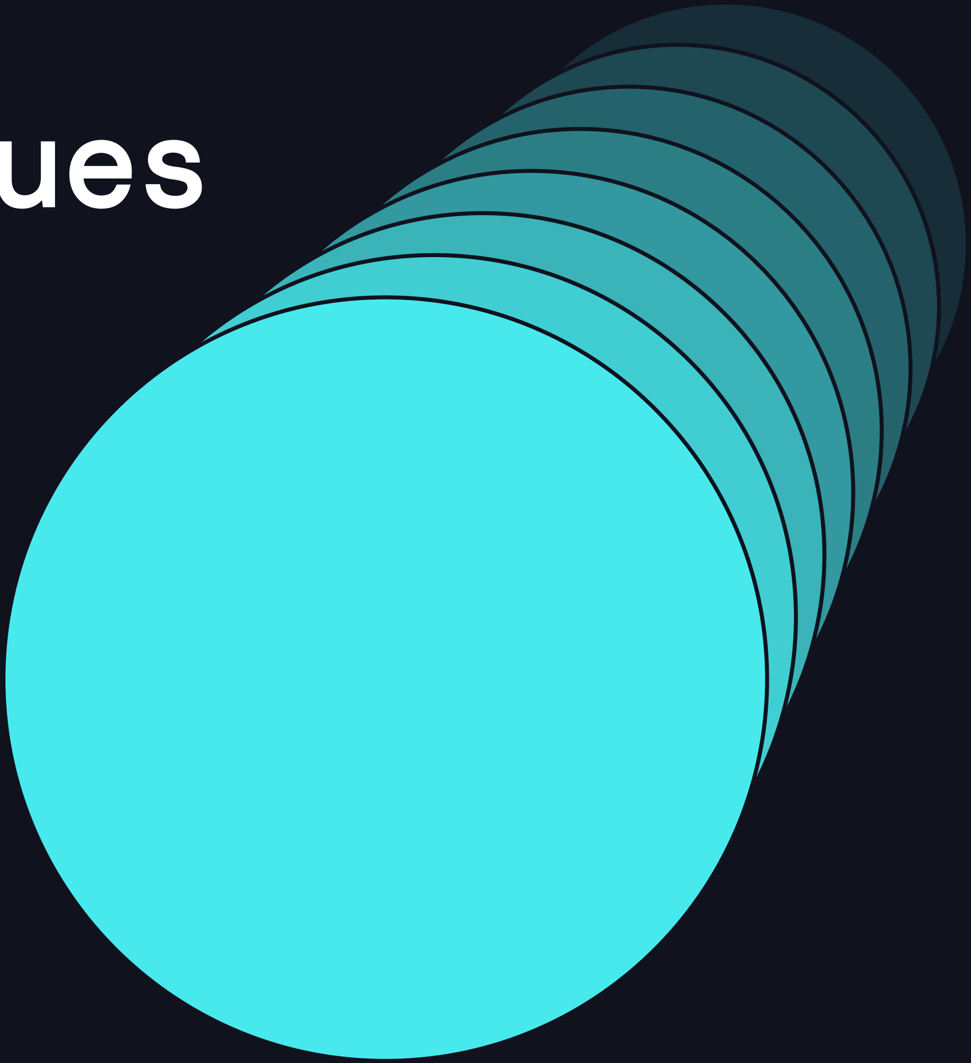
DATA+AI SUMMIT
BY databricks

Practical techniques for applying data quality in the lakehouse



Liping Huang & Lara Rachidi

11 June 2024

Meet the Speakers

Liping Huang

Senior Solutions Architect

https://www.linkedin.com/in/lipinght/

Lara Rachidi

Solutions Architect

https://www.linkedin.com/in/lara-rachidi/

Agenda

- ❑ Six Dimensions of Data Quality
- ❑ Data Quality Management Lifecycle
- ❑ Crawl
- ❑ Walk
- ❑ Run
- ❑ Example Medallion Architecture

Six dimensions model

Dimensions of Data Quality

Consistency

No conflicts in data

Accuracy

No erros in data

Validity

Conform to set formats

Completeness

No missing data

Uniqueness

No duplicates in data

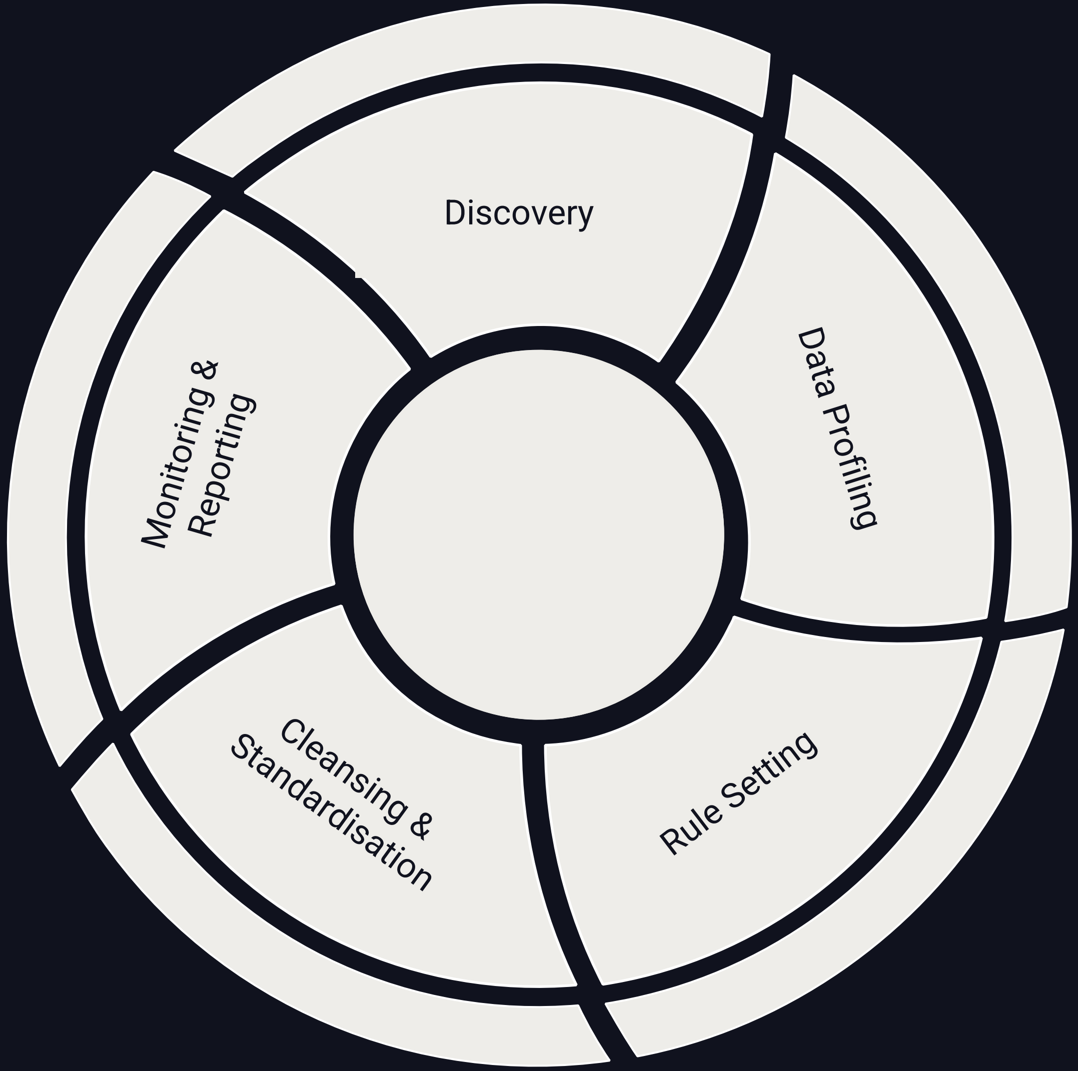
Timeliness

Up-to-date data

DATA+AI SUMMIT
BY databricks

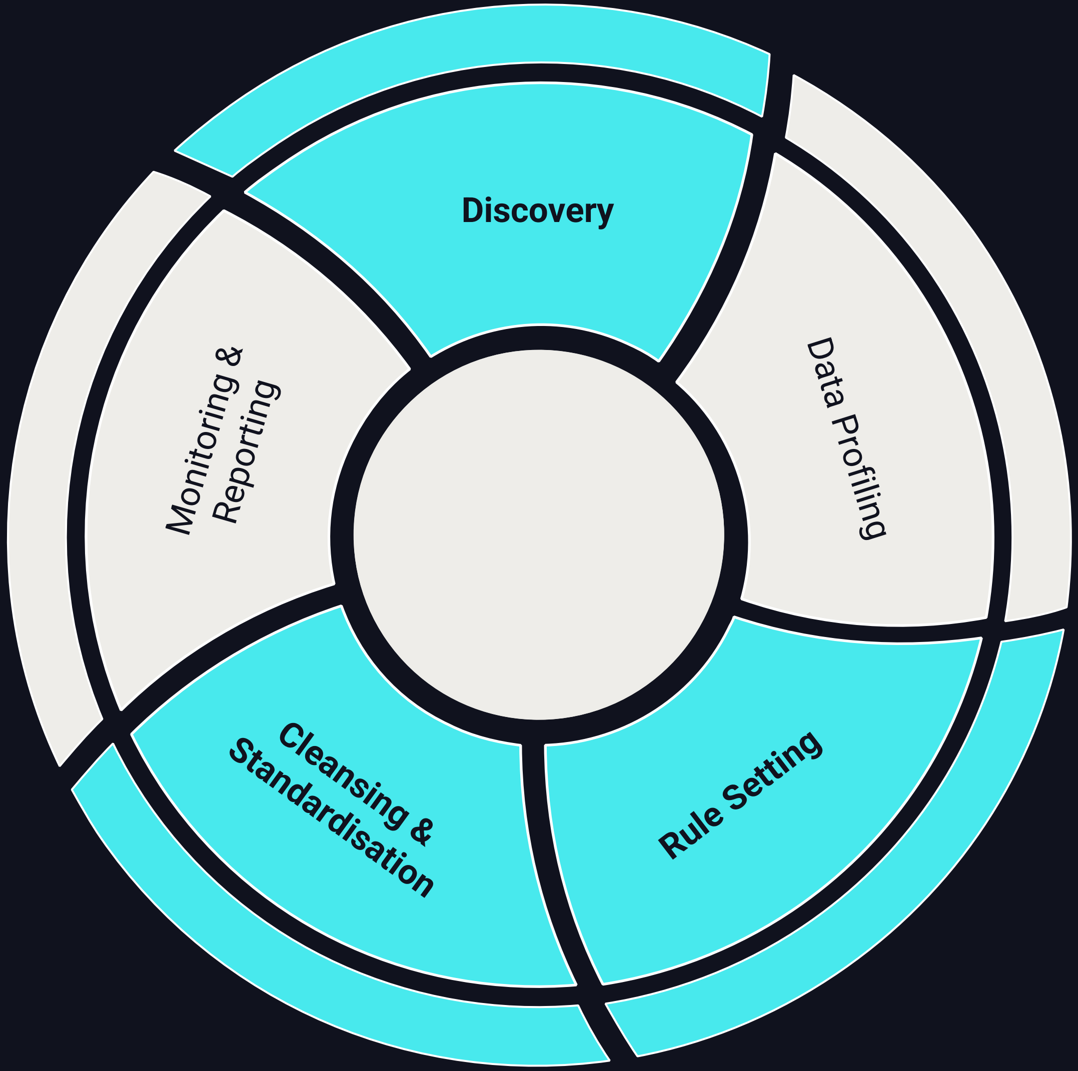
Data Quality Management Lifecycle

Data Quality Management Lifecycle



Crawl

Data Quality Management Lifecycle



Discovery

Data Profiling

Rule Setting

Cleansing & Standardisation

Monitoring & Reporting

Discovery

Discovery

Data quality is a team sport

Key Stakeholders

Business uers
Data scientists
Data engineers
Data analysts
Data stewards
Compliance officers
Executive stakeholders

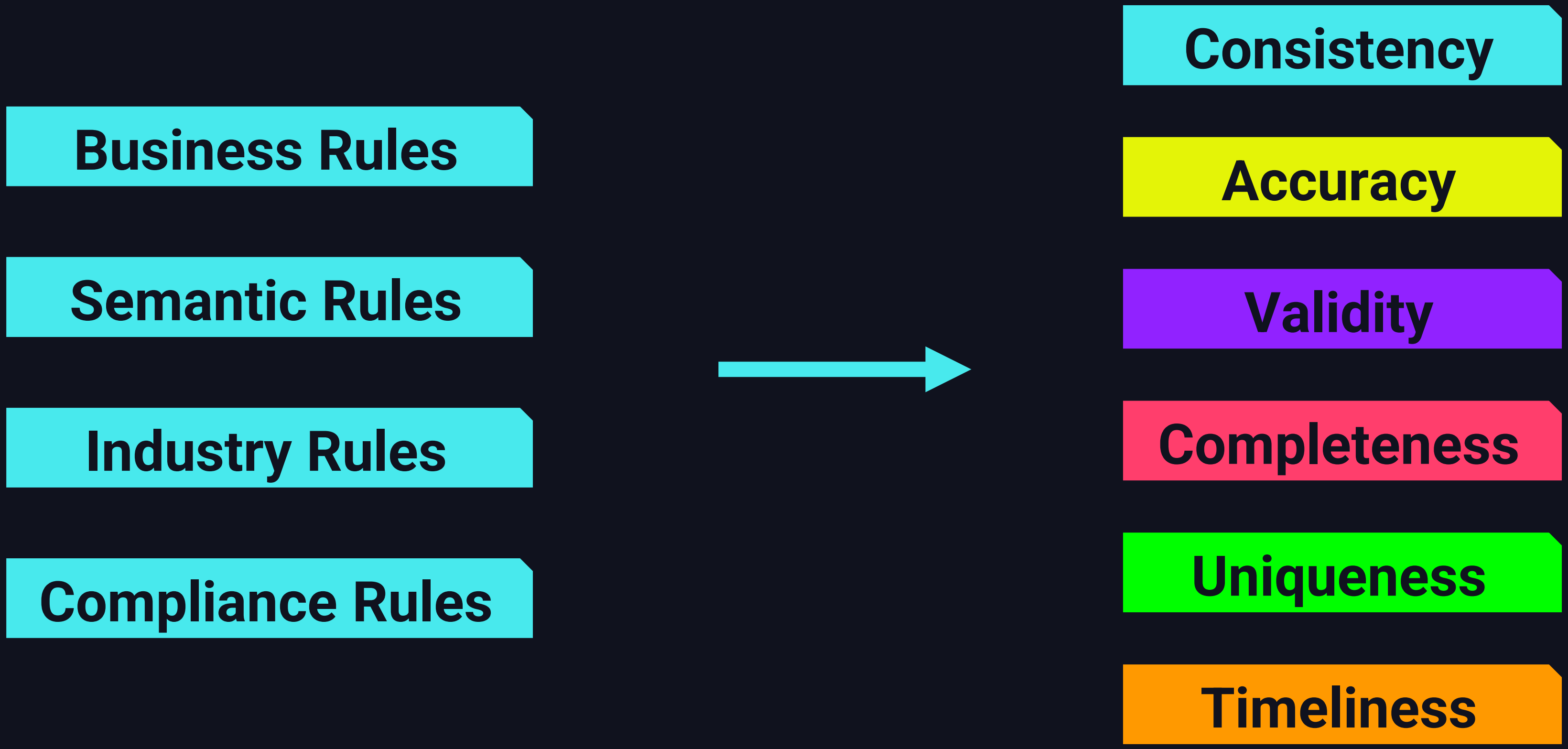
Requirement Gathering Techniques

Interviews
Surveys
Workshops
Obversations
Assessments

Rule Setting, Cleansing & Standardization

Data Quality Rule Setting

Different rules inform the requirements for six dimensions



Consistency

Detect inconsistencies

Values in different datasets are not conflicting

Causes

Data sources
Transformation logic
Imputation techniques
Metric defitions

Techniques

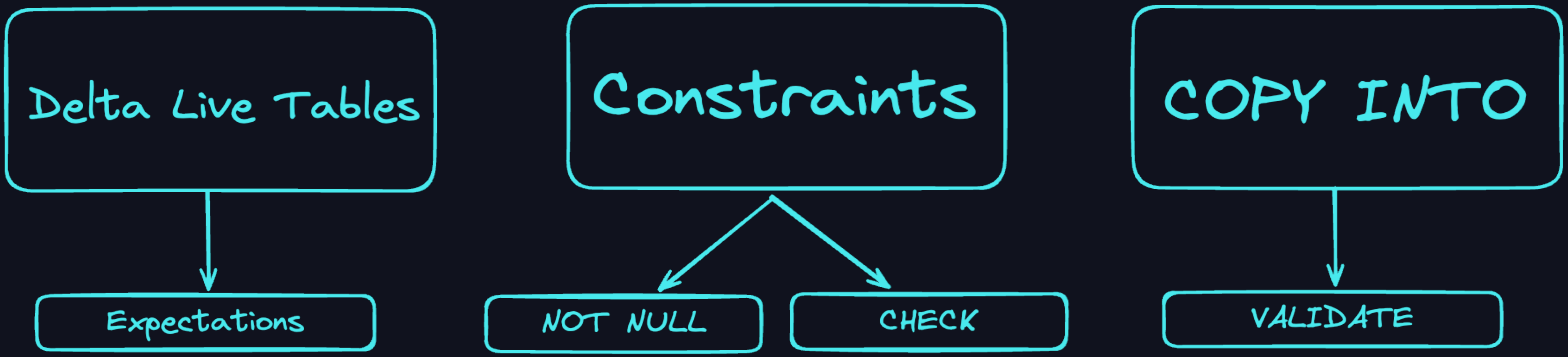
MERGE
JOIN

Detect inaccurate data

Set rules and identify inaccurate data

Accuracy

Validity

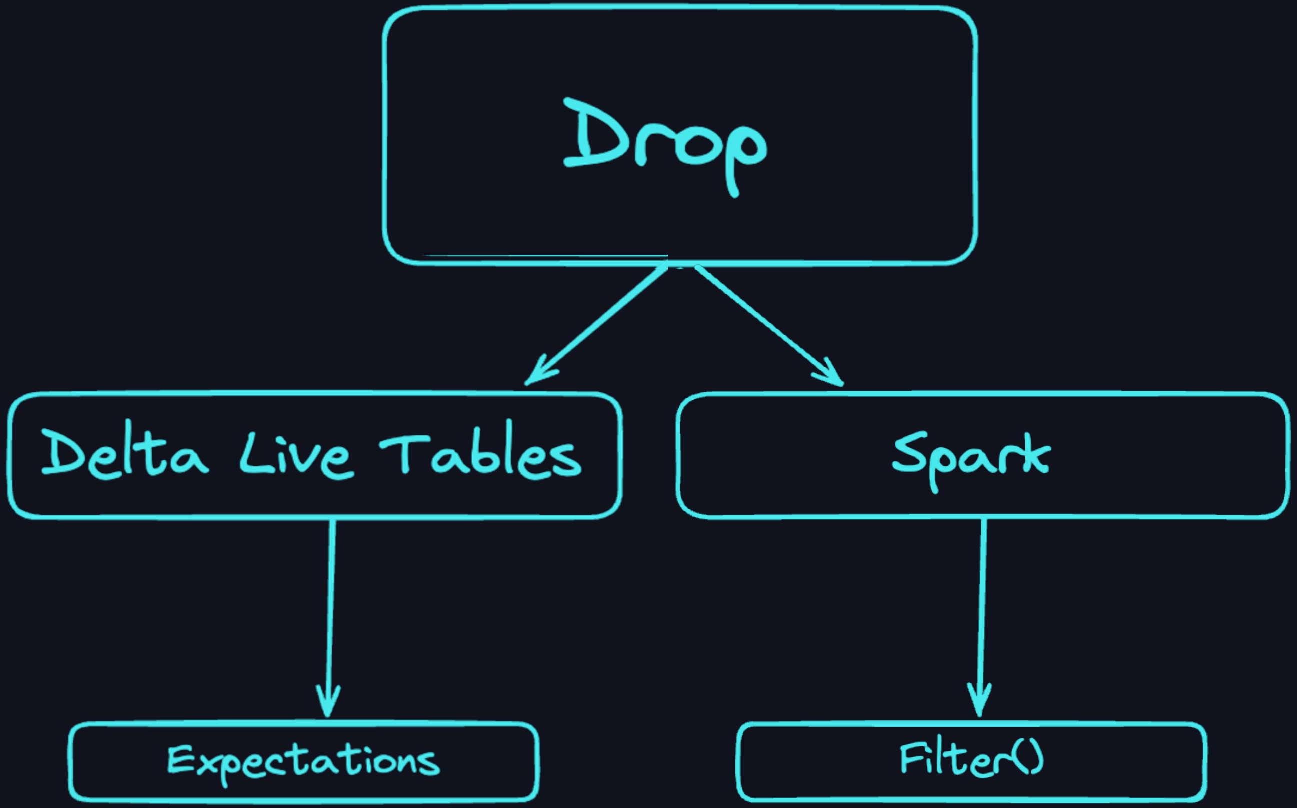


Accuracy

Consistency

Handle violations

Drop records failed against rule checks



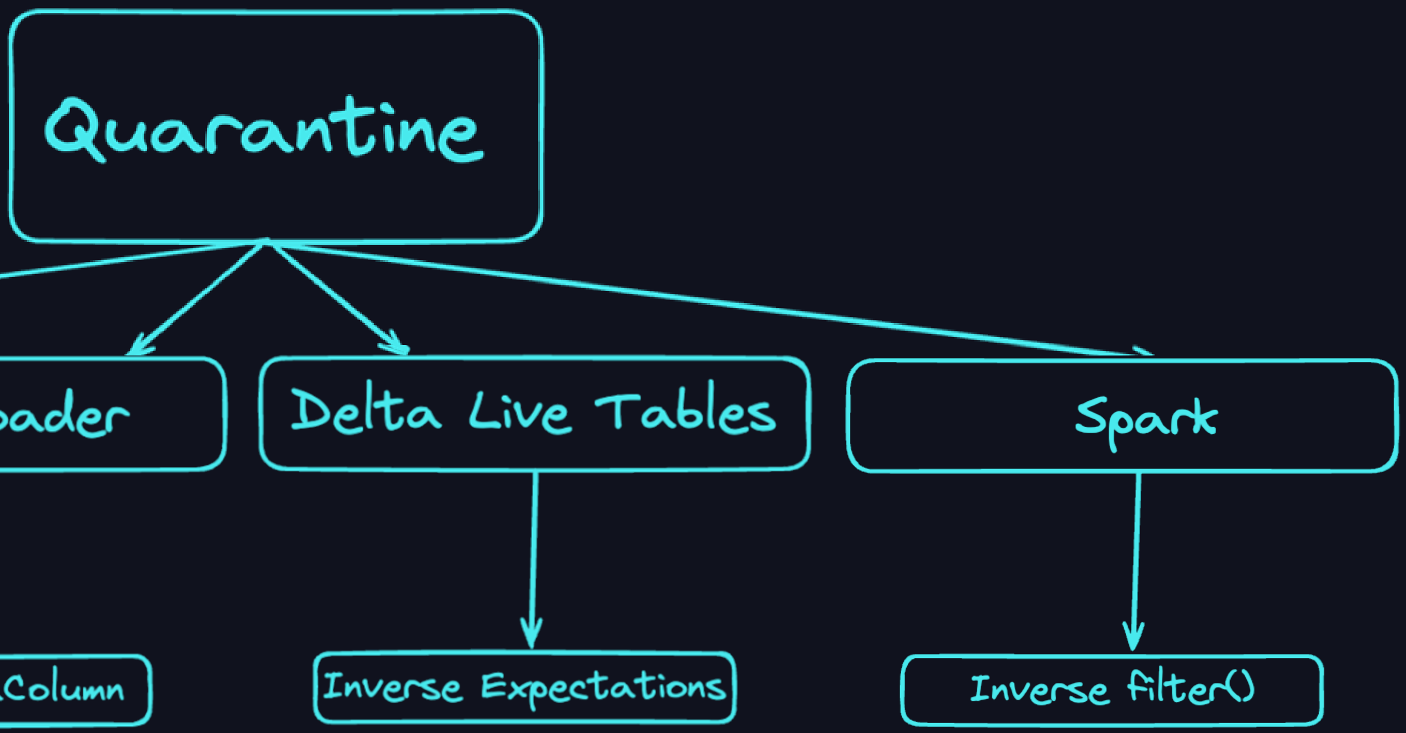
Handle violations (2)

Quarantine records failed against rule checks

Accuracy

Consistency

Validity



Validity

Schema Management

Schema Enforcement

Delta
Auto Loader

Schema Evolution

mergeSchema option
AutoMerge
Auto Loader

Schema Overwrite

overwriteSchema option

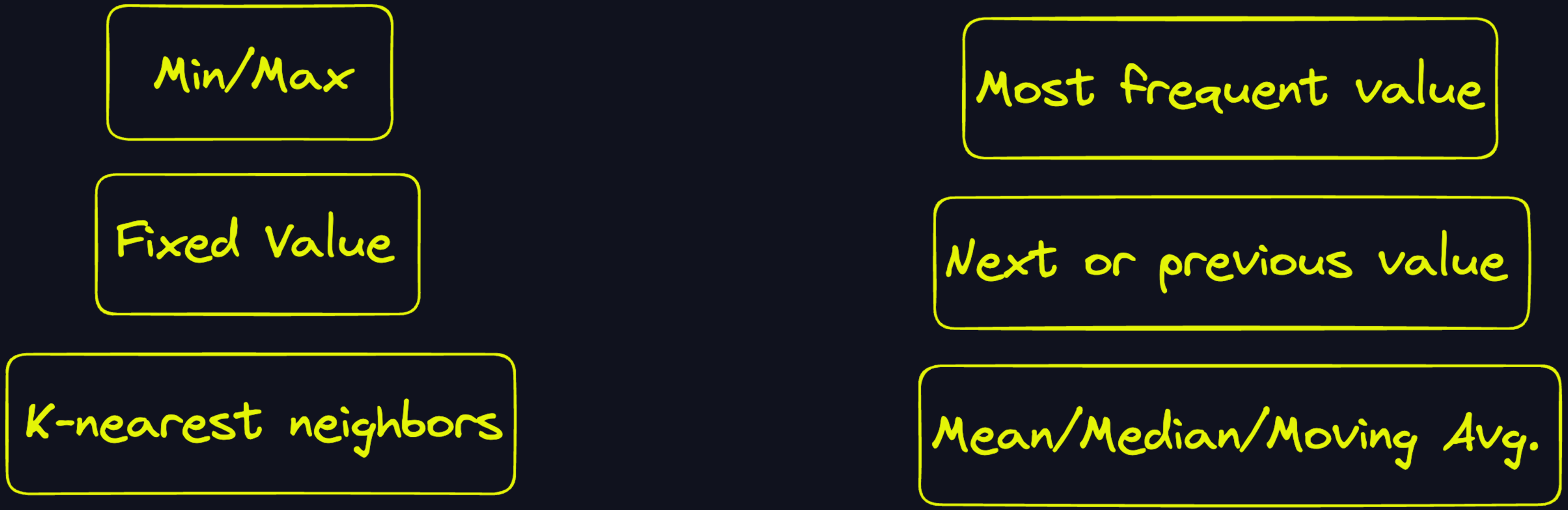
Schema Update

add new columns
add columns' comment
change column's ordering
replace columns
rename columns
drop columns
change columns' type

Impute Missing Data

Completeness

Impute missing records for data completeness



Min/Max

Fixed Value

K-nearest neighbors

Most frequent value

Next or previous value

Mean/Median/Moving Avg.

Drop Duplicates

Uniqueness

Remove duplicate records

Merge

distinct()

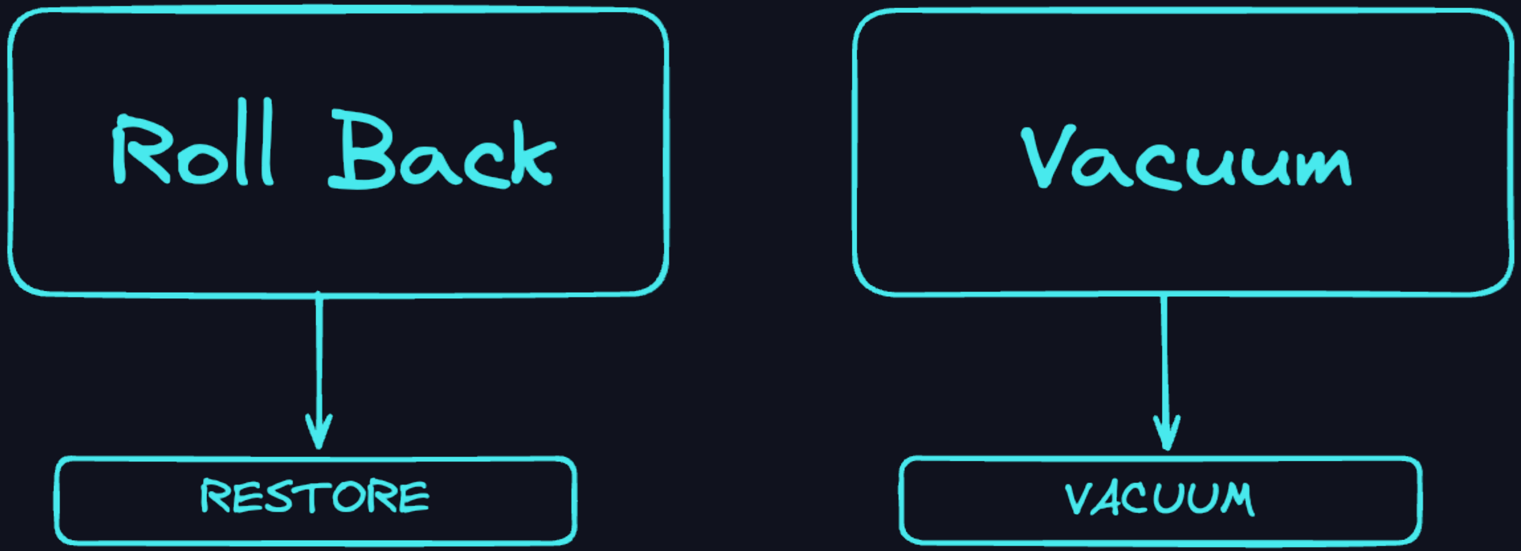
dropDuplicates

Ranking Window

Accuracy

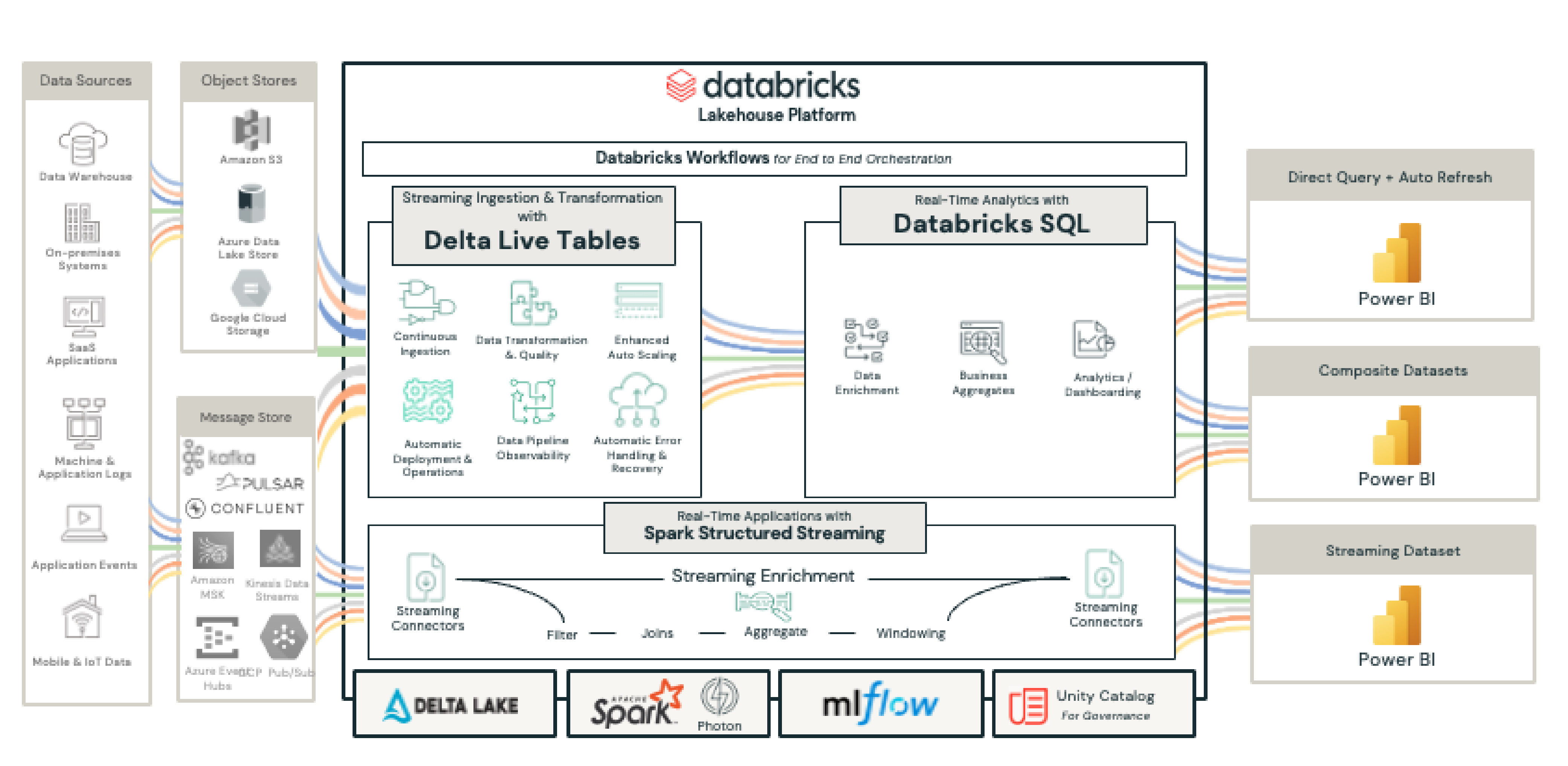
Rollback and Vacuum

Enable rollbacks but prevent access to invalid historical data



Structured Streaming and DLT

Timeliness



Data Sources

Data Warehouse

On-premises Systems

SaaS Applications

Machine & Application Logs

Application Events

Mobile & IoT Data

Object Stores

Amazon S3

Azure Data Lake Store

Google Cloud Storage

Message Store

kafka

PULSAR

CONFLUENT

Amazon MSK

Kinesis Data Streams

Azure EventGCP Pub/Sub Hubs

databricks

Lakehouse Platform

Databricks Workflows *for End to End Orchestration*

Streaming Ingestion & Transformation with

Delta Live Tables

Continuous Ingestion

Data Transformation & Quality

Enhanced Auto Scaling

Automatic Deployment & Operations

Data Pipeline Observability

Automatic Error Handling & Recovery

Real-Time Analytics with

Databricks SQL

Data Enrichment

Business Aggregates

Analytics / Dashboarding

Real-Time Applications with

Spark Structured Streaming

Streaming Connectors

Streaming Enrichment

Filter — Joins — Aggregate — Windowing

Streaming Connectors

DELTA LAKE

Spark

Photon

mlflow

Unity Catalog

For Governance

Direct Query + Auto Refresh

Power BI

Composite Datasets

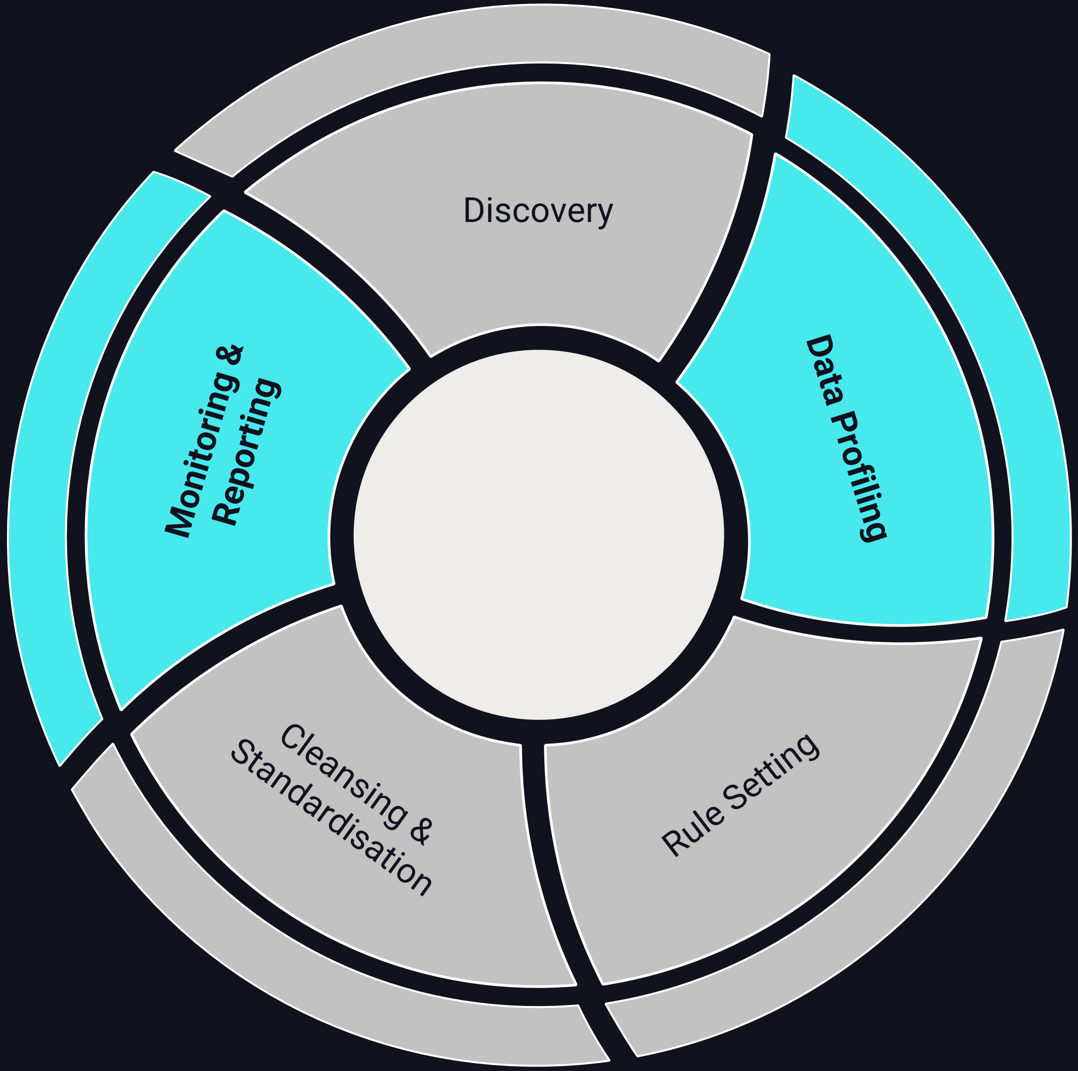
Power BI

Streaming Dataset

Power BI

Walk

Data Quality Management Lifecycle

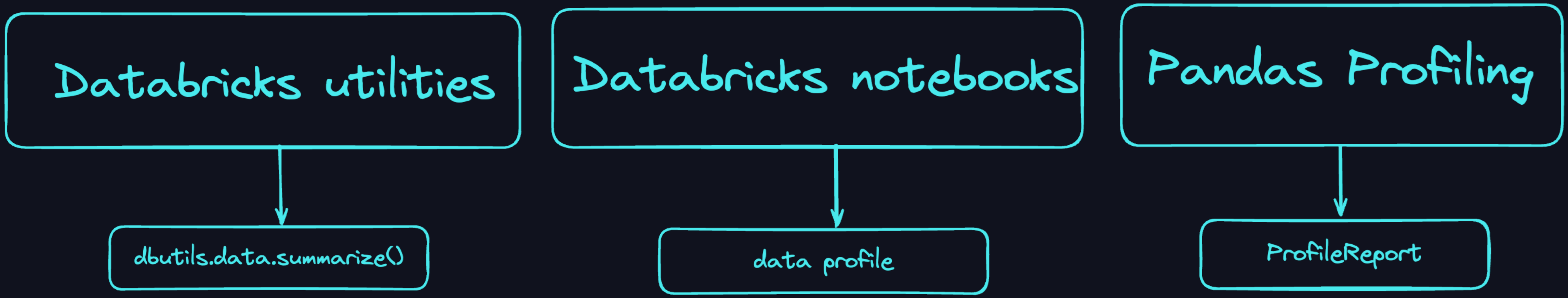


Data Profiling

Data Profiling

All Dimensions

Identify trends and patterns in data and spot outliers and anomalies



Databricks utilities

dbutils.data.summarize()

Databricks notebooks

data profile

Pandas Profiling

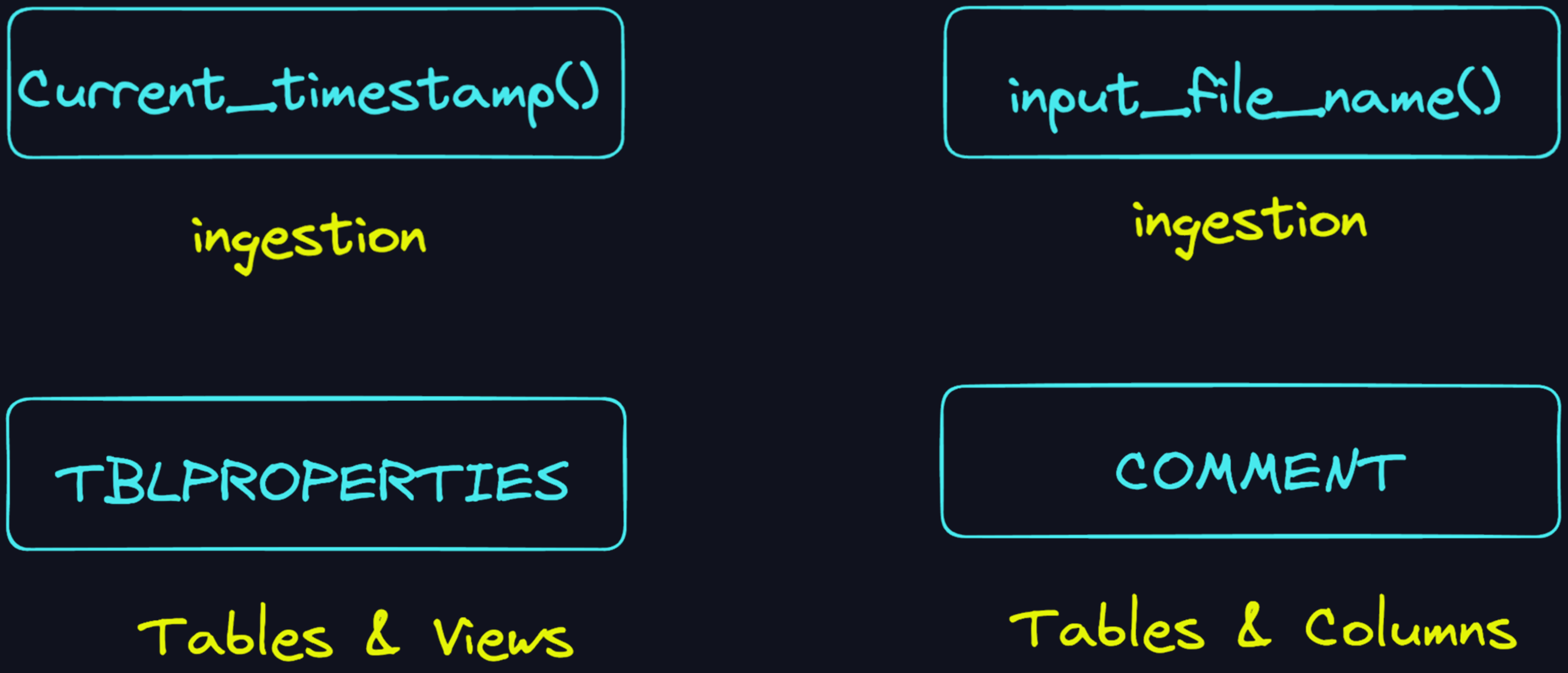
ProfileReport

Metadata Management

Data Augmentation

All Dimensions

Add additional metadata to facilitate discovery



Current_timestamp()

ingestion

input_file_name()

ingestion

TBLPROPERTIES

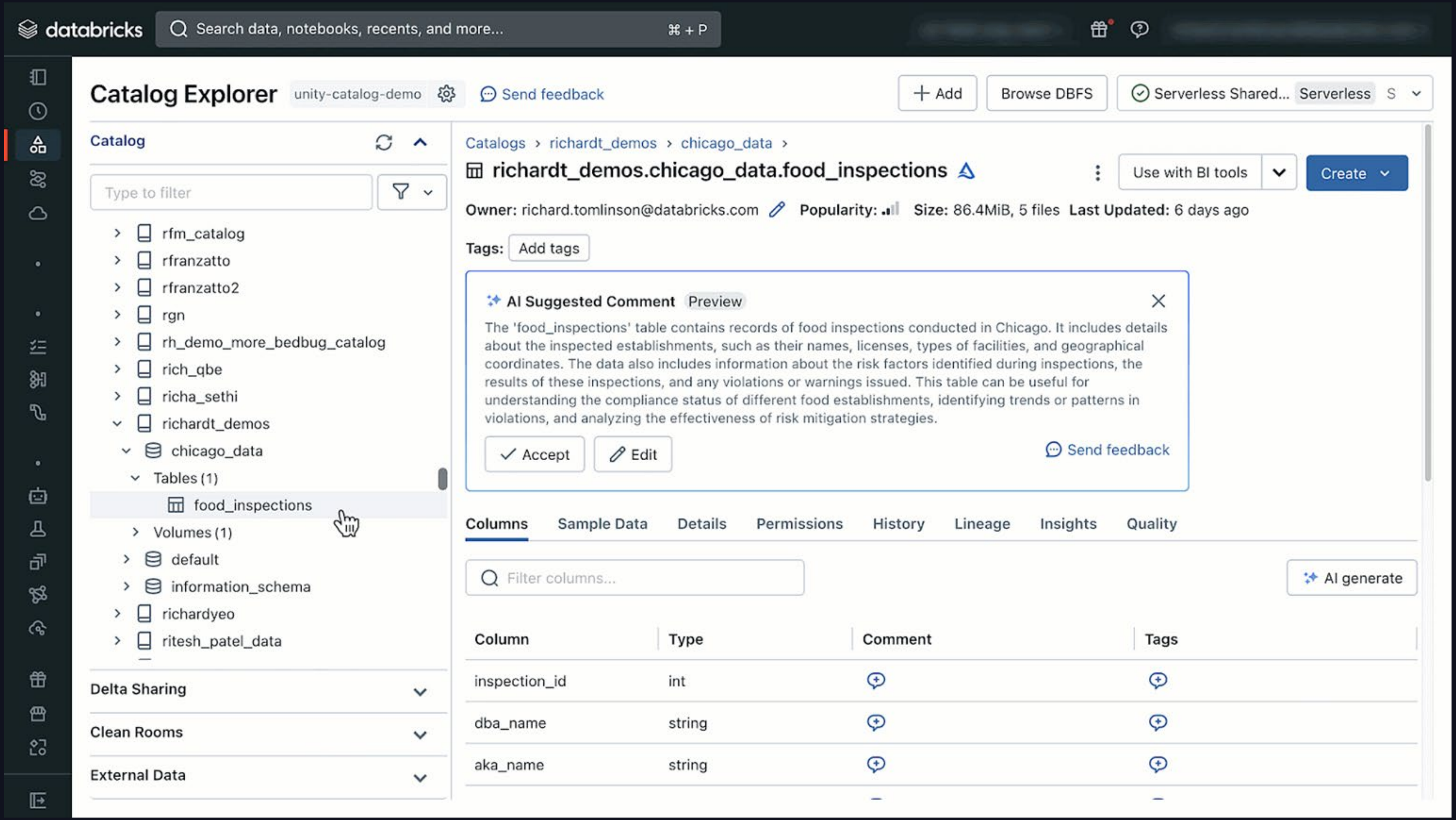
Tables & Views

COMMENT

Tables & Columns

AI-generated Descriptions and Comments

All Dimensions



richardt_demos.chicago_data.food_inspections

Owner: richard.tomlinson@databricks.com Popularity: Size: 86.4MiB, 5 files Last Updated: 6 days ago

Tags:

AI Suggested Comment Preview

The 'food_inspections' table contains records of food inspections conducted in Chicago. It includes details about the inspected establishments, such as their names, licenses, types of facilities, and geographical coordinates. The data also includes information about the risk factors identified during inspections, the results of these inspections, and any violations or warnings issued. This table can be useful for understanding the compliance status of different food establishments, identifying trends or patterns in violations, and analyzing the effectiveness of risk mitigation strategies.

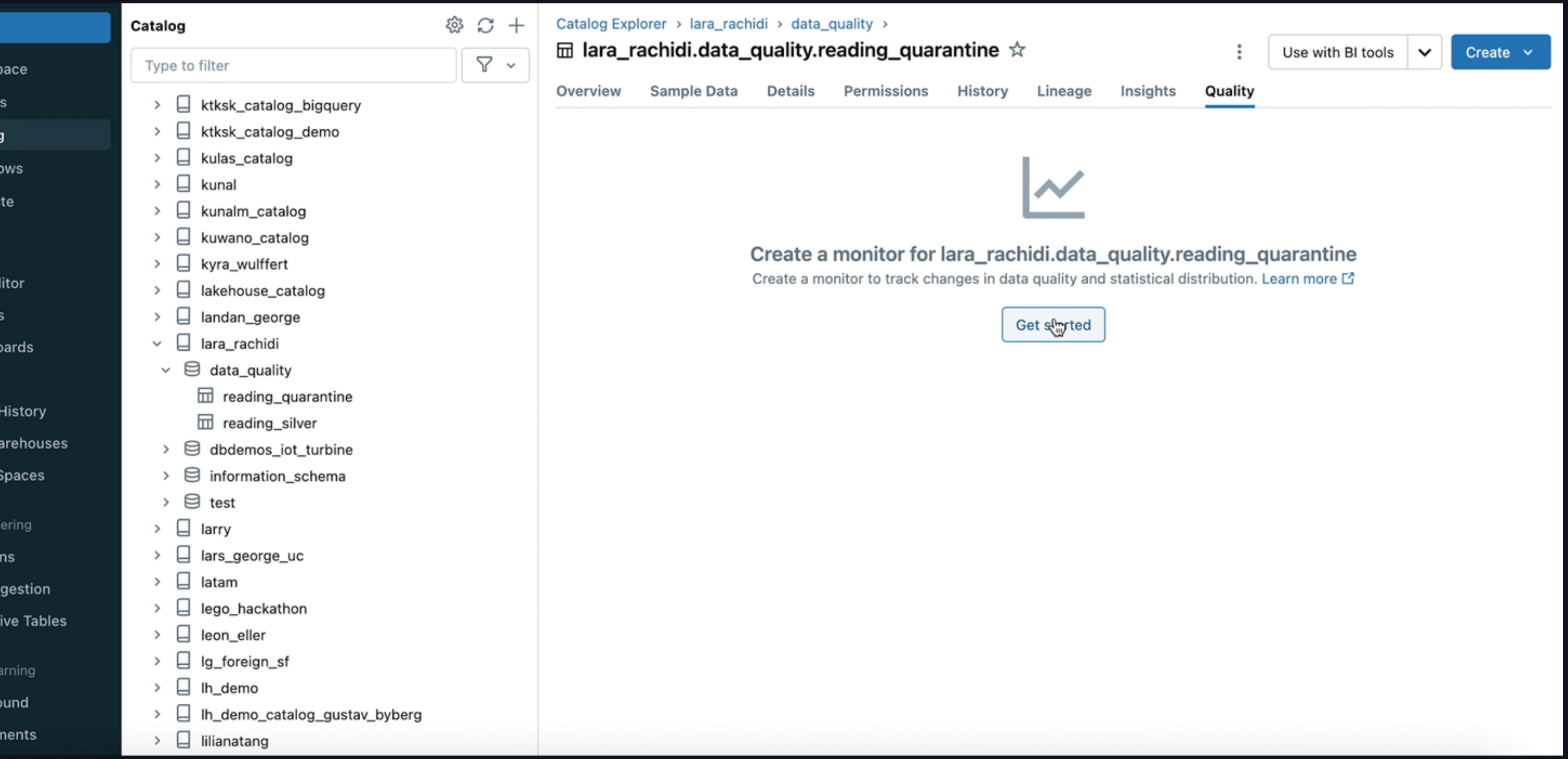
Column	Type	Comment	Tags
inspection_id	int		
dba_name	string		
aka_name	string		

DATA+AI SUMMIT

Monitoring & Reporting

Lakehouse Monitoring

Monitor data quality and model quality over time



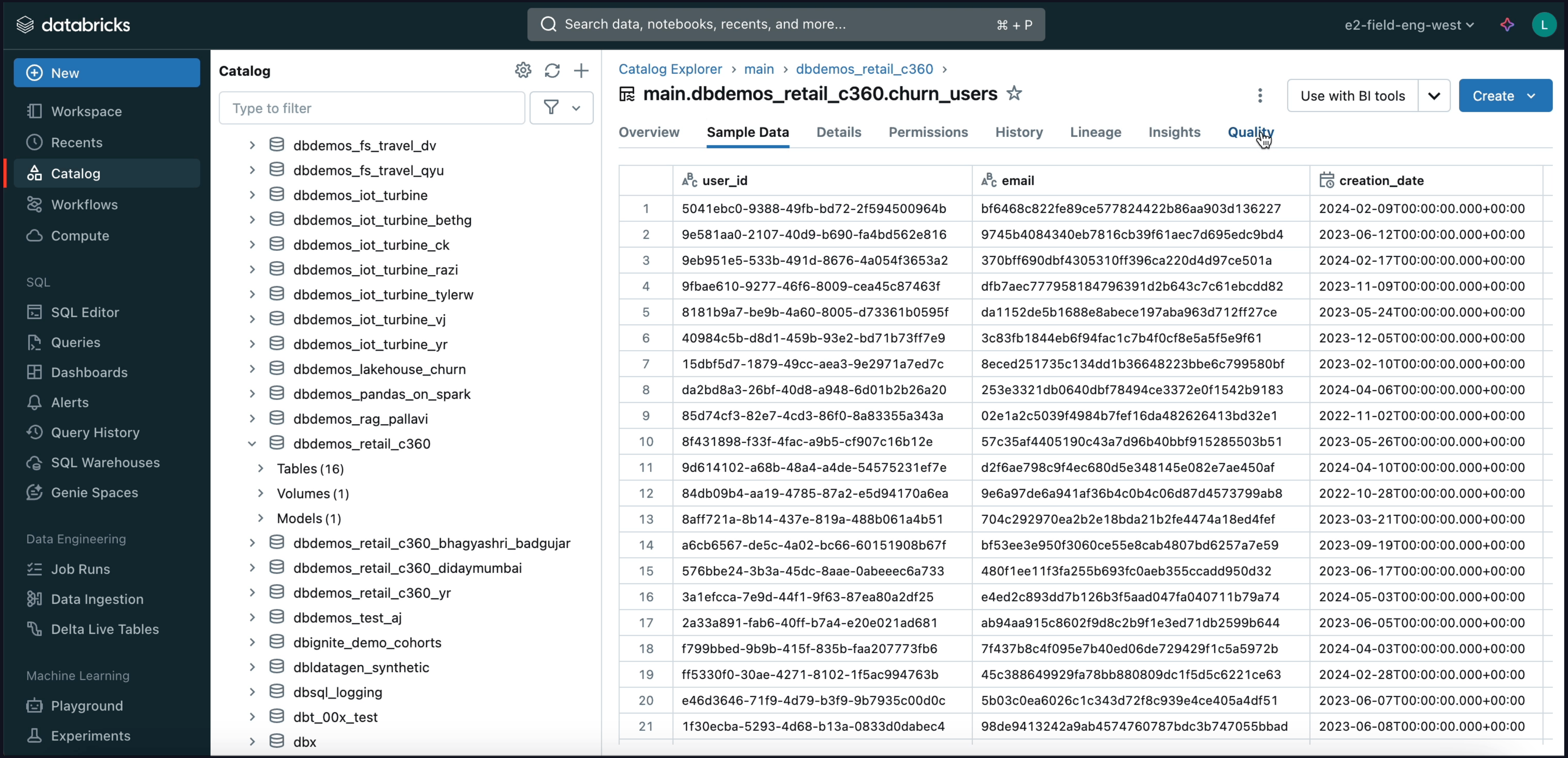
Create a monitor for lara_rachidi.data_quality.reading_quarantine

Create a monitor to track changes in data quality and statistical distribution. Learn more

- □ Time Series
- □ Model Inference
- □ Snapshot
- □ Optional: baseline table

Lakehouse Monitoring

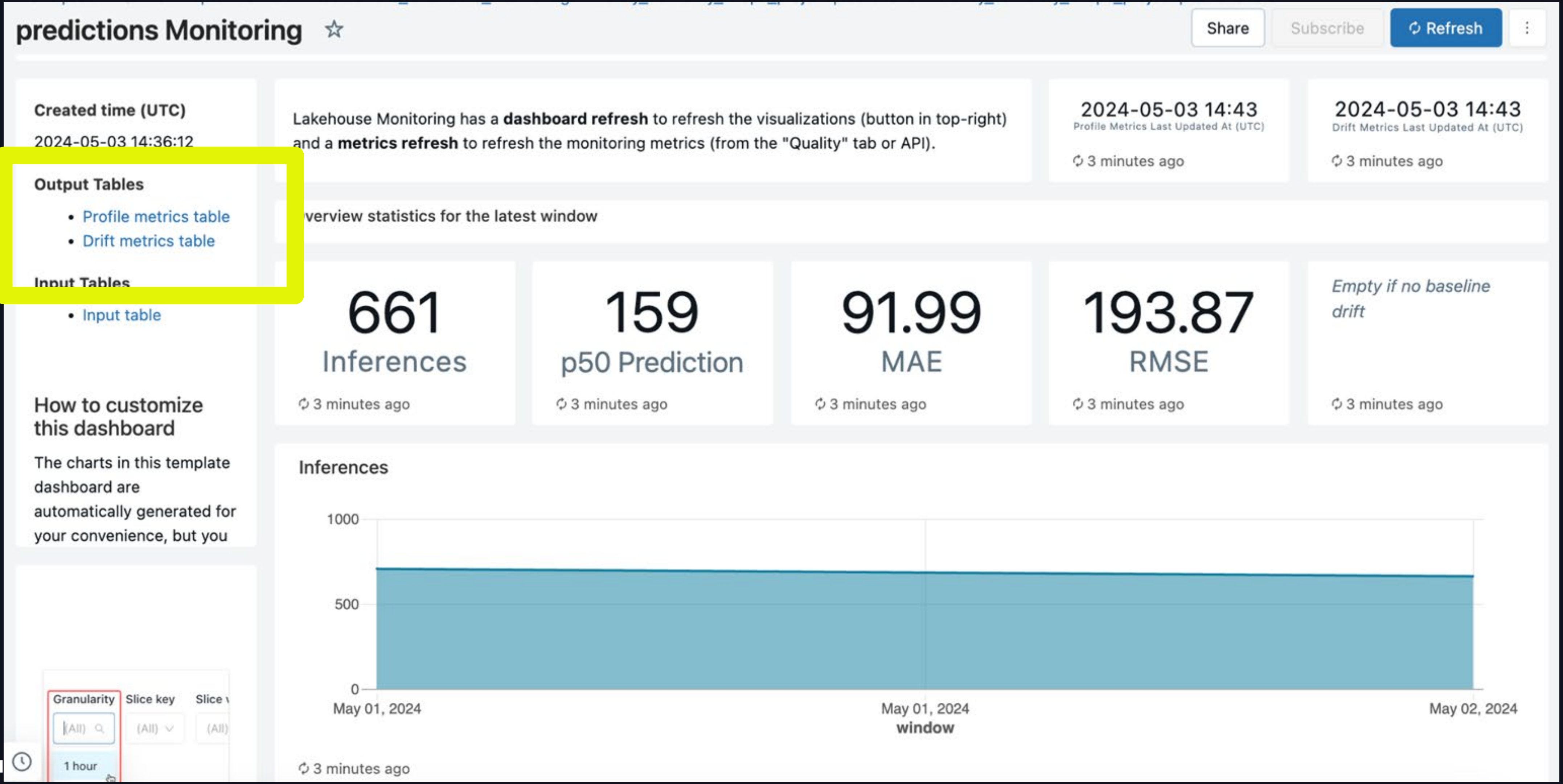
Display automatically computed metrics in a fully customizable dashboard



	user_id	email	creation_date
1	5041ebc0-9388-49fb-bd72-2f594500964b	bf6468c822fe89ce577824422b86aa903d136227	2024-02-09T00:00:00.000+00:00
2	9e581aa0-2107-40d9-b690-fa4bd562e816	9745b4084340eb7816cb39f61aec7d695edc9bd4	2023-06-12T00:00:00.000+00:00
3	9eb951e5-533b-491d-8676-4a054f3653a2	370bff690dbf4305310ff396ca220d4d97ce501a	2024-02-17T00:00:00.000+00:00
4	9fbae610-9277-46f6-8009-cea45c87463f	dfb7aec777958184796391d2b643c7c61ebcdd82	2023-11-09T00:00:00.000+00:00
5	8181b9a7-be9b-4a60-8005-d73361b0595f	da1152de5b1688e8abece197aba963d712ff27ce	2023-05-24T00:00:00.000+00:00
6	40984c5b-d8d1-459b-93e2-bd71b73ff7e9	3c83fb1844eb6f94fac1c7b4f0cf8e5a5f5e9f61	2023-12-05T00:00:00.000+00:00
7	15dbf5d7-1879-49cc-aea3-9e2971a7ed7c	8eced251735c134dd1b36648223bbe6c799580bf	2023-02-10T00:00:00.000+00:00
8	da2bd8a3-26bf-40d8-a948-6d01b2b26a20	253e3321db0640dbf78494ce3372e0f1542b9183	2024-04-06T00:00:00.000+00:00
9	85d74cf3-82e7-4cd3-86f0-8a83355a343a	02e1a2c5039f4984b7fef16da482626413bd32e1	2022-11-02T00:00:00.000+00:00
10	8f431898-f33f-4fac-a9b5-cf907c16b12e	57c35af4405190c43a7d96b40bbf915285503b51	2023-05-26T00:00:00.000+00:00
11	9d614102-a68b-48a4-a4de-54575231ef7e	d2f6ae798c9f4ec680d5e348145e082e7ae450af	2024-04-10T00:00:00.000+00:00
12	84db09b4-aa19-4785-87a2-e5d94170a6ea	9e6a97de6a941af36b4c0b4c06d87d4573799ab8	2022-10-28T00:00:00.000+00:00
13	8aff721a-8b14-437e-819a-488b061a4b51	704c292970ea2b2e18bda21b2fe4474a18ed4fef	2023-03-21T00:00:00.000+00:00
14	a6cb6567-de5c-4a02-bc66-60151908b67f	bf53ee3e950f3060ce55e8cab4807bd6257a7e59	2023-09-19T00:00:00.000+00:00
15	576bbe24-3b3a-45dc-8aae-0abeeec6a733	480f1ee11f3fa255b693fc0aeb355ccadd950d32	2023-06-17T00:00:00.000+00:00
16	3a1efcca-7e9d-44f1-9f63-87ea80a2df25	e4ed2c893dd7b126b3f5aad047fa040711b79a74	2024-05-03T00:00:00.000+00:00
17	2a33a891-fab6-40ff-b7a4-e20e021ad681	ab94aa915c8602f9d8c2b9f1e3ed71db2599b644	2023-06-05T00:00:00.000+00:00
18	f799bbed-9b9b-415f-835b-faa207773fb6	7f437b8c4f095e7b40ed06de729429f1c5a5972b	2024-04-03T00:00:00.000+00:00
19	ff5330f0-30ae-4271-8102-1f5ac994763b	45c388649929fa78bb880809dc1f5d5c6221ce63	2024-02-28T00:00:00.000+00:00
20	e46d3646-71f9-4d79-b3f9-9b7935c00d0c	5b03c0ea6026c1c343d72f8c939e4ce405a4df51	2023-06-07T00:00:00.000+00:00
21	1f30ecba-5293-4d68-b13a-0833d0dabec4	98de9413242a9ab4574760787bdc3b747055bbad	2023-06-08T00:00:00.000+00:00

Dashboard and metrics tables

Summary statistics and statistics related to data drift over time stored in UC



predictions Monitoring ☆

Created time (UTC)

2024-05-03 14:36:12

Output Tables

- Profile metrics table
- Drift metrics table

Input Tables

- Input table

How to customize this dashboard

The charts in this template dashboard are automatically generated for your convenience, but you

Lakehouse Monitoring has a **dashboard refresh** to refresh the visualizations (button in top-right) and a **metrics refresh** to refresh the monitoring metrics (from the "Quality" tab or API).

2024-05-03 14:43
Profile Metrics Last Updated At (UTC)
3 minutes ago

2024-05-03 14:43
Drift Metrics Last Updated At (UTC)
3 minutes ago

verview statistics for the latest window

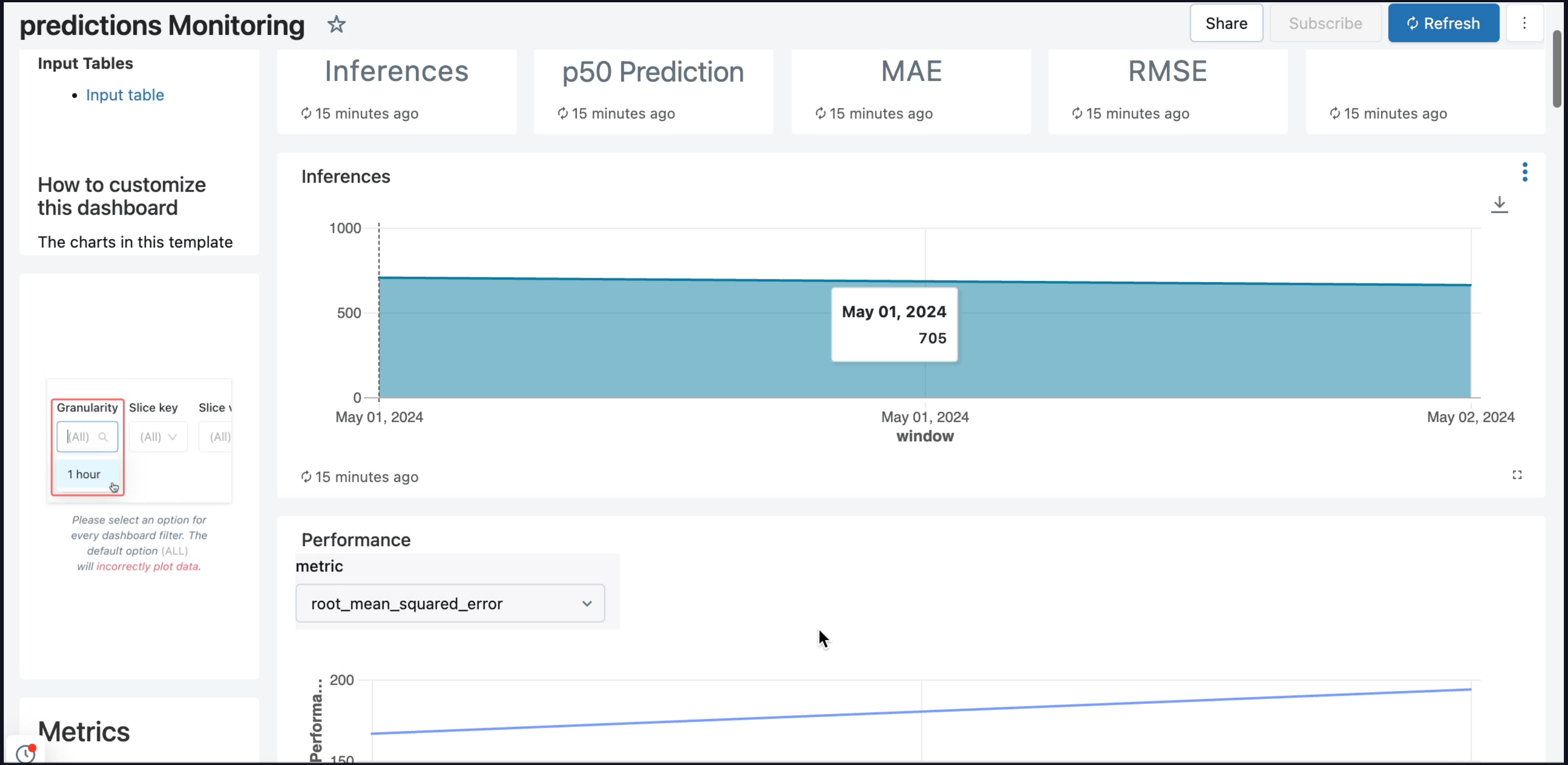
661	159	91.99	193.87	*Empty if no baseline drift*
Inferences	p50 Prediction	MAE	RMSE	
3 minutes ago	3 minutes ago	3 minutes ago	3 minutes ago	3 minutes ago

Inferences

3 minutes ago

Alerts with Lakehouse Monitoring

Set up alerts for reporting



predictions Monitoring ☆

Share | Subscribe | Refresh

Input Tables

- Input table

How to customize this dashboard

The charts in this template

Granularity: (All) — 1 hour | Slice key: (All) | Slice v: (All)

Please select an option for every dashboard filter. The default option (ALL) will incorrectly plot data.

Metrics

Inferences — 15 minutes ago | p50 Prediction — 15 minutes ago | MAE — 15 minutes ago | RMSE — 15 minutes ago | 15 minutes ago

Inferences

15 minutes ago

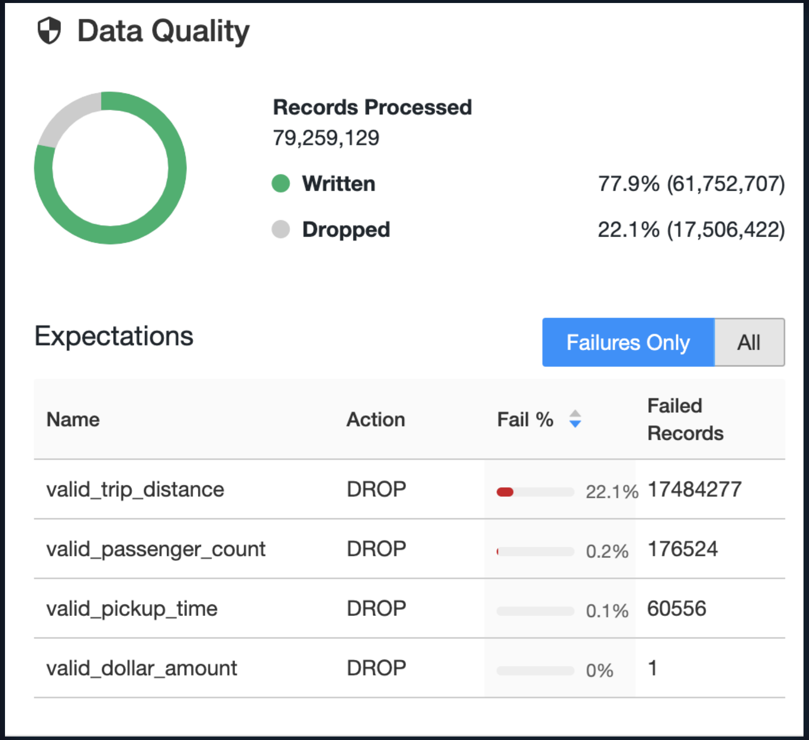
Performance

metric: root_mean_squared_error

Delta Live Tables Monitoring

All Dimensions

Data quality metrics in DLT are captured in event logs and can be reported using DBSQL



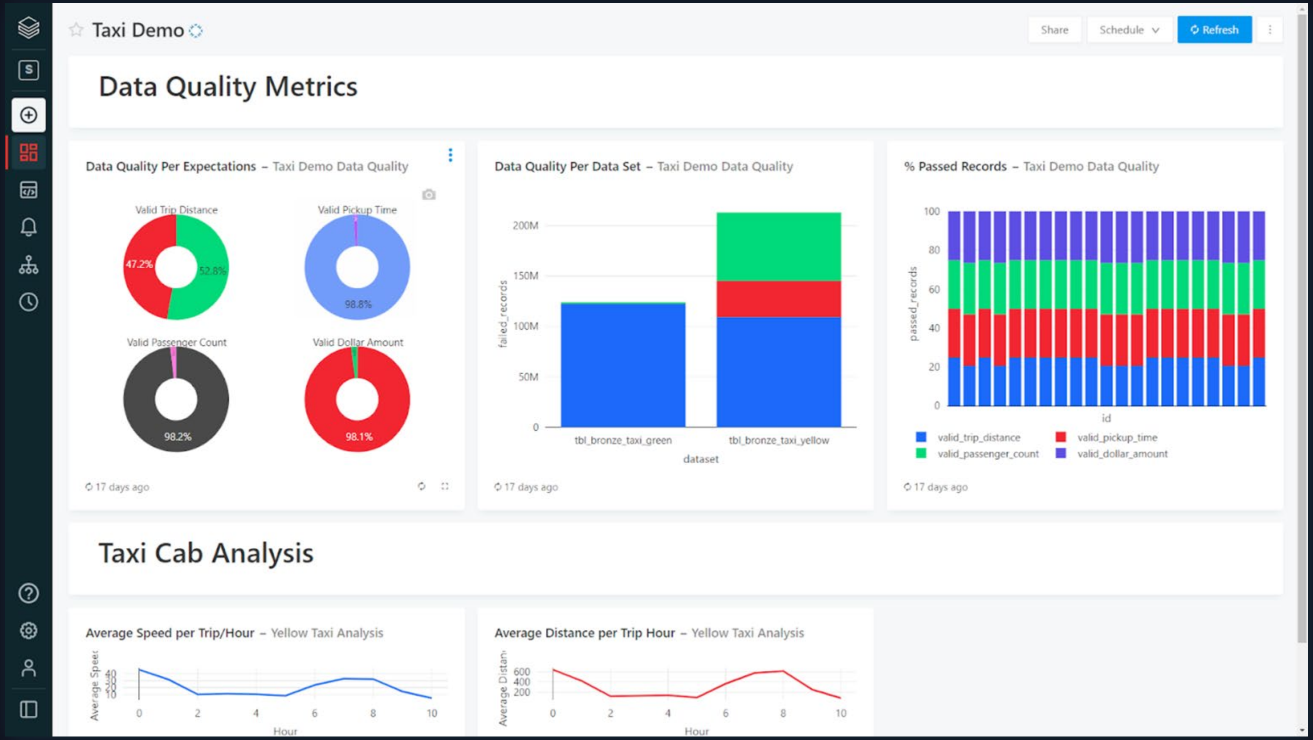
Data Quality

Records Processed
79,259,129

● **Written**	77.9% (61,752,707)
● **Dropped**	22.1% (17,506,422)

Expectations

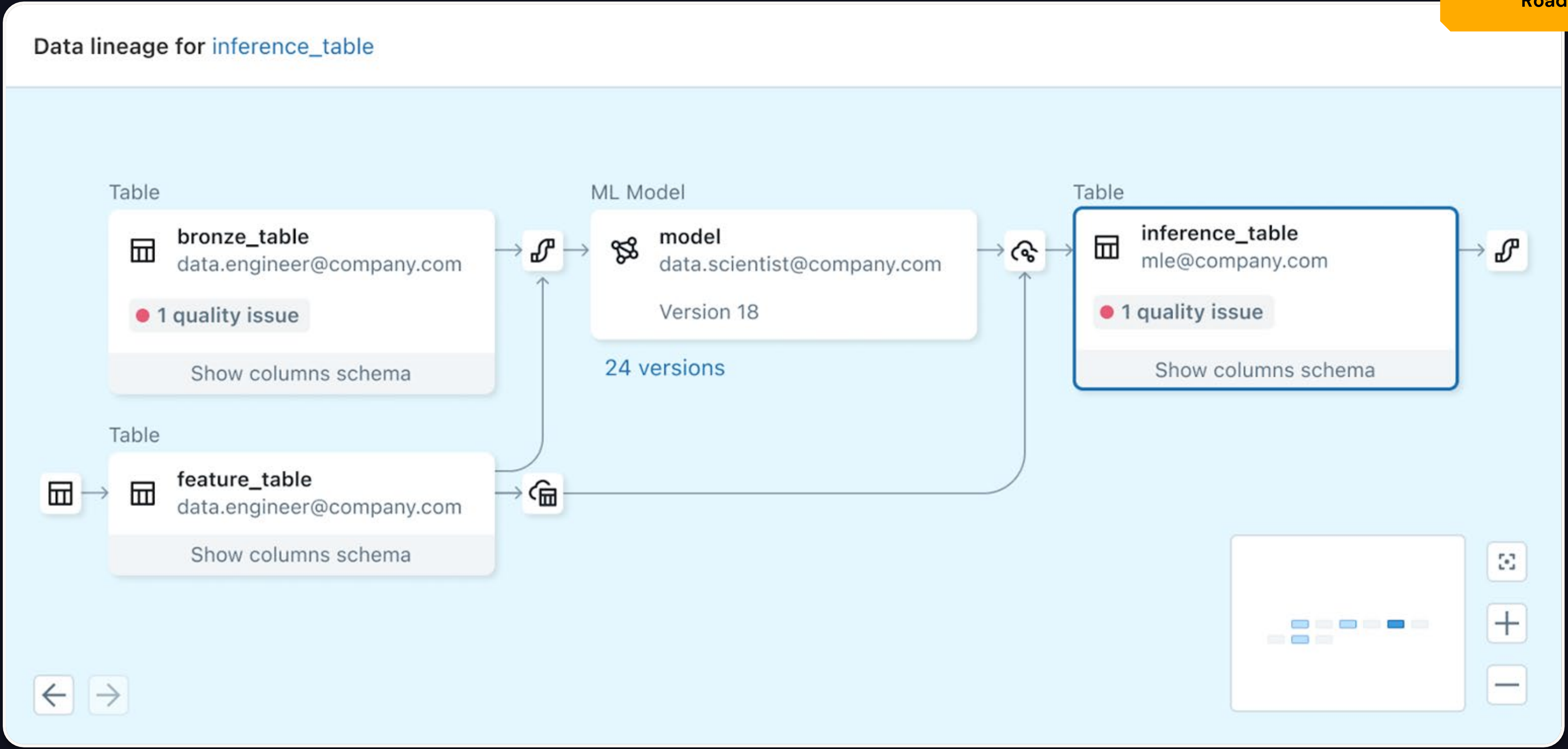
Name	Action	Fail %	Failed Records
valid_trip_distance	DROP	22.1%	17484277
valid_passenger_count	DROP	0.2%	176524
valid_pickup_time	DROP	0.1%	60556
valid_dollar_amount	DROP	0%	1



Root Cause Analysis leveraging Data Lineage

All Dimensions

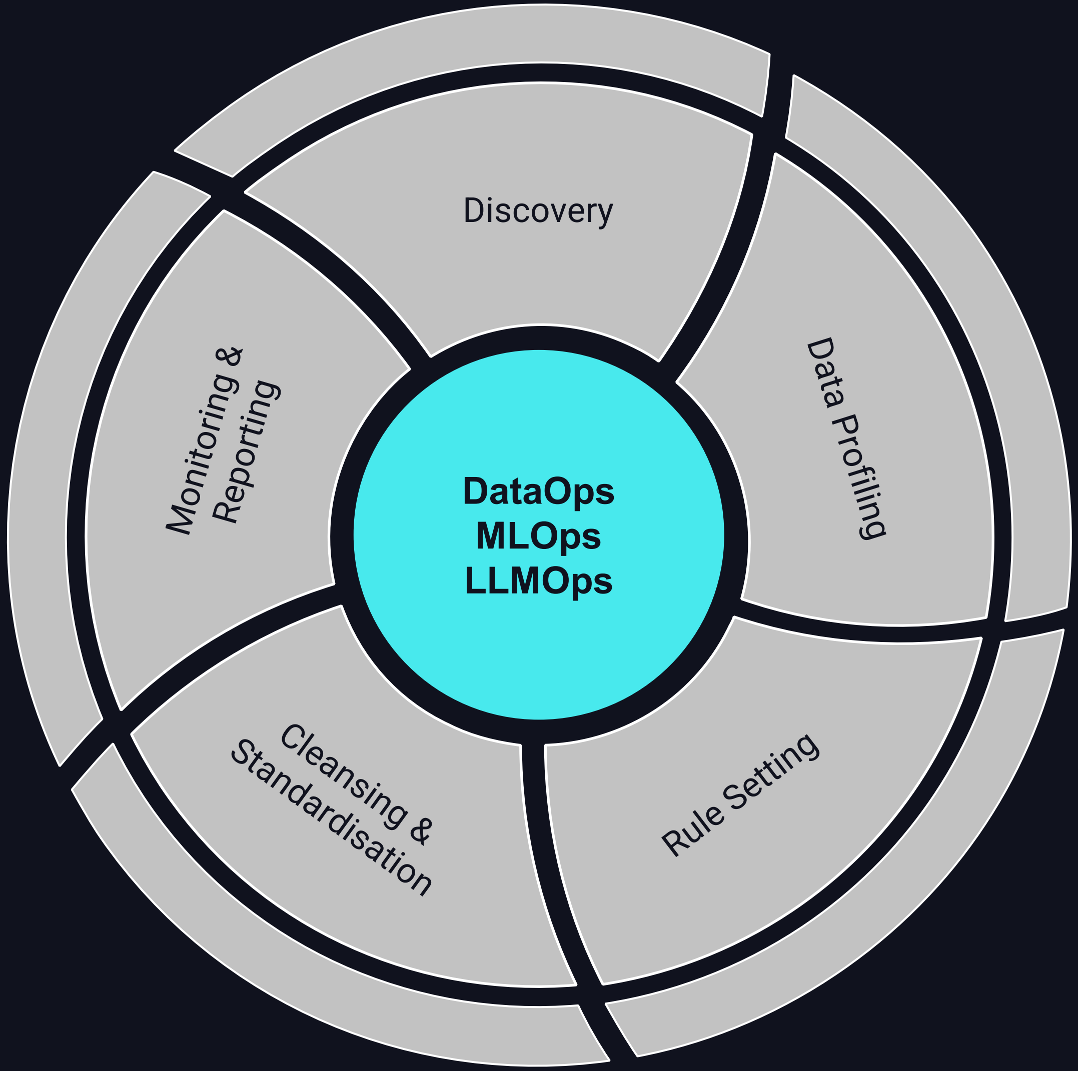
Roadmap



Data lineage for inference_table

Run

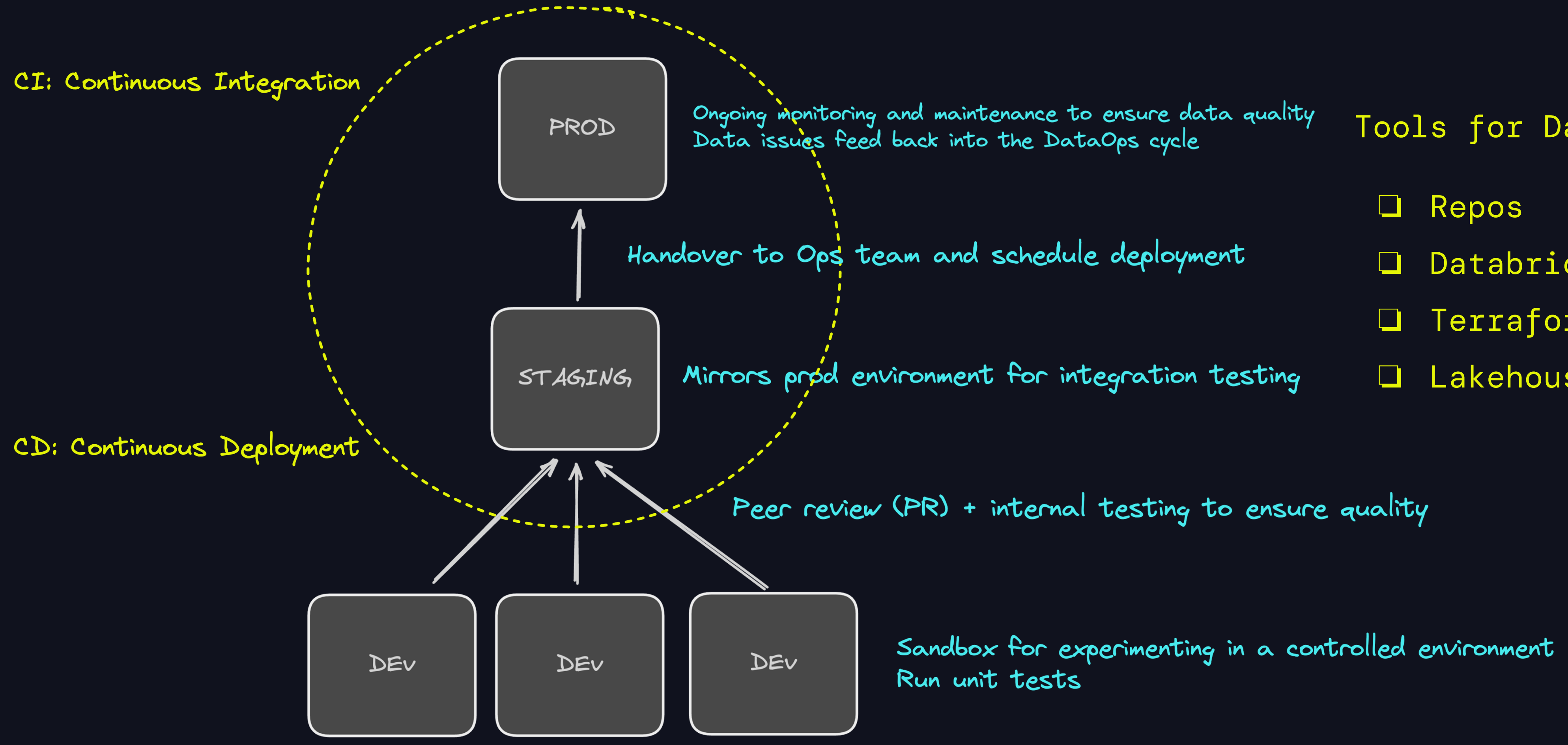
Data Quality Management Lifecycle



DataOps, MLOps, LLMOps

Ensuring high quality with CI/CD

Manual processes are error -prone, impacting data quality



Ongoing monitoring and maintenance to ensure data quality
Data issues feed back into the DataOps cycle

Handover to Ops team and schedule deployment

Mirrors prod environment for integration testing

Peer review (PR) + internal testing to ensure quality

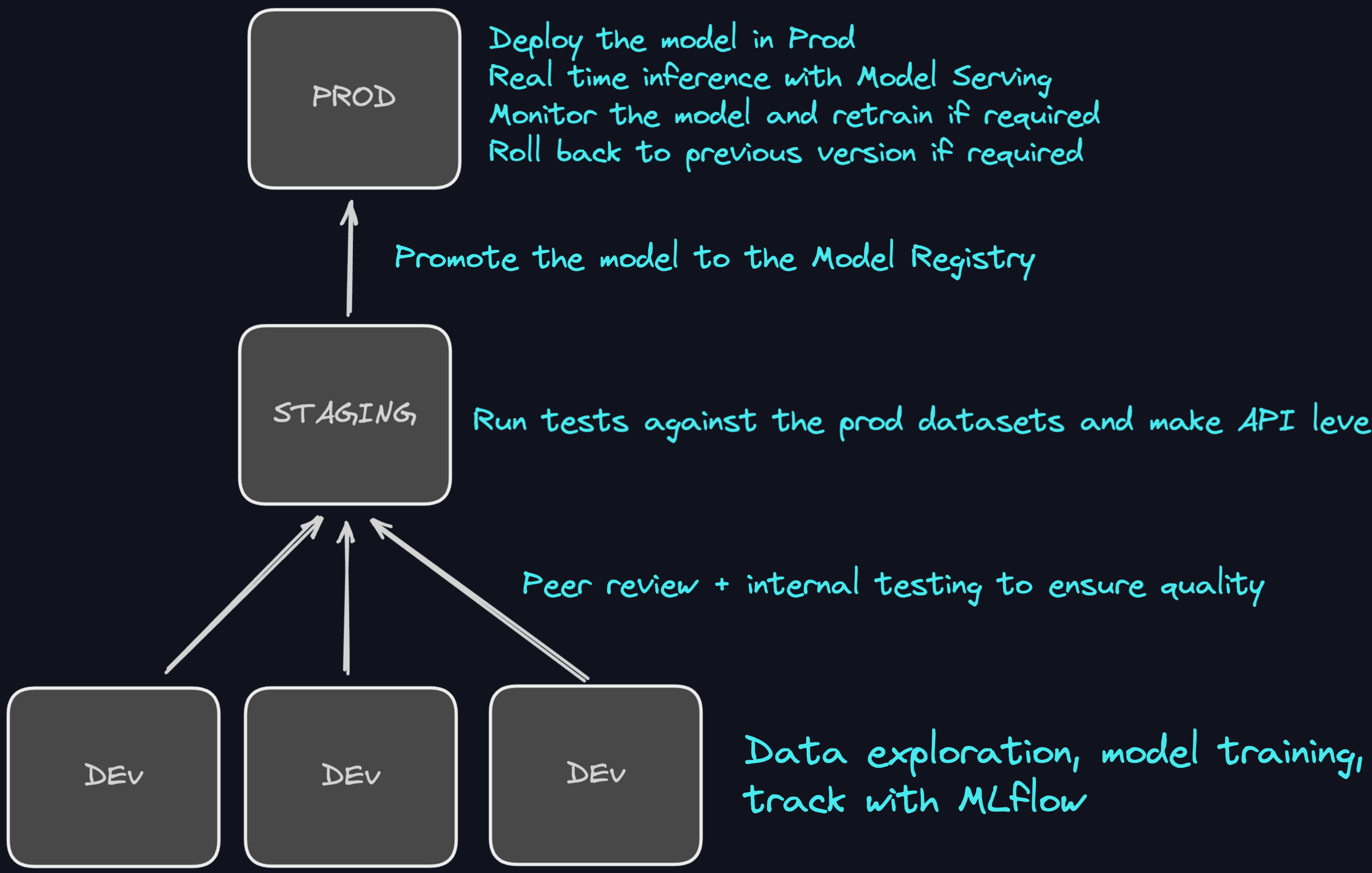
Sandbox for experimenting in a controlled environment
Run unit tests

Tools for DataOps:

- ❑ Repos
- ❑ Databricks Asset Bundles
- ❑ Terraform
- ❑ Lakehouse Monitoring

MLOps to ensure quality ML models

Progress your code, data, models and pipelines through quality thresholds



Deploy the model in Prod
Real time inference with Model Serving
Monitor the model and retrain if required
Roll back to previous version if required

Promote the model to the Model Registry

Run tests against the prod datasets and make API level tests

Peer review + internal testing to ensure quality

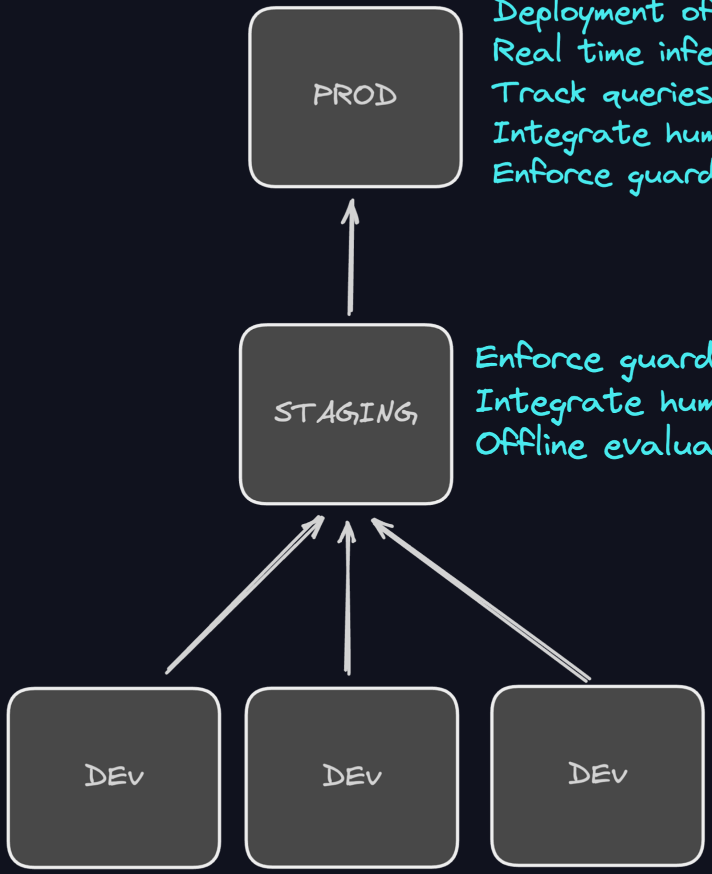
Data exploration, model training,
track with MLflow

Tools:

❑ Repos
❑ Databricks Asset Bundles
❑ Terraform
❑ Lakehouse Monitoring
❑ MLflow

LLMOps to ensure quality GenAI apps

Progress your code, data, models and pipelines through quality thresholds



Deployment of LLM pipelines packaged using MLflow
Real time inference with Model Serving or AI Gateway
Track queries and responses using System Tables
Integrate human feedback (offline & online)
Enforce guardrails, e.g. against toxic responses

Enforce guardrails, e.g. against toxic responses
Integrate human feedback (offline & online)
Offline evaluation using LLM as a judge

Store prompt and pipelines in a chain
MLflow: prompt iteration, pipeline packaging, tracking queries and responses
Enforce guardrails, e.g. filter toxic responses
Offline eval by integrating human feedback or using LLM as a judge
Low quality? Consider RAG or fine tuning instead of prompt engineering

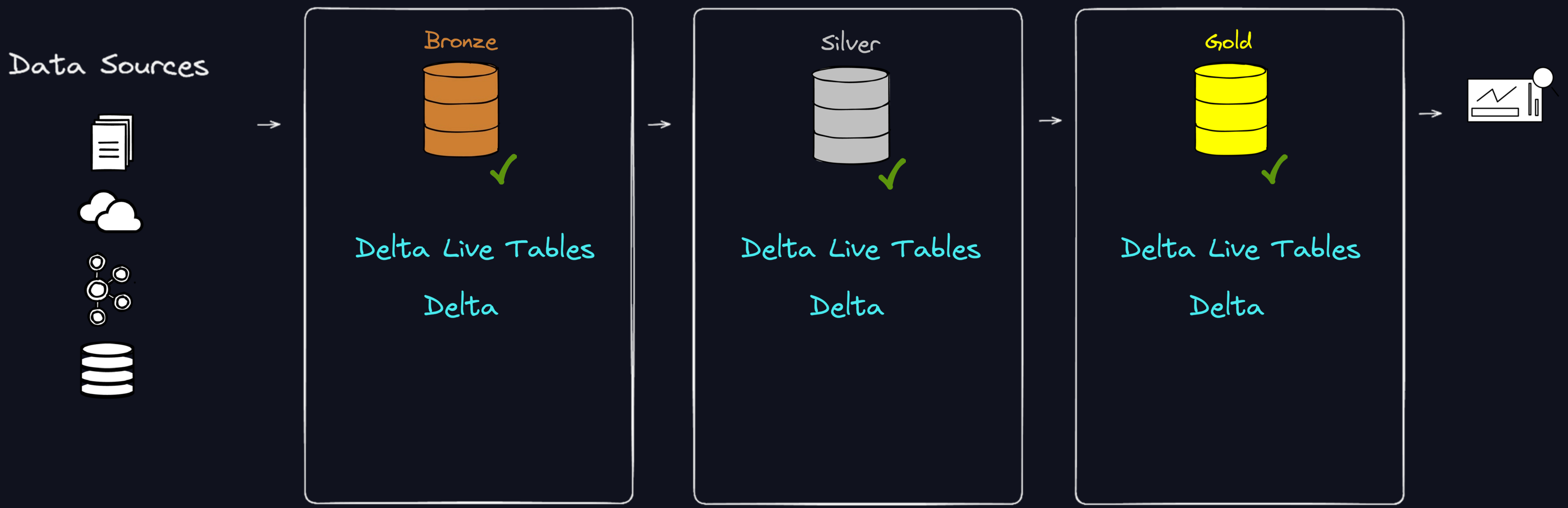
Tools:

- ❏ MLflow
- ❏ LlamaIndex, LangChain
- ❏ Hugging Face
- ❏ DSPy

Putting it all together

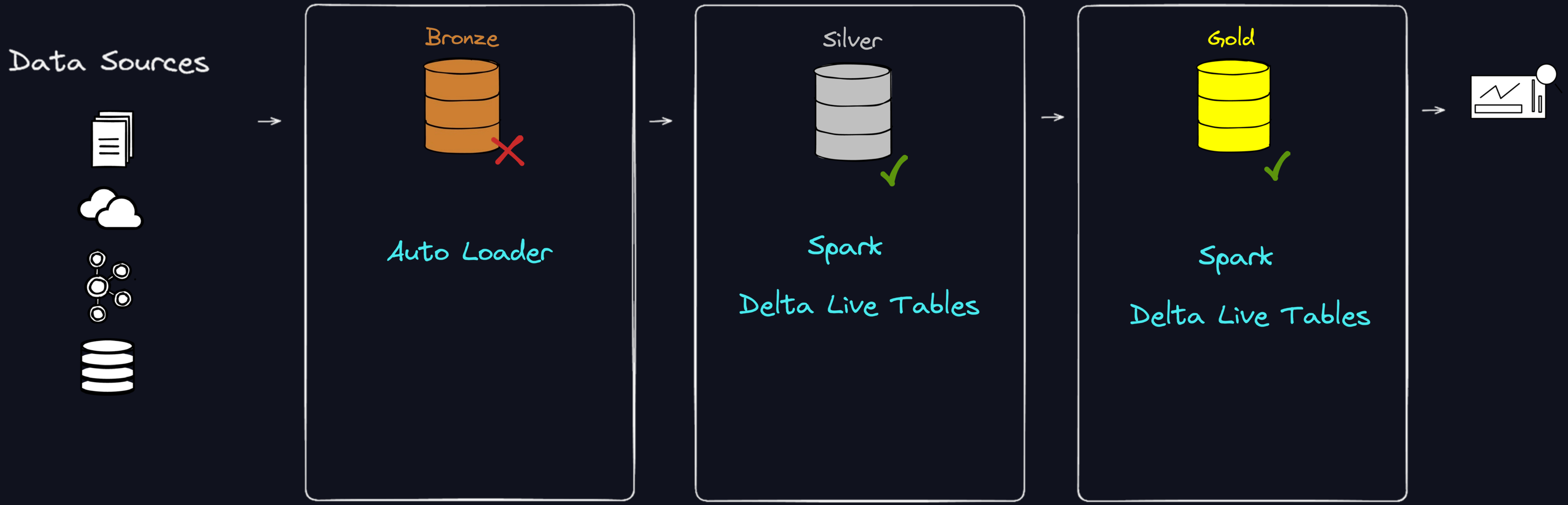
Medallion Architecture

Detecting inconsistencies and inaccuracies



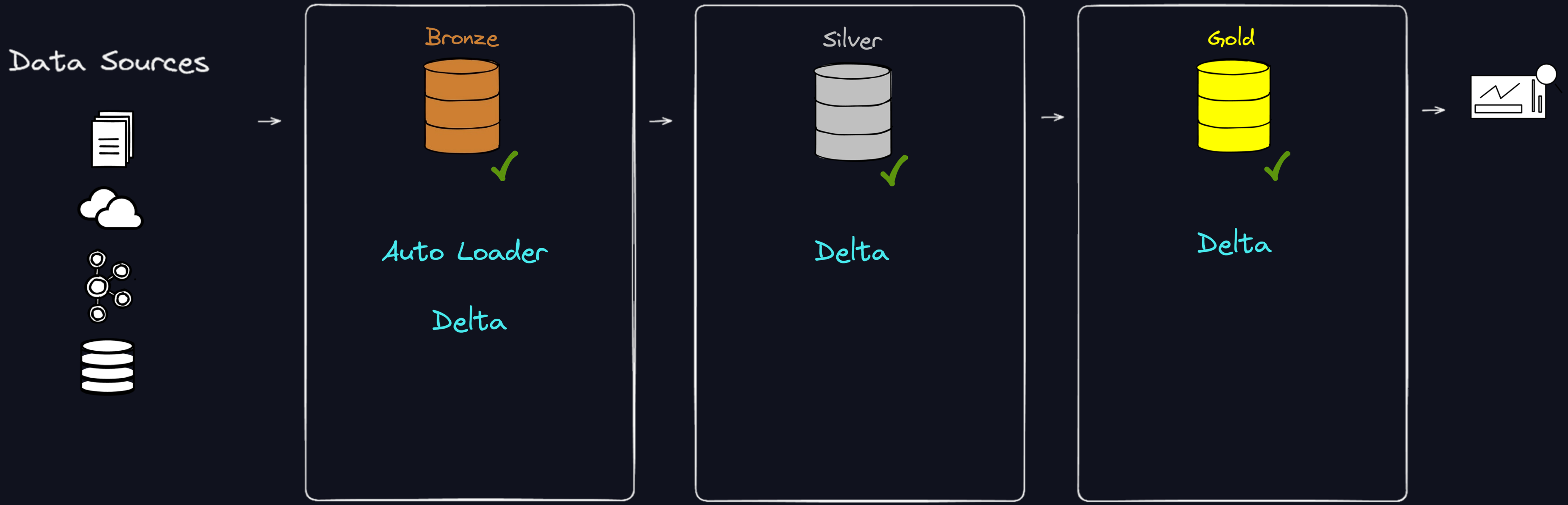
Medallion Architecture

Handling Violations



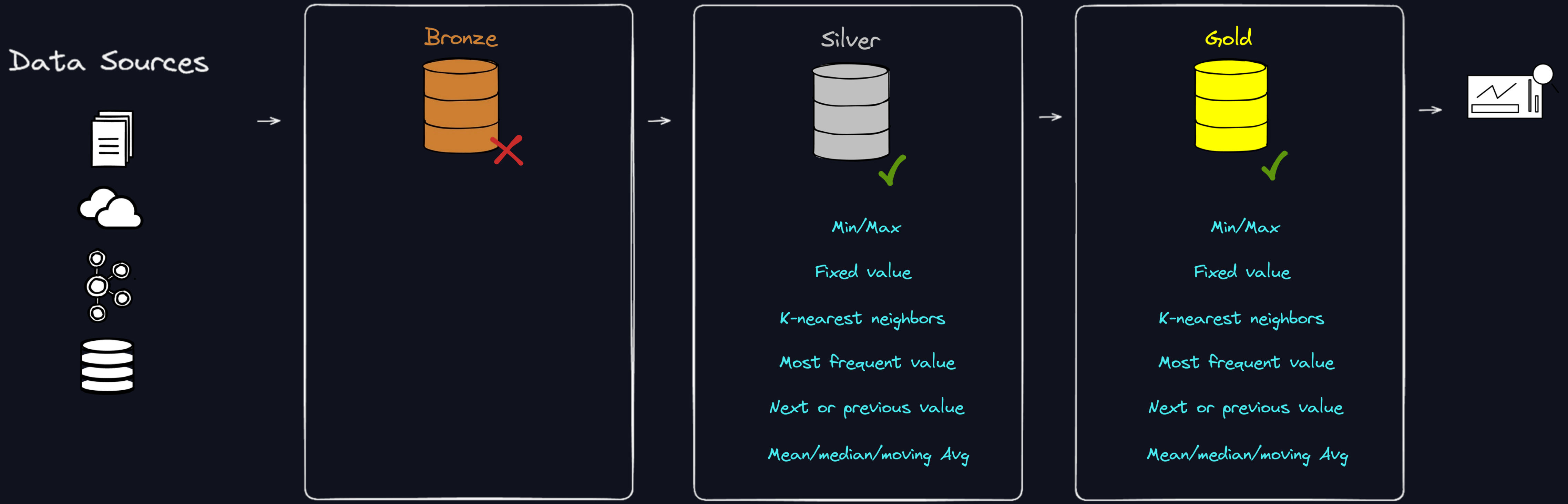
Medallion Architecture

Schema management: enforcement, evolution, overwrite, update



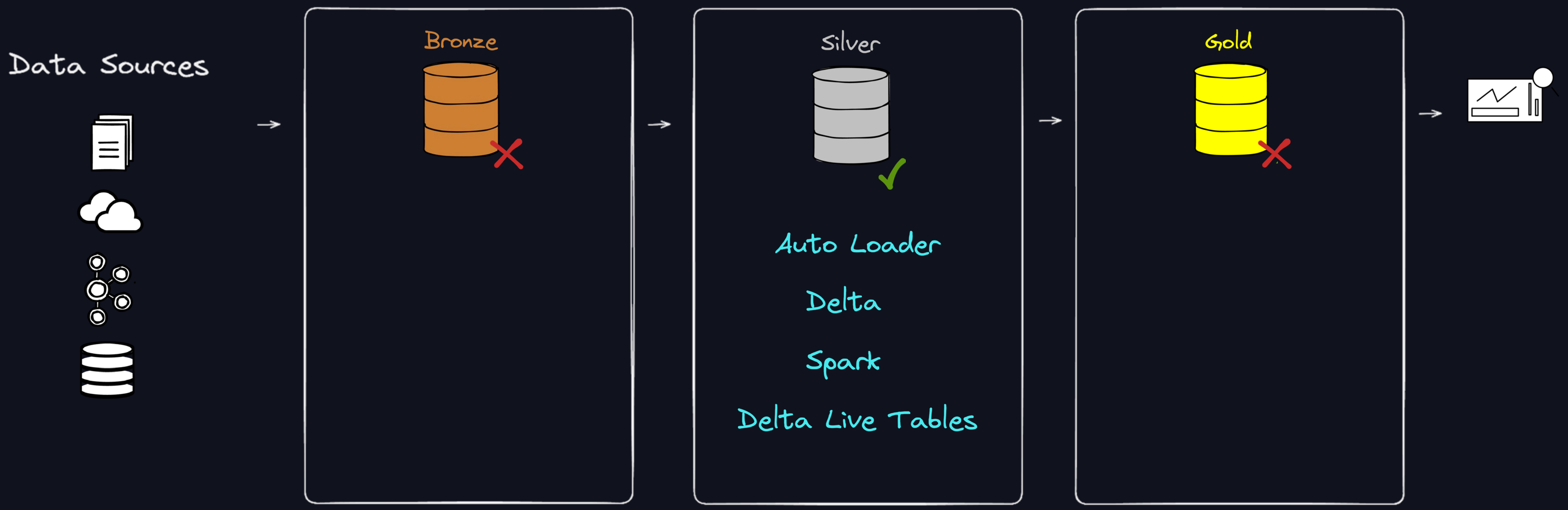
Medallion Architecture

Impute missing data



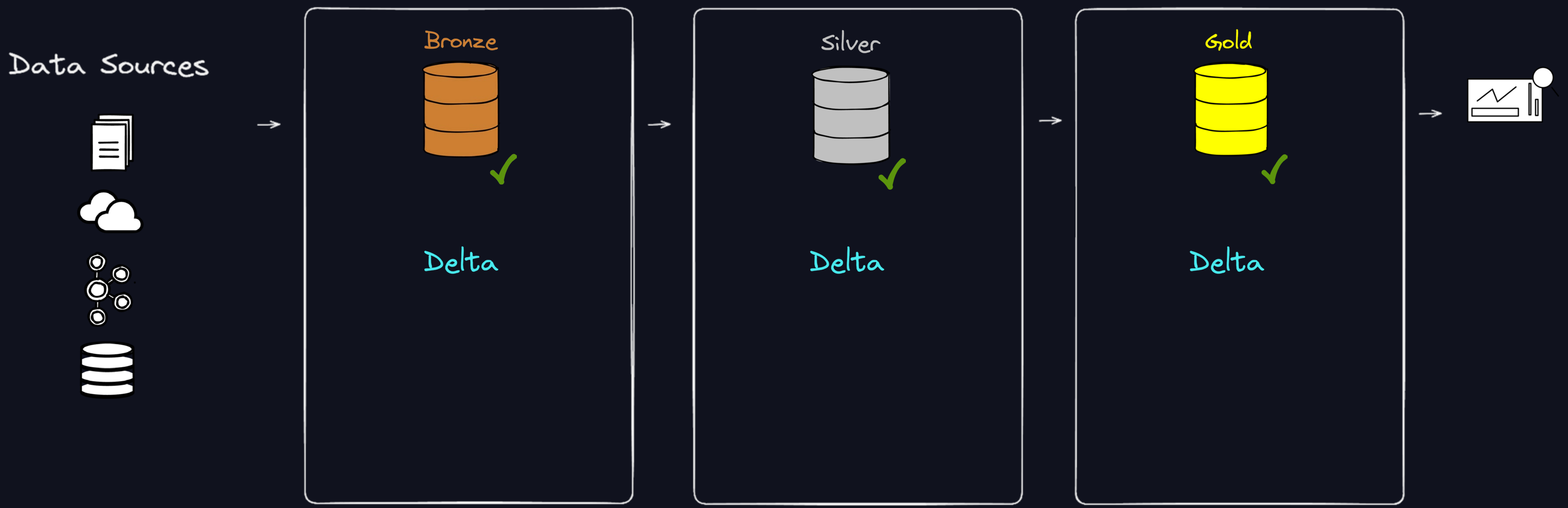
Medallion Architecture

Removing duplicate records



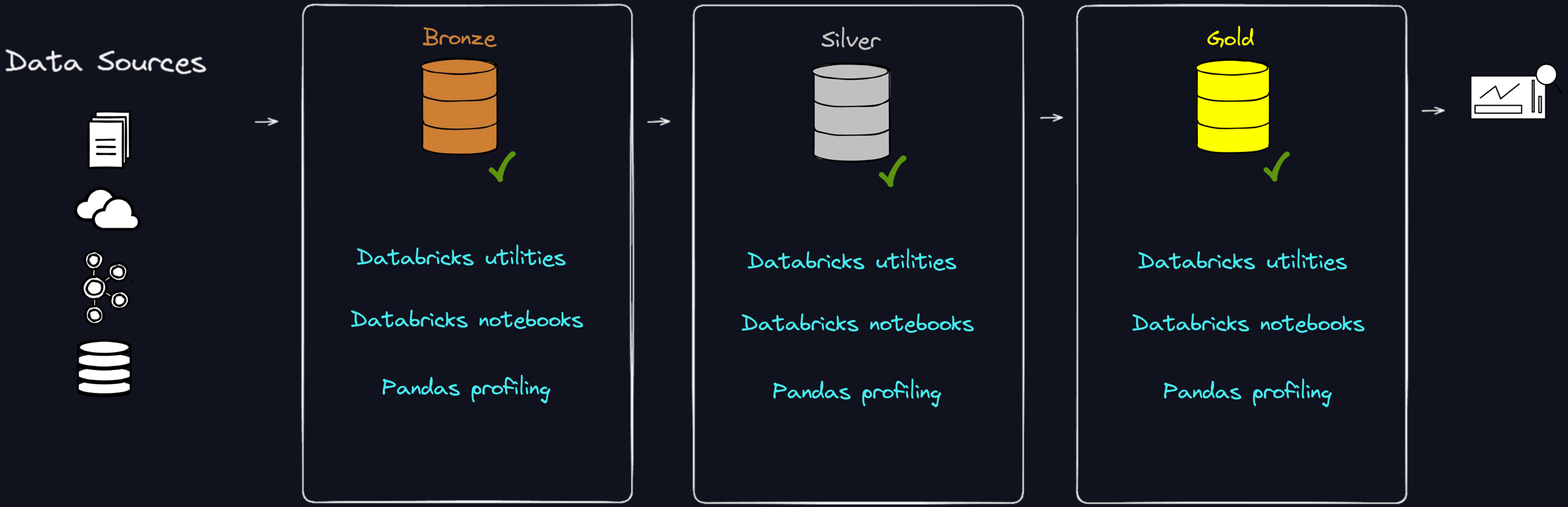
Medallion Architecture

Enabling rollbacks but preventing access to invalid historical data



Medallion Architecture

Profiling data



Data Sources

Bronze

Databricks utilities

Databricks notebooks

Pandas profiling

Silver

Databricks utilities

Databricks notebooks

Pandas profiling

Gold

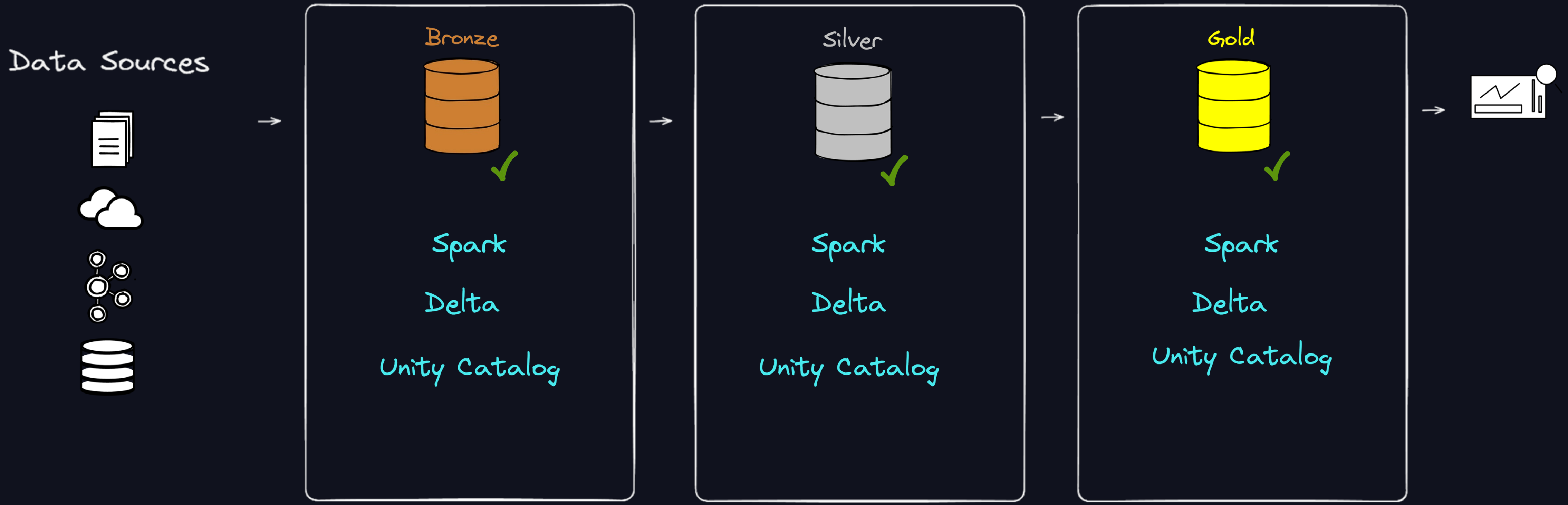
Databricks utilities

Databricks notebooks

Pandas profiling

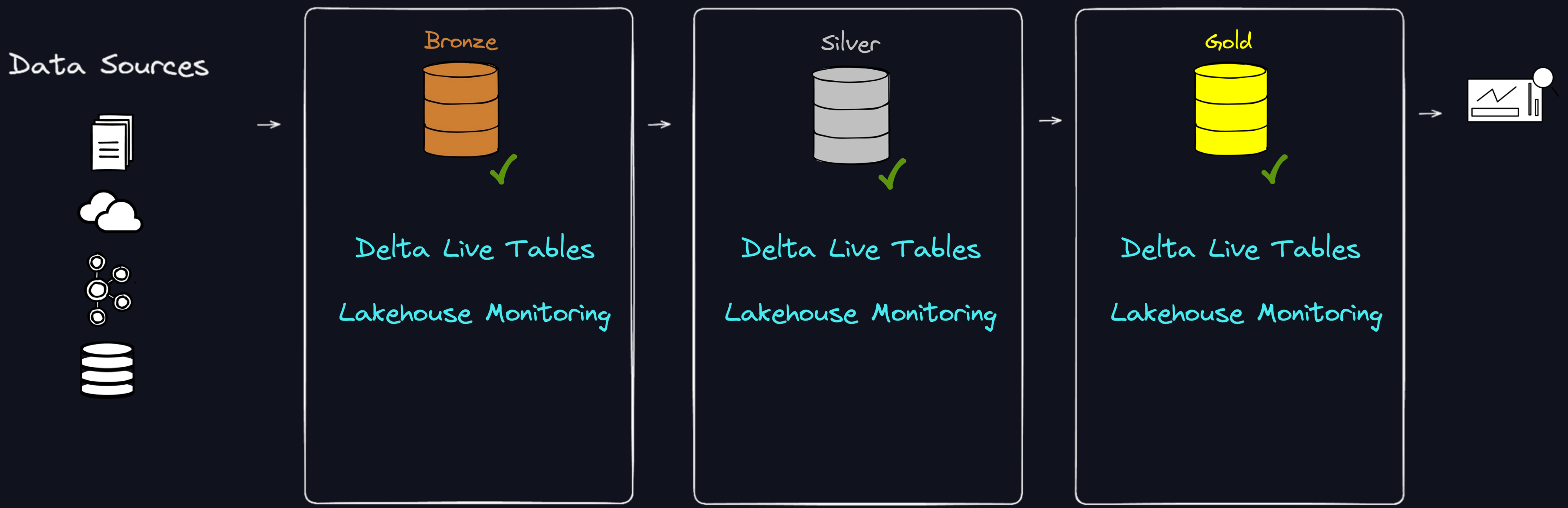
Medallion Architecture

Augmenting data



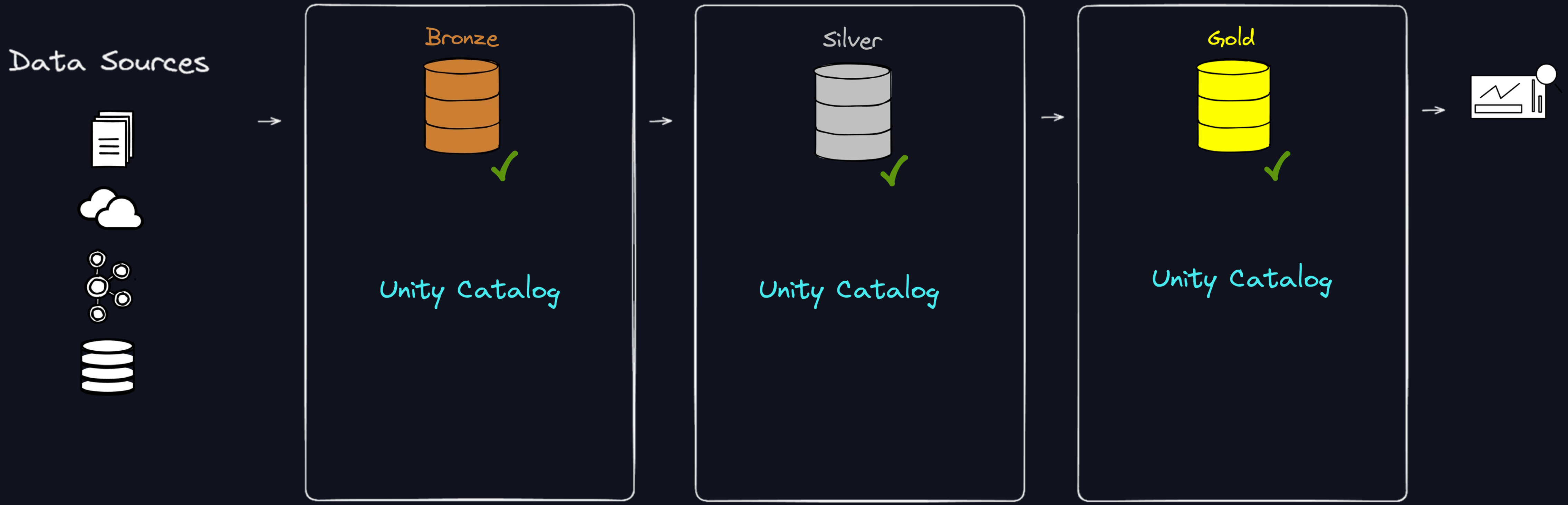
Medallion Architecture

Monitoring and alerting



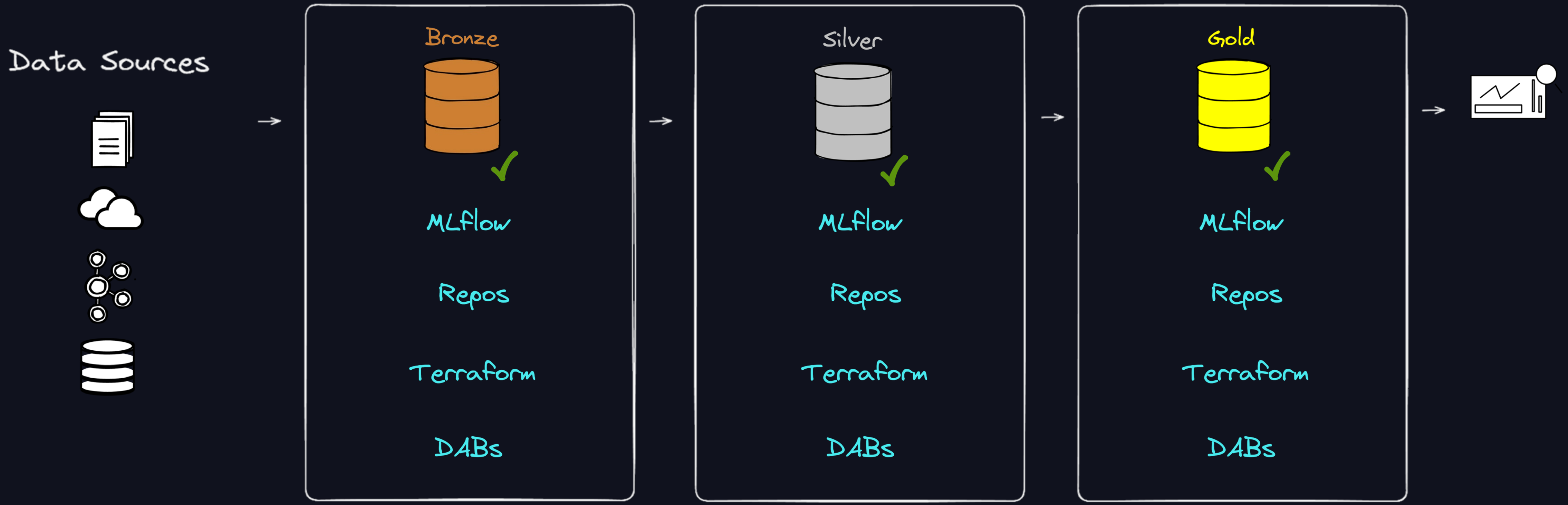
Medallion Architecture

Conducting root cause analysis



Medallion Architecture

Implementing DataOps, MLOps, LLMOps



Key Takeaways

❑ Data quality is the foundation of everything

❑ Get the basics right: Data Quality Management With Databricks

❑ Start with requirements, work towards automation, iterate!

Data Quality at Data+AI Summit

- Sponsored by: IBM | Building Better Data Quality with IBM Data Fabric and Databricks: Tuesday, Jun 11, 3:40 PM - 4:00 PM PDT
- Sponsored by: Datafold | Shifting Data Quality to the Left: Automating Data Testing on Databricks: Tuesday, Jun 11, 4:00 PM - 4:40 PM PDT
- LLM Evaluation: Auditing Fine-Tuned LLMs for Guaranteed Output Quality: Wednesday, Jun 12, 1:40 PM - 2:20 PM PDT
- Building High-Quality and Trusted Data Products with Databricks: Wednesday, Jun 12, 12:30 PM - 1:10 PM PDT
- Lakehouse Monitoring GA: Profiling, Diagnosing, and Enforcing Data Quality with Intelligence: Thursday, Jun 13, 11:20 AM - 12:00 PM PDT