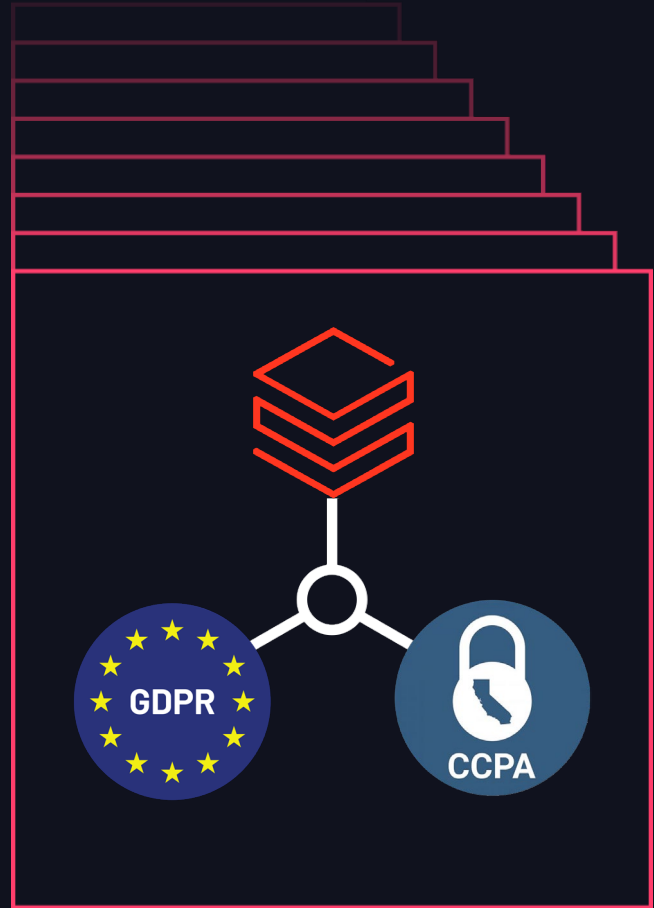



BECOMING GDPR AND CCPA COMPLIANT USING DLT

—
Marcin Wojtyczka
June 2024



About Me

- Senior Resident Solutions Architect at  databricks
- Open Source Contributor (eg. UCX, DQX, Terraform, ...)
- I help customers with:
 - Data Architecture & Engineering
 - Data Governance & Data Quality
 - Designing & Building reusable Frameworks and Tools
- Outside work expedition sailor



Agenda

1. GDPR and CCPA regulations and why You should be compliant
2. Approach to handle "Right to be Forgotten"
3. Challenges to implement "Right to be Forgotten"
4. Solutions using Delta Live Table (DLT)

Privacy and Data Security Laws

Requirements with broad application to any "Personal Identifiable Data" (PII)



GDPR (General Data Protection Regulation)

- **Who is regulated:** Data processing by businesses located in the EU and any organization that offer goods or services to individuals in the EU
- **Who is protected:** EU citizens and residents
- **What is protected:** PII data of EU citizens and PII data processed by EU businesses
- **Requirements:** Right to be Forgotten, Data Retention, Right of Access, ...



CCPA (California Consumer Privacy Act)

- **Who is regulated:** Organizations doing businesses in California with annual revenue \geq \$25M or process info of \geq 50,000 people
- **Who is protected:** Consumers who are California citizens
- **What is protected:** Information that identifies, relates to, or links to customers
- **Requirements:** Right to Delete, Data Retention, Right of Access, ...

Why this matters?

Non-Compliance can be very very costly \$\$\$

Why this matters?

Facebook case

In May 2023, Meta was issued a \$1.2 billion fine for relating to GDPR breaches following an inquiry into its Facebook service. This fine, was imposed for Meta's transfers of personal data to the U.S. on the basis of standard contractual clauses (SCCs).

Why this matters?

Criteo case

French advertising technology company Criteo was fined €40 million for GDPR breaches related to targeted advertising.

In addition to other allegations, the company was accused of non-compliance with the **Right to be Forgotten** requirement.

Why this matters?

Axpo Italia case

In September 2023, Axpo Italia, the producer and trader of renewable energy, was fined €10 million for processing outdated and inaccurate customer data.

Cost of doing nothing

Non-compliance can lead to significant fines and other legal consequences

- **Fines:** up to 4% of annual global turnover or €20 million, whichever is greater (GDPR)
- **Legal Challenges:** Consumer Lawsuits
- **Business Disruptions:** Systems to be taken offline
- **Loss of Customer Trust and Reputation**

Approach to handle "Right to be Forgotten"

Techniques

- Anonymization
- Pseudonymisation
- Data Masking
- Complete Erasure of PII data (safest)

Incomplete or wrongfully conducted anonymization processes can result in the re-identification of individuals!

Delta Live Tables (DLT)

Declarative framework that creates Delta tables and keeps them up to date

The screenshot displays the Delta Live Tables SQL Pipeline interface. At the top, the pipeline is titled "Delta Live Tables SQL Pipeline" and is currently in "Production" mode. The status is "Completed" as of 2/25/2022, 9:25:02 AM. The interface shows a data flow diagram with several tables and views, including v_raw_yellow_taxi, tbi_bronze_taxi_y..., tbi_silver_yellow..., tbi_gold_taxi_for..., v_ref_taxi_zone_l..., v_ref_taxi_payme..., tbi_silver_taxi_pa..., tbi_gold_union_taxi, v_ref_taxi_rate_e..., tbi_silver_taxi_ra..., tbi_silver_green..., and tbi_bronze_taxi_g... Each table/view shows completion status and data counts.

On the right side, there is a "Data Quality" section with a donut chart showing 77.9% (61,752,707) Written and 22.1% (17,506,422) Dropped. Below this is a "Expectations" table:

Name	Action	Fail %	Failed Records
valid_trip_distance	DROP	22.1%	17484277
valid_passenger_count	DROP	0.2%	176524

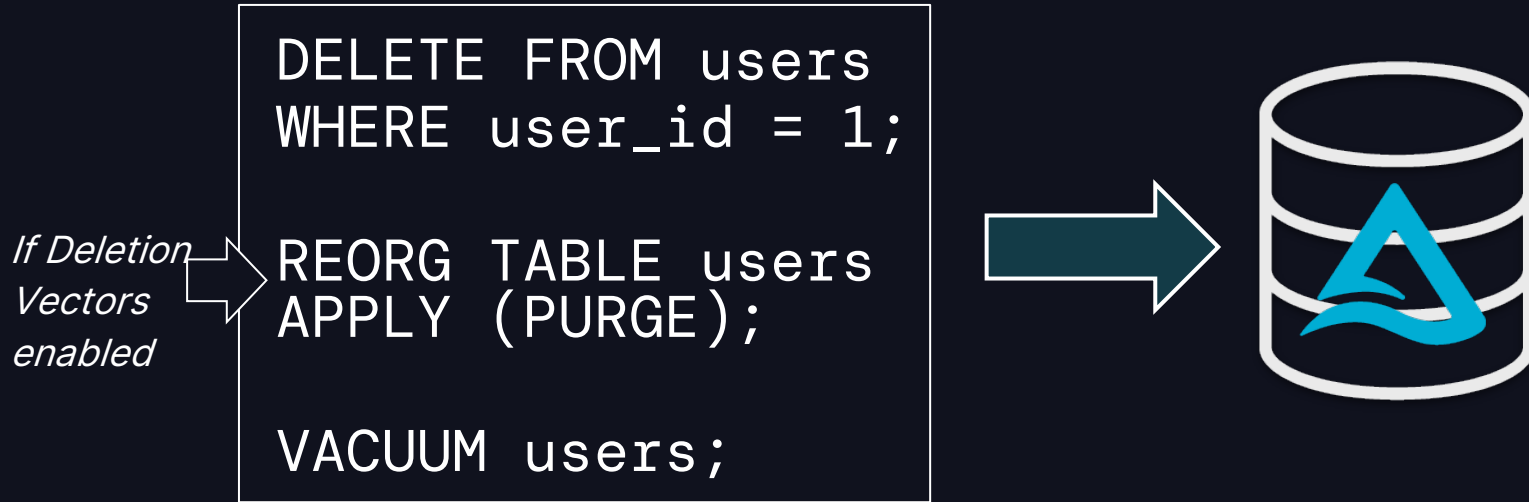
At the bottom, there is a log of completed flows:

Time	Flow Name	Status
4 minutes ago	flow_progress	Flow 'tbi_gold_taxi_for_analysis' has COMPLETED.
4 minutes ago	flow_progress	Flow 'tbi_silver_taxi_payments' has COMPLETED.
4 minutes ago	flow_progress	Flow 'tbi_silver_taxi_rates' has COMPLETED.
4 minutes ago	flow_progress	Flow 'tbi_gold_union_taxi' has COMPLETED.
4 minutes ago	update_progress	Update '1d1ad5' is COMPLETED.

Point Deletes in the Data Lakehouse

Delta Lake

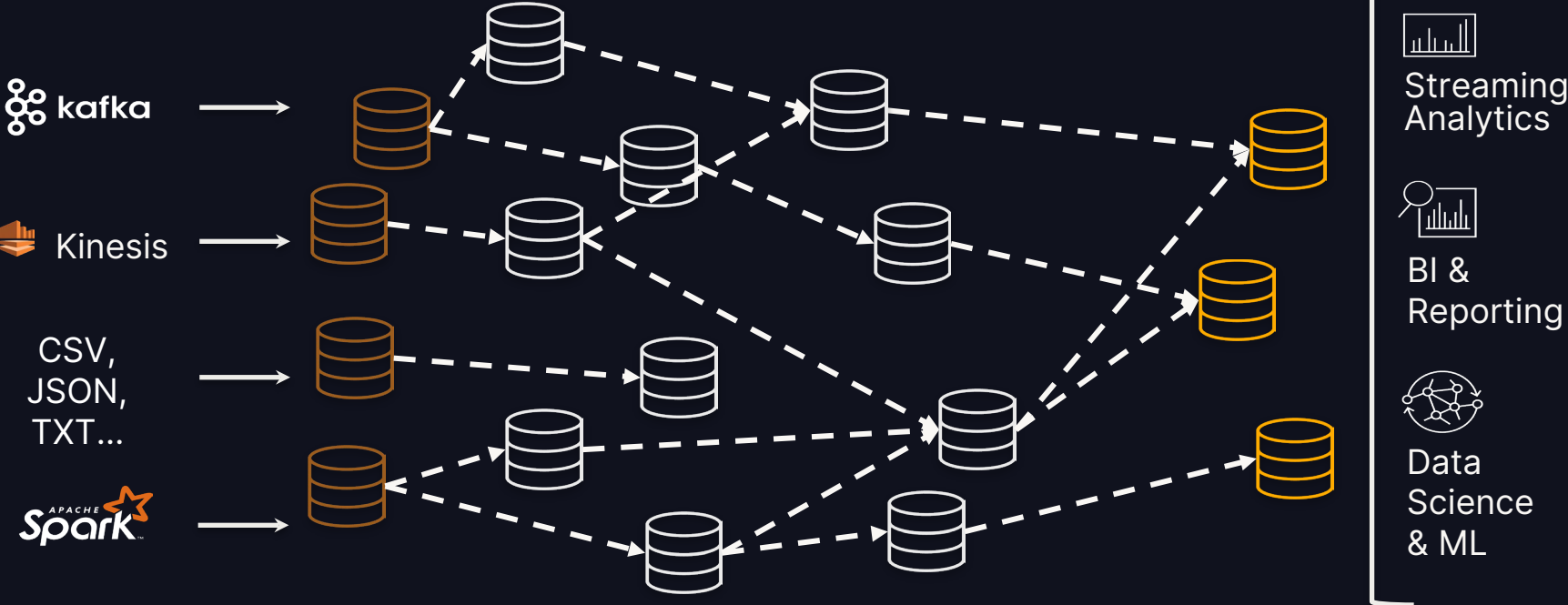
Efficient point deletes thanks to ACID transactions and Deletion Vectors:



*Even faster with Data Skipping, Liquid Clustering, or Z-Order
(eg. on fields used during delete operations)*

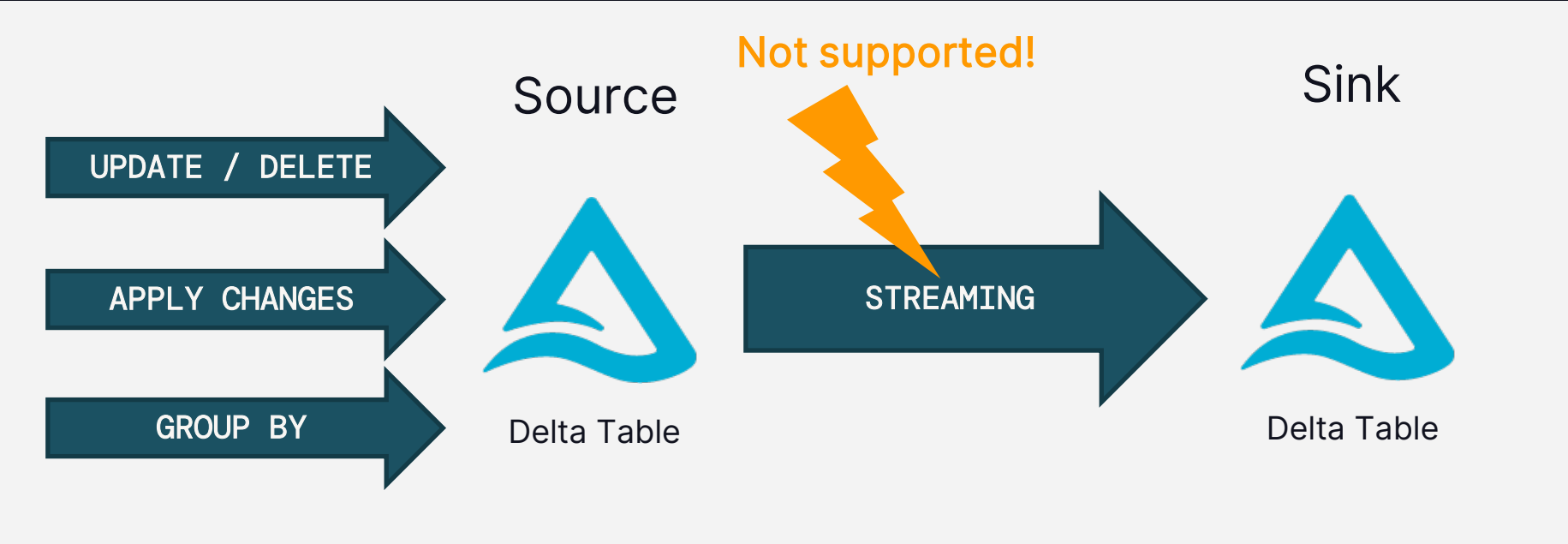
Challenges to implement "Right to be Forgotten"

The reality is not so simple



Challenges to implement "Right to be Forgotten"

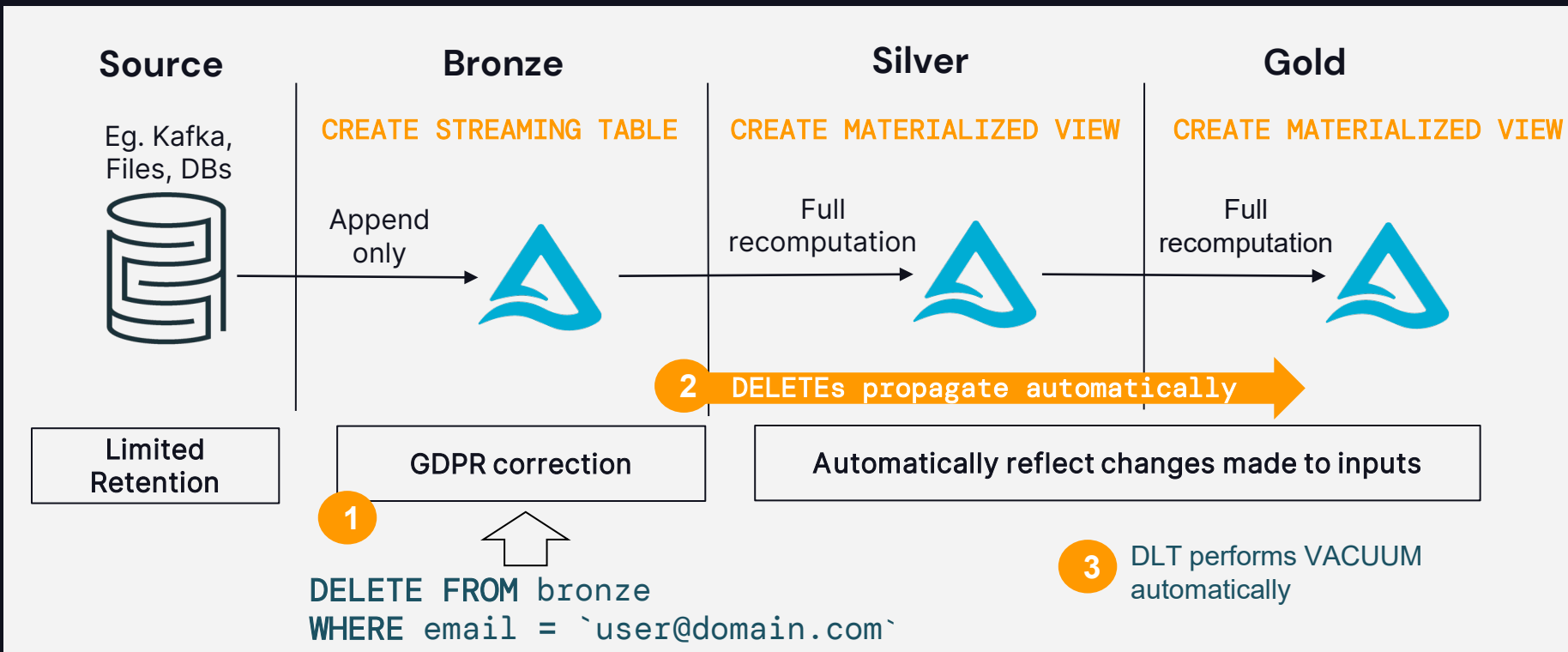
DLT Streaming



What can we do about it?

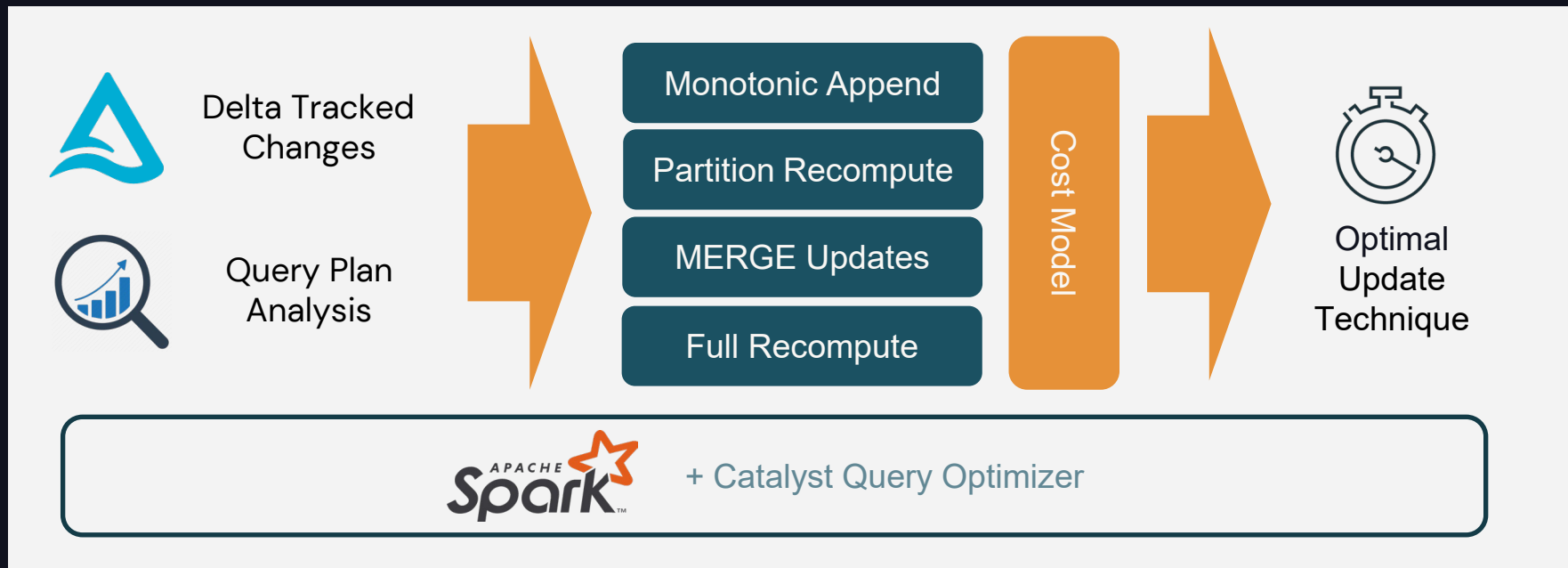
Solution 1 - Full Recomputation

Streaming Tables for Bronze and Materialized Views afterward



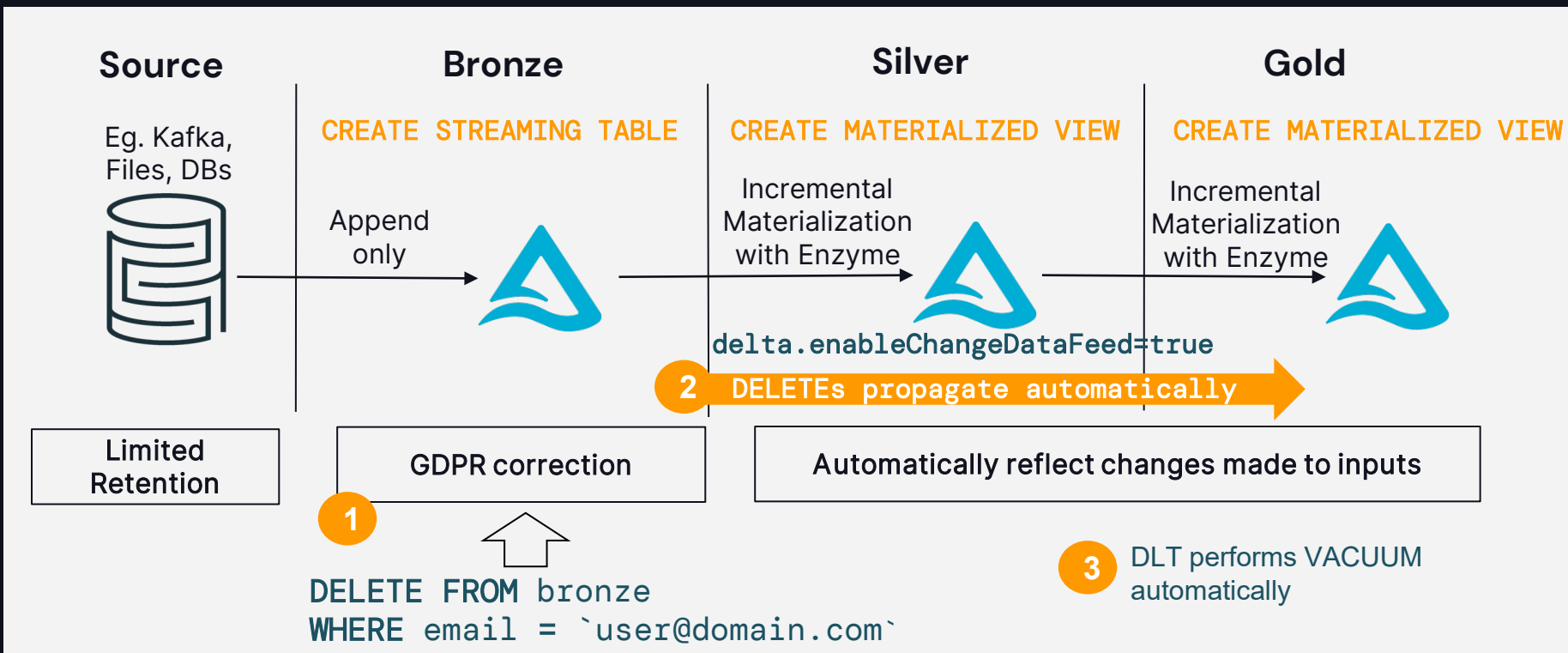
Enzyme (Incremental Refresh) to the rescue

Automatically select the best strategy to refresh materialized views



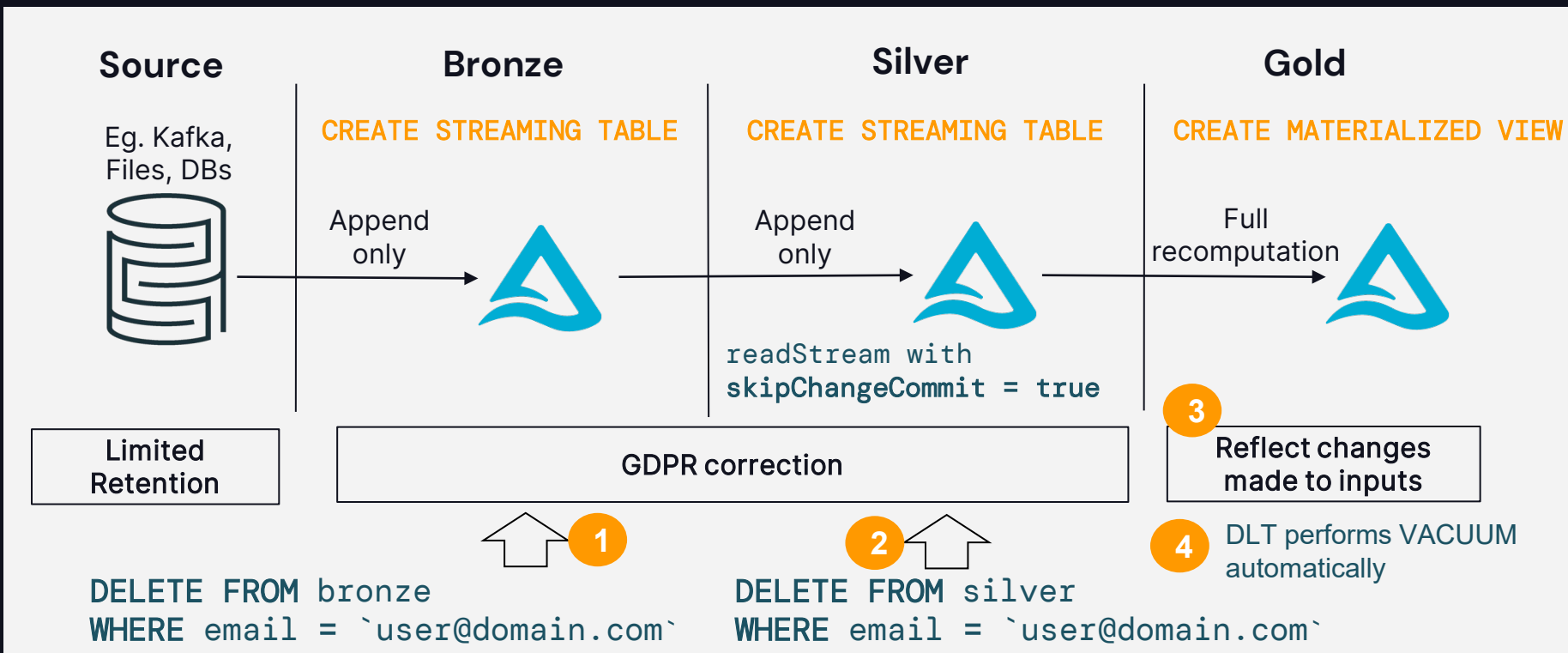
Solution 2 - With Incremental Refresh

Streaming Tables for Bronze and Materialized Views with Enzyme afterward



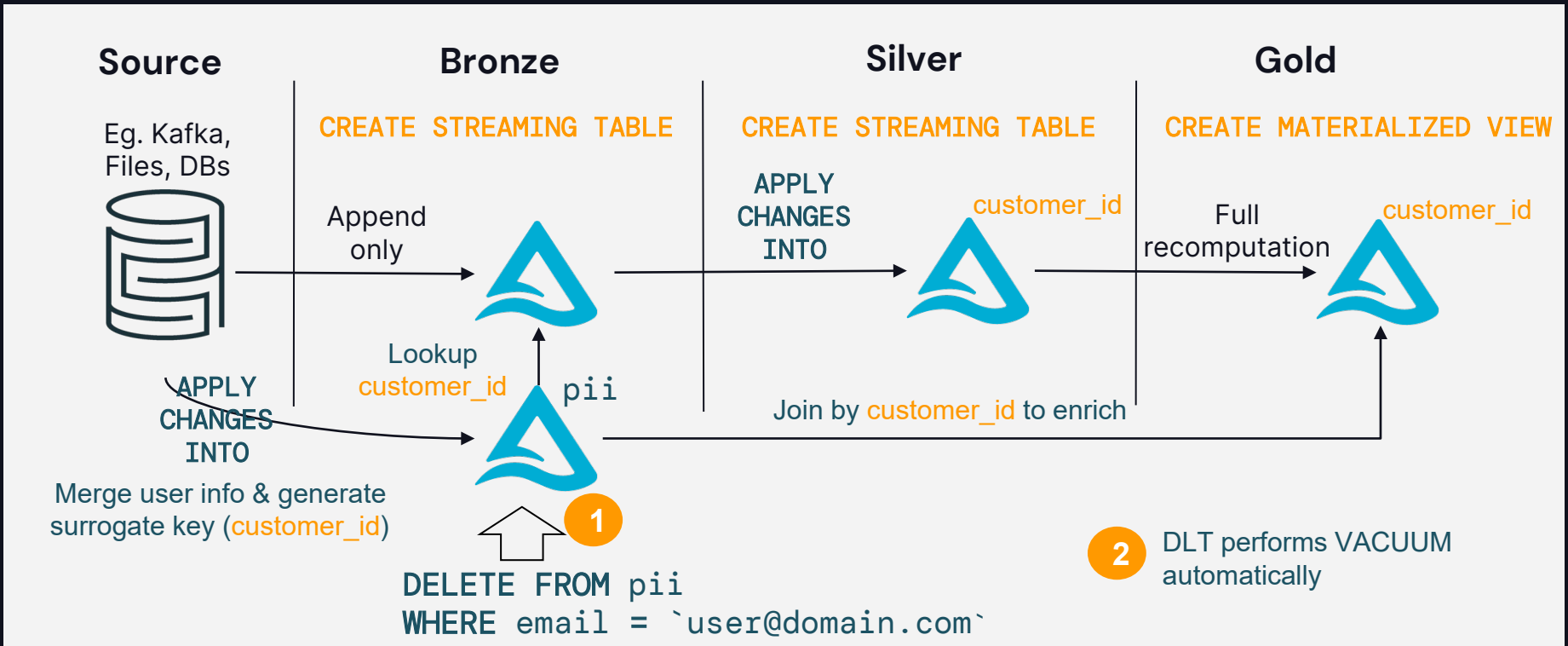
Solution 3 - For append only sinks

Streaming Tables for Bronze & Silver and Materialized Views afterward



Solution 4 - For Most Scenarios

Separate PII data from the rest of the data



Summary of Solutions

Solution	Consider using when



Summary of Solutions

Solution	Consider using when
1 - Streaming Tables for Bronze and Materialized Views afterward	<ul style="list-style-type: none">● Type of query used is not supported by Enzyme optimization (Incremental Refresh), otherwise, use Solution 2● Full recomputation of tables is acceptable



Summary of Solutions

Solution	Consider using when
1 - Streaming Tables for Bronze and Materialized Views afterward	<ul style="list-style-type: none">● Type of query used is not supported by Enzyme optimization (Incremental Refresh), otherwise, use Solution 2● Full recomputation of tables is acceptable
2 - Streaming Tables for Bronze and Materialized Views with Enzyme afterward	<ul style="list-style-type: none">● Type of query used is supported by Enzyme optimization● Full recomputation of tables is not acceptable due to cost and latency requirements

Summary of Solutions

Solution	Consider using when
1 - Streaming Tables for Bronze and Materialized Views afterward	<ul style="list-style-type: none">● Type of query used is not supported by Enzyme optimization (Incremental Refresh), otherwise, use Solution 2● Full recomputation of tables is acceptable
2 - Streaming Tables for Bronze and Materialized Views with Enzyme afterward	<ul style="list-style-type: none">● Type of query used is supported by Enzyme optimization● Full recomputation of tables is not acceptable due to cost and latency requirements
3 - Streaming Tables for Bronze & Silver and Materialized Views afterward	<ul style="list-style-type: none">● Type of query used is not supported by Enzyme optimization, otherwise, use Solution 2● Full recomputation of Silver tables is not acceptable and queries in the Silver layer are run in append mode

Summary of Solutions

Solution	Consider using when
1 - Streaming Tables for Bronze and Materialized Views afterward	<ul style="list-style-type: none">● Type of query used is not supported by Enzyme optimization (Incremental Refresh), otherwise, use Solution 2● Full recomputation of tables is acceptable
2 - Streaming Tables for Bronze and Materialized Views with Enzyme afterward	<ul style="list-style-type: none">● Type of query used is supported by Enzyme optimization● Full recomputation of tables is not acceptable due to cost and latency requirements
3 - Streaming Tables for Bronze & Silver and Materialized Views afterward	<ul style="list-style-type: none">● Type of query used is not supported by Enzyme optimization, otherwise, use Solution 2● Full recomputation of Silver tables is not acceptable and queries in the Silver layer are run in append mode
4 - Separate PII data from the rest of the data	<ul style="list-style-type: none">● Full recomputation of tables is not acceptable● Designing a new data model● Managing lots of tables and needing a simple method to make the whole system compliant● Needing to be able to reuse most of the data (using surrogate key) while being compliant with regulations

Thank you for attending!

More on the subject in the Databricks blog:
[Handling "Right to be Forgotten" in GDPR and CCPA using Delta Live Tables \(DLT\)](#)