Processing Streaming Data Frames

Janani Ravi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Triggers and output modes

Selections, projections, and aggregations on DataFrames

Running SQL queries on streaming DataFrames

Triggers and Output Modes

Trigger

Events that determine when transformations on accumulated input data need to be re-performed. Each trigger event emits new data into the Result Table

Trigger

Events that determine when transformations on accumulated input data need to be re-performed. Each trigger event emits new data into the Result Table

Trigger

Events that determine when transformations on accumulated input data need to be re-performed. Each trigger event emits new data into the Result Table

Trigger

Events that determine when transformations on accumulated input data need to be re-performed. Each trigger event emits new data into the Result Table

Result Table

Executing a query on input data generates the Result Table. Rows in the Result Table are written out to an external data sink

Types of Triggers

Default

Fixed interval micro-batch

One-time micro-batch

Continuous with fixed checkpoint interval

Micro-batch Processing Mode

Default

Fixed interval micro-batch

One-time micro-batch

Continuous with fixed checkpoint interval

Continuous Processing Mode

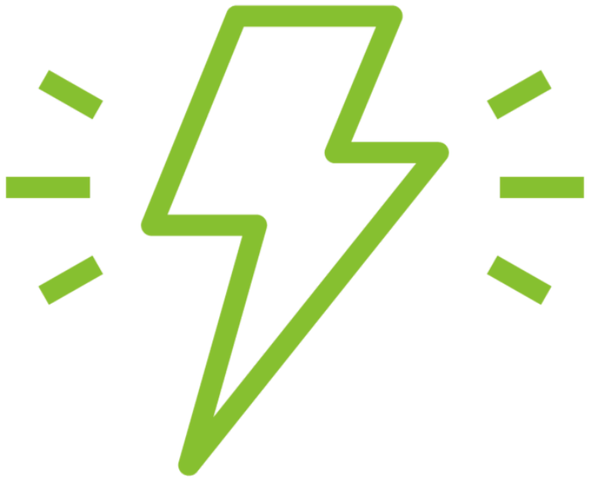
Default

Fixed interval micro-batch

One-time micro-batch

Continuous with fixed checkpoint interval

Default

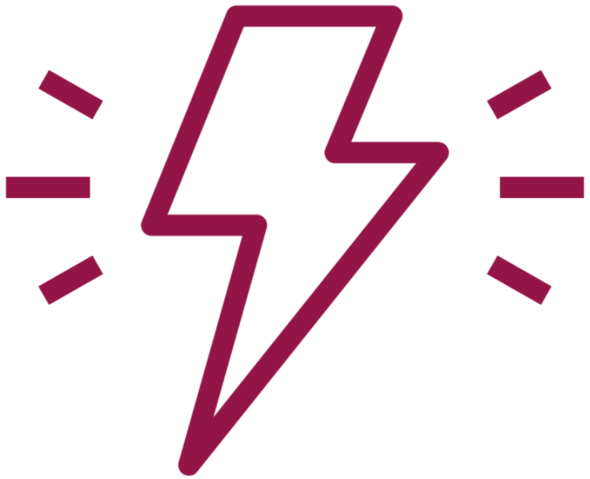


Used when no trigger setting specified

Query executed in micro-batch mode

Each new micro-batch generated when previous one completes processing

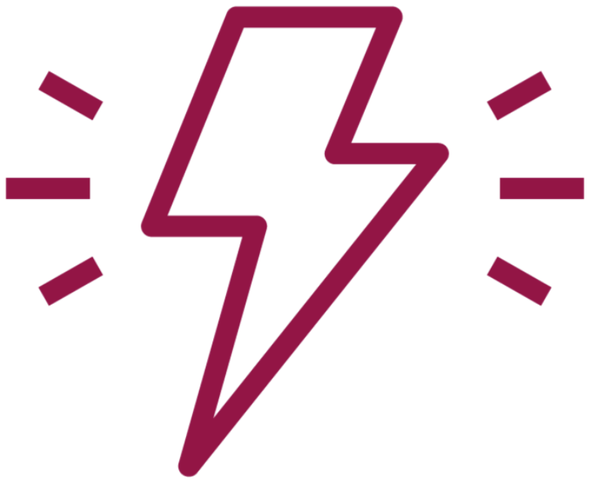
Fixed Interval Micro-batch



Micro-batch kicked off at user-specified intervals

If no data available no processing

Fixed Interval Micro-batch



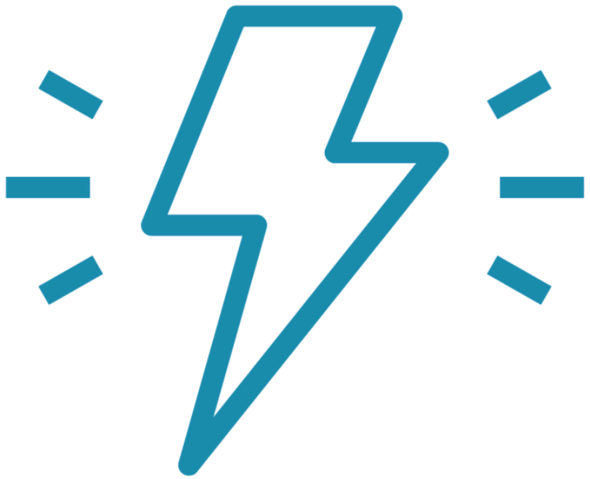
If previous micro-batch completes within the interval:

- engine waits till interval is over

If previous micro-batch takes longer than specified interval:

- next micro-batch starts as soon as data arrives

One-time Micro-batch



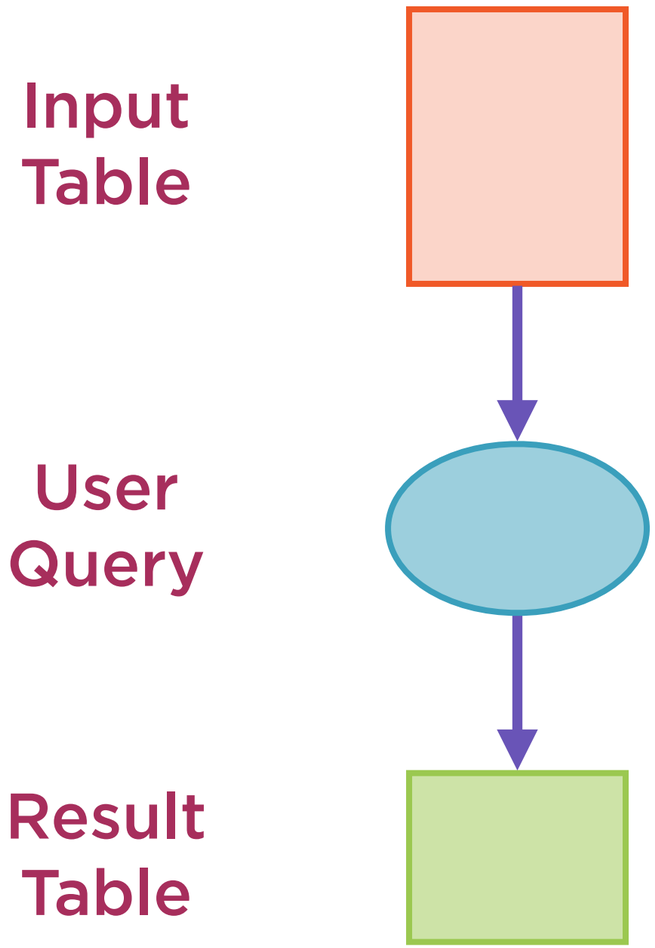
Execute only one micro-batch to process all available data

Once processed query will stop

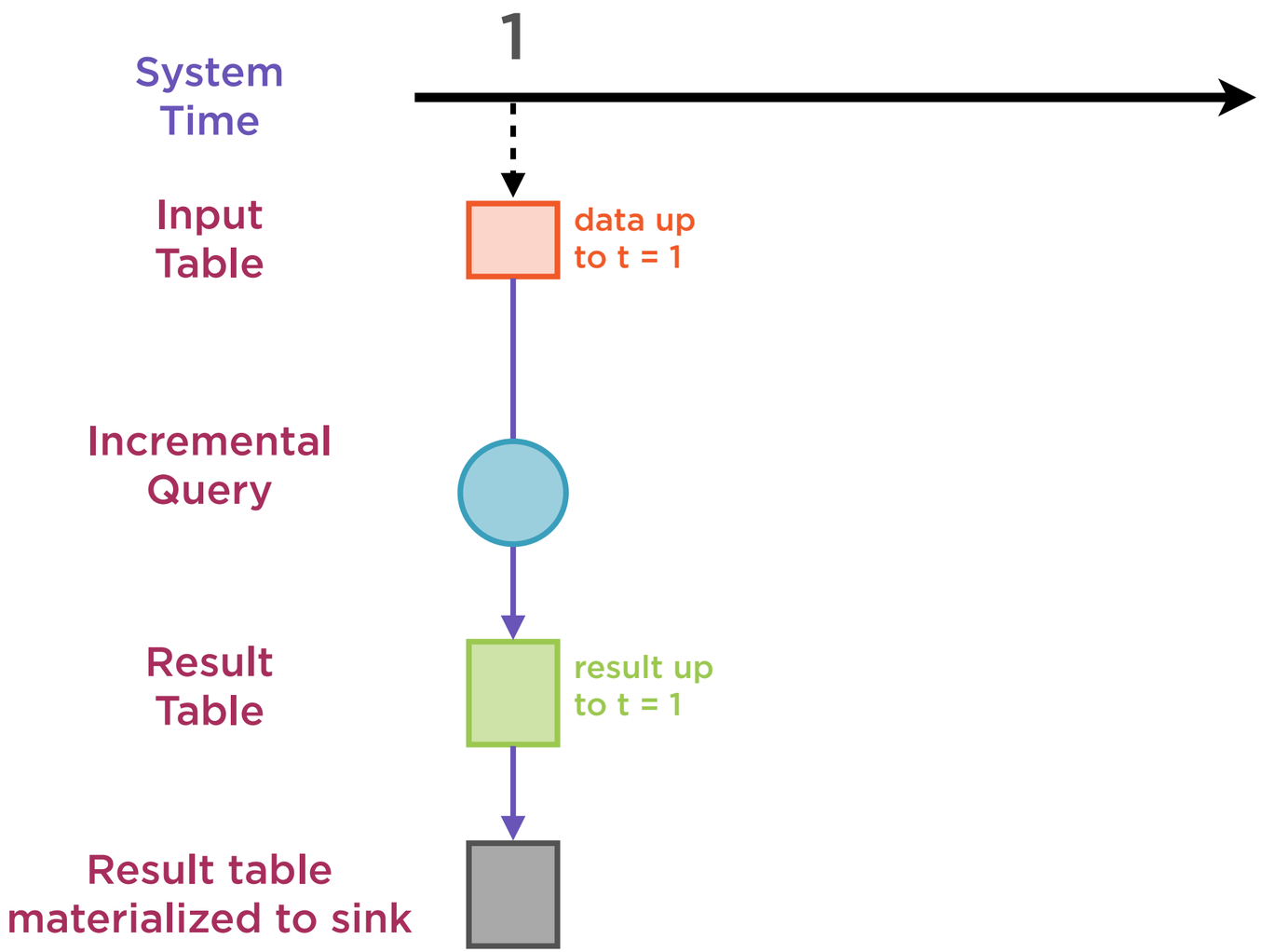
Used when cluster periodically spun up to process data since last period

May result in significant cost savings

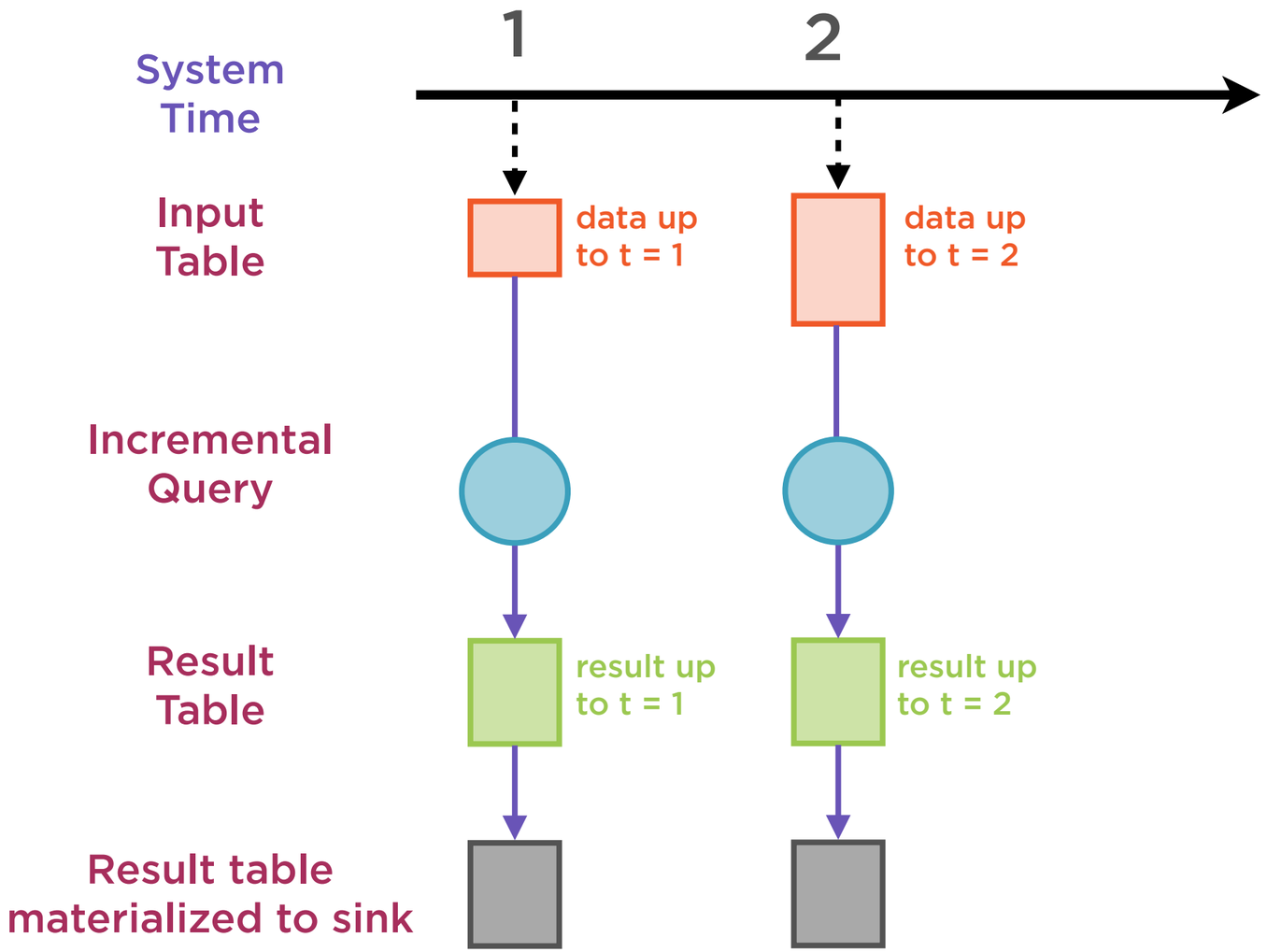
Result Table



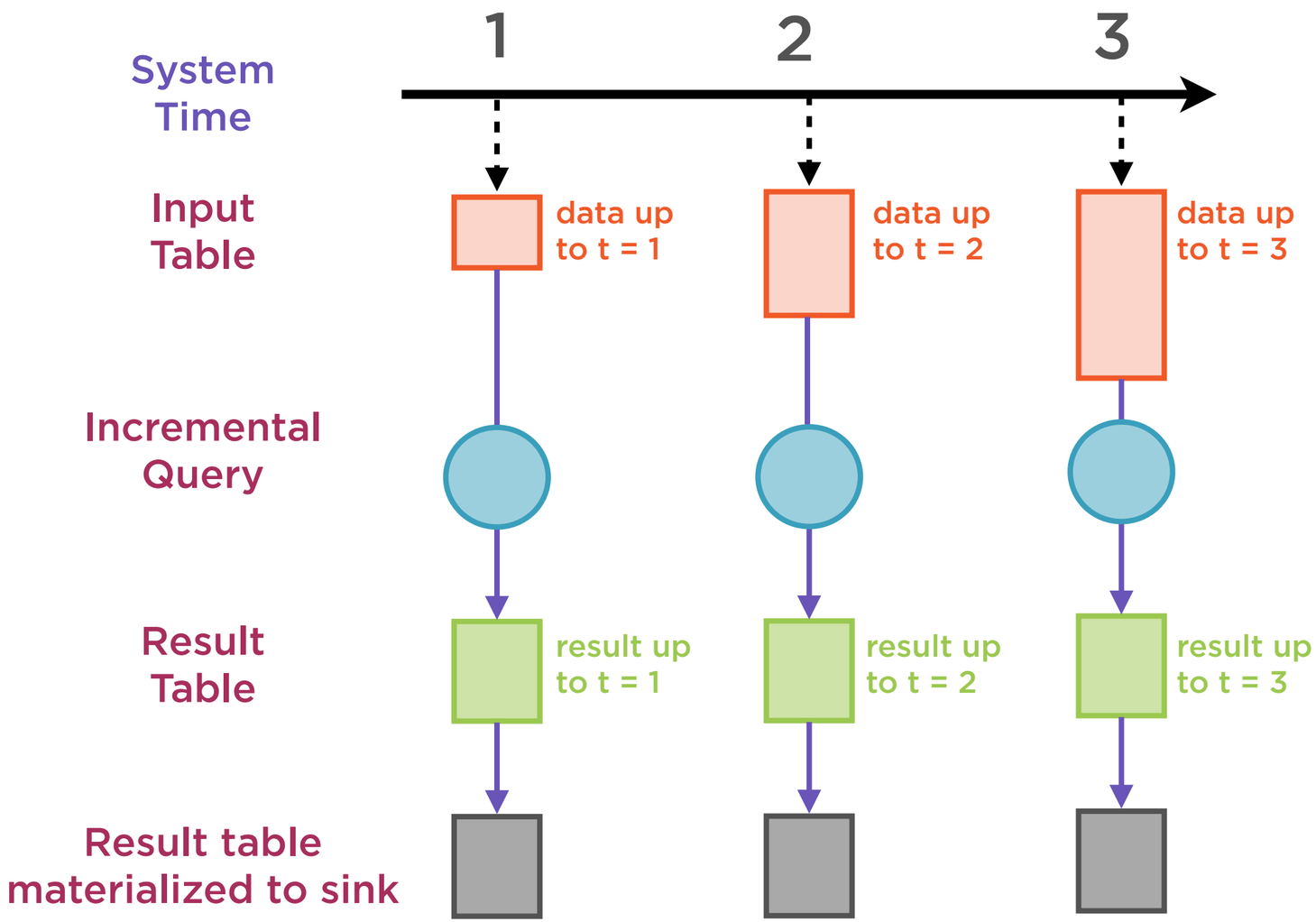
Result Table



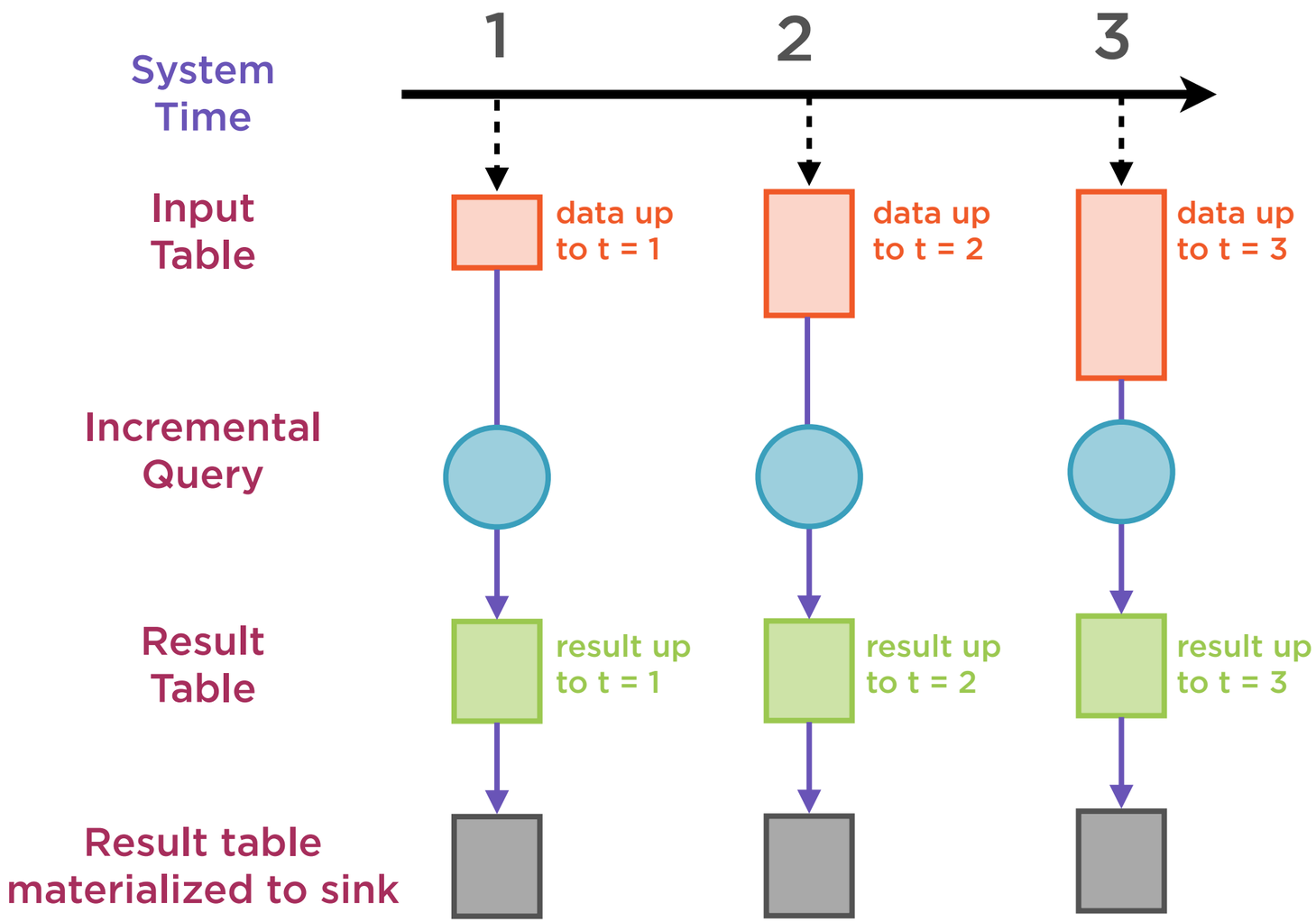
Result Table



Result Table



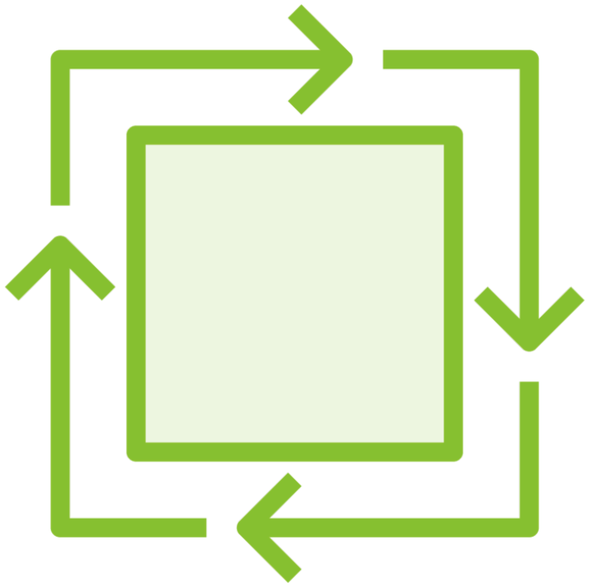
Result Table



When writing to the sink the entire Result Table is not materialized

What is written out depends on the **mode**

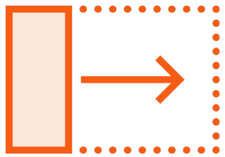
Output Modes



Determines what Result Table rows get sent to storage

- Update mode
- Append mode
- Complete mode

Output Modes



Update mode - only Result Table rows updated since last trigger

Even previous results will be updated in case of aggregations

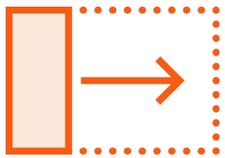
Append mode - only Result Table rows appended since last trigger

Previous (existing) output rows cannot change

Complete mode - entire updated Result Table is sent across

Storage connector must decide how to use all that data

Output Modes



Update mode

Selections, projections, and aggregations

Append mode

Selections, projections, aggregations not supported

Complete mode

Selections, projections, aggregations, ordering

Demo

Performing selection and projection operations on input data using DataFrames and SQL

Demo

Performing aggregation operations on input data using DataFrames and SQL

Demo

Exploring triggers in Spark

Unsupported Operations on DataFrames

Unsupported Operations

A small number of operations are not supported by Streaming DataFrames

- Multiple streaming aggregations
- Limit and take first N rows
- Distinct operations
- Some types of outer joins

Unsupported Operations

Sorting on Streaming DataFrames is allowed only

- After an aggregation
- And in Complete Output Mode

Summary

Triggers and output modes

Selections, projections, and aggregations on DataFrames

Running SQL queries on streaming DataFrames

Up Next:
Performing Windowing Operations on Streams