Buổi 3:

Spark SQL (tiếp) **và Spark Structured Streaming**

- Using SQL in SparkSQL
- Spark Catalog
- Stream Processing Concept
- Spark Structured Streaming Fundamentals
- Xây dựng ứng dụng xử lý luồng dữ liệu với Spark Structured Streaming

Nguyễn Chí Thanh – Ban Quản trị Dữ liệu

1.
Using SQL in SparkSQL

SQL in SparkSQL

- **Bản chất DF, DS giống như các bảng dữ liệu truyền thống**
  - Được tổ chức theo hàng, cột.
  - Có schema xác định

- **Ngoài việc sử dụng SparkSQL APIs, Spark hỗ trợ thực thi các câu lệnh SQL như một RDBMS thông thường**
  - *select*
  - *where*
  - *group*
  - *agg functions*
  - *order*

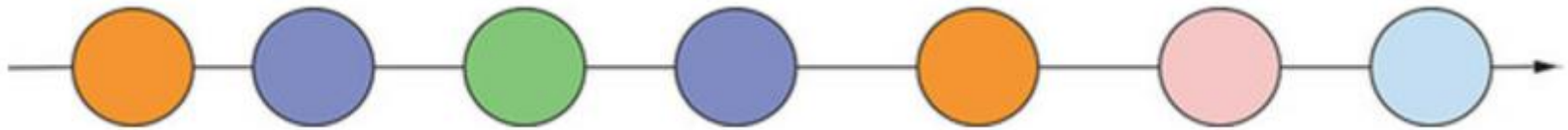
- **Cần có thành phần quản lý các bảng.**

Spark Catalog

- Quản lý các "bảng" mà bản chất là DataFrame hoặc DataSet.
- Thông tin quản lý chỉ là Metadata.
- Hỗ trợ chia sẻ thông tin "bảng" để thực thi SQL, thông qua Spark Session.

2.
Stream Processing

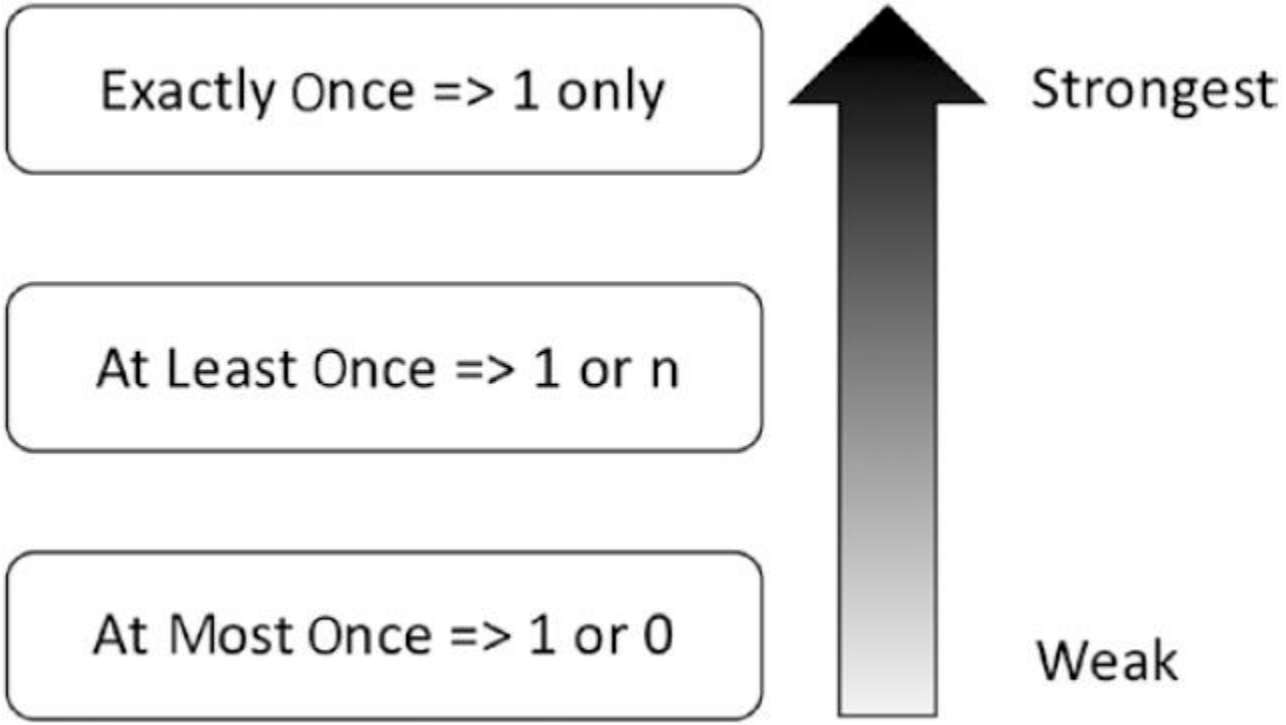
Stream Processing



- Xử lý dữ liệu ngay khi nó xuất hiện trong hệ thống.
- Có khả năng làm việc với luồng dữ liệu vô hạn.
- Quá trình tính toán xảy ra liên tục, ứng dụng hoạt động 247.
- Yêu cầu cao về tính ổn định, khả năng chịu lỗi.
- Hệ thống cần có thiết kế tốt, đồng bộ từ phần cứng đến phần mềm.

Data Delivery Semantics

- At most 1: Nhiều nhất 1, cũng có thể không có bản ghi nào được đưa vào để xử lý.
- At least 1: Ít nhất 1, có thể đẩy lặp nhưng không mất mát dữ liệu.
- Exactly 1: Chính xác 1 bản ghi, đảm bảo không lặp nhưng cũng không thiếu.



Stream Processing Concept (1)

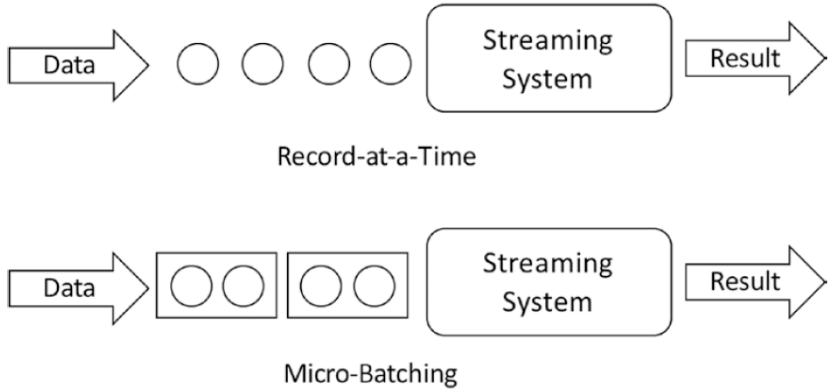
Record-at-a-Time

Micro-Batching

Stream Processing Concept (2)

- **Record-at-a-time: Xử lý dữ liệu ngay khi nó xuất hiện:**
  - Yêu cầu rất cao về công nghệ xử lý, thiết kế ứng dụng, phần mềm, phần cứng, cách thức triển khai, vận hành.
  - Độ trễ nhỏ, đem lại trải nghiệm khách hàng tốt

- **Micro-batching: Chia dữ liệu thành các lô nhỏ, theo khoảng thời gian và xử lý**
  - Thiết kế và triển khai dễ dàng hơn so với mô hình trên.
  - Khâu xử lý vẫn có độ trễ nhất định.



Stream Process at Viettel (1)

- Xử lý khoảng ~50 - 60 tỉ message/ ngày
- Mô hình micro-batching.
- Input rate cao, ~ 1.5 triệu bản ghi/s
- Khối lượng xử lý 1 batch ~250-300 triệu record.
- Sử dụng ~1.1TB memory
- Hoạt động 24/7

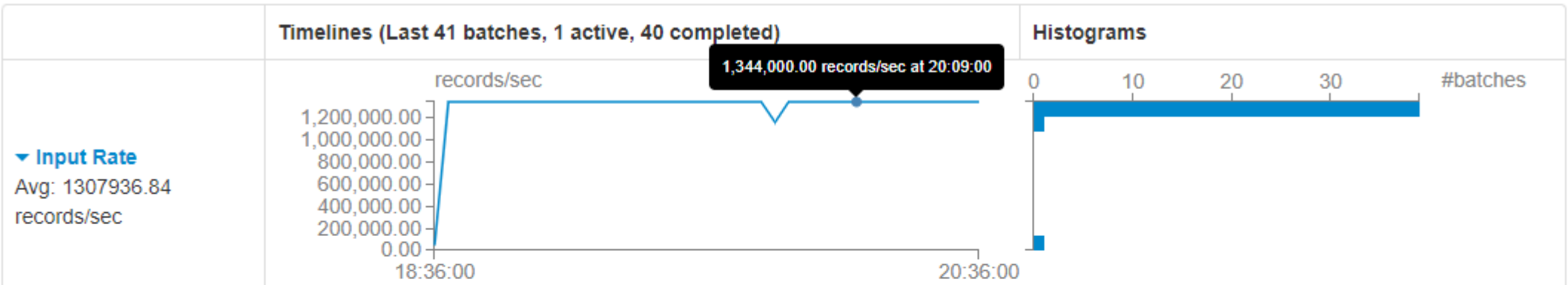
Stream Process at Viettel (2)

Streaming Statistics

Running batches of **3 minutes** for **2 hours 16 seconds** since **2020/11/19 18:35:55** (**40** completed batches, **9652573846** records)

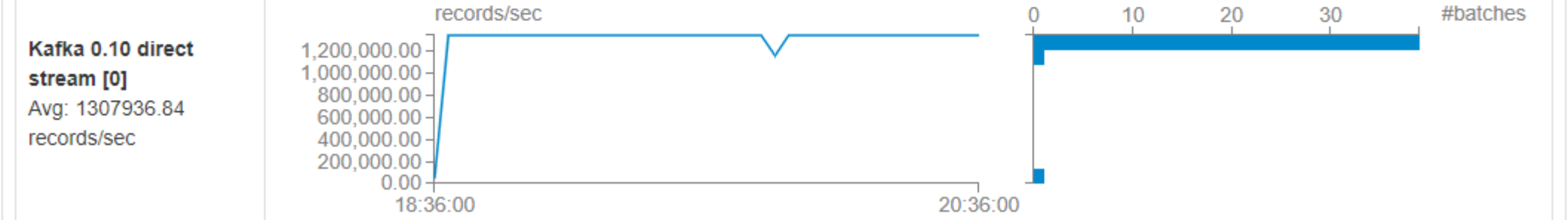
Timelines (Last 41 batches, 1 active, 40 completed)

Histograms



▼ **Input Rate**
Avg: 1307936.84
records/sec

Status	Executor ID / Host	Last Error Time	Last Error Message
-	-	-	-



Kafka 0.10 direct stream [0]
Avg: 1307936.84
records/sec

Stream Process at Viettel (3)

Active Batches (1)

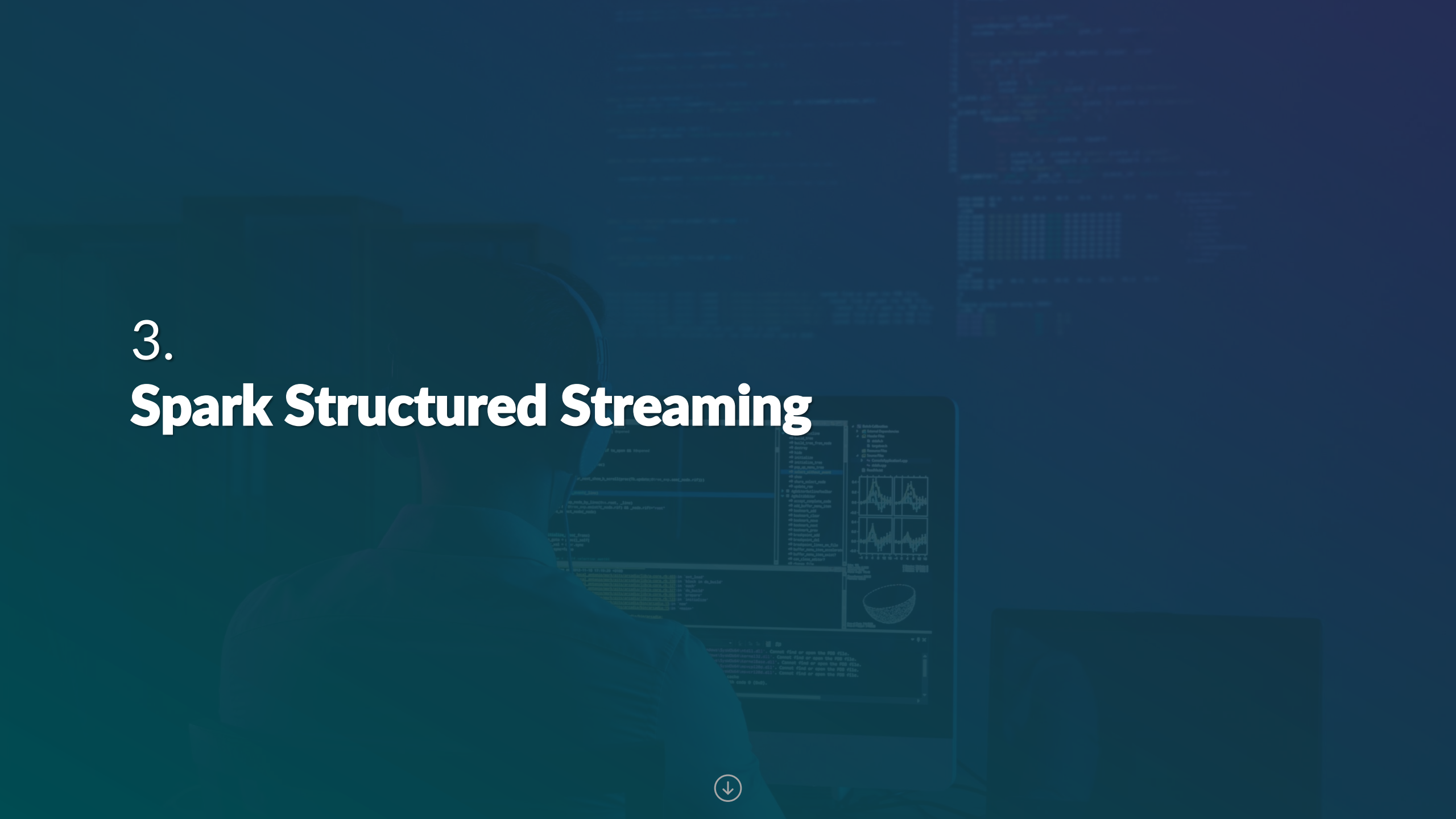
Batch Time	Records	Scheduling Delay (?)	Processing Time (?)	Output Ops: Succeeded/Total	Status
2020/11/19 20:42:00	241920000 records	0 ms	-	0/1 (1 running)	processing

Completed Batches (last 42 out of 42)

Batch Time	Records	Scheduling Delay (?)	Processing Time (?)	Total Delay (?)	Output Ops: Succeeded/Total
2020/11/19 20:39:00	241920000 records	1 ms	2.6 min	2.6 min	1/1
2020/11/19 20:36:00	241920000 records	1 ms	2.6 min	2.6 min	1/1
2020/11/19 20:33:00	241920000 records	1 ms	2.6 min	2.6 min	1/1
2020/11/19 20:30:00	241920000 records	1 ms	2.6 min	2.6 min	1/1
2020/11/19 20:27:00	241920000 records	1 ms	2.5 min	2.5 min	1/1
2020/11/19 20:24:00	241920000 records	1 ms	2.6 min	2.6 min	1/1
2020/11/19 20:21:00	241920000 records	1 ms	2.6 min	2.6 min	1/1
2020/11/19 20:18:00	241920000 records	1 ms	2.6 min	2.6 min	1/1
2020/11/19 20:15:00	241920000 records	0 ms	2.6 min	2.6 min	1/1
2020/11/19 20:12:00	241920000 records	5 ms	2.6 min	2.6 min	1/1
2020/11/19 20:09:00	241920000 records	0 ms	2.6 min	2.6 min	1/1
2020/11/19 20:06:00	241920000 records	1 ms	2.6 min	2.6 min	1/1
2020/11/19 20:03:00	241920000 records	0 ms	2.7 min	2.7 min	1/1

THẢO LUẬN

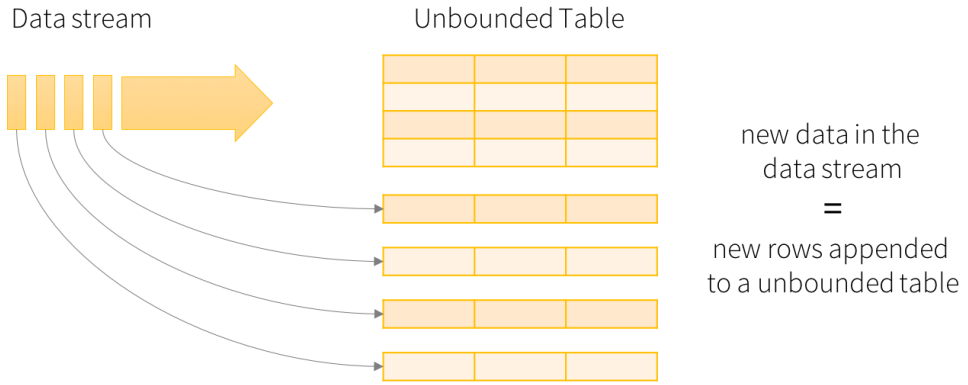
- Các nguồn dữ liệu dạng stream tại đơn vị.
- Nhu cầu với các ứng dụng stream processing.
- Các giải pháp hiện nay đã áp dụng và còn đang nghiên cứu?



3.
Spark Structured Streaming

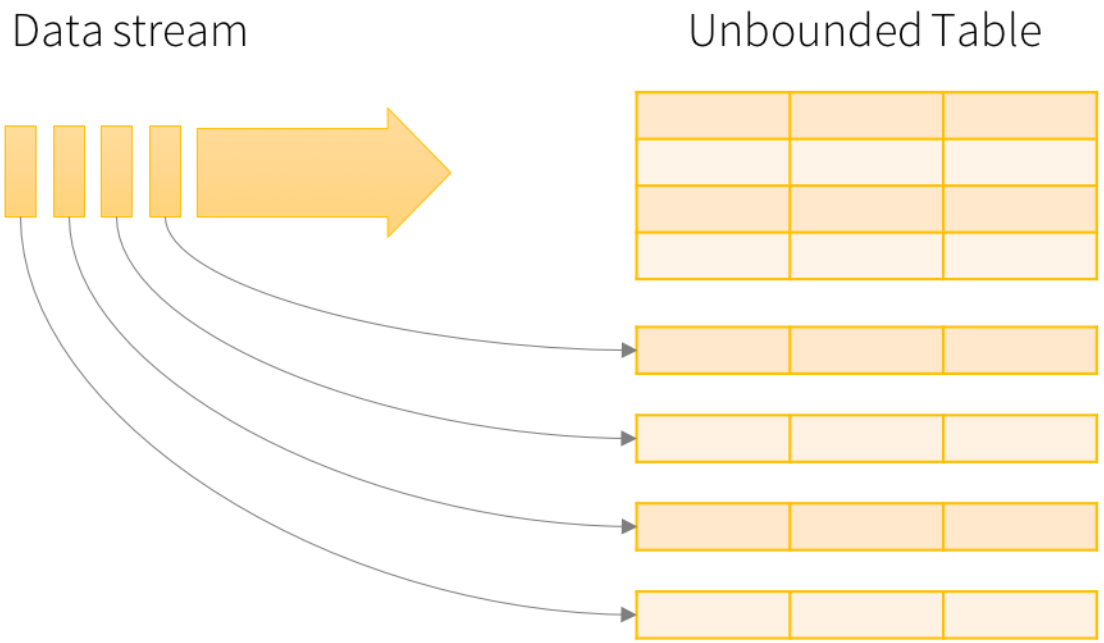
Concept

- Coi dữ liệu stream như là 1 bảng không giới hạn (Unbounded Table)
- Dữ liệu bên trong vẫn là DataFrame.
- Có thể sử dụng SparkSQL APIs với dữ liệu Stream tương tự như dữ liệu tĩnh.
- Phiên bản hiện tại đã hỗ trợ cả mô hình event-at-a-time và micro-batching



Data stream as an unbounded table

Concept (2)

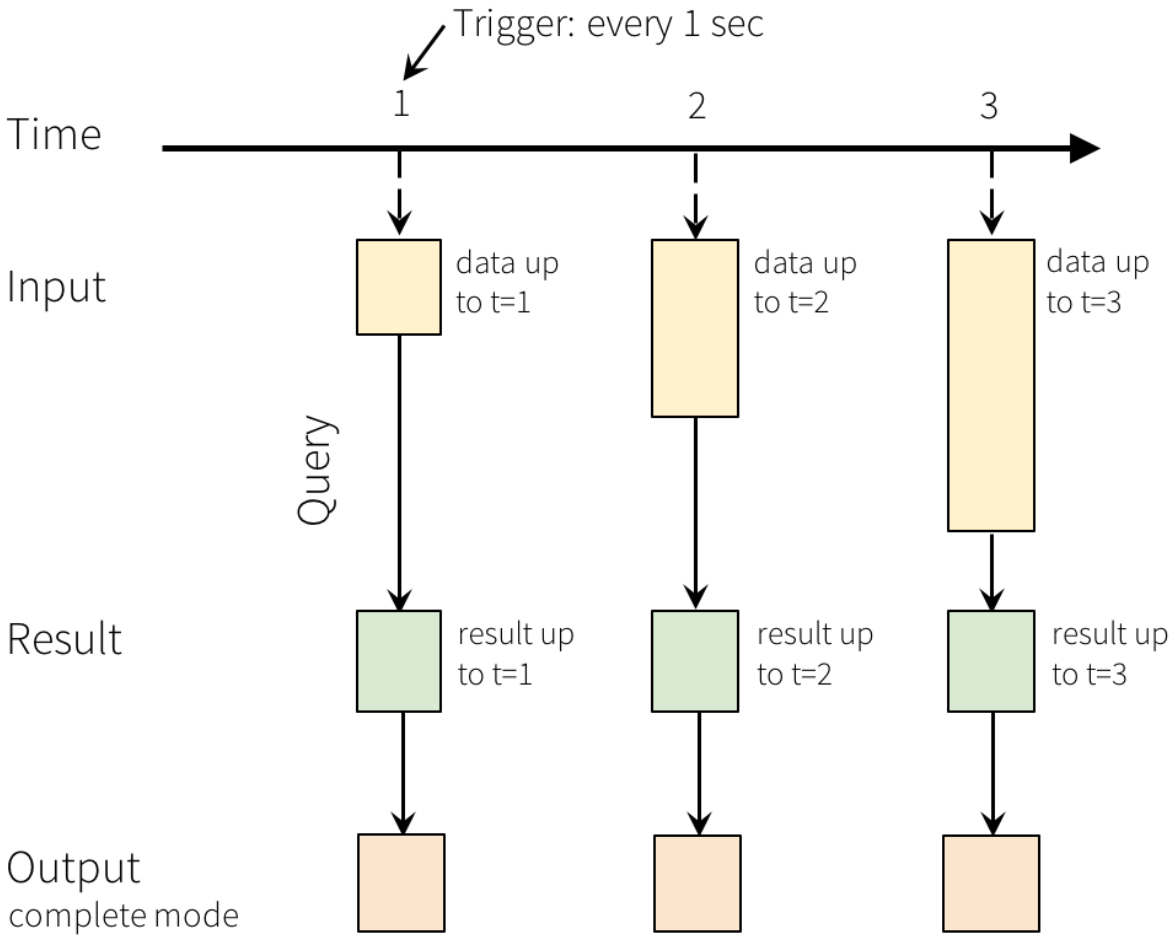


new data in the
data stream
=
new rows appended
to a unbounded table

Data stream as an unbounded table

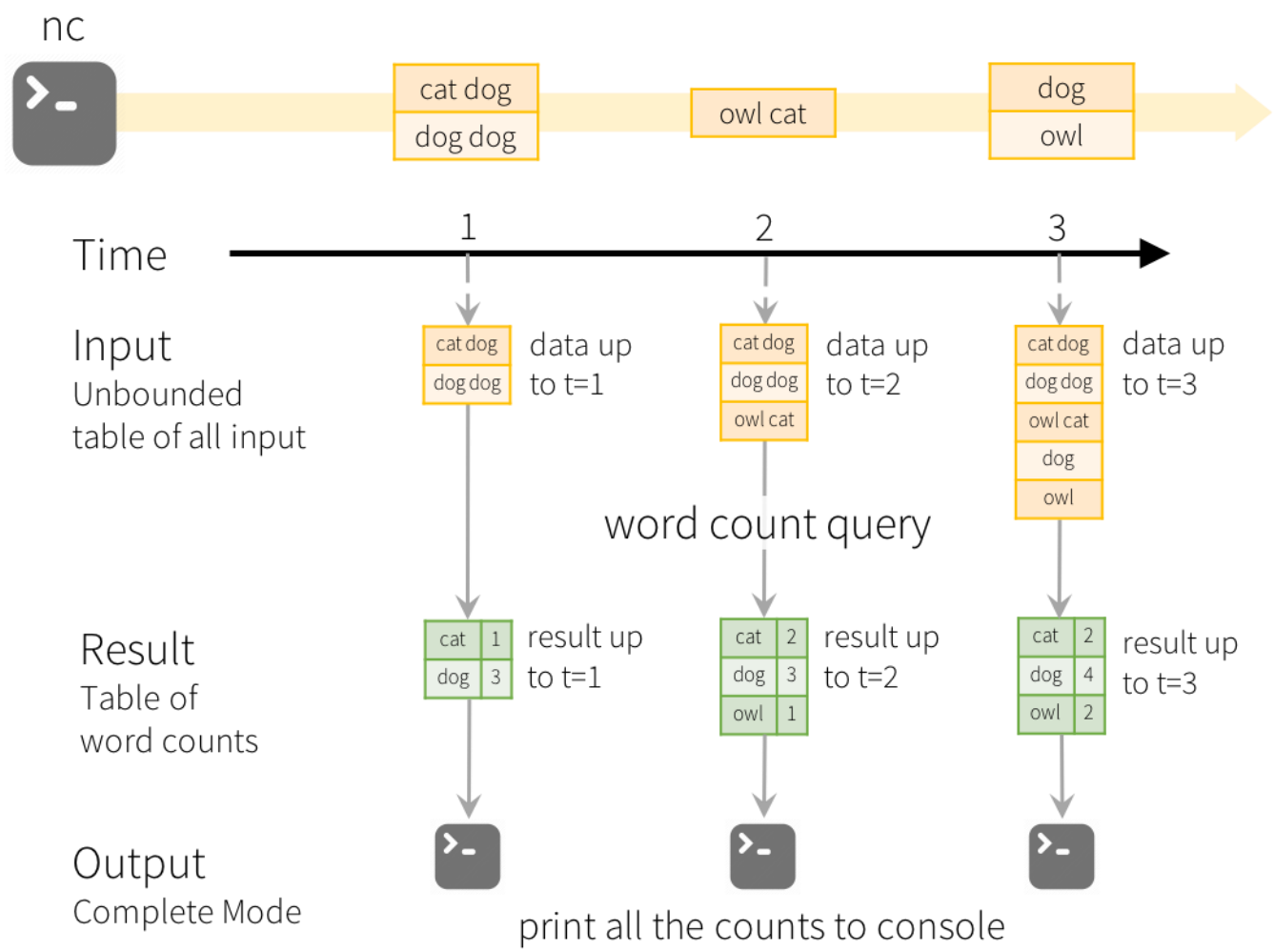
Concept (3)

- Ứng dụng "*liên tục*" nhận dữ liệu từ nguồn.
- Logic xử lý được thực hiện theo trigger đã cấu hình



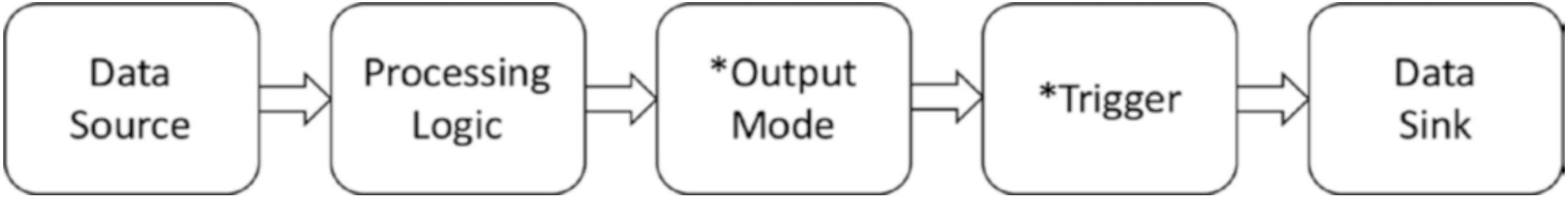
Programming Model for Structured Streaming

Concept (4)



Model of the Quick Example

Spark Structured Streaming Application



- Datasource: Nguồn dữ liệu
- Processing Logic: Logic xử lý dữ liệu
- Output Mode: Cách ghi dữ liệu ra bên ngoài
- Trigger: Quy định khi nào thì thực thi logic xử lý
- Data Sink: Đầu ra của dữ liệu

4.
Thực hành