

Analyzing IoT Data Using Azure Stream Analytics

GVHD:Ths. Hà Lê Hoài Trung



Nội dung

- 01** Giới thiệu bài toán
- 02** Cơ sở lý thuyết
- 03** Mô hình dữ liệu
- 04** Demo

1. Giới thiệu bài toán

Ngày nay, lượng dữ liệu thời gian thực khổng lồ được tạo ra bởi các ứng dụng kết nối, thiết bị và cảm biến Internet of Things (IoT), cùng với nhiều nguồn khác. Sự phát triển của các nguồn dữ liệu trực tuyến đã khiến khả năng tiêu thụ và đưa ra quyết định thông tin từ những dữ liệu này gần như ngay lập tức trở thành một yêu cầu hoạt động đối với nhiều tổ chức.

Azure Stream Analytics cung cấp một bộ xử lý dữ liệu thời gian thực dựa trên đám mây, chúng ta có thể sử dụng để lọc, tổng hợp và xử lý luồng dữ liệu thời gian thực từ các nguồn khác nhau.



2.Cơ sở lý thuyết

a.Giới thiệu chung

Azure Stream Analytics là một dịch vụ trong nền tảng đám mây của Microsoft Azure, được thiết kế để xử lý và phân tích dữ liệu dòng (streaming data) từ nhiều nguồn khác nhau như cảm biến, máy chủ, thiết bị IoT và ứng dụng trực tuyến. Dịch vụ này cung cấp khả năng xử lý dữ liệu thời gian thực, đồng thời và có khả năng mở rộng linh hoạt để xử lý các luồng dữ liệu lớn.

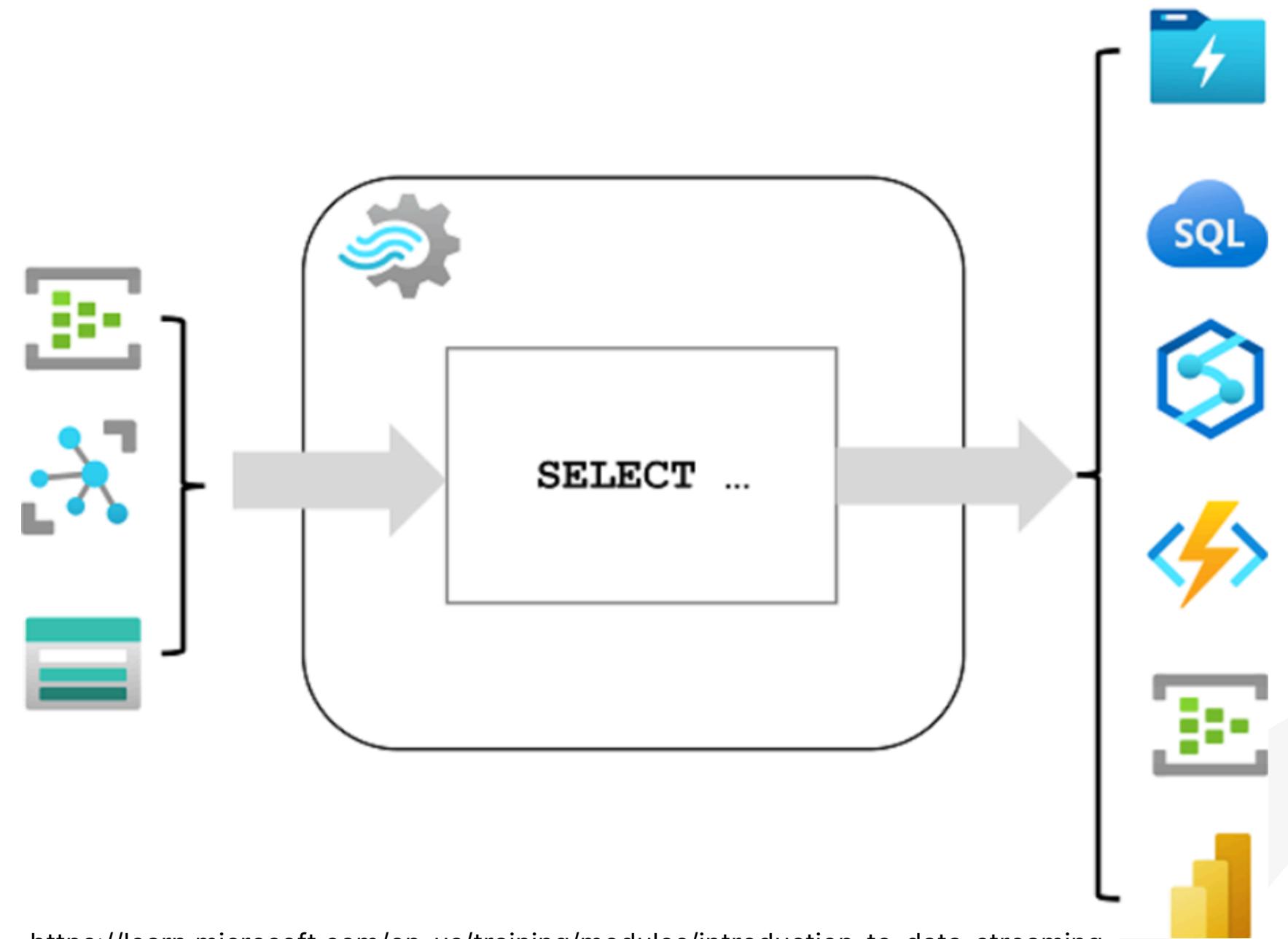
Azure Stream Analytics



2. Cơ sở lý thuyết

a. Giới thiệu chung

- Dữ liệu đầu vào được nhập từ: Azure event hub, Azure IoT Hub, or Azure Storage blob container
- Xử lý dữ liệu bằng cách sử dụng các câu truy vấn để chọn, chiếu, hoặc tổng hợp dữ liệu.
- Kết quả dữ liệu đầu ra được ghi chằng hạn như là Azure Data Lake Gen 2, Azure SQL Database, Azure Synapse Analytics, Azure Functions, Azure event hub, Microsoft Power BI,...



<https://learn.microsoft.com/en-us/training/modules/introduction-to-data-streaming>

2. Cơ sở lý thuyết

a. Giới thiệu chung

Một số tính năng của Azure Stream Analytics:

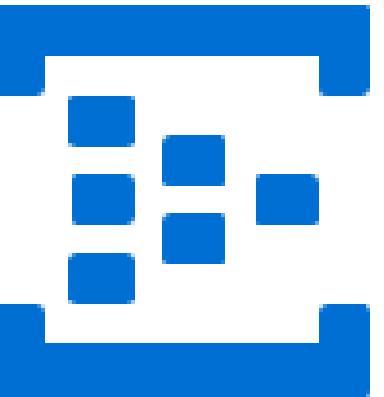
- Exactly Once Event Processing
- At-Least-Once Event Delivery
- Recovery Capabilities
- Checkpointing
- Azure Stream Analytics là một platform-as-a-service(PaaS) nên nó cung cấp một mô hình lập trình linh hoạt và có độ tin cậy cao, và có hiệu suất cao, bởi vì cho phép tính toán trong bộ nhớ. Sử dụng SQL cho ngôn ngữ truy vấn.



b. Đầu vào(Inputs)

Azure Stream Analytics có thể nhận dữ liệu đầu vào từ nhiều nguồn khác nhau như:

- Azure Event Hubs.
- Azure IoT Hub.
- Azure Blob storage.
- Azure Data Lake Storage Gen2.
- Ngoài ra thì cũng có thể định nghĩa các đầu vào tham chiếu(reference inputs) dùng để nhập dữ liệu tinh bô sung dữ liệu theo thời gian thực.



Event hub



IoT hub



Azure Data Lake Storage



Azure Blob Storage

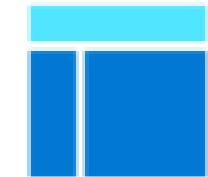
c. Đầu ra(Outputs)

Là đích đến mà kết quả của quá trình truyền tải dữ liệu được gửi tới. Azure Stream Analytics hỗ trợ một loạt các kết quả đầu ra:

- Duy trì kết quả của quá trình Stream để có thể thực hiện phân tích thêm, bằng cách lưu chúng vào trong Data Lake, hoặc kho dữ liệu.
- Hiển thị trực quan hóa dữ liệu luồng dữ liệu theo thời gian thực, bằng cách thêm dữ liệu vào tập dữ liệu trong Power BI.
- Tạo ra các bộ lọc hay tóm tắt để có thể xử lý tiếp theo. Có thể viết kết quả vào trong Event Hub.



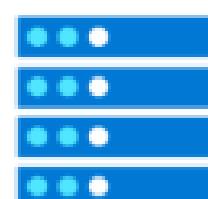
Alerts and actions
Event Hubs, Service Bus,
Azure Functions etc



Dynamic Dashboarding
Power BI



Data Warehousing
Azure Synapse
Analytics



Storage/ Archival
SQL DB, Azure Data Lake Gen 1 &
Gen 2, Cosmos DB, Blob storage, etc

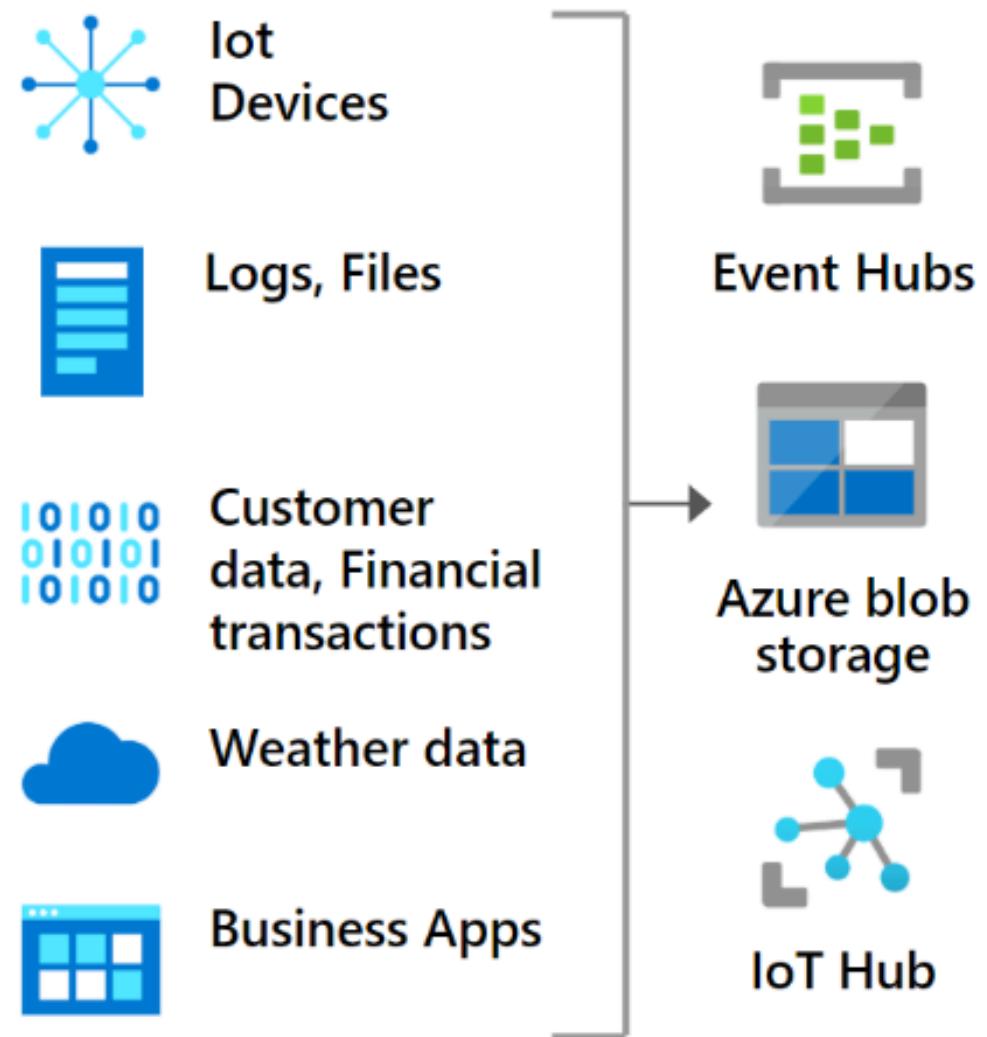
d. Truy vấn(Query)

Logic xử lý dòng được đóng gói trong một truy vấn. Truy vấn được định nghĩa bằng cách sử dụng các câu lệnh SQL SELECT để lấy các trường dữ liệu từ(FROM) một hoặc nhiều nguồn đầu vào, lọc hoặc tổng hợp dữ liệu và ghi kết quả vào(INTO) một đầu ra.

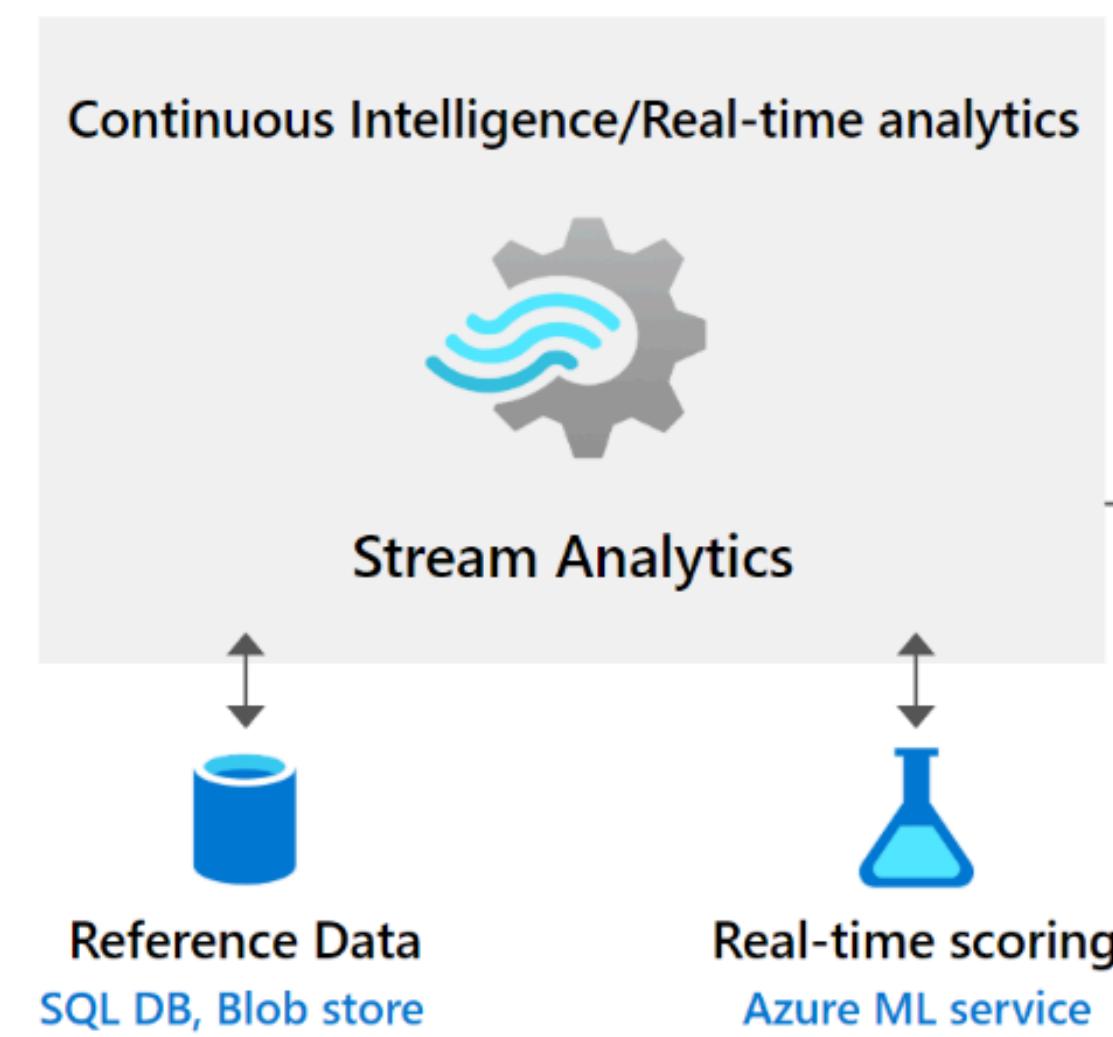
```
SELECT observation_time, weather_station,  
temperature  
INTO cold-temp  
FROM weather-events TIMESTAMP BY  
observation_time  
WHERE temperature < 0
```



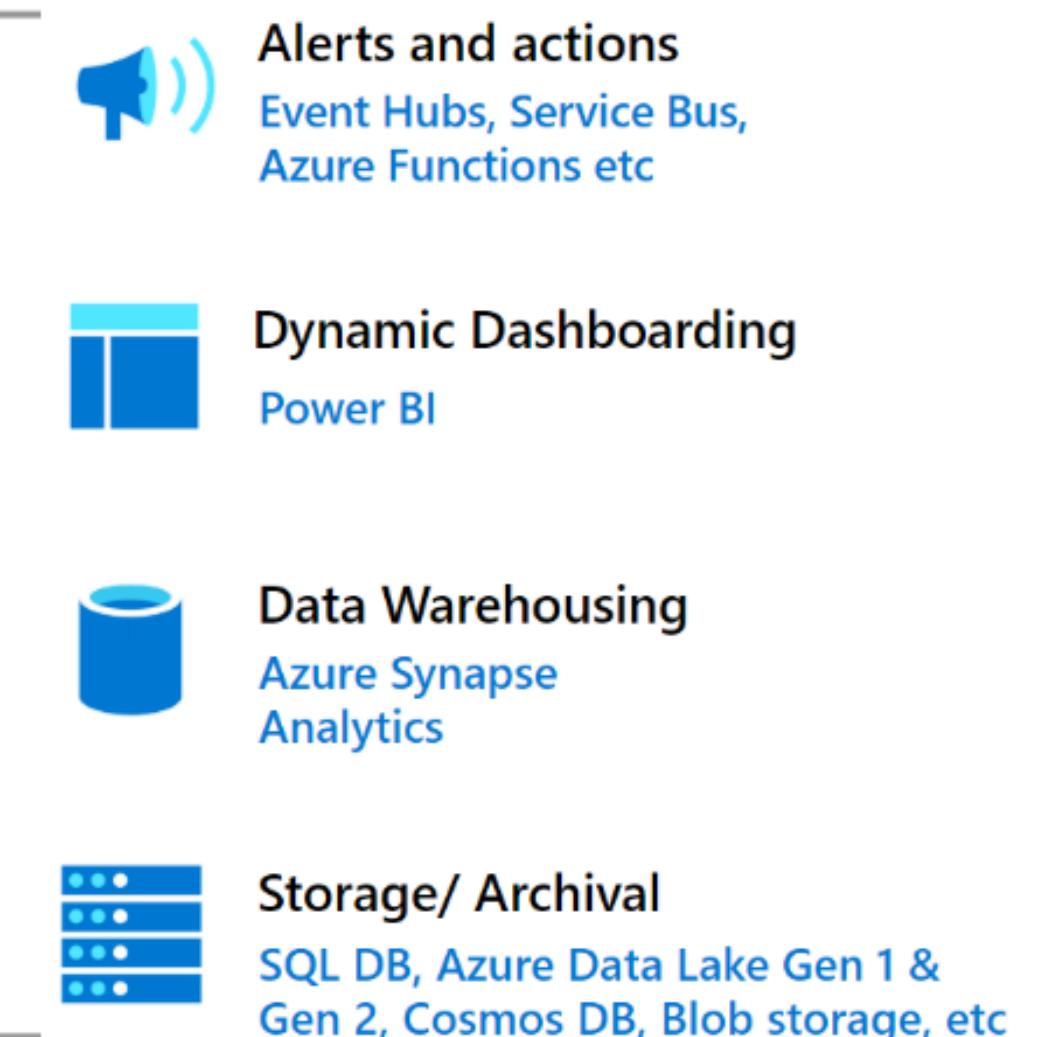
Ingest



Analyze



Deliver



<https://learn.microsoft.com/en-us/azure/stream-analytics/stream-analytics-introduction>

e. Window function

Mục tiêu phổ biến quá trình truyền dữ liệu là tổng hợp các sự kiện vào các khoảng thời gian, hoặc cửa sổ thời gian.

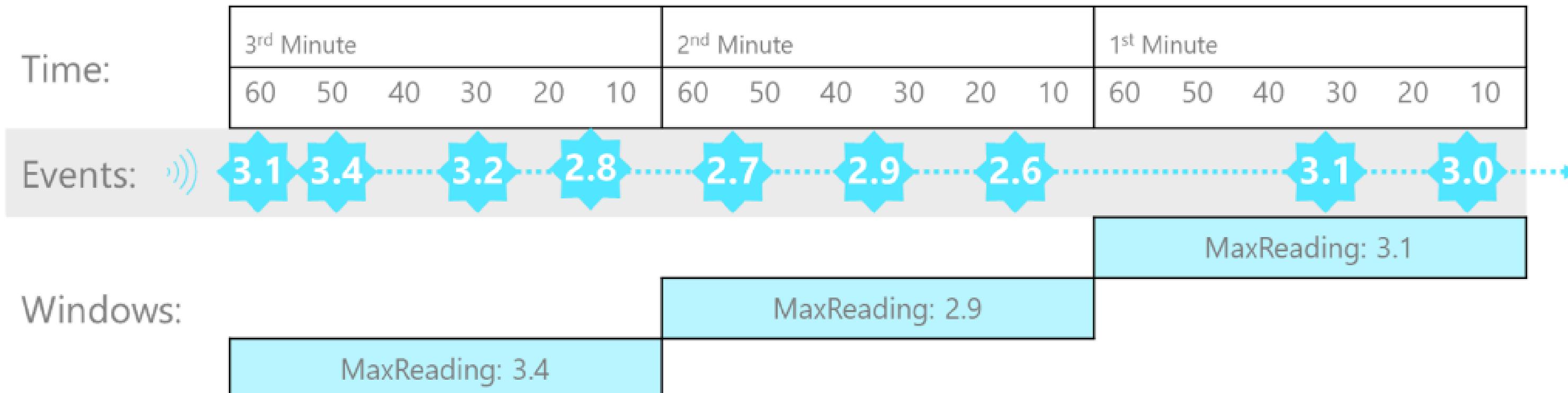
Azure Stream Analytics hỗ trợ nguyên bản cho năm loại hàm cửa sổ thời gian. Các hàm này cho phép xác định các khoảng thời gian mà dữ liệu được tổng hợp trong một truy vấn. Các hàm cửa sổ được hỗ trợ bao gồm:

- Tumbling,
- Hopping,
- Sliding,
- Session
- Snapshot.



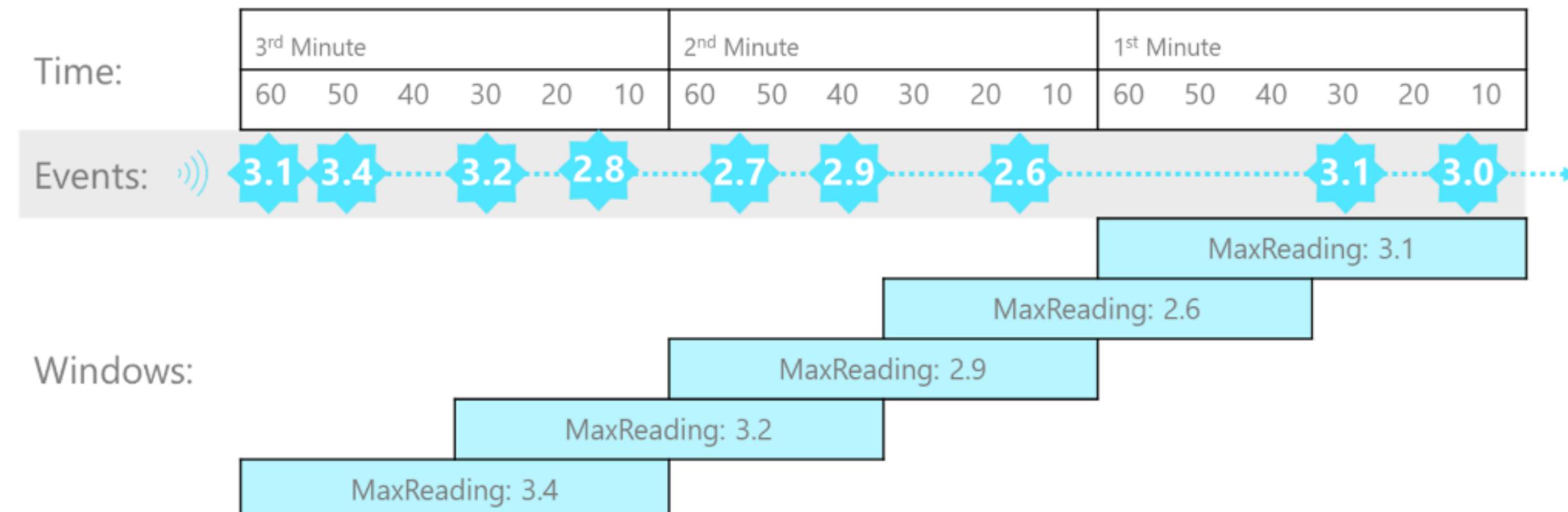
Tumbling

```
SELECT DateAdd(minute,-1,System.TimeStamp) AS WindowStart,  
       System.TimeStamp() AS WindowEnd,  
       MAX(Reading) AS MaxReading  
INTO  
       [output]  
FROM  
       [input] TIMESTAMP BY EventProcessedUtcTime  
GROUP BY TumblingWindow(minute, 1)
```



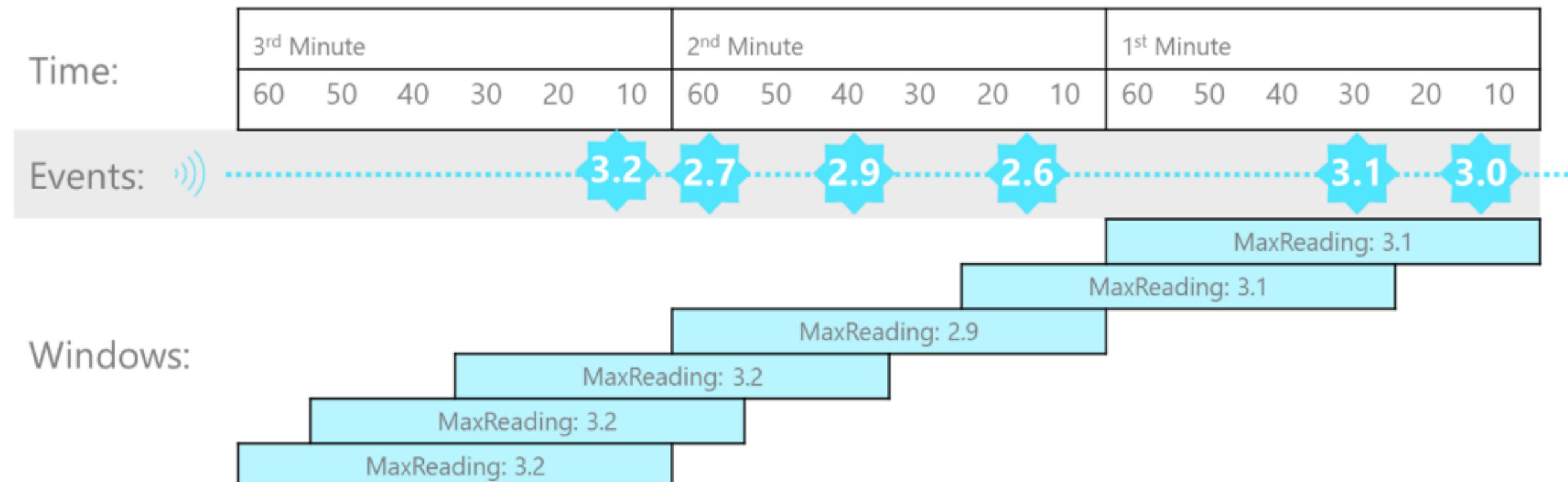
Hopping

```
SELECT DateAdd(second,-60,System.TimeStamp) AS WindowStart,  
System.TimeStamp() AS WindowEnd,  
MAX(Reading) AS MaxReading  
INTO  
[output]  
FROM  
[input] TIMESTAMP BY EventProcessedUtcTime  
GROUP BY HoppingWindow(second, 60, 30)
```



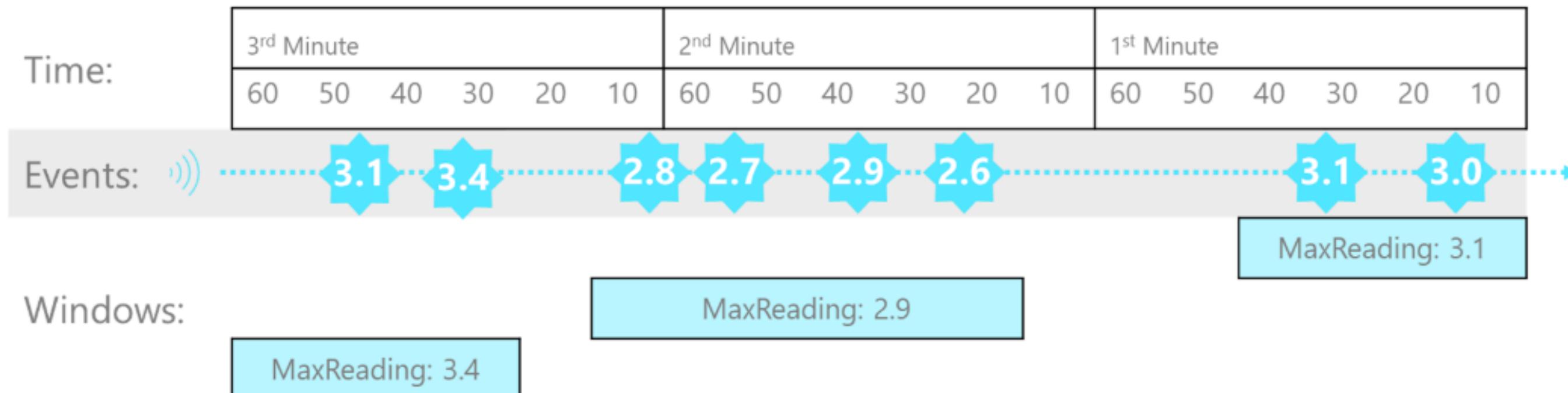
Sliding

```
SELECT DateAdd(minute,-1,System.TimeStamp) AS WindowStart,  
       System.TimeStamp() AS WindowEnd,  
       MAX(Reading) AS MaxReading  
INTO  
       [output]  
FROM  
       [input] TIMESTAMP BY EventProcessedUtcTime  
GROUP BY SlidingWindow(minute, 1)
```



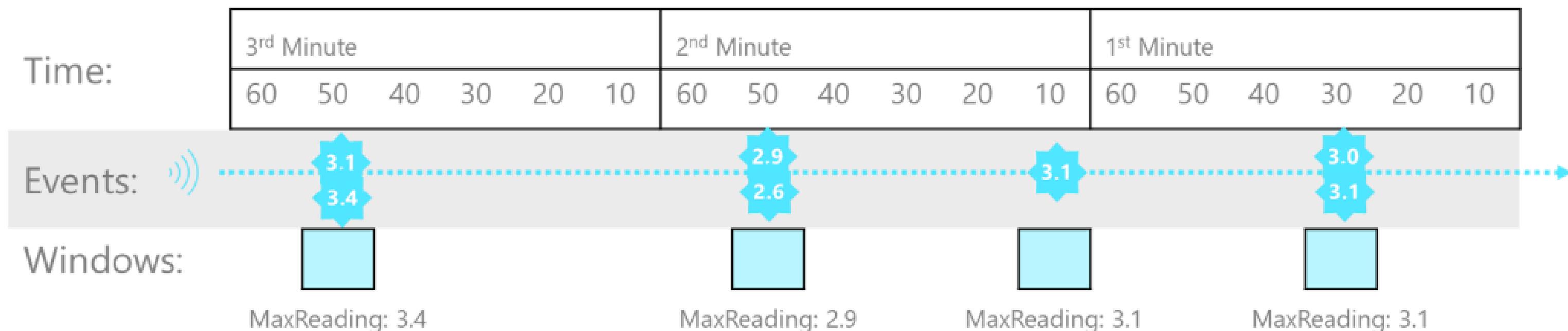
Session

```
SELECT DateAdd(second,-60,System.TimeStamp) AS WindowStart,  
       System.TimeStamp() AS WindowEnd,  
       MAX(Reading) AS MaxReading  
INTO  
       [output]  
FROM  
       [input] TIMESTAMP BY EventProcessedUtcTime  
GROUP BY SessionWindow(second, 20, 60)
```



Snapshot

```
SELECT System.TimeStamp() AS WindowTime,  
       MAX(Reading) AS MaxReading  
  INTO  
    [output]  
   FROM  
    [input] TIMESTAMP BY EventProcessedUtcTime  
 GROUP BY System.Timestamp()
```



Định nghĩa câu truy vấn

```
SELECT  
    EventEnqueuedUtcTime AS ReadingTime,  
    SensorID,  
    ReadingValue  
INTO  
    [synapse-output]  
FROM  
    [streaming-input] TIMESTAMP BY EventEnqueuedUtcTime
```

Định nghĩa câu truy vấn

```
SELECT  
    EventEnqueuedUtcTime AS ReadingTime,  
    SensorID,  
    ReadingValue  
INTO  
    [synapse-output]  
FROM  
    [streaming-input] TIMESTAMP BY EventEnqueuedUtcTime  
WHERE ReadingValue < 0
```

Định nghĩa câu truy vấn

```
SELECT  
    DateAdd(second, -60, System.TimeStamp) AS StartTime,  
    System.TimeStamp AS EndTime,  
    SensorID,  
    MAX(ReadingValue) AS MaxReading  
INTO  
    [synapse-output]  
FROM  
    [streaming-input] TIMESTAMP BY EventEnqueuedUtcTime  
GROUP BY SensorID, TumblingWindow(second, 60)  
HAVING COUNT(*) >= 1
```

Quản lý hiệu suất

Streaming units (SU)

Event Ordering

Tài liệu tham khảo

<https://learn.microsoft.com/en-us/training/paths/implement-data-streaming-with-asa/>

Azure Storage, Streaming, and Batch Analytics
A guide for data engineers - Richard L.Nuckolls

**CẢM ƠN THẦY VÀ CÁC BẠN
ĐÃ CHÚ Ý LẮNG NGHE**