**Explore common elements of stream processing architecture**

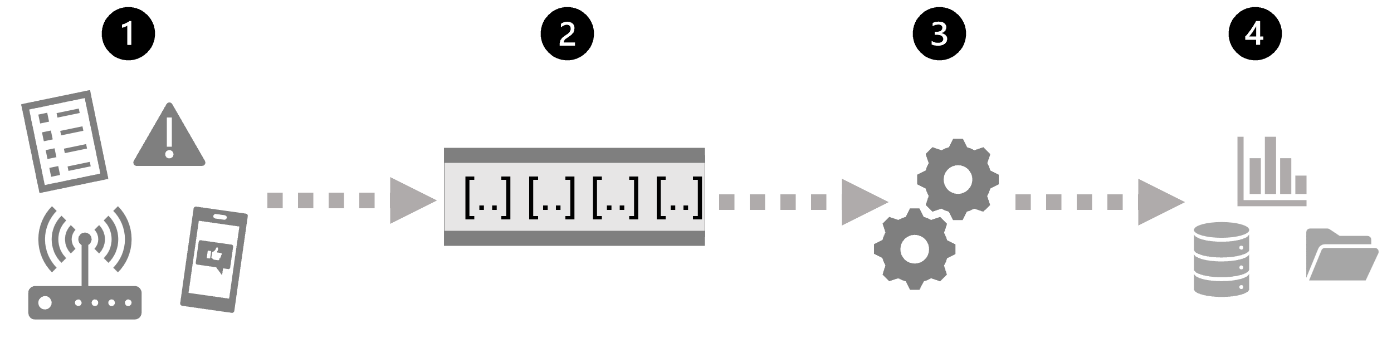
Completed100 XP

* 4 minutes

There are many technologies that you can use to implement a stream processing solution, but while specific implementation details may vary, there are common elements to most streaming architectures.

**A general architecture for stream processing**

At its simplest, a high-level architecture for stream processing looks like this:



1. An event generates some data. This might be a signal being emitted by a sensor, a social media message being posted, a log file entry being written, or any other occurrence that results in some digital data.
2. The generated data is captured in a streaming *source* for processing. In simple cases, the source may be a folder in a cloud data store or a table in a database. In more robust streaming solutions, the source may be a "queue" that encapsulates logic to ensure that event data is processed in order and that each event is processed only once.
3. The event data is processed, often by a perpetual query that operates on the event data to select data for specific types of events, project data values, or aggregate data values over temporal (time-based) periods (or *windows*) - for example, by counting the number of sensor emissions per minute.
4. The results of the stream processing operation are written to an output (or *sink*), which may be a file, a database table, a real-time visual dashboard, or another queue for further processing by a subsequent downstream query.

**Real-time analytics in Azure**

Microsoft Azure supports multiple technologies that you can use to implement real-time analytics of streaming data, including:

* **Azure Stream Analytics**: A platform-as-a-service (PaaS) solution that you can use to define *streaming jobs* that ingest data from a streaming source, apply a perpetual query, and write the results to an output.
* **Spark Structured Streaming**: An open-source library that enables you to develop complex streaming solutions on Apache Spark based services, including **Azure Synapse Analytics**, **Azure Databricks**, and **Azure HDInsight**.
* **Azure Data Explorer**: A high-performance database and analytics service that is optimized for ingesting and querying batch or streaming data with a time-series element, and which can be used as a standalone Azure service or as an **Azure Synapse Data Explorer** runtime in an Azure Synapse Analytics workspace.

***Sources* for stream processing**

The following services are commonly used to ingest data for stream processing on Azure:

* **Azure Event Hubs**: A data ingestion service that you can use to manage queues of event data, ensuring that each event is processed in order, exactly once.
* **Azure IoT Hub**: A data ingestion service that is similar to Azure Event Hubs, but which is optimized for managing event data from *Internet-of-things* (IoT) devices.
* **Azure Data Lake Store Gen 2**: A highly scalable storage service that is often used in *batch processing* scenarios, but which can also be used as a source of streaming data.
* **Apache Kafka**: An open-source data ingestion solution that is commonly used together with Apache Spark. You can use Azure HDInsight to create a Kafka cluster.

***Sinks* for stream processing**

The output from stream processing is often sent to the following services:

* **Azure Event Hubs**: Used to queue the processed data for further downstream processing.
* **Azure Data Lake Store Gen 2** or **Azure blob storage**: Used to persist the processed results as a file.
* **Azure SQL Database** or **Azure Synapse Analytics**, or **Azure Databricks**: Used to persist the processed results in a database table for querying and analysis.
* **Microsoft Power BI**: Used to generate real time data visualizations in reports and dashboards.

**Khám phá các yếu tố phổ biến của kiến trúc xử lý theo luồng**

Có nhiều công nghệ bạn có thể sử dụng để triển khai giải pháp xử lý theo luồng, nhưng trong khi các chi tiết triển khai cụ thể có thể khác nhau, thì hầu hết các kiến trúc xử lý theo luồng đều có các yếu tố chung.

**Kiến trúc tổng hợp cho xử lý theo luồng**

Secara đơn giản nhất, kiến trúc cấp cao cho xử lý theo luồng trông giống như thế này:

1. **Một sự kiện tạo ra một số dữ liệu.** Đây có thể là tín hiệu phát ra từ cảm biến, tin nhắn trên mạng xã hội được đăng, mục nhập tệp nhật ký được viết hoặc bất kỳ sự kiện nào khác dẫn đến một số dữ liệu kỹ thuật số.
2. **Dữ liệu được tạo được ghi lại trong một nguồn theo luồng để xử lý.** Trong các trường hợp đơn giản, nguồn có thể là một thư mục trong kho lưu trữ dữ liệu đám mây hoặc bảng trong cơ sở dữ liệu. Trong các giải pháp xử lý theo luồng mạnh mẽ hơn, nguồn có thể là một "đám đợi" bao gồm logic để đảm bảo dữ liệu sự kiện được xử lý theo thứ tự và mỗi sự kiện chỉ được xử lý một lần.
3. **Dữ liệu sự kiện được xử lý**, thường là bởi một truy vấn liên tục hoạt động trên dữ liệu sự kiện để chọn dữ liệu cho các loại sự kiện cụ thể, giá trị dữ liệu dự án hoặc tổng hợp các giá trị dữ liệu theo các khoảng thời gian (theo thời gian) (hoặc **cửa sổ**) - ví dụ, bằng cách đếm số lần phát ra cảm biến mỗi phút.
4. **Kết quả của hoạt động xử lý theo luồng được ghi vào một đầu ra (hoặc sink)**, có thể là tệp, bảng cơ sở dữ liệu, bảng điều khiển trực quan thời gian thực hoặc một hàng đợi khác để xử lý thêm bởi truy vấn hạ lưu tiếp theo.

**Phân tích theo thời gian thực trong Azure**

Azure của Microsoft hỗ trợ nhiều công nghệ mà bạn có thể sử dụng để triển khai phân tích theo thời gian thực của dữ liệu theo luồng, bao gồm:

* **Azure Stream Analytics:** Giải pháp nền tảng dịch vụ (PaaS) mà bạn có thể sử dụng để xác định **作业 theo luồng (streaming jobs)** thu thập dữ liệu từ nguồn theo luồng, áp dụng truy vấn liên tục và ghi kết quả vào đầu ra.
* **Spark Structured Streaming:** Thư viện mã nguồn mở cho phép bạn phát triển các giải pháp xử lý theo luồng phức tạp trên các dịch vụ dựa trên Apache Spark, bao gồm Azure Synapse Analytics, Azure Databricks và Azure HDInsight.
* **Azure Data Explorer:** Dịch vụ cơ sở dữ liệu và phân tích hiệu suất cao được tối ưu hóa cho việc nhập và truy vấn dữ liệu theo lô hoặc theo luồng với yếu tố theo chuỗi thời gian, có thể được sử dụng như một dịch vụ Azure độc lập hoặc là **Azure Synapse Data Explorer runtime** trong không gian làm việc Azure Synapse Analytics.

**Nguồn cho xử lý theo luồng**

Các dịch vụ sau đây thường được sử dụng để nhập dữ liệu cho xử lý theo luồng trên Azure:

* **Azure Event Hubs:** Dịch vụ nhập dữ liệu mà bạn có thể sử dụng để quản lý hàng đợi dữ liệu sự kiện, đảm bảo mỗi sự kiện được xử lý theo thứ tự, chính xác một lần.
* **Azure IoT Hub:** Dịch vụ nhập dữ liệu tương tự như Azure Event Hubs, nhưng được tối ưu hóa để quản lý dữ liệu sự kiện từ các thiết bị Internet of Things (IoT).
* **Azure Data Lake Store Gen 2:** Dịch vụ lưu trữ có khả năng mở rộng cao thường được sử dụng trong các tình huống xử lý theo lô, nhưng cũng có thể được sử dụng làm nguồn dữ liệu theo luồng.
* **Apache Kafka:** Giải pháp nhập dữ liệu mã nguồn mở thường được sử dụng cùng với Apache Spark. Bạn có thể sử dụng Azure HDInsight để tạo cụm Kafka.

**Sink cho xử lý theo luồng**

Kết quả đầu ra từ xử lý theo luồng thường được gửi đến các dịch vụ sau:

* **Azure Event Hubs:** Được sử dụng để xếp hàng dữ liệu đã xử lý để xử lý thêm ở hạ lưu.
* **Azure Data Lake Store Gen 2 hoặc Azure Blob storage:** Được sử dụng để duy trì kết quả đã xử lý dưới dạng tệp.
* **Azure SQL Database hoặc Azure Synapse Analytics, hoặc Azure Databricks:** Được sử dụng để duy trì kết quả đã xử lý trong bảng cơ sở dữ liệu để truy vấn và phân tích.
* Microsoft Power BI: Được sử dụng để tạo các hình ảnh hóa dữ liệu theo thời gian thực trong báo cáo