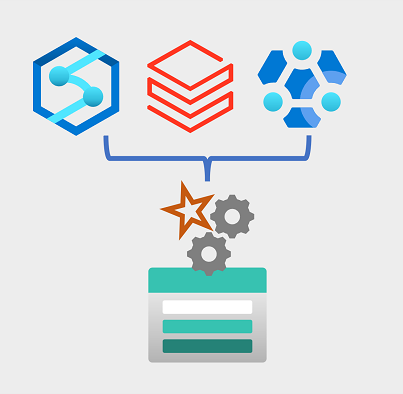
**Use Azure Data Lake Storage Gen2 in data analytics workloads**

Completed100 XP

* 5 minutes

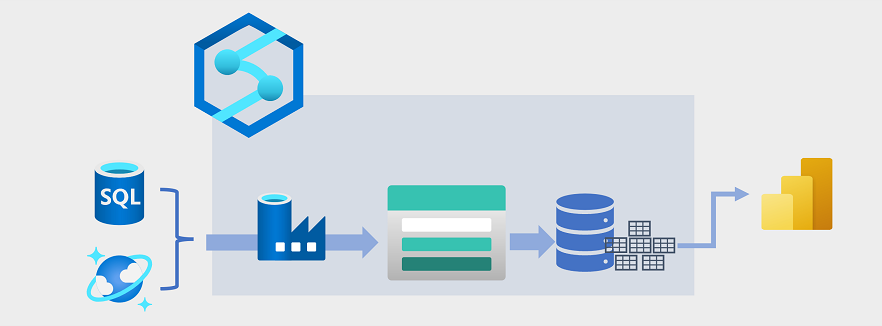
Azure Data Lake Store Gen2 is an enabling technology for multiple data analytics use cases. Let's explore a few common types of analytical workload, and identify how Azure Data Lake Storage Gen2 works with other Azure services to support them.

**Big data processing and analytics**



Big data scenarios usually refer to analytical workloads that involve massive *volumes* of data in a *variety* of formats that needs to be processed at a fast *velocity* - the so-called "three v's". Azure Data Lake Storage Gen 2 provides a scalable and secure distributed data store on which big data services such as Azure Synapse Analytics, Azure Databricks, and Azure HDInsight can apply data processing frameworks such as Apache Spark, Hive, and Hadoop. The distributed nature of the storage and the processing compute enables tasks to be performed in parallel, resulting in high-performance and scalability even when processing huge amounts of data.

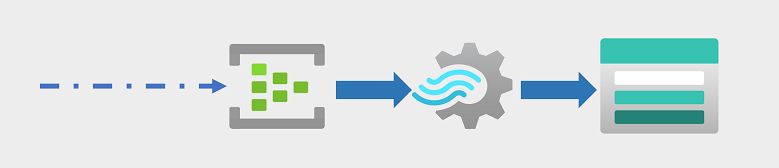
**Data warehousing**



Data warehousing has evolved in recent years to integrate large volumes of data stored as files in a data lake with relational tables in a data warehouse. In a typical example of a data warehousing solution, data is extracted from operational data stores, such as Azure SQL database or Azure Cosmos DB, and transformed into structures more suitable for analytical workloads. Often, the data is staged in a data lake in order to facilitate distributed processing before being loaded into a relational data warehouse. In some cases, the data warehouse uses *external* tables to define a relational metadata layer over files in the data lake and create a hybrid "data lakehouse" or "lake database" architecture. The data warehouse can then support analytical queries for reporting and visualization.

There are multiple ways to implement this kind of data warehousing architecture. The diagram shows a solution in which Azure Synapse Analytics hosts *pipelines* to perform *extract, transform, and load* (ETL) processes using Azure Data Factory technology. These processes extract data from operational data sources and load it into a data lake hosted in an Azure Data Lake Storage Gen2 container. The data is then processed and loaded into a relational data warehouse in an Azure Synapse Analytics dedicated SQL pool, from where it can support data visualization and reporting using Microsoft Power BI.

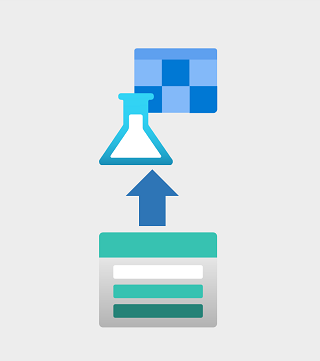
**Real-time data analytics**



Increasingly, businesses and other organizations need to capture and analyze perpetual streams of data, and analyze it in real-time (or as near to real-time as possible). These streams of data can be generated from connected devices (often referred to as *internet-of-things* or *IoT* devices) or from data generated by users in social media platforms or other applications. Unlike traditional *batch processing* workloads, streaming data requires a solution that can capture and process a boundless stream of data events as they occur.

Streaming events are often captured in a queue for processing. There are multiple technologies you can use to perform this task, including Azure Event Hubs as shown in the image. From here, the data is processed, often to aggregate data over temporal windows (for example to count the number of social media messages with a given tag every five minutes, or to calculate the average reading of an Internet connected sensor per minute). Azure Stream Analytics enables you to create *jobs* that query and aggregate event data as it arrives, and write the results in an output *sink*. One such sink is Azure Data Lake Storage Gen2; from where the captured real-time data can be analyzed and visualized.

**Data science and machine learning**



Data science involves the statistical analysis of large volumes of data, often using tools such as Apache Spark and scripting languages such as Python. Azure Data Lake Storage Gen 2 provides a highly scalable cloud-based data store for the volumes of data required in data science workloads.

Machine learning is a subarea of data science that deals with training predictive models. Model training requires huge amounts of data, and the ability to process that data efficiently. Azure Machine Learning is a cloud service in which data scientists can run Python code in notebooks using dynamically allocated distributed compute resources. The compute processes data in Azure Data Lake Storage Gen2 containers to train models, which can then be deployed as production web services to support predictive analytical workloads.

**Sử dụng Azure Data Lake Storage Gen2 trong khối lượng công việc phân tích dữ liệu**

Azure Data Lake Storage Gen2 là nền tảng hỗ trợ cho nhiều trường hợp sử dụng phân tích dữ liệu. Chúng ta hãy khám phá một vài loại khối lượng công việc phân tích phổ biến và xác định cách Azure Data Lake Storage Gen2 hoạt động với các dịch vụ Azure khác để hỗ trợ chúng.

**1. Xử lý và phân tích dữ liệu lớn (Big data processing and analytics)**

Các tình huống dữ liệu lớn thường đề cập đến khối lượng công việc phân tích liên quan đến khối lượng dữ liệu khổng lồ ở nhiều định dạng khác nhau, cần được xử lý với tốc độ nhanh - cái gọi là "ba V" (volume - dung lượng, variety - đa dạng, velocity - tốc độ). Azure Data Lake Storage Gen2 cung cấp kho lưu trữ dữ liệu phân tán, an toàn và có khả năng mở rộng, trên đó các dịch vụ dữ liệu lớn như Azure Synapse Analytics, Azure Databricks và Azure HDInsight có thể áp dụng các khung xử lý dữ liệu như Apache Spark, Hive và Hadoop. Bản chất phân tán của lưu trữ và tính toán xử lý cho phép các tác vụ được thực hiện song song, mang lại hiệu suất cao và khả năng mở rộng ngay cả khi xử lý khối lượng dữ liệu khổng lồ.

**2. Kho dữ liệu (Data warehousing)**

Kho dữ liệu đã phát triển trong những năm gần đây để tích hợp khối lượng lớn dữ liệu được lưu trữ dưới dạng file trong hồ dữ liệu với các bảng quan hệ trong kho dữ liệu. Trong một ví dụ điển hình về giải pháp kho dữ liệu, dữ liệu được trích xuất từ các kho lưu trữ dữ liệu hoạt động, chẳng hạn như Azure SQL Database hoặc Azure Cosmos DB, và được chuyển đổi thành các cấu trúc phù hợp hơn cho khối lượng công việc phân tích. Dữ liệu thường được dàn dựng trong hồ dữ liệu để tạo điều kiện xử lý phân tán trước khi được tải vào kho dữ liệu quan hệ. Trong một số trường hợp, kho dữ liệu sử dụng các bảng ngoài để định nghĩa lớp siêu dữ liệu quan hệ trên các file trong hồ dữ liệu và tạo ra kiến trúc "data lakehouse" hoặc "lake database" lai. Sau đó, kho dữ liệu có thể hỗ trợ các truy vấn phân tích cho báo cáo và trực quan hóa.

Có nhiều cách để triển khai kiến trúc kho dữ liệu kiểu này. Biểu đồ cho thấy giải pháp trong đó Azure Synapse Analytics lưu trữ các pipeline để thực hiện các quy trình trích xuất, chuyển đổi và tải (ETL) bằng công nghệ Azure Data Factory. Các quy trình này trích xuất dữ liệu từ các nguồn dữ liệu hoạt động và tải dữ liệu đó vào hồ dữ liệu được lưu trữ trong vùng chứa Azure Data Lake Storage Gen2. Sau đó, dữ liệu được xử lý và tải vào kho dữ liệu quan hệ trong một nhóm SQL chuyên dụng của Azure Synapse Analytics, từ đó có thể hỗ trợ trực quan hóa dữ liệu và báo cáo bằng Microsoft Power BI.

**3. Phân tích dữ liệu theo thời gian thực (Real-time data analytics)**

Ngày càng có nhiều doanh nghiệp và tổ chức cần thu thập và phân tích các luồng dữ liệu liên tục, đồng thời phân tích chúng theo thời gian thực (hoặc gần thời gian thực nhất có thể). Các luồng dữ liệu này có thể được tạo ra từ các thiết bị được kết nối (thường được gọi là thiết bị Internet vạn vật (IoT)) hoặc từ dữ liệu do người dùng tạo trên các nền tảng truyền thông xã hội hoặc các ứng dụng khác. Không giống như khối lượng công việc xử lý hàng loạt (batch processing) truyền thống, dữ liệu theo luồng đòi hỏi một giải pháp có thể thu thập và xử lý luồng dữ liệu sự kiện vô hạn khi chúng xảy ra.

Các sự kiện theo luồng thường được ghi lại trong hàng đợi để xử lý. Có nhiều công nghệ bạn có thể sử dụng để thực hiện tác vụ này, bao gồm Azure Event Hubs như được hiển thị trong hình ảnh. Từ đây, dữ liệu được xử lý, thường là để tổng hợp dữ liệu theo các cửa sổ thời gian (ví dụ: đếm số tin nhắn truyền thông xã hội có thẻ nhất định cứ 5 phút hoặc tính toán mức đọc trung bình của cảm biến được kết nối Internet mỗi phút). Azure Stream Analytics cho phép bạn tạo các tác vụ để truy vấn và tổng hợp dữ liệu sự kiện khi nó đến và ghi kết quả vào một sink đầu ra. Một trong những sink như vậy là Azure Data Lake Storage Gen2; từ đó dữ liệu theo thời gian thực được captur có thể được phân tích và trực quan hóa.

**Khoa học dữ liệu và học máy (Data science and machine learning)**

Khoa học dữ liệu liên quan đến phân tích thống kê khối lượng lớn dữ liệu, thường sử dụng các công cụ như Apache Spark và ngôn ngữ kịch bản như Python. Azure Data Lake Storage Gen2 cung cấp kho lưu trữ dữ liệu dựa trên đám mây có khả năng mở rộng cao cho khối lượng dữ liệu cần thiết trong các khối lượng công việc khoa học dữ liệu.

Học máy là một lĩnh vực con của khoa học dữ liệu liên quan đến việc đào tạo các mô hình dự đoán. Việc đào tạo mô hình đòi hỏi một lượng lớn dữ liệu và khả năng xử lý dữ liệu đó một cách hiệu quả. Azure Machine Learning là dịch vụ đám mây cho phép các nhà khoa học dữ liệu chạy mã Python trong sổ tay bằng cách sử dụng tài nguyên tính toán phân tán được phân bổ động. Các quy trình xử lý dữ liệu trong các vùng chứa Azure Data Lake Storage Gen2 để đào tạo mô hình, sau đó có thể được triển khai dưới dạng dịch vụ web sản xuất để hỗ trợ các khối lượng công việc phân tích dự đoán.

Đúng vậy, trong một số trường hợp, kho dữ liệu sử dụng các bảng ngoài để định nghĩa lớp siêu dữ liệu quan hệ trên các file trong hồ dữ liệu và tạo ra kiến trúc "data lakehouse" hoặc "lake database" lai.

**Kiến trúc "data lakehouse"** kết hợp các khía cạnh tốt nhất của cả hồ dữ liệu (data lake) và kho dữ liệu (data warehouse) để cung cấp một nền tảng linh hoạt và mạnh mẽ cho việc phân tích dữ liệu.

**Lợi ích của kiến trúc "data lakehouse":**

* **Khả năng mở rộng:** Có thể lưu trữ và xử lý lượng dữ liệu lớn một cách hiệu quả.
* **Linh hoạt:** Hỗ trợ nhiều loại dữ liệu và định dạng dữ liệu khác nhau.
* **Hiệu suất cao:** Cung cấp hiệu suất truy vấn nhanh cho các phân tích dữ liệu.
* **Bảo mật mạnh mẽ:** Bảo vệ dữ liệu khỏi truy cập trái phép.
* **Dễ sử dụng:** Dễ dàng quản lý và truy cập dữ liệu.

**Cách thức hoạt động của kiến trúc "data lakehouse":**

1. Dữ liệu được tải vào hồ dữ liệu.
2. Các bảng ngoài được sử dụng để định nghĩa lớp siêu dữ liệu quan hệ trên các file trong hồ dữ liệu.
3. Dữ liệu có thể được truy vấn bằng các công cụ truy vấn SQL tiêu chuẩn.

**Kiến trúc "lake database"** tương tự như kiến trúc "data lakehouse", nhưng nó tập trung nhiều hơn vào việc cung cấp một nền tảng cho các phân tích dữ liệu SQL.

**Lợi ích của kiến trúc "lake database":**

* **Hiệu suất truy vấn cao:** Cung cấp hiệu suất truy vấn nhanh cho các truy vấn SQL.
* **Dễ sử dụng:** Dễ dàng sử dụng cho các nhà phân tích dữ liệu và nhà khoa học dữ liệu.
* **Khả năng mở rộng:** Có thể mở rộng để đáp ứng nhu cầu phân tích dữ liệu ngày càng tăng.

**Kiến trúc "data lakehouse" và "lake database"** là những lựa chọn tốt cho các tổ chức cần một nền tảng linh hoạt, mạnh mẽ và có thể mở rộng cho việc phân tích dữ liệu.

**Ví dụ về các công cụ hỗ trợ kiến trúc "data lakehouse" và "lake database":**

* Azure Data Lake Storage Gen2
* Azure Synapse Analytics
* Amazon S3
* Amazon Redshift
* Google Cloud Storage
* Google BigQuery

s