Data integration at scale with Azure Data Factory or Azure Synapse Pipeline

ADF

trigger the batch movement of data

set up a regular schedule

ADF provides a cloud-based data integration service that ORCHESTRATES the movement and transformation of data between various data stores and compute resources.

also provides rich visualizations to display the lineage and dependencies between your data pipelines, and monitor all your data pipelines from a single unified view to easily pinpoint issues and setup monitoring alerts.

Much of the functionality of Azure Data Factory appears in Azure Synapse Analytics as a feature referred to as Pipelines, which enables you to integrate data pipelines between SQL Pools, Spark Pools and SQL Serverless, providing a one stop shop for all your analytical needs.

Describe data integration patterns

Data integration firstly involves the collection of data from one or more sources. Optionally, it typically then includes a process where the data may be cleansed and transformed, or perhaps augmented with additional data and prepared. Finally, the amalgamated data is stored in a data platform service that handles the type of analytics that you want to perform. This process can be automated by Azure Data Factory in a pattern know as Extract, Transform and Load (ETL).

Evolution from ETL to ELT

The benefit of ELT is that you can store data in its original format, be it JSON, XML, PDF, or images. In ELT, you define the data's structure during the transformation phase, so you can use the source data in multiple downstream systems.

A Modern Data Warehouse is a centralized data store that provides descriptive analytics and decision support services across the whole enterprise using structured, unstructured, or streaming data sources. Data flows into the warehouse from multiple transactional systems, relational databases, and other data sources on a periodic basis. The stored data is used for historical and trend analysis reporting. The data warehouse acts as a central repository for many subject areas and contains the "single source of truth."

Azure Data factory is typically used to automate the process of extracting, transforming, and loading the data through a batch process against structured and unstructured data sources.

An Azure subscription might have one or more Azure Data Factory instances. Azure Data Factory is composed of four core components. These components work together to provide the platform on which you can compose data-driven workflows with steps to move and transform data.

Data Factory supports a wide variety of data sources that you can connect to through the creation of an object known as a , which enables you to ingest the data from a data source in readiness to prepare the data for transformation and/or analysis.

**Datasets** represent data structures within the data store that is being referenced by the Linked Service object.

**Activities** typically contain the transformation logic or the analysis commands of the Azure Data Factory’s work. Activities includes the Copy Activity that can be used to ingest data from a variety of data sources. It can also include the Mapping Data Flow to perform code-free data transformations. It can also include the execution of a stored procedure, Hive Query, or Pig script to transform the data.

Multiple activities can be logically grouped together with an object referred to as a **Pipeline**, and these can be scheduled to execute, or a trigger can be defined that determines when a pipeline execution needs to be kicked off. There are different types of triggers for different types of events.

