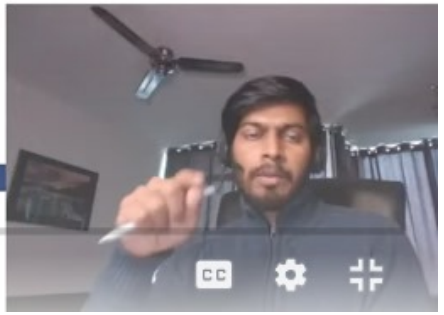
What is Azure Data Factory - ADF ?

- It is the cloud-based ETL and data integration service that allows you to create data-driven workflows for orchestrating data movement and transforming data at scale. Using Azure Data Factory, you can create and schedule data-driven workflows (called pipelines) that can ingest data from disparate data stores.

- You can build complex ETL processes that transform data visually with data flows or by using compute services such as Azure HDInsight Hadoop, Azure Databricks, and Azure SQL Database.

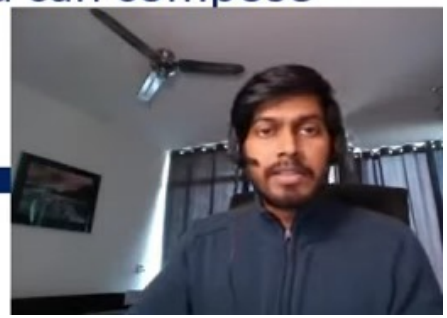


Top-level concepts - ADF

Azure Data Factory is composed of below key components.

- Pipelines
- Activities
- Datasets
- Linked services
- Data Flows
- Integration Runtimes
- Triggers

These components work together to provide the platform on which you can compose data-driven workflows with steps to move and transform data.



Pipeline

A data factory might have one or more pipelines. A pipeline is a logical grouping of activities that performs a unit of work. Together, the activities in a pipeline perform a task.

Example - A pipeline can contain a group of activities that ingests data from an Azure blob, and then runs a Hive query on an HDInsight cluster to partition the data.

Datasets and Linked services

Datasets represent data structures within the data stores, which simply point to or reference the data you want to use in your activities as inputs or outputs.

Linked services are much like connection strings, which define the connection information that's needed for Data Factory to connect to external resources. Think of it this way: a linked service defines the connection to the data source, and a dataset represents the structure of the data. For example, an Azure Storage-linked service specifies a connection string to connect to the Azure Storage account. Additionally, an Azure blob dataset specifies the blob container and the folder that contains the data.

Integration Runtime

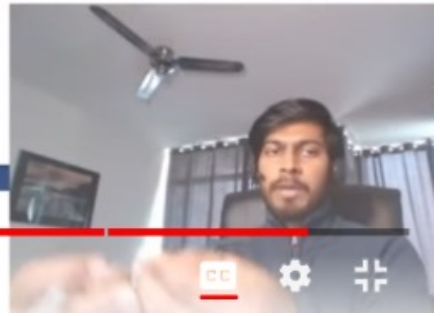
In Data Factory, an activity defines the action to be performed. A linked service defines a target data store or a compute service. An integration runtime provides the bridge between the activity and linked Services. It's referenced by the linked service or activity, and provides the compute environment where the activity either runs on or gets dispatched from. This way, the activity can be performed in the region closest possible to the target data store or compute service in the most performant way while meeting security and compliance needs.

again then we have integration runtime
so this is also very important concept

Triggers

Triggers represent the unit of processing that determines when a pipeline execution needs to be kicked off. There are different types of triggers for different types events.

but for now just remember trigger is mainly used to execute your pipeline

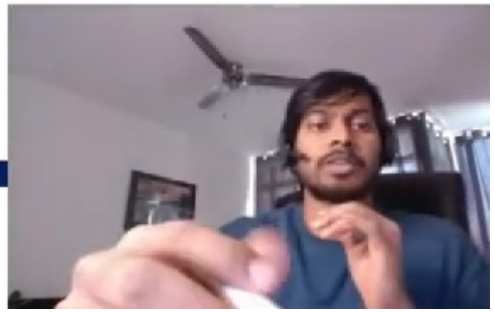


Integration Run Time

Integration runtime is the compute infrastructure used by Azure Data Factory (ADF) to provide various data integration capabilities across different network environments. There are three types of integration runtimes offered by Data Factory:

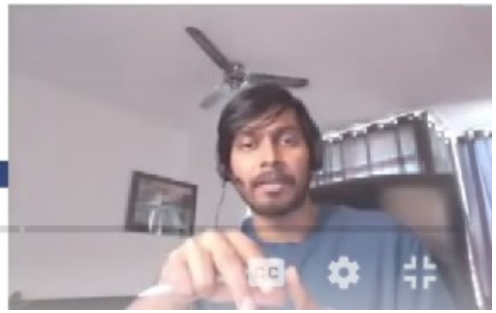
Types of Integration Run Times

- Azure integration runtime
- Self-hosted integration runtime
- Azure-SQL Server Integration Services (SSIS) integration runtime



Types of Integration Run Times

- Azure integration runtime - The compute resource for an Azure integration runtime is fully managed elastically in Azure
- Self-hosted integration runtime - A self-hosted integration runtime can run copy activities between a cloud data store and a data store in a private network
- Azure-SQL Server Integration Services (SSIS) integration runtime - Azure-SSIS IR is a fully managed cluster of Azure virtual machines (VMs or nodes) dedicated to run your SSIS packages



Pipeline and Activities

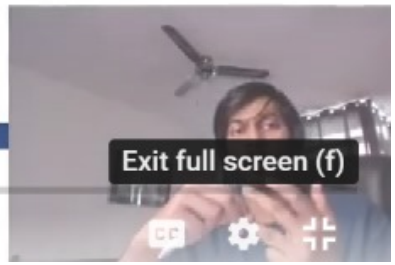
- A data factory might have one or more pipelines. A pipeline is a logical grouping of activities that performs a unit of work. Together, the activities in a pipeline perform a task.

- Example - A pipeline can contain a group of activities that ingests data from an Azure blob, and then runs a Hive query on an HDInsight cluster to partition the data.

- Activities represent a processing step in a pipeline. For example, you might use a copy activity to copy data from one data store to another data store.

Types of Activities

- Data movement activities
- Data transformation activities
- Control flow activities

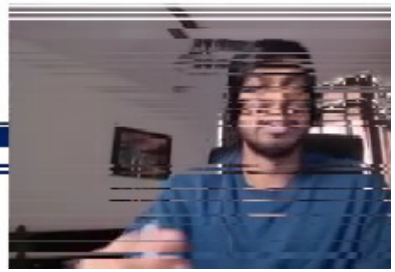


Triggers

Triggers are another way that you can execute a pipeline run. Triggers represent a unit of processing that determines when a pipeline execution needs to be kicked off

Types of Triggers

- Schedule trigger: A trigger that invokes a pipeline on a wall-clock schedule.
- Tumbling window trigger: A trigger that operates on a periodic interval, while also retaining state.
- Event-based trigger: A trigger that responds to an event.

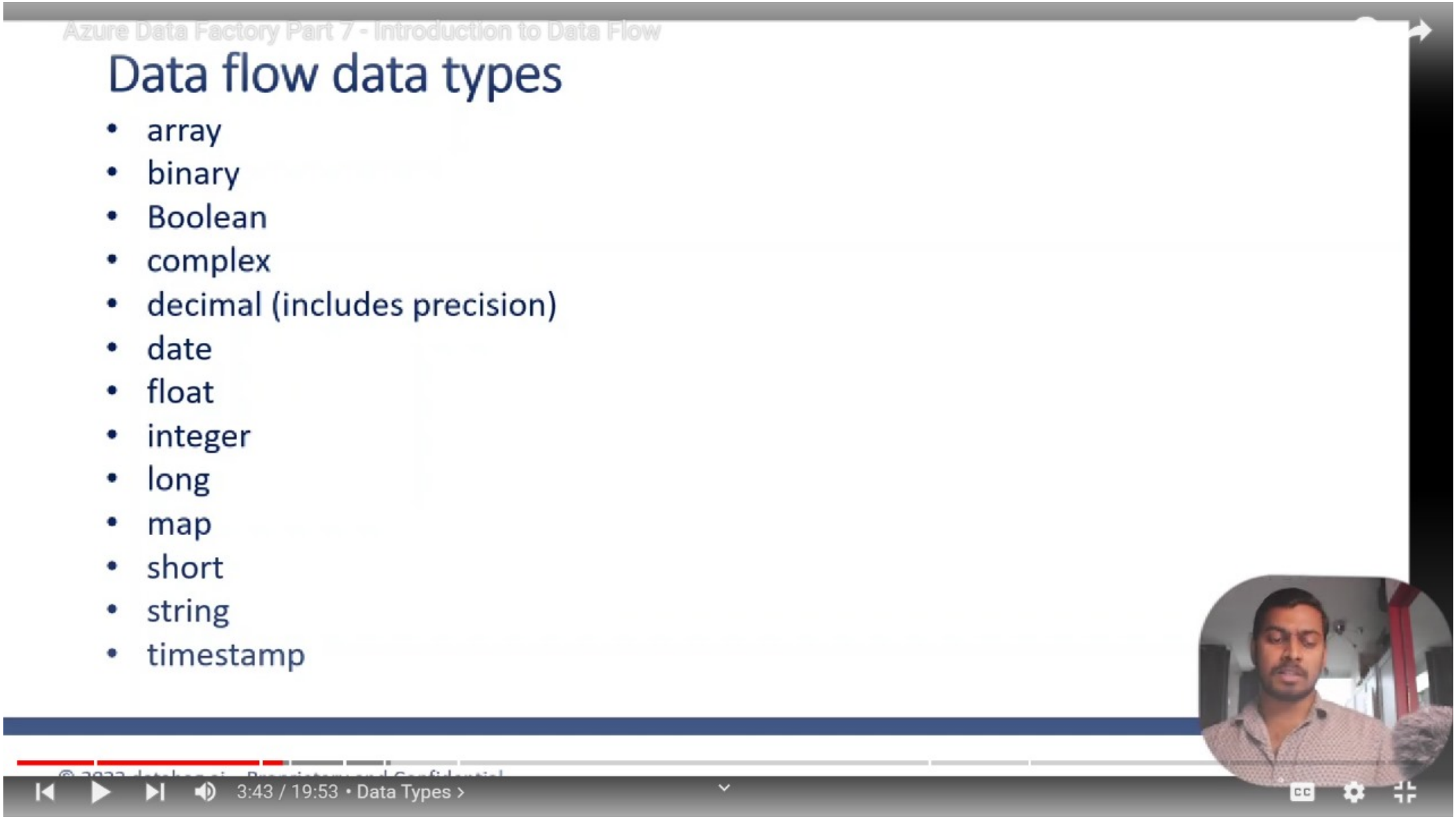


Azure Data Factory Part 7 - Introduction to Data Flow

Mapping Data Flows

Mapping data flows are visually designed data transformations in Azure Data Factory. Data flows allow data engineers to develop data transformation logic without writing code. The resulting data flows are executed as activities within Azure Data Factory pipelines

Mapping data flows provide an entirely visual experience with no coding required. Your data flows run on ADF-managed execution clusters for scaled-out data processing. Azure Data Factory handles all the code translation, path optimization, and execution of your data flow jobs

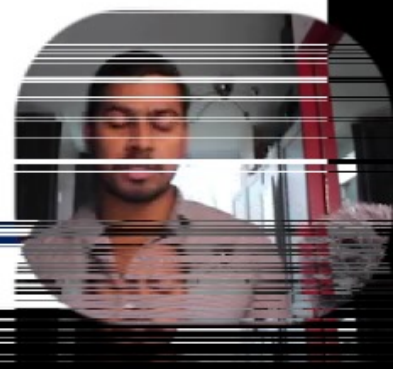

Azure Data Factory Part 7 - Introduction to Data Flow
Data flow data types
array
binary
Boolean
complex
decimal (includes precision)
date
float
integer
long
map
short
string
timestamp
3:43 / 19:53 • Data Types ›

Data flow activity

Mapping data flows are operationalized within ADF pipelines using the data flow activity. All a user has to do is specify which integration runtime to use and pass in parameter values.

Debug mode

Debug mode allows you to interactively see the results of each transformation step while you build and debug your data flows. The debug session can be used both in when building your data flow logic and running pipeline debug runs with data flow activities.



Monitoring data flows

Mapping data flow integrates with existing Azure Data Factory monitoring capabilities.

The Azure Data Factory team has created a performance tuning guide to help you optimize the execution time of your data flows after building your business logic