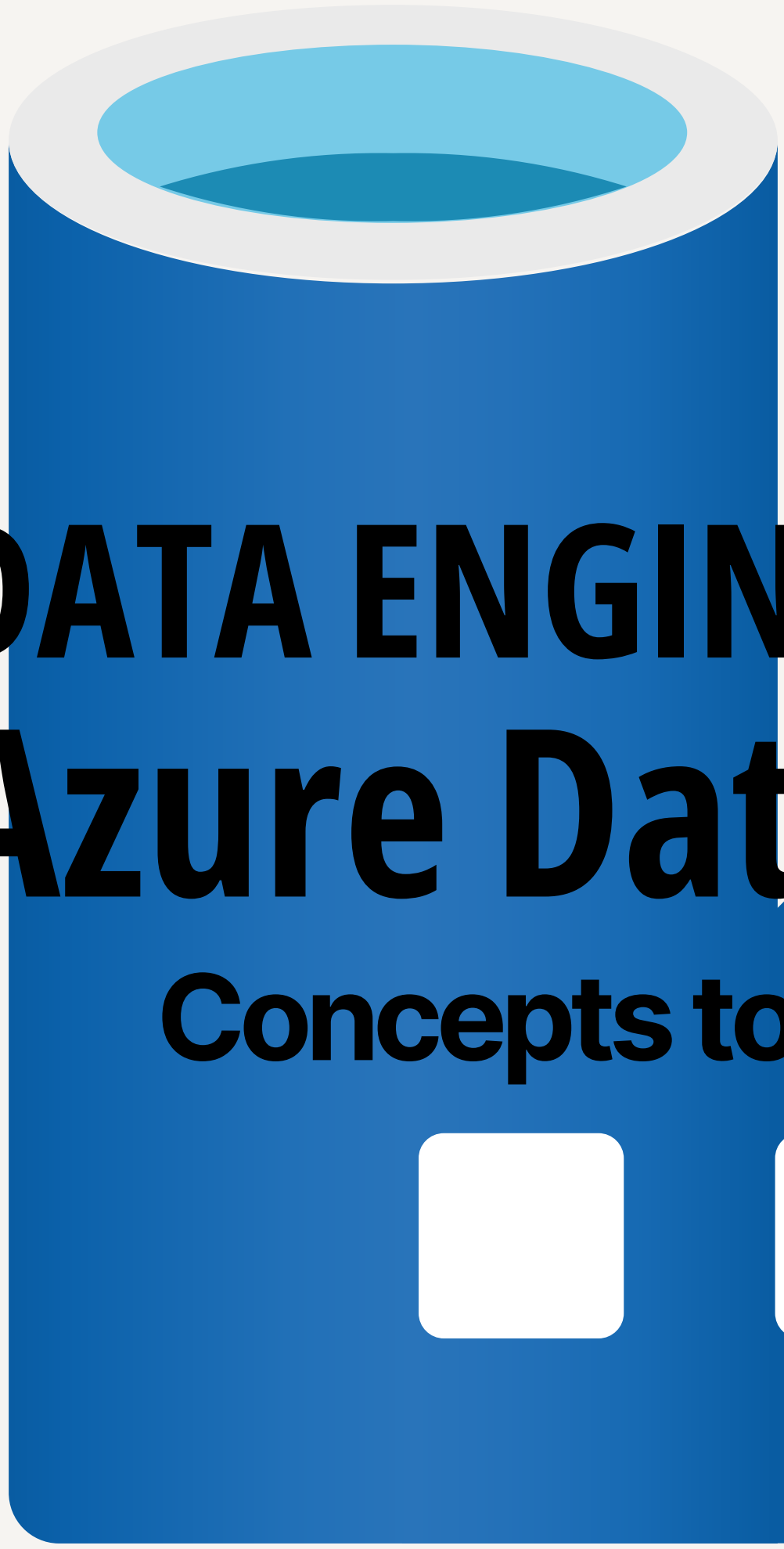
Data Engineering 101: Azure Data Factory



DATA ENGINEERING 101

Azure Data Factory

Concepts to get started

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Data Factory

A cloud-based data integration service that allows creation, scheduling, and orchestration of data workflows.

Building ETL processes, data migration, data transformation.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Pipeline

A logical grouping of activities that together perform a task. Pipelines allow activities to be linked together.

ETL pipelines, data migration pipelines, orchestrating data workflows.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Activity

Represents a single step in a pipeline. Types include data movement, data transformation, and control activities.

Copy activity for data movement, data transformation using Azure Data Lake Analytics, Databricks, or custom scripts.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Linked Service

Defines the connection information needed for Data Factory to connect to external resources.

Connecting to Azure Blob Storage, SQL Database, Cosmos DB, REST APIs.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Dataset

Represents the data structure within the linked data stores. Used in activities for reading or writing data.

CSV files in Blob Storage, tables in SQL Database, documents in Cosmos DB.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Mapping Data Flow

A visually designed data transformation feature in ADF that allows for code-free data transformations.

Data aggregation, data joins, data filtering, and transformation logic applied to incoming data streams.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Wrangling Data Flow

A feature that allows users to perform data transformation using the Power Query Editor within ADF.

Data wrangling, cleaning, reshaping, and enrichment using Power Query interface.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Control Flow

Orchestrates how activities are executed within a pipeline. Includes conditional, looping, and branching constructs.

If activities, ForEach loops, Until loops, Switch case activities.

Shwetank Singh
GritSetGrow - GSGLearn.com

Triggers

Used to schedule or trigger pipelines based on time or events.

Schedule-based triggers, tumbling window triggers, event-based triggers.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Integration Runtime (IR)

The compute infrastructure used by ADF to perform data movement and transformation.

Azure Integration Runtime, Self-hosted Integration Runtime, and Azure-SSIS Integration Runtime.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Data Movement

The process of moving data from source to destination. ADF supports various connectors and protocols.

Copy data from on-premises SQL Server to Azure SQL Database, transfer files between Blob Storage and Data Lake.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Parameterization

Allows the creation of dynamic, reusable pipelines. Parameters can be passed to pipelines, datasets, and linked services.

Passing different file names or table names to a single pipeline.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Monitoring

ADF provides a monitoring dashboard to view the status of pipeline runs, debug runs, and triggers.

Monitoring pipeline execution success/failure, reviewing activity logs, and diagnosing issues.

Shwetank Singh
GritSetGrow - GSGLearn.com

Git Integration

Allows integration with Git repositories for version control of ADF pipelines and assets.

Using GitHub or Azure Repos for managing ADF assets, promoting code changes across environments.

Shwetank Singh
GritSetGrow - GSGLearn.com

Debugging

ADF offers debugging capabilities to test pipeline activities before publishing.

Running pipelines in debug mode, testing activities with test data.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Data Flow Debug

Specific to Mapping Data Flows, allows for debugging of data transformation logic.

Viewing row-level data as it passes through transformation steps, testing transformations with sample data.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Global Parameters

Reusable parameters available across pipelines in a Data Factory instance.

Defining a global connection string or file path that can be used by multiple pipelines.

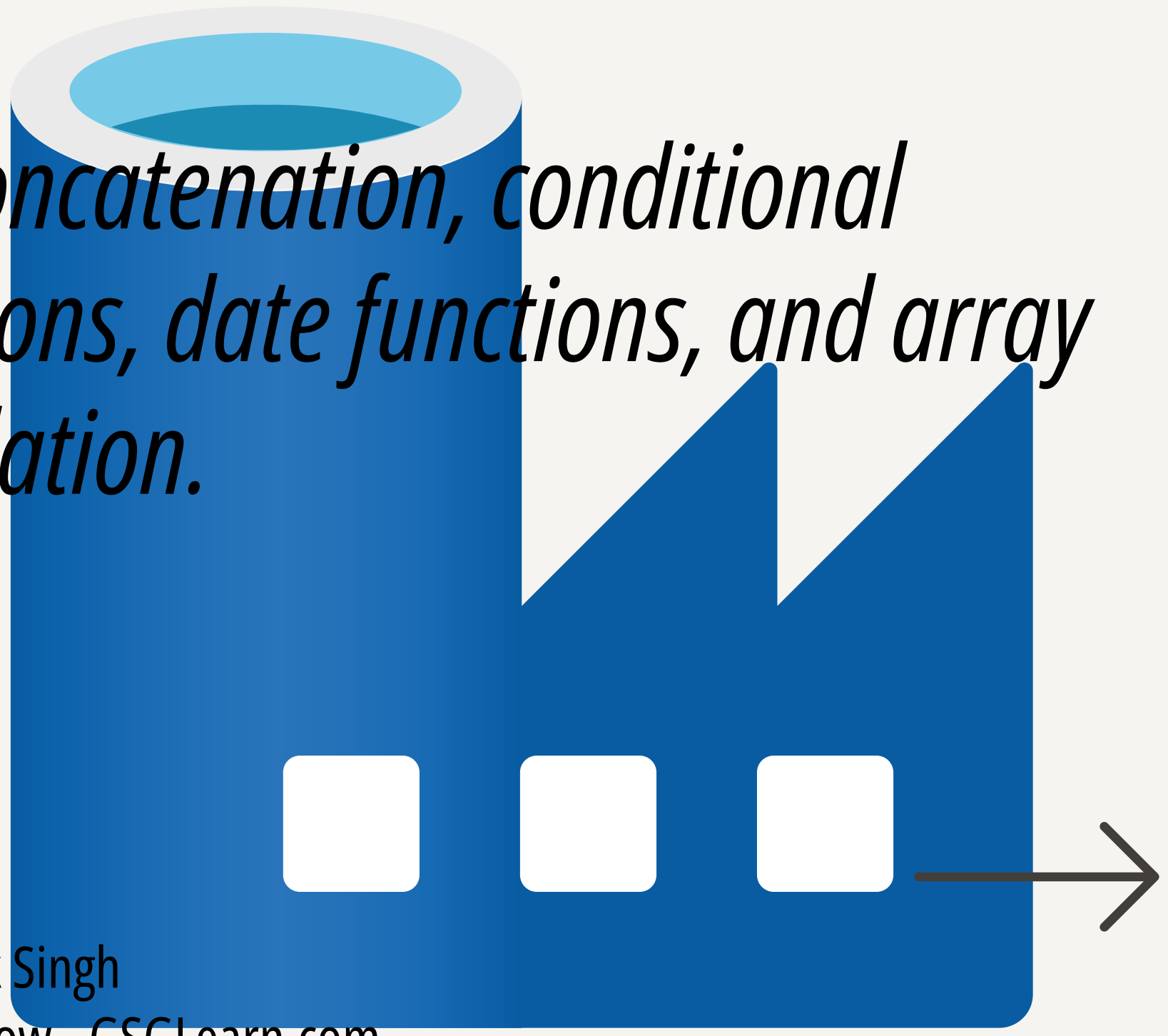
Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Expression Language

ADF's expression language allows dynamic content and logic within activities and datasets.

String concatenation, conditional expressions, date functions, and array manipulation.



Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Azure Key Vault Integration

Allows secure storage and retrieval of secrets, keys, and certificates used by ADF.

Securely connecting to databases using ts stor ey Vault, accessing ys sec

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

SSIS Integration

ADF allows running SQL Server Integration Services (SSIS) packages in the cloud with full compatibility.

Migrating on-premises SSIS packages to run in ADF using Azure-SSIS Integration Runtime.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Data Preview

Provides a snapshot view of the data in the dataset before execution, specific to Mapping Data Flows.

Previewing data after applying a transformation to validate the results.

Shwetank Singh
GritSetGrow - GSGLearn.com

Pipeline Parameters

Allows defining parameters within a pipeline to make it more flexible and reusable.

Creating a parameterized pipeline for copying data between different storage accounts or databases.

Shwetank Singh
GritSetGrow - GSGLearn.com

ADF SDKs and APIs

Provides programmatic access to create, manage, and monitor ADF resources using various languages.

Automating pipeline deployment using Azure SDKs for .NET, Python, or REST API.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Power Query

A low-code/no-code feature that allows data transformation within ADF using a familiar Power Query interface.

Data wrangling, simple transformations, data cleanup, and enrichment

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

If Condition

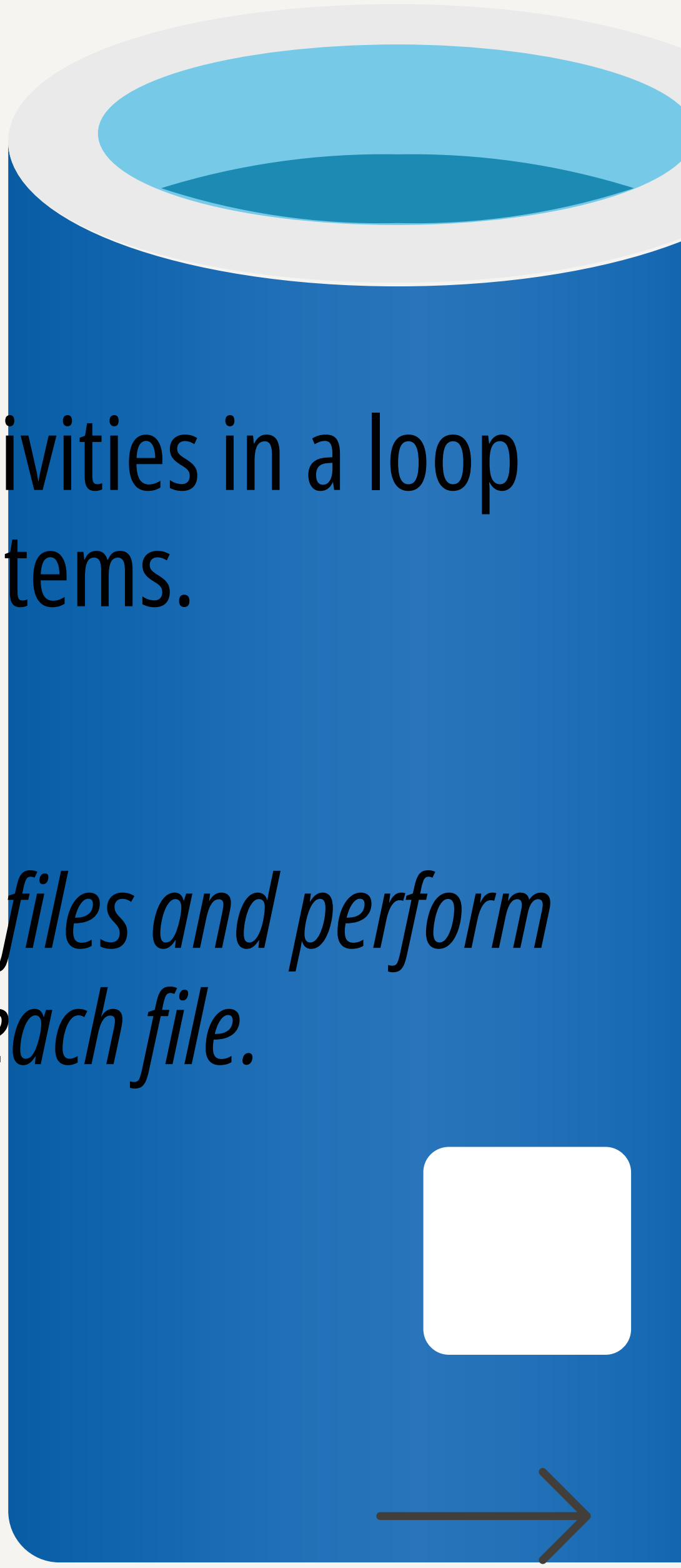
Allows conditional logic within a pipeline to execute activities based on a Boolean expression.

Execute activity A if condition is true, otherwise execute activity B.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

ForEach



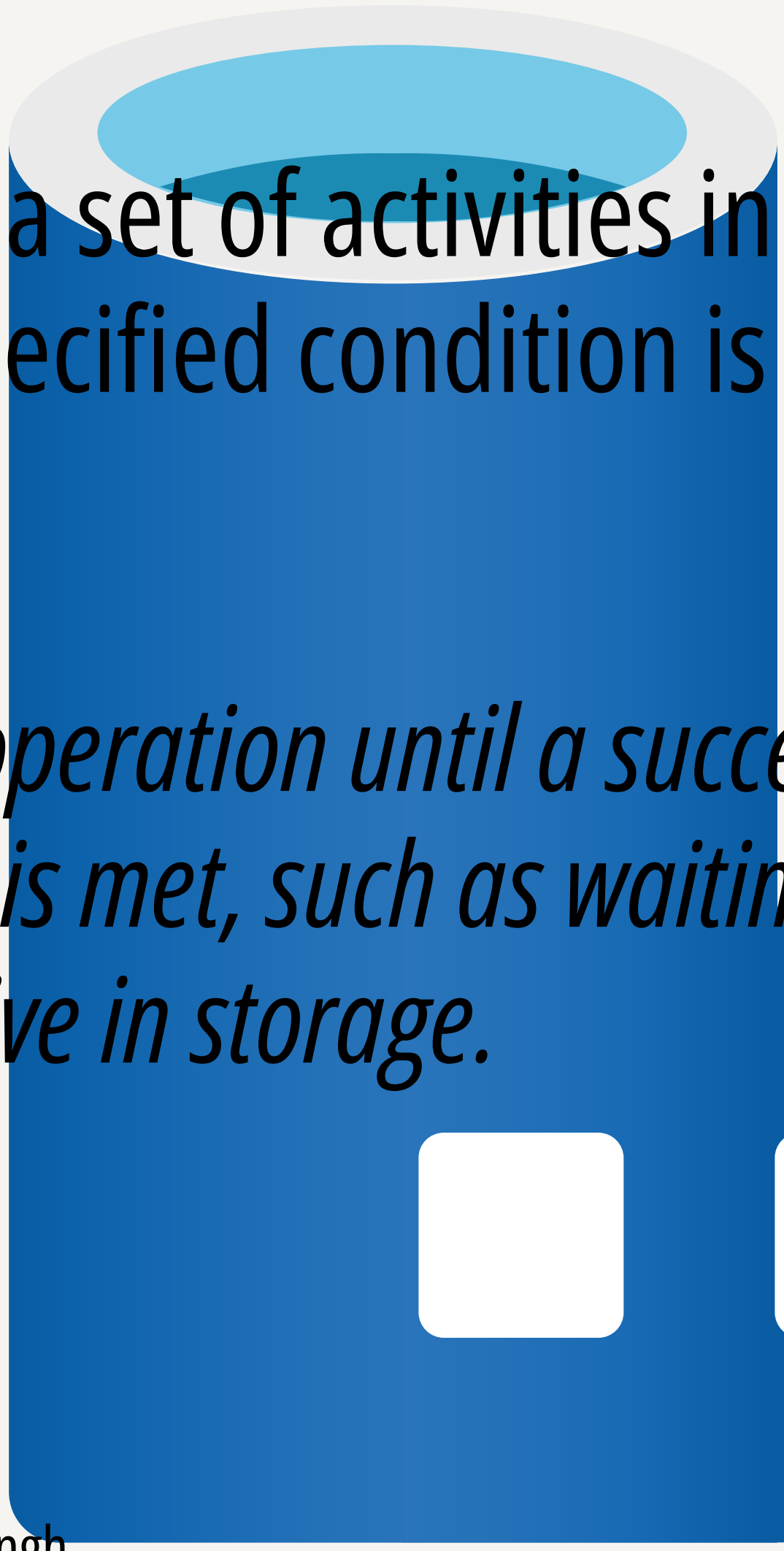
Executes a set of activities in a loop over a collection of items.

Loop through a list of files and perform a copy operation for each file.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Until



Executes a set of activities in a loop until a specified condition is met.

Retry an operation until a success condition is met, such as waiting for a file to arrive in storage.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Wait

Pauses pipeline execution for a specified period.

Wait for 10 minutes before proceeding to the next activity.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Execute Pipeline

Invokes another pipeline from within a pipeline.

Modularize complex workflows by calling sub-pipelines.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Switch

Allows branching logic based on a value. Similar to a switch-case statement in programming.

Execute different activities based on the value of a dataset field or parameter.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Set Variable

Sets the value of a variable that can be used later in the pipeline.

Set a variable with the current timestamp for logging purposes.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Append Variable

Appends values to an array variable.

Collect and store file names processed during pipeline execution

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Copy Data

Moves data from a source to a destination. Supports over 90 data connectors.

Copy data from SQL Server to Azure Data Lake Storage, transfer files from SFTP to Blob Storage.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Lookup

Retrieves data from a dataset and makes it available for subsequent activities.

Look up a value in a SQL table to use in a conditional activity.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Get Metadata

Retrieves metadata information (e.g., file size, last modified date) from data in a dataset.

Check the size of a file before processing it.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Mapping Data Flow

Provides a code-free experience for transforming data at scale.

Join, filter, aggregate, and perform co[redacted]ansfor[redacted]on dat[redacted] vi[redacted]signed[redacted]w.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Wrangling Data Flow

Allows for data preparation and transformation using Power Query.

Data cleanup, normalization, and transformation using Power Query language.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Data Flow Debug

Enables testing and debugging of Mapping Data Flows with real data.

Test and valid sforma inspecting rov ata du design phase.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Databricks Notebook

Executes an Azure Databricks notebook as part of a pipeline.

Run co ata tra tions, machi ing mo ustom logic in Databricks.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Azure HDInsight

Runs jobs on an Azure HDInsight cluster, including Hive, Pig, Spark, etc.

Process big data using Hadoop or Spark, run Hive queries, or Pig scripts.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Azure Machine Learning

Invokes an Azure Machine Learning pipeline endpoint to run ML models.

[redacted] machi[redacted]ing mo[redacted] or scor[redacted]ine as [redacted] data processing workflow.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Stored Procedure

Executes a stored procedure in a relational database.

Run a stored procedure in Azure SQL Database or SQL Server to perform complex database operations.

Shwetank Singh
GritSetGrow - GSGLearn.com

Web Activity

Makes a call to a REST endpoint.

Trigger a web service, pass data between ADF and external systems via REST APIs.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Azure Function

Executes an Azure Function as part of a pipeline.

Run serverless code in response to event triggers, run lightweight transformations, or trigger external processes.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Execute SSIS Package

Executes an SSIS package stored in Azure SQL Database, Azure File Storage, or SSISDB.

Migrate and run existing SSIS ETL packages in the cloud using Azure-SSIS Integration Runtime.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Custom Activity

Runs custom code on an Azure Batch Service.

Execute complex transformations or log[redacted]sn't su[redacted]d by a[redacted]e-box[redacted]ies usi[redacted]om co[redacted]

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Aggregate Data Flow

Summarizes data by performing aggregation operations.

Calculate sum, average, count, min, max, etc., on a dataset.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Conditional Split Data Flow

Routes data rows to different streams based on a condition.

Filter data into different outputs based on conditional

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Derived Column Data Flow

Adds or replaces columns in the dataset with derived values.

Create new columns by applying expressions or transformations on existing columns.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Filter Data Flow

Filters rows in the dataset based on a Boolean expression.

move u rows f taset b a cond

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Join Data Flow

Combines two datasets based on a common key or condition.

Perform inner, outer, left, or right joins on datasets

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Lookup Data Flow

Performs a lookup of data in another dataset and returns the matching rows.

[redacted] data [redacted] column[redacted] another dataset based on a key.

Shwetank Singh
GritSetGrow - GSGLearn.com

Select Data Flow

Chooses which columns to include or exclude in the output.

Remove unnecessary columns from the output.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Sort Data Flow

Orders the rows in a dataset based on specified columns.

Support data [redacted] performi[redacted] operations like windowing.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Union Data Flow

Combines rows from two or more datasets into a single dataset.

Merg[redacted]s with th[redacted]chema.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Window Data Flow

Performs window functions, such as ranking or calculating moving averages, based on a specified window.

Calculate running totals, moving averages, or rank rows within partitions.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Pivot Data Flow

Converts rows into columns based on an aggregation function.

Transfor rical dat ore suitable format for analysis.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Unpivot Data Flow

Converts columns into rows to normalize the data.

Flatten[illegible]able into[illegible]ble for easier processing.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Azure Data Factory

Exists Data Flow

Filters rows based on the existence of corresponding rows in another dataset.

Keep only the r[redacted]ave ma[redacted] entries in another dataset.

Shwetank Singh
GritSetGrow - GSGLearn.com

THaNK you