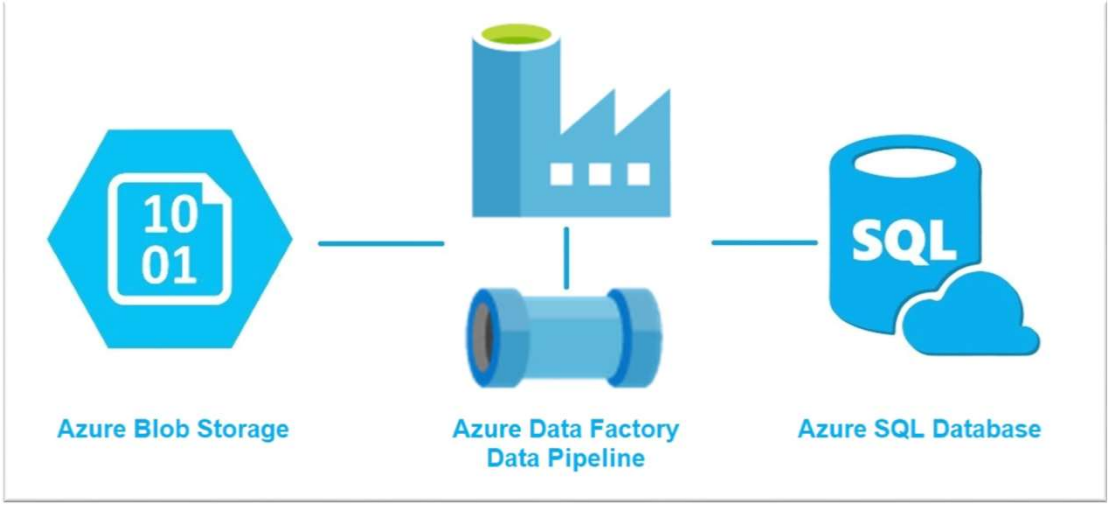
Azure Data Factory

Azure Data Factory

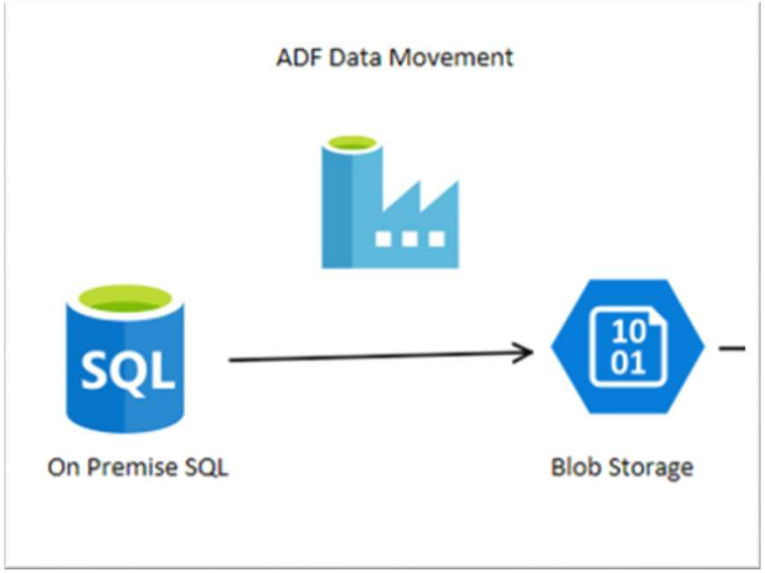
- In the world of big data, raw, unorganized data is often stored.
- However, on its own, raw data doesn't have the proper context or meaning to provide meaningful insights to analysts, data scientists, or business decision makers.
- Azure Data Factory is a managed cloud service that's built for extract-transform-load (ETL), extract-load-transform (ELT), and data integration projects.



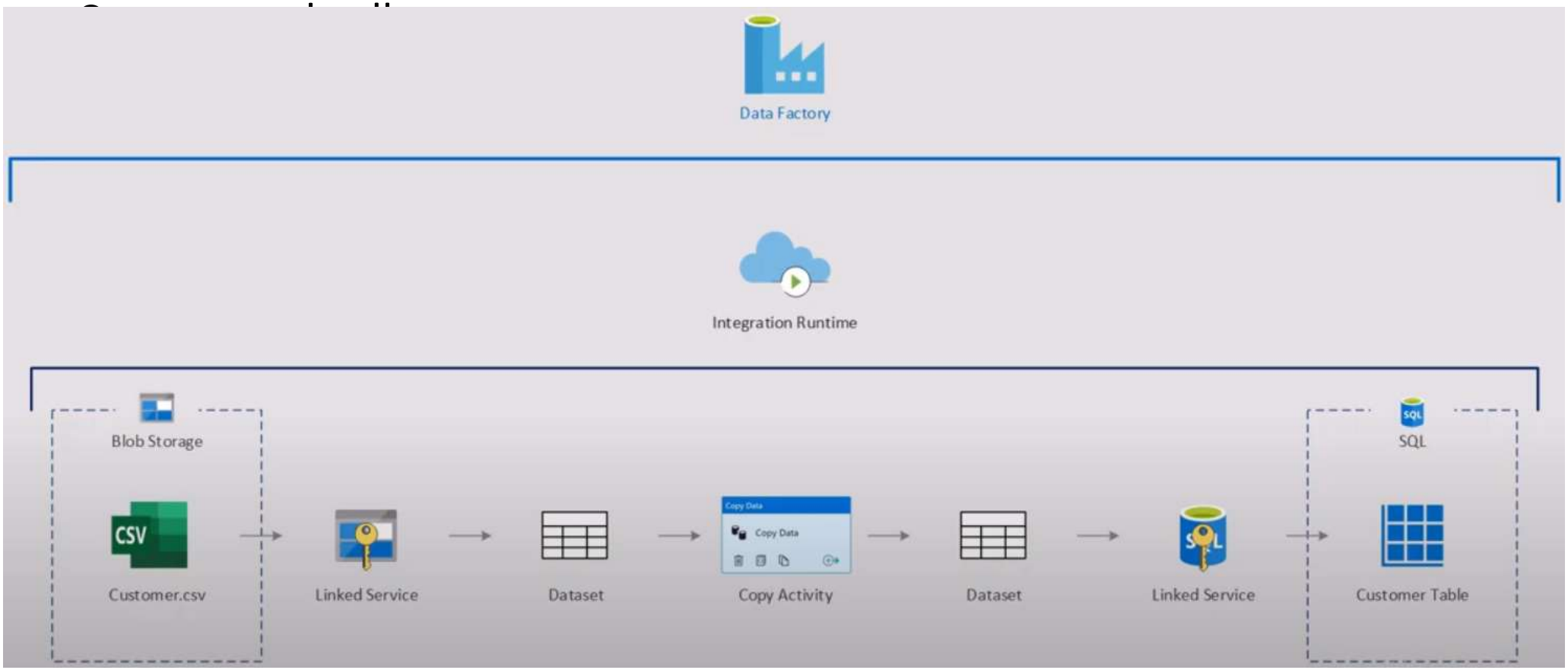
Azure Data Factory

- Cloud-based ETL and data integration service allows to create data-driven workflows
- Can ingest data from disparate data stores.
- Can build complex ETL processes that transform data

- Ultimately, through Azure Data Factory, raw data can be organized into meaningful data stores for better business decisions.



How does it work?



How does it work?

- Connect and collect
- Transform and enrich
- CI/CD and publish
- Monitor

Top-level concepts

- Azure Data Factory is composed of many key components.
- These components work together to provide the platform
- Pipeline
  - A logical grouping of activities that performs a unit of work.
  - Together, the activities in a pipeline perform a task.
  - For 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.
  - The benefit of this is that the pipeline allows you to manage the activities as a set instead of managing each one individually.

Top-level concepts

- Mapping data flows
  - Create and manage graphs of data transformation logic that you can use to transform any-sized data.
  - You can build-up a reusable library of data transformation routines and execute those processes in a scaled-out manner from your ADF pipelines
- Activity
  - 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.
- Datasets
  - Datasets represent data structures within the data stores

Top-level concepts

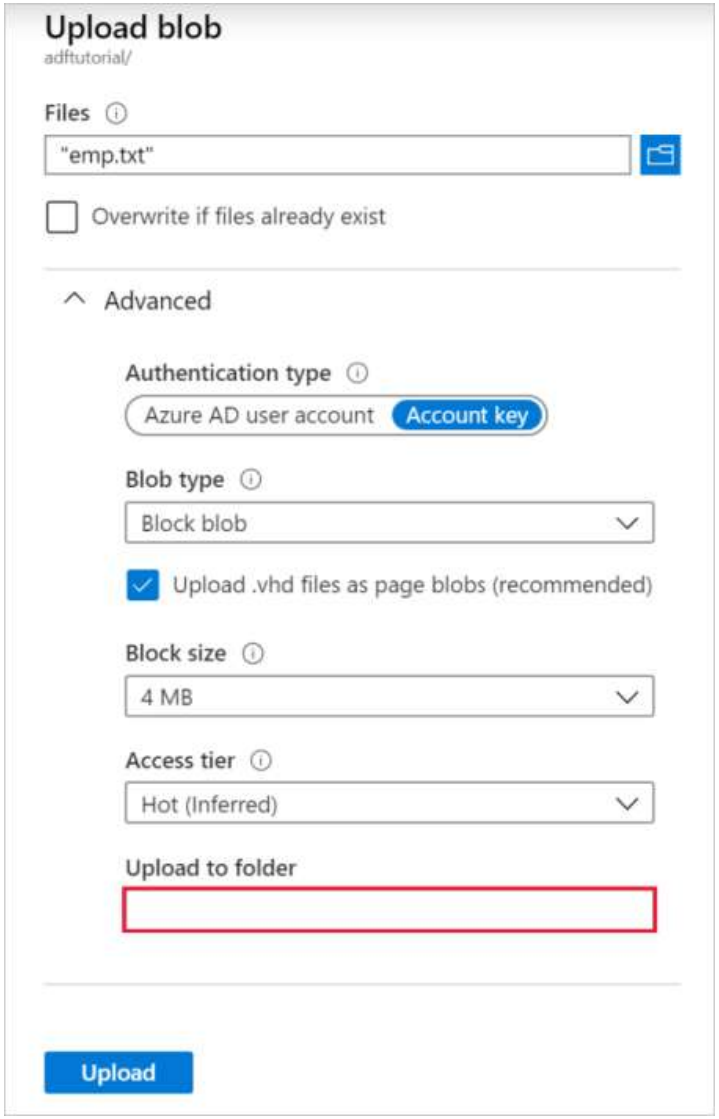
- Linked services
  - Linked services are much like connection strings
- Triggers
  - Determines when a pipeline execution needs to be kicked off.
  - There are different types of triggers for different types of events.
- Pipeline runs
  - A pipeline run is an instance of the pipeline execution.
  - Pipeline runs are typically instantiated by passing the arguments to the parameters that are defined in pipelines.

Top-level concepts

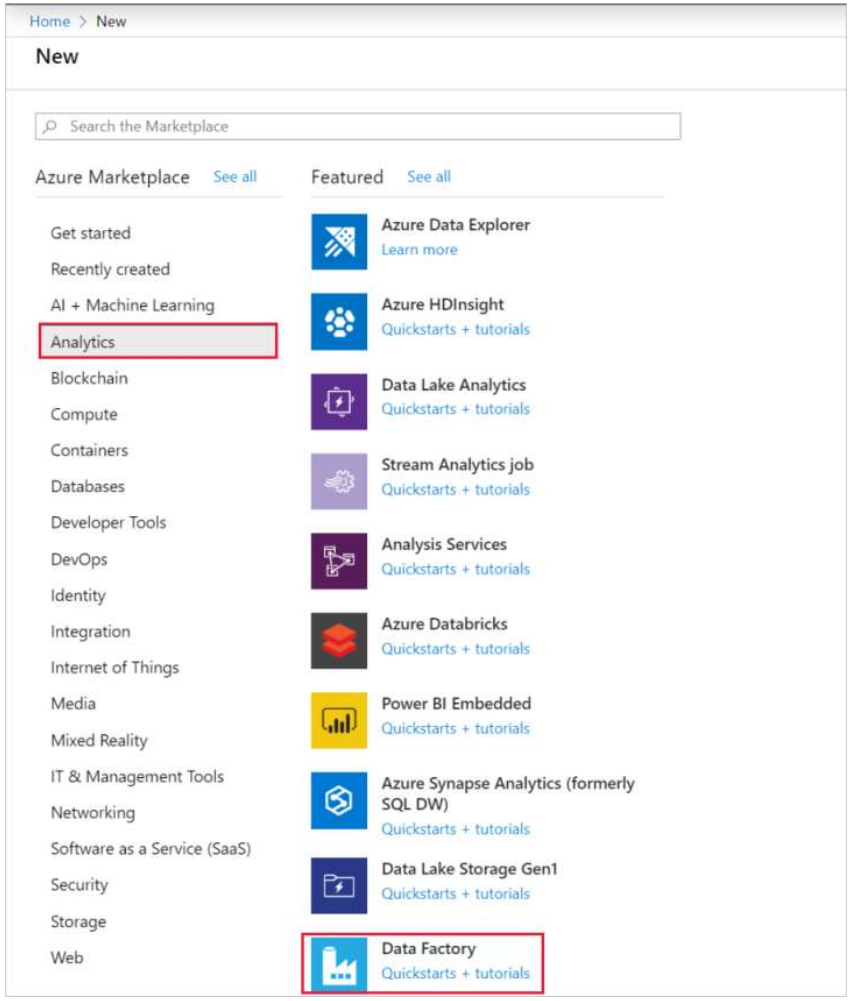
- Parameters
  - Parameters are key-value pairs of read-only configuration.
  - Parameters are defined in the pipeline
- Control flow
  - Control flow is an orchestration of pipeline activities that includes chaining activities in a sequence, branching, defining parameters and passing arguments
- Variables
  - Variables can be used inside of pipelines to store temporary values

Using the Azure Data Factory UI

- You use a general-purpose Azure Storage account (specifically Blob storage) as both source and destination data stores
- If you don't have a general-purpose Azure Storage account, Create a storage account
- Get the storage account name
- Create a blob container
- Add an input folder and file for the blob container

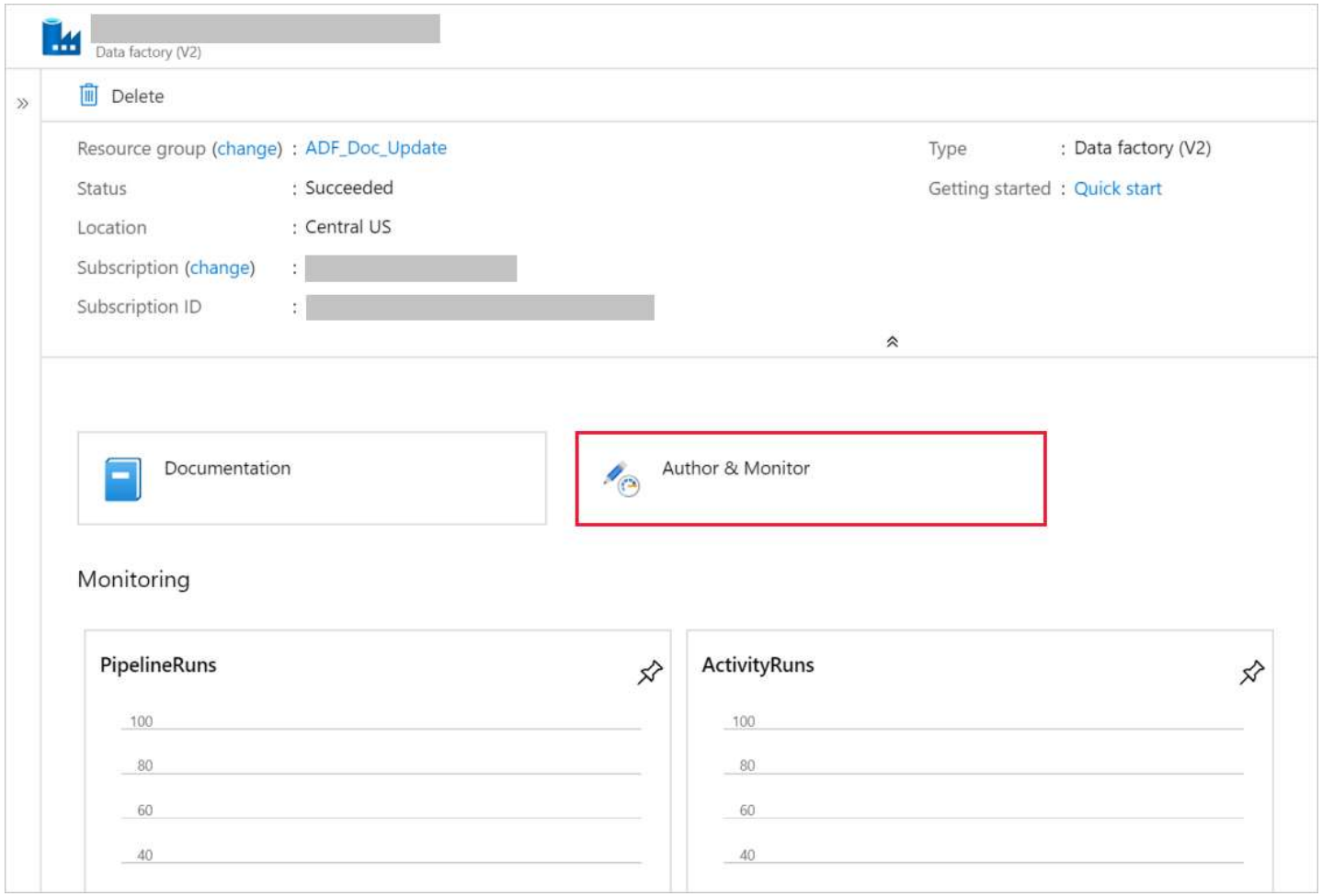


Create a data factory



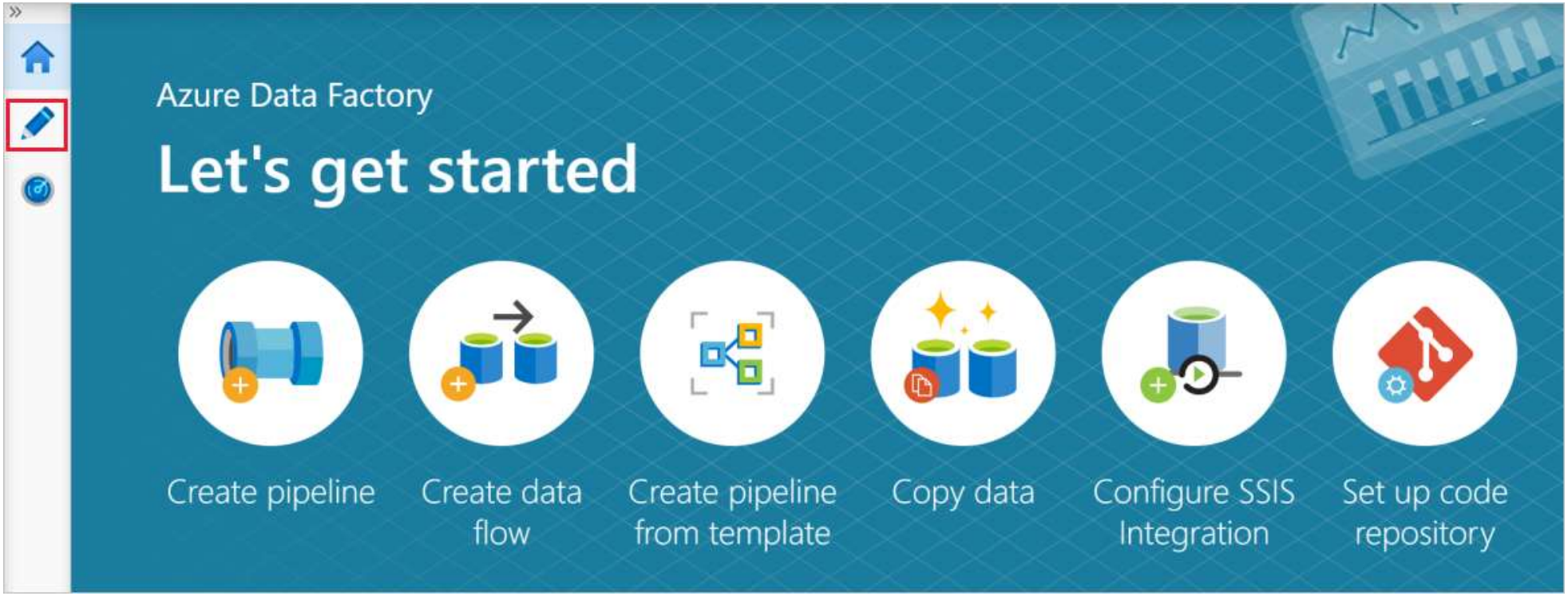
Create a data factory

- Select the Author & Monitor tile to start the Azure Data Factory user interface (UI) application on a separate tab.



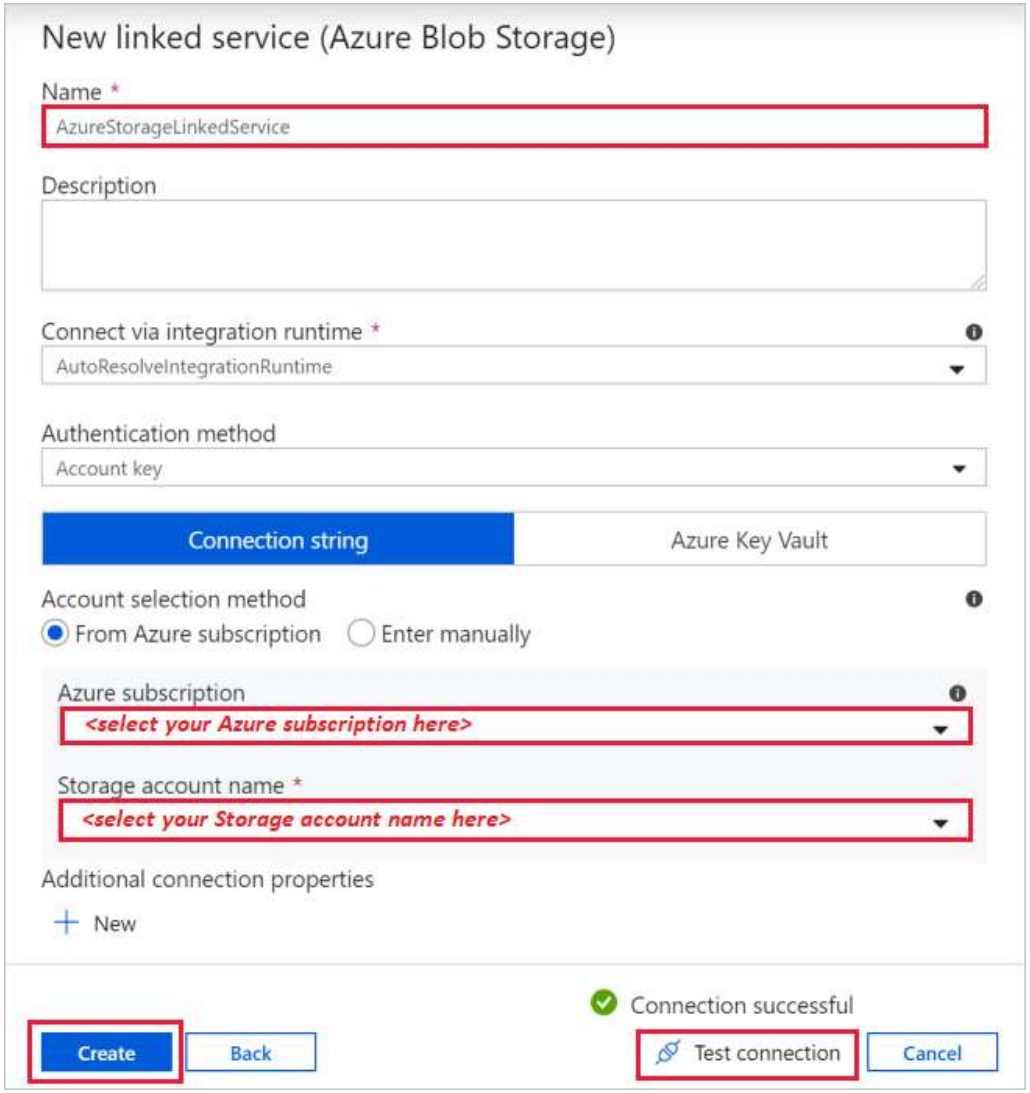
Create a data factory

- On the Let's get started page, switch to the Author tab in the left panel.



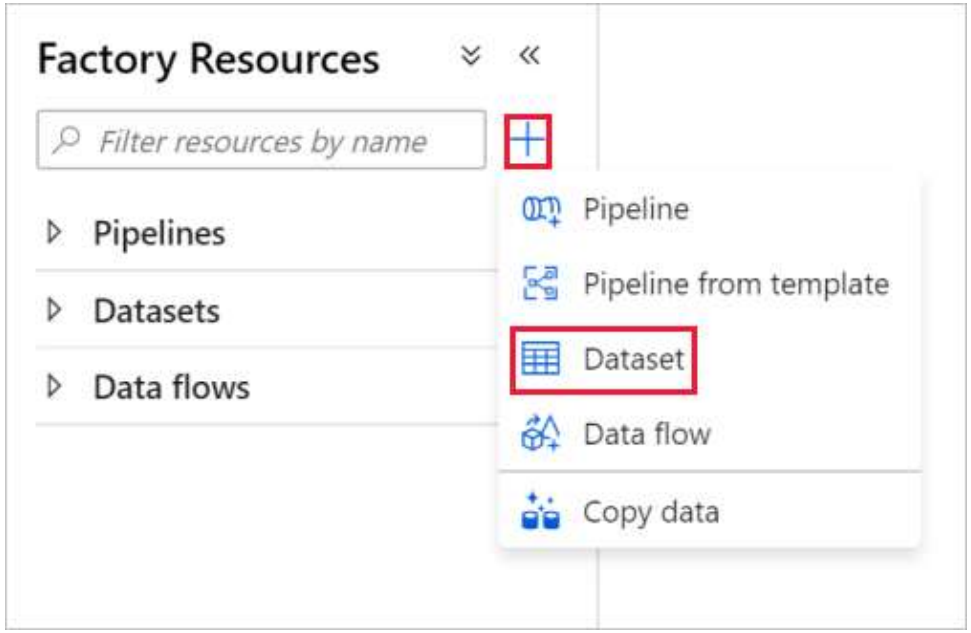
Create a linked service

- To link your Azure Storage account to the data factory.
- The linked service has the connection information that the Data Factory service uses at runtime to connect to it.



Create datasets

- In this procedure, you create datasets:
  - InputDataset and
  - OutputDataset.
- They refer to the Azure Storage linked service that you created.
- In the linked service settings, you specified the Azure Storage account that contains the source data.



In the source dataset settings, you specify where exactly the source data resides (blob container, folder, and file).

In the sink dataset settings, you specify where the data is copied to (blob container, folder, and file).

Create a pipeline

- In this procedure, you create and validate a pipeline with a copy activity that uses the input and output datasets.
- The copy activity copies data from the file you specified in the input dataset settings to the file you specified in the output dataset settings.

Thanks