Week V

Microsoft Azure

Intermediate Data Engineer in Azure

Trainer: Balazs Balogh

CUBIX
INSTITUTE OF TECHNOLOGY

Azure – General information

Microsoft Azure is a comprehensive cloud computing platform and service created by Microsoft. It provides a broad range of cloud services, including computing, storage, networking, and analytics, allowing businesses and developers to build, deploy, and manage applications and services through Microsoft-managed data centers.

It has a wide range of services: Azure offers over 200 services categorized into the following:

- **Compute:** Virtual Machines, Azure Kubernetes Service, Azure Functions (serverless computing).
- **Storage:** Blob Storage, Azure Data Lake, Queue Storage.
- **Networking:** Virtual Networks, Load Balancers, Azure VPN Gateway.
- **Databases:** Azure SQL Database, Cosmos DB, Azure Database for MySQL/PostgreSQL.
- **AI & Machine Learning:** Azure Cognitive Services, Azure ML.
- **Big Data & Analytics:** Azure Synapse Analytics, Data Factory, Databricks
- **IoT Services:** Azure IoT Hub, Azure Digital Twins.
- **DevOps:** Azure DevOps, GitHub integration, Azure Pipelines.

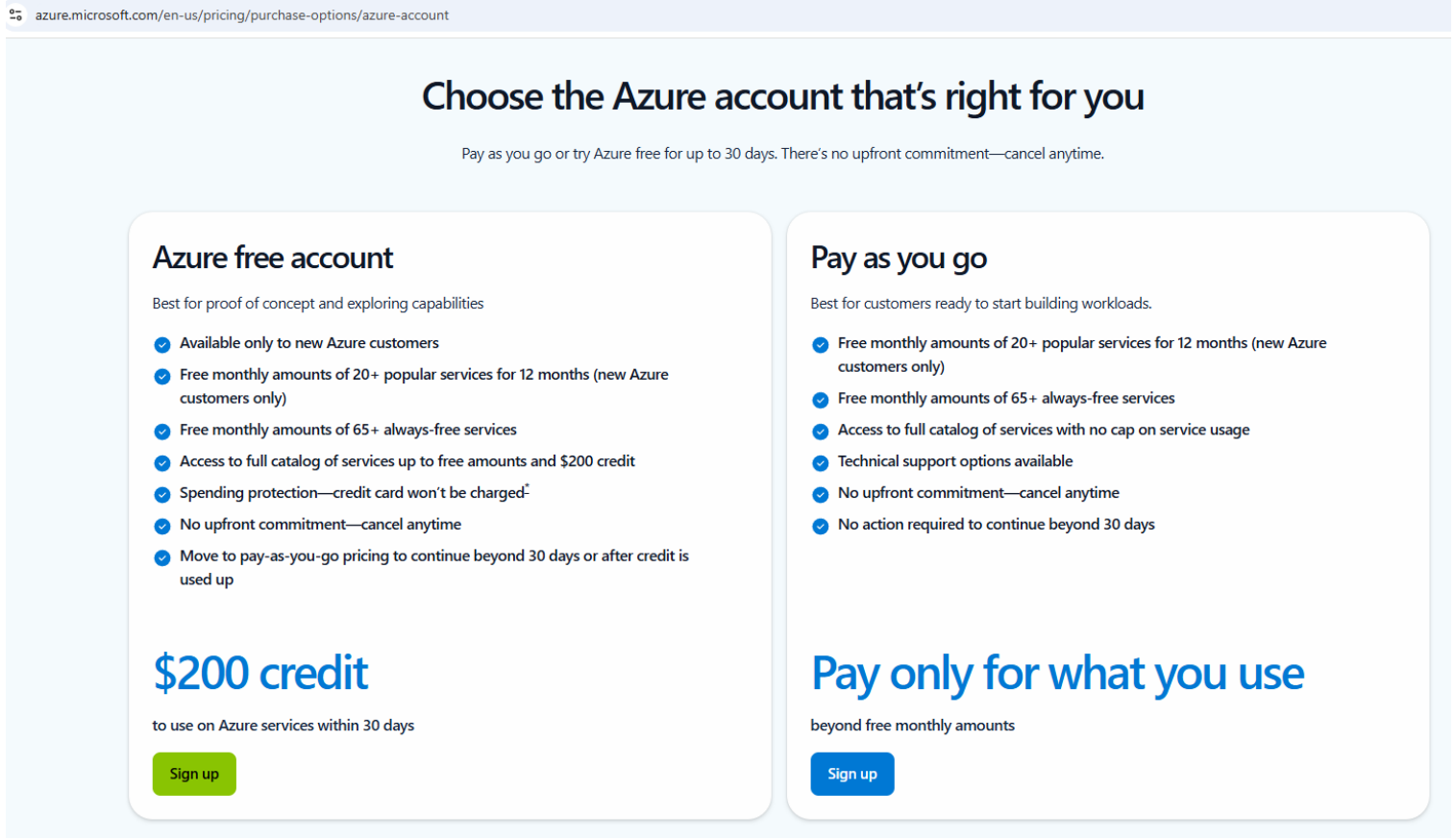
Free tiers:

There are two groups of free services:

- Popular services are free the first 12 months.
- 65+ other services are free always

Azure – Creating an account

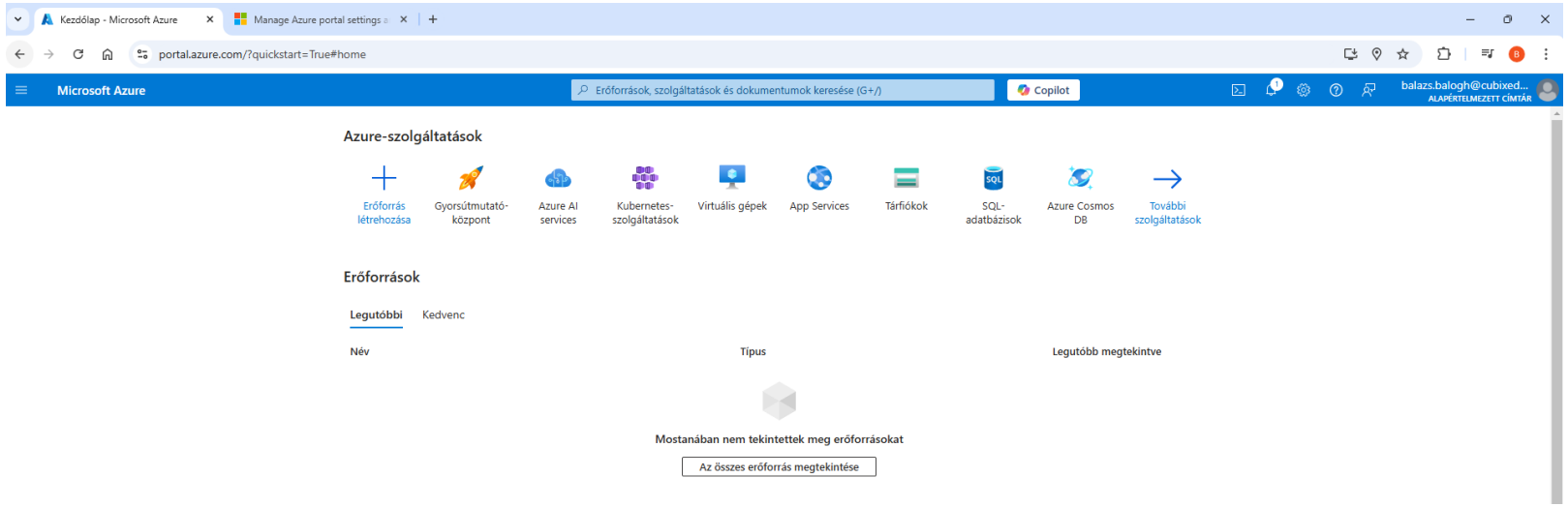
If you don’t have a Microsoft account, create one here.
After that, you are straight in the Azure registration, go through it. **If you are new to Azure, you’ll get 200$ credit for the first 30 days!**



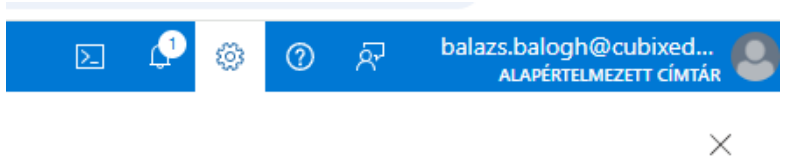
CUBIX
INSTITUTE OF TECHNOLOGY

Azure – Creating an account

After it’s done, you are in (you should set up the two-factor authentication).



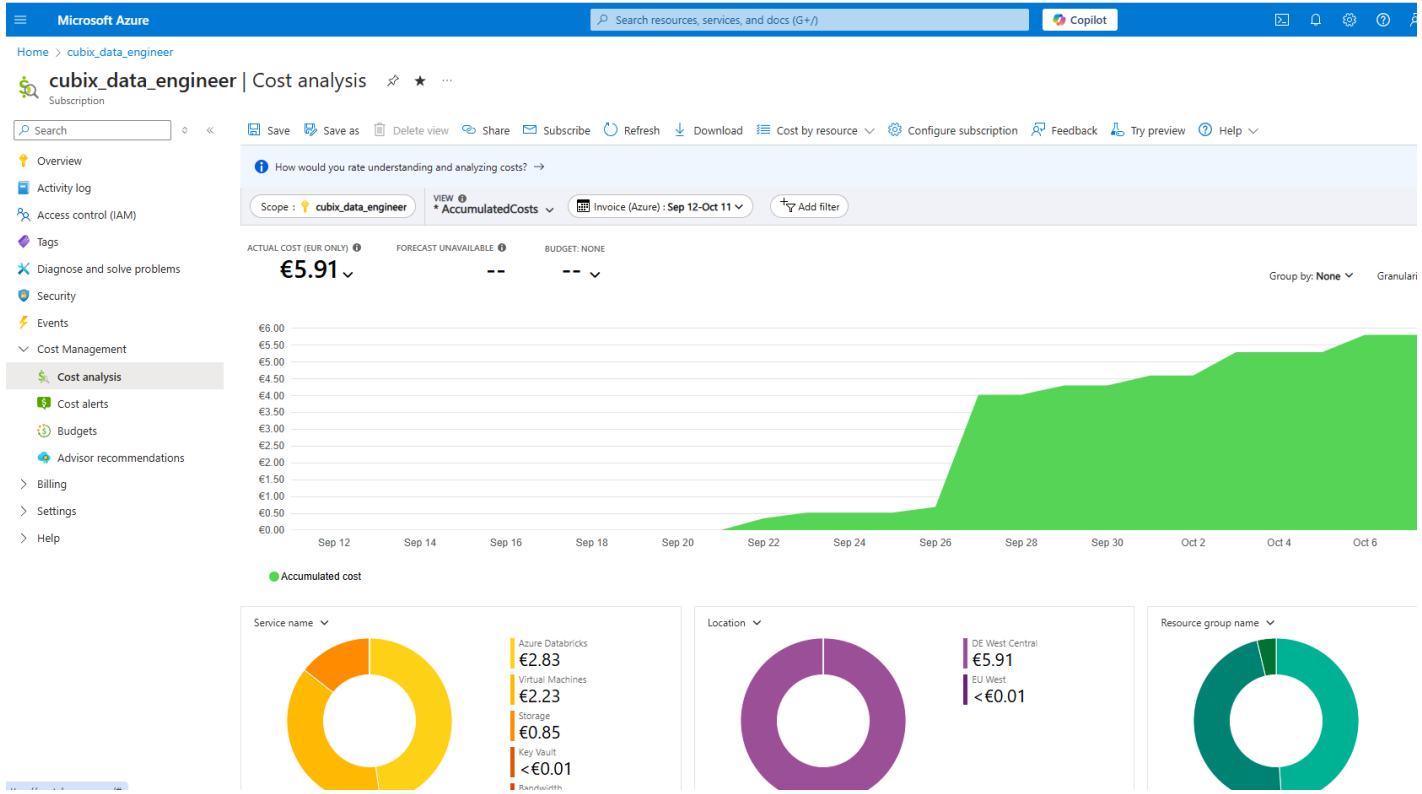
If you are comfortable in English, change here:

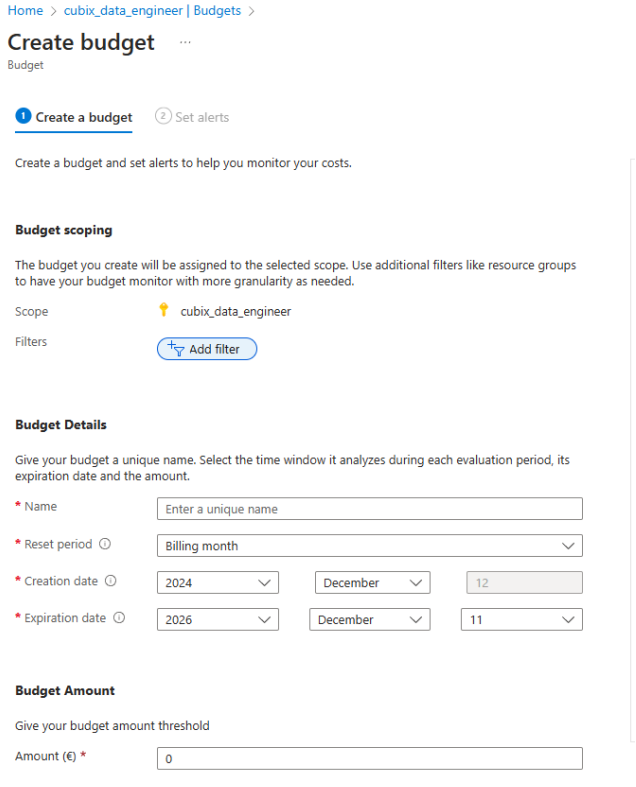


CUBIX
INSTITUTE OF TECHNOLOGY

Azure – Cost analysis

It is important to keep track your costs in Azure.
I advise to set up a budget, which will alert you in email, if you go beyond your threshold (for instance 5 euros per month)

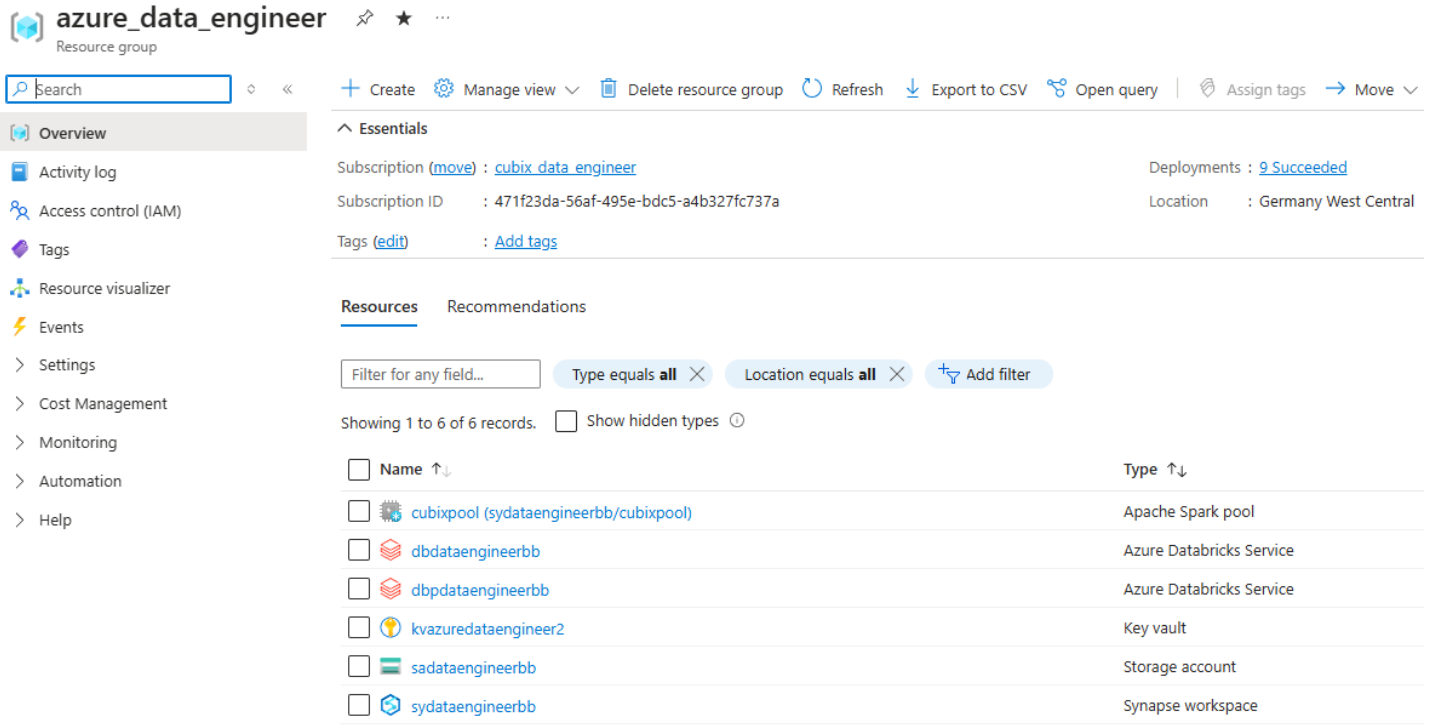




CUBIX
INSTITUTE OF TECHNOLOGY

Azure – Resource Groups

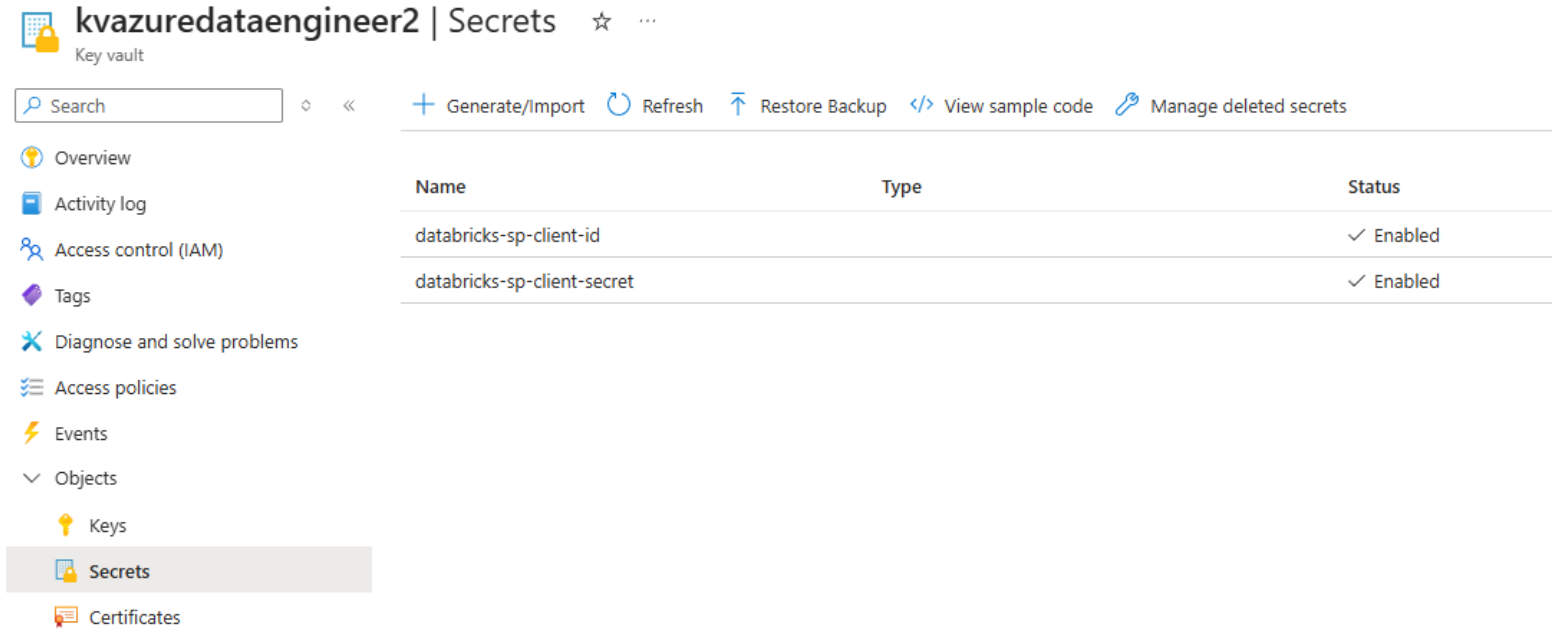
An **Azure Resource Group** is a fundamental organizational unit in Microsoft Azure, designed to manage and group related resources in a cloud solution.
It acts as a container that holds multiple Azure resources, such as virtual machines, storage accounts, databases, web apps, or networking resources, that share a common lifecycle, permissions, and management.



CUBIX
INSTITUTE OF TECHNOLOGY

Azure – Key Vault

Key Vault's purpose is to securely store and manage sensitive information such as secrets, keys, and certificates.
It is designed to help organizations enhance security and control access to their application secrets and encryption keys in a centralized, secure manner.



CUBIX
INSTITUTE OF TECHNOLOGY

Azure – Storage – Data Lake Storage

Azure Data Lake Storage Gen 2 **(ADLS Gen2)** is a highly scalable, secure, and performant data storage service designed for big data analytics. It combines the scalability and cost-effectiveness of Azure Blob Storage with hierarchical file system capabilities.

Key Features

- **Scalability and Cost-Effectiveness:** Built on Azure Blob Storage, it can handle massive amounts of data at a lower cost.
- **Hierarchical Namespace:** Supports directories and file-like structures for efficient data organization and access, enabling faster queries and analytics.
- **Compatibility with Big Data Frameworks:** Works seamlessly with Apache Hadoop, Spark, and other big data tools via HDFS interfaces.
- **Performance:** Optimized for high-performance analytics, especially for batch processing and interactive queries.
- **Security and Compliance:** Provides role-based access control (RBAC), integration with Azure Active Directory, and encryption at rest and in transit.
- **Access Tiers:** Offers hot, cool, and archive storage tiers to optimize costs based on data access frequency.

Azure Storage offers different access tiers, which allows you to store blob object data in the most cost-effective manner possible. The available access tiers include:

- **Hot:** Optimized for storing data that's accessed frequently.
- **Cool:** Optimized for storing data that's infrequently accessed. Data is stored for at least 30 days.
- **Cold tier:** Optimized for storing data that is infrequently accessed or modified. Data is stored for at least 90 days. The cold tier has lower storage costs and higher access costs compared to the cool tier.
- **Archive:** Optimized for storing data that's rarely accessed. The data is stored for at least 180 days with flexible latency requirements, on the order of hours.

Azure – Synapse

Azure Synapse Analytics is an integrated analytics service that brings together big data and data warehousing capabilities. It allows users to ingest, prepare, manage, and serve data for business intelligence (BI) and machine learning (ML) seamlessly.

Key Features

- **Unified Analytics:** Combines data integration, big data processing, and enterprise data warehousing in a single platform.
- **Dedicated and Serverless Options:** Offers **Dedicated SQL Pools** (traditional data warehouse) for predictable performance and **Serverless SQL Pools** for on-demand query processing They are basically parquet (csv, etc.) files, what you can query.
- **Data Integration:** Built-in data pipelines powered by Azure Data Factory simplify data ingestion and transformation.
- **Big Data Support:** Works natively with Azure Data Lake Storage Gen2, enabling massive-scale analytics on unstructured and structured data.
- **Advanced Security:** Integration with Azure Active Directory (AAD), encryption, and features like private endpoints for secure data access.
- **Integration with BI and ML Tools:** Connects seamlessly with tools like Power BI, Azure Machine Learning, and third-party analytics tools.
- **T-SQL Support:** Enables familiar SQL-based querying, making it accessible to data analysts and engineers.
- **Real-Time Analytics:** Processes streaming data using Azure Stream Analytics or Spark in Synapse.

Advantages Over Traditional DWH

- **Unified Platform**: Combines data lake, data warehouse, and big data processing in one.
- **Cost Flexibility**: Pay-as-you-go model for serverless and reserved capacity for dedicated pools.
- **Scalability**: Supports petabyte-scale storage and query processing.

Azure – Data Factory / Synapse / Fabric

These three services often cause a confusion, because seemingly they all for the same use-case, especially **Data Factory** and **Synapse**.

Data Factory is a cloud-based **data integration service** that allows you to create, schedule, and orchestrate data workflows at scale. It's primarily used for data movement, data transformation, and data orchestration tasks.
ADF enables you to ingest data from various sources, transform it as needed, and load it into target data stores.
Azure Synapse Analytics is a cloud-based **analytics service** that integrates big data and data warehousing capabilities. It provides functionalities for data integration, enterprise data warehousing, and big data analytics.

Azure Synapse Analytics can be considered **superset over ADF** as it contains Data pipelines, Notebooks, Databases etc.
Therefore **ADF** is just a **subset** as compared to Synapse.

Microsoft Fabric is an end-to-end SaaS offering that brings together several data and analytics workloads under one roof.

These workloads **include**: Data Factory, Synapse Data Warehouse, Synapse Data Engineering, Synapse Data Science, Synapse Real-Time Analytics, Power BI, and Data Activator.

According to Microsoft, Fabric **offers**:
“Full-service capabilities including data movement, data lakes, data engineering, data integration, data science, real-time analytics, and business intelligence—backed by a shared platform for data security, governance, and compliance. So, your organization no longer needs to stitch together individual analytics services from multiple vendors. Instead, use a streamlined solution that’s easy to connect, onboard, and operate.”

Microsoft Fabric can be seen as an evolution of Azure Synapse Analytics, or Synapse 3.0, though currently Synapse is the most widespread service.

Azure - Databricks

Databricks is a cloud-based unified analytics platform designed to simplify big data and machine learning workflows.
It integrates with Apache Spark to process massive datasets and supports data engineering, data science, and machine learning on a single collaborative platform.

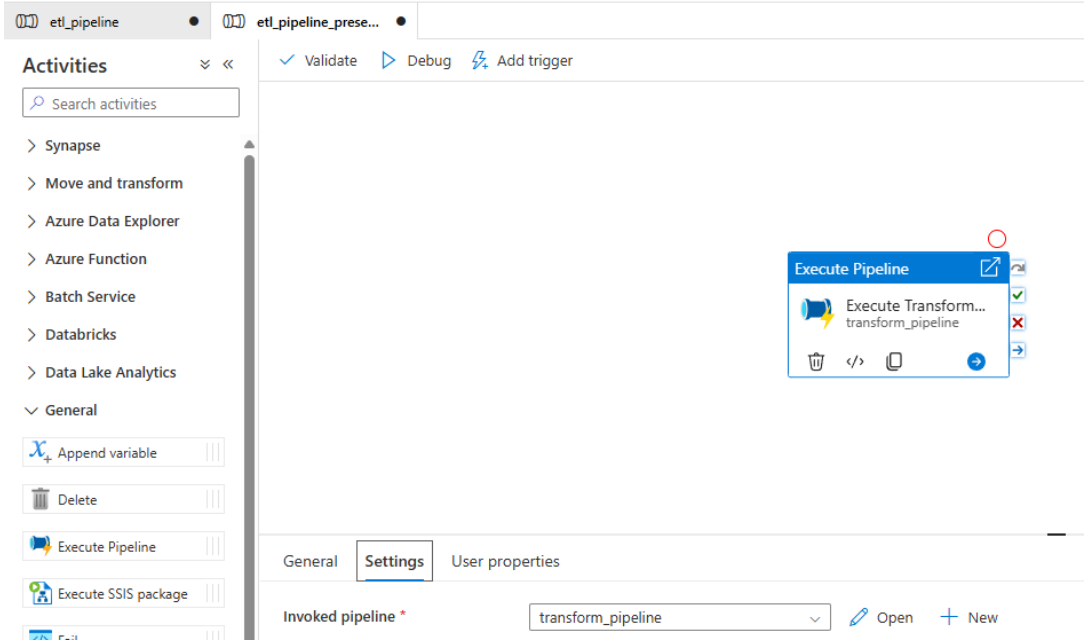
Key Features

- **Unified Workspace:** Combines data engineering, data science, machine learning, and business analytics in one environment.
- **Apache Spark Integration:** Built on Apache Spark, enabling distributed computing for massive datasets and parallel processing.
- **Multi-Language Support:** Supports Python, SQL, Scala, R, and Java, making it flexible for diverse teams.
- **Delta Lake:** Adds reliability to data lakes with ACID transactions, schema enforcement, and time travel for versioning.
- **Interactive Notebooks:** Collaborative notebooks for real-time editing, visualization, and code execution.
- **Machine Learning:** Built-in libraries for feature engineering, model training, and hyperparameter tuning.
- **AutoML:** Simplifies machine learning by automatically generating models for quick experimentation.
- **Integration with Azure:** Fully integrates with Azure Data Lake, Synapse, and Azure Machine Learning.
- **Scalability:** Auto-scaling clusters allow handling workloads of any size, from small data to petabytes.
- **Security and Governance:** Includes Unity Catalog for centralized access control, data lineage, and role-based permissions.

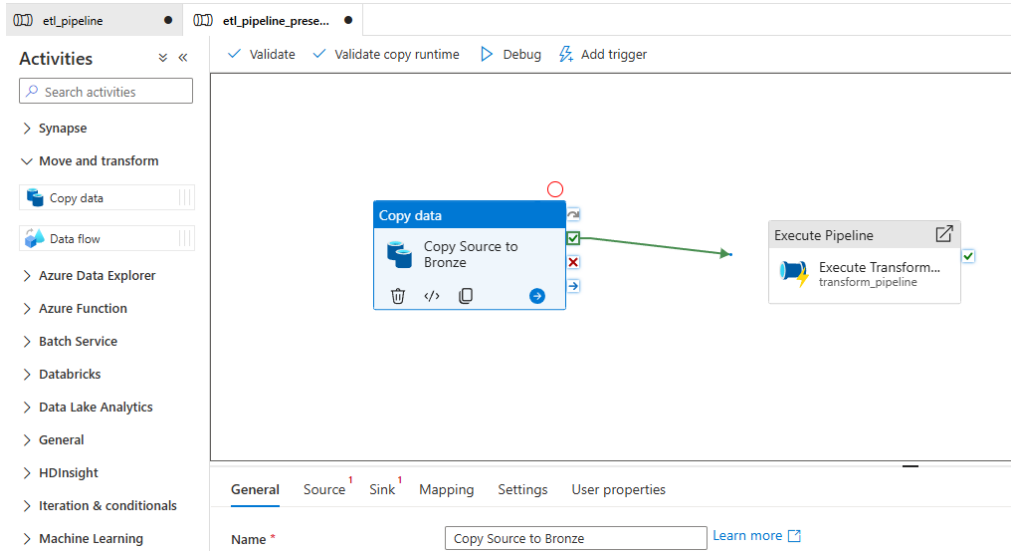
Azure – Synapse – Video “Synapse pipeline parameters”

The creation of the **etl_pipeline** cannot be seen on the video, it starts from adding the parameters to it.
Here are the steps to build the pipeline. After you can continue with the video.

1. Create a new pipeline, call it **etl_pipeline**, add an **Execute pipeline** activity (it’s under General), set it’s **Name** to **“Execute Transform Pipeline”**, and under **Settings** choose your **transform_pipeline** as **the Invoked pipeline**:



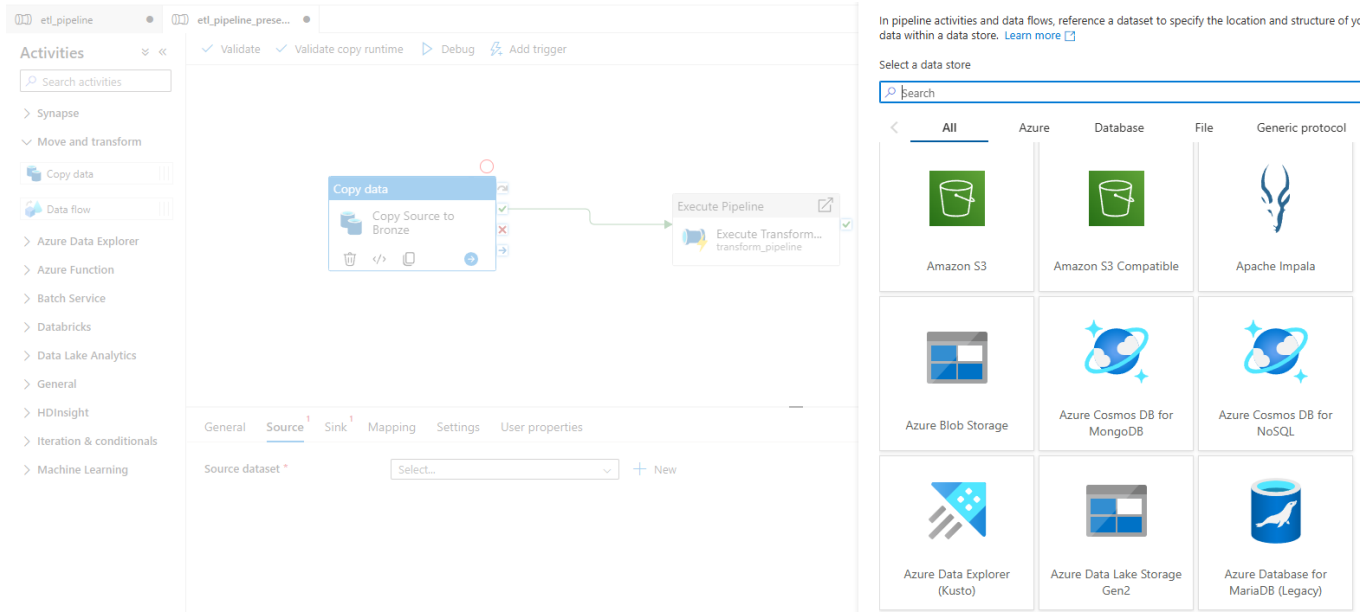
2. Then create a **Copy data,** name it **“Copy Source to Bronze”** and grab the green check mark and connect it with the **Execute Pipeline.**



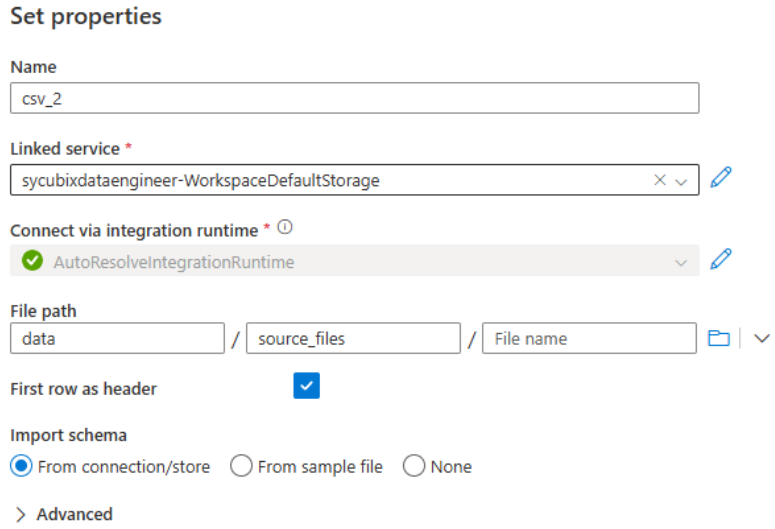
CUBIX
INSTITUTE OF TECHNOLOGY

Azure – Synapse – Video “Synapse pipeline parameters”

3. Setting up the **Copy Source to Bronze** activity first with the **Source** tab. Click on **+ New** at the **Source dataset** and select **Azure Data Lake Storage Gen 2**, then select **Delimited text**, as our source file will be a csv.



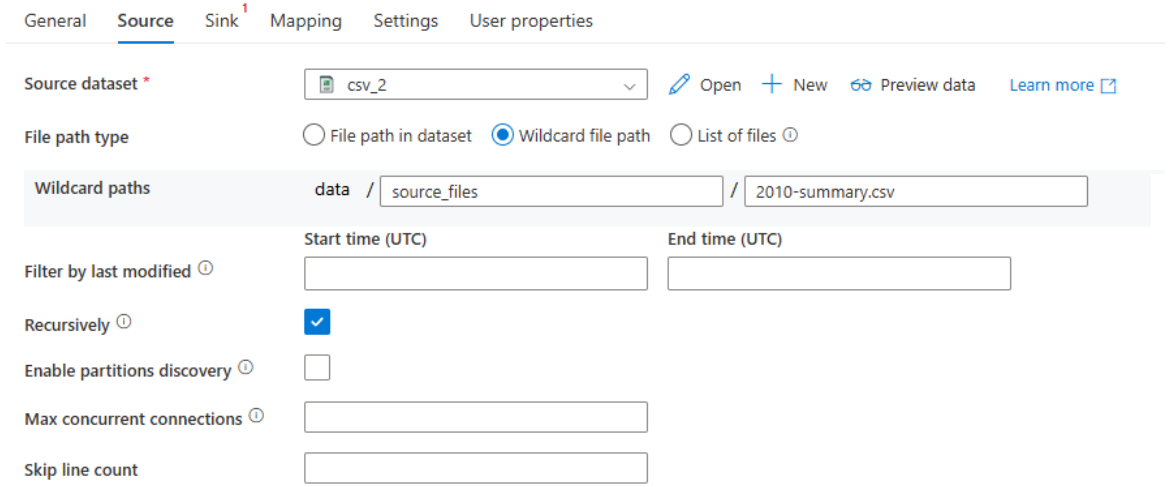
Name it **“csv”** and select the **WorkspaceDefaultStorage.** Then browse your **“source_files”** folder, **First row as header** is ticked in.

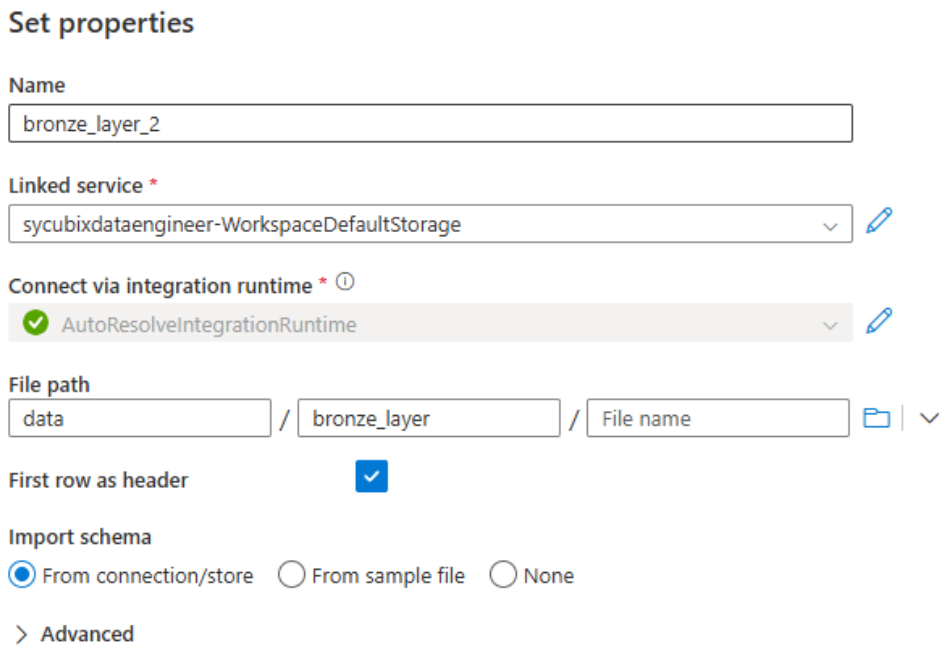


CUBIX
INSTITUTE OF TECHNOLOGY

Azure – Synapse – Video “Synapse pipeline parameters”

3. Returning to the **Copy Source to Bronze** activity’s main page now **Source** should look like this:





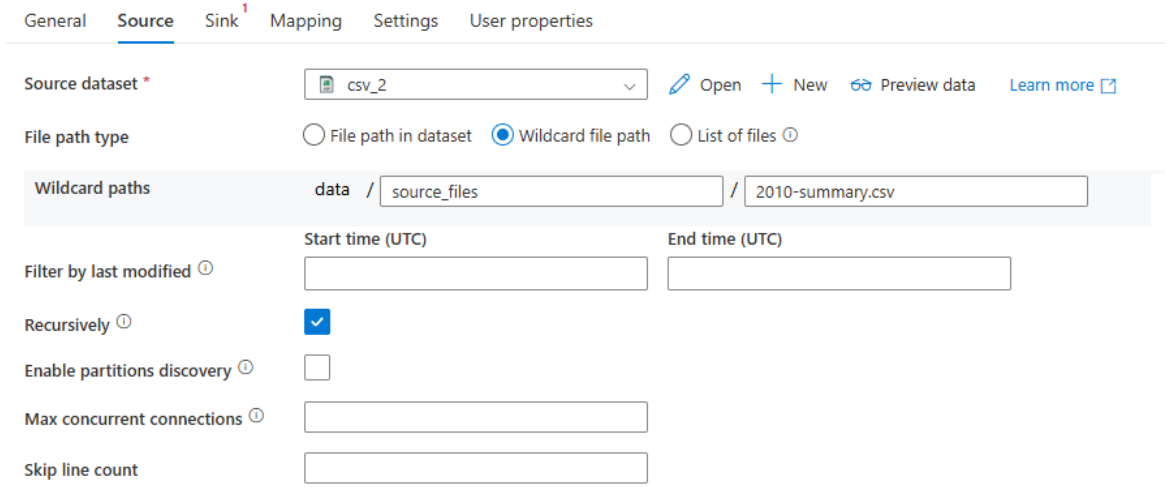
4. Let’s create the **Sink** part, create a new **Sink dataset**, choose **Data Lake Storage Gen 2**, set it’s name to **“bronze_layer”** and the **Linked service** to the **WorkspaceDefaultStorage**. **File path** will be **data / bronze_layer**.

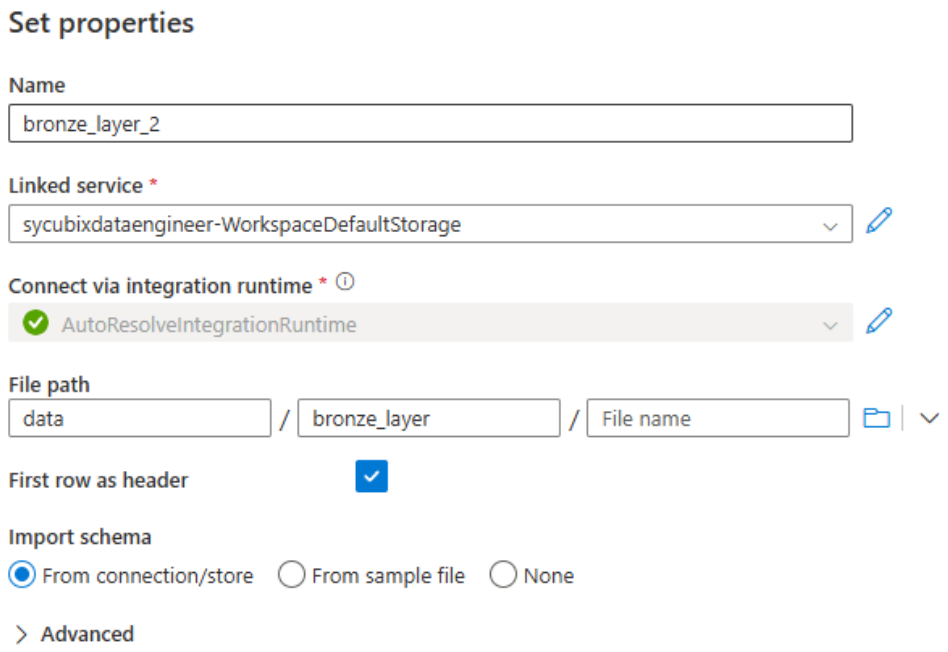
5. After returning to the **Copy Source to Bronze** again, change the **Copy behaviour** to **Preserve hierarchy**, and the **File extension** to **.csv**.

CUBIX
INSTITUTE OF TECHNOLOGY

Azure – Synapse – Video “Synapse pipeline parameters”

3. Returning to the **Copy Source to Bronze** activity’s main page now **Source** should look like this:





4. Let’s create the **Sink** part, create a new **Sink dataset**, choose **Data Lake Storage Gen 2**, set it’s **Name** to **“bronze_layer”** and the **Linked service** to the **WorkspaceDefaultStorage**. **File path** will be **data / bronze_layer**.

5. After returning to the **Copy Source to Bronze** again, change the **Copy behaviour** to **Preserve hierarchy**, and the **File extension** to **.csv**.

These were all the steps, now test it, click on **Debug**.
The pipeline will **extract** the **2010-summary.csv** from your **source_files** folder, **copy** it to the **bronze_layer** folder, then the notebook will create a **transformed** parquet from it and **load** it into the **silver_layer** folder.

Congratulations, you have your first ETL process in Synapse!

CUBIX
INSTITUTE OF TECHNOLOGY