Fachbereich Mechatronik und
Maschinenbau

Hochschule Bochum
Bochum University
of Applied Sciences

Praktikum CPS

Versuch 2: „ETL Process“

1. Microsoft Azure: Portal and DevOps

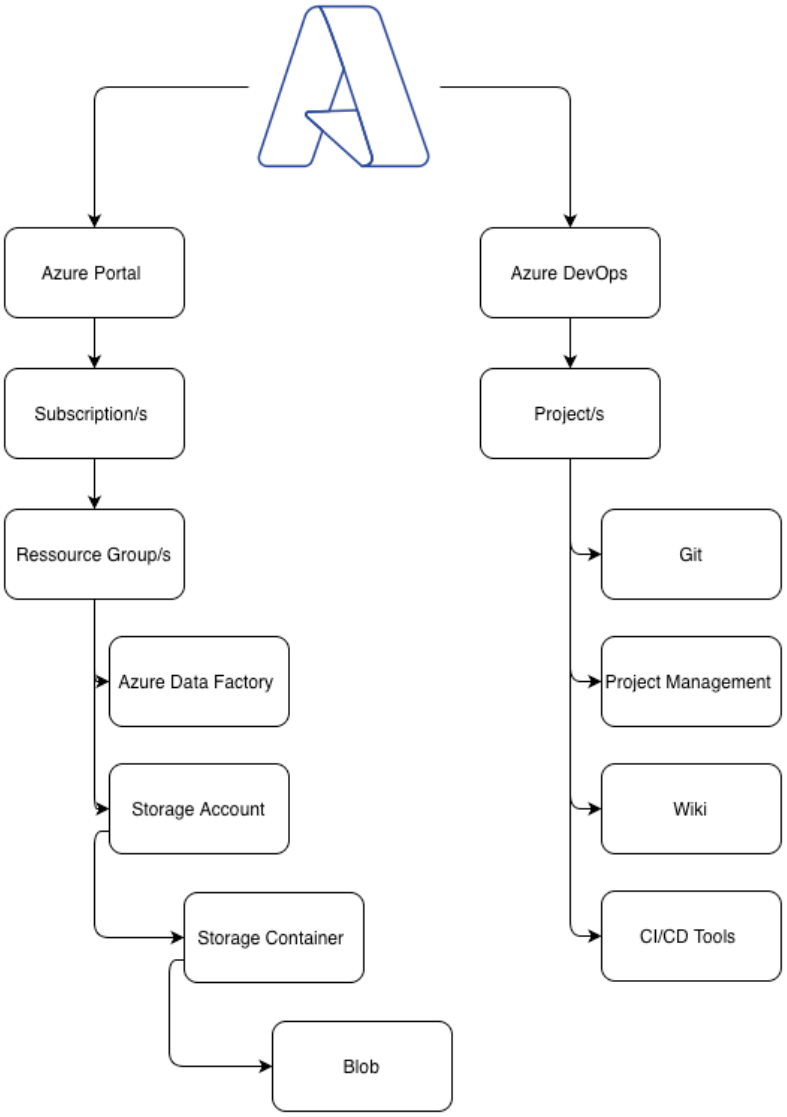

Abbildung 1 - Azure Cloud structure

2. IaS: Azure Data Factory Deployment

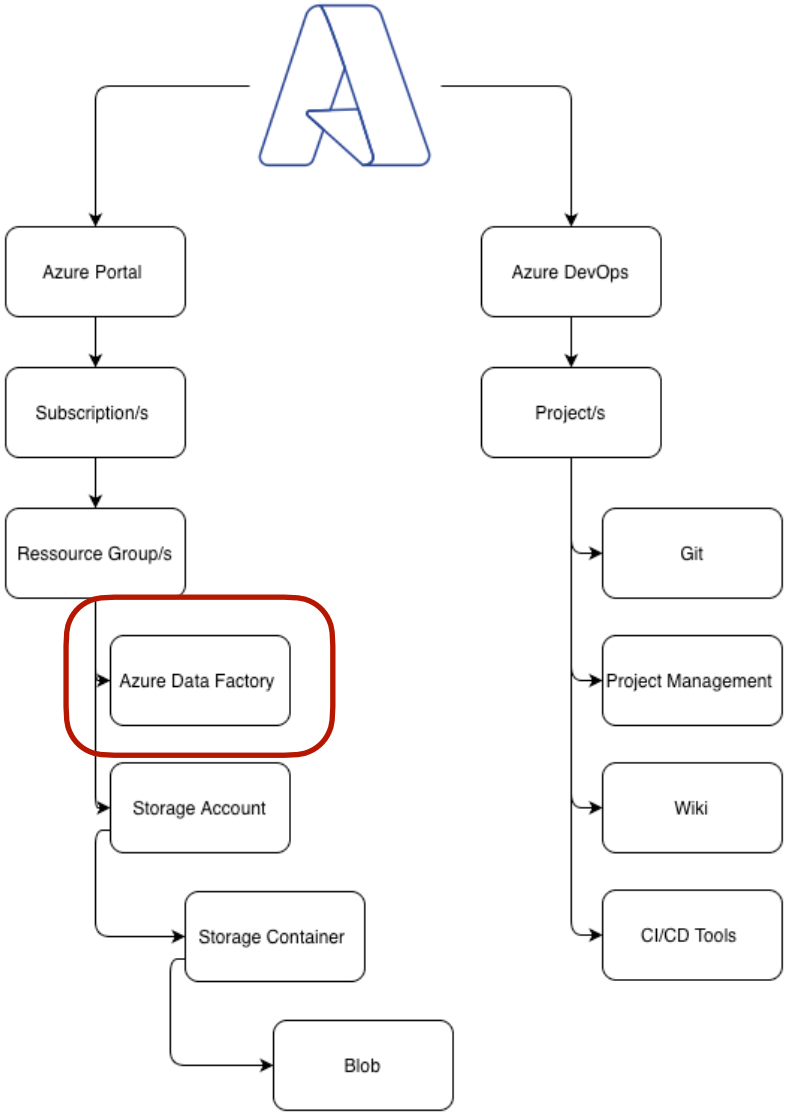

Abbildung 2 - Benötigte Azure Ressource

3. ADF: ETL Pipeline

Im letzten Versuch haben wir Azure Data Factory erstellt. Jetzt suchen Sie bitte in Portal nach dieser Ressource und klicken Sie auf „**Launch Workspace**“.

Wie bereist besprochen, dient ADF als eine Umgebung für die Entwicklung von ETL-Processes. Eine ETL Pipeline wird verwendet, um Daten von den Quellen zu laden und in den Senke zu speichern. Dabei können unterschiedliche Systeme als Source und Sink verwendetet werden.

Eine ETL Pipeline besteht grundsätzlich aus den folgenden Teilen: **Linked Service**, **Dataset** und **Pipeline Activities**.

1. Linked Service beschreibt, **wo** die Daten abgeholt bzw. gespeichert werden müssen. (Im Klartext: Connection)
2. Dataset beschreibt, **wie** oder **welche** Daten abgeholt bzw. gespeichert werden müssen. (Im Klartext: Query)
3. Pipeline Activity beschreibt, **was** mit den Daten geschehen werden muss.

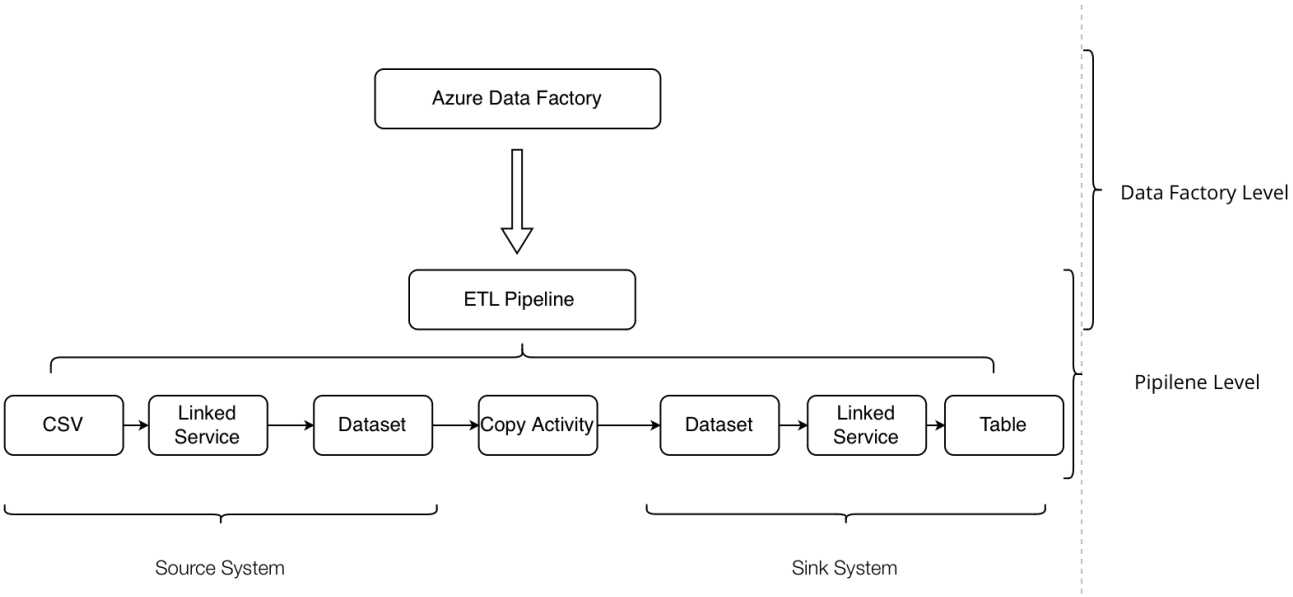

Abbildung 3 - Data Factory Struktur

3.1. Linked Services

Linked Services können als die Verbindungsstring oder die Autorisierung angesehen werden, die die Pipeline verwendet, um sich mit einem Source-System zu verbinden. Ein Linked Service kann Anmeldeinformationen wie Benutzername und Passwort für die Autorisierung in einem Quell- oder Senksystem enthalten. Als Quell- oder Senksystem können Blob Storage, SAP System oder Business Warehouse (BW) dienen

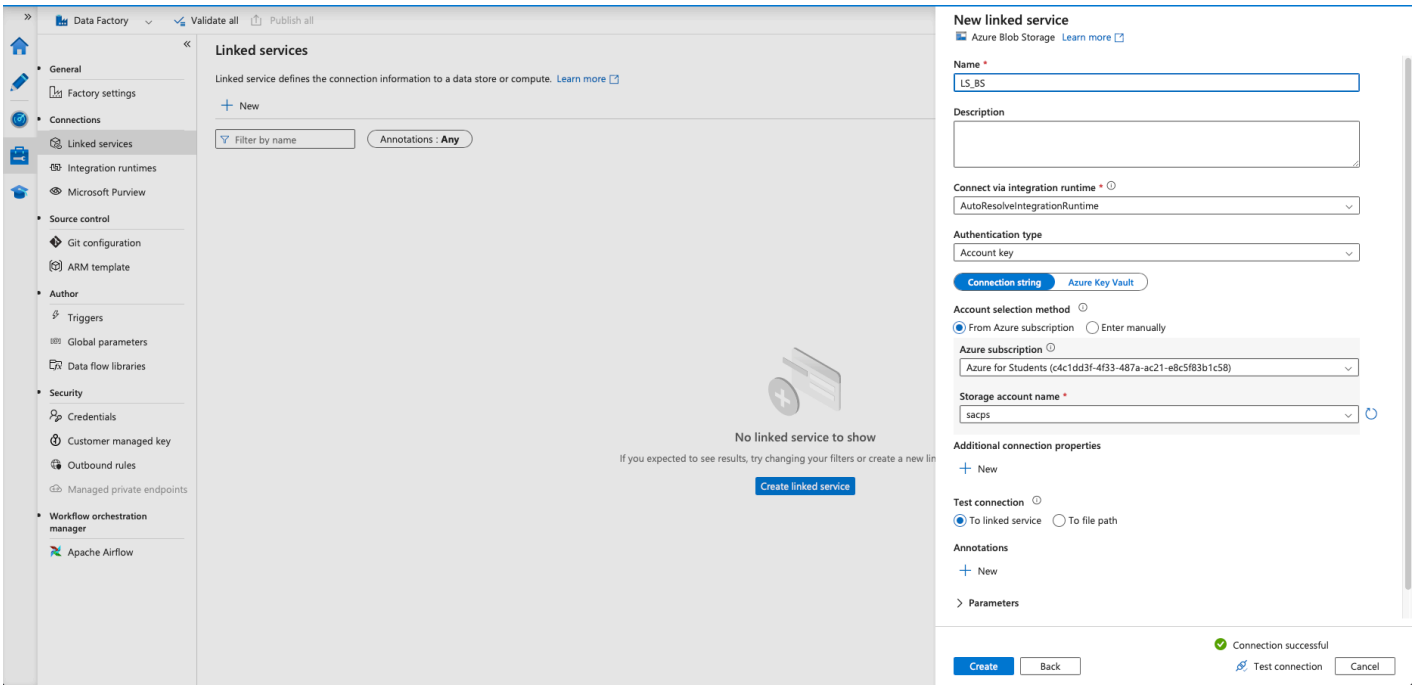

Abbildung 4 - Linked Service

3.2. Dataset

Ein **Dataset** kann als die Quelle und das Senke einer Pipeline Aktivität betrachtet werden. Eine Pipeline kann mehrere Datasets erhalten. Dataset ist also ein Begriff, um die Tabellen und Datenspeicher zu beschreiben.

In der Regel haben ein Dataset und ein Linked Service **Eins-zu-Eins-Beziehung**.

Ein mögliche Szenario:

Wenn Daten aus einem SAP-System extrahiert werden sollen, kann ein SAP-System Tausende von Tabellen enthalten. Im **Linked Service** gibt der Entwickler die IP-Adresse des Systems und die Autorisierungsdaten an. Im **Dataset** sollten jedoch die Namen der Tabellen angegeben werden, die extrahiert werden sollen.

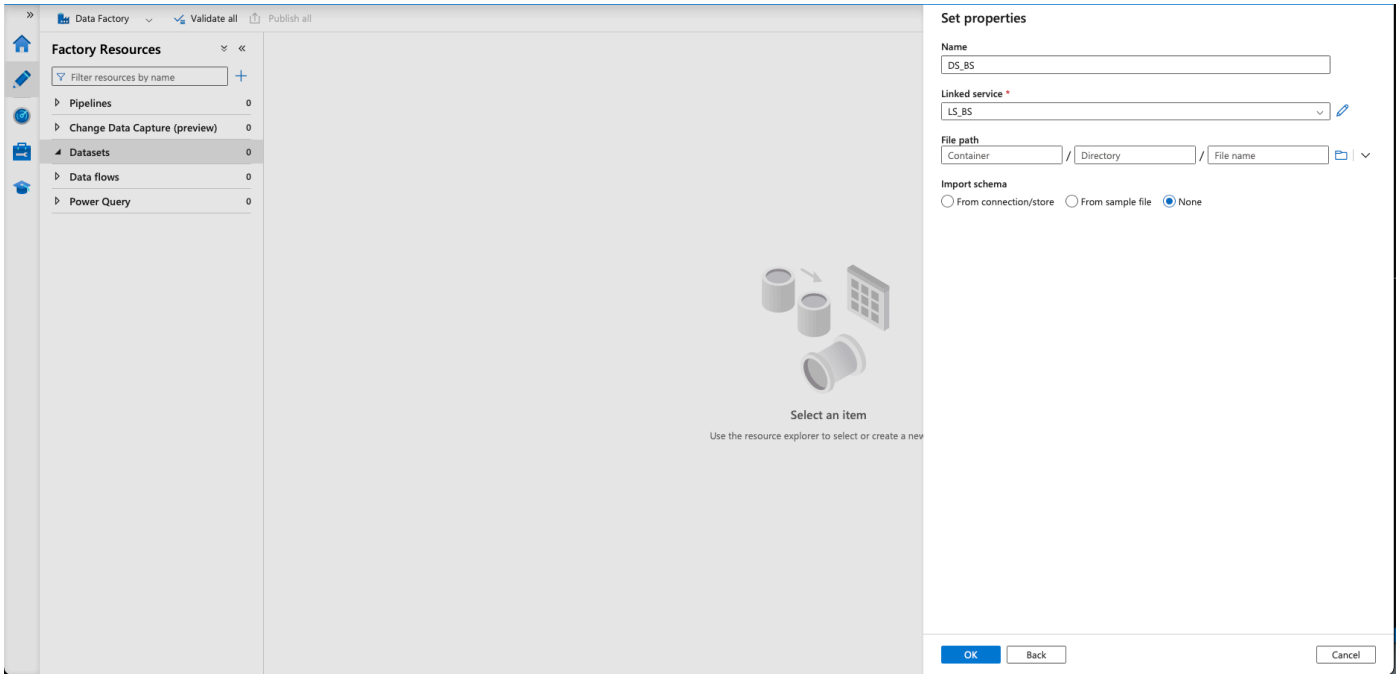

Abbildung 5 - Dataset

3.3. Pipelines Activity

Pipeline Aktivität steht für **T**ransformation in der **ETL** Prozessen.

Je nach Anforderungen kann dies unterschiedlich sein, von metadata-driven-processing bis einfachen Copy Activity.

Wir werden uns eine einfache **Copy Activity** beschäftigen.

4. Entwicklung erster ETL Pipeline

1. Öffnen Sie Data Factory.

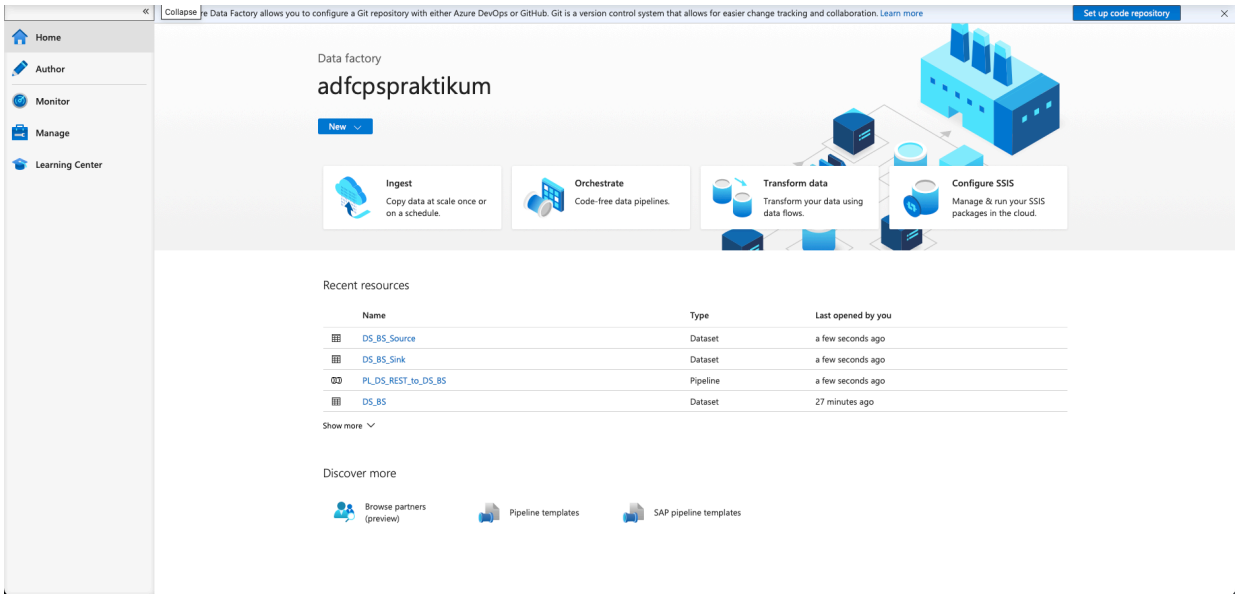

Abbildung 6- Data Factory Homepage

2. Klicken Sie auf ‚Manage' -> ‚Linked Services' -> ‚New', um ein neues Linked Service zu erstellen.
    1. Als ‚Data Store' wählen Sie ‚Azure Blob Storage' aus.
    2. Verwenden Sie den folgende Name: ‚**LS_BS**'
    3. Wählen Sie aus dem Drop-Down Menu ‚**Azure Subscription**' und ‚**Storage account name**'
    4. Schließlich klicken Sie auf ‚**Test connection**'. Wenn alles richtig gemacht wurde, wird eine grüne Meldung mit ‚connection successful' angezeigt. Das bedeutet, dass Ihre Data Factory erfolgreich Verbindung zu Ihren Storage hergestellt hat.

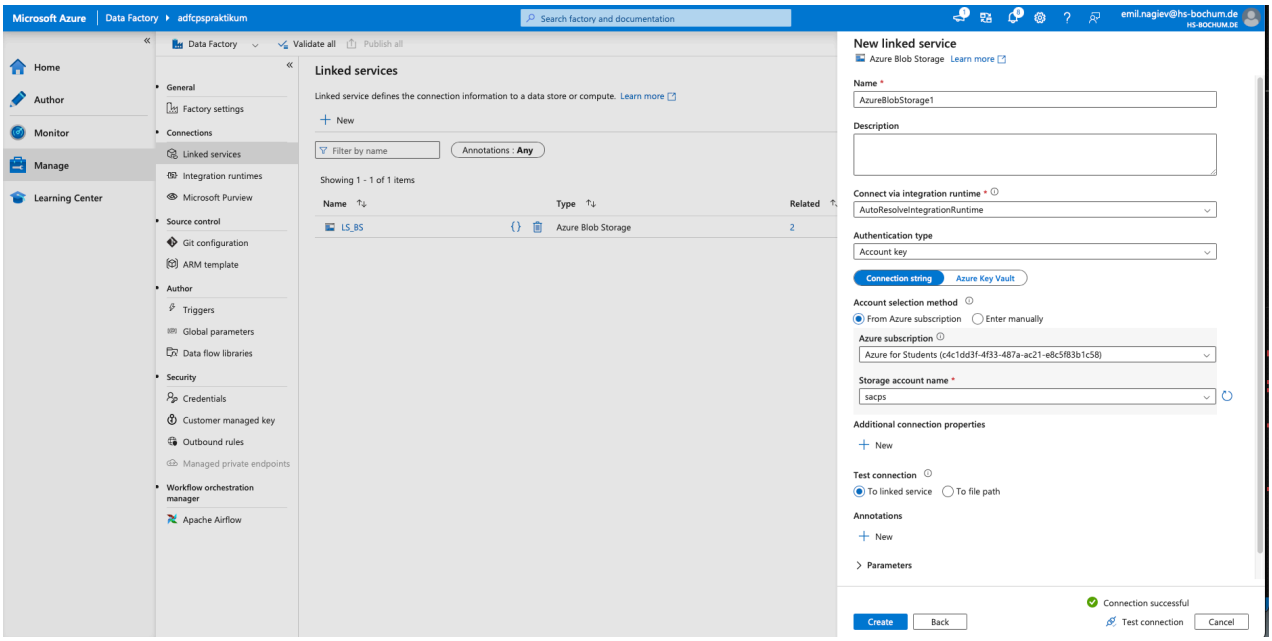

Abbildung 7 - Linked Service

3. Wechseln Sie zur Seite ‚**Autor**‘ und unter ‚**Factory Resources**‘ finden Sie Unterseite ‚**Datasets**‘. Erstellen Sie ein neues Dataset.
    1. Als ‚**Data Store**‘ wählen Sie ‚**Azure Blob Storage**‘ aus.
    2. Als ‚**Format type**’ wählen Sie **CSV Format**
    3. Geben Sie folgenden Namen - „**DS_BS_Source**“.
    4. Aus dem Drop-Down Menu wählen Sie den Linked service, den Sie im vorherigen Schritt erstellt haben.
    5. Unter **Path** geben Sie den Namen von **sourcecontcps** und unter File name - **sensor.csv** ein und klicken Sie auf ‚Ok‘.
    6. Erstellen Sie nun das zweite Dataset. Als den Namen verwenden Sie - **DS_BS_Sink** und als Containername in Path geben Sie - **sinkcontcps**.

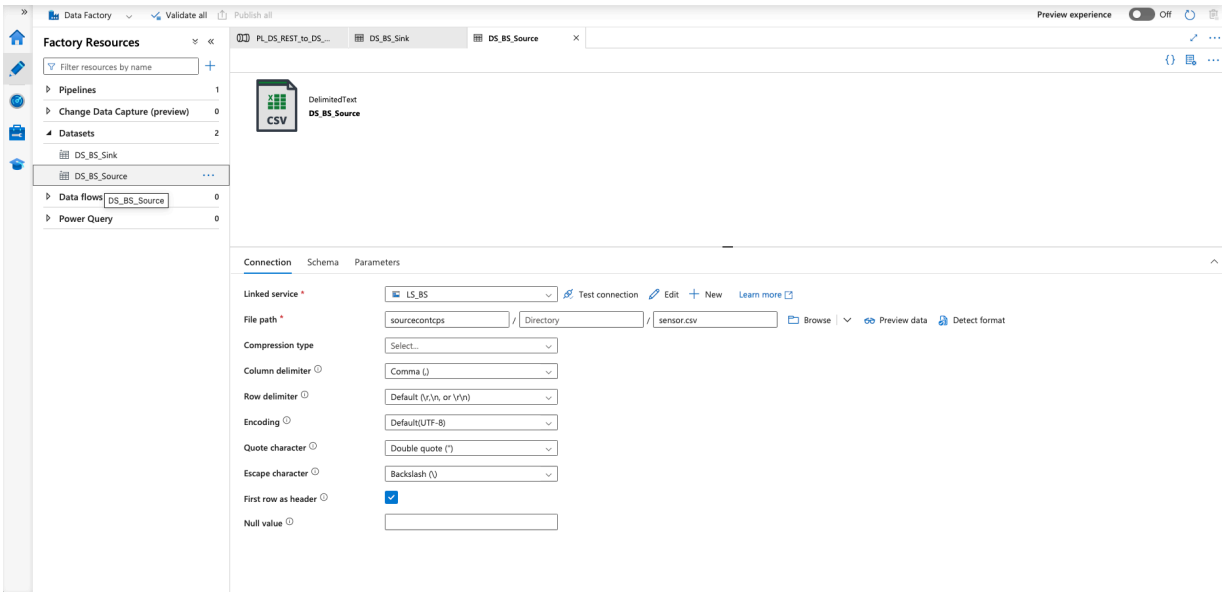

Abbildung 8 -Dataset

4. Unter ‚Pipelines' erstellen Sie eine neue Pipeline.
    1. Der Pipelinename sollte wie folgt definiert werden - „**PL_DS_REST_to_DS_BS**"
    2. Als Activities wählen Sie ‚**Copy Data**'
    3. Als Source-dataset wählen Sie aus dem Drop-Down Menu ‚DS_BS_Source'. Wenn an dieser Stelle alles richtig gemacht wurde, können Sie auf ‚Preview data' klicken um ein Snapshot von dem sensor.csv File anzuzeigen.
    4. Als Sink-dataset wählen Sie aus dem Drop-Down Menu ‚**DS_BS_Sink**'.
5. In der oberen Leiste (Data Factory Level) finden Sie zwei Button ‚**Validate all**' und ‚**Publish all**', klicken Sie auf beide, um Ihre Änderungen zu speichern.
6. Nun können Sie die Pipeline ausführen, indem Sie auf ‚Add trigger' -> ‚Trigger now' klicken.
7. Die Pipeline wird ausgeführt, und die .csv-Datei wird von den Source Container zum Sink Container kopiert.
8. Unter dem Verzeichnis ‚**Monitor**' können Sie alle Pipeline-Ausführungen in Data Factory überwachen. Klicken Sie auf Ihre ausgeführte Pipeline und dann neben ‚Copy data' auf ‚Details'(Brillen Symbol). In diesem Fenster werden alle Details zu dieser Ausführung angezeigt. Hier sehen Sie, dass 1 Datei gelesen und gespeichert wurde.

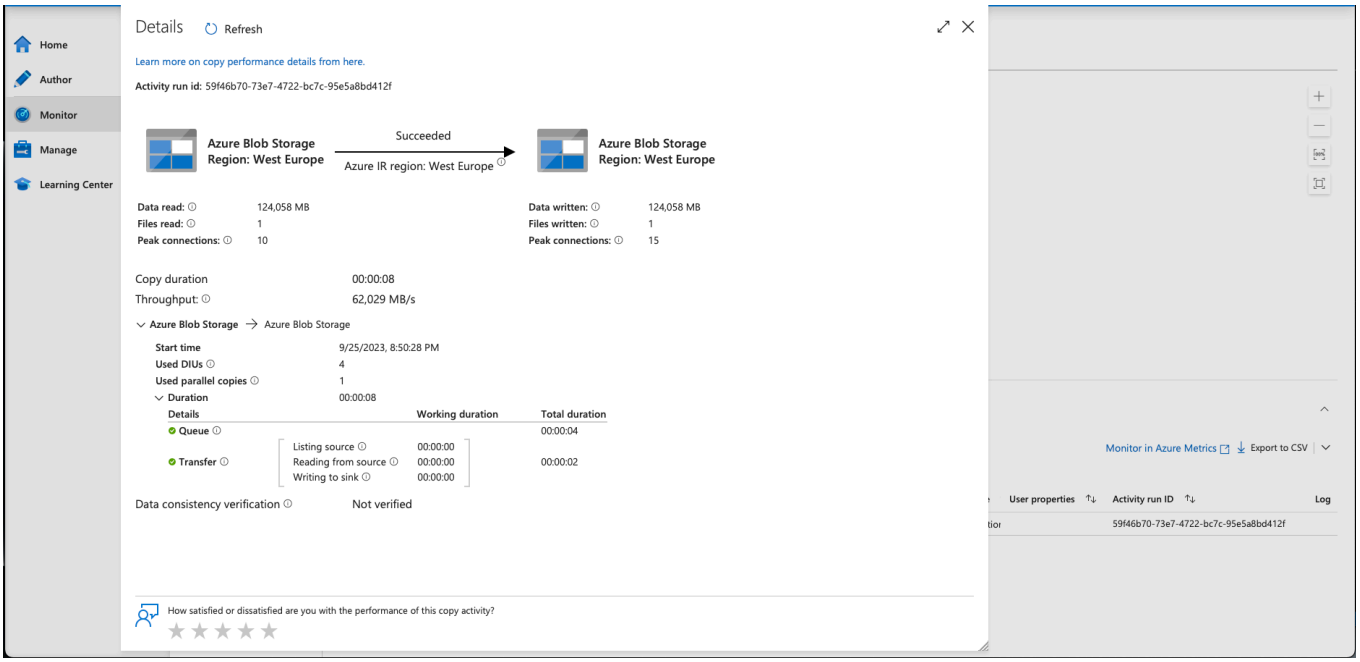

Abbildung 9 - Details zur Ausführung

9. Gehen Sie zum Portal und überprüfen Sie, ob die Datei sensor.csv im Sink-Container vorhanden ist.