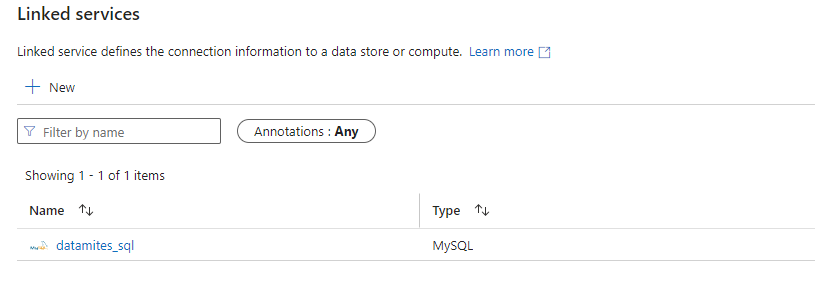
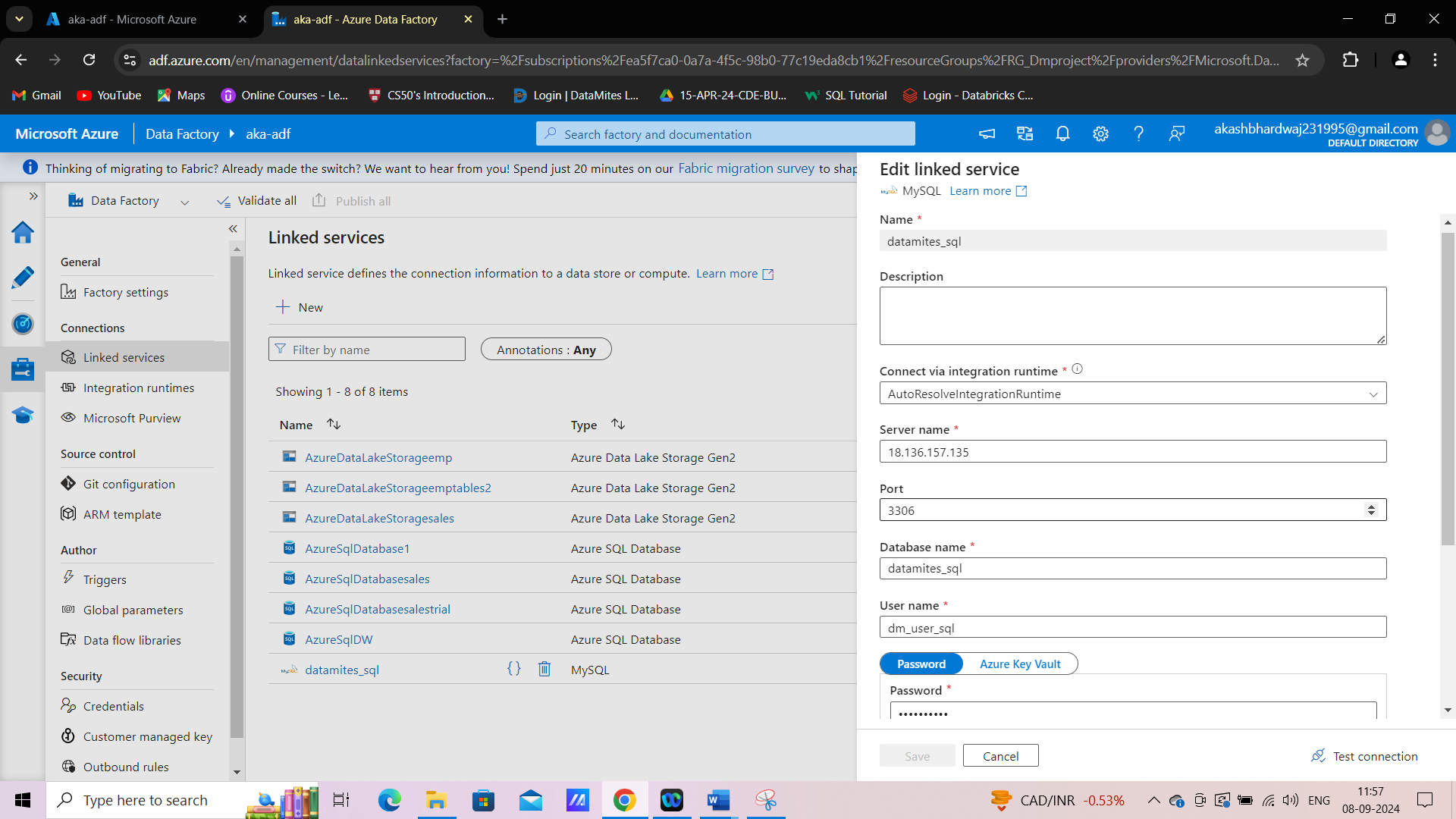
**Extraction:**

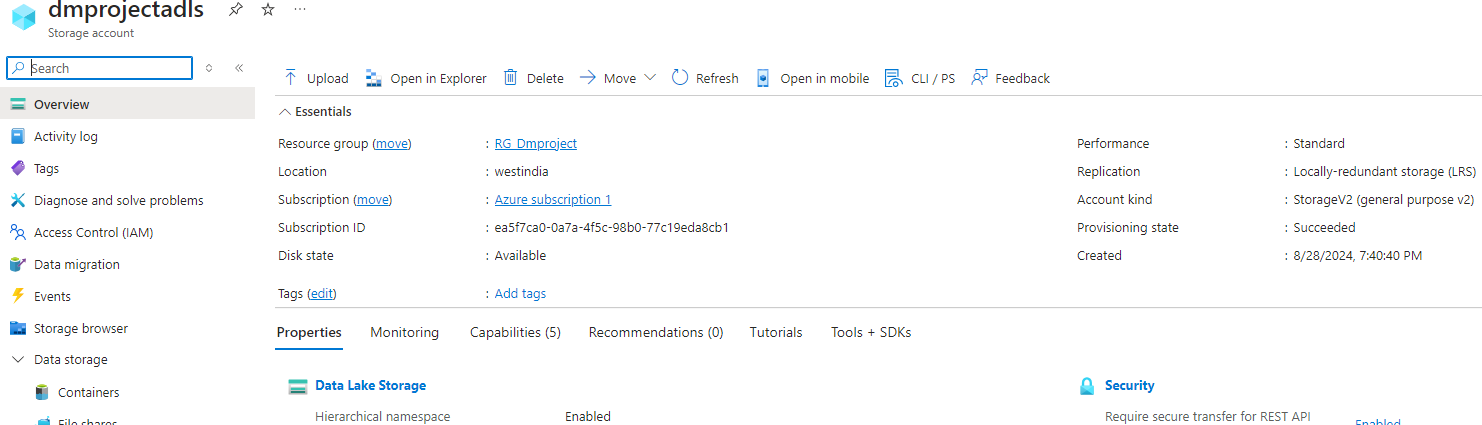
Steps performed to extract the data from Datamites MySql server:

1. Created an Azure Data Factory resource.
2. Created a MySQL linked service using UI.

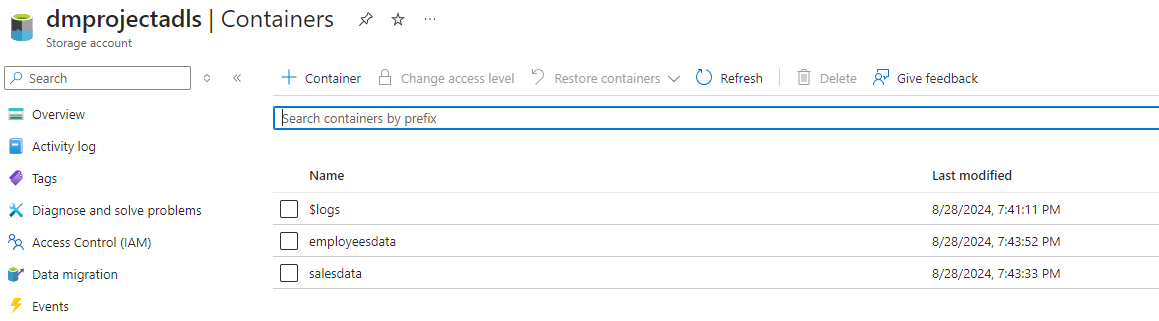




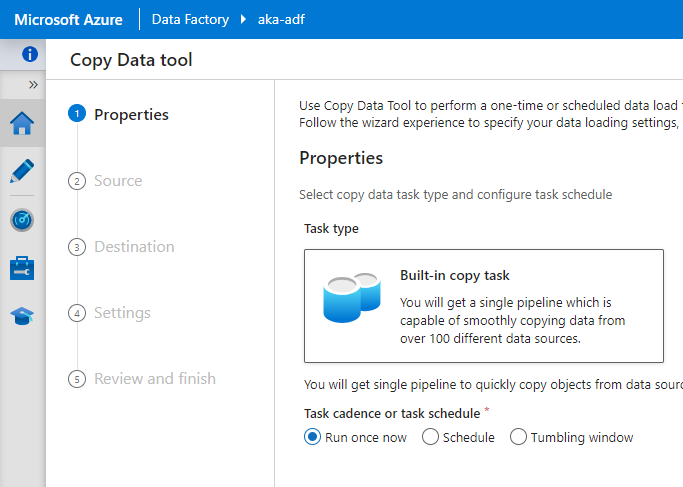
1. Created an Azure Data Lake Storage service account.



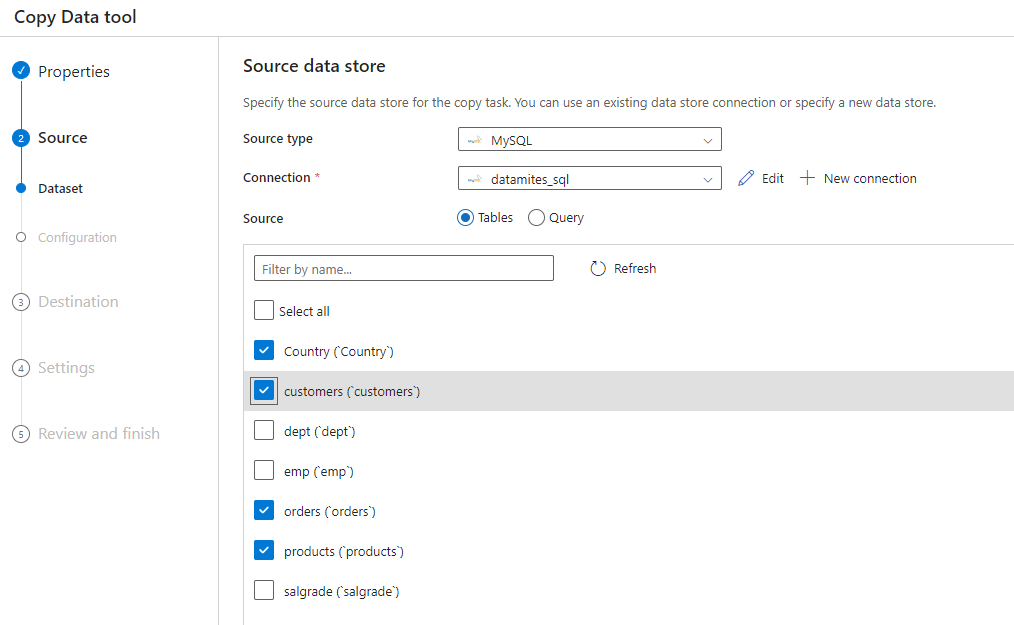
1. Inside this storage account we created two containers named “employeesdata” and “salesdata”.

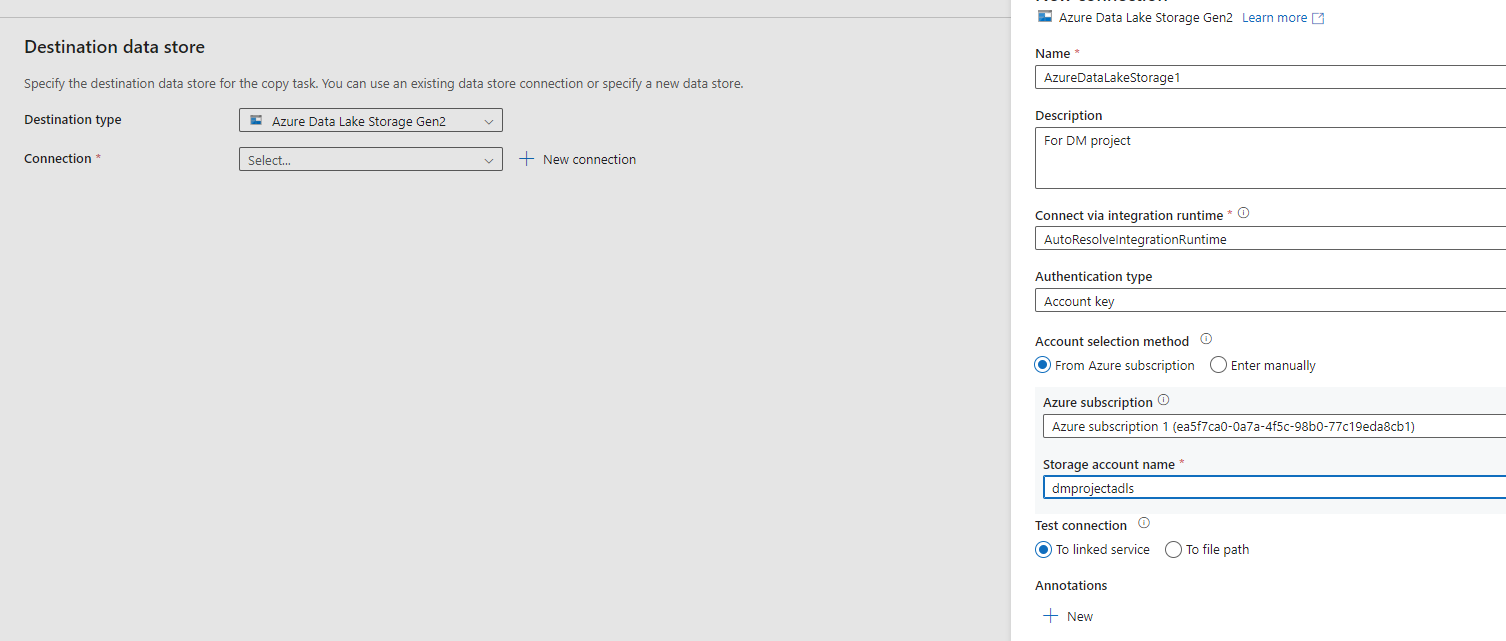


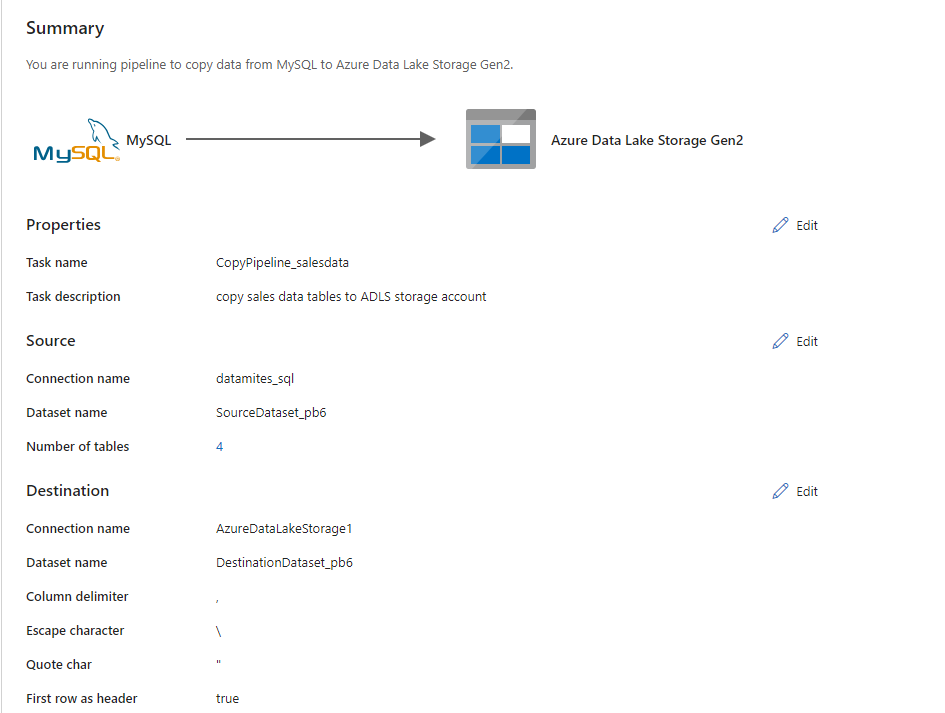
1. Used the “Copy data tool” of ADF to move the tables from Datamites MySQL server to ADLS storage container by converting them into CSV format and inferring source schema:

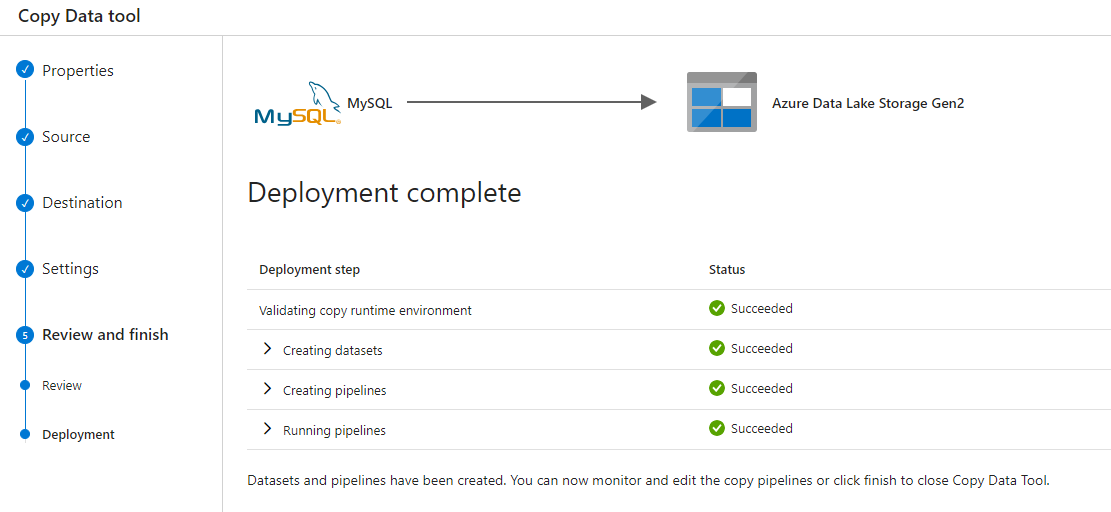


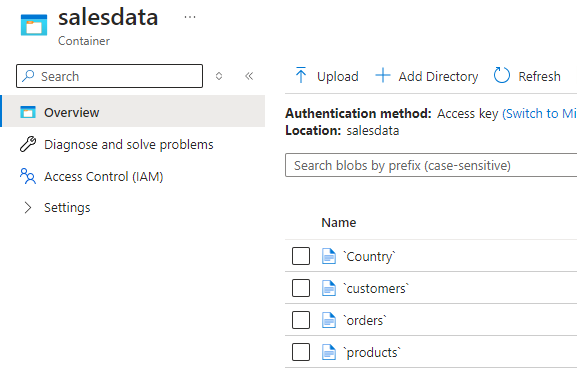
1. **For sales dataset:**



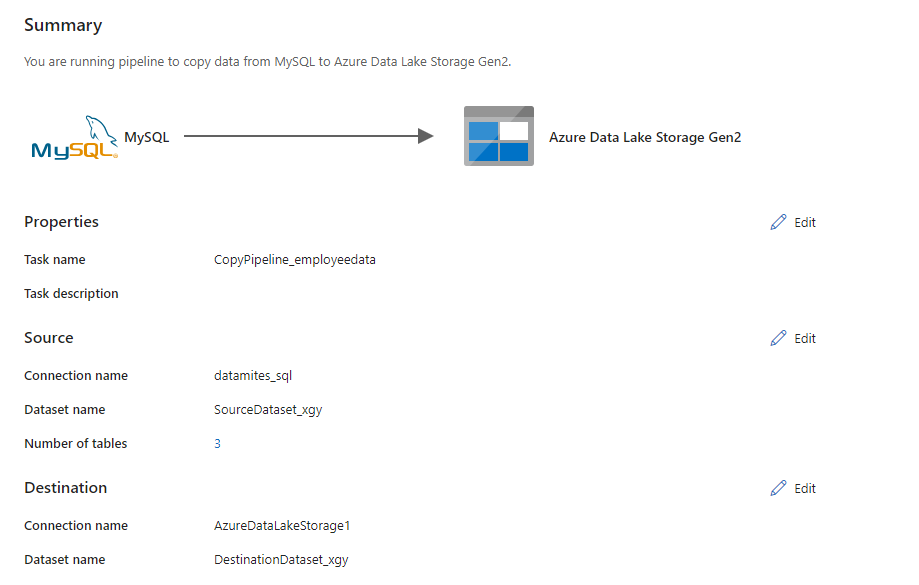


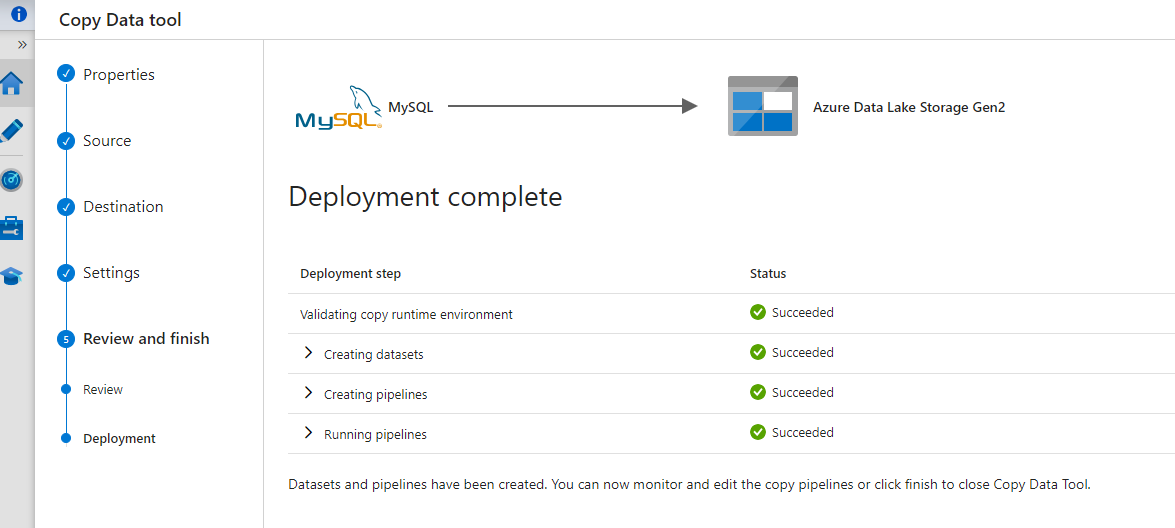


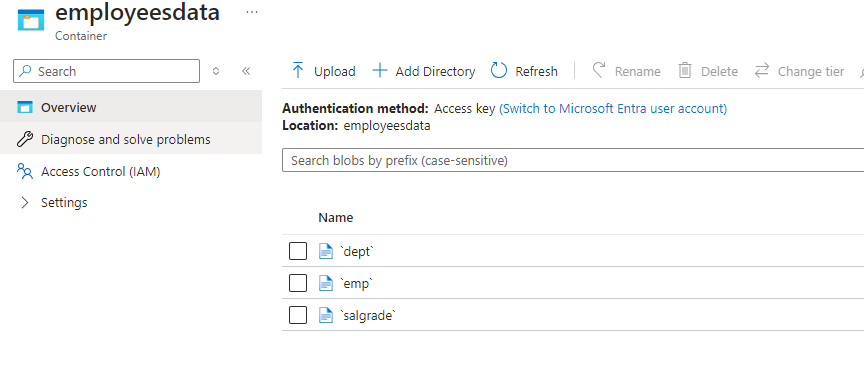




1. **Performed the same steps to copy employees’ data tables to ADLS:**

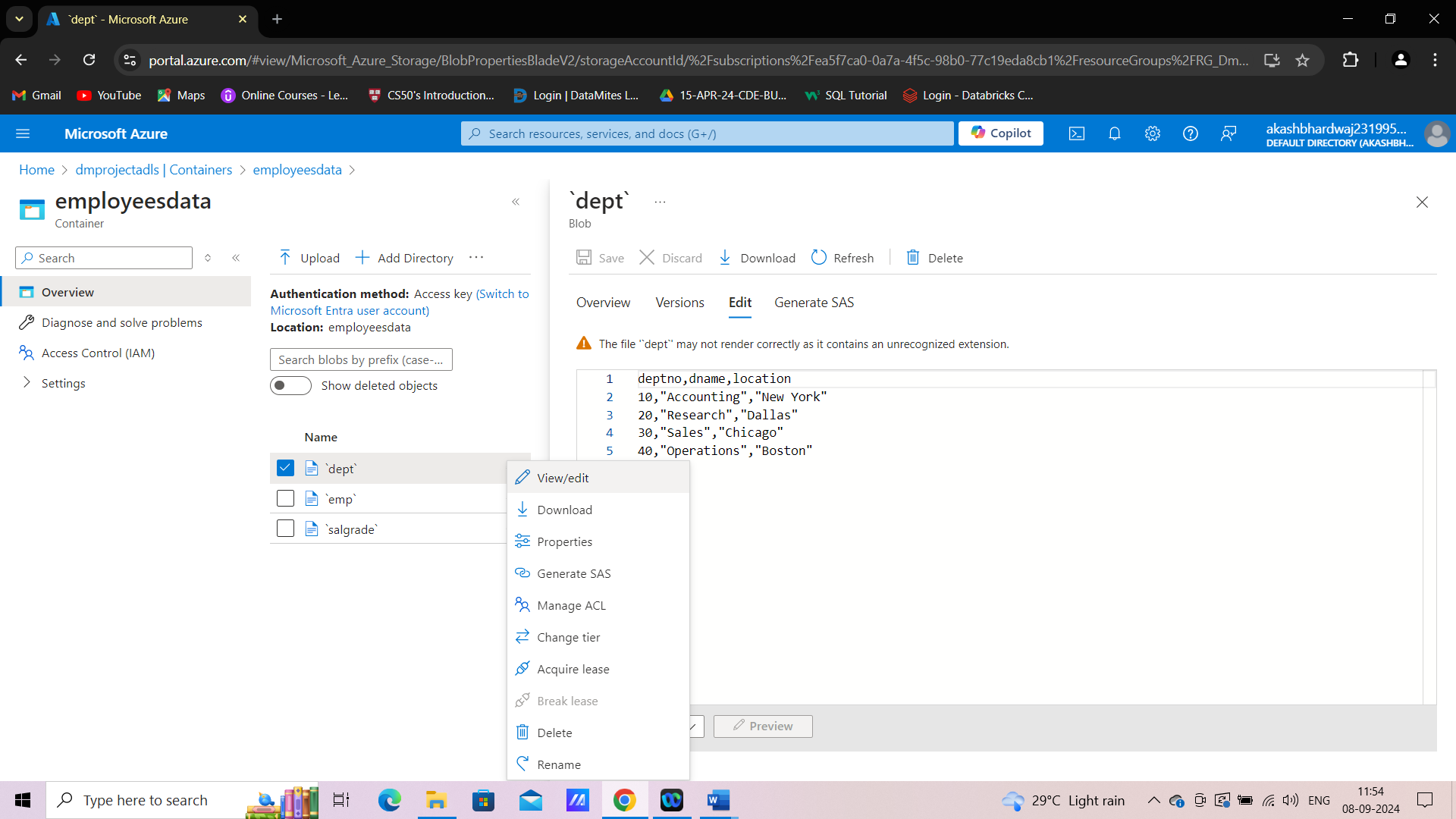




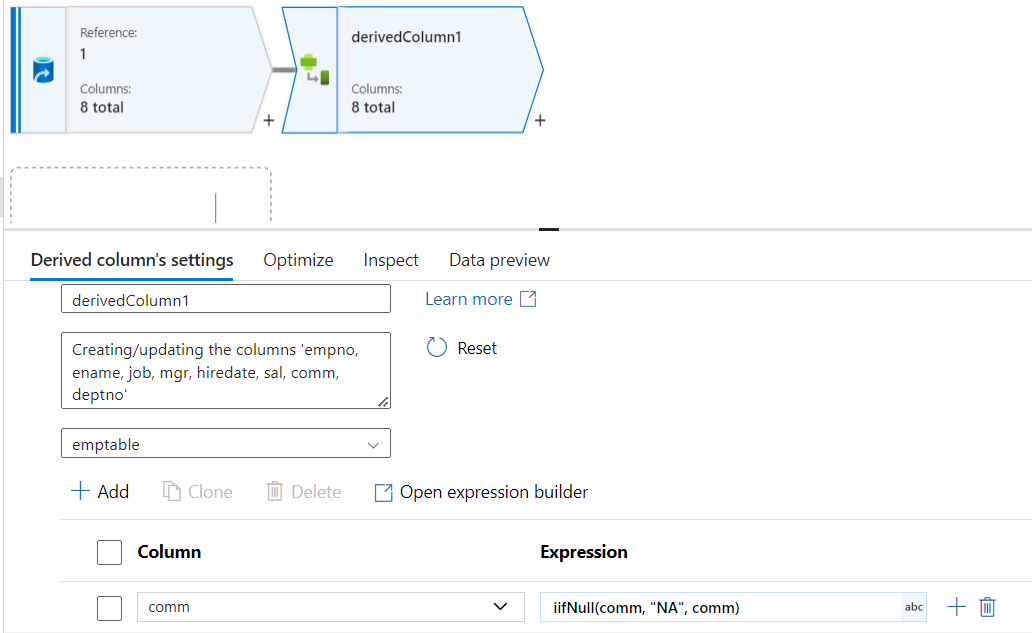


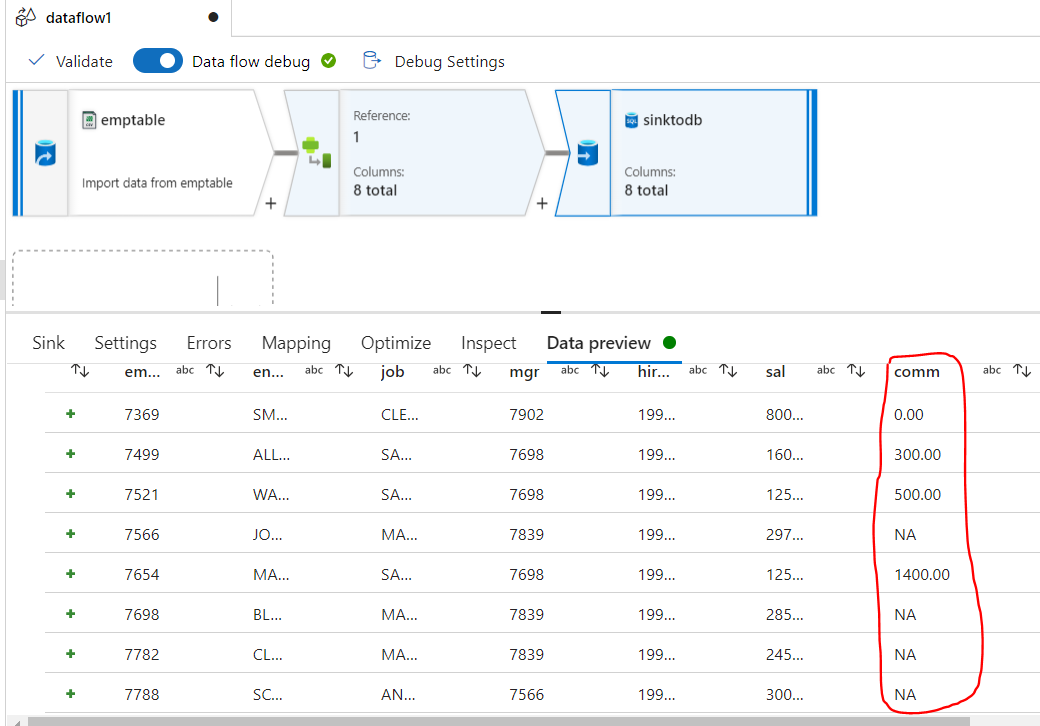
**Transformation:**

1. Once these tables were moved into the ADLS containers, I removed the rows with NULL values only using the “Edit” option as shown in below screenshot:



1. For replacing the NULL values in “comm” column of the employees table, I created a dataflow in ADF and used the Derived Column Transformation as shown below:





**Script:**

source(output(

        empno as string,

        ename as string,

        job as string,

        mgr as string,

        hiredate as string,

        sal as string,

        comm as string,

        deptno as string

    ),

    allowSchemaDrift: true,

    validateSchema: false,

    ignoreNoFilesFound: false) ~> emptable

emptable derive(comm = iifNull(comm, "NA", comm)) ~> derivedColumn1

derivedColumn1 sink(allowSchemaDrift: true,

    validateSchema: false,

    deletable:false,

    insertable:true,

    updateable:false,

    upsertable:false,

    format: 'table',

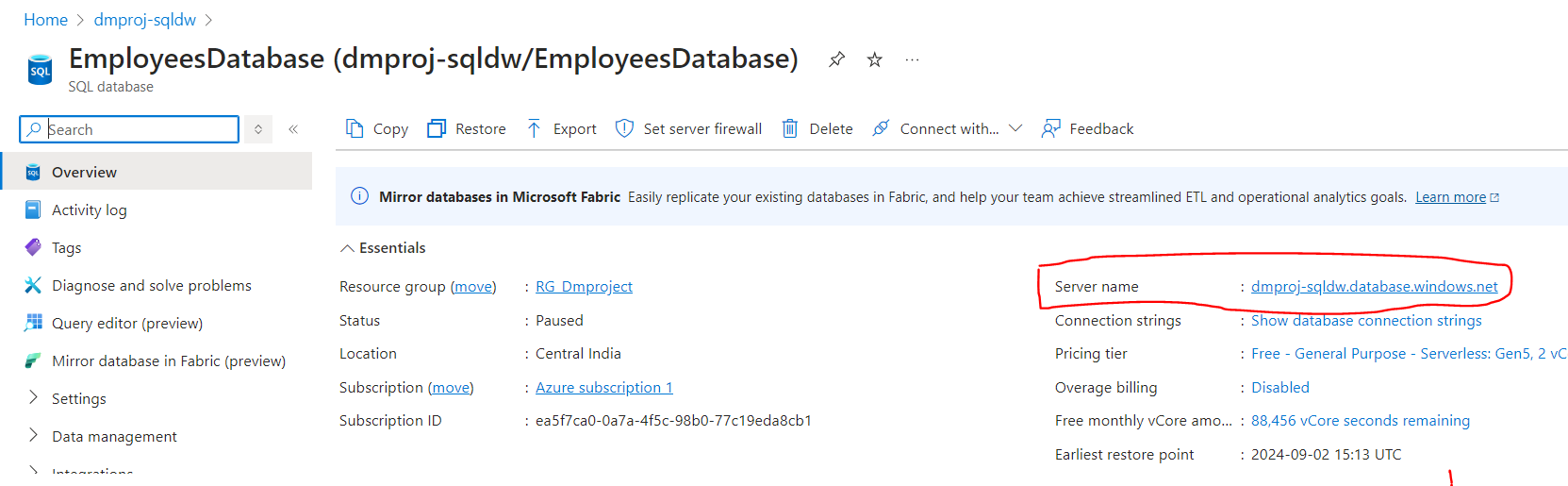
    skipDuplicateMapInputs: true,

    skipDuplicateMapOutputs: true,

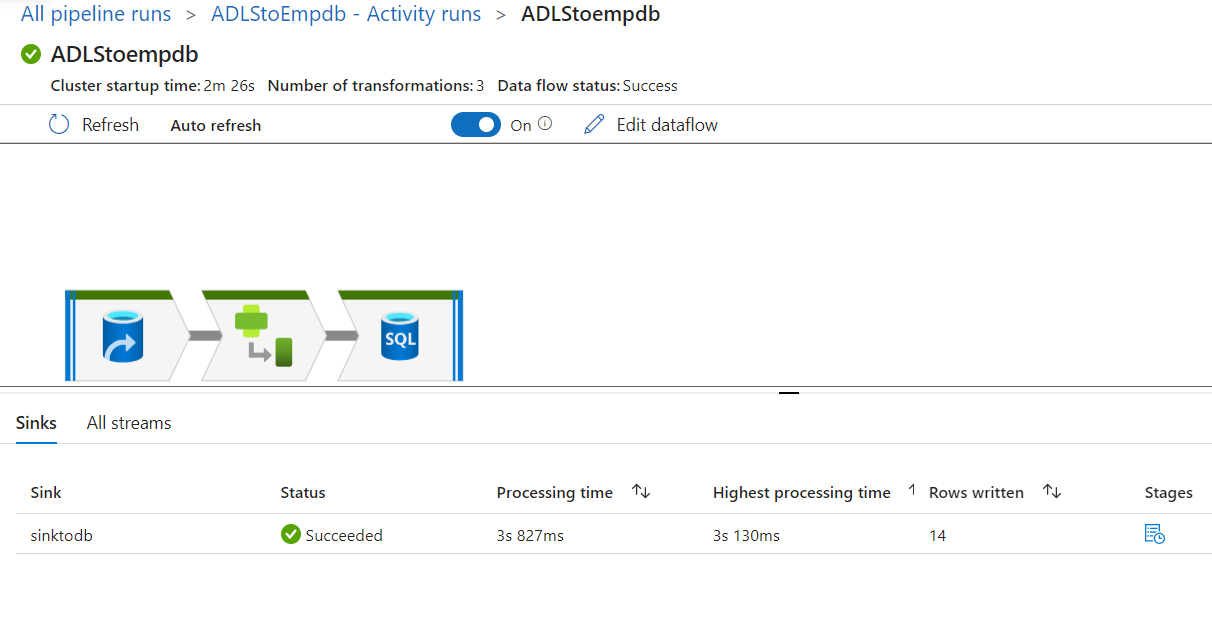
    errorHandlingOption: 'stopOnFirstError') ~> sinktodb

**Loading the data into data warehouse:**

1. Created Azure SQL database which will hold the enterprise data warehouse in this scenario. I created two databases named “**Employees database**” and “**Sales database**” which are hosted on the same Azure SQL server named “**dmproj-sqldw**”.

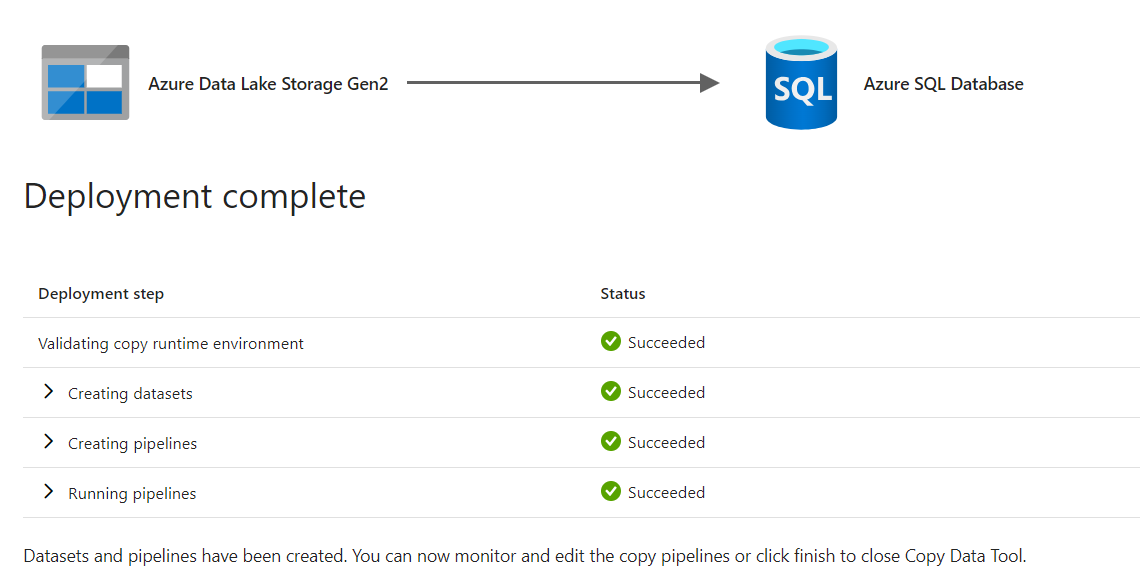


1. After the transformation applied in step 7, the employees’ table was moved to the **Employees database** by setting it as the sink in the dataflow in ADF.

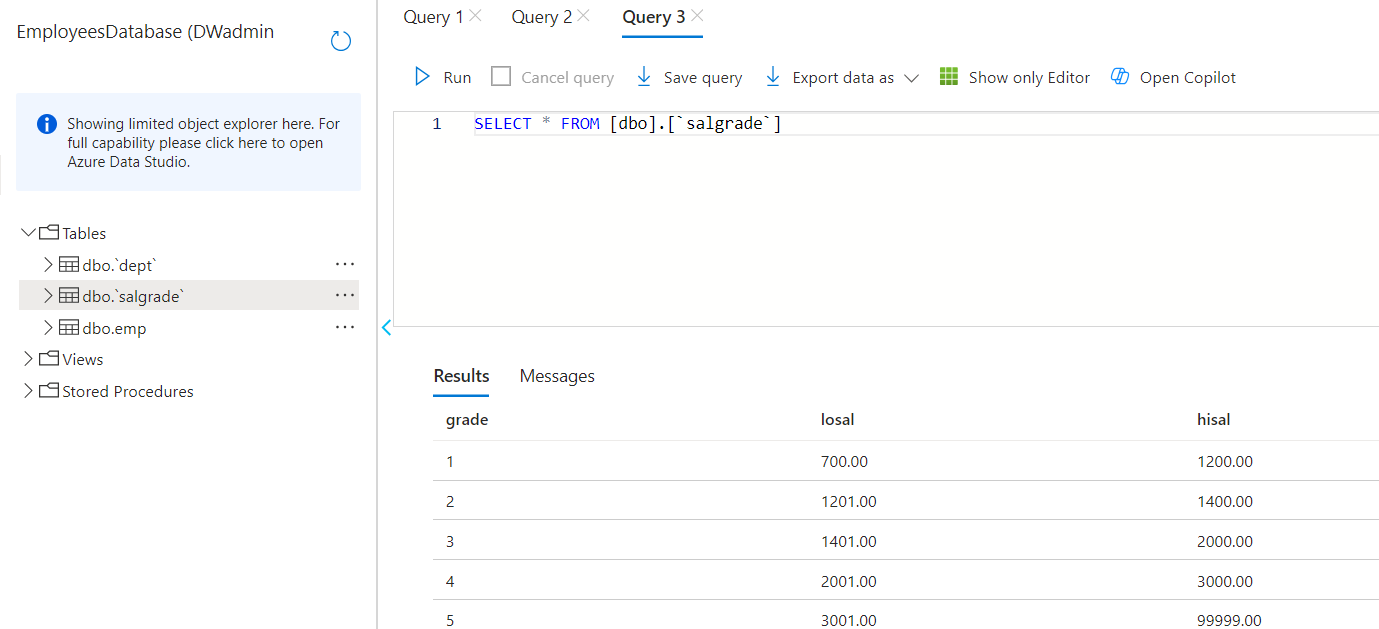


1. Copied the “dept” table to employees database using the ADF copy tool:

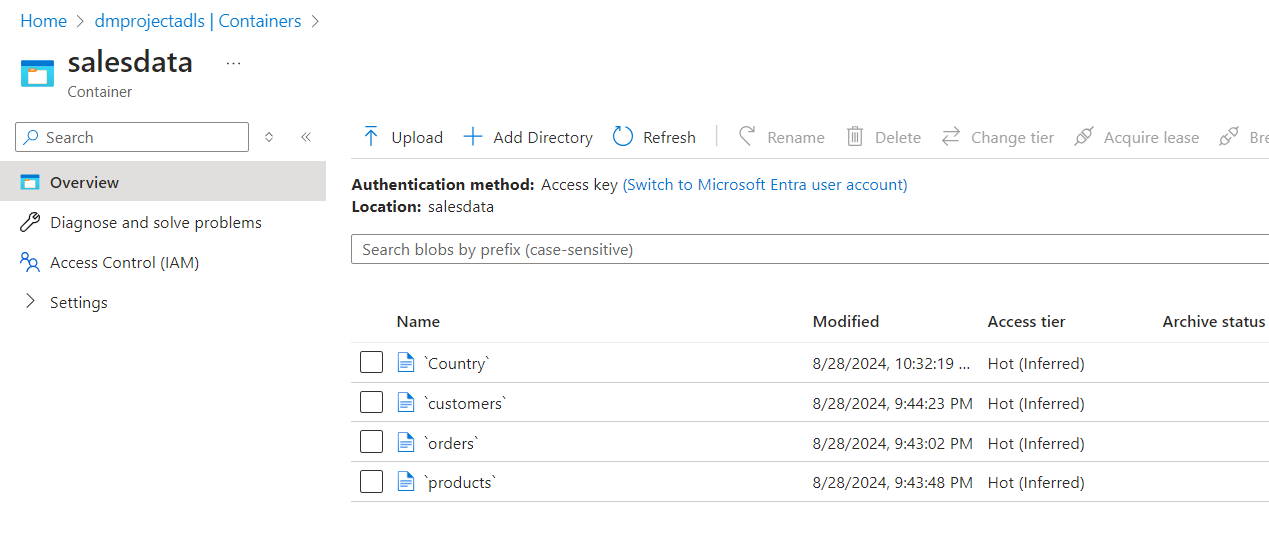


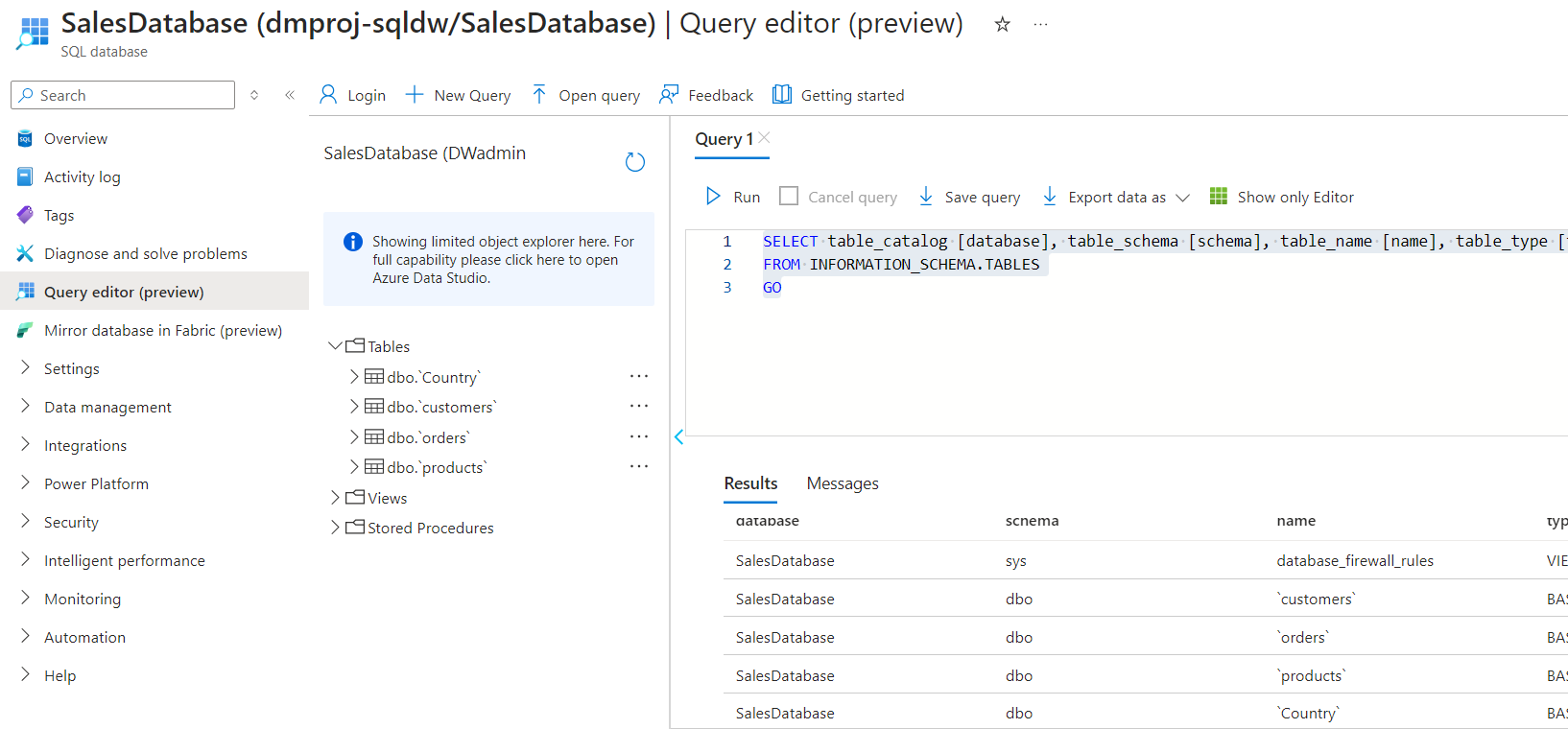


1. Copied the “salgrade” table following same steps:

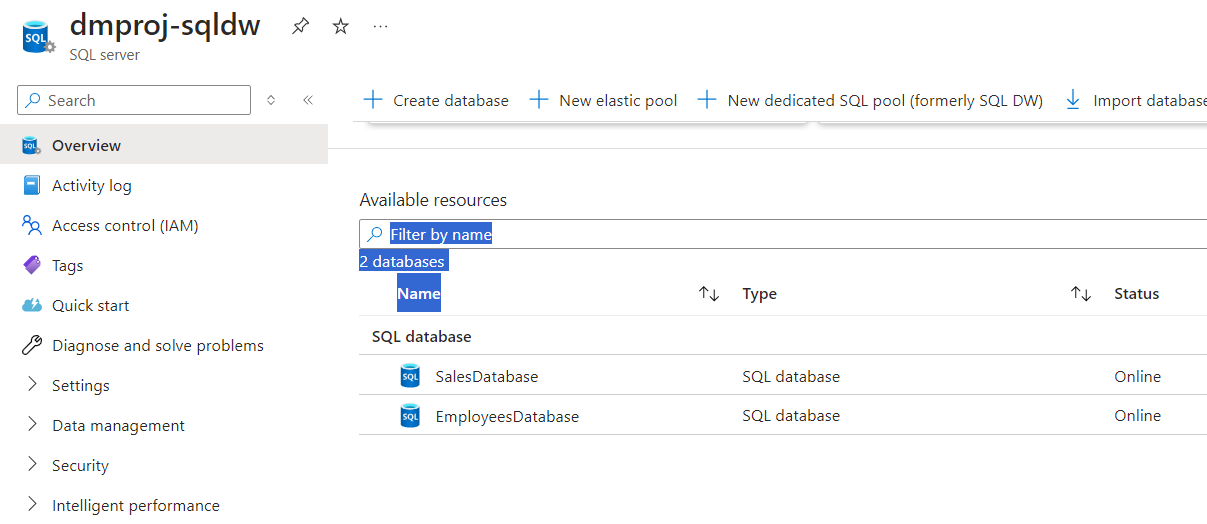


1. The same steps were followed to copy the tables inside “salesdata” container in ADLS storage to “Sales database”.





The SQL server hosts both the databases which can be accessed via Azure Data Studio. This will be the data warehouse for this project.



**Queries:**

1. Top 3 customers belong to which city.

Ans. The top 3 customers belong to **Los Angeles, Fort Lauderdale and Henderson.**

>> Added primary key in customers table:

ALTER TABLE [dbo].[`customers`]

ALTER COLUMN customer\_id nvarchar(450) NOT NULL;

ALTER TABLE [dbo].[`customers`]

ADD PRIMARY KEY (customer\_id);

>> Foreign key created on the order table:

ALTER TABLE [dbo].[`orders`]

ALTER COLUMN customer\_id nvarchar(450) NOT NULL;

ALTER TABLE [dbo].[`orders`]

ADD FOREIGN KEY (customer\_id)

REFERENCES [dbo].[`customers`] (customer\_id)

select top 3 SUM(sales\*quantity) AS totalamount, customer\_name, city from [dbo].[`orders`] join [dbo].[`customers`] on [dbo].[`orders`].customer\_id=[dbo].[`customers`].customer\_id group by city, customer\_name order by totalamount DESC;

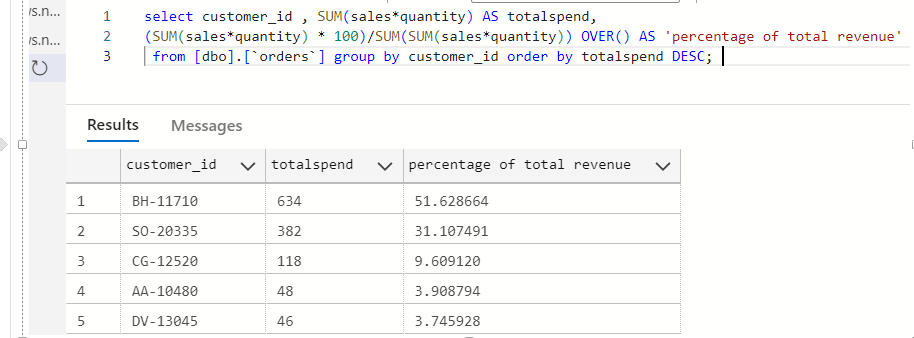


1. What percentage of top customers are responsible for 50% of the revenue generated.

Ans. select customer\_id , SUM(sales\*quantity) AS totalspend,

(SUM(sales\*quantity) \* 100)/SUM(SUM(sales\*quantity)) OVER() AS 'percentage of total revenue'

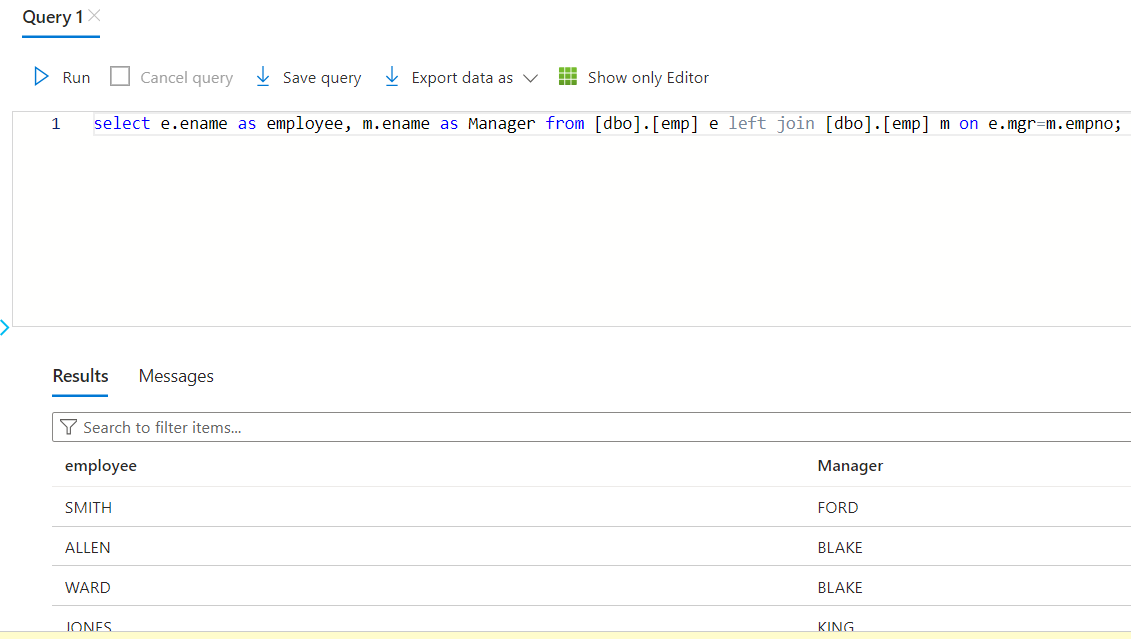
from [dbo].[`orders`] group by customer\_id order by totalspend DESC;



So, 20% of top customers are responsible for 50% of the revenue generated.

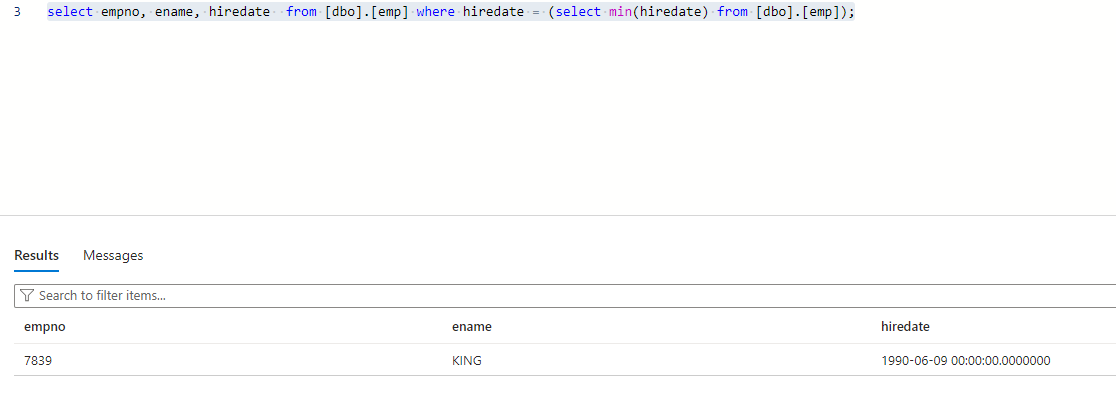
1. Explain the hierarchy of the employee in the emp table using SQL query.

Ans. select e.ename as employee, m.ename as Manager from [dbo].[emp] e left join [dbo].[emp] m on e.mgr=m.empno;



1. Who was the first employee to get hired in the organization?

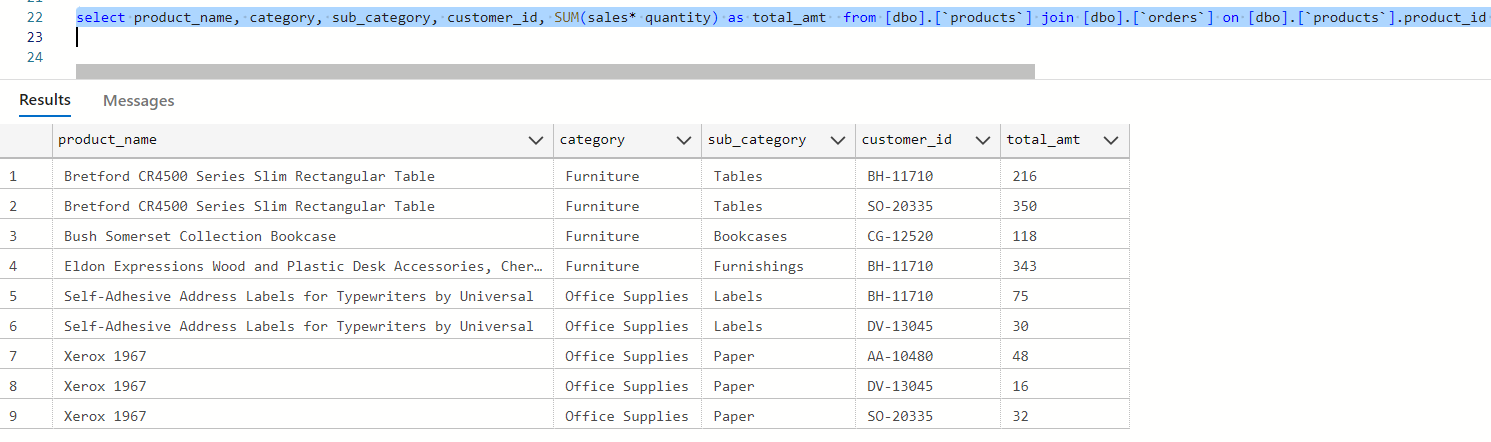
Ans. select empno, ename, hiredate from [dbo].[emp] where hiredate = (select min(hiredate) from [dbo].[emp]);



King was the first employee to get hired in the organization.

1. If the organization wants to offer discounts on which products they can offer, and to which customers?

Ans. select product\_name, category, sub\_category, customer\_id, SUM(sales\* quantity) as total\_amt  from [dbo].[`products`] join [dbo].[`orders`] on [dbo].[`products`].product\_id = [dbo].[`orders`].product\_id group by product\_name, category, sub\_category, customer\_id;



Performed a left join after this to find out the products in similar categories which can be discounted for potential customers:

>> Discount can be offered on Chromcraft Rectangular Conference Tables from furniture category to customers with ID BH-11710 and SO-20335.

 >> Also, binders can be discounted for customers buying paper, i.e., for customers with IDs AA-10480, DV-13045 and SO-20335.

**Reference:**

[**https://learn.microsoft.com/en-us/azure/architecture/example-scenario/data/small-medium-data-warehouse**](https://learn.microsoft.com/en-us/azure/architecture/example-scenario/data/small-medium-data-warehouse)