DRAFT VERSION

Empowering Manufacturing Industries with Industrial Data Lake solution on Azure cloud

Essential Azure Software Components and Associated tasks

A Technical Project Done

By

V. Mageshwaran

Dataever Consulting

Table of Contents

1. Introduction

Manufacturing industries face many data management challenges namely Integration, Quality, accuracy, Volume, Security, Storage and Management, Compliance and Standards, Governance and Change Management.

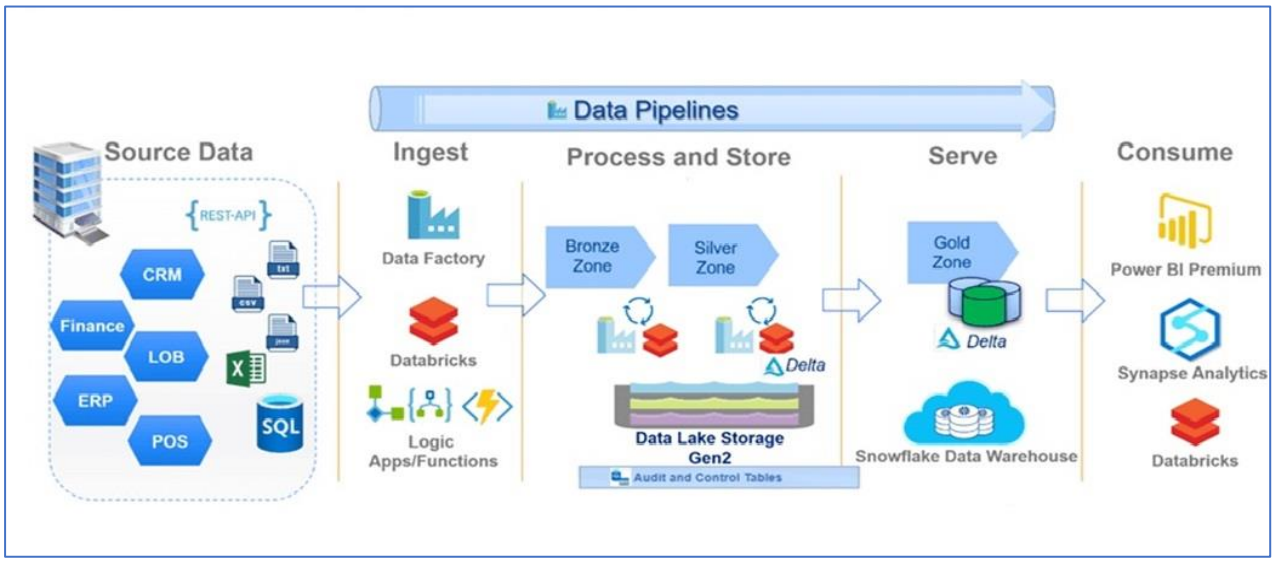
Few select Popular manufacturing industries data management trends are Industrial Internet of Things(IIoT), Big Data Analytics, Artificial Intelligence and Machine Learning, Digital Twins, Augmented Reality (AR) and Virtual Reality (VR), Sustainability and Energy Management

2. Data lake solution

Data lakes are distributed data stores that can hold very large volumes of diverse data. They can be used to store different types of data such as structured SQL, semi-structured, unstructured email, images, streaming sensor data, and so on.

A data lake solution usually comprises a storage layer, a compute layer, and a serving layer. The compute layers on cloud include **Extract, Transform, Load (ETL); Batch; or Stream** processing.

The following image shows a generic data lake architecture



2.1 Data Lake zones

A data lake can be broadly segregated into three zones where different stages of the processing take place, outlined as follows:

1. **Landing Zone or Raw Zone:** This is where the raw data is ingested from different input sources. In this report, we may refer this as bronze zone.
2. **Transformation Zone:** This is where the batch or stream processing happens. The raw data gets converted into a more structured and **business intelligence (BI)**-friendly format. In this report, we may refer this as silver zone.

3. **Serving Zone:** This is where the curated data that can be used to generate insights and reports are stored and served to BI tools. The data in this zone usually adheres to well-defined schemas. We may refer this as gold zone.

2.2 Data Lake use case scenarios

1. Data that is too big to be stored in traditional structured storage systems like data warehouse or SQL databases
2. Raw data that needs to be stored for further processing, such as an ETL system or a batch processing system
3. Storing continuous streaming data such as **Internet of Things (IoT)** data, sensor data, tweets, and so on for low latency, high throughput streaming scenarios
4. Storing processed data for advanced tasks such as ad hoc querying**, Deep learning, machine learning (ML)**, and data exploration.

3. Industrial Data Lake

Industrial Data Lake is a custom data lake solution meant for industries. Manufacturing industries produce a vast voluminous data regularly. Until recently, this data mostly went unused, but today's advanced AI and ML analytics can analysis to generate valuable real-time insights, support efficiency improvements, process improvements, enable predictive maintenance, data driven decisions, reduce unplanned downtime and more.

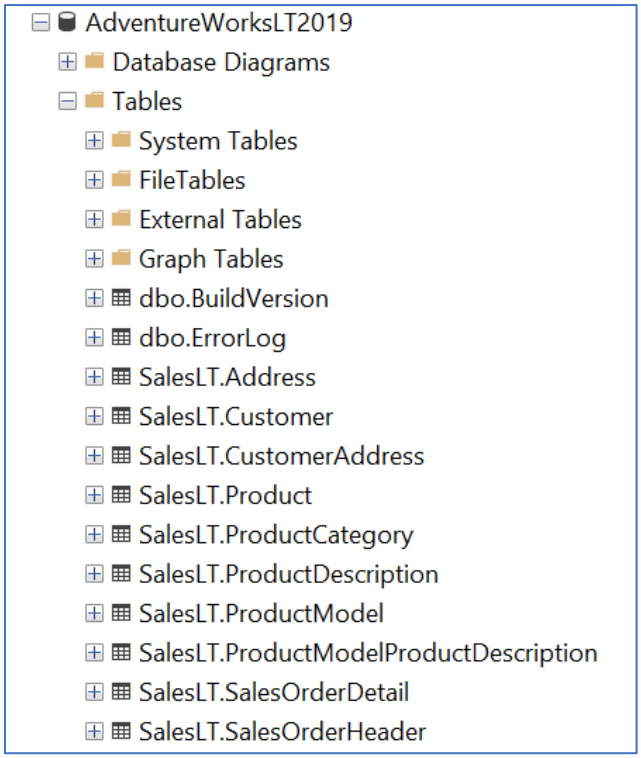
4. Our Goal

The primary goal of this project is to build industrial end-to-end. Data Lakes on Azure cloud. For this purpose, use an opensource dataset. In the process of building a data lake on Azure, we also want to create an implementation guide. Such a guide can be used in future for similar application involving different datasets. This will save time and effort.

5. Data sets

For our work we have taken publicly available AdventureWorksLT2017 and AdventureWorksLT2019. This can be downloaded from

https://learn.microsoft.com/en-us/sql/samples/adventureworks-install-configure?view=sql-server-ver16&tabs=ssms



This dataset has two schemas, for example dbo. A sample dataset section in shown below. But we use for our project only SalesLT schema.

	SalesOrderID	SalesOrderDetailID	OrderQty	ProductID	UnitPrice	UnitPriceDiscount	LineTotal	rowguid	ModifiedDate
1	71774	110562	1	836	356.898	0.00	356.898000	E3A1994C-7A68-4CE8-96A3-77FDD3BBD730	2008-06-01 00:00:00.000
2	71774	110563	1	822	356.898	0.00	356.898000	5C77F557-FDB6-43BA-90B9-9A7AEC55CA32	2008-06-01 00:00:00.000
3	71776	110567	1	907	63.90	0.00	63.900000	6DBFE398-D15D-425E-AA58-88178FE360E5	2008-06-01 00:00:00.000
4	71780	110616	4	905	218.454	0.00	873.816000	377246C9-4483-48ED-A5B9-E56F005364E0	2008-06-01 00:00:00.000
5	71780	110617	2	983	461.694	0.00	923.388000	43A54BCD-536D-4A1B-8E69-24D083507A14	2008-06-01 00:00:00.000
6	71780	110618	6	988	112.998	0.40	406.792800	12706FAB-F3A2-48C6-B7C7-1CCDE4081F18	2008-06-01 00:00:00.000
7	71780	110619	2	748	818.70	0.00	1637.400000	B12F0D3B-5B4E-4F1F-B2F0-F7CDE99DD826	2008-06-01 00:00:00.000
8	71780	110620	1	990	323.994	0.00	323.994000	F117A449-039D-44B8-A4B2-B12001DACC01	2008-06-01 00:00:00.000
9	71780	110621	1	926	149.874	0.00	149.874000	92E5052B-72D0-4C91-9A8C-42591803667E	2008-06-01 00:00:00.000
10	71780	110622	1	743	809.76	0.00	809.760000	8BD33BED-C4F6-4D44-84FB-A7D04AFCD794	2008-06-01 00:00:00.000
11	71780	110623	4	782	1376.994	0.00	5507.976000	686999FB-42E6-4D00-9A14-83FFA86833E3	2008-06-01 00:00:00.000
12	71780	110624	2	918	158.43	0.00	316.860000	82940B03-C70B-4183-8660-6B3418908429	2008-06-01 00:00:00.000
13	71780	110625	4	780	1391.994	0.00	5567.976000	644B0CD6-B2C3-4E4D-AB43-091C2EF6C829	2008-06-01 00:00:00.000
14	71780	110626	1	937	48.594	0.00	48.594000	7F5FEB17-8EF4-4236-9F1C-15046D9638F0	2008-06-01 00:00:00.000
15	71780	110627	6	867	41.994	0.00	251.964000	AC78838D-B503-41A5-9791-480E528F028C	2008-06-01 00:00:00.000
16	71780	110628	1	985	112.998	0.40	67.798800	2C10A282-A13D-442A-8F45-F4D6B23A7D9C	2008-06-01 00:00:00.000
17	71780	110629	2	989	323.994	0.00	647.988000	654FB79E-70DF-4B92-9832-9FA67013215B	2008-06-01 00:00:00.000
18	71780	110630	3	991	323.994	0.00	971.982000	3D6CA7AB-055E-4536-8940-76234CC9BCDE	2008-06-01 00:00:00.000
19	71780	110631	1	992	323.994	0.00	323.994000	560FEEE1-DD54-4C34-ABB1-4F8841D0AA41	2008-06-01 00:00:00.000
20	71780	110632	2	993	323.994	0.00	647.988000	19570052-4023-4658-BC56-DC5C619BD00E	2008-06-01 00:00:00.000
21	71780	110633	2	984	112.998	0.40	135.597600	27562675-F8C3-4A38-BD9E-B366B83E5204	2008-06-01 00:00:00.000
22	71780	110634	3	986	112.998	0.40	203.396400	E193CE39-EF33-4969-87B1-468D2F7B48AD	2008-06-01 00:00:00.000
23	71780	110635	3	987	112.998	0.40	203.396400	E38E076F-5072-437A-A771-ADA53B5AB803	2008-06-01 00:00:00.000
24	71780	110636	2	981	461.694	0.00	923.388000	26C00B7D-6E19-4FBF-B9F1-23C2609E8893	2008-06-01 00:00:00.000
25	71780	110637	3	982	461.694	0.00	1385.082000	6666A81B-90A1-4204-A39E-9F660CA43E5F	2008-06-01 00:00:00.000

6. Pre requisites

1. Azure cloud A/C
2. A Local system on prem with
    - Power BI Desktop
    - MS SQL
    - SSMS

7. Data Lake Solution Architecture Overview

The well know Extract-Transform-Load (ETL) principle forms the basis for the solution architecture. In addition, data preparation, cloud configuration and reporting are also part of further strengthens of the solution architecture.

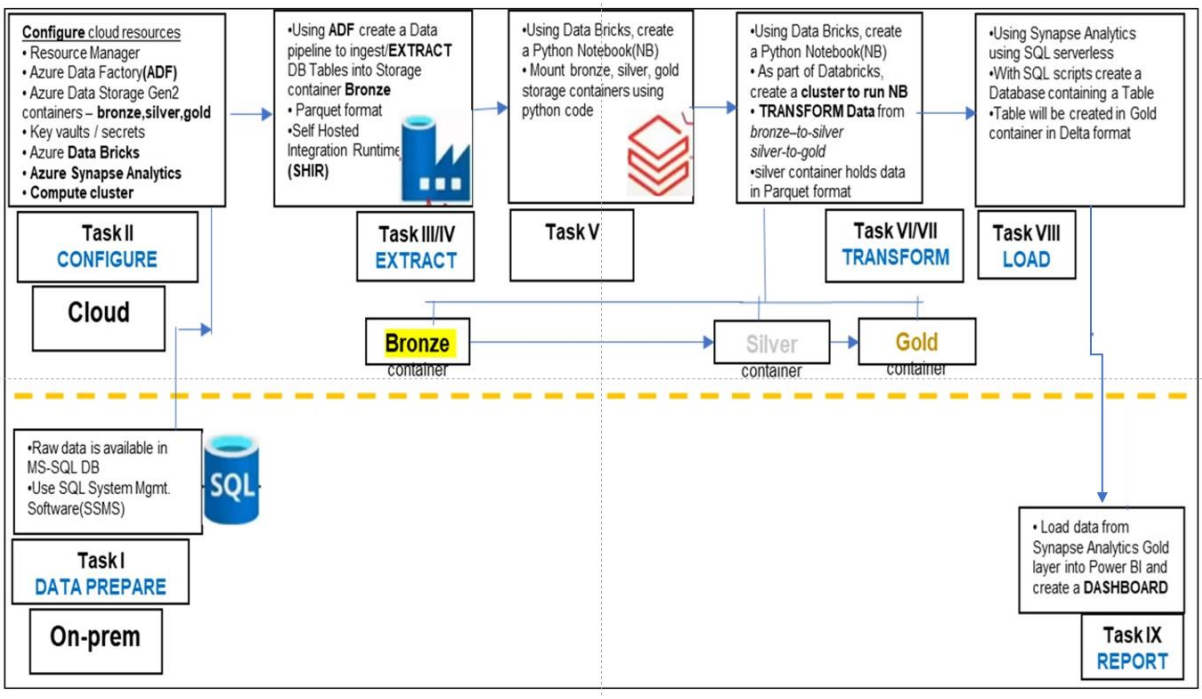


Figure 1: Solution architecture diagram

8. Software building blocks

8.1 Azure storage account

An Azure storage account contains all of your Azure Storage data objects: blobs, files, queues, and tables. The storage account provides a unique namespace for your Azure Storage data that is accessible from anywhere in the world over HTTP or HTTPS

8.2 Azure cloud A/C

The Azure cloud platform is more than 200 products and cloud services designed to help to build new technology solutions. Build, run, and manage applications across multiple clouds, on-premises, and at the edge, with the tools and frameworks of your choice.

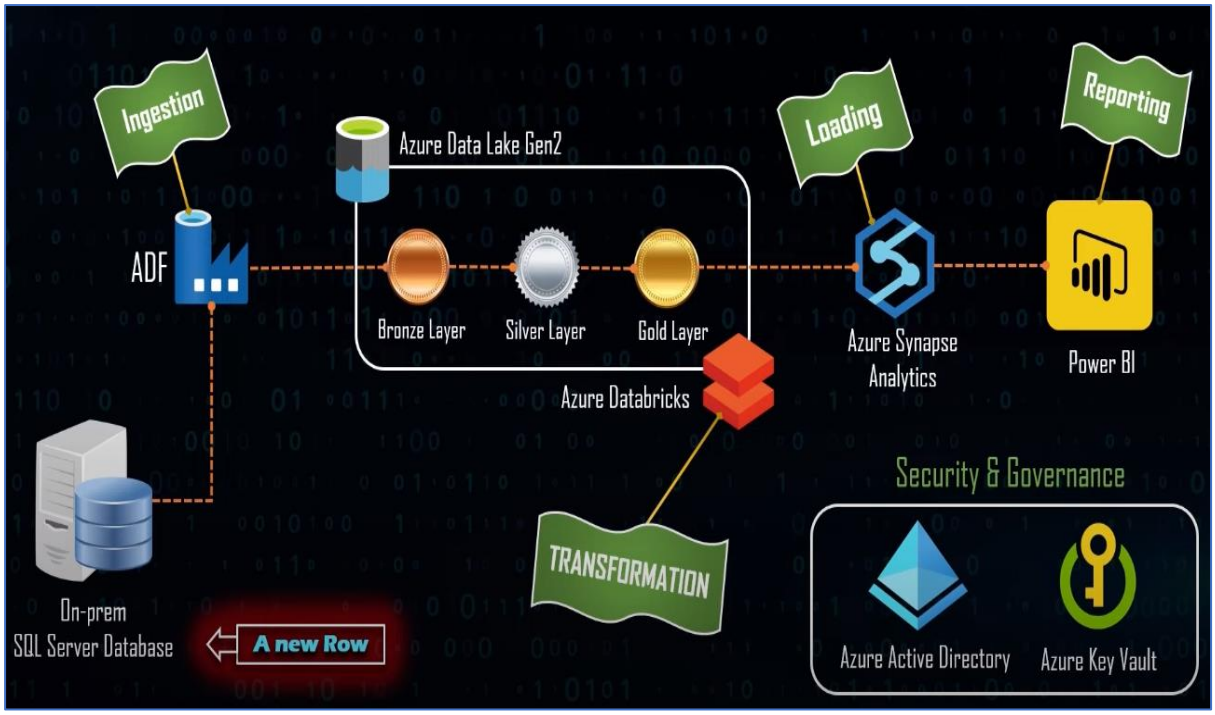


Figure 2: Azure data transformation of tool components

8.3 Azure Data Lake Storage Gen 2 (ADSL Gen2)

Azure Data Lake Storage Gen2 is a set of capabilities dedicated to big data analytics, built on Azure Blob Storage. Data Lake Storage Gen2 converges the capabilities of Azure Data Lake Storage Gen1 with Azure Blob Storage. This storage helps to store file system type and SQL-type data. While configuring storage account on azure, one has to choose ‘Enable hierarchical name space ‘option. This will make the storage as ADSL Gen2.

8.4 Azure Data Factory

Azure Data Factory is Azure's cloud ETL service for scale-out serverless data integration and data transformation. The current version is V2

8.5 Data Factory Pipeline

A pipeline is a logical grouping of activities that together perform a task. For example, a pipeline could contain a set of activities that ingest and clean log data, and then kick off a mapping data flow to analyse the log data. The pipeline allows you to manage the activities as a set instead of each one individually. You deploy and schedule the pipeline instead of the activities independently.

The activities in a pipeline define actions to perform on your data. For example, you can use a copy activity to copy data from SQL Server to an Azure Blob Storage. Then, use a data flow activity or a Databricks Notebook activity to process and transform data from the blob storage to an Azure Synapse Analytics pool on top of business intelligence reporting solutions

8.6 Linked Service

Linked services are much like connection strings, which define the connection information needed for the service to connect to external resources. Think of it this way: the dataset represents the structure of the data within the linked data stores, and the linked service defines the connection to the data source. For example, an Azure Storage linked service links a storage account to the service. An Azure Blob dataset represents the blob container and the folder within that Azure Storage account that contains the input blobs to be processed.

Here is a sample scenario. To copy data from Blob storage to a SQL Database, you create two linked services: Azure Storage and Azure SQL Database. Then, create two datasets: Azure Blob dataset (which refers to the Azure Storage linked service) and Azure SQL Table dataset (which refers to the Azure SQL Database linked service). The Azure Storage and Azure SQL Database linked services contain connection strings that the service uses at runtime to connect to your Azure Storage and Azure SQL Database, respectively.

8.7 Azure Functions

Azure Function is a serverless compute service that enables user to run event-triggered code without having to provision or manage infrastructure. Being as a trigger-based service, it runs a script or piece of code in response to a variety of events.

8.8 Azure Key Vault

Azure Key Vault is a cloud service that provides a secure store for secrets. You can securely store keys, passwords, certificates, and other secrets. Azure key vaults may be created and managed through the Azure portal. In this quick start, you create a key vault, then use it to store a secret.

8.9 Azure SQL DB

Part of the Azure SQL family, Azure SQL is a fully managed relational database service built for the Azure cloud. Build your next app with the assistance of a fully managed SQL database with built-in AI capabilities, auto-scaling, and backups

8.10. Store secrets in Key Vault and access them

Key Vault provides secure storage of generic secrets, such as passwords and database connection strings. All secrets in Key Vault are stored encrypted. Key Vault encrypts secrets at rest with a hierarchy of encryption keys, with all keys in that hierarchy protected by modules that are FIPS 140-2 compliant.

8.11. SSMS

SQL Server Management Studio (SSMS) is an integrated environment for managing any SQL infrastructure. SSMS is a tool to write SQL queries, stored procedures, and basically play with structured data.

8.12 Azure Databricks

Azure Databricks is another Spark distribution that can provide limited analytics store capabilities via its in-memory stores and wide column store support. It also supports SQL-Like interfaces.

8.13 Azure Synapse Analytics

Synapse Analytics provides both SQL pools and spark pools. Serverless SQL pools can be used for ad hoc querying. Spark pools, on the other hand, can support analytical workload through their in-memory store and wide column store support. Both support SQL/SQL-like interface

8.14Features of Azure data bricks and azure synapse analytics

Feature	Databricks	Synapse Analytics
Overview	Unified data analytics platform powered by Apache Spark.	Integrated analytics service combining big data and data warehousing.
Core Technology	Apache Spark	SQL Data Warehouse, Apache Spark, Data Explorer
Data Storage	Delta Lake, Parquet, ORC, Avro, etc.	Azure Data Lake Storage, SQL pools, Cosmos DB
Integration	Integration with various data sources (Azure, AWS, GCP, on-premises)	Deep integration with Azure ecosystem (Data Factory, Power BI, Azure ML)
Data Lake Integration	Delta Lake integration for efficient data lake management	Direct integration with Azure Data Lake Storage (ADLS)

9. Data insights using Power-BI

After processing the dataset using ETL method, it is very important to present the data insights using different powerful visualization techniques. Power bi software tool is very helpful in this process

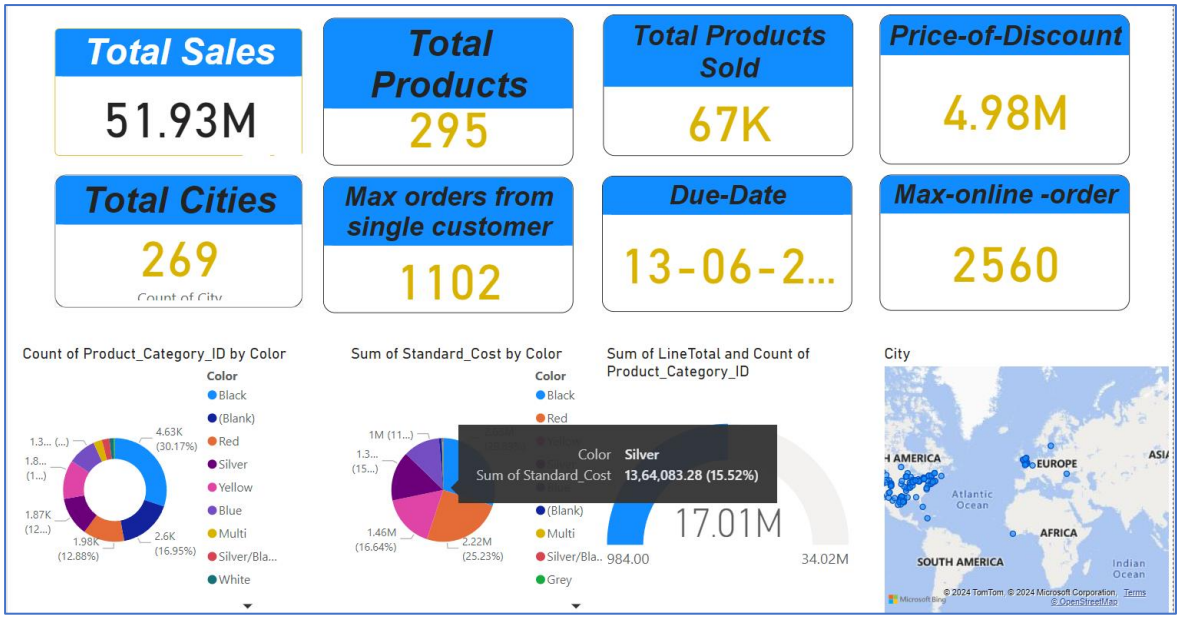


Figure 3: If use of many charts as create a one visualzaition of Power BI. Total Sales of Products,dute Date , Max-of-online orders,etc…

Overall Sales Performance: The total sales value is significant, showing robust sales performance.

Product Diversity: A large number of products (295) have been sold, indicating a diverse product range.

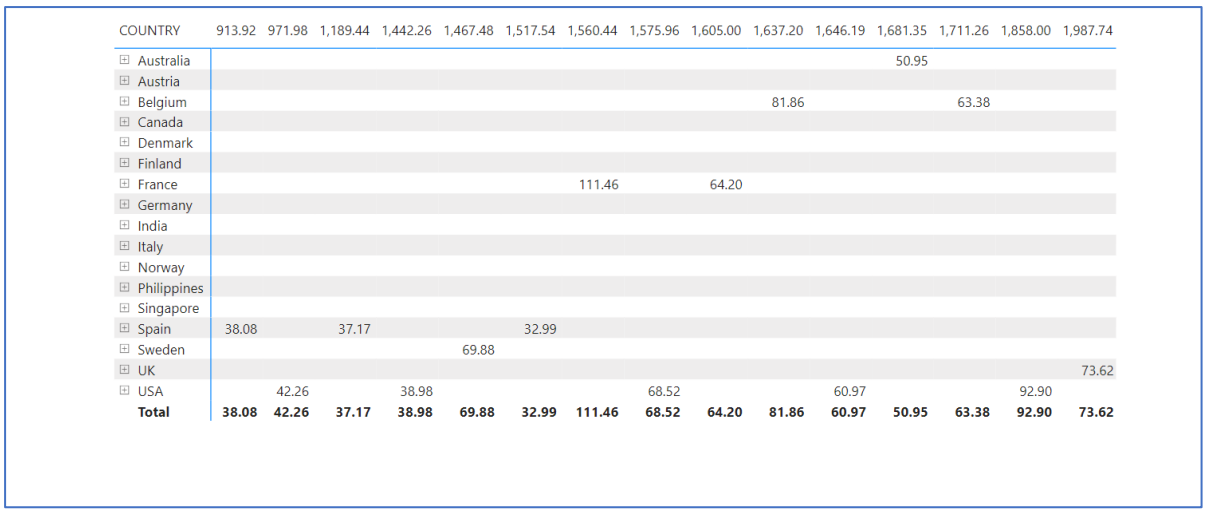
Discount Impact: The total discount value (4.98M) shows the impact of discounts on sales.

Geographical Reach: Sales have occurred in 269 different cities, showing a wide geographical reach.

Customer Engagement: A single customer placing 1102 orders indicates high customer engagement and loyalty.

Colour Analysis: The distribution of products and costs by color provides insights into popular product variations.

Order Trends: The pie charts and map visualization help in understanding the trends and geographical distribution of sales.



COUNTRY	913.92	971.98	1,189.44	1,442.26	1,467.48	1,517.54	1,560.44	1,575.96	1,605.00	1,637.20	1,646.19	1,681.35	1,711.26	1,858.00	1,987.74
Australia												50.95			
Austria															
Belgium										81.86			63.38		
Canada															
Denmark															
Finland															
France							111.46		64.20						
Germany															
India															
Italy															
Norway															
Philippines															
Singapore															
Spain	38.08		37.17			32.99									
Sweden					69.88										
UK															73.62
USA		42.26		38.98				68.52			60.97			92.90	
Total	**38.08**	**42.26**	**37.17**	**38.98**	**69.88**	**32.99**	**111.46**	**68.52**	**64.20**	**81.86**	**60.97**	**50.95**	**63.38**	**92.90**	**73.62**

*Figure 4: To creat table chart**.** If using PowerBI employe Data table as number of country,number of citys. The columns from left to right seem to represent numerical values associated with each country. These could be financial figures, statistics, or other measurable data*

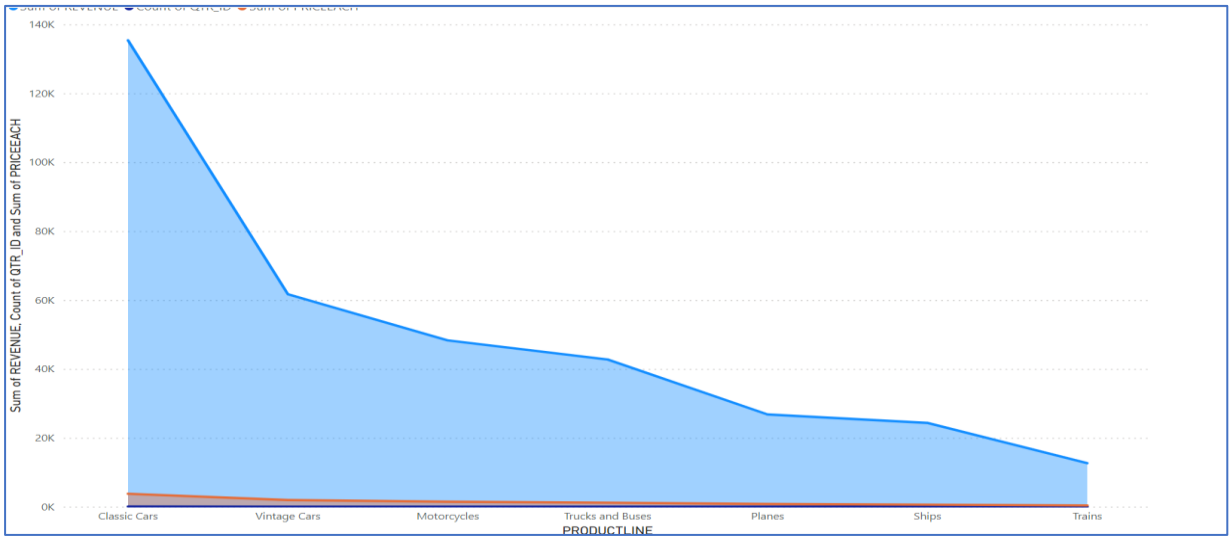


*Figure 5**:** To create the area chart using Power BI. To use the area chart as sum of revenue,product of quentity-id,**etc**…*

- *The product lines listed from left to right are: Classic Cars, Vintage Cars, Motorcycles, Trucks and Buses, Planes, Ships, and Trains*
- *This axis shows the sum of revenue, which is a measure of the total income generated from sales for each product line.*
- *Data Series (Blue Area)::The blue area represents the sum of revenue for each product line.*
- *The height of the blue area indicates the revenue amount for each category.*
- *Classic Cars have the highest revenue, while Trains have the lowest.*

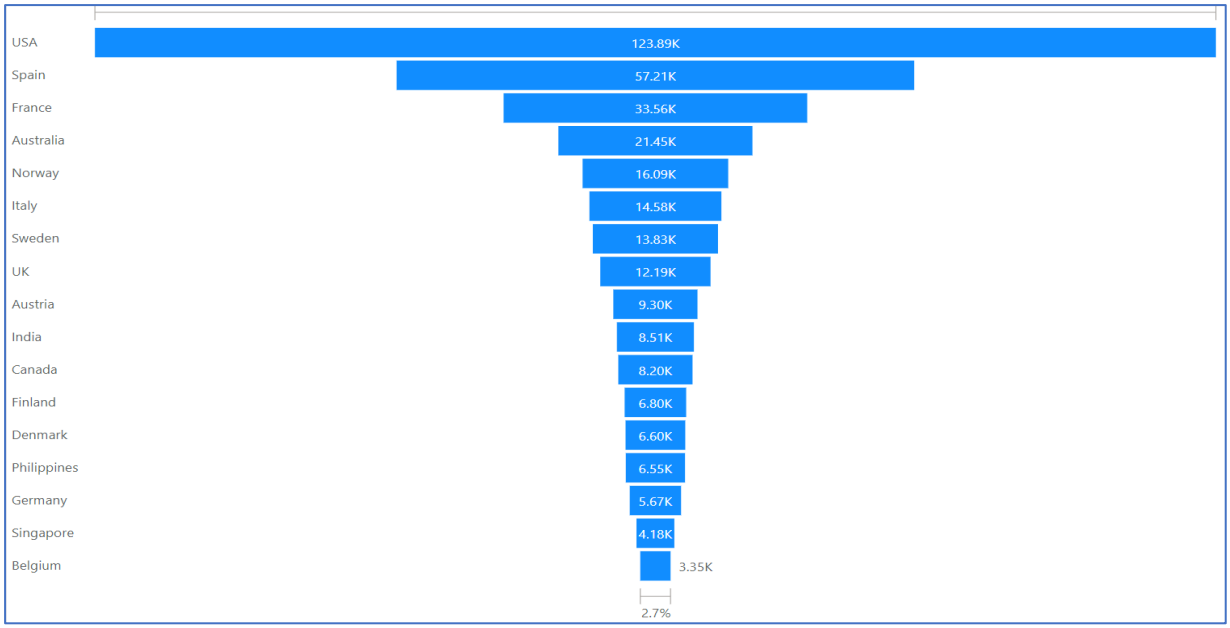


Figure 6: Funnel as used to country of profit. A funnel as be most highest country of Revenue in the world

- *The length of each bar visually represents the value associated with each country.The USA has the longest bar, indicating the highest value, while Belgium hastheshortest bar, indicating the lowest value among the listed countries.*

- *At the bottom of the chart, there is a reference line marked "2.7%". This might indicate a threshold, average, or a percentage of the total value represented by the countries*

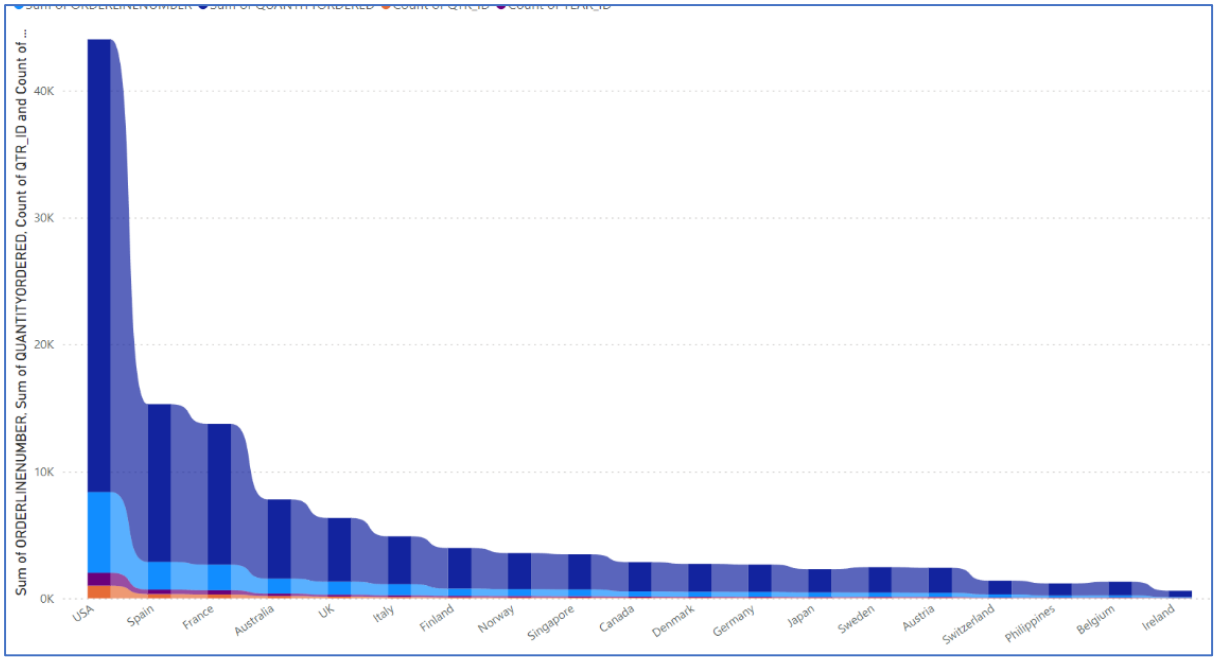


Figure 7: Ribbon chart as be use to sum of orders,count of this year,etc..

- *The USA has a particularly high sum of ORDERLINENUMBER and QUANTITYORDERED, suggesting a large volume of orders.*
- *Spain, France, and Australia also show relatively high values but with a noticeable drop compared to the USA.*
- *Other countries like the UK, Italy, and Finland have moderate values.*
- *Smaller contributions from countries like Ireland, Belgium, and the Philippines indicate lower activity or volume in these metrics.*

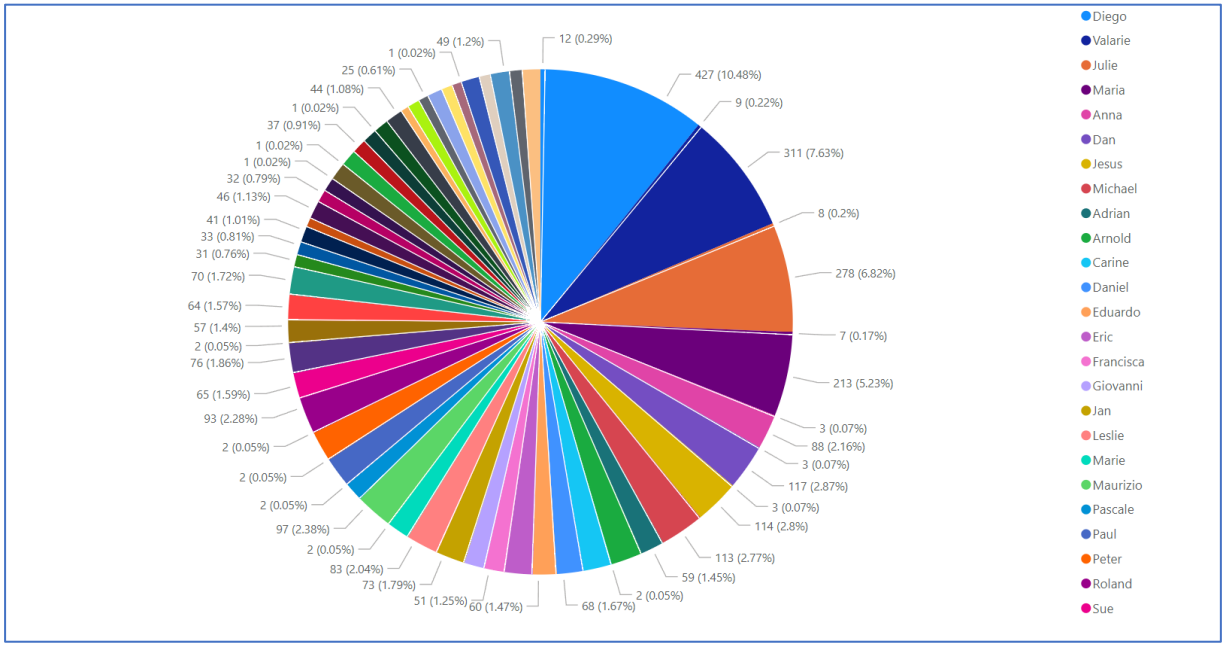


Figure 8: Pie chart use sum of quantity orders,sales of orders,etc…

- *Categories with slightly lower percentages include Carine (1.67%), Daniel (1.45%), Eduardo (1.26%), Eric (1.08%), Francisca (0.91%), Giovanni (0.79%), Jan (0.76%), Leslie (0.57%), Marie (0.56%), Maurizio (0.45%), Pascale (0.41%), Paul (0.35%), Peter (0.32%), Roland (0.29%), and Sue (0.29%).*
- *Minor Categories:There are several categories with very small percentages, including those with less than 1% share, such as Arnold (0.17%), Carine (0.17%), and other minor contributions.*
- *The chart highlights that Diego, Valarie, Julie, and Maria have the most significant shares in the distribution, making them key players or categories of interest.*

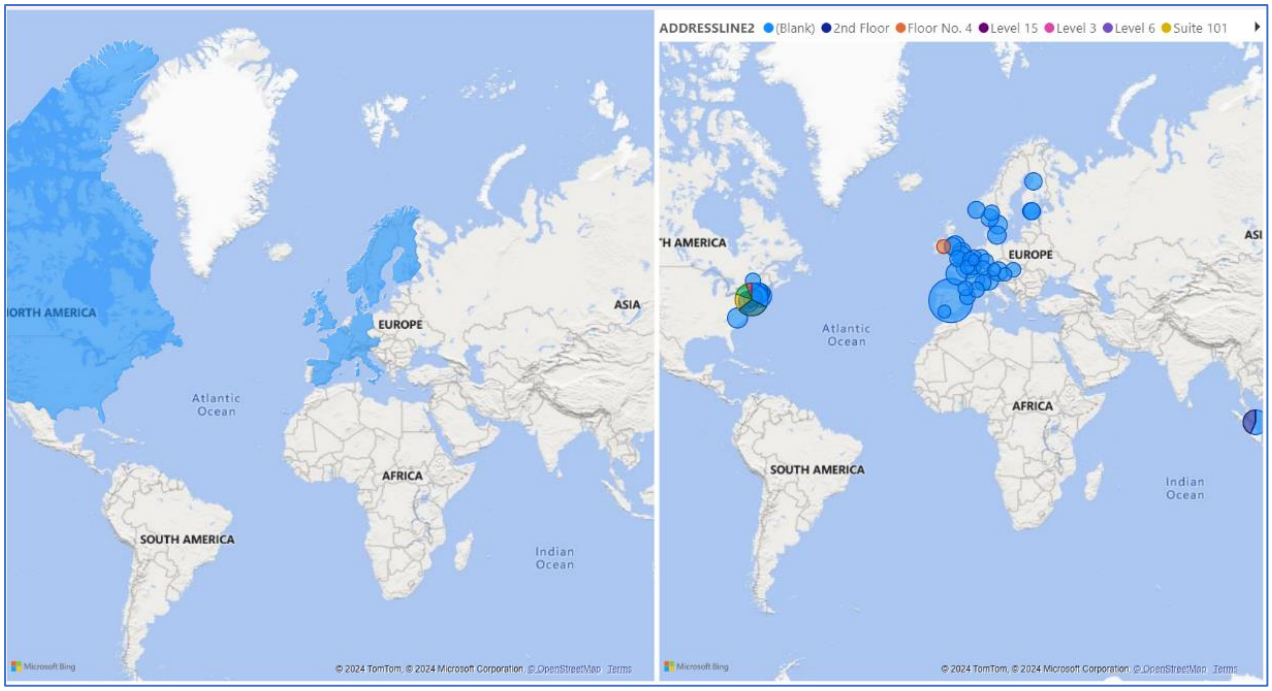


*Figure 9: Map showing as how many countries are purcheses of product**.** Map as showing a location of product sales of countries*

- *There is a high concentration of data points in Europe, especially in Western and Central Europe.*
- *There are also several data points in the eastern part of North America.*
- *The color-coded dots suggest varied types of address line information, potentially indicating different levels or floors in buildings.The high concentration in specific regions suggests that these areas have more detailed address information available or more activity.*
- *Provides an overview of regions with significant activity or data, highlighting North America and Europe.*
- *Offers a detailed view of specific locations with varying address information, showing a concentration in Europe and parts of North America.*

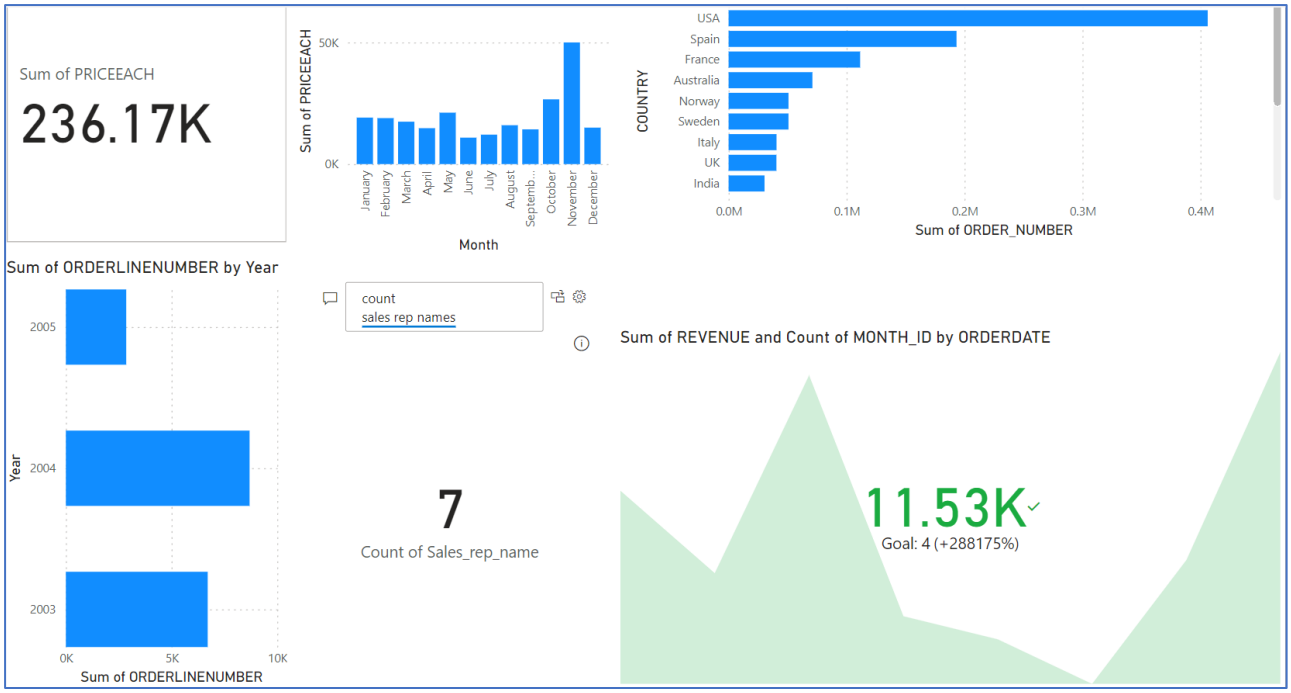


Figure 10: If using of many charts as creat a one Dashboard.sum of country,year of sales,etc..

The total sales value *price each is substantial*.

- *November is a peak month for sales value.*
- *The USA leads in order numbers, highlighting a significant market.*
- *2004 was a peak year for order line numbers.*
- *There are 7 distinct sales representatives.*
- *The revenue trend is visualized over time, with a significant achievement of the set goal.*

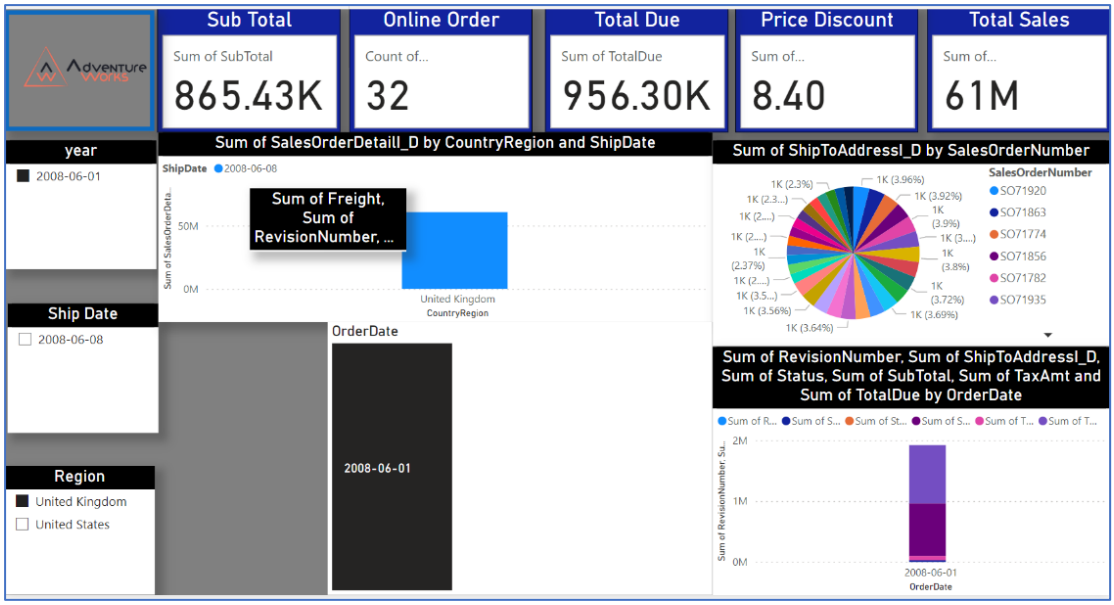


Figure 11:.online orders, price Discount,Total Sales,etc…

Sales Performance-The total sales value and other key metrics like subtotal, total due, and price discount provide a snapshot of overall sales performance.

Online Orders-Indicates the count of online orders, which could be important for e-commerce analysis.

Geographical Distribution-The bar chart by country region and ship date highlights the sales distribution geographically and temporally.

Order Analysis-The pie chart and stacked bar chart allow for a deeper analysis of specific sales orders and various sales metrics over time.

Figure 12: Dashboard

Price-Pie chart showing the distribution of costs between the sum of StandardCost (129.28K, 37.05%) and the sum of ListPrice (219.66K, 62.95%).

*Total Due, Tax Amount, and Sub Total**:**A donut chart illustrating the relationship between SubTotal (865.43K, 45%), TotalDue (956.30K, 50%), and a smaller component (69.23K, 3.66%).*

*Freight**-**Semi-circle chart showing a value of 21.64K out of a possible 43.27K.Sum of TaxAmt by Country/Region and City:*

Map-A map visualization highlighting the tax amounts by different cities (El Cajon, El Segundo, Elgin).

10. Step by step procedure for key tasks

Here we present cookbooks or step by step procedures for accomplishing key data engineering tasks given below

- Task 1- preparing SQL DB. This is performed on-prem
- Task 2 - Resource Group creation and other cloud configuration(see *Figure:1)*
- Task 2.2 -Key Vault and Secret creation (*see Figure :2)*
- Task 3 - How to create Azure Data Factory and configure (see *Figure:3,4,5,6)*
- Task 4- Azure Data Factory configure and Extract (see *Figure:7,8,9,10,11)*
- Task 5-Data bricks configuration bronze, Silver, mount the containers
- Task 6-Using Data Bricks create python notes connect to cluster
- Task 7-Data Transform from bronze–to-silver, silver-to-gold
- Task 8-Data Load (Azure Synapse Analytics)
- Task *9*-Data Reporting using Power BI

Step	Action	Remarks
Step 1	Download 'AdventureWorksLT2019.bak' file from web & copy to C ->Program Files->Microsoft SQL server-> MSSQL16.MSSQLSERVER->MSSQL->Backup	Download link https://learn.microsoft.com/en-us/sql/samples/adventureworks-install-configure?view=sql-server-ver16&tabs=ssms
Step 2	Login to SSMS as **sa** user	SSMS - SQL Server Management Studio
Step 3	Select Database and Right click on it and from the drop-down menu, select the option 'Restore Files & File Groups'	The following video explains the steps https://www.youtube.com/watch?v=ntsigyCkCas
Step 4	In the 'Destination to Restore …' Box, fill the Database Name as 'AdventureWorksLT2019'	
Step 5	In the 'Source for Restore' Box In the right side of 'From Device' box, **click …** Window with Name 'Select Backup Devices' pop up.	
Step 6	Click ADD	
Step 7	Select the file of your interest such as 'AdventureWorksLT2019.bak'	
Step 8	Press OK	
Step 9	Follow further instructions	
Step 10	Refresh and Verify the Database is displayed. Also Check all the Tables are available	

Table 1: Task 1, preparing SQL DB. This is performed on-prem

Step	Action	Comments
Step 1	Go to Azure Portal	
Step 2	You will see a Resource Group Called Student-rg-n3. If not create one	
Step 3	Click ‘Student-rg-n3’ it will take you to the respective resource group	We will use the existing RG to create an end-to-end project

Table 2: Task 2 – Resource Group creation (see *Figure:1)*

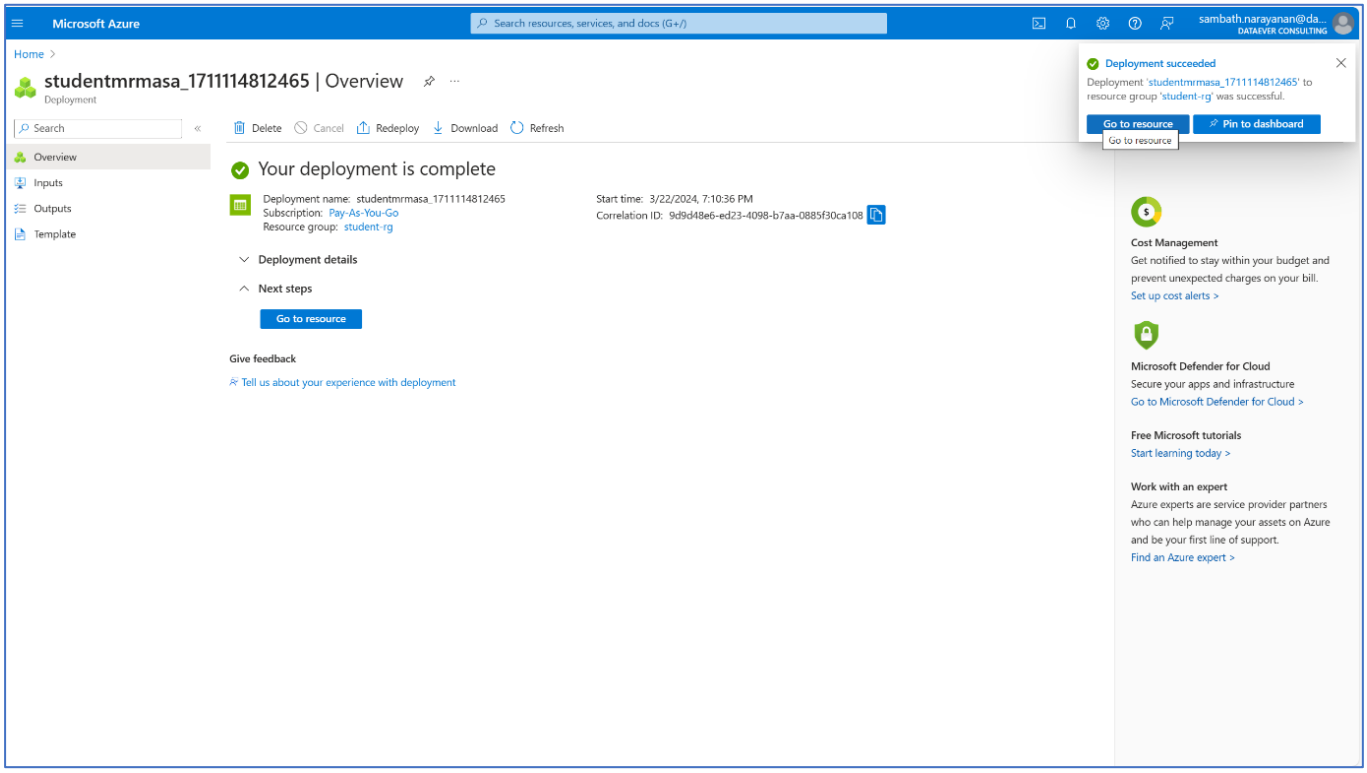


Figure 1: Resource Group created on Azure cloud

Step	Action	Comment
Step 1	Click + Create	
Step 2	Go to Market Place	
Step 3	Search for Key vault	
Step 4	Click Create	
Step 5	Fill in Subscription, Resource Group Name,	
Step 6	Give a Key vault Name as 'Student-sn020-kv4'	
Step 7	Choose Region as South India	
Step 8	Click Next	
Step 9	Choose Vault Access Policy	
Step 10	Choose other default values such as Public Network Access	
Step 11	Choose Secrets	
Step 12	Click + Generate/Import	
Step 13	Fill the name as 'login'	encrypted
Step 14	Fill the secret value as 'sn020'	Instead of direct login, safe way
Step 15	Leave other option as they are	
Step 16	Click create	
Step 17	Click + Generate/Import	
Step 18	Fill the name as 'password'	
Step 19	Fill the secret value as '*******'	
Step 20	Leave other options as they are	
Step 21	Click create	
Step 22	Now if you select secrets, you will see two secrets – login & password	
Step 23	Check & import Datasets Establish connection between Azure Cloud and SQL server	
Step 24		By checking SSMS you can find, there two schemas. One **dbo** and other is **SalesLT**
Step 25	Run the following if user for DB is not created already "CREATE LOGIN sn020 WITH PASSWORD '******' Create user sn020 for login sn020"	Keep always checking whether DB you are using is the correct one To check user already created for DB, check with following script
Step 26	How to give db_data reader permission to the sn020 user? Choose the username in the left panel. Right click the username. Check for properties. Click Membership. Select db_datareader	Now we can connect and access these Tables

*Table 3:*Task 2.2 – Key Vault and Secret creation (*see Figure :2)*

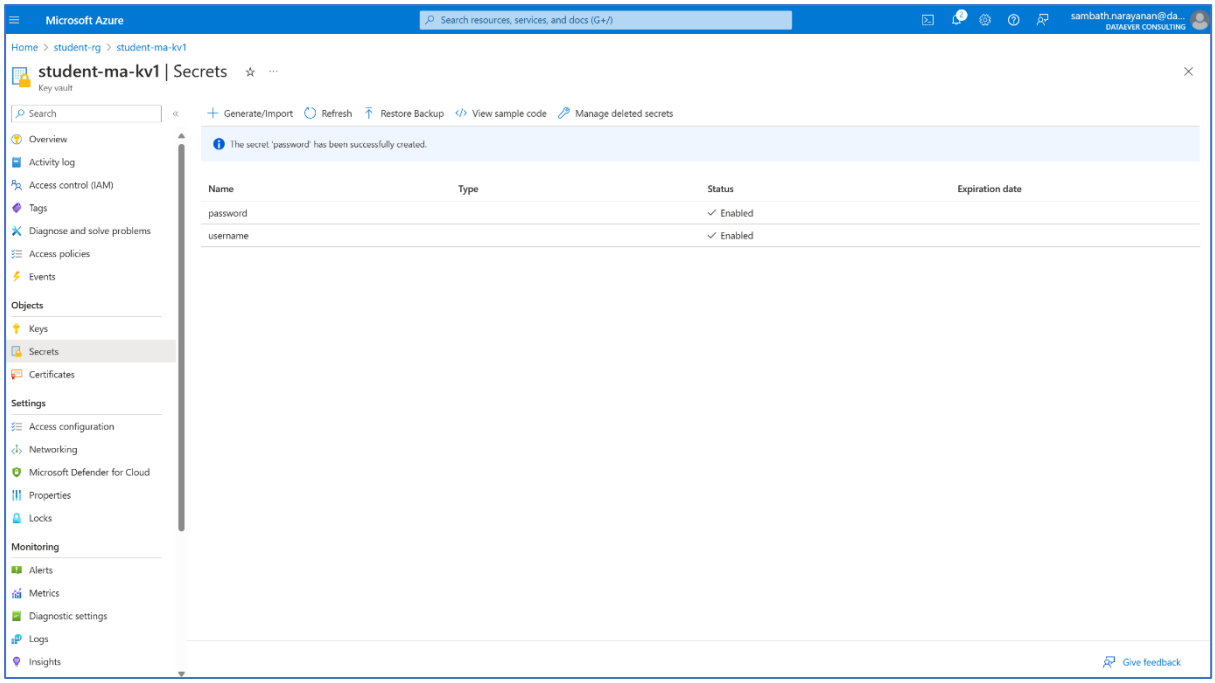


Figure 2: Key Vault created Secret name and password

Steps	Action	Comment
Step 1	Go to Resource Group student-rg-n3	
Step 2	Click + Create, Marketplace page will open	
Step 3	Search for 'Azure Data Factory'	
Step 4	Click the panel with Heading 'Data Factory'	
Step 5	Click Create	
Step 6	Fill Instance Details such as Name = **student-sn020-adf4** Region = **South India** Then Click **Next**	
Step 7	Leave Networking as it is and Click Next	
Step 8	Leave Advanced Tab as it is and Click Next	
Step 9	If you want fill Tags, otherwise click Next and go to the next tab	
Step 10	Next Tab is Review+ Create Click Create	Deployment in progress message will be followed by Your Deployment Complete
Step 11	Click Go to Resource Group	
Step 12	Click the just created Azure Data Factory '…' You will see the Button 'Launch Studio'	
Step 13	Click Launch Studio This will take you to Azure Data Factory workspace	No relation between ADF & on Prem SQL In order to establish connection, we have to use **Integration Runtime**
Step 14	You need to install S**elf-hosted integration Runtime** on the device where SQL DB is present	
Step 15	Click the Manage Tab in the left	
	It will open General Box. In that you will find 'Integration Runtime' Click the same	
Step 16	Click **+ New** You can find different integration Runtimes	What already exists is auto resolve integration runtime This can't be used for connecting with On-prem system
Step 17	Click **Azure, Self-hosted Box** & hit Continue	

Step 18	You will now find Two Boxes. One Azure. Another Self-Hosted. Now You click Self-hosted	
Step 19	Click Continue	
Step 20	Give some Name as follows **SHIR**	You can optionally give description Used to connect SQL server
Step 21	After that you see Create at the Bottom page. Click **Create**	In the Top Right you will get a pop-up with message **Successfully Saved**
Step 22		Actually, you will find two options 1. Express setup 2. Manual Setup **Manual**: download application and after install, copy the keys and paste If already running, stop and uninstall on the on-prem system
Step 23	After successful uninstall, click the Express setup link on Azure Portal. This will download a file to local system. Then you use that to install, by clicking that file Keep watching the status of Express Setup Four steps	
Step 24	After Successful Install, you will get Integration Runtime (self-hosted) “SHIR” is successfully installed on your computer	While installing authorization, ensure the local system clock is correct Also launch SHIR locally to verify whether it is Referring to the correct ADF name which we created **‘student-sn020-adf4’**
Step 25	Click Close	
Step 26	Go back to Azure Portal- Azure Data Factory	
Step 27	Click Close	You will now see in the Integration Run Time Page, SHIR is Running
Step 28	In ADF, left panel, choose click **Author - Pipelines**	
Step 29	Click + next to Rectangle Box	
Step 30	Again, you will see Pipeline-Pipeline, you click the same	

Step 31	In the Right Column of the screen, you will see the Properties Panel	
Step 32	Give the Name ‘copy_pipeline4’. You will now see the name of ‘copy_pipeline4’ in **Factory Resources** panel.	In the top corner you can see a small Blue square Box (Properties). If you click the Properties Panel will be minimized
Step 33	Next to Factory Resources Panel, you will see Activities Panel. In that search for Copy data activity	
Step 34	Now you drag & drop the copy data activity into the white canvass in the middle of the screen. You will see a BOX with title ‘Copy data’	Below that canvass, you will now see Tabs like ‘General -Source-Sink-Mapping-Settings-User properties’
Step 35	We will copy SalesLT2019 Table from AdventreWorksLT2019	
Step 36	Click General in the bottom Box & Give the name as ‘Copy address table4’	
Step 37	Click ‘Source’ in the bottom Box & You will see source dataset box appearing. Next to that box ‘+ New’ option will be available Click ‘+ New’	
Step 38	You will now see the ‘New data set’ page in the right side of Window	
Step 39	In the search Bar, search for SQL Server. You will get SQL Server icon/thumbnail	
Step 40	Now you click SQL icon and click ‘Continue’ Button you see at the bottom page	
Step 41	Now a new ‘Set properties ‘page will appear. Give Name as ‘address4’	
Step 42	In the ‘Linked service’ Box, a drop-down arrow will be available. Click the arrow	Connection String. 2 connect to DB we need a connect string
Step 43	You will see ‘New’ option. Click ‘New’	Connection String
Step 44	Now ‘New linked service’ page will be displayed. In that you see input data box expecting values for Name, Description, Connection via integration runtime, Server Name, DB name, Authentication Type…	*Some can be selected from drop-down option*

Step 45	Name = onpremsqlserver4	You can skip description. Boxes with * are compulsory
Step 46	Connect Via integration Runtime has a dropdown arrow ‘v’. When clicked, you can see SHIR.	
Step 47	Select SHIR	
Step 48	Server Name = SQLNODE2	This is the Local Host server name
Step 49	Database Name = AdventureWorksLT2019	Value can be taken from SSMS
Step 50	Authentication Type = SQL Authentication	Other option is Window authentication
Step 51	User name = sn020	It can be accessed from key vault also. But here we are directly giving
Step 52	Password This secret. We have to take this from ‘Azure Key Vault’	There are two buttons: Password ‘Azure Key vault’
Step 53	Click on ‘Azure Key Vault’ and get secret encrypted from KV	KV = Key Vault
Step 54	You will see the input boxes appearing for AKV linked service* Secret name * Secret version	**Our understanding** Whenever two separate software/devices are involved, one needs a ***Linked service***
Step 55	Clicking ‘AKV linked service, drop-down ‘v’ gives an option for ‘+New’	
Step 56	Click on ‘+New’	
Step 57	Now a new page with title ‘New Linked Service’ gets opened. You need to fill correct values in the input boxes	
Step 58	Name = AzureKeyVault4	
Step 59	The Azure Key vault selection Method comes pre-selected as ***From Azure Subscription***	
Step 60	The drop down shows available subscriptions. Choose your preferred subscription	
Step 61	Azure Key Vault Name = **student-sn020-kv4**	This is selected from drop down menu option
Step 62	The authentication method input Box comes pre-selected with System assigned managed identity	When you create any resource, you get an identity with that. Object id. When you use this to connect it will use Azure Key Vault
Step 63	Clicking on Test Connection should show Connection Successful message	This used the object id to connect to Azure Key vault.
Step 64	Now we have given all the inputs. Now you click the **Create** Button at the Bottom	A pop-up will display a message saying It will be created
Step 65	In the beginning, Secret Name = (Loading Failed) will appears	As you are aware, we have added two secrets already. 1. Username 2. Password

	This is because the Reader permission is not granted	
Step 66	We need to configure additional things in KV site.	
Step 67	Go to Azure Key Vault	
Step 68	In the left column panel, you will see ‘Access Policies’ option	
Step 69	Click on Access Policies	You will see
Step 70	Click + Add has drop down menu. One of them is *Add Role Assignment*	
Step 71	In that click & choose ‘Add Role Assignment’	
Step 72	A list of Roles displayed. In that select ‘Key Vault Secrets Officer’	
Step 73	Follow instructions such as Click Next	
Step 74	A new page Titled ‘Add Role Assignment’ appears. In that few options in horizontal Tabs are shown	
Step 75	In that select ‘Members’	
Step 76	Against Members, there will be ‘+Select ‘members link. Click that link	
Step 77	Now in the Right side of the screen a new page with title ‘Select Members’ appears	
Step 78	In that choose the displayed Member Name as ‘Sambath. Narayanan…’	
Step 79	In the Bottom, there will be a ‘Select’ Button. Click Select	
Step 80	Click Next	
Step 81	Click Review + Assign	
Step 82	In the Right Top corner of the screen, a pop-up appears with the message “Added Role Assignment “	
Step 83	Click on Access Policies to verify whether the User ‘SAMBATH NARYANAN …’ is added	
Step 84	After verifying …	
Step 85	Again, In the left column panel, you will see ‘Access Policies’ option	
Step 86	Click Access Policies and a new Page will appear	
Step 87	In that new page click ‘+Create’	
Step 88	Do you see a ‘Create an access policy’ page which has the permission Tab.	

Step 89	Click Permissions and you will get more options such as a. Key Permission b. Secret Permission c. Certificate Permissions	
Step 90	Select all check-boxes under Secret Permissions. In the Bottom 'Click Next'	You have now necessary permission to access KV
Step 91	Click 'Next 'Button at the Bottom. You will see Green Tick adjacent to Permissions	
Step 92	Under Principal Tab, in the Search Bar, search for **'student-sn020-adf4'**	
Step 93	Click the 'student-sn020-adf4' entry from the list. And click Next at the Bottom	Application Details are optional
Step 94	Click Next	
Step 95	Click 'Review + Create'	
Step 96	Click 'Create'	
Step 97	"Updating the KV student-sn020-kv4' message will appear with Green-Tick on Top Right of the screen	
Step 98	Now ADF has secret permission for all secrets Corresponding to ADF	
Step 99	You will see Secrets Are Enabled	
Step 100	Go To ADF 'student-sn020-adf4'. You will see the previous 'New linked service' page	We will now test the secrets by Refreshing the page
Step 101	In the Secret Name drop-down options, you select the password	
Step 102	After that ADF can read the password	
Step 103	Click always encrypted **certificate trusted.** Click the test connection. Connection was success	

Table 4: Task 2.3 – How to create Azure Data Factory and configure (see *Figure:3,4,5,6)*

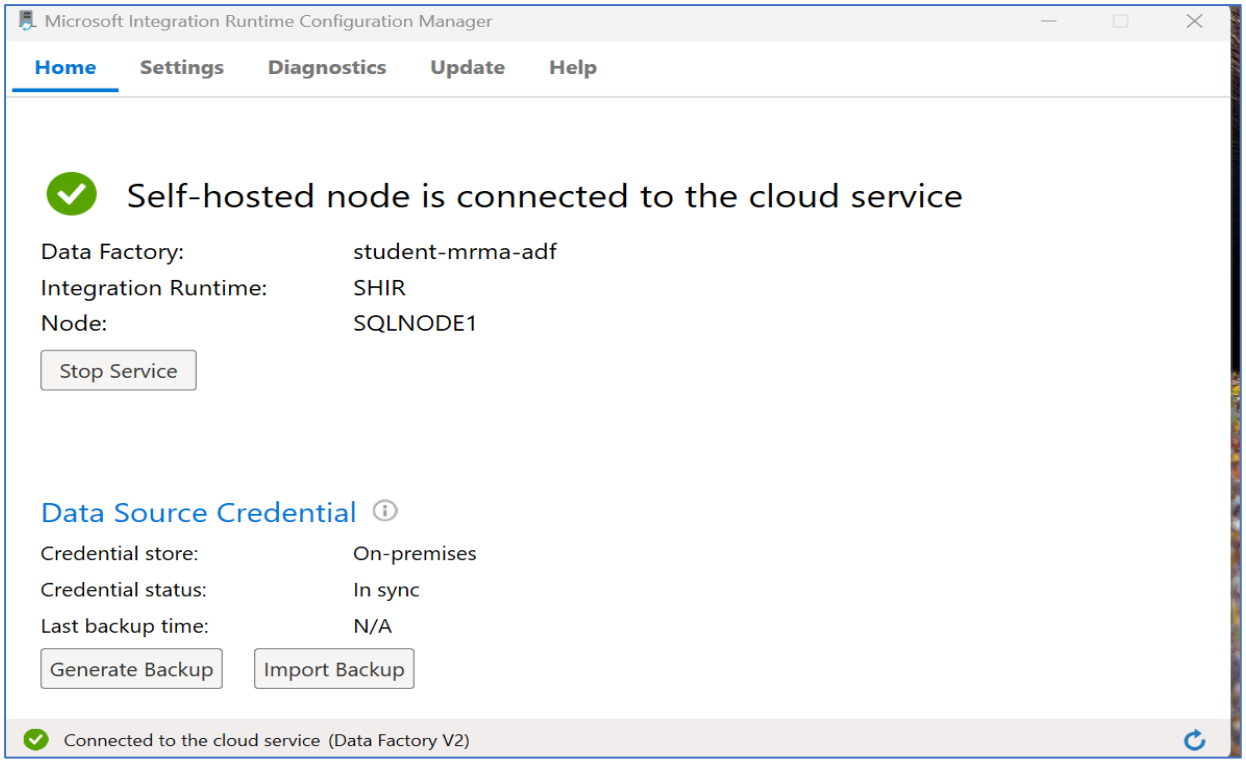


Figure 3: Data Factory Self-Hosted Integration run time connected to cloud service

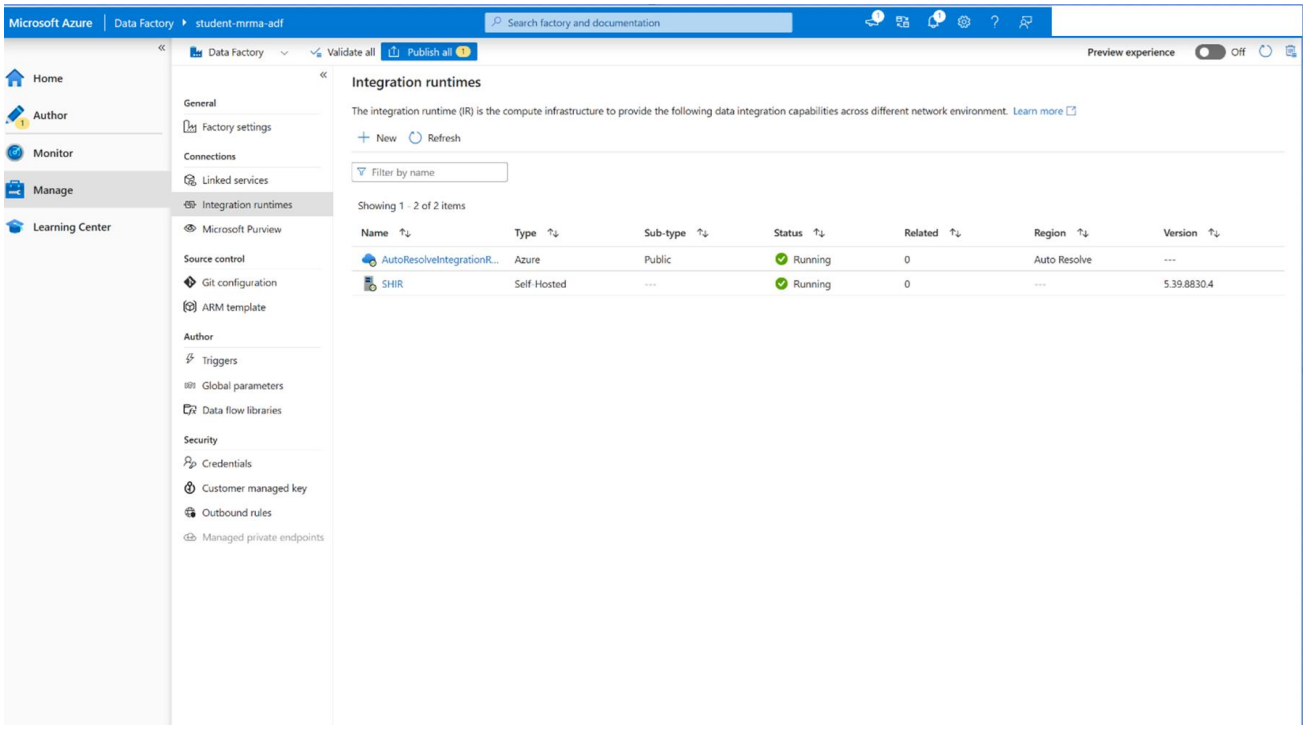


Figure 4: Data Factory Self-Hosted Integration run time as running

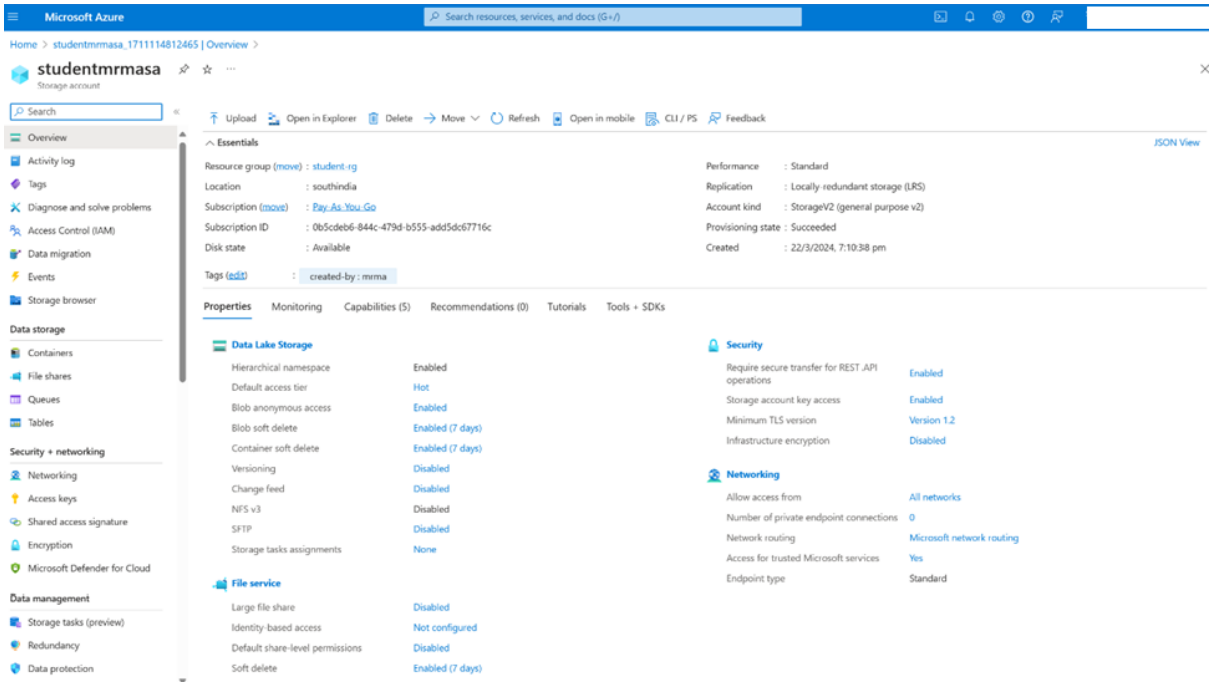


Figure 5: Azure Storage Account will be created

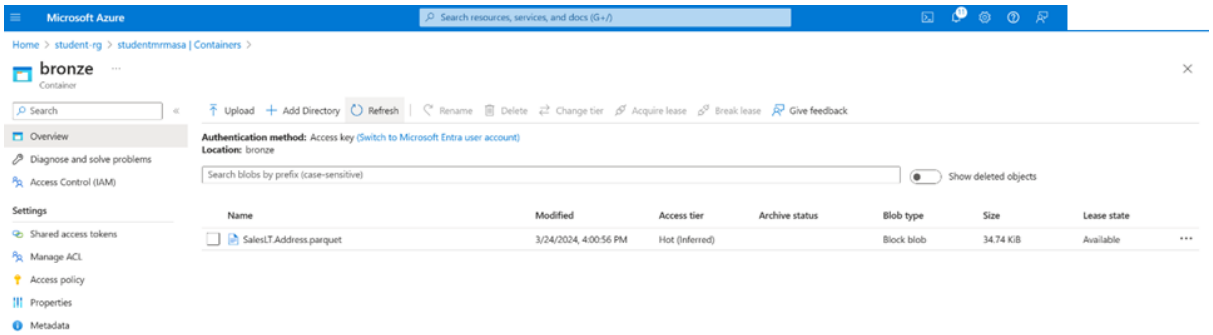


Figure 6: Go to Azure Storage Account will see a container name as SalesLT

Step	Action	Remark
Step 1	Open just created ‘student-sn020-adf4’	
Step 2	Go to Author and click + at the top of Second Column titled ‘Factory Resources’	
Step 3	The Column page at the right side with title ‘**Properties**’ opens	
Step 4	In that fill the Name Box with ‘Copy_all_tables4’	
Step 5	Click Properties icon on the Top. The column closes and you see a new pipeline named ‘copy_all_tables4’ gets created	
Step 6	Open SCMS and select Adv…LT2019 & click ‘New Query’ to open a Query Window	

Step 7	Type the following Query
	
	`USE AdventureWorksLT2019;`
	
	`SELECT s.name AS SchemaName,`
	`t.name AS TableName`
	`FROM sys.tables t`
	`INNER JOIN sys.schemas s ON`
	`t.schema_id = s.schema_id`
	`WHERE s.name = 'SalesLT';`
		When you run you will see the following
		
		`SchemaName TableName`
		`1 SalesLT Address`
		`2 SalesLT Customer`
		`3 SalesLT CustomerAddress`
		`…`
		`…`
		`10 SalesLT SalesOrderHeader`

Step 8	Go to ADF, go to third column, in Activities, fill the search box with ‘lookup ‘	We will copy all 10 Tables listed above
Step 9	Drag the ‘lookup’ and put it there in the canvas	Below the canvas you see, horizontal options as given below *General Settings User props*
Step 10	Under General option, change the Name to ‘Look for all Tables4’	
Step 11	Click ‘Settings’ Fill-in the Source dataset box by clicking The drop-down Box Select +New	We will use the same SQL query
Step 12	Another Window open titled ‘New Data set’ with option thumb nails for few datasets names	
Step 13	Select ‘SQL server’	
Step 14	Click ‘Continue’	
Step 15	Another page titled ‘Set Properties’ opens	
Step 16	Fill the Name Box with ‘SqlDBTables4’	
Step 17	Click drop-down in linked service*	

Step 18	Select ‘onpremaqlserver4’	No Table name is elected explicitly, leave as it is
Step 19	Click OK	
Step 20	Under setting, against Use query, select ‘Query’ button and a box opens below	
	In that Box paste the same query from SSMS as in step 7 above SELECT s.name AS SchemaName, t.name AS TableName FROM sys.tables t INNER JOIN sys.schemas s ON t.schema_id = s.schema_id WHERE s.name = 'SalesLT';	We copy multiple Tables by running this query Check this step by selecting preview data, unselect tick mark from First Row Only option
Step 21	To run the lookup activity, **click** the top option called as **‘Debug’** Now will you see your lookup activity is created.	The lookup activity will use the same query to query against the SQL database The run the lookup activity is debug option you see the things as output window You Can see the succeeded
Step 22	Check the and see the output and input by clicking suitable small icon left of Type = lookup	You will see Output in the form of Jason structure You will see input form of SQL query
Step 23	In the ‘Activities’ col search for ‘for each’ and drag and drop the same into canvas	
Step 24	Again, you see horizontal options like General Settings Activities user props	
Step 25	Select general & fill the Name as ‘For Each schema tables4 ‘	
Step 26	In the canvas **connect** ‘Lookup Activity’ □ ‘ForEachSchemaTable’ Box	For is for each loop you will be using in the C or Python 1.Iterates thru a list 2.each item has schema name and table name 3. this will run 10 times one by one 4. because there are 10 items
Step 27	Go to Settings tab	
Step 28	In the options, you have ‘items. In that Box below choose ‘Add dynamic content’ option	

Step 29	Then a page with title ‘Pipeline expression builder’ appears	
Step 30	Below the box, under ‘Activity outputs’	It will get all outputs in the activity. E.g. cl
Step 31	Click on look for all tables	
Step 32	You will see in the Top Box, a SQL code appears @activity (‘look for all tables’). output	This is called dynamic expression
Step 33	Modify above script as follows `@activity (‘look for all tables’). output. Value`	For each can iterate thru the loop one by one and get the schema name and Table Name one by one. We can do copy operation
Step 34	At the Bottom, you press OK button	
Step 35	Select the Activities Tab	
Step 36	You see pencil icon against ‘For Each’ case	
Step 37	Click Pencil	
Step 38	It will take you inside the For Each Loop	We are now inside For Each, you can view
Step 39	Go to Activities Column, and search for ‘Copy Data’	
Step 40	Drag & Drop Copy data into the canvass	
Step 41	Update the Name of this activity to ‘Copy Each Table’	
Step 42	Click Configure & input Source Option values	
Step 43	Source Dataset Box should be filled with	
Step 44	Click + New	
Step 45	After that, A page with title ‘New dataset’ appears	
Step 46	Search SQL server	
Step 47	Click SQL server	
Step 48	Click Continue	
Step 49	A new page titled ‘Se properties’ appears	
Step 50	Input the Name Box with ‘SqlServerCopy4’	
Step 51	After, Select & input ‘Linked service’ *	
Step 52	Select ‘onpremsqlserver4’ from drop-down.	This was previously created ‘onpremsqlserver4’
Step 53	In the Box ‘Connect the integration runtime*’ No need to input any value	
Step 54	Click OK & you will get back to the Canvas and options below	

Step 55	Now select **'Source'** again	
Step 56	Fill the Use query Box by Clicking Query button	
Step 57	Below that **Query*** Box opens	
Step 58	Click **'Add dynamic content'** Below the Box	
Step 59	A new page titled 'Pipe line expression builder appears'	
Step 60	Below Box several options, such Foreach iterator, … appears	
Step 61	In the Box type the following query @{concat('SELECT * FROM ', item().schemaname, '.', item().TableName)}	One by one iteratively until all the Tables are copied
Step 62	Next select the **'Sink'** option	
Step 63	Adjacent to the 'Sink dataset*' Box, there is **+New** option	
Step 64	Click **+New**	
	A new page titled 'New dataset' opens which has thumb nails for Blob storage, Cosmos DB, and so on	
Step 65	Select 'Azure Data Lake Gen2' and hit Continue	
Step 66	You will now get another new page titled 'Select format'	
Step 67	You see Thumbnails of formats listed, you select 'Parquet' and hit Continue	
Step 68	You now see a new page titled 'Set properties'	
Step 69	Fill the Name Box with 'parquetTables4'	
Step 70	Then Click the Linked service*	
Step 71	Choose the on which we have already created namely, 'AzureDataLakeStorage1'	
Step 72	In the ensuing options, against File path, click the Browse icon to select the '**bronze'** container.	In general, When you put the file into bronze container, you need to follow the folder structure as bronze/Schema/Tablename/Tablename.parquet Specifically for our case bronze/SalesLT/Address/Address.parquet
Step 73	Click OK and you will be at the previous page titled 'Set properties'	

Step 74	Click OK you will be at the canvas page	How to configure this in our Azure Data Factory
Step 75	Below the canvas, there are options such as General Source Sink ….	
Step 76	In that select sink option	
Step 77	You will Sink dataset* Box and others	
Step 78	Next to the above box, pencil icon Open	
Step 79	Click the Pencil	
Step 80	Then a New Canvas titled, paquestTables4 appears	
Step 81	Below that Box you see Connection Schema Parameters	
Step 82	Click Parameters and +New	
Step 83	Fill the ensuing boxes as follows	
Step 84	Under Name fill schema name	
Step 85	Again click +New	
Step 86	Fill the ensuing boxes as follows	
Step 87	Under Name fill table name	
Step 88	Above the canvas you will see few activities listed. One of those is ‘copy_all_tables4’	
Step 89	Click copy_all_tables4	
Step 90	Now you will see ‘Copy Each Table ‘Box in the canvas’ and below with option tabs General Source Sink …	
Step 91	Click the Sink option	
Step 92	You will now see schema name and table name	
Step 93	Now for filling schema name Value, click ‘Add dynamic content’ link	
Step 94	After that you will see a page titled Pipeline expression builder	
Step 95	Again, below the Box, click For Each Iterator	
Step 96	You then click ‘For Each Schema’ Table link below search Box	
Step 97	Then Add dynamic content Box gets filled automatically with the script @item()	
Step 98	Modify the above script as follows @item().SchemaName	
Step 99	After that click OK	
Step 100	This takes you to the previous page on which you this time select and input for ‘table name’	

Step 101	Now for filling table name Value, click ‘Add dynamic content’ link	
Step 102	After that you will see a page titled Pipeline expression builder	
Step 103	Again, below the Box, click For Each Iterator	
Step 104	You then click ‘For Each Schema’ Table link below search Box	
Step 105	Then Add dynamic content Box gets filled automatically with the script @item()	
Step 106	Modify the above script as follows @item().TableName	
Step 107	After that click OK	
Step 108	This will take you to the previous page	We use these two parameter values to create the folder structure
Step 109	Click Sing option, select Connection, you will see file path	Bronze/ Directory/Filename
Step 110	Click the Directory Box	
Step 111	Click Add dynamic content	
Step 112	The page titled ‘Pipeline expression builder’ appears	
Step 113	Below the Parameter option, you see Schemaname Tablename	
Step 114	In the Top big box paste a code as follows **@{concat(dataset().schemaname, '/', dataset().tablename)}**	For directory structure. Based on the parameters we have defined dataset for this s
Step 115	Click OK	
Step 116	This leads to previous page with parquet icon in canvas	
Step 117	When you click Connection some values in the box will be already populated. You need to fill the File name as	
Step 118	Click the File name	
Step 119	Click Add dynamic content	
Step 120	The page titled ‘Pipeline expression builder’ appears	
Step 121	Below the Parameter option, you see Schema name Table name	
Step 122	In the Top big box paste a code as follows **@{concat(dataset().tablename,'.parquet')}**	For File Name structure. Based on the parameters we have defined dataset for this as
Step 123	Click OK	

Step 124	This leads to page with parquet icon in canvas	
Step 125	From Top of the Canvas activity menus, choose ‘copy_all_tables’	
Step 126	Now below the canvas you will see options horizontally listed as General Source Sink …	
Step 127	Now click sink	
Step 128	In the canvas section click ‘copy_all_tables4’ link	
Step 129	The canvas section displays the pipelines- ‘LoopUp’ & ‘ForEach’	
Step 130	This is time now to Publish All it will open page title publish all with list of things to publish	
Step 131	Click Publish Button below	
Step 132	We can run the pipeline and see whether it is able to copy all the tables or not	
Step 133	From the Top of canvas Menu, choose Add trigger and in the activity box heading, ‘Trigger now’	
Step 134	Now a new page title; Pipeline run’ opens	
Step 135	Click OK	Now it is going to run the pipeline
Step 136	You will see the previous canvas page	
Step 137	Running message pop-up window appears on top-right corner of screen	
Step 138	Click Montor Option from the left-most column-panel	
Step 139	New page titled ‘Pipeline runs’ appears which displays the status of the run	
Step 140	In that click copy_all_tables pipeline	
Step 141	Individual activities are displayed	
Step 142	Now the status column should display Succeeded for all Activity names	10foreach activity Each for each activity trying to copy one table
Step 143	Goto storage container & Refresh with folder name SalesLT, click schema name, you will see the Tables	

Table 5: Task 3 – How to create Azure Data Factory and configure (see *Figure:7,8,9,10,11)*

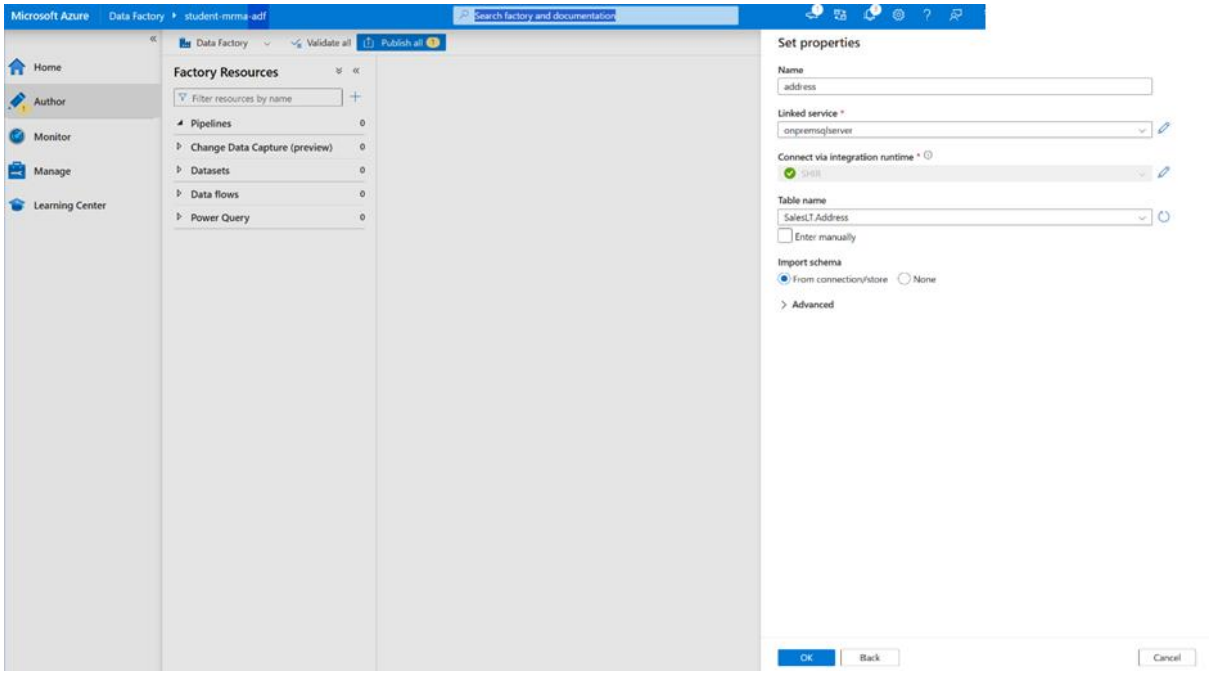


Figure 7: Azure Data factory Pipeline. Connected to Self-Hosted Integration run time

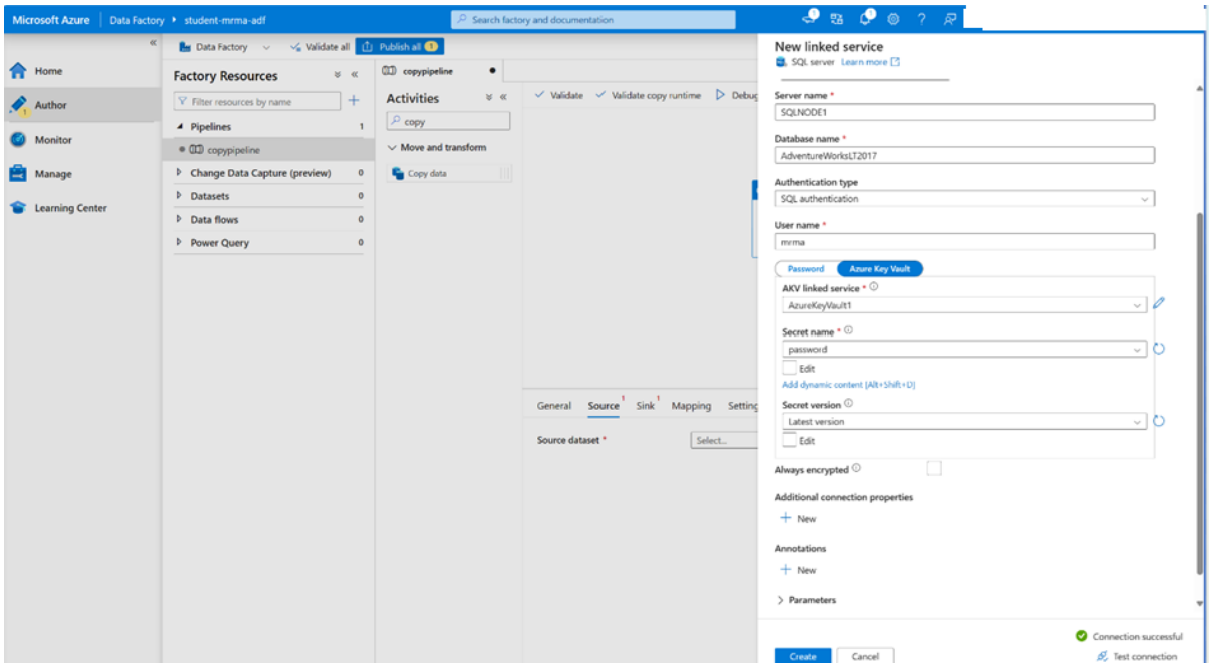


Figure 8: Azure Data factory Pipeline connected to SQL server

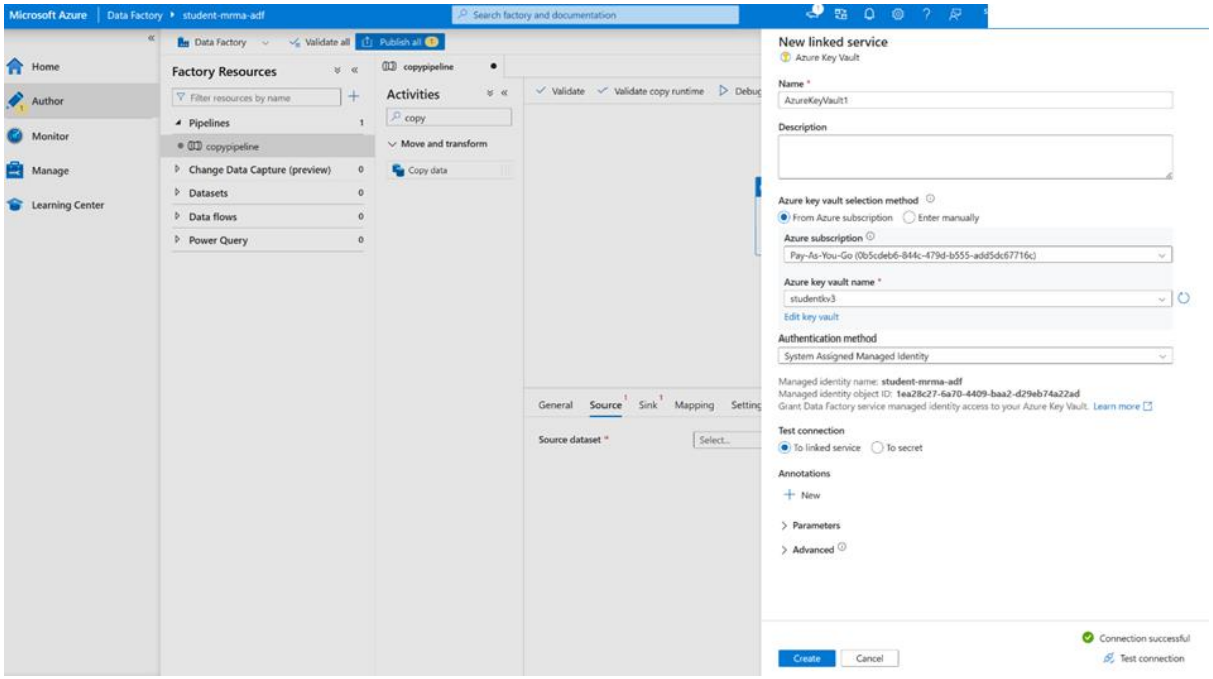


Figure 9: Azure Data factory Pipeline connected to Key Vault

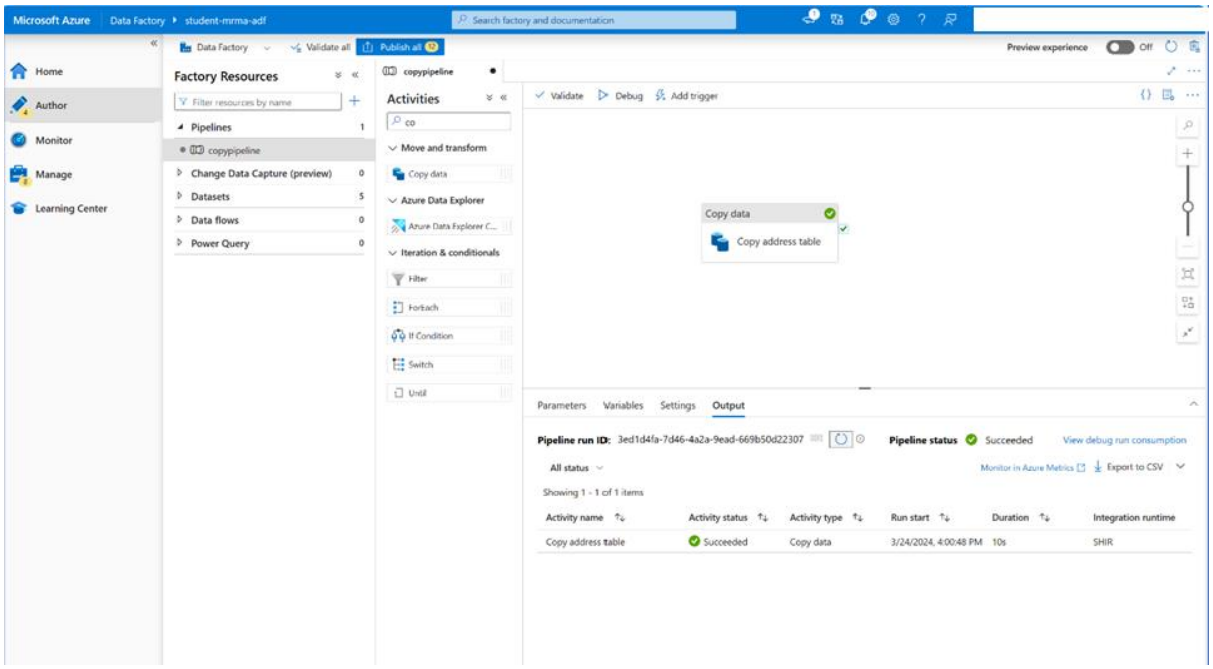


Figure 10: Azure Data factory Pipeline will be Succeeded

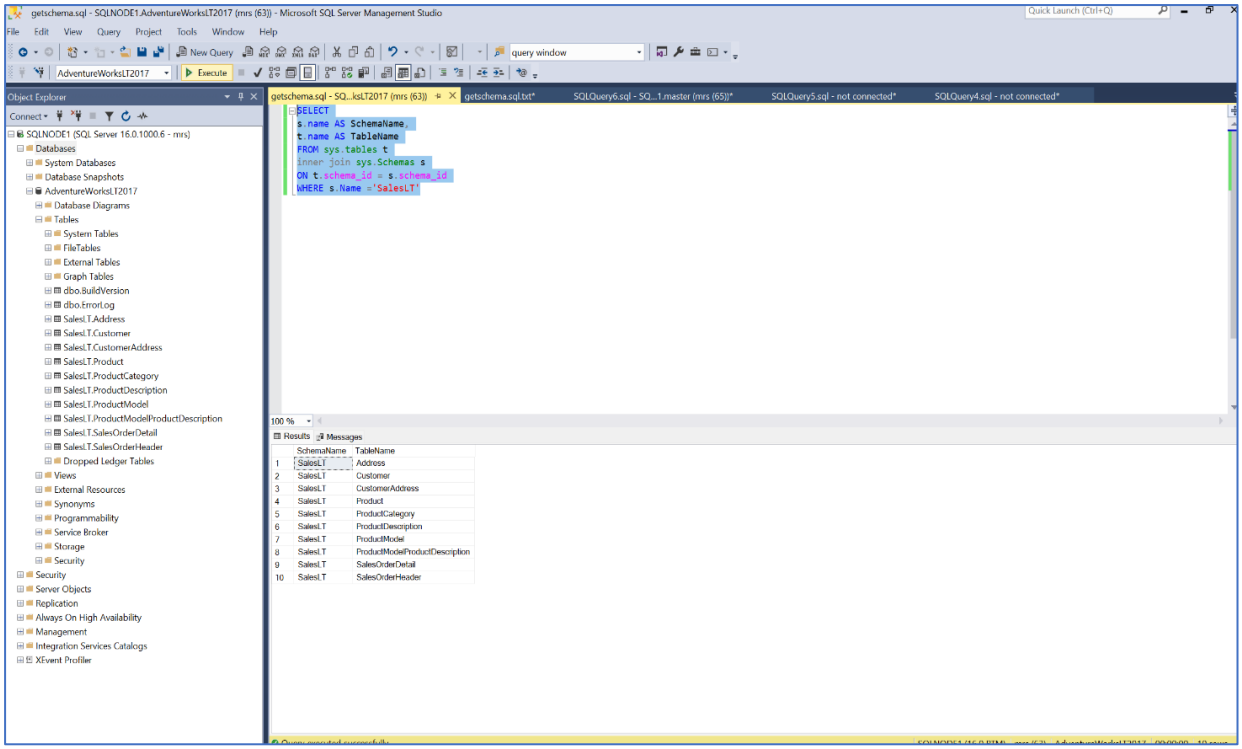


Figure 11: SSMS to create a new query of SalesLT

Step 1	Create Azure Data Bricks with name ‘student-sn020-adb4’	This step itself has multiple sub steps. Refer to PART 5.1
Step 2	First click Data Bricks and launch workspace button	
Step 3	New Workspace opens	All data transformation logic is developed here. Similar to ADF workspace for pipeline
Step 4	Click Compute in the left panel	Workspace is used for creating NB
Step 5	Click Create Compute	DB & Tables can be Create
Step 6	Name of the compute Cluster – give name as data_transformation	Compute Tab-important- – to create Compute cluster Power, use this – also called Sparc clusters
Step 7	Go with all defaults such as Policy, nodes are single node, access mode is single user, Data Bricks run time version	Workflow – for creating jobs – but we will not use
Step 8	Select Node Type as default	
Step 9	Terminate after 15 mins	Less mins saves money
Step 10	Click Advance option	
Step 11	Enable **Credential pass through**	With this Data Bricks can directly access Blob storage
Step 12	Configure Role Assignment as Storage Blob Data Contributor	
Step 13	Go to Resource Group	
Step 14	Click storage account	
Step 15	Click Access Control IAM	
Step 16	Click +Add	
Step 17	You will see two opts: 1. Add Role assignment 2. Add co-admin	
Step 18	Click 1. Add Role Assign	
Step 19	Choose ‘Storage B Blob data contributor	
Step 20	Click Next	
Step 21	Click Select Members link	
Step 22	New Page titled Select Members shows up	
Step 23	Select the right User Name ‘Sambath ….’	Manage identity
Step 24	Click Select at the Bottom	
Step 25	Click Next	
Step 26	Click Next	
Step 27	Click Review + Assign	
Step 28	Click 1. Add Role Assign	
Step 29	Choose ‘Storage Blob data contributor	
Step 30	Click Next	
Step 31	Click Select Members link	
Step 32	New Page titled Select Members shows up	
Step 33	Select the right User Name ‘student-sn020-adf4’	
Step 34	Click Select at the Bottom	
Step 35	Click Next	
Step 36	Click Next	

Step 37	Click Review + Assign	
Step 38	We will see role assignment is created	
Step 39	Click Create Cluster	You get Message saying Cluster is created, it will take 5 mins
Step 40	In the left panel, click Workspace	
Step 41	Workspace column page appears. In that you see Shared	
Step 42	Select Shared	
Step 43	A white workspace page opens with Create Button on the right Top corner.	
Step 44	Click the Create Button	
Step 45	Drop down with option shows up	
Step 46	From that drop-down, select Notebook	
Step 47	A NB with title untiled Notebook opens	
Step 48	Change the Name of the Notebook 'storagemount'	
Step49	Choose the language from drop-down next to Name of the NB	
Step 50	Choose Python	
Step 51	Ensure that you run the Note Book on right Cluster, 'data_transformation'	
Step52	Open the Microsoft website https://learn.microsoft.com/en-us/azure/databricks/archive/credential-passthrough/adls-passthrough	1. Access Azure Data Lake Storage using Azure Active Directory credential passthrough(legacy) –
Step53	This link has How to Mount instruction	
Step54	Copy the Python code available under section Mount Azure Datalike storage to DBFS using credential pass through	
Step55	Go to Azure DB NB	
Step56	Paste the just copied code into the NB cell	
Step57	Modify the python code	
Step58	Update the path	
Step59	Replace the container name as 'bronze'	
Step60	Replace the storage A/C name e as''studentsn020sa4'	We have to access container from the mount point
Step61	Change mount point name = "/mnt/bronze",	
Step62	In top-corner, there is a white window, click drop-down and choose the cluster with name 'data_transformation'	After which cluster will be running and attached
Step63	Run the cell	Code will run
Step64	After cell runs Successfully, you will get output as True	Means Successfully mounted the bronze container We can use this to access all data inside bronze
Step65	Validate by running	Will list all data

	dbutils.fs.ls("/mnt/bronze") in the new cell	
Step66	Output will be Bronze file details with 'SalesLT' will be displayed	
Step67	Validate by running dbutils.fs.ls("/mnt/bronze/SalesLT/") in the new cell	
Step68	It will list all the Tables inside SalesLT	Compare this output with storage container list
Step69	Like we mounted bronze, we will mount silver and gold containers also	Since we use credentials-passthrough, it is not mandatory to mount, it is enough to give full path pf the container. But we are mounting only
Step70	Validate by running dbutils.fs.ls("/mnt/silver") in the new cell	Output of this run should be True
Step71	Validate by running dbutils.fs.ls("/mnt/gold") in the new cell	Output of this run should be True
Step72	We can use bronze container data read and do data transform to silver container in Data Bricks	I level transformation
Step73	Part 5 Over	

Table 6: TASK-4 To do Data Transform bronze – Silver, mount the containers

Step 1	As a check compare MS SQL tables and Cloud silver container contents are identical	The real data transformation will be minimal here
Step 2	Use Databricks to d transformation of the data in bronze container	
Step 3	Actual Transformation is Day-time format to Date format	First level of transformation
Step 4	Go to Databricks workspace	
Step 5	Create two NB s. L1. ‘Bronze to silver’ L2. ‘Silver to bronze’	
Step 6	Type the code into the Note Book	The code to be typed in the Data Bricks Note Book is given below
Step 7	When you run the full code(bronze-to-silver) in the NB, from ‘display(df)’ code you will get the output similar to the one shown in Table 10	
Step 8	When you run the full code(silver-to-gold) in the NB, from ‘display(df)’you will get the following output similar to the one shown in Table 11	

Table 7: TASK- 5 – TRANSFORM Data from bronze–to-silver, silver-to-gold

AddressID	AddressLine1	AddressLine2	City	StateProvince	CountryRegion	PostalCode	rowguid	ModifiedDate
9	8713 Yosemite Ct.	null	Bothell	Washington	United States	98011	268af621-76d7-4c78-9441-144fd139821a	2006-07-01T00:00:00Z
11	1318 Lasalle Street	null	Bothell	Washington	United States	98011	981b3303-aca2-49c7-9a96-fb670785b269	2007-04-01T00:00:00Z
25	9178 Jumping St.	null	Dallas	Texas	United States	75201	c8df3bd9-48f0-4654-a8dd-14a67a84d3c6	2006-09-01T00:00:00Z
28	9228 Via Del Sol	null	Phoenix	Arizona	United States	85004	12ae5ee1-fc3e-468b-9b92-3b970b169774	2005-09-01T00:00:00Z
32	26910 Indela Road	null	Montreal	Quebec	Canada	H1Y 2H5	84a95f62-3ae8-4e7e-bbd5-5a6f00cd982d	2006-08-01T00:00:00Z
185	2681 Eagle Peak	null	Bellevue	Washington	United States	98004	7bccf442-2268-46cc-8472-14c44c14e98c	2006-09-01T00:00:00Z
297	7943 Walnut Ave	null	Renton	Washington	United States	98055	52410da4-2778-4b1d-a599-95746625ce6d	2006-08-01T00:00:00Z
445	6388 Lake City Way	null	Burnaby	British Columbia	Canada	V5A 3A6	53572f25-9133-4a8b-a065-102ff35416ee	2006-09-01T00:00:00Z
446	52560 Free Street	null	Toronto	Ontario	Canada	M4B 1V7	801a1dfc-5125-486b-aa84-ccbd2ec57ca4	2005-08-01T00:00:00Z
447	22580 Free Street	null	Toronto	Ontario	Canada	M4B 1V7	88cee379-dbb8-433b-b84e-a35e09435500	2006-08-01T00:00:00Z
448	2575 Bloor Street East	null	Toronto	Ontario	Canada	M4B 1V6	2df6d0ad-0926-4f34-a450-9b1083150cbf	2007-08-01T00:00:00Z
449	Station E	null	Chalk Riber	Ontario	Canada	K0J 1J0	8b5a7729-cb75-4303-a607-7f9793b4d94f	2005-08-01T00:00:00Z
450	575 Rue St Amable	null	Quebec	Quebec	Canada	G1R	5f3c345a-6475-41d5-b17b-db8d27733bb1	2006-09-01T00:00:00Z
451	2512-4th Ave Sw	null	Calgary	Alberta	Canada	T2P 2G8	49644f1e-6f90-46d9-8dbb-9db15f0ef7ec	2006-12-01T00:00:00Z

Table: 10 Sample output of executing **bronze–to-silver NB**

SalesOrde	RevisionN	OrderDate	DueDate	ShipDate	Status	OnlineOrd	SalesOrde	PurchaseC	AccountNi	Customer	ShipToAd	BillToAddi	ShipMeth	CreditCar	SubTotal	TaxAmt	Freight	TotalDue	Comment	rowguid	ModifiedDate
71774	2	01-06-2008	13-06-2008	08-06-2008	5	FALSE	SO71774	PO348186	10-4020-0	29847	1092	1092	CARGO TR	null	880.3484	70.4279	22.0087	972.785	null	89e42cdc-	08-06-2008
71776	2	01-06-2008	13-06-2008	08-06-2008	5	FALSE	SO71776	PO199521	10-4020-0	30072	640	640	CARGO TR	null	78.81	6.3048	1.9703	87.0851	null	8a3448c5-	08-06-2008
71780	2	01-06-2008	13-06-2008	08-06-2008	5	FALSE	SO71780	PO196041	10-4020-0	30113	653	653	CARGO TR	null	38418.69	3073.495	960.4672	42452.65	null	a47665d2-	08-06-2008
71782	2	01-06-2008	13-06-2008	08-06-2008	5	FALSE	SO71782	PO193721	10-4020-0	29485	1086	1086	CARGO TR	null	39785.33	3182.826	994.6333	43962.79	null	f1be45a5-	08-06-2008
71783	2	01-06-2008	13-06-2008	08-06-2008	5	FALSE	SO71783	PO193431	10-4020-0	29957	992	992	CARGO TR	null	83858.43	6708.674	2096.461	92663.56	null	7db2329e	08-06-2008
71784	2	01-06-2008	13-06-2008	08-06-2008	5	FALSE	SO71784	PO192851	10-4020-0	29736	659	659	CARGO TR	null	108561.8	8684.947	2714.046	119960.8	null	ca31f324-	08-06-2008
71796	2	01-06-2008	13-06-2008	08-06-2008	5	FALSE	SO71796	PO170521	10-4020-0	29660	1058	1058	CARGO TR	null	57634.63	4610.771	1440.866	63686.27	null	917ef5ba-	08-06-2008
71797	2	01-06-2008	13-06-2008	08-06-2008	5	FALSE	SO71797	PO165011	10-4020-0	29796	642	642	CARGO TR	null	78029.69	6242.375	1950.742	86222.81	null	bb3fee84-	08-06-2008
71815	2	01-06-2008	13-06-2008	08-06-2008	5	FALSE	SO71815	PO130211	10-4020-0	30089	1034	1034	CARGO TR	null	1141.578	91.3263	28.5395	1261.444	null	2aa5f39b-	08-06-2008
71816	2	01-06-2008	13-06-2008	08-06-2008	5	FALSE	SO71816	PO129921	10-4020-0	30027	1038	1038	CARGO TR	null	3398.166	271.8533	84.9541	3754.973	null	e3c189e7-	08-06-2008
71831	2	01-06-2008	13-06-2008	08-06-2008	5	FALSE	SO71831	PO102951	10-4020-0	30019	652	652	CARGO TR	null	2016.341	161.3073	50.4085	2228.057	null	625d76fc-	08-06-2008
71832	2	01-06-2008	13-06-2008	08-06-2008	5	FALSE	SO71832	PO103531	10-4020-0	29922	639	639	CARGO TR	null	35775.21	2862.017	894.3803	39531.61	null	addb8620	08-06-2008
71845	2	01-06-2008	13-06-2008	08-06-2008	5	FALSE	SO71845	PO269711	10-4020-0	29938	1020	1020	CARGO TR	null	41622.05	3329.764	1040.551	45992.37	null	e68f7ee9-	08-06-2008
71846	2	01-06-2008	13-06-2008	08-06-2008	5	FALSE	SO71846	PO237813	10-4020-0	30102	669	669	CARGO TR	null	2453.765	196.3012	61.3441	2711.41	null	a86d90ad	08-06-2008
71856	2	01-06-2008	13-06-2008	08-06-2008	5	FALSE	SO71856	PO165301	10-4020-0	30033	1090	1090	CARGO TR	null	602.1946	48.1756	15.0549	665.4251	null	05fee073-	08-06-2008

Table: 11 Sample output of executing **silver-to-gold NB**

Step	Action	Remark
Step 1	Goto ADF	
Step 2	You select the Pipeline which you already created before namely '**Look for all Tables'** & '**For Each Schema Table'**	
Step 3	Goto **Manage** Option in the left panel	
Step 4	Under General, you select 'Linked Services' option	
Step 5	In the Linked Service Canvass, click the '+ New' button	
Step 6	Then you get New linked service panel with various 'Data Store' options	
Step 7	In the search button search for' Azure Data Bricks'	
Step 8	Next to Data Store Tab, you see Compute Tab. Click Compute Tab	
Step 9	Select Azure Data Bricks	
Step 10	Click Continue Button at the bottom	
Step 11	Then you get a new linked service panel	
Step 12	In the appropriate Box, Fill-in Data Bricks Workspace as 'Student-sn020-adb4'	Name & Integration Runtime options are already selected/filled-in
Step 13	In the appropriate Box, Fill-in' Select Cluster' as 'Existing Interactive Cluster'	
Step 14	In the appropriate Box, Fill-in Authentication Type as 'Access Token Method'	Using Key Vault, you can provide Access Token
Step 15	You create the Access Token	
Step 16	Data Bricks-User Name (top right) -User settings(click)	
Step 17	In User Setting window, you get Access Tokens and other Options Tabs. Choose Access Token – Generated new Token(button)	
Step 18	Generate new token 'Pop-Up' shows up In that fill suitable comment – Life Time – 90 days(default)	
Step 19	Click Generate-Button at the bottom	
Step 20	Now a New Pop-Up showing 'Your Token Generated Successfully'	See this Token Once. So, keep a copy safely & press Done
Step 21	Use the Just generated Token and add it to Azure Key Vault as a secret	
Step 22	Go to Key vault, select new secret for the Token	
Step 23	Create New Secret – click '+ Generate/import'	
Step 24	Fill following Name = dbwtoken Secret = XXXXXXXXXXX Click Create Button	You copied this in Step 20
Step 25	Pop-up Notification indicating status of our ops – 'Creating The Secret dbwtoken'	

Step 26	Goto ADF	
Step 27	Click Azure Key Vault – Select Secret name – in the drop down you find the just created Token as ‘dbwtoken’	Now we can access the AKV token safely
Step 28	Another option as choose from existing clusters and click the drop down for seeing the already created cluster in Data Bricks. Select Cluster as ‘SAMBATH parthasarathy’ cluster	
Step 29	Test the Connection & Click Create Button	Connection Success message should come
Step 30	You see the Status Notification indicating the progress of Cluster Creation	
Step 31	Click Publish all and **Publish Button** to Save all the changes	
Step 32	You see the pop-up indicating “Publishing Completed”	
Step 33	Goto Author Tab and	
Step 34	We have so far created Copy for All & put in the bronze container	
Step 35	Start NB activity for starting bronze-silver & silver - gold	
Step 36	Goto Author-Factory Resources activities - copy_all_tables - Activities	
Step 37	In search Bar look for Notebook	
Step 38	From two types 1. Synapse and 2 Databricks, choose Data Brick Note Book	
Step 39	Drag the Notebook and put it in Canvass, Click Notebook	
Step 40	Under General Tab Fill in Name Box =’ bronze-to-sliver’	
Step 41	In the pipeline canvas connect the arrow ‘ForEachTable’ to ‘Notebook’	This means Output of ‘ForEach Schema’ activity into input of ‘Bronze to Silver’ NB
Step 42	Next Click ‘Azure Data Bricks Tab	
Step 43	Select the link service connection name as ‘AzureDataBricks1’	This name comes from earlier creation activities
Step 44	Next Click Setting Tab	
Step 45	To Fill Notebook Path, click Browse button next to the Box	
Step 46	Browse Page Opens and you see Root Folder with 1. Repos 2. Shared 3. Users listed	
Step 47	Click 2. Shared	
Step 48	Next you can see three folders 1. Bronze to silver 2. Silver to gold 3. Storage mount	
Step 49	Choose the 1. BronzetoSilver	
Step 50	Goto Author-Factory Resources activities - copy_all_tables - Activities	
Step 51	In search Bar look for Notebook	

Step 52	From two types 1. Synapse and 2 Databricks, choose Data Brick Note Book	
Step 53	Drag the Notebook and put it in Canvass, Click Notebook	
Step 54	Under General Tab Fill in Name Box =' sliver-to-gold'	
Step 55	In the pipeline canvas connect the arrow 'ForEachTable' to 'Notebook'	
Step 56	Next Click 'Azure Data Bricks Tab	
Step 57	Select the link service connection name as 'AzureDataBricks1'	
Step 58	Next Click Setting Tab	
Step 59	To Fill Note Book Path, click Browse button next to the Box	
Step 60	Browse Page Opens and you see Root Folder with 1. Repos 2. Shared 3. Users listed	
Step 61	Click 2. Shared	
Step 62	Next you can see three folders 1. Bronze to silver 2. Silver to gold 3. Storage mount	
Step 63	Choose the 2. Silver to gold and click on OK button	
Step 64	After the above step in the ADF Workspace Canvas, you will see Four Tasks connected as pipeline. And the last two tasks on the right are Data Bricks Notebooks	
Step 65	To click on the top of publishing button and you will see a new page mane as publish all. click publish button.	
Step 66	You see the pop-up indicating "Publishing Completed"	
Step 67	The top of option as **validate, Debug, Add trigger.**	
Step 68	Click the Add trigger button and select as Trigger now. You will see a new page as pipeline run. click the **ok button**	Observe the Pop-Up window on Top-Right, saying" Running"
Step 69	Goto Monitor Tab in the left Panel, you will see "Activity Runs" Canvas at the bottom half of screen -showing Activity Details	
Step 70	Hit Refresh Option at the Top to see the current running progress status	
Step 71	Go to Storage Account- bronze-container-SalesLT to check the data	New Parquet file is loaded & it is copied recently
Step 72	IN real time we could monito Data Bricks	
Step 73	Against activity – there is an eye-glasses -icon – click that to see Details	
Step 74	A window opens and provide the link. If you open the link, you can see the progress of Data transfer done by the Note Book code	This is helpful in monitoring the runs in Real Time

		If some bug, we will come to know which cell is giving problem
Step 75	Go to Azure Data Bricks Studio. Refresh Azure pipeline again	
Step 76	Like we checked silver container in Step 71, we can do the same for silver container	By doing this you can see the 2 parts of parquet file – since we processed the same data twice
Step 77	Go to ADF studio – Refresh – Pipeline run would have been successful. The data has been transferred to gold container in the most curated form.	Because we use **Delta format** – only data recently added will be seen
Step 78	Thus, you can use ADF& Data Bricks to Transform the data	

Table 8:TASK- 6 – DATA TRANSFORM 3

Step	Action	Remark
	Creating Synapse Analytics Workspace Resource(Steps 1- 10)	
Step 1	Go to azure Resource group name as student-rg-n3.	First few steps for creating Synapse are from WAFA
Step 2	On top of screen there is an option button - +create. Click that	
Step 3	You can see a New Page named as Market Place	
Step 4	Search for Azure Synapse Analytics Workspace	
Step 5	New page titled Create Synapse workspace opens	
Step 6	It has Options in the form of Tabs such as Basics Security…	
Step 7	In Basics Fill all the necessary boxes (indicated by *) Subscription= your sub, Resource Group = student-rg-n3, Workspace name =**student-sn020-asws** Region = South India Storage Account Name =studentsn020sa4 filesystem name = bronze	
Step 8	Click the Review+Create Button	Perform validation – once successful – it will
Step 9	Click Create button	
Step 10	Now a Pop-up comes up indicating the Successful Deployment completion	
Step 11	Goto Synapse Workspace & click open ‘student-sn020-asws’	
Step 12	New page titled **student-sn020-asws** opens	
Step 13	Below the Getting Started line, there is Box titled ‘Open Synapse Studio’ Click the Box/Open button	
Step 14	Now a new Tab-page titled Synapse Analytics Workspace - **student-sn020-asws** opens	Synapse Is built on top of ADF – Synapse looks similar to ADF
Step 15	Now create a Database in Azure synapse analytics.	
Step 16	Select a data tab in left side panel.	
Step 17	Open a new panel name as data and click the **+button**	
Step 18	To click and create a SQL database	
Step 19	Go to manage tab and click the SQL pool button and you can see your SQL pool status is online	Then you will check your build-in are status are offline or online
Step 21	Opens a new column panel in right side, with name as ‘Create SQL database’	The serverless SQL

		database uses the build-in pool to process the data
Step 22	Undersee Select SQL pool type, Serverless is pre-selected And leave that as it is	
Step 23	Fill the DB Name in the Box as 'gold_db'	
Step 24	Click Create Button	DB will get created
Step 25	Go to left panel name 'Data'	
Step 26	Click the just-created 'gold_db'	
Step 27	Look for drop-down menu with options such as External Table External Resources Views Schemas Security	Serverless data will be in Data Lake We will be using built-in SQL to query
Step 28	We know Data Lake has direct connection to Synapse	If I switch from Workspace to link, you will see Data Lake, it is already linked to Synapse
Step 29	If you click 'student-sn020-asws' Drop down menu will show containers we created Bronze gold silver other data inside gold container can checked	Query data lake from synapse serverless SQL
Step 30	Go to gold-SalesLT- Select address & right click You can see a new panel named as 'Select Top 100 rows Here you can choose the file type format as Delta Click on Apply Button	<u>Feature demo</u> Go to gold-SalesLT- Select address & right click You can see a new panel named as 'Select Top 100 rows Here you can choose the file type

		format as Delta Click on Apply Button

Step 31	New Tab/page Titled SQL script1 gets opened
	(a script is created for you) sample given below
	
	`CREATE VIEW address`
	`AS`
	`SELECT`
	`*`
	`FROM`
	`OPENROWSET(`
	`BULK`
	`'https://studentsn020sa4.dfs.core.windows.net/gold/SalesLT/Address/`
	`',`
	`FORMAT = 'DELTA'`
	`) AS [result]`	

Step 32	Click the >Run option	
Step 33	It runs the script and shows the output in Result Tab in the bottom portion of the screen	

Step 34	Now modify the generated script as
	`CREATE VIEW address`
	`AS`
	`SELECT`
	`*`
	`FROM`
	`OPENROWSET(`
	`BULK`
	`'https://studentsn020sa4.dfs.core.windows.net/gold/SalesLT/Address/`
	`',`
	`FORMAT = 'DELTA'`
	`) AS [result]`	

Step 35	Run the Query and ensure it runs Success	
Step 36	Then switch-back to WS and refresh the database. If then expand the view to see an address view if we created recently.	
Step 37	See (…) next to dbo. address & Click (…)	
Step 38	New SQL scripts show in work-area	
Step 39	Choose New SQL script > Select TOP 100 Rows	
Step 40	New Script/Query displayed in the Work-area	
Step 41	➢ Run the thus generated Query	
Step 42	At the bottom-half of screen, Address Table is displayed	When the data changes in the Data Lake, this view automatically changes Cool
Step 43	We need to do this for all Tables by creating a pipeline using Synapse Analytics	Pipeline can be created in

		Synapse or in ADF
Step 44	Go to left panel Select Develop Tab	
Step 45	You see Develop Option & a + button next to that	
Step 46	Click on the + Button	

Step 47	There will be an option called import the following looking Query (stored procedure) from the local folder
	`USE gold_db`
	`GO`
	`CREATE OR ALTER PROCEDURE CreateSQLServerlessView_gold`
	`@ViewName NVARCHAR(100)`
	`AS`
	`BEGIN`
	`DECLARE @statement NVARCHAR(MAX)`
	`SET @statement = N'CREATE OR ALTER VIEW ' + QUOTENAME(@ViewName) + N' AS`
	`SELECT * FROM`
	`OPENROWSET(`
	`BULK ''https://studentsn020sa4.dfs.core.windows.net/gold/SalesLT/' + @ViewName + N'/'' ,`
	`FORMAT = ''DELTA''`
	`) AS [result]'`
	`EXEC (@statement)`
	`END`
	`GO`	

Step 48	Run the Stored Procedure Query & ensure it is run to Success	
Step 49	No Publish the New Script and other changes.	Look for Successful Publish Status on Top Corner
Step 50	Goto Manage Tab. Select the Link Service connection. Click on New Button	
Step 51	You can see New Page Named “ New Linked Service”	
Step 52	Search for “Azure SQL DB”	
Step 53	Select Azure SQL DB and Continue	
Step 54	New Page opens with Title “New linked service”, Azure SQL DB Fill all Boxes correctly	
Step 55	Name as “ serverlesssqldb” Account Selection = Fully qualified domain = Database name = gold_db Authentication Type = System Assigned Managed Identity	It will use sambath. Narayanan account to connect to this DB
Step 51	To get Fully qualified domain, do the following	
Step 52	Goto Synapse WS>Properties>Serverless SQL endpoint copy the endpoint name Paste this endpoint name in the box in Step 55	
Step 53	After completing all Boxes in Step 55, check the test connection	
Step 54	The connection is successful and click the create button. **serverlessSQLbd** will be created	
Step 55	Click the publish button. Publish completed	

Step 56	Go to integrate in left side panel and click the **+** button. Ten you will click a pipeline button.	
Step 57	Then you will see a sub panel name as Activities	
Step 58	If then you search a get "**Metadata** and drag and drop the get Metadata in canvas ". if the below options are general, setting, user properties.	
Step 59	To give a name as Get **Tablenames** and go to settings select new button.	
Step 60	Then you see a new panel name as new integration dataset. If you search **Azure data lake storage gen2** select and click the continue button	
Step 61	After then you see a new panel name as Select Format. if you will select the **Binary** and click the continue button.	
Step 62	To see a new panel name as set properties then you give a name as **gold tables**	
Step 63	Then you will see a linked service connection click the drop-down option as default linked service connection name as '**student-sn020-asws'**	
Step 64	Now we will select the browser. If you see a new panel as Browse and select the gold and click the continue button. Now we will create	
Step 65	Then fill the field list as click as **+new** and the drop-down box is filling name as **child items**	
Step 66	To click the "debug" button and you can see an output as successful	
Step 67	At the activity panel as search a **for each** activity and drag and drop the canvas. If then we will connection of get Table names to for each table.	
Step 68	To modified name as for each table name and go to setting tab. Then you see options are item. Select the box of "ADD dynamic content".	
Step 69	If you can see A new page name as pipeline expression builder a below options are Get Table name select and modified the query as **@activity('Get Tablenames').output.childitems** To click the ok button	
Step 70	After you can see a for each table in canvas at page and see a small **pencil** icon clicked an if show a **new canvas page**	
Step 71	Click the activity search bar search a **Stored procedure** and drag and drop of the canvas page	
Step 72	Go to settings tab below option select a linked service dropdown of box as click the **serverlessSQLdb**	
Step 74	Select a Store procedure name click the drop down showing as, if we already created SQL **serverlessview_gold**	
Step 75	To select a store procedure parameter a click on new button gives a name as View Name another box as type as fill **string**. If you correctly fill it.	
Step 76	To fill value of box as "Add dynamic content" and you can see a new panel name as pipeline expression builder .to select below the option as for 'each table name' and modified the query as **@item().name**	

	To click on ok button	
Step 77	To give a pipeline name as **create view**	
Step 78	To click a publish button. publish completed	
Step 79	To click on this, add trigger and select a Trigger now an click the ok button	
Step 80	Go to monitor tab in the left side pencil. If then you can see a pipeline is currently running. if they click on this refresh button	
Step 81	The pipeline is running successfully	
Step 82	Go to data tab in left side panel an select a gold_db (…) a refresh the gold_db and you can see a view table as drop down as generated a view of all tables	

Table 9: TASK-7 – DATA Load (Azure Synapse Analytics)

Step	Action	Remark
	Installing Power BI on-prem	
Step 1	Check whether you have Power BI installed on your system. If installed, uninstall and do a fresh install.	Power BI tool provided by MS – can create interactive dashboards and visuals
Step 2	Go to Chrome browser and search for power BI	
Step 3	Then you can see a power bi-data visualization and click it.	
Step 4	If you will be seeing a new page name as Microsoft power BI platform	
Step 5	You can see as options are power bi, product, etc.…	
Step 6	Click the products and drop down of options as power bi	
Step 7	Select the power bi and click the Desktop	
Step 8	Then you can see a new page as power bi Desktop	
Step 9	You can click the option as see download or language options	
Step 10	Then you can see a new page as Microsoft Power BI Desktop	
Step 11	They You can click the Download button. if you can see a new sub tab name, as Choose the download you want	
Step 12	You can see options are **PBIDesktopSetup.exe, PBIDesktopSetup_x64.exe**	
Step 13	You can select **PBIDesktopSetup_x64.exe** and download **button** **PBIDesktopSetup_x64.exe as Downloading completed**	
Step 14	You can search and select Power BI click and install	
Step 15	Go to Power BI Desktop	
Step 15	If they you can see a Power BI Desktop	The using of power BI desktop to connect with the serverless SQL database to get all the data
Step 16	You can see a Home tab in the top of options. you can click on the **Get data** and dropdown options are **excel workspace, power bi datasets, etc..**	
Step 17	Click the dropdown option as click on more button.	
Step 18	If you can see a new tab name, as Get Data and you. we see a more option is All, File, Azure, etc…	
Step 19	Click the Azure button and you can see a new panel name as Azure	
Step 20	Click the azure and you will see right side of sub tab azure synapse work analytics SQL	
Step 21	Then you can select a dropdown box as connect button click	
Step 22	If you can see a new sub tab name as SQL server data base Fill the box	server name as end to end point
Step 23	Go to synapse workspace click left side panel name as properties then you can dropdown small box name as sqlserverless SQL end point and copy the end point	
Step 24	Go to power bi and past the end point as server box and next box name as database	gold_db
Step 25	If you fill the box as gold_db	
Step 26	After you can see adventure worksLT2019 tables	
Step 27	Click all tables and dropdown button name as load	

Step 28	AdventureworksLT2019 data be loaded in power bi	
Step 29	If then create a dashboard	
Step 30	You can see as various options right side as visualizations	
Step 31	Click the card. Select product-id and see a total count of product will be show	
Step 32	Click the card. Select sum of sub total can be showing	
Step 33	Click the card. Select List price and you will seaing a sum of list price	
Step 34	Click the pie chart. Select list of price and standard price	
Step 35	Click the Donut chart. Select subtotal,Taxamount, Total duean you will see a prices	
Step 36	Click the Gauge. Select sum of Freight an you will see freight	
Step 37	Click the Map. Select some of country, City, Tax amount and see a tax of amount Country	
Step 38	Now visualization as be completed	

Table 10: TASK-8– DATA Reporting using Power BI

11. Bronze to silver Python code to run on Data bricks Notebook

```
# Databricks notebook source

dbutils.fs.ls('mnt/bronze/SalesLT/')

# COMMAND ----------

dbutils.fs.ls('mnt/silver/')

# COMMAND ----------

input_path = '/mnt/bronze/SalesLT/Address/Address.parquet'

# COMMAND ----------

df=spark.read.format('parquet').load(input_path)

# COMMAND ----------

display(df)

# COMMAND ----------

from pyspark.sql.functions import from_utc_timestamp, date_format

from pyspark.sql.types import TimestampType

df = df.withColumn("ModifiedDate",
date_format(from_utc_timestamp(df["ModifiedDate"].cast(TimestampType
()),"UTC"),"yyyy-MM-dd"))

# COMMAND ----------

display(df)

# COMMAND ----------

# MAGIC %md

# MAGIC ## Doing transformation for all tables

# MAGIC


# COMMAND ----------

 table_name = []

for i in dbutils.fs.ls('mnt/bronze/SalesLT/'):

   table_name.append(i.name.split('/')[0])
```

```
# COMMAND ----------

table_name

# COMMAND ----------

display(df)

# COMMAND ----------

from pyspark.sql.functions import from_utc_timestamp, date_format

from pyspark.sql.types import TimestampType

for i in table_name:

    path = '/mnt/bronze/SalesLT/' + i + '/' + i + '.parquet'

    df = spark.read.format('parquet').load(path)

    column = df.columns

    for col in column:

        if "Date" in col or "date" in col:

            df = df.withColumn(col,
date_format(from_utc_timestamp(df[col].cast(TimestampType()),
"UTC"), "yyy-MM-dd"))

    output_path = '/mnt/silver/SalesLT/' + i + '/'

    df.write.format('delta').mode("overwrite").save(output_path)

# COMMAND ----------

display(df)
```

12. Silver to gold Python code to run on Data bricks Notebook

```
# Databricks notebook source

dbutils.fs.ls('mnt/silver/SalesLT/')


# COMMAND ----------


dbutils.fs.ls('mnt/gold/')


# COMMAND ----------


input_path = ' /mnt/silver/SalesLT/Address/'


# COMMAND ----------


df = spark.read.format('delta').load('/mnt/silver/SalesLT/Address')


# COMMAND ----------


display(df)



# COMMAND ----------


from pyspark.sql import SparkSession

from pyspark.sql.functions import col, regexp_replace


column_names = df.columns
```

```
for old_col_name in column_names:

    new_col_name = "".join(["_" + char if char.isupper() and not old_col_name[i - 1].isupper() else char for i, char in enumerate(old_col_name)]).lstrip("_")

    df = df.withColumnRenamed(old_col_name, new_col_name)


# COMMAND ----------


display(df)


# COMMAND ----------


# MAGIC %md

# MAGIC # Doing transformation for all tables (changing column names)

# MAGIC

# MAGIC


# COMMAND ----------


  table_name = []

for i in dbutils.fs.ls('mnt/silver/SalesLT/'):

   table_name.append(i.name.split('/')[0])


# COMMAND ----------


table_name


# COMMAND ----------
```

```
display(df)


# COMMAND ----------


for name in table_name:

    path = '/mnt/silver/SalesLT/' + name

    print(path)

    df = spark.read.format('delta').load(path)

    column_names = df.columns

    for old_col_name in column_names:

        new_col_name = "".join(["_" + char if char.isupper() and
old_col_name[i - 1].isupper() else char for i, char in
enumerate(old_col_name)]).lstrip("_")

        df = df.withColumnRenamed(old_col_name, new_col_name)

    output_path = '/mnt/gold/SalesLT/' + name + '/'

    df.write.format('delta').mode("overwrite").save(output_path)


# COMMAND ----------


display(df)
```

13. Troubleshooting

This solution architecture consists of many software tools and services, spread across on Prem and cloud. In addition, usage of each components involves multiple steps, some times even in the form of pipelines. Due to this reason, we came across, multiple errors. In this section we have presented troubleshooting of few critical errors which are given below the **Table 11**

Description	**Resolution**	**Technology**	**R a**
Error 2*(b)*	Copying few Tables succeeded while other copies failed	**Added the missing @ in SQL script syntax** - We are using SQL scripts to automate. There were some syntax errors.	M H in a ru m
Error 3 *(b)*	The migration process failed throwing error	**Corrected the SQL script syntax error- missing space after FROM** @{concat('SELECT * FROM ', item().SchemaName, '.', item().TableName)}	A D Fa ry S qu fe es sy S
			B n C D pi n A D Fa ry
		(a). The Integration (self-hosted) Run Time node has encountered an error during registration. The integration (Self-hosted) node failed to connect to the cloud service due to network connectivity issues. Check network connectivity issues.	
		(b). Failure happened on 'Source' side. ErrorCode=SqlOperationFailed,'Type=Microsoft.DataTransfer.Common.Shared.HybridDeliveryException,Message=A database operation failed with the following error: 'Incorrect syntax near 'FROMSalesLT'.',Source=,''Type=System.Data.SqlClient.SqlException,Message	

		=Incorrect syntax near 'FROMSalesLT'.,Source=.Net SqlClient Data Provider,SqlErrorNumber=102,Class=15,ErrorCode=-2146232060,State=1,Errors=[{Class=15,Number=102,State=1,Message=Incorrect syntax near 'FROMSalesLT'.,},],'	
		(c).ErrorCode=AdlsGen2OperationFailed,'Type=Microsoft.DataTransfer.Common.Shared.HybridDeliveryException,Message=ADLS Gen2 operation failed for: Operation returned an invalid status code 'Conflict'. Account: 'studentsn020sa3'. FileSystem: 'bronze'. Path: 'SalesLT/item().TableName/item().TableName.parquet'. ErrorCode: 'LeaseNotPresentWithLeaseOperation'. Message: 'The lease ID is not present with the specified lease operation.'. RequestId: '9edb3332-901f-001a-2b6a-8026cd000000'. TimeStamp: 'Wed, 27 Mar 2024 17:17:38 GMT'..,Source=Microsoft.DataTransfer.ClientLibrary,''Type=Microsoft.Azure.Storage.Data.Models.ErrorSchemaException,Message=Operation returned an invalid status code 'Conflict',Source=Microsoft.DataTransfer.ClientLibrary	

	Data Bricks Error Trouble Shoot - AnalysisException: A schema mismatch detected	I had to add this code in the last cell for it get rid of the following error
		
		`# Enable schema migration for other operations`
		`spark.conf.set("spark.databricks.delta.schema.autoMerge.enabled", "true")`
		`------------------`
		AnalysisException: A schema mismatch detected when writing to the Delta table (Table ID: e76fc263-4221-42ea-a6ba-355b2a6ba0e5). To enable schema migration using DataFrameWriter or DataStreamWriter, please set: '.option("mergeSchema", "true")'. For other operations, set the session configuration spark.databricks.delta.schema.autoMerge.enabled to "true". See the documentation specific to the operation for details	

Table 11: Troubleshooting critical errors

14. References

1. https://www.rst.software/blog/introduction-to-data-lakes-how-to-deploy-them-in-the-cloud
2. https://medium.com/@coreytalkscode/data-lakes-a-comprehensive-guide-to-understanding-features-pros-and-cons-for-effective-data-cec7341fbb25
3. https://precog.co/glossary/big-data-in-manufacturing/
4. https://precog.co/glossary/industrial-data-lake/
5. Access Azure Data Lake Storage using Azure Active Directory credential passthrough(legacy) – https://learn.microsoft.com/en-us/azure/databricks/archive/credential-passthrough/adls-passthrough
6. Technology Blogs by SAP – Build an Azure Data Factory Pipeline with the ODBC Driver for ABAP-Frank-Martin https://community.sap.com/t5/technology-blogs-by-sap/build-an-azure-data-factory-pipeline-with-the-odbc-driver-for-abap/ba-p/13612960
7. https://learn.microsoft.com/en-us/training/modules/get-started-with-**power-bi**/?WT.mc_id=powerbi_home_inproduct_introcard
8. Azure data Engineer Associate Certification Guide, Newton Alex, 2022