

Azure Synapse Workspace

Hands on Lab Guide

실습 환경 설정

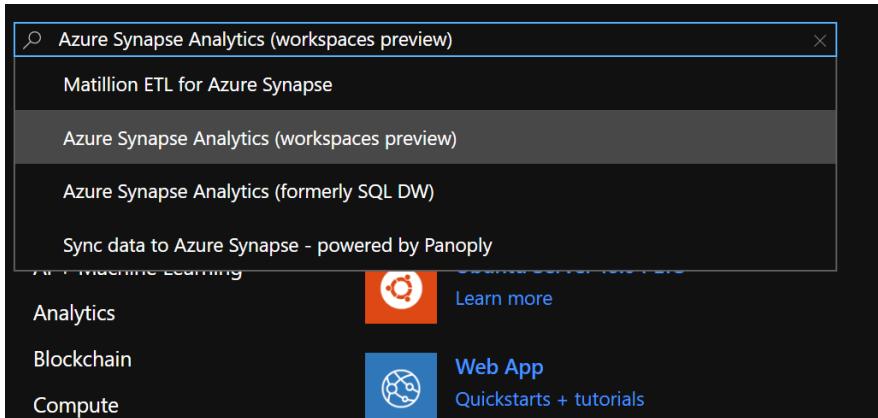
요구 사항

1. Microsoft Azure 구독 (Microsoft 이외의 구독은 유료 구독이어야합니다).
2. **중요** :이 실습의 OAuth 2.0 액세스 구성 요소를 완료하려면 Azure 구독 내에 Azure Active Directory 내에 앱 등록 및 서비스 주체를 만들 수 있는 권한이 있어야합니다.

실습 전에

작업 1 : Azure Synapse Workspace 생성

1. Azure Portal 에 접속하여 Azure Synapse Workspace 를 입력합니다.



2. Region 은 Southeast Asia 를 선택하고 필요한 정보를 입력합니다.

Create Synapse workspace

* Basics * Security + networking Tags Summary

Create a Synapse workspace to develop an enterprise analytics solution in just a few clicks.

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all of your resources.

Subscription * ⓘ Microsoft Azure Internal Consumption

Resource group * ⓘ (New) Kyowon_HOL Create new

Workspace details

Name your workspace, select a location, and choose a primary Data Lake Storage Gen2 file system to serve as the default location for logs and job output.

Workspace name * kyowonsynapse

Region * Southeast Asia

Select Data Lake Storage Gen2 * ⓘ

From subscription Manually via URL

Account name * ⓘ Create new

File system name * ⓘ Create new

3. Data Lake Storage Gen2 를 새로 만들기 위해 Create New 를 선택하고 필요 정보들을 입력합니다.

4. 파일 시스템 명도 마찬가지로 정해주고 아래와 같이 체크박스에 체크합니다

Subscription * ⓘ Microsoft Azure Internal Consumption

Resource group * ⓘ (New) Kyowon_HOL Create new

Workspace details

Name your workspace, select a location, and choose a primary Data Lake Storage Gen2 file system to serve as the default location for logs and job output.

Workspace name * kyowonsynapse ✓

Region * Southeast Asia

Select Data Lake Storage Gen2 * ⓘ From subscription Manually via URL

Account name * ⓘ (New) kyowonadlsgen2 Create new

File system name * ⓘ (New) kyowonsynapse Create new

Assign myself the Storage Blob Data Contributor role on the Data Lake Storage Gen2 account 'kyowonadlsgen2'.

ⓘ We will automatically grant the workspace identity data access to the specified Data Lake Storage Gen2 account, using the Storage Blob Data Contributor role. To enable other users to use this storage account after you create your workspace, you must grant them the Storage Blob Data Reader role.

5. Security + Networking 템에서 어드민 계정 정보를 입력합니다.

* Basics * **Security + networking** Tags Summary

Configure security options and networking settings for your workspace.

SQL administrator credentials

Provide credentials that can be used for administrator access to the workspace's SQL pools. If you don't provide a password, one will be automatically generated. You can change the password later.

Admin username * sqladminuser

Password ✓

Confirm password ✓ Pass

Workspace managed identity

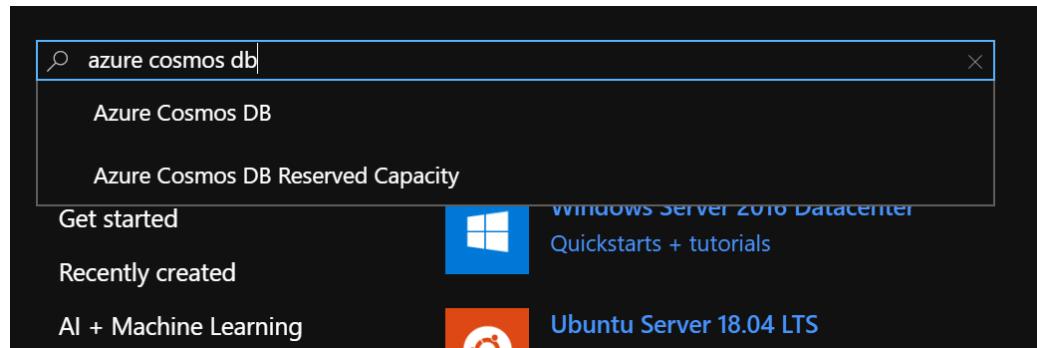
Choose whether you'd like your workspace's managed identity to be automatically granted CONTROL permission (inherits all permissions). This applies when using SQL pools and SQL on-demand within the current workspace.

Grant CONTROL to the workspace's managed identity on all SQL pools and SQL on-demand. ⓘ

6. Review + Create 하여 생성을 완료합니다.

작업 2 : Azure Cosmos DB 생성

1. Create Resource 를 선택하여 Azure Cosmos DB 를 입력합니다.



2. Azure Cosmos DB 를 선택하고 생성합니다.
3. API 는 Core(SQL)을 선택하고, Region 은 Southeast Asia 혹은 Korea Central 을 선택합니다.

Project Details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * Microsoft Azure Internal Consumption

Resource Group * Kyowon_HOL

Create new

Instance Details

Account Name * kyowoncosmos

API * Core (SQL)

Notebooks (Preview) On

Location * (Asia Pacific) Korea Central

With Azure Cosmos DB free tier, you will get 400 RU/s and 5 GB of storage for free in an account. You can enable free tier on up to one account per subscription. Estimated \$24/month discount per account.

Apply Free Tier Discount

Account Type Production

Geo-Redundancy

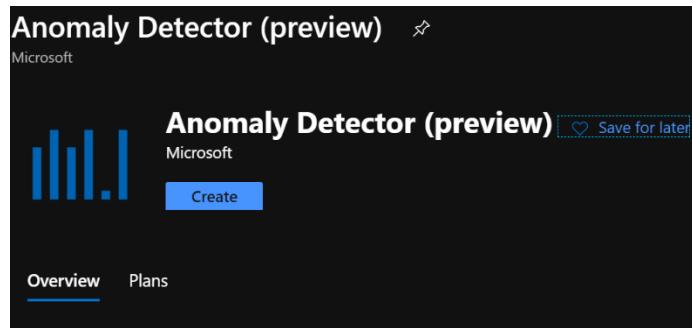
Multi-region Writes

*Up to 33% off multi-region writes is available to qualifying new accounts only. Offer limited to accounts with both account locations and geo-redundancy, and applies only to multi-region writes in those same regions. Both Geo-Redundancy and Multi-region Writes must be enabled in account settings. Actual discount will vary based on number of qualifying regions selected.

- Review + Create 하여 생성을 완료합니다.

작업 3 : Anomaly Detector (Cognitive Service) 생성

- Create Resource에서 Anomaly Detector를 입력하여 설치를 진행합니다.



- 아래와 같이 적절한 정보를 입력하고 생성을 진행합니다.

Name *kyowonanomaly

Subscription *Microsoft Azure Internal Consumption

Location *(US) East US

Pricing tier (View full pricing details) *Free F0 (10 Calls per second, 20K Transactions per month)

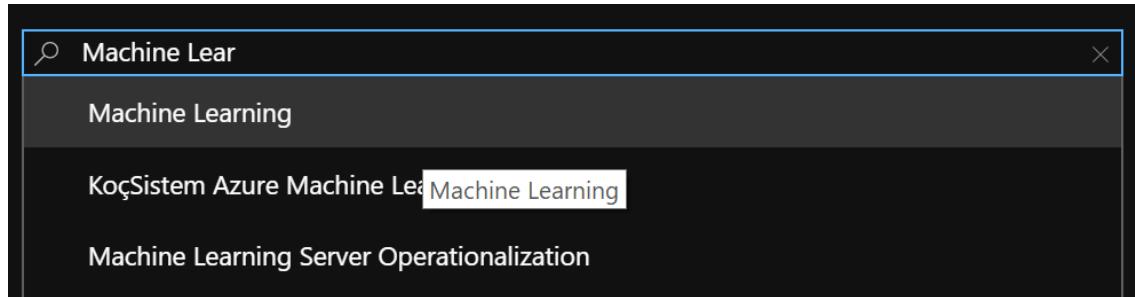
Resource group *Kyowon_HOL
[Create new](#)

I confirm I have read and understood the notice below. *

Previews are made available to you on the condition that you agree to the [Supplemental Terms of Use for](#)

작업 4 : Azure Machine Learning Workspace 생성

1. Create Resource에서 Machine Learning 를 입력하여 설치를 진행합니다.



2. 아래와 같이 적절한 값을 넣고 생성을 진행합니다.

Machine Learning

Create a machine learning workspace

Basics Tags Review + create

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * ⓘ Microsoft Azure Internal Consumption

Resource group * ⓘ Kyowon_HOL

Create new

Workspace details

Specify the name, region, and edition for the workspace.

Workspace name * ⓘ kyowonml

Region * ⓘ Southeast Asia

Workspace edition * ⓘ Basic

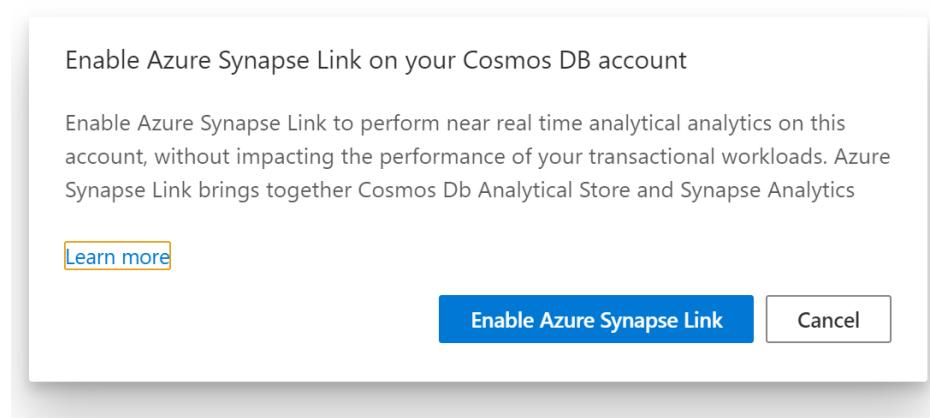
Info For your convenience, these resources are added automatically to the workspace, if regionally available: Azure Storage, Azure Application Insights, Azure Key Vault

실습 랩 가이드

작업 1 : Azure Cosmos DB Enable Synapse Link

1. [Azure Portal](#)에서 Azure Cosmos DB 계정으로 이동한 다음 왼쪽 메뉴에서 Data Explorer를 선택합니다.
2. 아래와 같이 Enable Azure Synapse Link(Preview)를 클릭합니다.

The screenshot shows the Azure Cosmos DB Data Explorer interface. At the top, there's a navigation bar with 'Dashboard > Resource groups > Kyowon_HOL >'. Below it is the account name 'kyowoncosmos | Data Explorer' and 'Azure Cosmos DB account'. On the left, a sidebar lists various options: Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Quick start, Notifications, and Data Explorer (which is selected and highlighted in grey). In the main content area, there's a 'Search (Ctrl+/' input field, a 'New Container' dropdown, and a button labeled 'Enable Azure Synapse Link (Preview)'. This last button is circled in red to indicate it's the target of the first step in the task.



3. Setting 아래의 Feature에서 Azure Synapse Link 항목이 아래와 같이 바뀌면 적용된 것입니다.

Feature	Status
Azure Synapse Link	Enrolled

작업 2 : Database 및 Collection 생성

1. [Azure Portal](#)에서 Azure Cosmos DB 계정으로 이동한 다음 왼쪽 메뉴에서 Data Explorer 를 선택합니다.
2. 아래와 같이 DB ID 를 CosmosDemo 로 입력하고 Autoscale 4000 을 선택, Container ID 를 IoTDeviceInfo 로 입력 후 Partition Key 는 /id 를 입력합니다.

* Database id ⓘ
 Create new Use existing
 CosmosDemo

Provision database throughput ⓘ
 * Throughput (autoscale) ⓘ
 Autoscale Manual
 Provision maximum RU/s required by this resource. Estimate your required RU/s with [capacity calculator](#).

Max RU/s
 4000
 Your database throughput will automatically scale from **400 RU/s** (**10% of max RU/s**) - **4000 RU/s** based on usage.

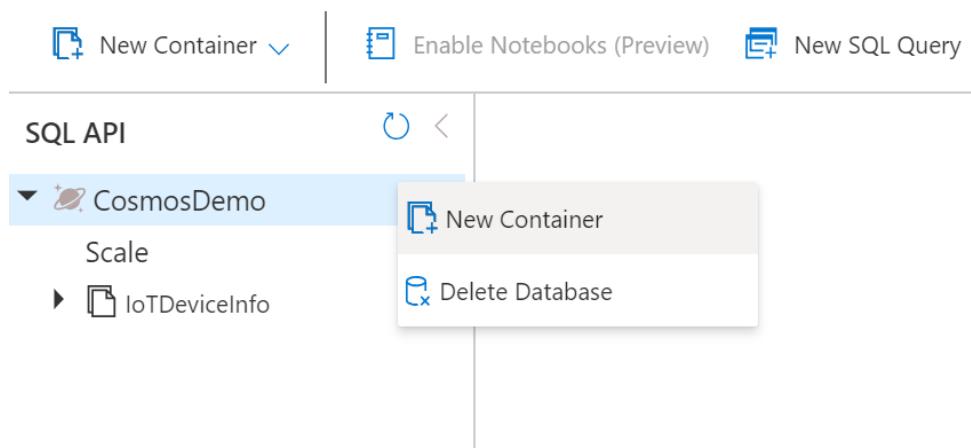
After the first 40 GB of data stored, the max RU/s will be automatically upgraded based on the new storage value. [Learn more](#).

Estimated monthly cost (USD): **\$35.04 - \$350.40** (1 region, 400 - 4000 RU/s, \$0.00012/RU)

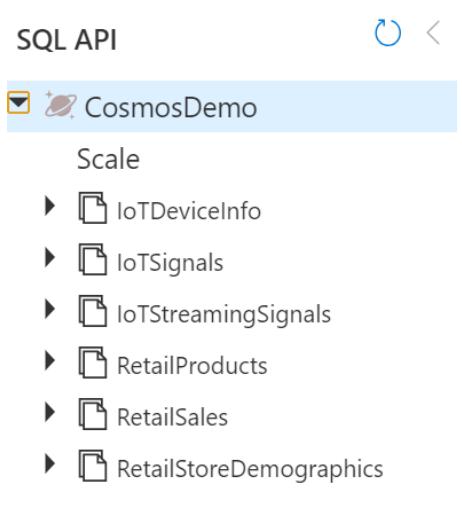
* Container id ⓘ
 IoTDeviceInfo

* Partition key ⓘ
 /id
 My partition key is larger than 100 bytes

3. CosmosDemo DB 를 선택하고 New Container 를 선택합니다.



4. Collection ID 에 IoTSignals 를 입력, Partition Key 에 마찬가지로 /id 를 입력합니다.
5. Collection ID 에 IoTStreamingSignals 를 입력, Partition Key 에 마찬가지로 /id 를 입력하고, 여기서는 Analytical Store 를 Off 를 선택합니다.
6. Collection ID 에 RetailProducts 를 입력, Partition Key 에 마찬가지로 /id 를 입력합니다.
7. Collection ID 에 RetailSales 를 입력, Partition Key 에 마찬가지로 /id 를 입력합니다.
8. Collection ID 에 RetailStoreDemographics 를 입력, Partition Key 에 마찬가지로 /id 를 입력합니다.

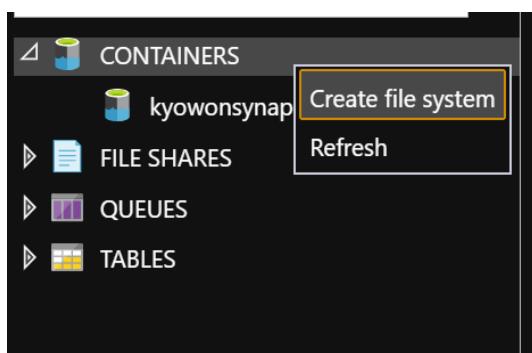


작업 3 : ADLS Gen2

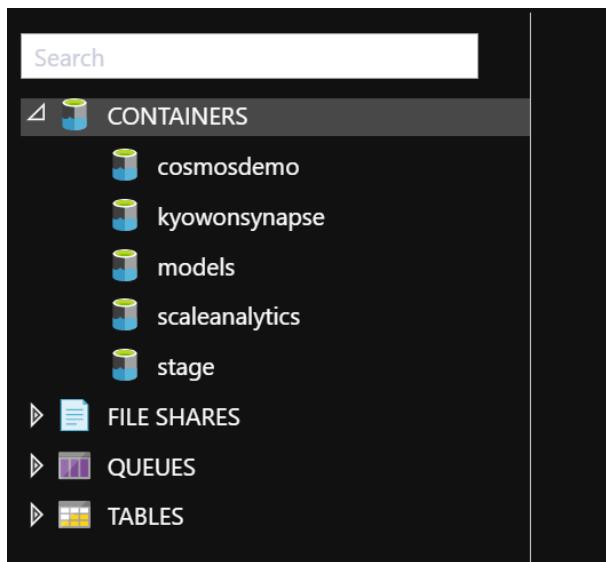
1. [Azure Portal](#)에서 생성된 ADLS Gen2 Storage Account를 선택합니다.
2. 아래와 같이 Storage Explorer를 선택합니다.

The screenshot shows the Azure Storage Explorer (preview) interface. At the top, it displays the path: Dashboard > Resource groups > Kyowon_HOL >. The main title is "kyowonadlsgen2 | Storage Explorer (preview)". On the left, there's a sidebar with a search bar and several navigation items: Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Data transfer, Events, Storage Explorer (preview) (which is highlighted), Settings, and Access keys. The main pane shows a list of storage resources: CONTAINERS, FILE SHARES, QUEUES, and TABLES. Each category has a corresponding icon and a small downward arrow indicating they can be expanded.

3. Container를 선택 후 마우스 우클릭하여 Create File System을 선택합니다.



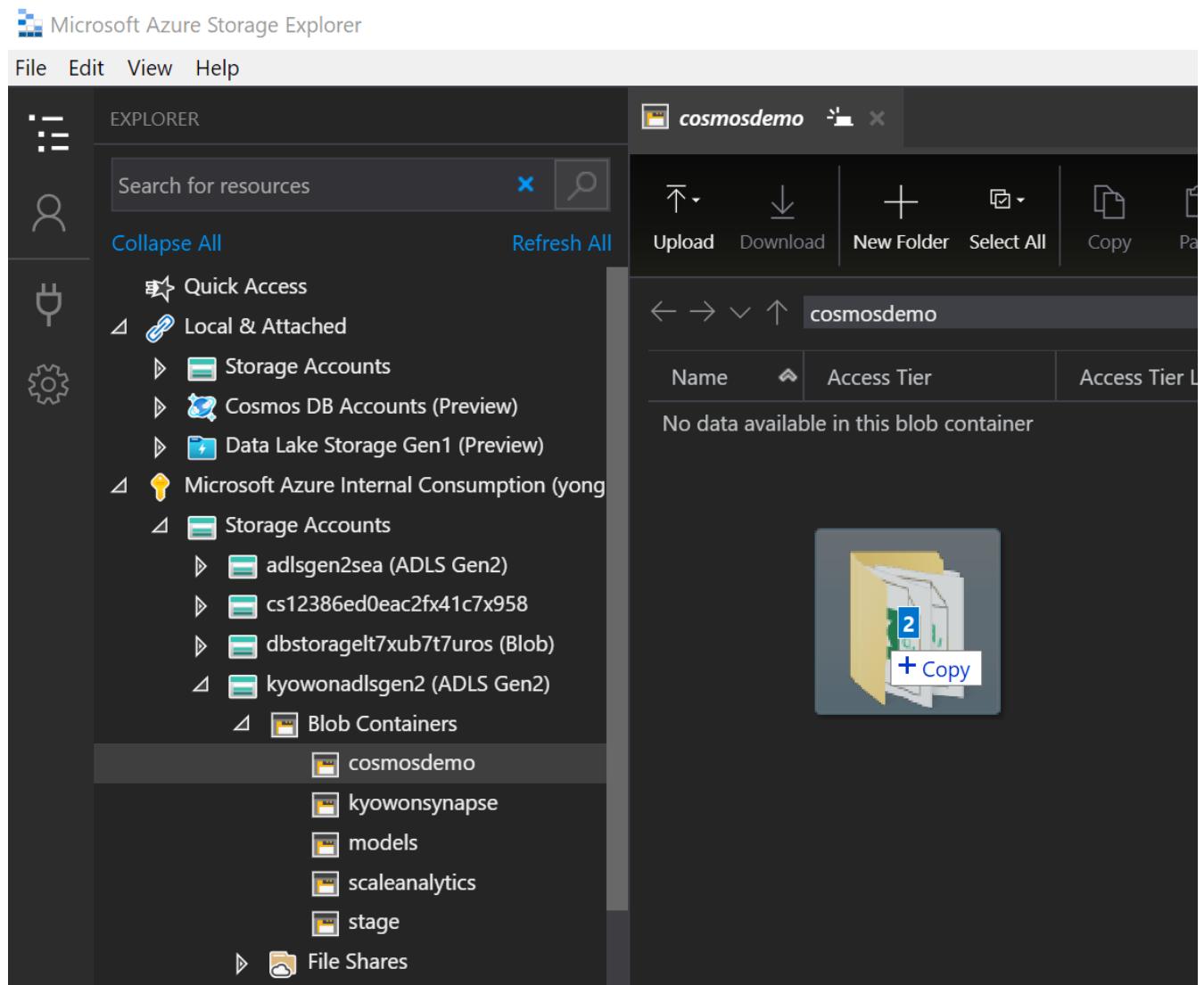
4. cosmosdemo 라고 이름을 입력합니다.
5. 마찬가지로 'models' 컨테이너를 생성합니다.
6. 마찬가지로 'scaleanalytics' 컨테이너를 생성합니다.
7. 마찬가지로 'stage' 컨테이너를 생성합니다.
8. 마지막으로 'wwi-02' 컨테이너를 생성합니다.



9. Upload 버튼을 클릭하고 Storage Explorer 를 다운로드 받습니다.



10. Storage Explorer 를 실행하고 계정 연결 후 ADLS Gen2 를 선택합니다. 깃허브에서 다운로드 받은 Synapse_HOL\Synapse_HOL\LabFiles\ADLSGen2\cosmosdemo 안의 모든 파일을 cosmosdemo 컨테이너 안에 옮겨놓습니다.



11. Scaleanalytics 와 'wwi-02' 컨테이너에도 마찬가지로
Synapse_HOL\Synapse_HOL\LabFiles\ADLSGen2\scaleanalytics 와 wwi-02 안의 모든
파일을 옮겨놓습니다

작업 4 : Synapse Workspace

1. 새로 생성한 Azure Synapse Workspace 를 선택합니다.
2. 왼쪽 메뉴에서 sqlpool 을 선택합니다.

Dashboard > Resource groups > Kywon_HOL >

kyowonsynapse Synapse workspace

Search (Ctrl+ /) + New SQL pool + New Apache Spark pool Refresh

- Overview**
- Activity log
- Access control (IAM)
- Tags

Settings

- SQL Active Directory admin
- Properties
- Locks

Synapse resources

- SQL pools**
- Apache Spark pools

Security

- Firewalls
- Managed identities

Resource group (change) : Kywon_HOL
Status : Succeeded
Location : Southeast Asia
Subscription (change) : Microsoft Azure Internal Cons...
Subscription ID : 2386ed0e-ac2f-41c7-958a-739...
Managed virtual network : No
Managed Identity object ... : ecf88fa-0589-4424-a585-84f...
Workspace web URL : <https://web.azuresynthesize.net?>
Tags (change) : Click here to add tags

Available resources

Search to filter items...
Name
No pools provisioned

3. +New 를 선택하여 SQL Pool 을 새로 생성합니다. 이름을 넣고, DW100c 로 Performance Level 을 조절하여 줍니다.

Create SQL pool

Synapse

* Basics * Additional settings Tags Review + create

Create a SQL pool with your preferred configurations. Complete the Basics tab then go to Review + Create to provision with smart defaults, or visit each tab to customize. [Learn more](#)

SQL pool details

Name your SQL pool and choose its initial settings.

SQL pool name * ✓

Performance level

Estimated price ⓘ **Est. Cost Per Hour**
1349.58 KRW [View pricing details](#)

✓ Database
✓ The value
✓ Database
✓ No database

4. 마저 생성을 진행합니다.
5. 마찬가지로 왼쪽 메뉴에서 Apache Spark Pool 을 선택하고, +New 를 클릭합니다.
6. 아래와 같이 이름을 기입하고 Performance Level 을 조절합니다.

* Basics * Additional settings Tags Summary

Create a Synapse Analytics Apache Spark pool with your preferred configurations. Complete the Basics tab then go to Review + create to provision with smart defaults, or visit each tab to customize.

Apache Spark pool details

Name your Apache Spark pool and choose its initial settings.

Apache Spark pool name *

sparkpool

Node size family

MemoryOptimized

Node size *

Medium (8 vCPU / 64 GB)

Autoscale *

Enabled

Number of nodes *

3

Estimated price

Est. cost per hour
4588.57 to 7647.62 KRW

⚠️ Contact an **Owner** of the storage account, and verify that the following role assignments have been made:

- Assign the workspace MSI to the **Storage Blob Data Contributor** role on the storage account
- Assign you and other users to the **Storage Blob Data Contributor** role on the storage account

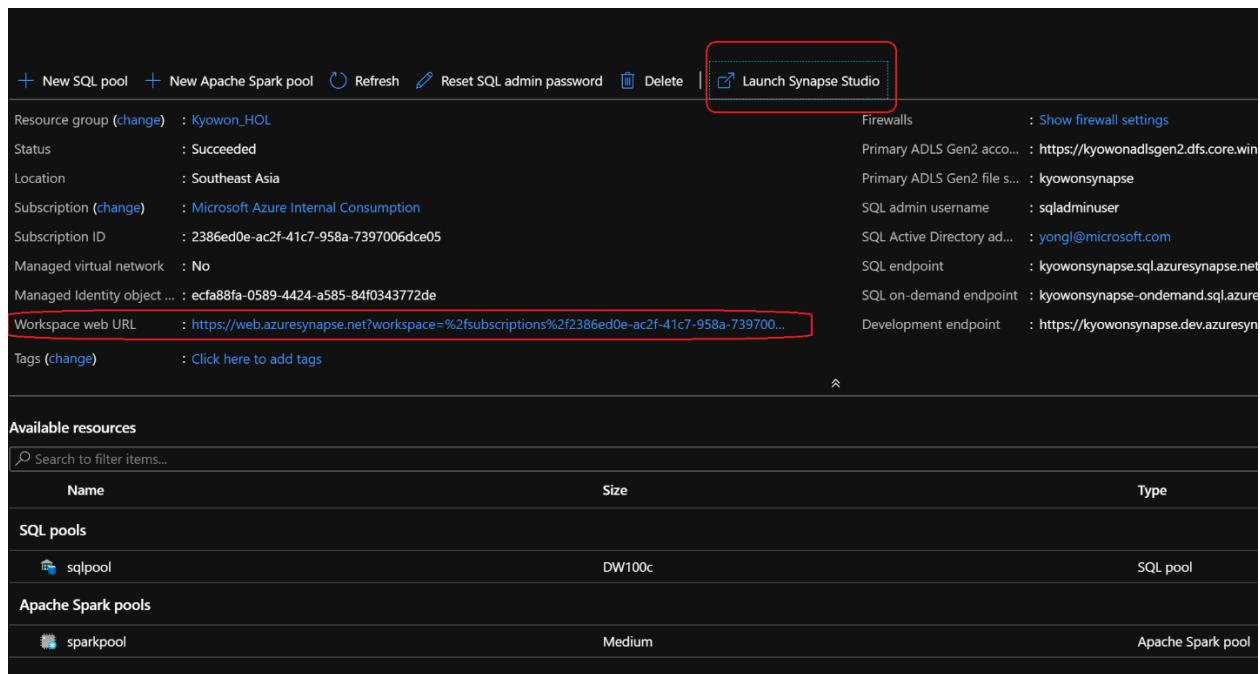
Once those assignments are made, the following Spark features can be used:
(1) Spark Library Management, (2) Read and Write data to SQL pool databases via the Spark SQL connector, and (3) Create Spark databases and tables

[Learn more](#)

7. Review + Create 하여 마저 생성을 진행합니다.

작업 5 : Synapse Studio SQL Analytics

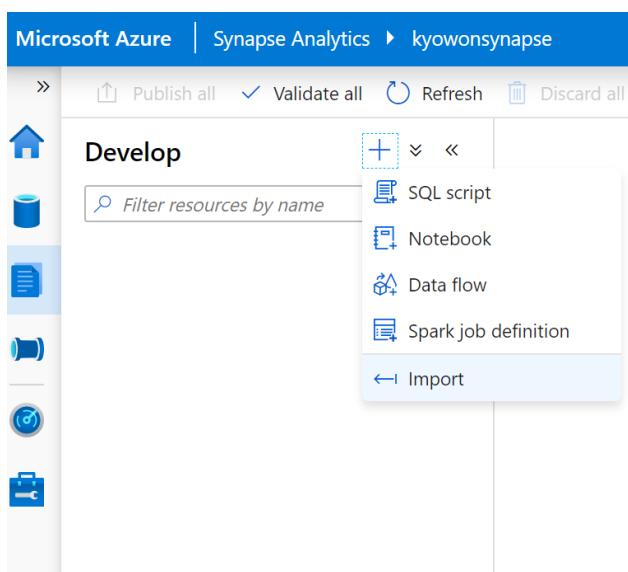
이번 작업에서는 MySQL Server VM(On-Premise로 가정)에 있는 DW Table 들의 Data 를 (**OLAP Data**) Migration 합니다. Azure Synapse Workspace Overview Page 에서 아래와 같이 Link 를 클릭하여 Synapse Studio 를 실행합니다.



The screenshot shows the Azure Synapse Workspace Overview page. At the top right, there is a 'Launch Synapse Studio' button, which is highlighted with a red box. Below the button, there is a list of workspace settings and a 'Available resources' section. The 'Available resources' section includes a table with columns for Name, Size, and Type, listing a SQL pool and an Apache Spark pool.

Name	Size	Type
sqlpool	DW100c	SQL pool
sparkpool	Medium	Apache Spark pool

1. 왼쪽 메뉴의 Develop 템을 선택하고 +를 선택하여 Import 를 실행합니다.



The screenshot shows the Microsoft Azure Synapse Analytics workspace. On the left, there is a vertical navigation bar with icons for Home, Databases, Pipelines, Data Flows, and Jobs. The 'Develop' icon is selected. In the center, there is a 'Develop' menu with options: '+', '<<', 'SQL script', 'Notebook', 'Data flow', 'Spark job definition', and 'Import'. The 'Import' option is highlighted with a blue box.

2. 다운로드 받은 Synapse_HOL\#Synapse_HOL\#LabFiles\#Scripts\#SQL 안의 SQL Script들을 선택하여 불러옵니다.
3. Publish All 을 클릭하여 Publish 합니다.

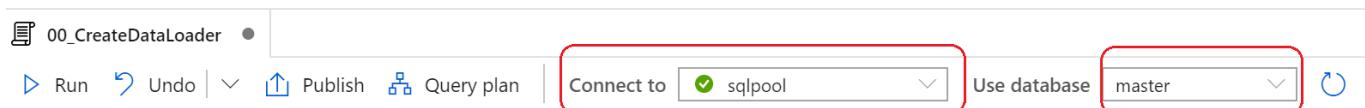
Publish all

You are about to publish all pending changes to the live environment. [Learn more](#)

Pending changes (7)

NAME	CHANGE	EXISTING
▲ SQL script		
00_CreateDataLoader	(New)	-
01_CreateDWTables	(New)	-
02_CreateCredential	(New)	-
03_CreateExternalTables	(New)	-
04_CreateTransactionTables	(New)	-
05_LoadData	(New)	-
06 OLTP_Table_Count	(New)	-

4. **(Optional for SSMS or Azure Data Studio)** Data Loader 계정 생성을 위해 00_CreateDataLoader 스크립트를 실행합니다. 먼저 sqlpool 을 연결하고, 실행할 때 위의 두 줄은 master DB에서 실행하고, 아래는 sqlpool 을 선택하여 실행합니다.

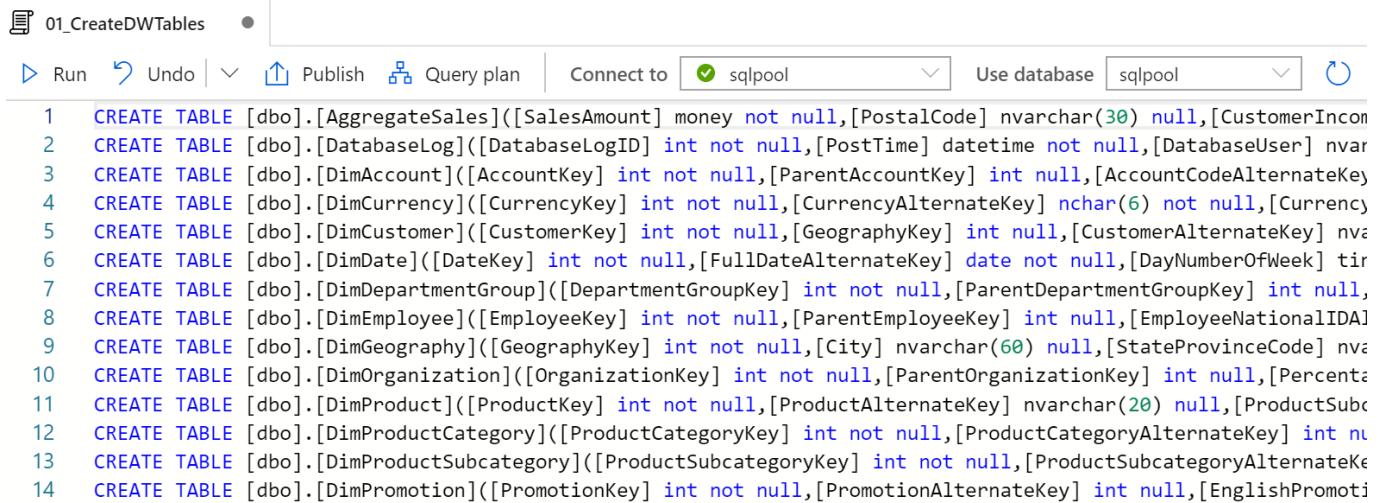


```

00_CreateDataLoader •
Run Undo | Publish Query plan | Connect to sqlpool | Use database master
1 --Execute this line from the master database
2 CREATE LOGIN dataloader WITH PASSWORD = 'Demo@pass123';
3 CREATE USER dataloader FOR LOGIN dataloader;
4
5 --Execute the remainder of these lines from the sqlpool database
6 CREATE USER dataloader FOR LOGIN dataloader;
7 GRANT CONTROL ON DATABASE::sqlpool TO dataloader;
8 EXEC sp_addrolemember 'largerc', 'dataloader';
9

```

5. 다음으로 MySQL Server VM 의 DW 테이블들의 데이터 이관을 위해 MySQL DW Table 스키마에 맞춰 SQL Pool 안에 테이블을 생성합니다. 01_CreateDWTables 스크립트를 마찬가지로 sqlpool 에 연결하고, sqlpool database 를 선택하고 실행합니다.



```

1  CREATE TABLE [dbo].[AggregateSales]([SalesAmount] money not null,[PostalCode] nvarchar(30) null,[CustomerIncon
2  CREATE TABLE [dbo].[DatabaseLog]([DatabaseLogID] int not null,[PostTime] datetime not null,[DatabaseUser] nvar
3  CREATE TABLE [dbo].[DimAccount]([AccountKey] int not null,[ParentAccountKey] int null,[AccountCodeAlternateKey]
4  CREATE TABLE [dbo].[DimCurrency]([CurrencyKey] int not null,[CurrencyAlternateKey] nchar(6) not null,[Currency
5  CREATE TABLE [dbo].[DimCustomer]([CustomerKey] int not null,[GeographyKey] int null,[CustomerAlternateKey] nva
6  CREATE TABLE [dbo].[DimDate]([DateKey] int not null,[FullDateAlternateKey] date not null,[DayNumberOfWeek] tir
7  CREATE TABLE [dbo].[DimDepartmentGroup]([DepartmentGroupKey] int not null,[ParentDepartmentGroupKey] int null,
8  CREATE TABLE [dbo].[DimEmployee]([EmployeeKey] int not null,[ParentEmployeeKey] int null,[EmployeeNationalIDA]
9  CREATE TABLE [dbo].[DimGeography]([GeographyKey] int not null,[City] nvarchar(60) null,[StateProvinceCode] nva
10 CREATE TABLE [dbo].[DimOrganization]([OrganizationKey] int not null,[ParentOrganizationKey] int null,[Percenta
11 CREATE TABLE [dbo].[DimProduct]([ProductKey] int not null,[ProductAlternateKey] nvarchar(20) null,[ProductSubc
12 CREATE TABLE [dbo].[DimProductCategory]([ProductCategoryKey] int not null,[ProductCategoryAlternateKey] int nu
13 CREATE TABLE [dbo].[DimProductSubcategory]([ProductSubcategoryKey] int not null,[ProductSubcategoryAlternateKe
14 CREATE TABLE [dbo].[DimPromotion]([PromotionKey] int not null,[PromotionAlternateKey] int null,[EnglishPromoti

```

6. 왼쪽 메뉴에서 Data 탭을 선택하고, Workspace 탭의 sqlpool에서 테이블이 생성되었는지 확인합니다.

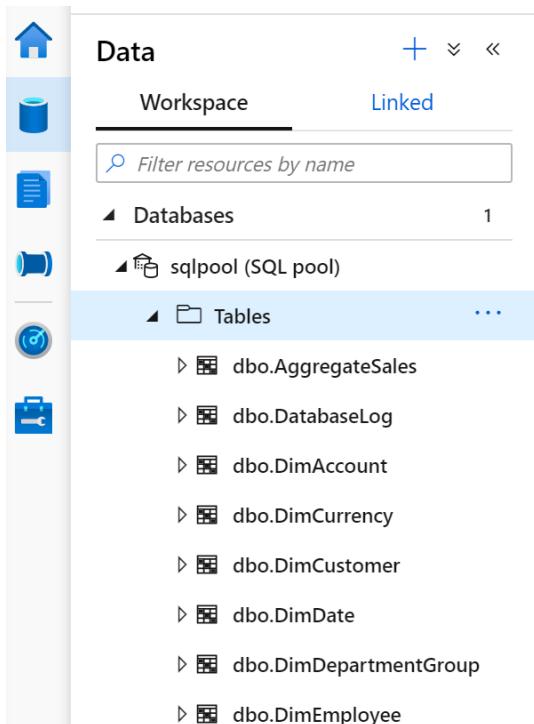
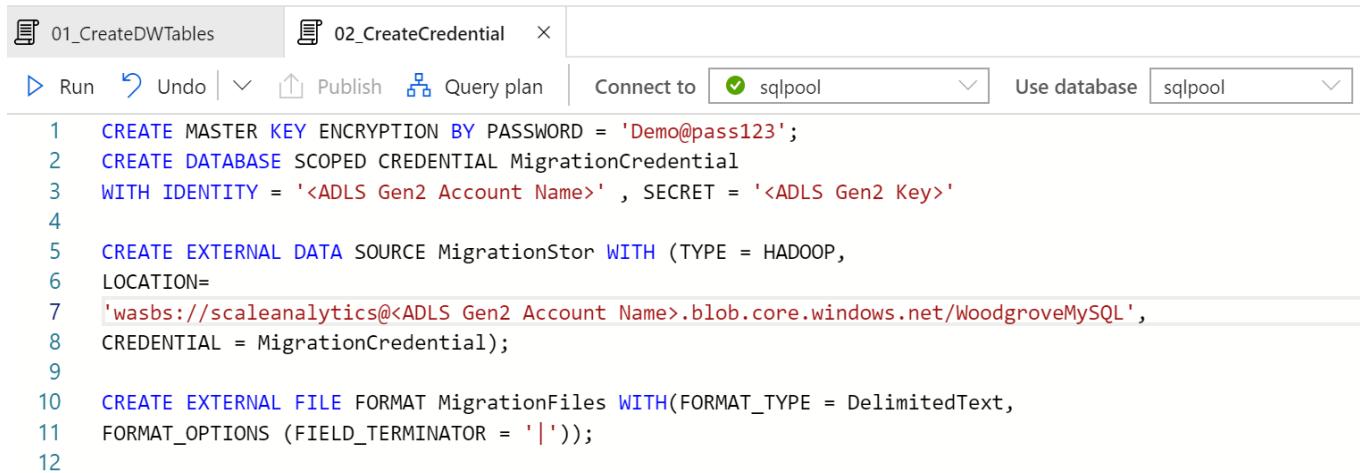


Table Name
dbo.AggregateSales
dbo.DatabaseLog
dbo.DimAccount
dbo.DimCurrency
dbo.DimCustomer
dbo.DimDate
dbo.DimDepartmentGroup
dbo.DimEmployee
dbo.DimGeography
dbo.DimOrganization
dbo.DimProduct
dbo.DimProductCategory
dbo.DimProductSubcategory
dbo.DimPromotion

7. 다시 Develop 탭으로 돌아와서, 02_CreateCredential 스크립트를 선택합니다. MySQL DW Table에서 Export 받은 파일들을 ADLS Gen2의 'scaleanalytics' 컨테이너 안에 WoodgroveMySQL 폴더 안에 옮겨놓았습니다. 여기에 접근하기 위한 Master Key, Credential, Data Source 와 파일 포맷을 정의하여 줍니다.

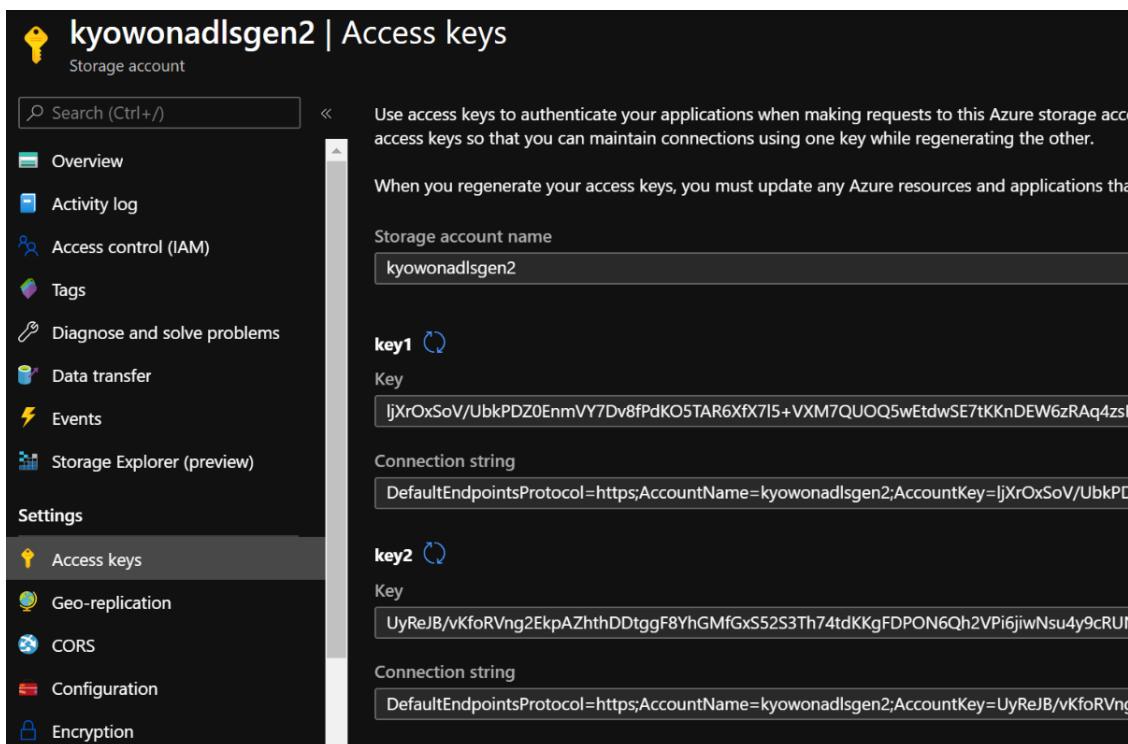


```

1 CREATE MASTER KEY ENCRYPTION BY PASSWORD = 'Demo@pass123';
2 CREATE DATABASE SCOPED CREDENTIAL MigrationCredential
3 WITH IDENTITY = '<ADLS Gen2 Account Name>' , SECRET = '<ADLS Gen2 Key>'
4
5 CREATE EXTERNAL DATA SOURCE MigrationStor WITH (TYPE = HADOOP,
6 LOCATION=
7 'wasbs://scaleanalytics@<ADLS Gen2 Account Name>.blob.core.windows.net/WoodgroveMySQL',
8 CREDENTIAL = MigrationCredential);
9
10 CREATE EXTERNAL FILE FORMAT MigrationFiles WITH(FORMAT_TYPE = DelimitedText,
11 FORMAT_OPTIONS (FIELD_TERMINATOR = '|'));
12

```

8. ADLS Gen2 Account Name과 Key를 확인하기 위해 Azure Portal의 Storage Account를 선택 후 Access Key 탭에서 확인하여 입력 후 실행합니다.



Access Key	Key	Connection String
key1	ljXrOxSoV/UbkPDZ0EnmVY7Dv8fPdKO5TAR6Xfx7I5+VXM7QUOQ5wEtdwSE7tKKnDEW6zRAq4zsF	DefaultEndpointsProtocol=https;AccountName=kyowonadlsgen2;AccountKey=ljXrOxSoV/UbkPDZ0EnmVY7Dv8fPdKO5TAR6Xfx7I5+VXM7QUOQ5wEtdwSE7tKKnDEW6zRAq4zsF
key2	UyReJB/vKfoRVng2EkpAZhthDDtggF8YhGMfGxS52S3Th74tdKKgFDPON6Qh2VPi6jiwNsu4y9cRUN	DefaultEndpointsProtocol=https;AccountName=kyowonadlsgen2;AccountKey=UyReJB/vKfoRVng2EkpAZhthDDtggF8YhGMfGxS52S3Th74tdKKgFDPON6Qh2VPi6jiwNsu4y9cRUN

9. 다음으로, 03_CreateExternalTables 스크립트를 실행하여 Export 받은 파일들을 직접 Link 하는 External Table 들을 생성합니다.
10. 왼쪽 메뉴에서 Data 탭을 선택하고, Workspace 탭의 sqlpool에서 External Table들이 생성되었는지 확인합니다. 일부 테이블에 쿼리를 실행하여 조회가 되는지 확인합니다.
11. 04_LoadData 스크립트를 실행하여, Migration 을 진행합니다.
12. 왼쪽 메뉴에서 Data 탭을 선택하고, Workspace 탭의 sqlpool에서 Table들의 Data가 잘 Migration 되었는지 일부 테이블에 쿼리를 실행하여 확인합니다.

작업 6 : Azure Synapse Ingest Wizard

이번 작업에서는 MySQL Server VM에 있는 Transaction Table(**OLTP Data**)을 Pipeline을 통해 Migration 합니다.

- 먼저 05_CreateTransactionTable 스크립트를 실행하여 테이블을 생성합니다.
- Azure Synapse Studio의 Home으로 돌아와 Ingest를 클릭합니다.

The screenshot shows the Azure Synapse workspace interface. At the top left, it says 'kyowonsynapse'. Below that are four main buttons: 'Ingest' (highlighted with a red box), 'Explore', 'Analyze', and 'Visualize'. To the right is a large circular graphic with a bar chart and network connections. Below the buttons is a section titled 'Resources' with 'Recent' and 'Pinned' tabs. Under 'Recent', there are five items: '01_CreateDWTables', '02_CreateCredential', '04_LoadData', '03_CreateExternalTables', and '05.CreateTransactionTable'. On the right side, there's a 'Useful links' section with links like 'Getting started', 'Synapse Analytics overview', 'Documentation', and 'Give feedback'.

- Initial_Copy 라고 Task Name을 입력합니다. Next.

The screenshot shows the 'Properties' step of the Ingest Wizard. On the left, a vertical navigation menu lists steps 1 through 6: 1 Properties, 2 Source, 3 Destination, 4 Settings, 5 Summary, and 6 Deployment. Step 1 is currently selected. The main area has a title 'Properties' and a sub-instruction 'Enter name and description for the copy data task.' Below this is a 'Task name *' field containing 'Initial_Copy'. Underneath is a 'Task description' field with a large empty text area. At the bottom, there are two radio buttons for 'Task cadence or task schedule': 'Run once now' (selected) and 'Run regularly on schedule'.

4. + Create New Connection 을 클릭합니다.

Properties
One time copy

Source
Connection
Dataset

Destination
Connection
Dataset

Settings

Summary

Deployment

Source data store
Specify the source data store for the copy task. You can use an existing data store connection or specify a new data store.

All Azure Database File Generic protocol NoSQL Services and apps

All Filter by name + Create new connection

CosmosDemo kyowonsynapse-
 WorkspaceDefaultStorage sqlpool

5. MySQL 을 입력하고 아래와 같이 선택합니다. Next

New linked service

mysql

All Azure Database File Generic protocol NoSQL Services and apps

Azure Database for MySQL MySQL

6. 아래와 같이 적절하게 입력합니다.

- Name: MySQL_Server
- Server Name: 52.187.2.170
- Port: 3306

- Database Name: cohooltp
- User Name: demouser
- Password: Demo@pass123

Name *MySQL_Server

Description

Connect via integration runtime *
AutoResolveIntegrationRuntime

Connection string Azure Key Vault

Server name *
52.187.2.170

Port
3306

Database name *
cohooltp

User name *
demouser

Password Azure Key Vault

Password *
.....

7. Test Connection 을 클릭하여 연결이 잘 되는지 확인하고 생성합니다.

Annotations

+ New

✓ Connection successful

Create

Back

🔗 Test connection

Cancel

8. 생성된 Data Source 를 선택합니다. Next.

The screenshot shows the 'Source data store' configuration screen. On the left, a sidebar lists steps 1 through 6: Properties, Source, Destination, Settings, Summary, and Deployment. Step 2, 'Source', is currently selected. The main area displays a list of data stores under 'All'. A blue highlight surrounds the 'MySQL_Server' entry, which is the target destination for the copy task. Other entries include 'CosmosDemo', 'kyowonsynapse-WorkspaceDefaultStorage', and 'sqlpo'.

9. Cardtransaction 테이블을 선택합니다. Next.

The screenshot shows the 'Source' configuration screen. The sidebar indicates step 2, 'Source', is selected. In the main area, the 'EXISTING TABLES' tab is active. A list of tables is shown, with the 'cardtransaction' table being selected, indicated by a checked checkbox. Other tables listed include 'address', 'buildversion', 'cardtransaction_backup', 'customer', and 'customeraddress'. Below this, a 'Preview' section shows two rows of sample data from the 'cardtransaction' table.

transactionID	accountID	transactionAmountUSD	transactionAmount	transactionCui
47F3C9CC-2948-4069-B7BC-260361539AAD	A914800993377925	86	86	USD
A949C89B-D346-4C42-AF18-4F75FA0DD82C	A844427439174139	0	0	USD

10. Apply Filter 항목에서는 그냥 Next로 넘어갑니다.

11. Destination Source는 sqlpool을 선택합니다. Next.

The screenshot shows the 'Copy data' interface in Azure Data Factory. On the left, a sidebar lists steps 1 through 6: Properties, Source, Destination, Settings, Summary, and Deployment. Step 2, 'Source MySQL', is selected. Step 3, 'Destination', is expanded, showing 'Connection' and 'Dataset'. Step 4, 'Settings', is also expanded. In the main area, the title 'Destination data store' is displayed with the instruction 'Specify the destination data store for the copy task. You can use an existing data store connection or specify a new data store.' Below this are tabs for All, Azure, Database, File, Generic protocol, NoSQL, and Services and apps. Under the All tab, there is a dropdown menu set to 'All', a search bar 'Filter by name', and a 'Create new connection' button. A list of data stores is shown, with 'sqlpool' highlighted in blue. Other items in the list include 'CosmosDemo' and 'kyowonsynapse-WorkspaceDefaultStorage'.

12. Table Mapping에서 dbo.CardTransaction 테이블을 선택합니다. Next.

Table mapping

For each table you have selected to copy in the source data store, select a corresponding table in the destination data store or specify the table mapping.

The screenshot shows the 'Table mapping' interface. On the left, under 'Source', the table 'cardtransaction' is listed. An arrow points from this table to the 'Destination' column. The 'Destination' column contains a dropdown menu labeled '-Select-' with a filter bar 'Filter...'. A list of destination tables is shown, including 'dbo.AggregateSales', 'dbo.AggregateSales_External', 'dbo.CardTransaction', 'dbo.DatabaseLog', 'dbo.DatabaseLog_External', 'dbo.DimAccount', 'dbo.DimAccount_External', and 'dbo.DimCurrency'. The table 'dbo.CardTransaction' is highlighted with a blue border. A callout box points to this highlighted table with the text 'dbo.CardTransaction'.

13.Column Mapping 에서는 그냥 Next 로 넘어갑니다.

14.Settings 화면에서 Staging Account Linked Service 의 +New 를 클릭합니다.

Settings

More options for data movement

Fault tolerance

 ▼ ⓘ

▲ Performance settings

Enable staging

 ⓘ

▲ Staging settings

Staging account linked service

 Select... ▼ ⓘ + New

Enable Compression

 ⓘ

▲ Advanced settings

Allow PolyBase

 ⓘ

Reject type

 Value ▼ ⓘ

Reject value

 0

Use type default

Data integration unit

 Auto ▼ ⓘ
 Edit

You will be charged # of used DIUs * copy duration * \$0.25/DIU-hour. Local currency subscription type. [Learn more](#)

Degree of copy parallelism

 ⓘ
 Edit

15. 이름에 StagingBlob 으로 입력하고 ADLS Gen2 를 선택합니다. 마찬가지로 Test Connection 을 클릭하여 연결이 잘 되는지 확인하고 생성을 진행합니다.

Name *StagingBlob

Description

Connect via integration runtime *
AutoResolveIntegrationRuntime

Authentication method
Account key

Connection string Azure Key Vault

Account selection method
From Azure subscription Enter manually

Azure subscription
Microsoft Azure Internal Consumption (2386ed0e-ac2f-41c7-958a-7397006dce05)

Storage account name *
kyowonadlsgen2

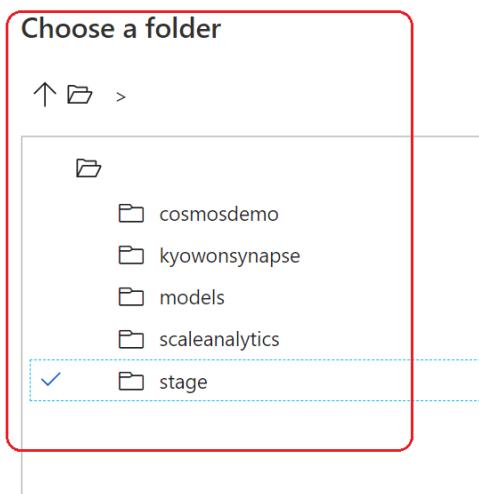
Additional connection properties
+ New

Test connection
To linked service To file path

Annotations
+ New

Advanced

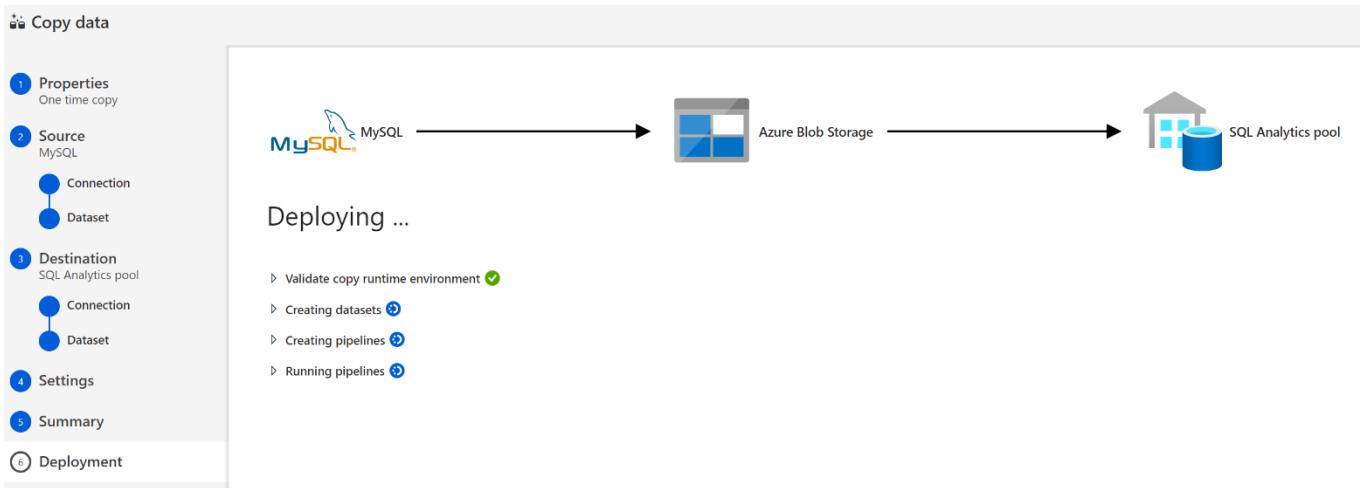
16. 바로 아래에 있는 Storage Path 에서 Browse 를 클릭하여 Stage 를 선택합니다.



17. Next 를 클릭하여 다음으로 넘어갑니다.

18. Summary Page에서 검토 후 이상업스면 Next로 넘어갑니다.

19. 아래와 같이 Pipeline이 Deploy 되고 자동으로 실행됩니다.



20. Monitor 버튼을 클릭하여 모니터 페이지로 와서, 실행된 Pipeline Activity의 Detail(안경 아이콘)을 클릭합니다.

Initial_Copy

The screenshot shows the 'Monitor' page for the 'Initial_Copy' pipeline. At the top, there are buttons for 'List' (selected) and 'Gantt'. Below are actions: Rerun, Rerun from activity, Rerun from failed activity, and Refresh. The main area shows the 'Activity runs' section for Pipeline run ID 583b49a6-0fb... . It lists one item: 'Copy_fq4'. A 'Details' button for this item is highlighted with a red box. Below the table, there are zoom controls (+, -, 100%, []).

ACTIVITY NAME	ACTIVITY TYPE	RUN START ↑↓	DURATION	STATUS	INTEGRATION RUNTIME
Copy_fq4	Copy	6/23/20, 6:19:04 AM	00:00:22	✓ Succeeded	DefaultIntegrationRuntime (Southe)

21. 아래와 같이 Detail 내용을 확인합니다.

Details ↻ Refresh ✖

Learn more on copy performance details from here. Provide feedback on performance.

Activity run id: 7791bd5c-5c2a-4276-a0ff-1732f28e533d

 MySQL → Succeeded → Azure Blob Storage → Succeeded →  Azure Synapse Analytics (formerly SQL DW) → Succeeded

Azure IR region: Southeast Asia Region: Southeast Asia Azure Synapse Region: Southeast Asia

MySQL

- Data read: 12.134 MB
- Rows read: 31,364
- Peak connections: 1

Azure Blob Storage

- Azure IR region: Southeast Asia

Azure Synapse Analytics (formerly SQL DW)

- Data written: 0 byte
- Rows written: 31,364
- Throughput: 730.882 KB/s

Copy duration 00:00:17

MySQL → Azure Blob Storage

Start time	6/23/20, 6:19:04 AM
Used DIUs	4
Used parallel copies	1
Duration	00:00:06
Details	Working duration Total duration
Queue (1)	00:00:00 00:00:02
Transfer (1)	[Time to first byte (1) 00:00:00 Reading from source (1) 00:00:01 Writing to sink (1) 00:00:00] 00:00:03

Azure Blob Storage → Azure Synapse Analytics (formerly SQL DW)

22. Develop 탭으로 돌아와서 06_OLTP_Table_Count 스크립트를 실행하여 총 Migration 건수를 확인합니다.

작업 7 : Azure Synapse Pipeline

이번 작업에서는 MySQL Server VM에 있는 Transaction Table의 Incremental Copy를 Pipeline을 통해 실행합니다.

1. Synapse Studio에서 왼쪽에 Orchestrate 탭을 선택하여 Refresh 합니다. 조금 전에 Ingest Wizard를 통해 생성된 Pipeline이 보일 것입니다.

The screenshot shows the Microsoft Azure Synapse Studio interface. The top navigation bar includes 'Microsoft Azure', 'Synapse Analytics', and the workspace name 'kyowonsynapse'. Below the navigation is a toolbar with 'Publish all', 'Validate all', 'Refresh', and 'Discard' buttons. On the left, there's a sidebar with icons for Home, Databases, Tables, Pipelines, Triggers, and Jobs. The main area is titled 'Orchestrate' and contains a search bar 'Filter resources by name'. Under the 'Pipelines' section, there is one item named 'Initial_Copy'.

2. 다음으로 +를 선택하고 Pipeline을 클릭하여 새로운 Pipeline을 생성합니다.

This screenshot shows the same Synapse Studio interface as above, but with a context menu open over the 'Pipelines' section. The menu has two options: 'Pipeline' and 'Copy data'. The 'Pipeline' option is highlighted, indicating it is the selected action for creating a new pipeline.

3. Property 탭에 이름에 Incremental_Copy 로 입력합니다.

Properties

General

i Choose a name for your pipeline.
This name can be updated at any time until it is published.

Name *

Description

Concurrency



Annotations

+ New

4. 다음으로, Activities 항목안에 General -> LookUp 을 Drag&Drop 하여 끌어옵니다.

The screenshot shows the Azure Data Factory Pipeline designer interface. On the left, there is a sidebar titled 'Activities' with a search bar and a list of activity types. The 'General' category is expanded, and the 'Lookup' activity is selected, highlighted with a red box. The main workspace shows a 'Lookup' activity named 'Lookup1' being placed on the canvas. Below the workspace, the 'General' tab of the 'Lookup1' activity settings is visible, showing fields for 'Name' (set to 'Lookup1'), 'Description', and 'Timeout' (set to '7.00:00:00').

5. LookUp1 의 Setting 탭으로 와서, 조금 전 Ingest Wizard 에서 생성된 Destination Data Set(sqlpool)을 선택합니다.

The screenshot shows the 'Settings' tab for a dataset. The 'Source dataset *' dropdown is open, displaying two options: 'Select...' and 'DestinationDataset_fq4'. The 'DestinationDataset_fq4' option is highlighted with a red box. Other tabs like 'General' and 'User properties' are visible at the top.

6. Use Query 에 Query 를 선택하고, 아래 Query 항목에 SELECT MAX([transactionDateTime]) AS transactionDateTime FROM [dbo].[CardTransaction]; 를 입력합니다.

The screenshot shows the 'Settings' tab for a dataset. The 'Source dataset *' dropdown is set to 'DestinationDataset_fq4'. Under the 'Use query' section, the 'Query' radio button is selected. The 'Query' text area contains the following SQL code:

```
SELECT MAX([transactionDateTime]) AS transactionDateTime
FROM [dbo].[CardTransaction];
```

The 'Preview data' button is highlighted with a red box. Other tabs like 'General' and 'User properties' are visible at the top.

7. Preview Data 를 클릭하여 값을 확인해봅니다.

Preview data

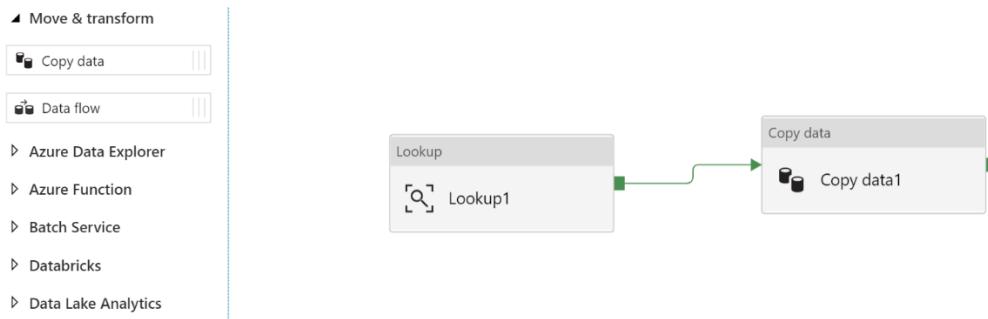
Linked service: sqlpool

Object:

transactionDateTime

2013-05-26T15:44:00

8. 다음은 Activities 항목의 Move & Transform 항목에서 Copy Data 를 Drag & Drop 하여 옮겨놓고, Lookup1 로부터 Output 을 연결시켜줍니다.



9. 다음은 Copy data1 을 선택하여 활성화 하고, Source 탭에서 Ingest Wizard 에서 생성된 SourceDataSet(MySQL)을 선택합니다. 쿼리에는 SELECT * FROM cohoOLTP.CardTransaction WHERE transactionDateTime > '@{activity('Lookup1').output.firstRow.transactionDateTime}'; 로 입력합니다.

This screenshot shows the 'Source' tab of an activity configuration. At the top, there are various icons for search, add, delete, and preview. Below that, tabs include General, Source (which is selected), Sink, Mapping, Settings, and User properties. Under 'Source dataset *', a dropdown menu is open, showing 'SourceDataset_fq4' as the selected option. Below it, a 'Query *' field contains the following SQL code:

```
SELECT * FROM cohoOLTP.CardTransaction WHERE transactionDateTime >
'@{activity('Lookup1').output.firstRow.transactionDateTime}';
```

There are also buttons for 'Open', 'New', and 'Preview data'.

10. 다음으로 Sink 탭으로 이동하여 목적지를 sqlpool 로 선택합니다.

This screenshot shows the 'Sink' tab of an activity configuration. The tabs at the top are General, Source, Sink (selected), Mapping, Settings, and User properties. Under 'Sink dataset *', a dropdown menu is open, showing 'Select...' and 'DestinationDataset_fq4' as options. There is also a '+ New' button.

11. 마지막으로 Setting 탭에서 Enable Staging에 체크 후 Ingest Wizard에서 했던 과정과 동일하게 설정합니다.

The screenshot shows the 'Settings' tab of a Data integration unit configuration. The 'Enable staging' checkbox is checked and highlighted with a red box. Other settings shown include 'Data integration unit' set to 'Auto', 'Degree of copy parallelism' with a checked 'Edit' button, and 'Fault tolerance'. Below these, under 'Staging settings', there is a section for 'Staging account linked service' (set to 'StagingBlob') with a 'Test connection' button, and a 'Storage Path' input field containing 'stage' with a 'Browse' button.

12. 이제 Publish All 을 클릭하여 Publish 합니다.

13. Add Trigger -> Trigger Now 를 클릭하여 바로 실행합니다.

The screenshot shows the triggers section of the Azure Data Factory interface. The 'Trigger now' option is selected and highlighted with a blue background. Below it is a 'New/Edit' button. At the bottom, there is a visual representation of a pipeline flow: a 'Lookup' activity (labeled 'Lookup1') connected by a green arrow to a 'Copy data' activity (labeled 'Copy data1').

14. 왼쪽의 모니터링 탭으로 이동하여 실행중인 파이프라인을 모니터 할 수 있습니다.

The screenshot shows the Microsoft Azure Synapse Analytics interface. On the left, there's a sidebar with icons for Home, Orchestration, Trigger runs, Integration runtimes, Activities (Apache Spark applications and SQL requests), and a briefcase icon. The 'Orchestration' section is expanded, showing 'Pipeline runs' which is selected and highlighted in blue. The main area is titled 'Pipeline runs' and shows a table of runs. The table has columns: PIPELINE NAME, RUN START, DURATION, TRIGGERED BY, STATUS, and PARAMETERS. There are two rows: 'Incremental_Copy' (Run Start: 6/23/20, 7:11:43 AM, Duration: 00:00:28, Triggered By: Manual trigger, Status: In progress) and 'Initial_Copy' (Run Start: 6/23/20, 6:19:00 AM, Duration: 00:00:26, Triggered By: Manual trigger, Status: Succeeded). At the top of the main area, there are filters for Time (Last 24 hours), Time zone (Seoul (UTC+9)), and Runs (Latest runs), along with buttons for Rerun, Cancel, Refresh, and Edit columns. A 'List' button is also present.

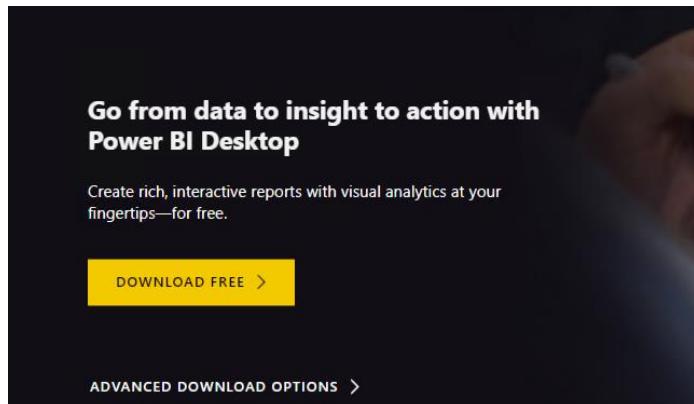
15. 바로 전 실행한 Incremental_Copy 를 선택하여 Activity 별 Detail 내용을 확인하실 수 있습니다.

The screenshot shows the Microsoft Azure Synapse Analytics interface. At the top, there's a toolbar with icons for adding (+), subtracting (-), and other operations. Below it, the title is 'Activity runs' and the pipeline run ID is listed as b86bb0a7-d01a-42f3-b0e1-90537b748469. There are filters for All status and a table showing activity runs. The table has columns: ACTIVITY NAME, ACTIVITY TYPE, RUN START, DURATION, STATUS, INTEGRATION RUNTIME, USER PROPERTIES, ERROR, and RUN ID. Two rows are shown: 'Copy data1' (Activity Type: Copy, Run Start: 6/23/20, 7:11:50 AM, Duration: 00:00:22, Status: Succeeded, Integration Runtime: DefaultIntegrationRuntime (Southeast Asia);DefaultInteg, Run ID: df743632-b) and 'Lookup1' (Activity Type: Lookup, Run Start: 6/23/20, 7:11:46 AM, Duration: 00:00:04, Status: Succeeded, Integration Runtime: AutoResolveIntegrationRuntime (Southeast Asia), Run ID: 840b55a9-7). Above the table, there's a diagram showing a 'Lookup' activity named 'Lookup1' connected by an arrow to a 'Copy data' activity named 'Copy data1'. Both activities have green checkmarks indicating success.

16. Develop 탭으로 돌아와서 06_OLTP_Table_Count 스크립트를 실행하여 총 Migration 건수를 확인합니다.

작업 8 : Power BI 로 Data Visualization.

1. 웹 브라우저에서 Power BI Desktop 다운로드 페이지 (<https://powerbi.microsoft.com/en-us/desktop/>) 로 이동합니다.
2. 페이지 중간에서 Download Free 링크를 선택합니다.



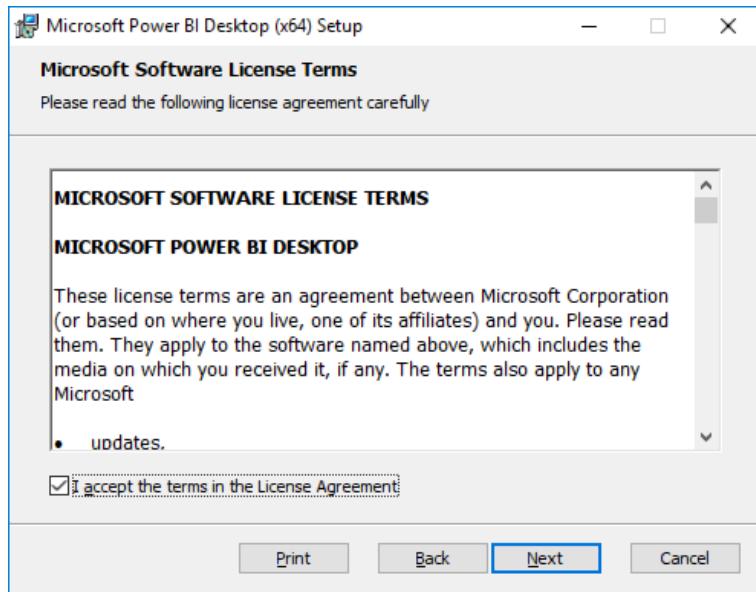
3. 설치파일을 실행합니다.



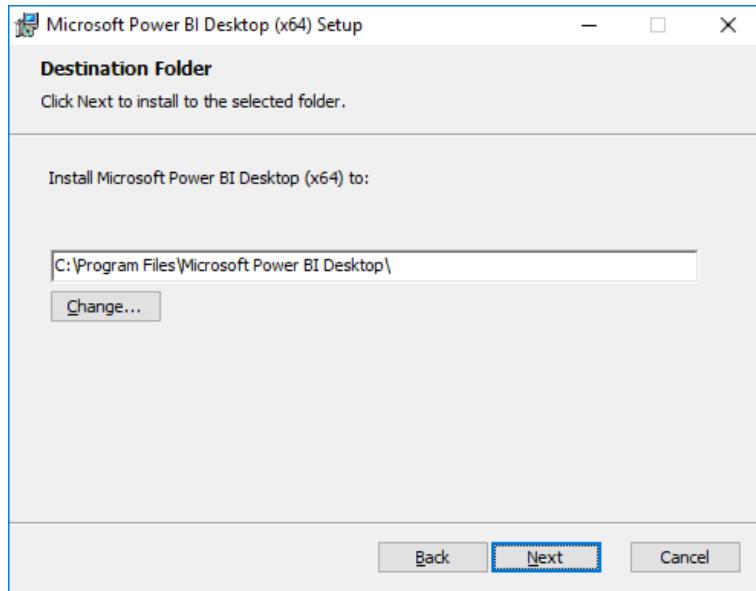
4. Welcome 화면에서 다음을 클릭합니다.



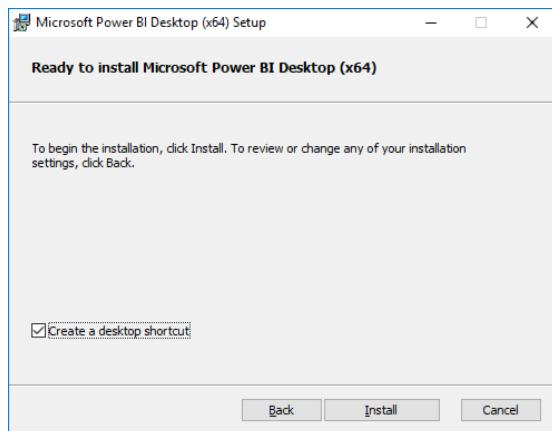
5. 사용권 계약에 동의하고 다음을 선택합니다.



6. 지정 폴더를 디폴트로 두고 다음을 선택합니다.



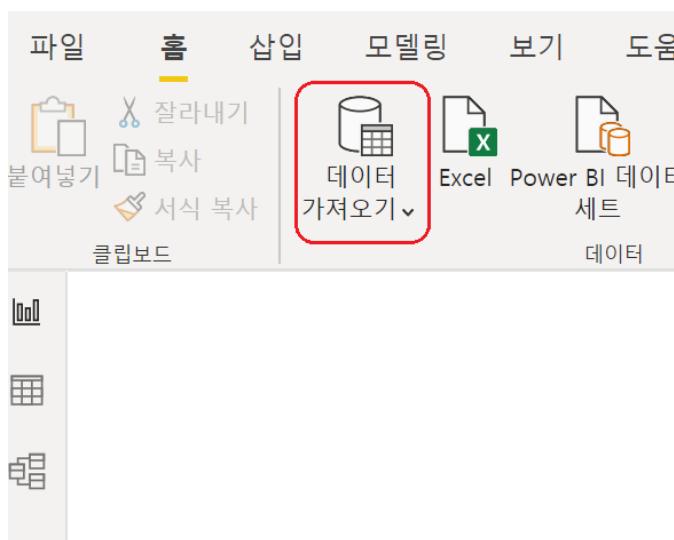
7. 바탕 화면 바로 가기 만들기 상자가 선택되어 있는지 확인하고 설치를 선택합니다.



8. Microsoft Power BI Desktop 시작이 선택되어 있는지 확인하고 마침을 선택합니다.



9. Power BI에서 데이터 가져오기를 클릭합니다.



10. Azure 를 선택 후 Azure SQL Database 를 선택합니다. 연결합니다.

데이터 가져오기

- 모두
- 파일
- 데이터베이스
- Power Platform
- Azure
- 온라인 서비스
- 기타

- Azure SQL Database
- Azure SQL Data Warehouse
- Azure Analysis Services 데이터베이스
- PostgreSQL용 Azure Database
- Azure Blob Storage
- Azure Table Storage
- Azure Cosmos DB
- Azure Data Lake Storage Gen2
- Azure Data Lake Storage Gen1
- Azure HDInsight(HDFS)
- Azure HDInsight Spark
- HDInsight Interactive Query
- Azure Data Explorer(Kusto)
- Azure Cost Management
- Azure Time Series Insights(베타)

인증된 커넥터

연결 취소

11. Azure Portal 의 Synapse Workspace 의 Overview Page 로 가서, SQL Endpoint 를 복사합니다.

The screenshot shows the Azure Synapse Workspace Overview page. In the center, there is a table of workspace settings. On the right side of the table, under the 'SQL endpoint' row, the value 'kyowonsynapse.sql.azuresynapse.net' is highlighted with a red box. To the right of this value is a 'Copy to clipboard' button. At the top of the page, there are navigation links for 'New SQL pool', 'New Apache Spark pool', 'Refresh', 'Reset SQL admin password', 'Delete', and 'Launch Synapse Studio'.

12. 복사한 값을 서버 항목에 붙여넣고, Direct Query 를 선택합니다.

SQL Server 데이터베이스

서버 ①

kyowonsynapse.sql.azuresynapse.net

데이터베이스(선택 사항)

데이터 연결 모드 ①

가져오기

DirectQuery

▷ 고급 옵션

확인

취소

13. Synapse Workspace 때 입력했던 Admin 계정으로 Database Login 을 하거나, Microsoft 계정으로 Login 을 하고 연결합니다.

14. DimGeography 테이블을 선택하고 로드 합니다.

탐색 창

The screenshot shows the Power BI Data Explorer interface. On the left, there is a sidebar titled "표시 옵션" (View Options) containing a list of tables. The "DimGeography" table is selected, indicated by a yellow border around its row in the list. On the right, the main area displays the "DimGeography" table data in a grid format. The columns are labeled "GeographyKey", "City", "StateProvinceCode", and "StateProvinceName". The data includes rows for various cities like Newcastle, East Brisbane, Basingstoke Hants, etc., with their corresponding state/province codes and names. At the bottom of the interface, there are buttons for "관련 테이블 선택" (Select Related Tables), "로드" (Load), "데이터 변환" (Data Transformation), and "취소" (Cancel).

GeographyKey	City	StateProvinceCode	StateProvinceName
10	Newcastle	NSW	New South Wales
22	East Brisbane	QLD	Queensland
228	Basingstoke Hants	ENG	England
514	Cheektowaga	NY	New York
95	Toronto	ON	Ontario
295	Baldwin Park	CA	California
487	Saint Louis	MO	Missouri
565	La Vergne	TN	Tennessee
119	Grevenbroich	BY	Bayern
475	Branch	MN	Minnesota
3	Darlinghurst	NSW	New South Wales
192	Paris	75	Seine (Paris)
482	Ferguson	MO	Missouri
138	Berlin	HH	Hamburg
293	Alpine	CA	California
507	Santa Fe	NM	New Mexico
12	North Sydney	NSW	New South Wales
352	Palo Alto	CA	California
474	Zeeland	MI	Michigan
569	Nashville	TN	Tennessee
255	London	ENG	England
500	Winston-Salem	NC	North Carolina
387	Englewood	CO	Colorado

15. 오른쪽 필드 항목에서 CountryRegionCode를 체크합니다. Map 시각화가 자동으로 시작될 것입니다.

The screenshot shows the Power BI Fields pane. On the left, there is a tree view of fields under the "DimGeography" table. The "CountryRegionCode" field is selected, indicated by a yellow border around its row in the list. At the bottom of the pane, there is a search bar with the placeholder "검색" (Search) and a "필드" (Fields) button.

16. 조금 더 구체적인 정보를 표시하기 위해, EnglishCountryRegionName 을 범례로, CountryRegion 을 크기로 Drag&Drop 합니다.

시각화 > 필드

검색

DimGeography

- City
- CountryRegionCode
- EnglishCountryRegionName
- FrenchCountryRegionName
- GeographyKey
- PostalCode
- SalesTerritoryKey
- SpanishCountryRegionName
- StateProvinceCode
- StateProvinceName

위치

범례

EnglishCountryRegionN

위도

여기에 데이터 필드 추가

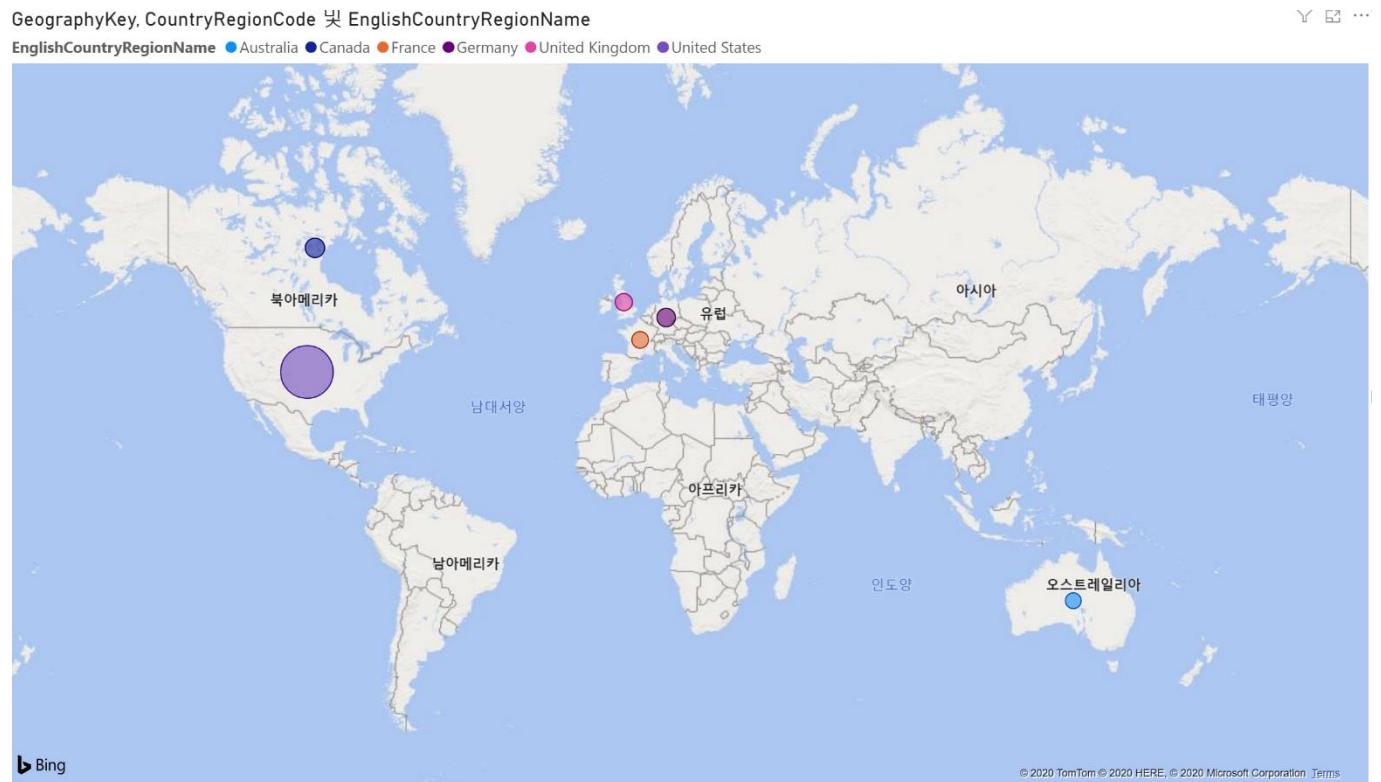
경도

여기에 데이터 필드 추가

크기

GeographyKey

17. 아래와 같은 시각화를 확인하실 수 있습니다.



작업 9: Create & Populate Customer Information Table

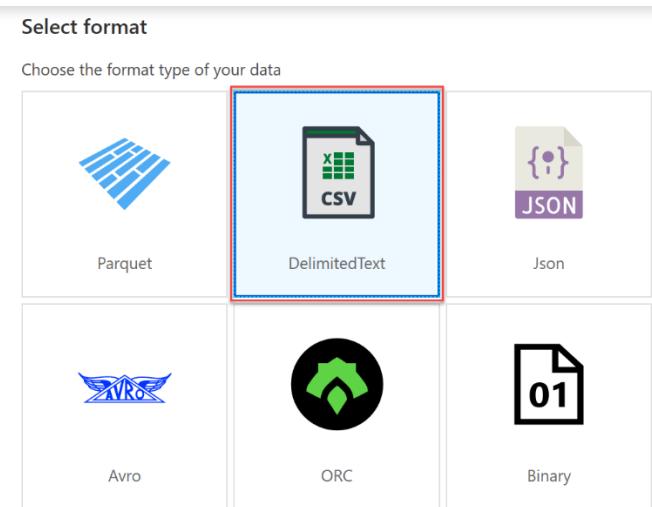
1. 왼쪽 메뉴의 Develop 탭을 선택하고 11_01_Create_Customer_Info_Table 스크립트를 열어 실행합니다.
2. Data 탭으로 이동하여 + 버튼을 누르고 Dataset을 생성합니다.

The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. On the left, there is a vertical navigation menu with icons for Home, Data, Develop, Orchestrate, Monitor, and Manage. The 'Data' option is highlighted with a red box. On the right, there is a main content area titled 'Data'. At the top of this area, there is a search bar with the placeholder 'Filter resources by name' and a '+' button with a red box around it. Below the search bar, there is a sub-menu with the title 'Dataset' and a red box around it. Underneath this, there are three categories: 'Storage accounts' (4 items), 'Databases' (7 items), and 'Datasets' (179 items). There are also 'Publish all', 'Validate all', 'Refresh', and 'Discard' buttons at the top of the main content area.

3. ADLS Gen2 를 선택합니다. Next

The screenshot shows the 'New dataset' creation dialog. At the top, it says 'Select a data store' and has a search bar. Below the search bar, there is a navigation bar with tabs: All, Azure, Database, File, Generic protocol, NoSQL, and Services and apps. The 'Azure' tab is selected and underlined. The main area displays a grid of data store options. The 'Azure Data Lake Storage Gen2' option is highlighted with a blue border, indicating it is selected. Other options include Azure Blob Storage, Azure Cosmos DB (MongoDB API), Azure Cosmos DB (SQL API), Azure Data Explorer (Kusto), Azure Data Lake Storage Gen1, and two other options whose icons are partially visible.

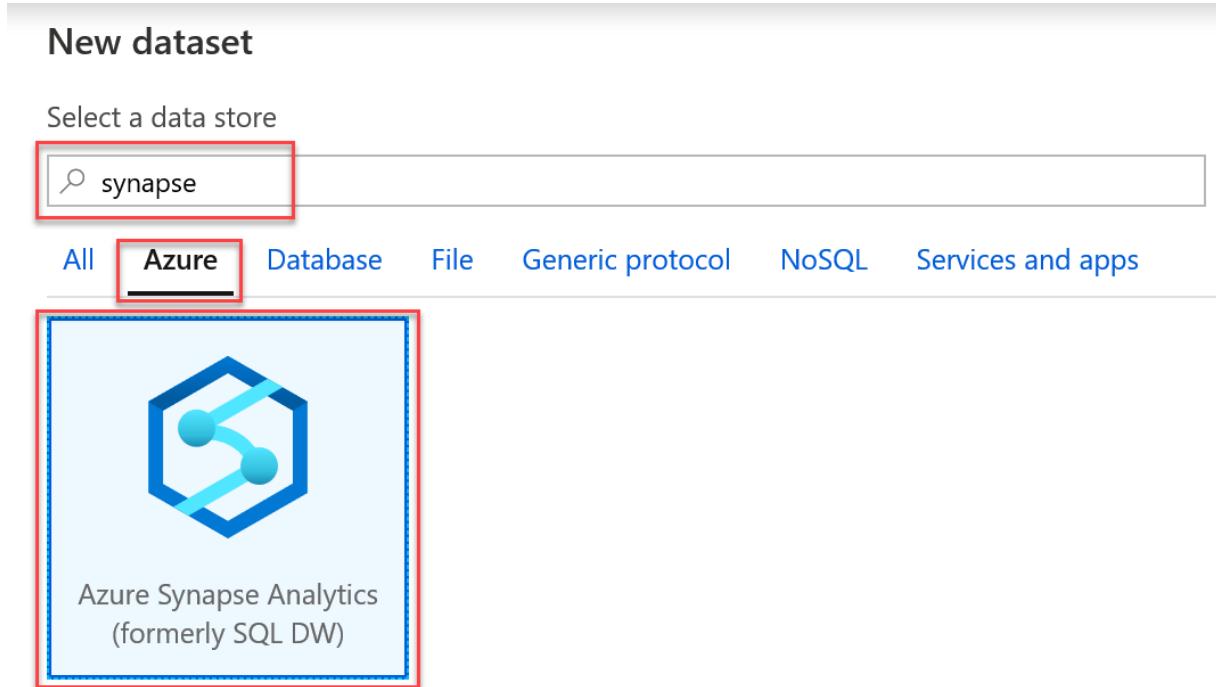
4. Delimited Text 를 선택합니다. Next



5. 아래와 같이 Property 를 설정합니다.

Field	Value
Name	Enter asamcw_customerinfo_csv
Linked service	Select <ADLS Gen2 Account Name>
File Path - Container	Enter wwi-02
File Path - Directory	Enter customer-info
File Path - File	Enter customerinfo.csv
First row as header	Checked
Import schema	Select From connection/store

6. 다시 + 버튼을 누르고 Data Set 을 생성합니다.
7. Azure 를 선택하고, **Azure Synapse Analytics (formerly SQL DW)**를 선택합니다.



8. 아래와 같이 Property 를 설정합니다.

Field	Value
Name	Enter asamcw_customerinfo_asa
Linked service	Select WorkspaceDefault
Table name	Select WWICustomerInfo .
Import schema	Select From connection/store

9. 테이블 이름에는 Edit 에 Check 하여 아래와 같이 입력합니다.

Set properties

Name
asamcw_customerinfo_asa

Linked service *
kyowonsynapse-WorkspaceDefaultSqlServer

Table name
dbo . WWICustomerInfo
 Edit

Import schema
 From connection/store None

10. DBName에는 sqlpool을 입력합니다.

11. Publish All하여 저장합니다.

12. Orchestrate 탭으로 이동하여 +를 누르고 새로운 Pipeline을 생성합니다.

13. 이름에 **Copy Customer Information**을 입력합니다.

Properties

General

i Choose a name for your pipeline.
This name can be updated at any time until it is published.

Name *

Copy Customer Information

Description

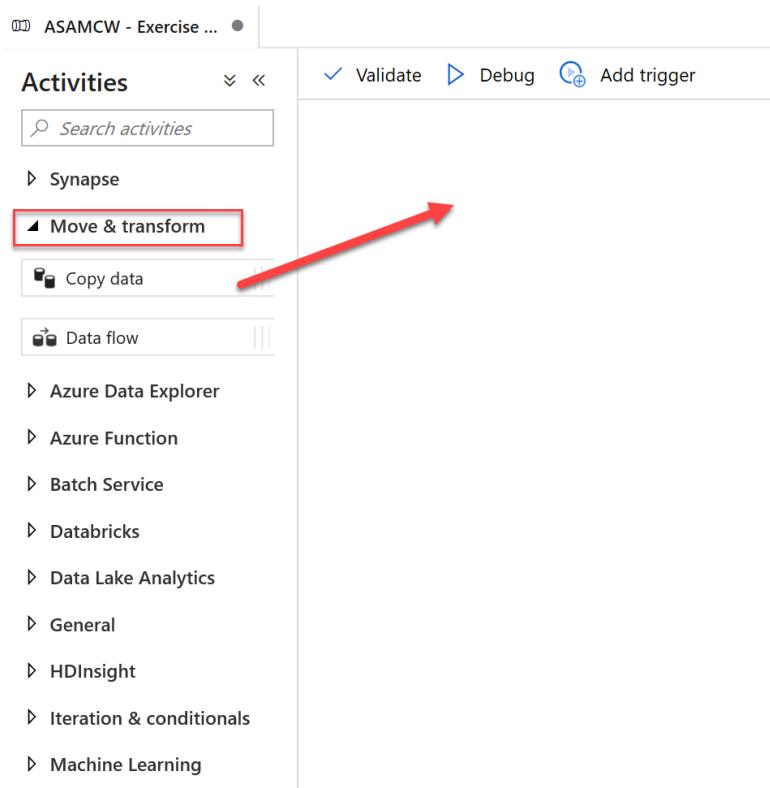
Concurrency

(i)

Annotations

 New

14. Move & Transform에서 Copy Data 를 Drag & Drop 합니다.



15. Copy Data Activity 를 선택하여 General 탭에서 이름을 입력합니다.

General Source Sink Mapping Settings User properties

Name *	<input type="text" value="Copy Customer Information"/>	Learn more
Description	<input type="text"/>	
Timeout	<input type="text" value="7.00:00:00"/>	
Retry	<input type="text" value="0"/>	
Retry interval	<input type="text" value="30"/>	
Secure output	<input type="checkbox"/>	
Secure input	<input type="checkbox"/>	

16. Source 탭에서 아래와 같이 미리 생성한 Data Set 을 선택합니다.

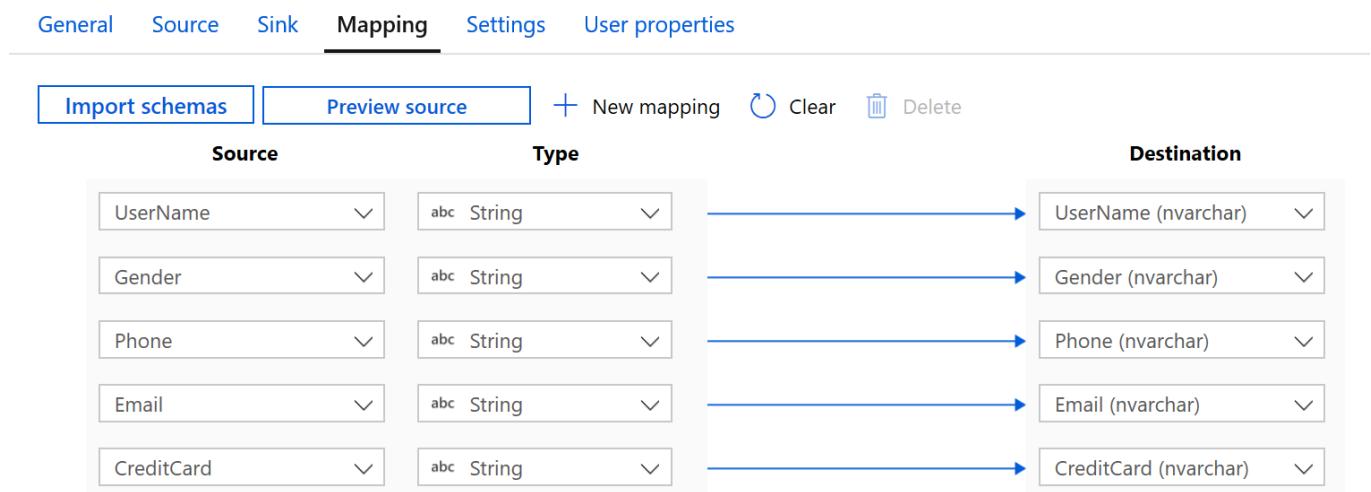
The screenshot shows the 'Source' tab configuration for a dataset. The 'Source dataset' dropdown is set to 'asamcw_customerinfo_csv'. The 'File path type' section has 'File path in dataset' selected. There are fields for 'Start time (UTC)' and 'End time (UTC)'. Under 'Filter by last modified', there is a checkbox for 'Recursively' which is checked. Other options like 'Enable partition discovery' and 'Max concurrent connections' are present. A 'Skip line count' field and an 'Additional columns' section with a 'New' button are also visible. A 'Preview data' button is located at the top right.

17. Sink 탭에서 마찬가지로 Destination Source 를 선택하고, Bulk Insert 를 선택합니다.

Pre-Copy Script 에는 truncate table dbo.WWICustomerInfo 를 입력합니다.

The screenshot shows the 'Sink' tab configuration. The 'Sink dataset' dropdown is set to 'asamcw_customerinfo_asa'. The 'Copy method' section has 'Bulk insert' selected. Under 'Table option', 'None' is chosen. In the 'Pre-copy script' section, the input field contains the SQL command 'truncate table dbo.WWICustomerInfo'. A note below says 'Add dynamic content [Alt+P]'. The 'General', 'Mapping', 'Settings', and 'User properties' tabs are also visible at the top.

18. Mapping 탭으로 이동하여 Import Schema 버튼을 클릭하고 확인합니다.



19. Publish All 하고 Trigger 하여 실행합니다.

20. Monitor 탭에서 수행 결과를 확인하고 11_02_Select_Customer_Info_Table 스크립트를 수행하여 결과를 확인합니다.

작업 10: Create & Populate Campaign Analytics Table

1. 왼쪽 메뉴의 Develop 탭을 선택하고 12_01_Create_Campaign_Analytics_Table 스크립트를 열어 실행합니다.
2. Data 탭으로 이동하여 + 버튼을 누르고 Data Set 을 생성합니다.
3. 이전과 동일하게 ADLS Gen2 를 선택 후 Delimited Text 를 선택합니다.
4. 아래와 같이 Property 를 입력합니다.

Field	Value
Name	Enter asamcw_campaignanalytics_csv
Linked service	Select <ADLS Gen2 Account Name>
File Path - Container	Enter wwi-02
File Path - Directory	Enter campaign-analytics
File Path - File	Enter campaignanalytics.csv
First row as header	Unchecked
Import schema	Select From connection/store

5. 다시 +를 누르고 Data Set 을 생성합니다.
6. 바로 전 작업과 동일하게 **Azure Synapse Analytics (formerly SQL DW)**를 선택합니다.

7. 아래와 같이 Property 를 입력합니다.

Field	Value
Name	Enter asamcw_campaignanalytics_asa
Linked service	Select WorkspaceDefault
Table name	Select WWICampaignAnalytics
Import schema	Select From connection/store

Connection Schema Parameters

Linked service * kyowonsynapse-WorkspaceDefaultS... Test connection Open New

Linked service properties ⓘ

NAME	VALUE
DBName	sqlpool

Table dbo . WWICampaignAnalytics Preview data Edit

8. Publish All 하여 저장합니다.

9. Develop 탭으로 돌아와서 +를 누르고 Data Flow 를 생성합니다.

Microsoft Azure | Synapse Analytics > workspace

Home Data Develop Orchestrate Monitor Manage

Develop

- SQL scripts
- Notebooks
- Data flows
- Power BI
- Import

Data flow

10. 이름에 **Campaign_Analytics_Data** 를 입력합니다.

Properties

General

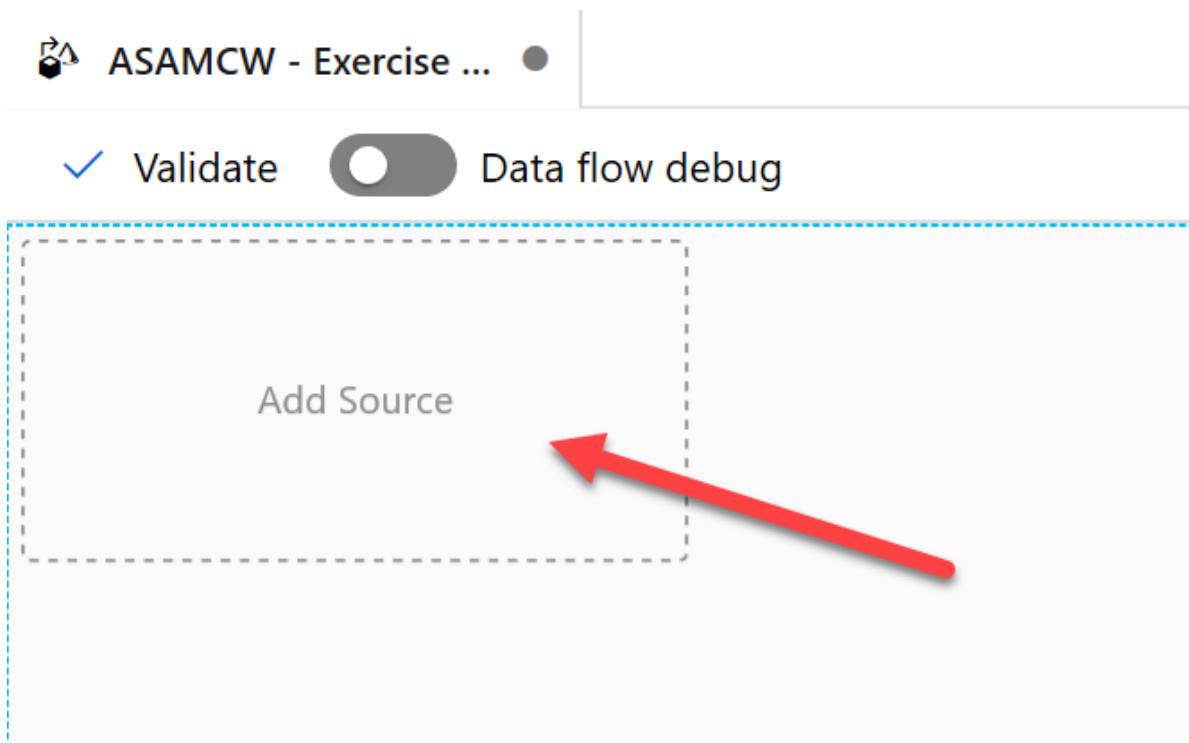
- i** Choose a name for your data flow.
This name can be updated at any time until it is published.

Name *

Campaign_Analytics_Data

Description

11. 화면에 보이는 Add Source 를 클릭합니다.



12. 아래와 같이 Property 를 입력합니다.

Field	Value
Output stream name	Enter campaignanalyticscsv
Source type	Select Dataset
Dataset	Select asamcw_campaignanalytics_csv .
Skip line count	Enter 1

Source settings Source options Projection Optimize Inspect Data preview

Output stream name * campaignanalyticscsv Learn more [🔗](#)

Source type * Dataset

Dataset * asamcw_campaignanalytics_csv [Test connection](#) [Open](#) [New](#)

Options

- Allow schema drift [ⓘ](#)
- Infer drifted column types [ⓘ](#)
- Validate schema [ⓘ](#)

Skip line count 1

Sampling * [Enable](#) [Disable](#) [ⓘ](#)

13. Data Flow 를 만들 때 데이터 미리보기 및 스키마 가져오기(프로젝션)와 같은 디버그를 설정하여 특정 기능을 사용할 수 있습니다만 시간적, 환경적 제약으로 인해 이런 과정은 Skip 합니다. Data Source 에는 일반적으로 정의 해줘야 하는 Schema 가 있으며, Data Flow 의 Designer Toolbar 도구모음에서도 Script 로 구성이 가능합니다.



14. 스크립트 안의 내용을 아래 스크립트로 수정합니다.

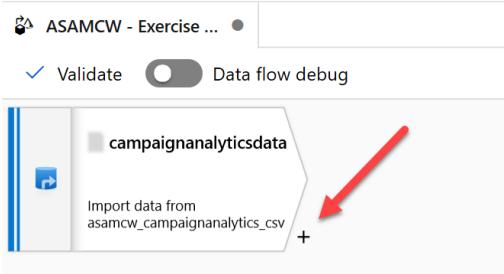
```
source(output(
    {__col0__} as string,
    {__col1__} as string,
    {__col2__} as string,
    {__col3__} as string,
    {__col4__} as string,
    {__col5__} as double,
    {__col6__} as string,
    {__col7__} as double,
    {__col8__} as string,
    {__col9__} as string
),
allowSchemaDrift: true,
validateSchema: false,
skipLines: 1) ~> campaignanalyticscsv
```

Note: 소스파일의 헤더가 잘못되어 매핑을 변경하고 있습니다.

15. Projection 탭으로 이동하여 아래와 같이 나타나는지 확인합니다.

Source settings	Source options	Projection	Optimize	Inspect	Data preview
Define default format		Detect data type	Import projection	Reset schema	
Column name	Type			Format	
__col0__	abc string	▼		Specify format	▼
__col1__	abc string	▼		Specify format	▼
__col2__	abc string	▼		Specify format	▼
__col3__	abc string	▼		Specify format	▼
__col4__	abc string	▼		Specify format	▼
__col5__	1.2 double	▼		Specify format	▼
__col6__	abc string	▼		Specify format	▼
__col7__	1.2 double	▼		Specify format	▼
__col8__	abc string	▼		Specify format	▼
__col9__	abc string	▼		Specify format	▼

16. 이제 아래 +를 클릭하여 Select 를 선택합니다.



17. Output Stream Name 에 를 **mapcampaignanalytics** 입력합니다.

18. Name as 아래에 있는 입력란에 아래와 같이 순서대로 입력합니다.

- Region
- Country
- ProductCategory
- CampaignName
- RevenuePart1
- Revenue
- RevenueTargetPart1
- RevenueTarget
- City
- State

Select settings Optimize Inspect Data preview

Output stream name * Learn more

Incoming stream *

Options Skip duplicate input columns
 Skip duplicate output columns

Input columns *

<input type="checkbox"/> campaignanalyticscsv's column	<input type="checkbox"/> Name as
abc_col0_	Region
abc_col1_	Country
abc_col2_	ProductCategory
abc_col3_	CampaignName
abc_col4_	RevenuePart1
1.2_col5_	Revenue
abc_col6_	RevenueTargetPart1
1.2_col7_	RevenueTarget
abc_col8_	City
abc_col9_	State

19. 다시 +를 클릭하여 Derived Column 을 선택합니다.

20. 이름을 **convertandaddcolumns** 로 입력합니다.

21. 아래를 참고하여 입력합니다.

Column	Expression	Description
Revenue	toDecimal(replace(concat(toString(RevenuePart1), toString(Revenue)), '\\', ','), 10, 2, '\$###,###.##')	Concatenate the RevenuePart1 and Revenue fields, replace the invalid \ character, then convert and format the data to a decimal type.
RevenueTarget	toDecimal(replace(concat(toString(RevenueTargetPart1), toString(RevenueTarget)), '\\', ','), 10, 2, '\$###,###.##')	Concatenate the RevenueTargetPart1 and RevenueTarget fields, replace the invalid \ character, then convert and format the data to a decimal type.
Analyst	iif(isNull(City), '', replace('DataAnalyst' + City, ' ', ''))	If the city field is null, assign an empty string to the Analyst field, otherwise concatenate DataAnalyst to the City value, removing all spaces.

Derived column's settings [Optimize](#) [Inspect](#) [Data preview](#)

Output stream name * [Learn more](#)

Incoming stream *

Columns * [①](#)

<input type="text" value="Revenue"/>	<input type="text" value="toDecimal(replace(concat(toString(RevenuePart1), toString(Revenue)), '\\', ','), 10, 2, '\$###,###.##')"/>	+ D B
<input type="text" value="RevenueTarget"/>	<input type="text" value="toDecimal(replace(concat(toString(RevenueTargetPart1), toString(RevenueTarget)), '\\', ','), 10, 2, '\$###,###.##')"/>	+ D B
<input type="text" value="Analyst"/>	<input type="text" value="iif(isNull(City), '', replace('DataAnalyst' + City, ' ', ''))"/>	abc + D B

22. 다시 +를 클릭하고 Select 를 선택합니다.

23. 이름에 **selectcampaignanalyticscolumns** 입력하고 **RevenuePart1** 과 **RevenueTargetPart1** 를 지웁니다.

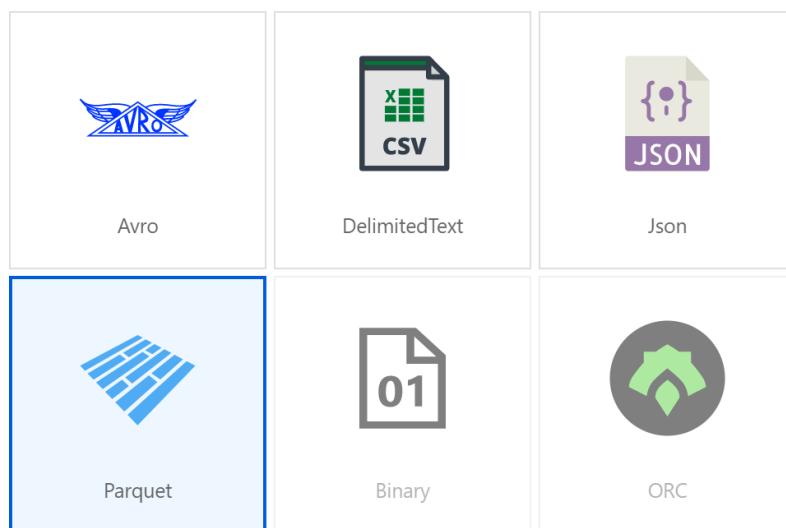
convertandaddcolumns's column	Name as
Region	Region
Country	Country
ProductCategory	ProductCategory
CampaignName	CampaignName
Revenue	Revenue
RevenueTarget	RevenueTarget
City	City
State	State
Analyst	Analyst

24. 마지막으로 +를 클릭하고 Sink 를 선택합니다.

25. 이름에 **campaignanalyticsasa** 를 입력하고 Destination Data Set 을 선택할 때 +New 를 눌러 새로 생성합니다. 이번에는 ADLS Gen2 의 Parquet 을 선택합니다.

Select format

Choose the format type of your data



26. 이름에 **asamcw_campaignanalytics_parquet** 를 입력하고 아래와 같이 Property 값을 넣어줍니다.

Set properties

i Choose a name for your dataset. This name can be updated at any time until it is published.

Name

Linked service *

File path

 / /  | 

Import schema

 From connection/store From sample file None Advanced

Sink Settings Mapping Optimize Inspect Data preview

Output stream name *	<input type="text" value="campaignanalyticsasa"/> 
Incoming stream *	<input type="text" value="selectcampaignanalyticscolumns"/> 
Sink type *	<input type="text" value="Dataset"/> 
Dataset *	 <input type="text" value="asamcw_campaignanalytics_parquet"/>  Test connection  Open  New
Options	<input checked="" type="checkbox"/> Allow schema drift  <input type="checkbox"/> Validate schema 

27. 위와 같이 된 것을 확인 후 Setting 탭으로 가서 File Name Option에서 **Output to one single file**을 선택합니다. 파일 명에는 **campaignanalytics.parquet**를 입력합니다.

Sink Settings Mapping Optimize Inspect Data preview

Clear the folder Add dynamic content [Alt+P]

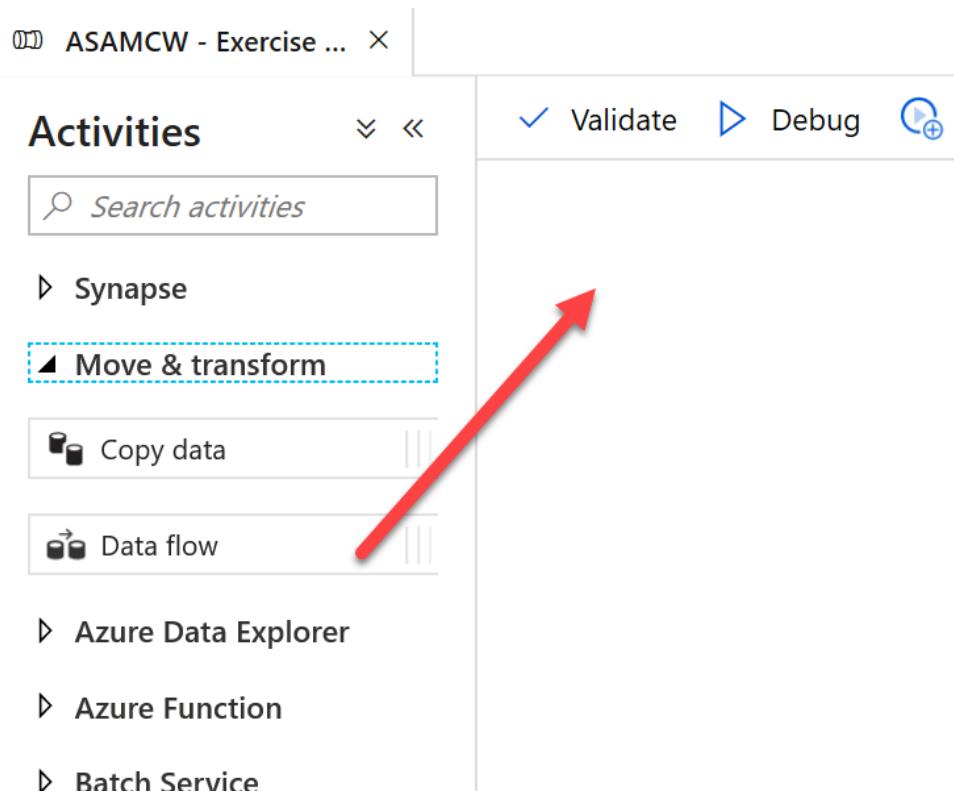
File name option * Default Pattern Per partition As data in column Output to single file

Output to single file * campaignanalytics.parquet ⓘ

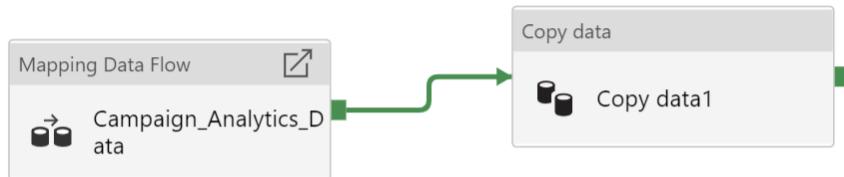
28. Publish All 하여 저장합니다.

29. Orchestrate 탭에서 +를 클릭 후 Pipeline을 생성합니다. 이름을 **Copy Campaign Analytics Data**로 입력합니다.

30. Data Flow 를 Drag & Drop 하고 방금 생성한 Data Flow를 선택합니다.



31. Copy Data 를 마찬가지로 Drag & Drop 하여 아래와 같이 이어줍니다.



32. Copy Data 를 선택하고, Source 탭으로 이동하여, **asamcw_campaignanalytics_parquet** 를 Source Data Set 으로 선택합니다. Wildcard file Path 를 선택 후 campaign-analytics 를 입력합니다.

General **Source** Sink Mapping Settings User properties

Source dataset *

File path type File path in dataset Wildcard file path List of files ⓘ

Wildcard paths wwi-02 / campaign-analytics / *.parquet

Filter by last modified Start time (UTC) End time (UTC)

Recursively ⓘ

Enable partition discovery

Max concurrent connections

33. Sink 탭으로 이동해서 **asamcw_campaignanalytics_asa** 를 선택하고 Bulk Insert 를 선택합니다.

General Source **Sink** Mapping Settings User properties

Sink dataset *

Copy method PolyBase ⓘ Copy command (Preview) ⓘ Bulk insert

Table option None Auto create table ⓘ

Pre-copy script

34. Publish All 하여 저장합니다.
35. Trigger Now 하여 실행하고 Monitoring 탭에서 확인합니다.
36. 작업이 완료되면 12_02_Count_Campaign_Analytics_Table 스크립트를 실행하여 결과를 확인합니다.

작업 11: Exploring Raw Data

1. 왼쪽 메뉴의 Data 탭을 선택하고 Storage 안에 **sale-small/Year=2010/Quarter=Q4/Month=12/Day=20101231**로 이동하여 파일을 우클릭 후 Select Top 100 rows SQL 스크립트를 실행합니다.

The screenshot shows the Azure Data Explorer interface. On the left, there's a navigation pane titled 'Data' with tabs for 'Workspace' and 'Linked'. Under 'Storage accounts', several accounts are listed, including 'kyowonsynapse (Primary - kyowo...)' and 'wwi-02' (which is highlighted with a red box). Under 'wwi-02', there are 'Cosmos DB' and 'Datasets'. In the main area, a file named 'sale-small-20101231' is selected. A context menu is open over this file, with the 'Select TOP 100 rows' option highlighted with a red box. The full path 'wwi-02 > sale-small > Year=2010 > Quarter=Q4 > Month=12 > Day=20101231' is also highlighted with a red box at the top of the menu.

2. SQL on-demand 에 연결된 것을 확인하고 실행합니다.

The screenshot shows the 'SQL on-demand' query editor. The query window contains the following T-SQL code:

```

Run Undo Publish Query plan Connect to SQL on-demand Use database master
1 SELECT
2     TOP 100 *
3 FROM
4     OPENROWSET(
5         BULK 'https://kyowonadlsgen2.dfs.core.windows.net/wwi-02/sale-small/Year=2010/Quarter=Q4/Month=12/Day=20101231/sale-small-20101231-snappy.parquet',
6         FORMAT='PARQUET'
7     ) AS [r];
8

```

3. 아래와 같이 SQL 을 수정하여 Grouping 과 Aggregation 을 수행합니다.

```

SELECT
    TransactionDate, ProductId,
    CAST(SUM(ProfitAmount) AS decimal(18,2)) AS [(sum) Profit],
    CAST(AVG(ProfitAmount) AS decimal(18,2)) AS [(avg) Profit],
    SUM(Quantity) AS [(sum) Quantity]
FROM
    OPENROWSET(
        BULK 'https://<ADLS Gen2 Account Name>.dfs.core.windows.net/wwi-02/sale-
small/Year=2010/Quarter=Q4/Month=12/Day=20101231/sale-small-20101231-snappy.parquet',
        FORMAT='PARQUET'
    ) AS [r] GROUP BY r.TransactionDate, r.ProductId;

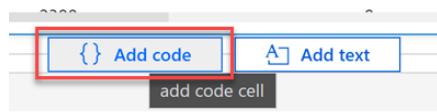
```

4. 다음은 같은파일에서 우클릭하여 새로운 노트북을 실행합니다.

The screenshot shows a file browser with a navigation bar at the top. Below the navigation bar is a breadcrumb path: 'wwi-02 > sale-small > Year=2010 > Quarter=Q4 > Month=12 > Day=20101231'. A red box highlights the file 'sale-small-20101231-snappy.parquet'. A context menu is open over this file, listing options: 'New SQL script', 'New notebook' (which is highlighted with a red box), 'Copy ABFSS path', 'Manage Access...', 'Rename...', 'Download', 'Delete', and 'Properties...'. The 'New notebook' option is the intended target for the user's action.

5. Run All 을 수행합니다.

6. Add Code 를 클릭하여 다음 Cell 에 아래와 같이 입력하여 Schema 를 확인합니다.



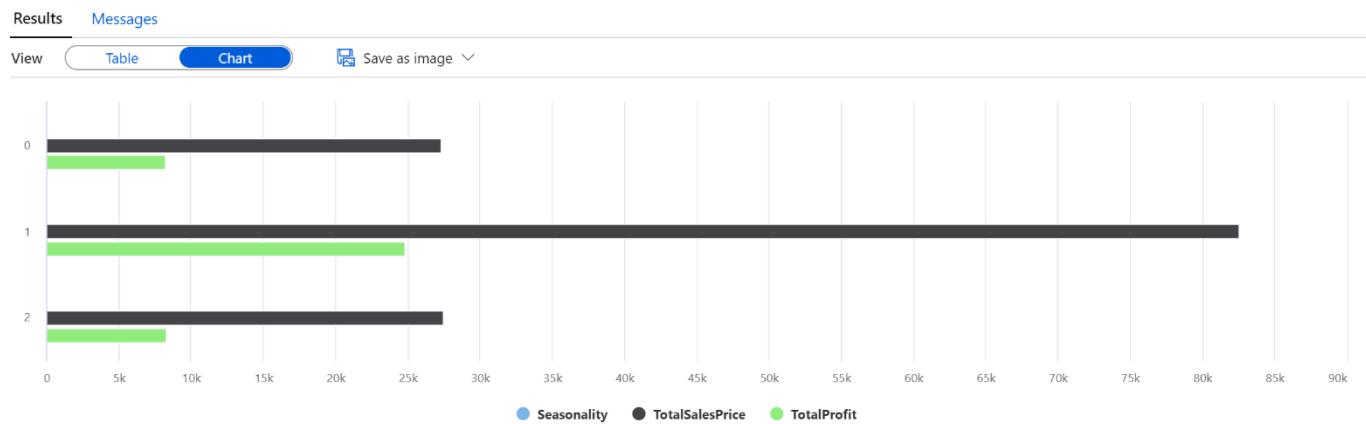
```
data_path.printSchema()
```

7. 다시 Add Code 하여 아래와 같이 입력하여 조금 전 SQL로 수행한 Script 와 결과를 비교해봅니다

```
from pyspark.sql import SparkSession
from pyspark.sql.types import *
from pyspark.sql.functions import *

profitByDateProduct = (data_path.groupBy("TransactionDate", "ProductId")
    .agg(
        sum("ProfitAmount").alias("(sum)ProfitAmount"),
        round(avg("Quantity"), 4).alias("(avg)Quantity"),
        sum("Quantity").alias("(sum)Quantity"))
    .orderBy("TransactionDate"))
profitByDateProduct.show(100)
```

8. Develop 탭의 13_01_Select_From_CSV ~ 13_03_Select_From_CSV SQL 스크립트를 열어 실행합니다.
9. 아래와 같이 Chart로 변경하여 다양하게 시각화를 해봅니다.



작업 12 : Azure Synapse Analytics Security

1. Develop 탭에서 14_01_Column_Level_Security SQL 스크립트를 열어 가이드에 따라 Step by Step 단계를 진행합니다.
2. Develop 탭에서 14_02_Row_Level_Security SQL 스크립트를 열어 가이드에 따라 Step by Step 단계를 진행합니다
3. Develop 탭에서 14_03_Dynamic_Data_Masking SQL 스크립트를 열어 가이드에 따라 Step by Step 단계를 진행합니다

작업 13 : Spark Pool Library Install

4. [Azure Portal](#)로 돌아가서 Synapse Workspace의 sparkpool을 선택합니다.
5. Sparkpool의 왼쪽 Package 탭에서 Upload Environment Config File을 선택합니다.

sparkpool (kyowonsynapse/sparkpool) | Packages

Apache Spark pool

Search (Ctrl+ /) <> Upload environment config file Refresh

Overview Activity log Access control (IAM) Tags

Settings

Packages Properties Locks

User-provisioned packages

Name
No user-provided packages currently uploaded. You can upload "environment config file".

Upload environment config file

6. 깃허브에서 다운로드 받은 Synapse_HOL\Synapse_HOL\LabFiles\Scripts\Notebooks\Env_Config 안에 있는 Environment_Config.txt 파일을 선택하고, 아래 체크박스에 체크하고 업로드합니다.

Force new settings on the Apache Spark pool (will immediately stop running Apache Spark applications).

Upload

작업 14: Azure Synapse Studio 의 Synapse Link 생성

1. Azure Synapse Workspace Overview Page 에서 아래와 같이 Link 를 클릭하여 Synapse Studio 를 실행합니다.

The screenshot shows the Azure Synapse Workspace Overview Page. At the top right, there is a button labeled "Launch Synapse Studio" with a red box drawn around it. Below the button, there is a table of workspace settings. One row in the table is highlighted with a red box, specifically the "Workspace web URL" row, which contains the URL <https://web.azuresynapse.net?workspace=%2fsubscriptions%2f386ed0e-ac2f-41c7-958a-7397006dce05>. The table includes columns for Resource group, Status, Location, Subscription, Subscription ID, Managed virtual network, Managed Identity object ID, Workspace web URL, and Tags. Below the table, there is a section titled "Available resources" with tables for "SQL pools" and "Apache Spark pools".

2. Data 탭을 선택하고 + 를 클릭하여 External Data Source 를 연결합니다.

The screenshot shows the Microsoft Azure Synapse Analytics workspace. On the left, there is a sidebar with icons for Home, Storage accounts, Pipelines, and Datasets. The main area has tabs for "Data" (selected), "Linked", and "Dataset". A context menu is open over the "Data" tab, with the "Connect to external data" option highlighted. Other options in the menu include "Publish all", "Validate all", "Refresh", and "Discard all".

3. Azure Cosmos DB(SQL API)를 선택합니다.

Connect to external data

Connect to supported data sources to explore them in the Data hub and make them available to pipelines as linked services.



4. 이름에 CosmosDemo로 입력하고 생성한 DB를 선택하여 연결합니다.

New linked service (Azure Cosmos DB (SQL API))

Name *
CosmosDemo

Description
[Empty text area]

Connect via integration runtime *
AutoResolveIntegrationRuntime

Connection string **Azure Key Vault** (Blue button)

Account selection method
 From Azure subscription Enter manually

Azure subscription
Microsoft Azure Internal Consumption (2386ed0e-ac2f-41c7-958a-...)

Cosmos DB account name *
kyowoncosmos

Database name *
CosmosDemo

Additional connection properties
+ New

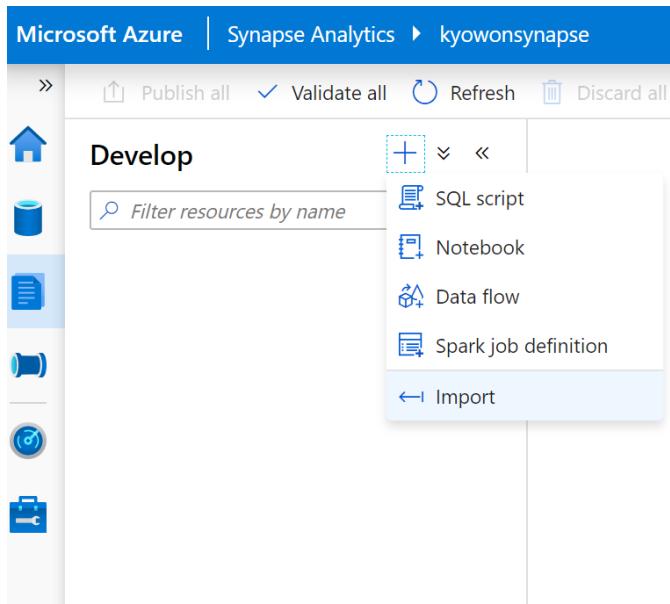
5. Refresh 하여 생성된 Link 를 확인합니다.

The screenshot shows the Azure Data workspace interface. At the top, there are buttons for Publish all, Validate all, Refresh, and Discard all. Below that, a navigation bar has 'Data' selected and tabs for 'Workspace' and 'Linked'. A search bar says 'Filter resources by name'. The main area lists resources under 'Linked': 'Storage accounts' (1 item), 'Cosmos DB' (1 item), and 'CosmosDemo (CosmosDemo)' (6 items). The 'CosmosDemo' item is expanded, showing 'RetailSales', 'IoTSignals', 'IoTDeviceInfo', 'RetailProducts', 'IoTStreamingSignals', and 'RetailStoreDemographics'. The 'Actions' button next to the CosmosDB entry is highlighted with a blue dashed box.

Category	Count
Storage accounts	1
Cosmos DB	1
CosmosDemo (CosmosDemo)	6

작업 15: Azure Synapse Workspace Spark Scenario

10. 왼쪽 메뉴의 Develop 탭을 선택하고 +를 선택하여 Import 를 실행합니다.



11. 다운로드 받은 Synapse_HOL\Synapse_HOL\LabFiles\Scripts\Notebooks 안의 노트북 Script 들을 선택하여 불러옵니다.

12. Publish All 을 클릭하여 Publish 합니다.

13. 00_CheckLibraries 노트북을 열고 실행하여 설치된 Library 들을 확인합니다.

```
+ Cell ▾ ▶ Run all ⏪ Undo | ▾ Publish Attach to sparkpool Language PySpark (Python) ▾
Cell 1
1 import pip #needed to use the pip functions
2 for i in pip.get_installed_distributions(local_only=True):print(i)
Command executed in 7mins 19s 191ms by yongl on 06-23-2020 09:52:44.072 +09:00
zipp 3.1.0
zict 1.0.0
xlwt 1.2.0
XlsxWriter 0.9.6
xlrd 1.0.0
wrapt 1.11.2
widgetsnbextension 3.5.1
wheel 0.30.0
Werkzeug 0.16.1
websocket-client 0.57.0
webencodings 0.5.1
wcwidth 0.2.4
```

14. [Azure Portal](#)에서 오늘 생성한 Resource Group에서 Cosmos DB를 찾아 선택합니다.

Showing 1 to 5 of 5 records. Show hidden types ⓘ

Name ↑↓	Type ↑↓
kyowonadlsgen2	Storage account
kyowoncosmos	Azure Cosmos DB account
kyowonsynapse	Synapse workspace
sparkpool (kyowonsynapse/sparkpool)	Apache Spark pool
sqlpool (kyowonsynapse/sqlpool)	SQL pool

15. Data Explorer 탭을 선택하여 CosmosDemo DB를 열고 각 Collection들의 Item 을 선택하여 Data 를 확인 할 수 있도록 합니다.

kyowoncosmos | Data Explorer

Azure Cosmos DB account

Search (Ctrl+ /) New Container Enable Notebooks (Preview)

Overview Activity log Access control (IAM) Tags Diagnose and solve problems Quick start Notifications Data Explorer

Settings Features Replicate data globally Default consistency

SQL API

CosmosDemo

- Scale
- IoTDeviceInfo
- IoTSignals
- Items
- Settings
- Stored Procedures
- User Defined Functions
- Triggers
- IoTStreamingSignals
- RetailProducts
- RetailSales
- RetailStoreDemographics

16. 이제 Develop 탭에서 01_CosmosDB_Configuration 노트북을 열고 <ADLS Gen2 Account Name> 부분을 수정해준 후 Run All 실행합니다.

17. 다시 Cosmos DB 의 Data Explorer 를 돌아와서 Data Ingestion 을 확인합니다.

The screenshot shows the Azure Cosmos DB Data Explorer interface. On the left, there's a sidebar with 'SQL API' and a tree view of databases and collections under 'CosmosDemo'. The 'IoTSignals' collection is selected. In the main area, there's a table titled 'Items' with columns 'id' and '/id'. One row is highlighted with a yellow box, showing the full JSON document content. The JSON document includes fields like 'dateTime', 'measureValue', 'unit', 'unitSymbol', 'deviceId', 'measureType', '_rid', '_self', '_etag', and '_attachments'. The JSON content is as follows:

```

1
2
3
4
5
6
7
8
9
10
11
12
13
14
{
    "dateTime": "2020-04-28 17:40:54",
    "measureValue": "4322",
    "unit": "Revolutions per Minute",
    "unitSymbol": "RPM",
    "id": "dca4c4d9-d503-4298-a1ce-850469b525d3",
    "deviceId": "dev-2",
    "measureType": "Rotation Speed",
    "_rid": "ePYRAJEonGcFAAAAAAAA==",
    "_self": "dbs/ePYRAJEonGc=/colls/ePYRAJEonGc=/docs/ePYRAJEonGcFAAAAAAAA==/",
    "_etag": "\\"1000f824-0000-2500-0000-5ef146c60000\\\"",
    "_attachments": "attachments/",
    "_ts": 1592870598
}

```

18. Synapse Studio 의 Develop 탭의 02_IoT_Data_Exploration 스크립트를 열고 Cell 4 까지 실행합니다.

1. Create Spark tables pointing to the Azure Cosmos DB Analytical Store collections using Azure Synapse Link

Cell 2

```
[ ] 1 %%sql
2 create database CosmosDemoIoT
```

Cell 3

```
[ ] 1 %%sql
2
3 create table if not exists CosmosDemoIoT.IoTSignals
4 using cosmos.olap
5 options(spark.synapse.linkedService 'CosmosDemo',
6 |||| spark.cosmos.container 'IoTSignals')
```

Cell 4

```
1 %%sql
2
3 create table if not exists CosmosDemoIoT.IoTDeviceInfo
4 using cosmos.olap
5 options(spark.synapse.linkedService 'CosmosDemo',
6 |||| spark.cosmos.container 'IoTDeviceInfo')
```

19. Data 탭을 선택하고 Workspace에서 CosmosDemoloT DB가 생성되었는지, 테이블도 생성되었는지 확인합니다.

The screenshot shows the Azure Data Explorer interface. On the left, the 'Data' tab is selected, showing the 'Workspace' section. Under 'Databases', 'cosmosdemoliot (Spark)' is expanded, revealing 'Tables' which contains 'iotdeviceinfo' and 'iotsignals'. To the right, a notebook cell titled '02_IoT_Data_Explora...' is open, showing the following code:

```

1. Create Spark tables pointing to the Azure
Cell 2
1 %%sql
2 create database CosmosDemoIoT
Cell 3
1 %%sql
2
3 create table if not exists (
4     using cosmos.olap

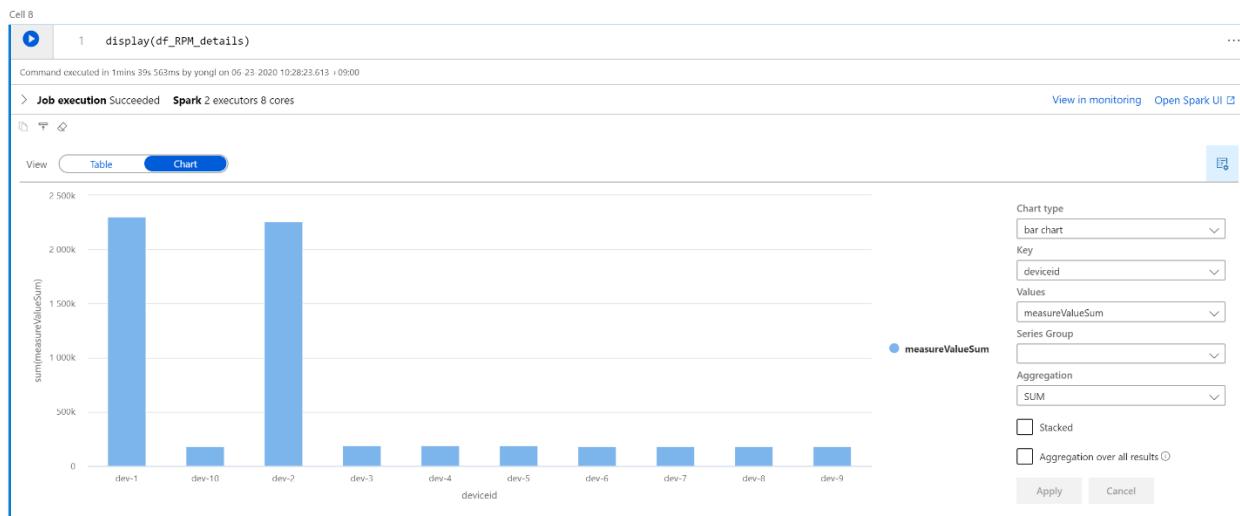
```

20. Cell 8 번까지 수행 후 결과를 확인합니다. 결과 창에서 Chart를 선택 후 옵션에서 Value 항목에 latitude 대신에 measurevaluesum을 선택합니다.

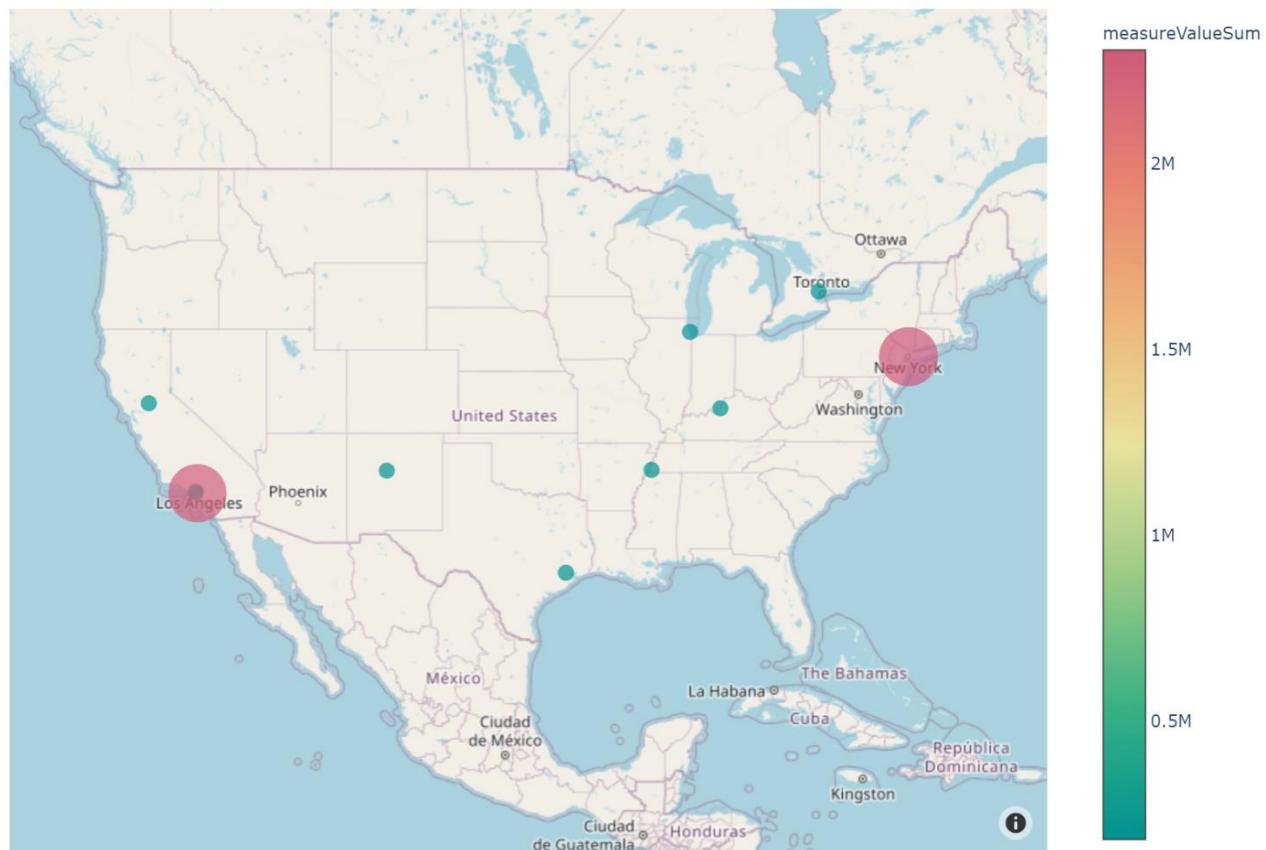
The screenshot shows the results of a query execution. The command was `display(df_RPM_details)` and it succeeded. The results are displayed in a table view. A 'Chart' button is highlighted with a red box. The table data is as follows:

deviceid	devicetype	location	latitude	longitude	measuretype	unitSymbol
dev-10	impulse turbine	Santa Monica, CA	34.011227	-118.499435	Rotation Speed	RPM
dev-3	impulse turbine	Chicago, IL	41.87864	-87.64025	Rotation Speed	RPM
dev-4	impulse turbine	Houston, TX	29.73222	-95.38752	Rotation Speed	RPM
dev-2	impulse-reaction turbine	Los Angeles, CA	33.944237	-118.3976	Rotation Speed	RPM
dev-5	impulse-reaction turbine	Toronto, Ontario	43.731525	-79.60837	Rotation Speed	RPM
dev-7	impulse-reaction turbine	Sacramento, CA	38.47561	-121.43124	Rotation Speed	RPM
dev-8	impulse turbine	Albuquerque, NM	35.107105	-106.57277	Rotation Speed	RPM
dev-1	impulse turbine	Manhattan, NY	40.7142	-74.0064	Rotation Speed	RPM

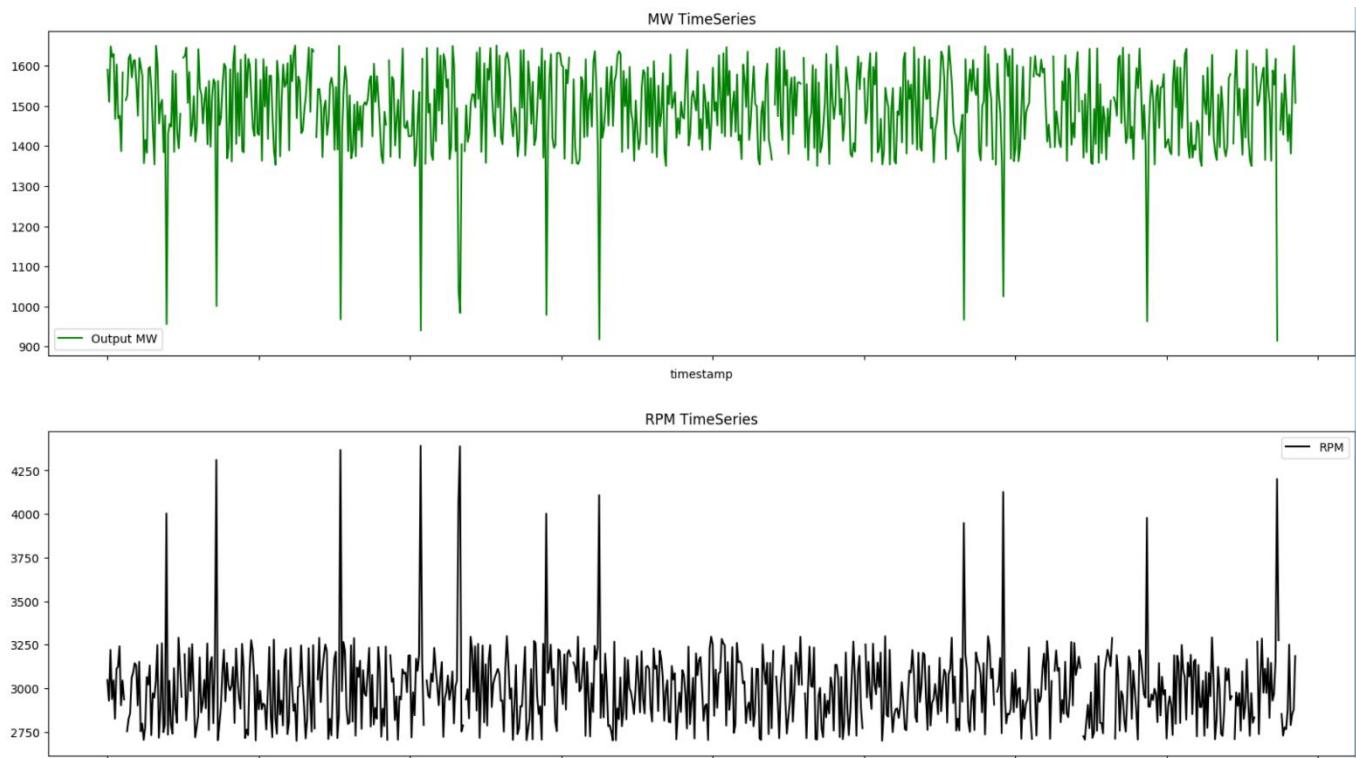
21. 아래와 같이 총 10 개의 IoT Device 중에 1 번과 2 번 Device 에 측정값 합(MeasureValueSum)이 상당히 높은것을 확인할 수 있습니다.



22. Cell 10 번을 수행하여 각 Device 별 Geographic Insight 를 확보합니다.



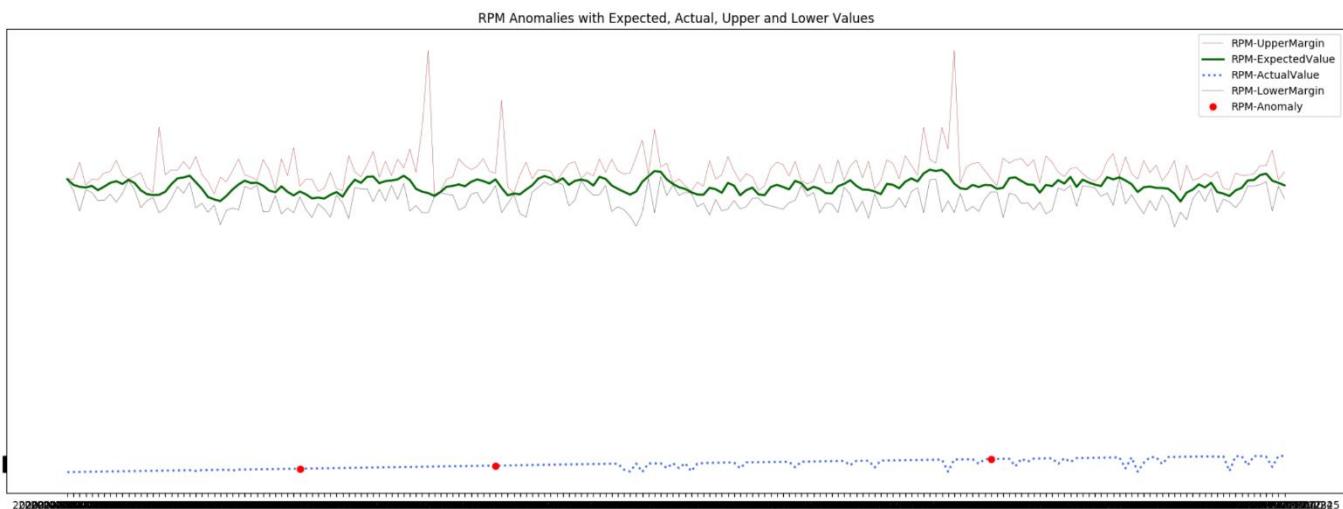
23. Cell 4 까지 수행하여, 이상값이 있는 Device 중 하나인 Dev-1에 대해 unitSymbol 을 기준으로 MW 와 RPM 에 Data 를 확인합니다.



24. Cell 6 번의 <Azure Anomaly Detector End Point> 부분을 적절하게 수정하고, Cell 7 번까지 수행하여 전체 Device 에 대해 unitSymbol 이 RPM 인 Data 에 대해 Anomaly Detector 로부터 받은 내용을 확인합니다.

Cell 7							
<pre>1 display(df_anomaly)</pre> <p>Command executed in 7s 772ms by yongj on 06-23-2020 10:52:20.111 +09:00</p> <p>> Job execution Succeeded Spark 2 executors 8 cores</p> <p>View in monitoring Open Spark UI</p>							
<p>View Table Chart</p>							
deviceId	measureType	.etag	value	grouping	SimpleDetectAnomalies 97521124...	anomalies	
7b65c...	dev-2	Rotation Speed	"1000952f-0000-2500-0000-5ef146...	3062	dev-2	<ul style="list-style-type: none"> ▼ {"isAnomaly":false,"isPositiveAnomaly":false,"isNegativeAnomaly":false,"period":0,"expectedValue":3062,"upperMargin":598000000000000,"lowerMargin":598000000000001} 	
7415...	dev-2	Rotation Speed	"1000e931-0000-2500-0000-5ef146...	3169	dev-2	<ul style="list-style-type: none"> ► {"isAnomaly":false,"isPositiveAnomaly":false,"isNegativeAnomaly":false,"period":0,"expectedValue":3169,"upperMargin":598000000000000,"lowerMargin":598000000000001} 	
13281...	dev-2	Rotation Speed	"10003d32-0000-2500-0000-5ef146...	2795	dev-2	<ul style="list-style-type: none"> ► {"isAnomaly":false,"isPositiveAnomaly":false,"isNegativeAnomaly":false,"period":0,"expectedValue":2795,"upperMargin":598000000000000,"lowerMargin":598000000000001} 	
fb4b...	dev-2	Rotation Speed	"10009928-0000-2500-0000-5ef146...	2848	dev-2	<ul style="list-style-type: none"> ► {"isAnomaly":false,"isPositiveAnomaly":false,"isNegativeAnomaly":false,"period":0,"expectedValue":2848,"upperMargin":598000000000000,"lowerMargin":598000000000001} 	

25. Cell 9 번을 수행하여 데이터를 Visualize 합니다.



26. 다음으로 04_Retail_Data_AutoML 노트북을 열어 Cell 4 번까지 실행합니다.

27. Data 탭으로 이동하여 Workspace에 CosmosDemoRetail DB 와 테이블들이 생성되었는지 확인합니다.

Data + ↴ <<

Workspace Linked

▲ Databases 4

- ▷ 🗁 sqlpool (SQL pool)
- ▷ 🗁 cosmosdemoiot (Spark)
- ◀ 🗁 **cosmosdemoretail** (Spark)
- ◀ └ Tables
 - ▷ 📄 retailproducts
 - ▷ 📄 retailsales
 - ▷ 📄 retailstoredemographics
- ▷ 🗁 default (Spark)

04_Retail_Data_Auto... X
+ Cell ↴ Run all ⏪ Undo ⏵ Publish ⏧ Attach

1. Create Spark tables pointing to the Azure CosmosDB

Cell 2

```
[ ] 1 %%sql
      2 create database CosmosDemoRetail
```

Cell 3

```
[ ] 1 %%sql
      2
      3 create table if not exists CosmosDemoRetail
      4 using cosmos.olap
      5 options(spark.synapse.linkedService 'CosmosDB'
      6           spark.cosmos.container 'RetailProducts')
```

28. Cell 5 번을 수행하고 Data 를 탐색합니다.
29. Cell 7 번의 필요한 정보들을 입력하고 수행하여 Azure ML Workspace 의 환경설정을 진행합니다.
30. Cell 9 번을 수행하여 Data Prep.을 진행합니다.
31. Cell 11 번을 수행하여 AutoML 을 수행하고 결과를 확인합니다. 30 분 가량 진행됩니다.

```

Running on local machine
Parent Run ID: AutoML_141283ae-566c-4259-b588-191b3292f3c4

Current status: DatasetFeaturization. Beginning to featurize the dataset.
Current status: DatasetFeaturizationCompleted. Completed featurizing the dataset.
Current status: DatasetCrossValidationSplit. Generating individually featurized CV splits.
Current status: DatasetFeaturization. Beginning to featurize the CV split.
Current status: DatasetFeaturizationCompleted. Completed featurizing the CV split.
Current status: DatasetFeaturization. Beginning to featurize the CV split.
Current status: DatasetFeaturizationCompleted. Completed featurizing the CV split.
Current status: DatasetFeaturization. Beginning to featurize the CV split.
Current status: DatasetFeaturizationCompleted. Completed featurizing the CV split.

*****
DATA GUARDRAILS:

TYPE: Frequency detection
STATUS: PASSED
DESCRIPTION: The time series was analyzed, all data points are aligned with detected frequency.

TYPE: Missing feature values imputation
STATUS: PASSED
DESCRIPTION: No feature missing values were detected in the training data.
Learn more about missing value imputation: https://aka.ms/AutomatedMLFeaturization

*****
Current status: ModelSelection. Beginning model selection.

*****
ITERATION: The iteration being evaluated.
PIPELINE: A summary description of the pipeline being evaluated.
DURATION: Time taken for the current iteration.
METRIC: The result of computing score on the fitted pipeline.
BEST: The best observed score thus far.

*****
```

ITERATION	PIPELINE	DURATION	METRIC	BEST
0	StandardScalerWrapper ElasticNet	0:00:20	0.0310	0.0310
1	StandardScalerWrapper ElasticNet	0:00:17	0.0324	0.0310
2	StandardScalerWrapper ElasticNet	0:00:19	0.0308	0.0308
3	StandardScalerWrapper ElasticNet	0:00:15	0.0310	0.0308
4	StandardScalerWrapper LassoLars	0:00:19	0.0323	0.0308
5	StandardScalerWrapper ElasticNet	0:00:21	0.0304	0.0304
6	StandardScalerWrapper RandomForest	0:01:37	0.0262	0.0262
7	StandardScalerWrapper ExtremeRandomTrees	0:00:14	0.0184	0.0184
8	StandardScalerWrapper DecisionTree	0:00:14	0.0040	0.0040
9	MinMaxScaler GradientBoosting	0:00:14	0.0215	0.0040
10	MinMaxScaler DecisionTree	0:00:14	0.0078	0.0040
11	RobustScaler LassoLars	0:00:15	0.0325	0.0040
12	MaxAbsScaler LightGBM	0:00:15	0.0121	0.0040
13	StandardScalerWrapper ElasticNet	0:00:14	0.0305	0.0040
14	MinMaxScaler RandomForest	0:00:23	0.0024	0.0024
15	StandardScalerWrapper LassoLars	0:01:41	0.0321	0.0024
16	StandardScalerWrapper ElasticNet	0:00:19	0.0325	0.0024
17	StandardScalerWrapper DecisionTree	0:00:14	0.0048	0.0024
18	MinMaxScaler DecisionTree	0:00:14	0.0157	0.0024
19	RobustScaler DecisionTree	0:00:19	0.0061	0.0024
20	StandardScalerWrapper RandomForest	0:00:31	0.0097	0.0024
21	RobustScaler RandomForest	0:00:23	0.0057	0.0024
22	MinMaxScaler LightGBM	0:00:32	0.0176	0.0024
23	MaxAbsScaler GradientBoosting	0:00:29	0.0115	0.0024
24	RobustScaler ExtremeRandomTrees	0:00:35	0.0106	0.0024
25	RobustScaler RandomForest	0:00:21	0.0160	0.0024
26	MaxAbsScaler LightGBM	0:00:20	0.0220	0.0024
27	StandardScalerWrapper ExtremeRandomTrees	0:00:20	0.0167	0.0024
28	MaxAbsScaler LightGBM	0:00:15	0.0214	0.0024
29	StandardScalerWrapper LightGBM	0:00:20	0.0102	0.0024
30	VotingEnsemble	0:03:12	0.0024	0.0024

32. Cell 13 번을 수행하여 AutoML 수행결과를 받아오고 Test Data로 Prediction 을 수행합니다.

33. Cell15 를 수행하여 결과를 Plotting 합니다.

[Test data scores]

```
root_mean_squared_error: 11.958
normalized_root_mean_squared_error: 0.007
normalized_root_mean_squared_log_error: 0.003
r2_score: 0.998
root_mean_squared_log_error: 0.015
explained_variance: 0.998
median_absolute_error: 0.426
mean_absolute_error: 3.304
mean_absolute_percentage_error: 0.862
normalized_mean_absolute_error: 0.002
spearman_correlation: 1.000
normalized_median_absolute_error: 0.000
```

