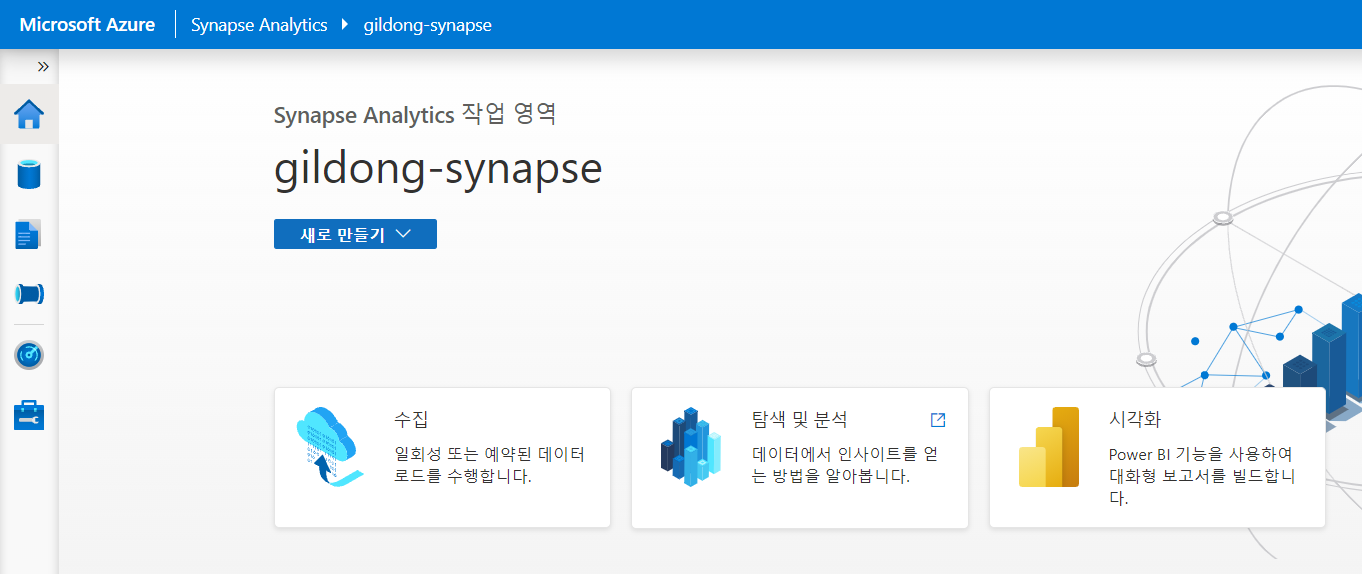
**Lab 7 – Synapse Spark**

# 요구 사항

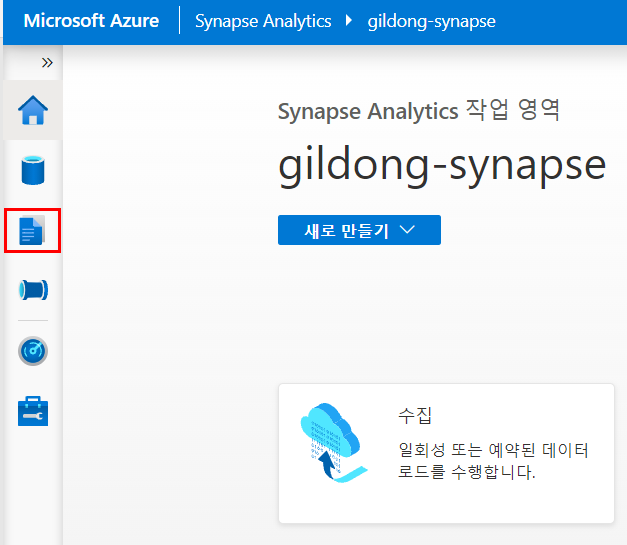
1. Microsoft Azure 구독 (Microsoft 이외의 구독은 유료 구독이어야 합니다.)
2. Azure Synapse Analytics, Azure Machine Learning Workspace

# Task 1 : Import Notebook files

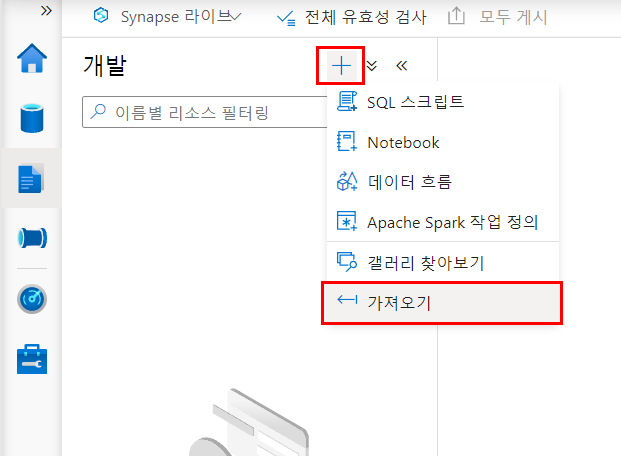
1. 구성해 놓은 **Synapse Analytics**에 접속합니다.



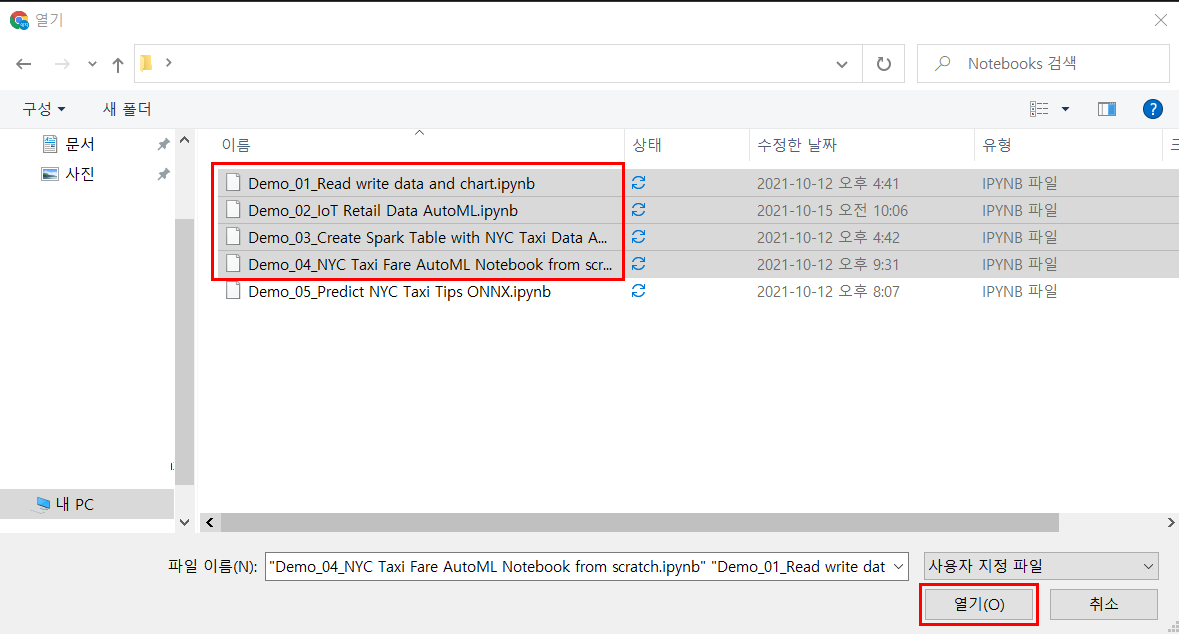
1. **개발(Develop) Hub**로 이동합니다.



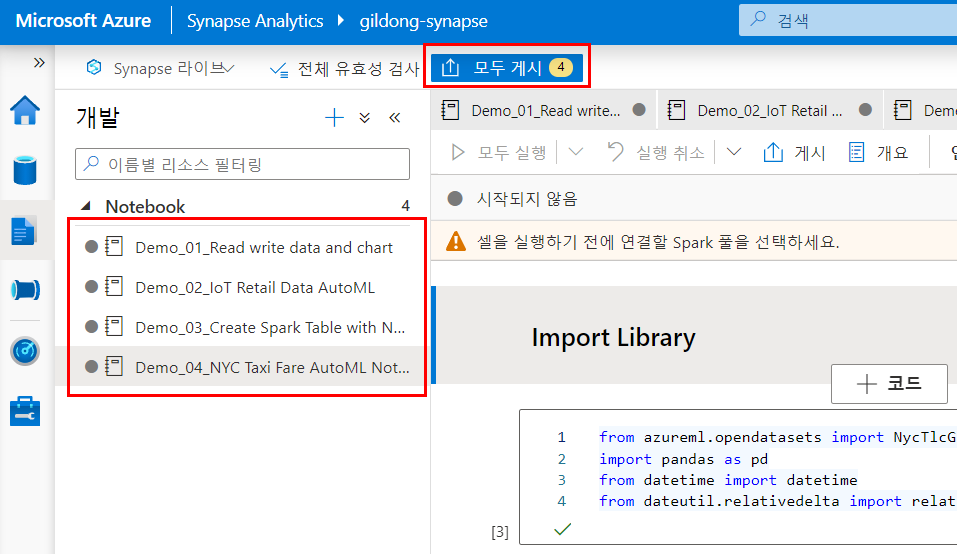
1. 상단의 **+** 표시를 클릭한 후 드롭다운 메뉴 중 **가져오기**를 클릭합니다.



1. **Demo01 ~ 04번** 파일을 선택하여 파일을 가져옵니다.



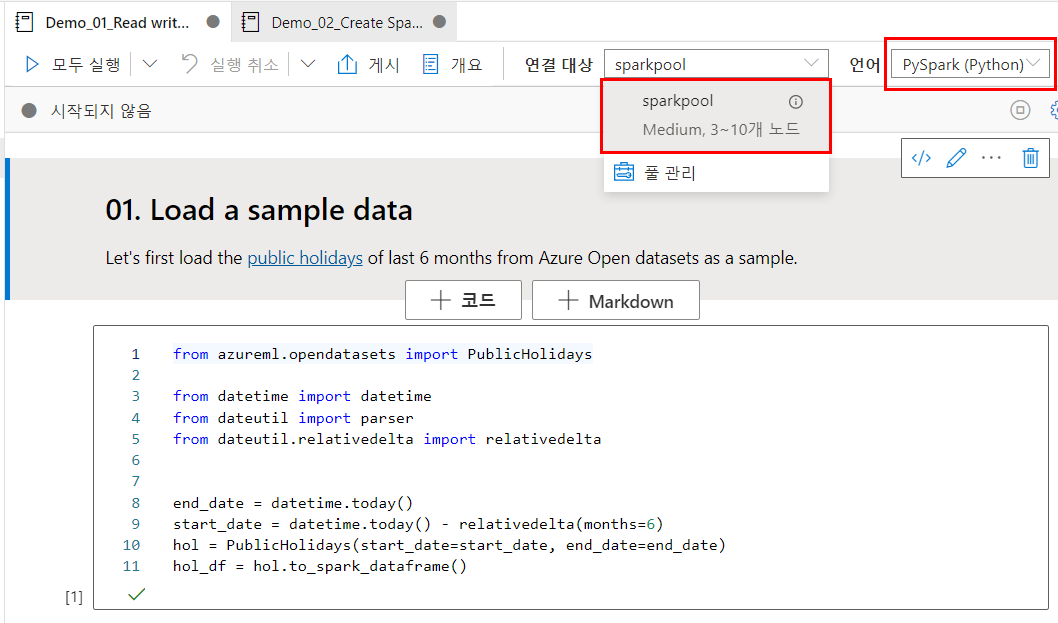
1. 파일들을 확인하고 **모두 게시**를 눌러 **현재까지의 상태(파일 가져오기)를 저장**합니다.

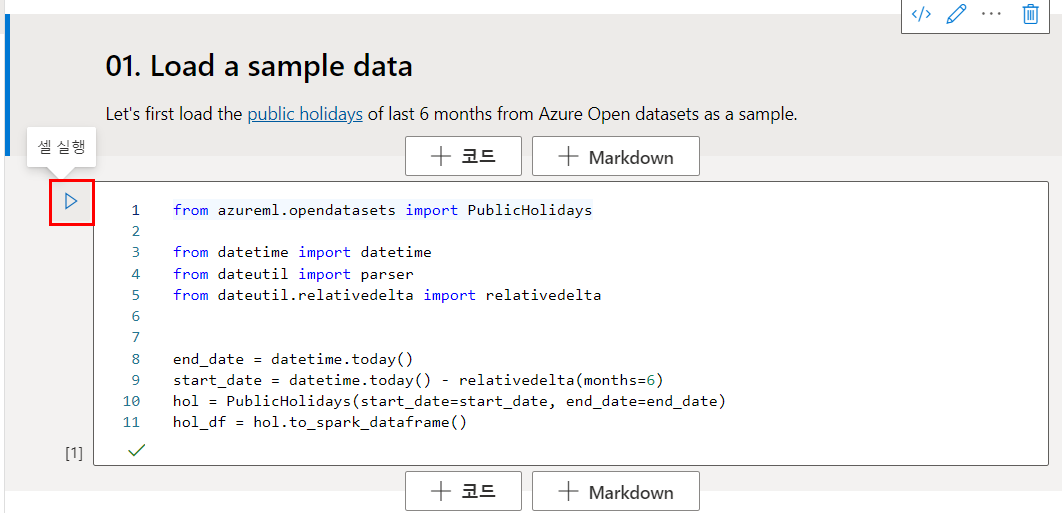


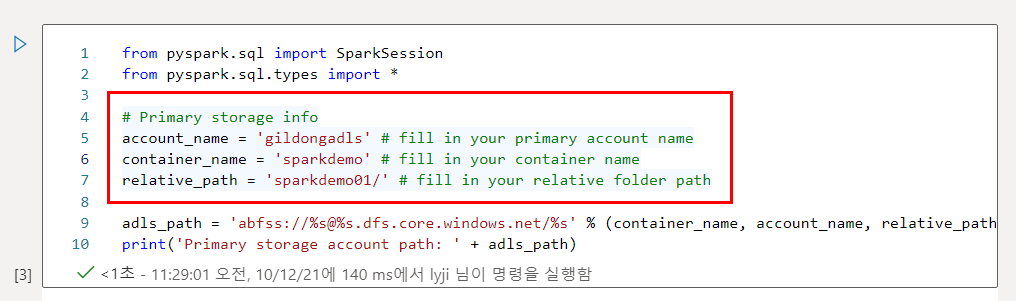


# Task 2 : Demo 01. Read Write Data & Chart

1. **01번 파일**을 선택합니다. 연결대상으로 생성한 **spark pool**을 선택합니다. 언어는 **PySpark**로 설정합니다.



1. **Demo 01**번 파일을 실행합니다. 각 셀마다 **실행** 단추를 눌러 실행하며 **2번 셀까지** 실행합니다. Spark pool세션이 시작되는데 시간이 몇 분 소요될 수 있습니다. 
2. **3번 셀**에 ADLS **account name, container name, relative path**를 아래와 같이 입력합니다.

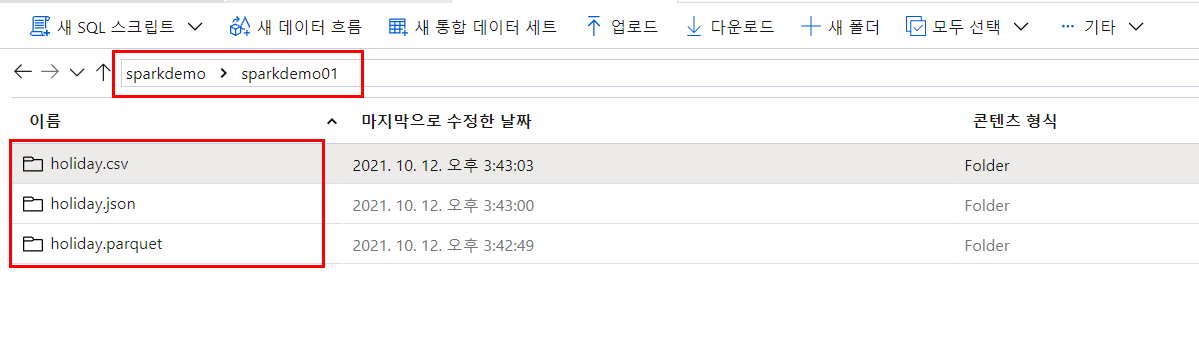


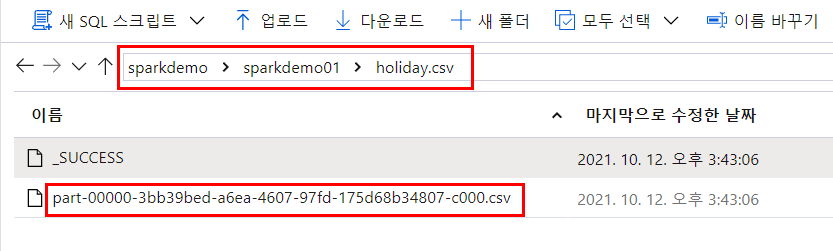
Account name : gildongadls

Container name: sparkdemo

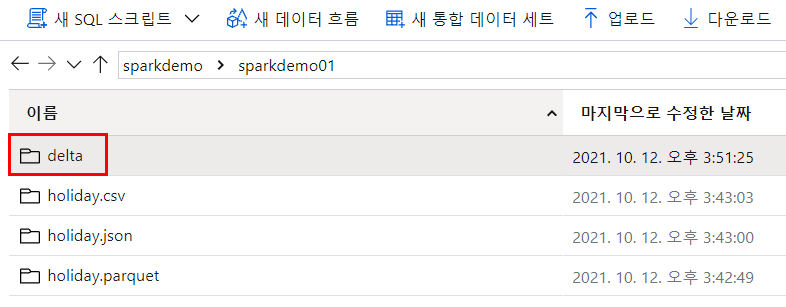
Relative path: sparkdemo01/

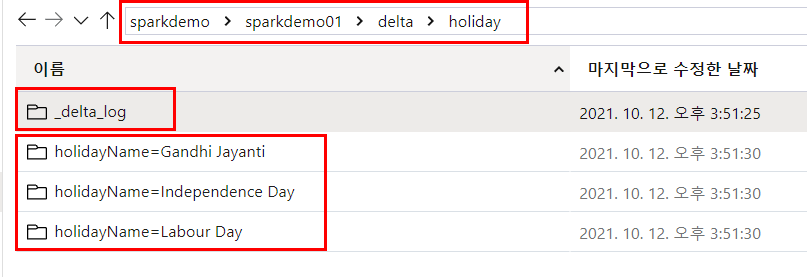
1. **5번 셀**까지 실행합니다. **ADLS**에 데이터가 저장되었는지 확인합니다. **데이터 허브에서 바로 연결된 ADLS를 조회**할 수 있으며, 아래와 같이 **.csv, .json, .parquet** 형식의 파일이 각 폴더에 생성된 것을 확인합니다.



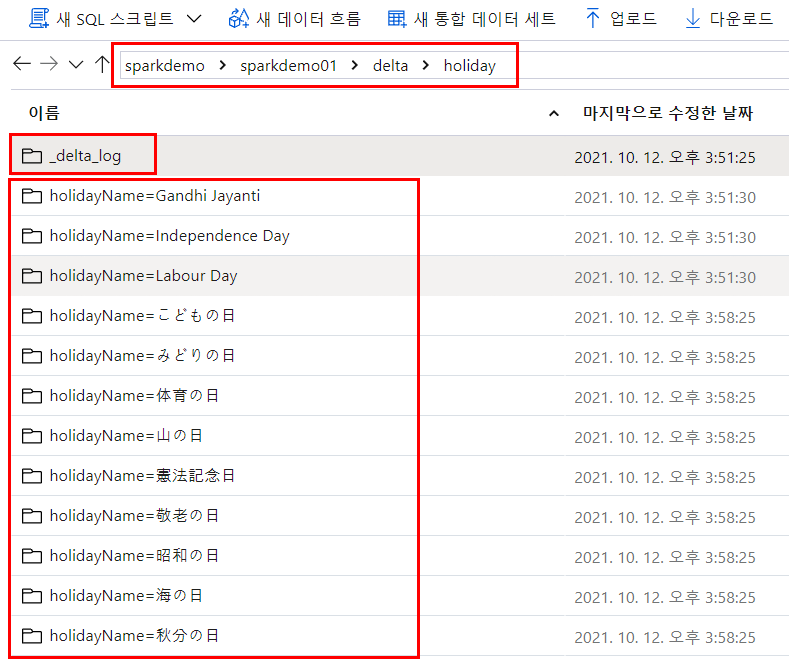


1. **6, 7, 8번 셀**을 실행하여 ADLS에서 다시 데이터를 읽어와 봅니다.
2. **9, 10, 11번 셀**을 실행하여 Delta Lake table을 만들고 저장합니다. **ADLS에서 Delta 테이블을 확인**합니다.

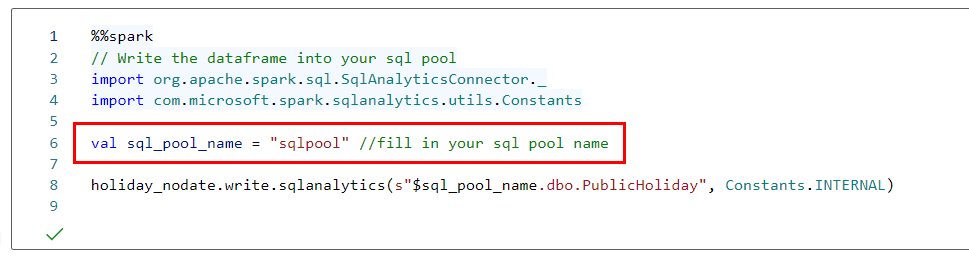




1. **12번 셀**을 실행해서 다시 delta table을 읽어옵니다.
2. **13번 셀**을 실행하고 다시 한번 **ADLS**를 확인합니다. **CountryRegionCode 가 JP**인 데이터로 **Overwrite된 결과**를 확인합니다.

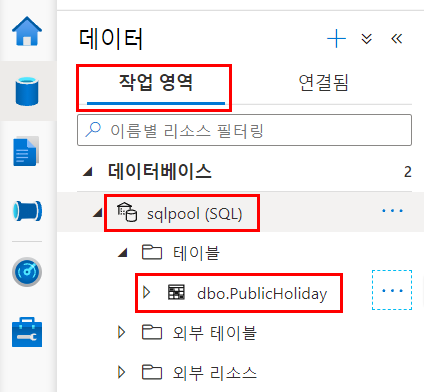


1. **14번 셀**을 실행해서 Overwrite된 버전을 읽어옵니다.
2. **18번 셀까지** 차례대로 실행해보면서 조건에 따라 **Delta table을 편집**해봅니다.
3. **20번 셀까지** 차례대로 실행합니다. **21번 셀**에 **SQL pool 이름**을 아래와 같이 입력합니다.

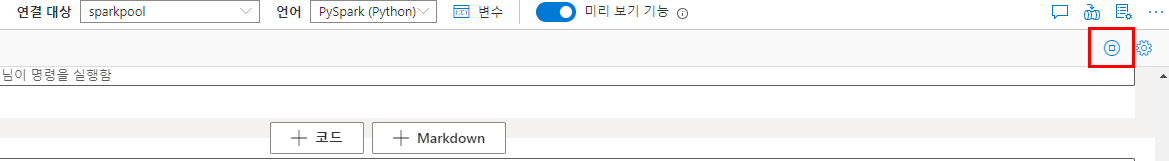


SQL pool name : sqlpool

1. **22번 셀까지** 차례대로 실행합니다. SQL pool에 데이터를 저장하고 읽어와 봅니다. **데이터 허브에서 SQL pool에 테이블이 생성된 것을 확인**합니다.

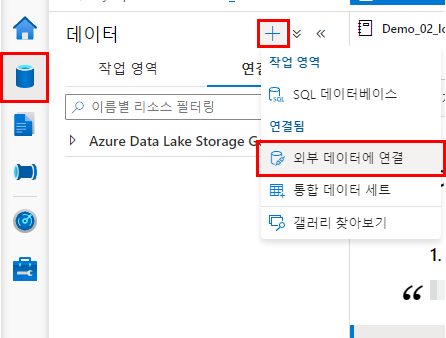


1. **나머지 셀**을 차례대로 실행합니다. 여러 **시각화 차트 결과**를 확인합니다.
2. 노트북을 닫기 전에 **세션 종료 버튼을 눌러 세션을 종료합니다.**

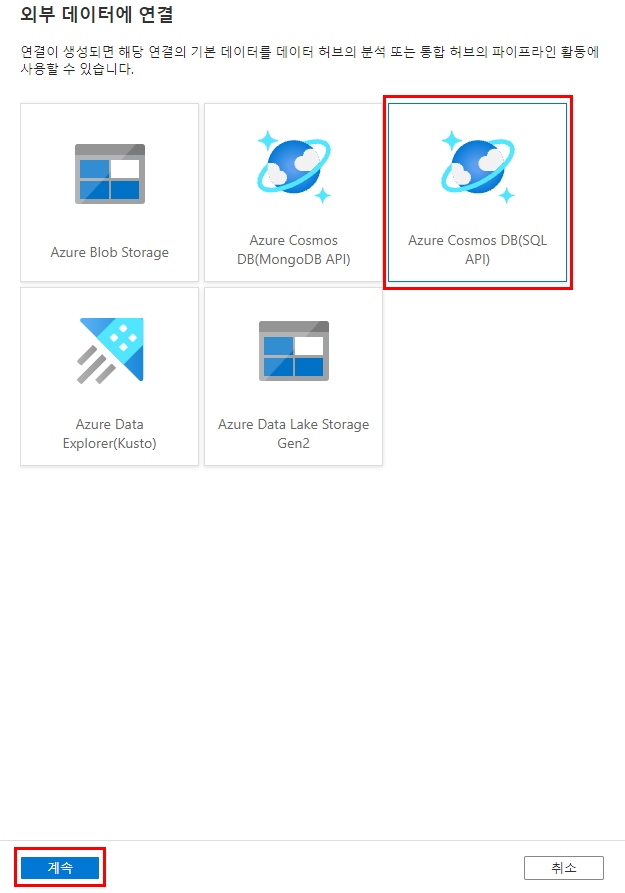


# Task 3 : Demo 02. IoT Retail Data + AutoML

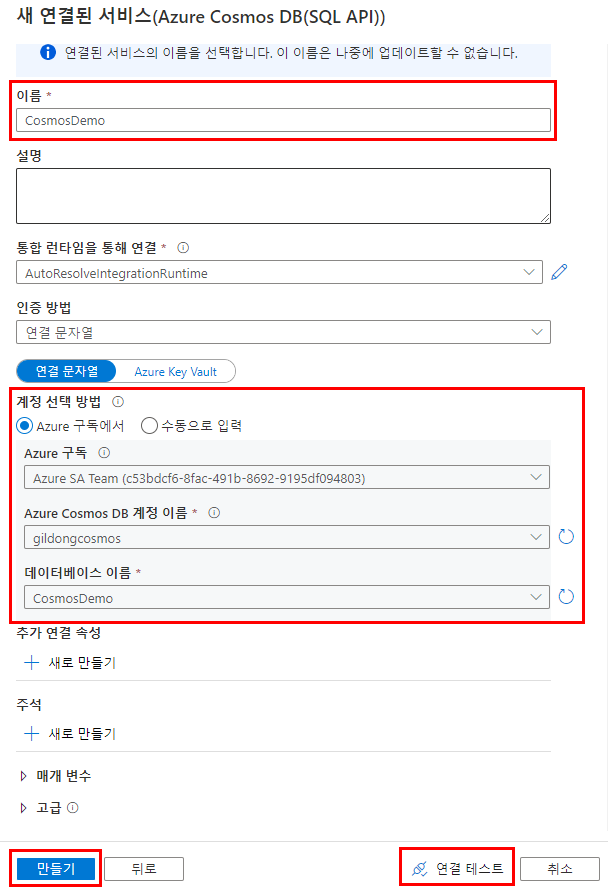
1. **데이터 허브**로 이동합니다. 상단의 **+**를 클릭하고 **외부 데이터에 연결**을 클릭합니다.



1. Azure Cosmos DB(SQL API)를 선택합니다. 계속을 클릭합니다.



1. **Linked Service**의 **이름과 Cosmos DB의 정보**를 입력합니다. **연결 테스트**를 진행하고 성공적으로 연결이 되는 것을 확인한 후, **만들기**를 클릭합니다.



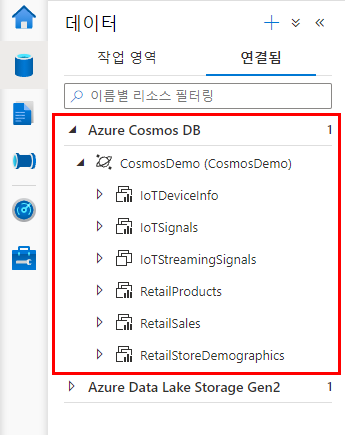
이름 : CosmosDemo

Azure 구독 : Cosmos DB를 만든 구독 선택(현재 구독)

Azure Cosmos DB 계정 이름 : gildongcosmos

데이터베이스 이름 : CosmosDemo

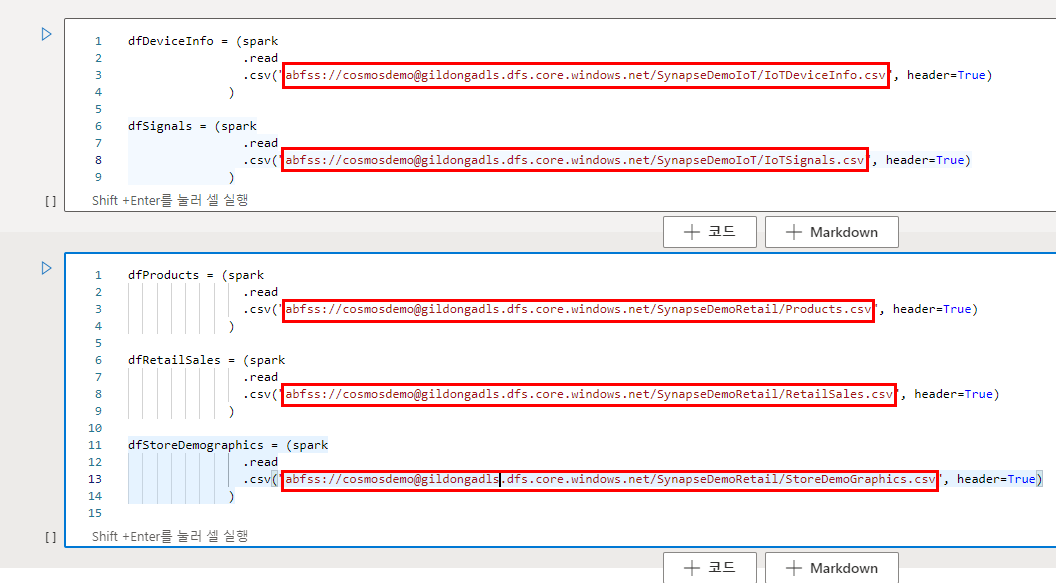
1. **Refresh**하여 생성된 Link를 확인합니다.



1. **Demo 02**번파일도 Demo 01 파일과 같은 단계를 거쳐서 실행합니다. 연결대상에 spark pool을 할당하고 실행하면서 결과를 조회합니다.



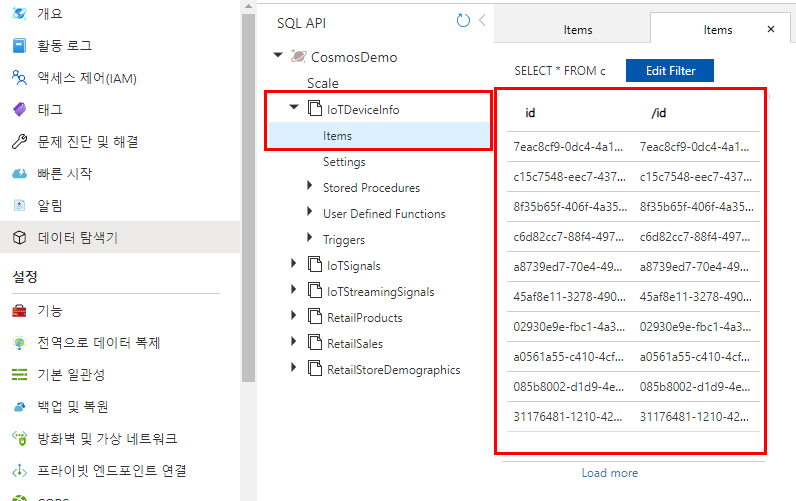
1. **2번 셀**까지 ADLS Gen2 Account를 변경합니다.



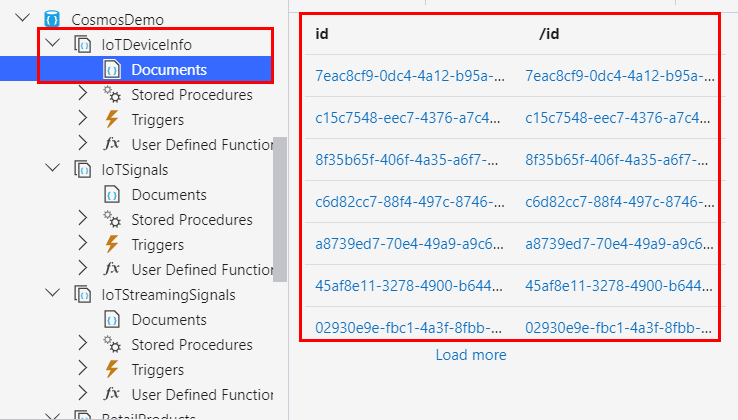
abfss://cosmosdemo@**gildongadls**.dfs.core.windows.net/SynapseDemoIoT/IoTDeviceInfo.csv

1. **4번 셀**까지 순차적으로 실행합니다. Cosmos DB의 데이터 탐색기 혹은 Storage Explorer에서 들어온 데이터를 확인합니다. **IoTDeviceInfo, IoTSignals, RetailProducts, RetailSales, RetailStoreDemographics**에 들어온 데이터를 확인합니다.

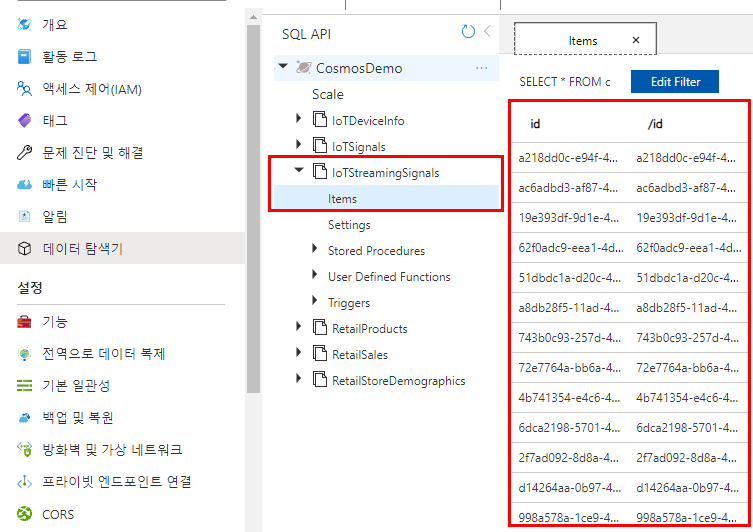
**[Portal의 데이터 탐색기]**



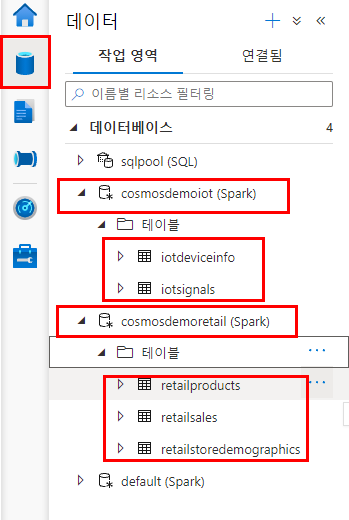
**[Storage Explorer]**



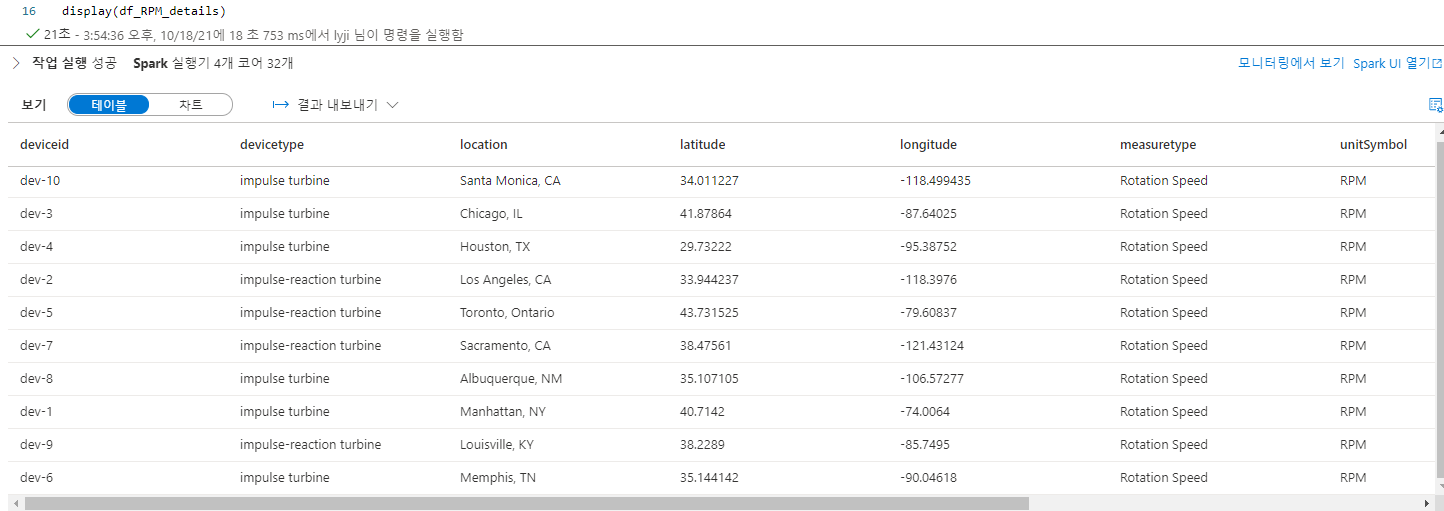
1. **5, 6번 셀**을 실행하여 Streaming 데이터 Insert를 준비합니다. **Cosmos DB의 IoTStreamingSignal 컨테이너의 비어있는 데이터를 확인**합니다.
2. **7번 셀**을 실행하고 바로 **Cosmos DB의 데이터를 확인**합니다. **초당 10개씩 데이터가 Insert**됩니다. 약 **2분정도 작업이 실행**됩니다.

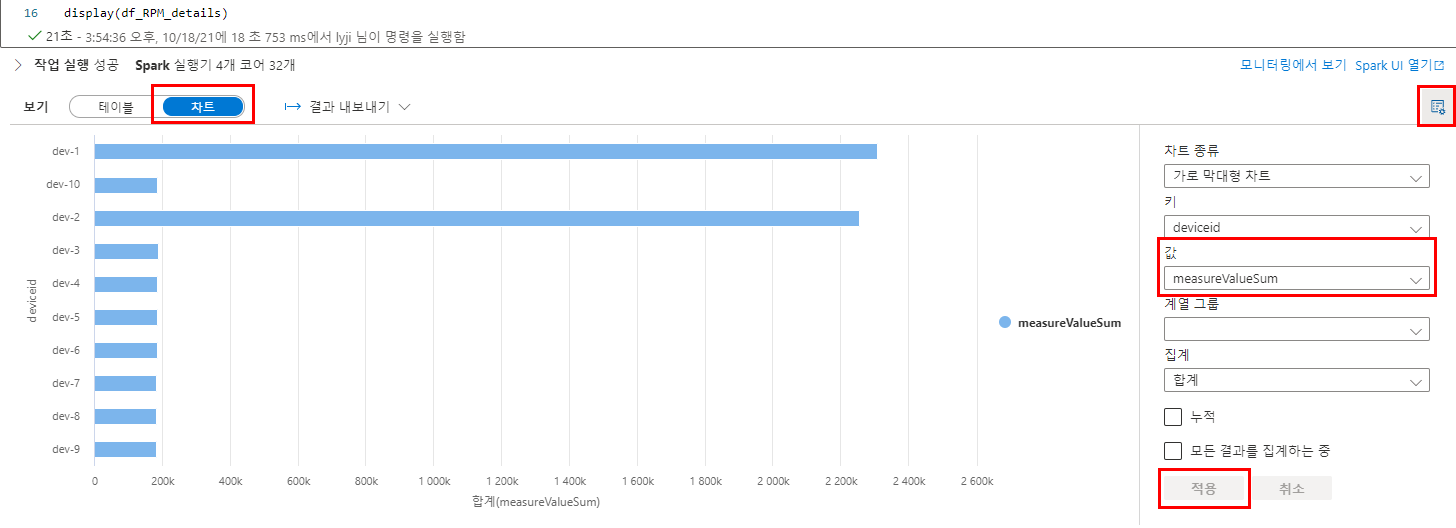


1. **8~14번 셀**을 실행합니다. 데이터베이스와 테이블을 생성합니다. 데이터 허브에서 결과를 확인합니다.

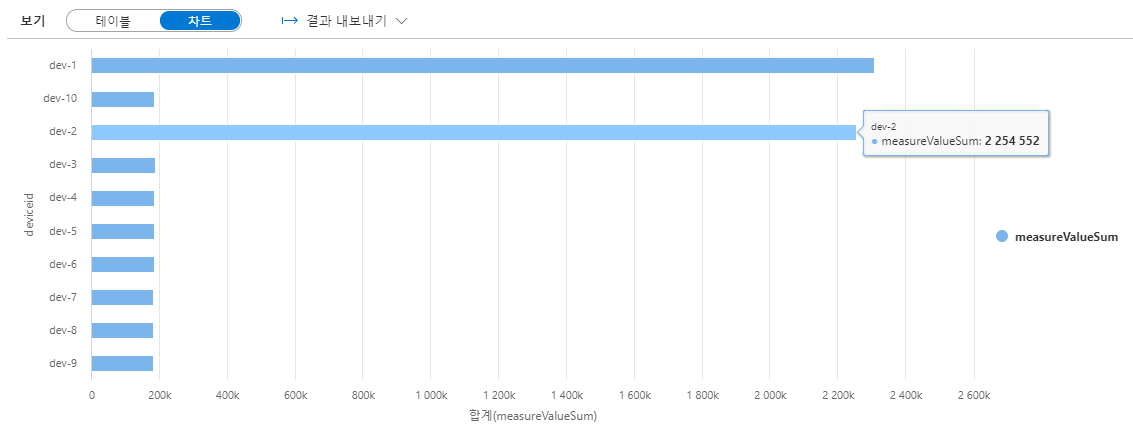


1. **15번 셀**을 실행합니다. 결과를 확인합니다. **보기 옵션을 차트로 변경**하고 설정 창에서 **값을 measureValueSum으로 변경**합니다. 적용을 클릭합니다. 손쉽게 조회 결과를 차트로 변환해볼 수 있습니다.

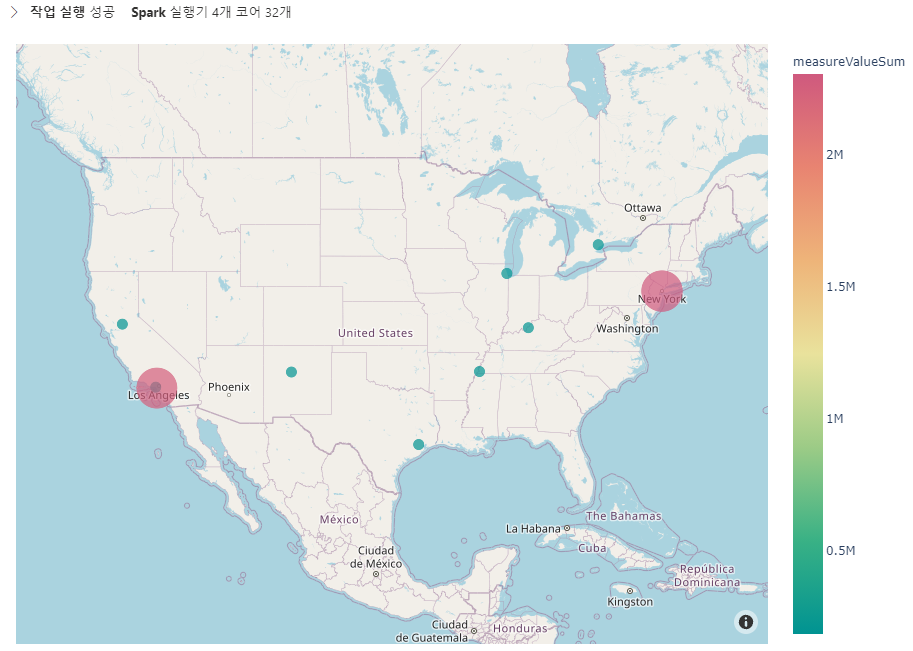




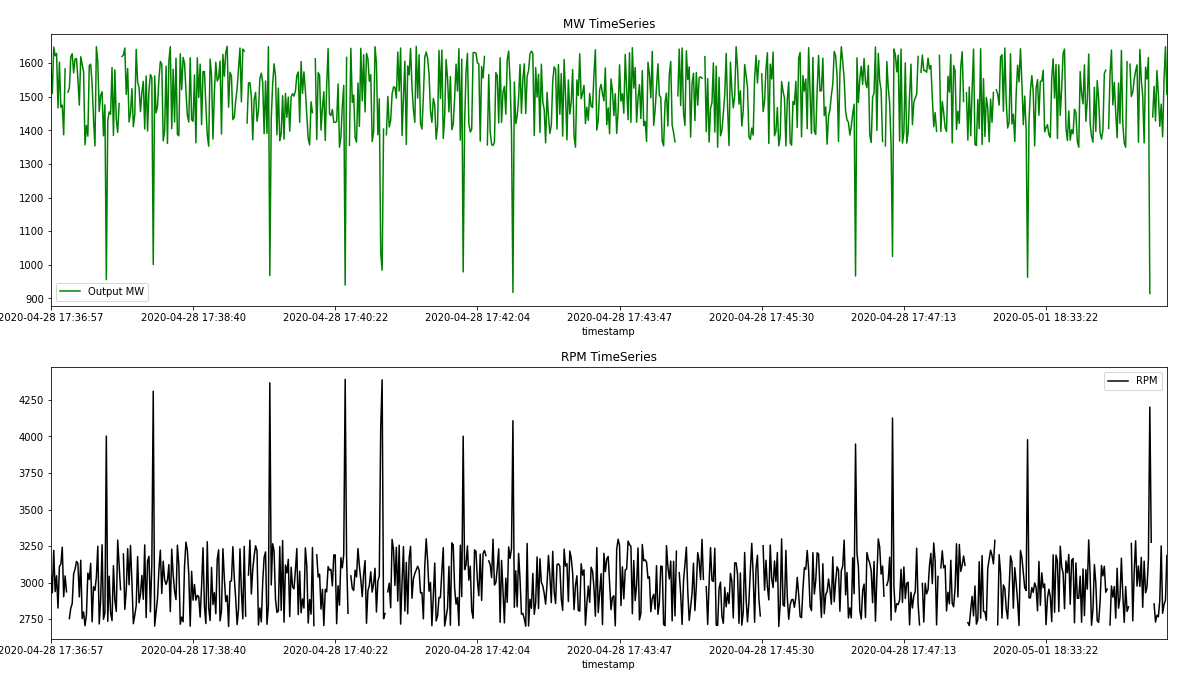
1. 조회결과, **Device ID** **1번과 2번**의 수치가 높은 것을 볼 수 있습니다.



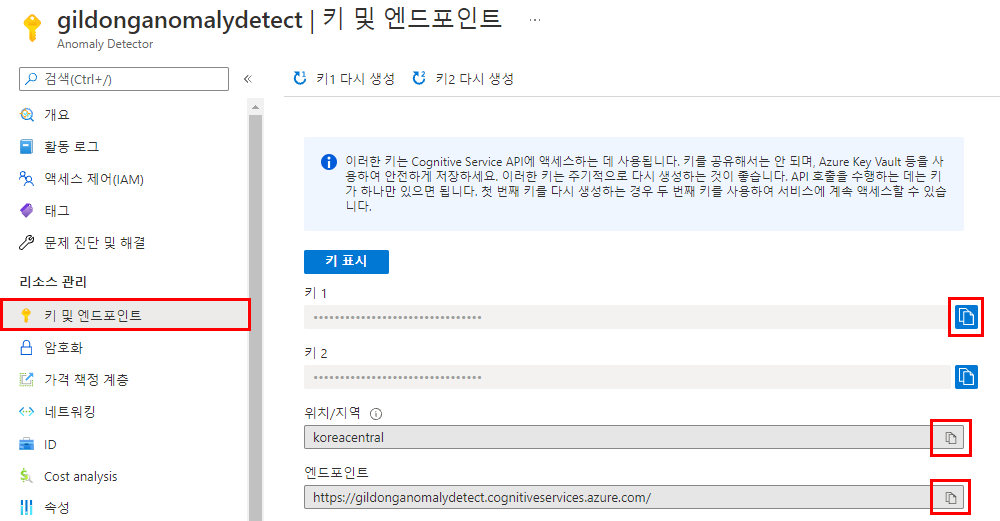
1. **16번 셀**을 실행하여 지리적 분포를 확인해봅니다. **뉴욕과 LA**에서 수치가 높게 나타납니다.

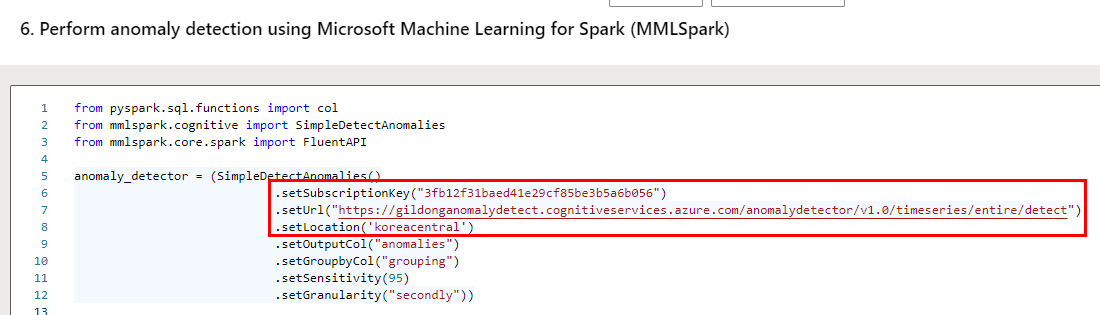


1. **17, 18번 셀**을 실행합니다. 수치가 높게 나온 디바이스 중 하나인 **Dev-1에 대해 unitSymbol을 기준으로 MW와 RPM에 Data를 확인**합니다.

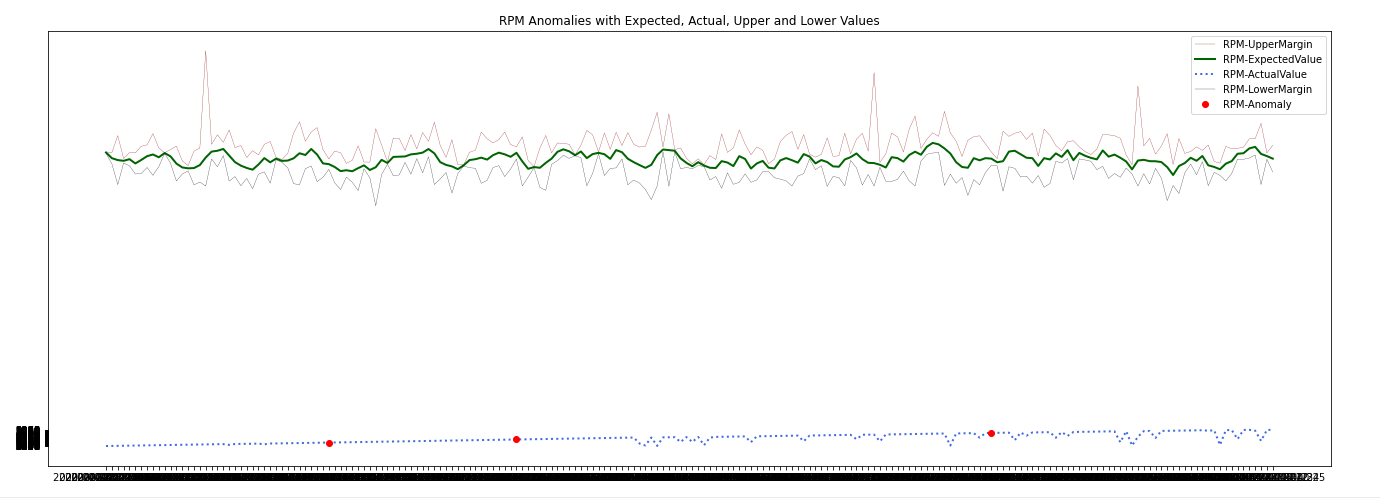


1. **19번 셀**을 실행하기 전, Anomaly Detector 부분을 적절하게 수정해야 합니다. **Anomaly Detector로 이동**합니다. **키 및 엔드포인트**에서 **키와 위치/지역, 엔드포인트**를 복사해서 아래와 같이 셀을 수정합니다.

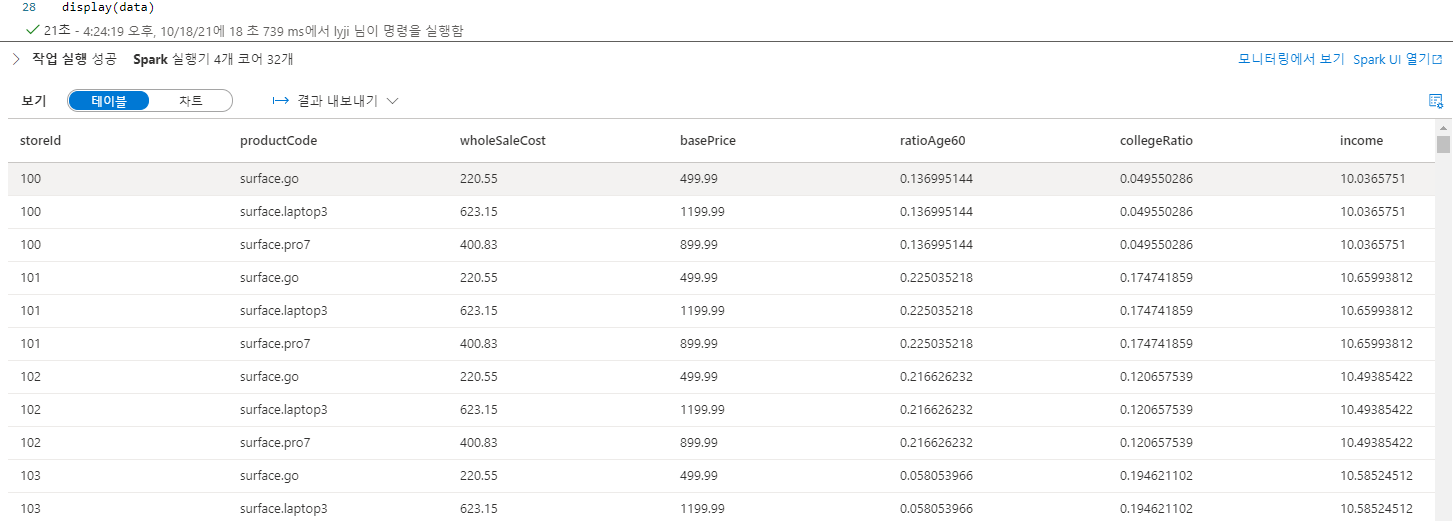




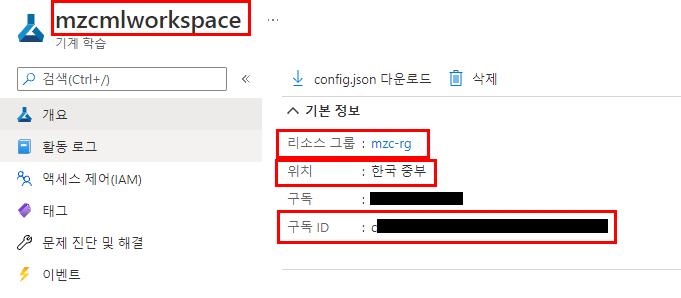
1. **19, 20번 셀**을 실행합니다. Anomaly Detector로부터 받은 결과를 조회합니다. **21번 셀**을 실행하여 결과를 시각화합니다. 결과는 아래와 같습니다.

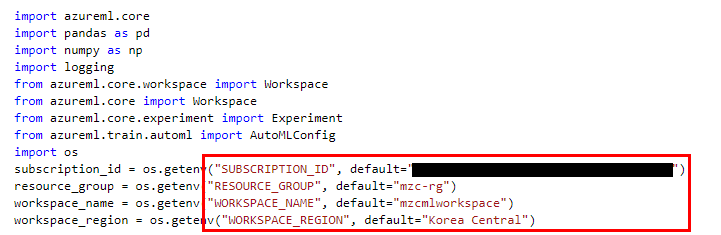


1. 다음은 Retail 데이터를 갖고 Machine Learning을 해보겠습니다. **22번 셀**을 실행하여 Retail 테이블을 조인합니다. 새로운 데이터셋을 생성합니다.



1. **23번 셀**을 실행하기 전 Azure ML로 이동합니다. **구독ID, 리소스 그룹, 워크스페이스 이름, 지역 정보**를 보고 셀에 입력합니다.



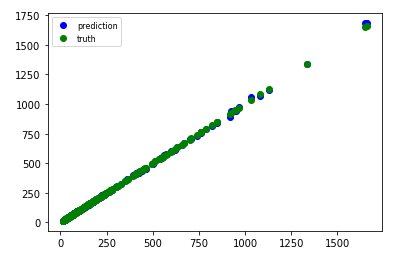


Resource Group: mzc-rg

Workspace Name: mzcmlworkspace

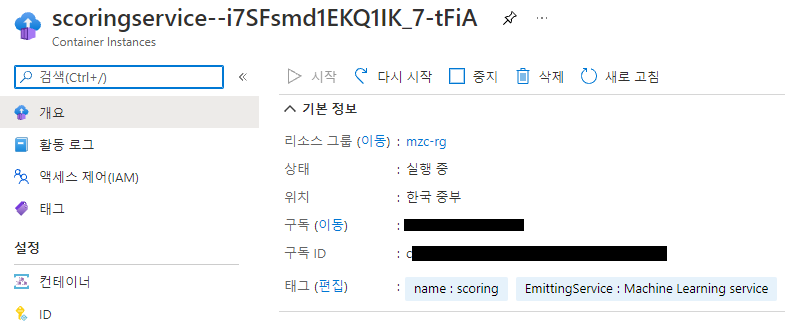
Workspace Region: Korea Central

1. **정보를 입력한 후 23번 셀을 실행**합니다. **24번 셀을 실행하여 Data Prep을 진행**하고 **25번 셀을 실행하여 모델을 빌드**합니다. 모델을 만드는 데 시간이 15분 정도 소요될 수 있습니다.
2. **26번 셀**을 실행하여 **Test Data로 Prediction을 수행**합니다.
3. **27번 셀**을 실행하여 결과를 **Plotting**합니다.

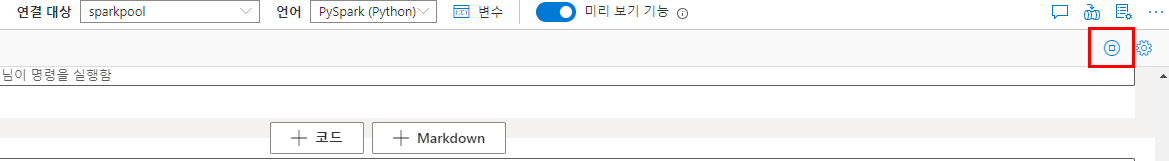


1. **28, 29번 셀**을 실행하여 모델을 컨테이너로 배포합니다. 컨테이너가 새로 생성됩니다.



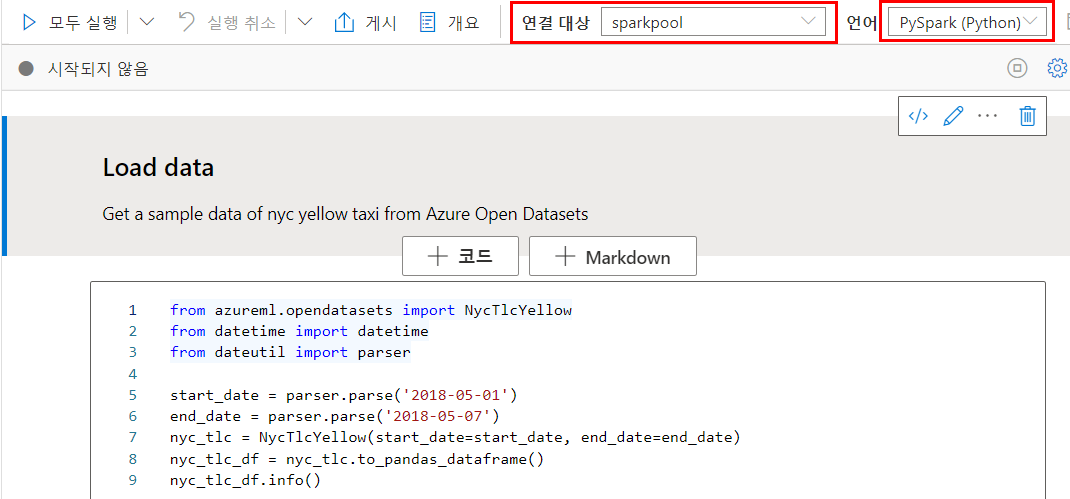


1. 나머지 셀들을 실행하여 **데이터베이스와 테이블을 정리**합니다.
2. 노트북을 닫기 전에 **세션 종료 버튼을 눌러 세션을 종료합니다.**

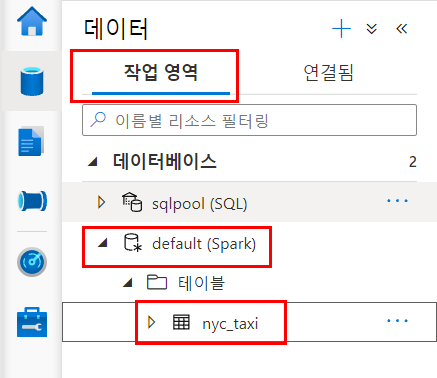


# Task 4 : Demo 03. Create Spark Table with NYC Taxi Data + AutoML

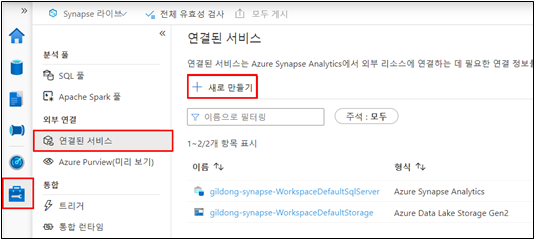
1. **Demo 03**번파일도 같은 단계를 거쳐서 실행합니다. 연결대상에 spark pool을 할당하고 실행하면서 결과를 조회합니다.



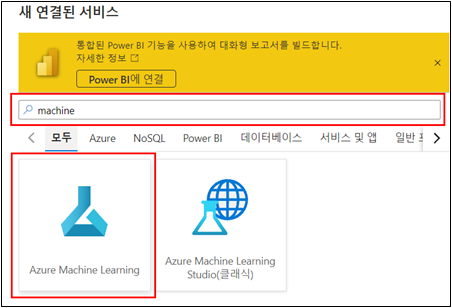
1. **2번 셀**까지 차례대로 실행하면서 **데이터를 로드**합니다.
2. **4번 셀**까지 차례대로 실행하면서 **데이터를 전처리**합니다.
3. **5번 셀**을 실행합니다**. Spark table로 데이터를 저장**합니다. **데이터 허브에서 확인**합니다.

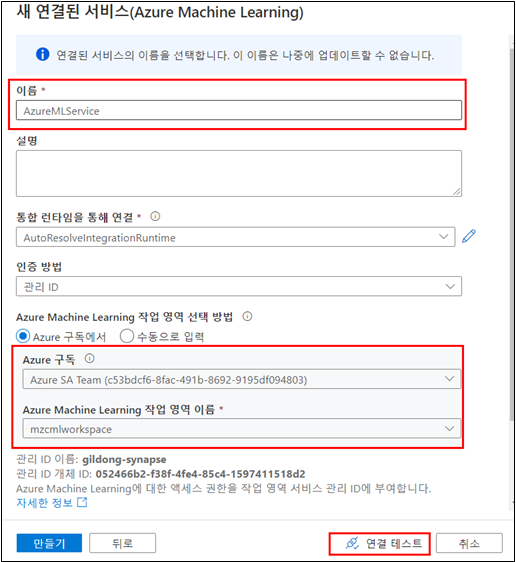


1. **관리 허브**로 이동합니다. **연결된 서비스**로 이동하여 **+새로 만들기**를 클릭합니다.



1. **Azure Machine Learning**을 검색하여 선택합니다. **계속**을 눌러 **연결된 서비스로 등록**합니다. **이름**과 **Azure Machine Learning Workspace**를 설정합니다. **연결테스트**를 눌러 연결이 잘 이루어지는지 테스트합니다.





이름: AzureMLService

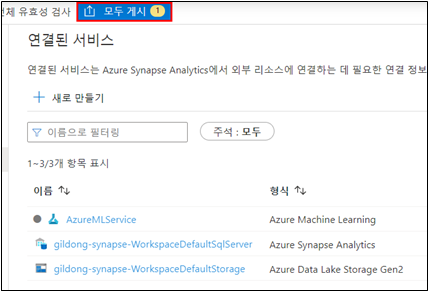
Azure 구독: 구독 선택

Azure Machine Learning 작업 영역 이름: 구독 선택 후 ML 워크스페이스 선택

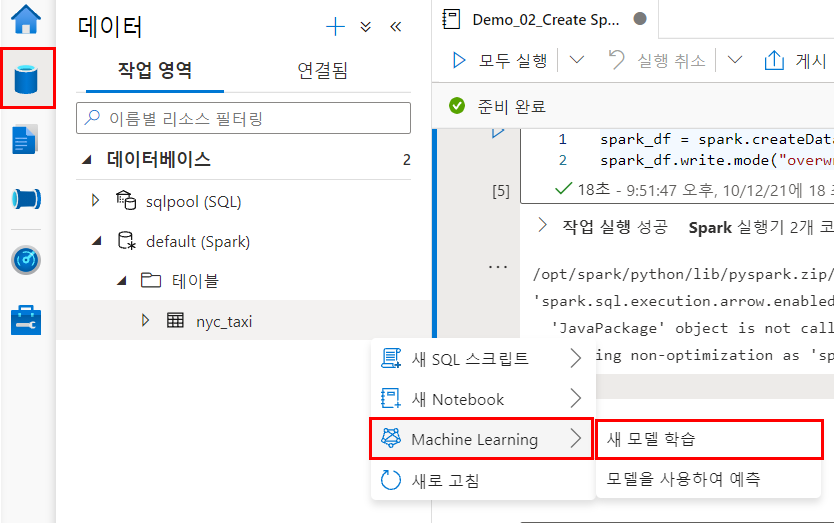
1. 연결 테스트를 성공하면 만들기를 눌러 생성합니다.



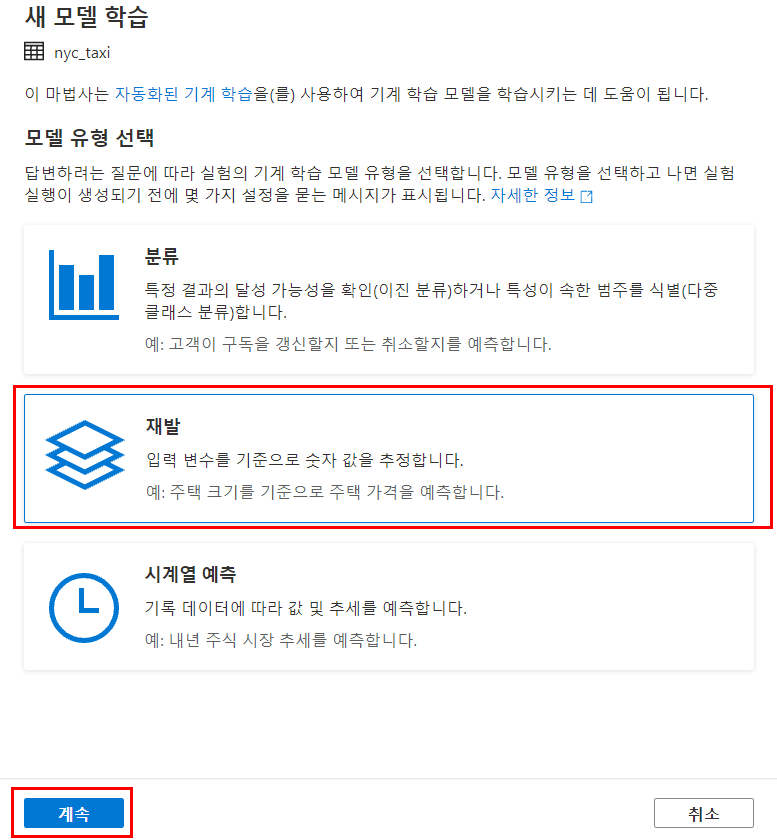
1. **모두 게시**를 눌러 현재 상태를 저장합니다.



1. **데이터 허브**로 이동합니다. 생성된 Spark 테이블 오른쪽 **점 세 개(…)**를 클릭합니다. **Machine Learning – 새 모델 학습**을 클릭합니다.



1. **Regression(회귀분석, 재발)**을 선택합니다. **계속**을 클릭합니다.



1. **Azure Machine Learning 작업 영역, 실험 이름, 최적 모델 이름**을 설정(기본값 유지)합니다. **대상 열은 fareAmount**를 선택합니다**. Spark 풀**을 선택하고 **계속**을 클릭합니다.



대상 열: fareAmount

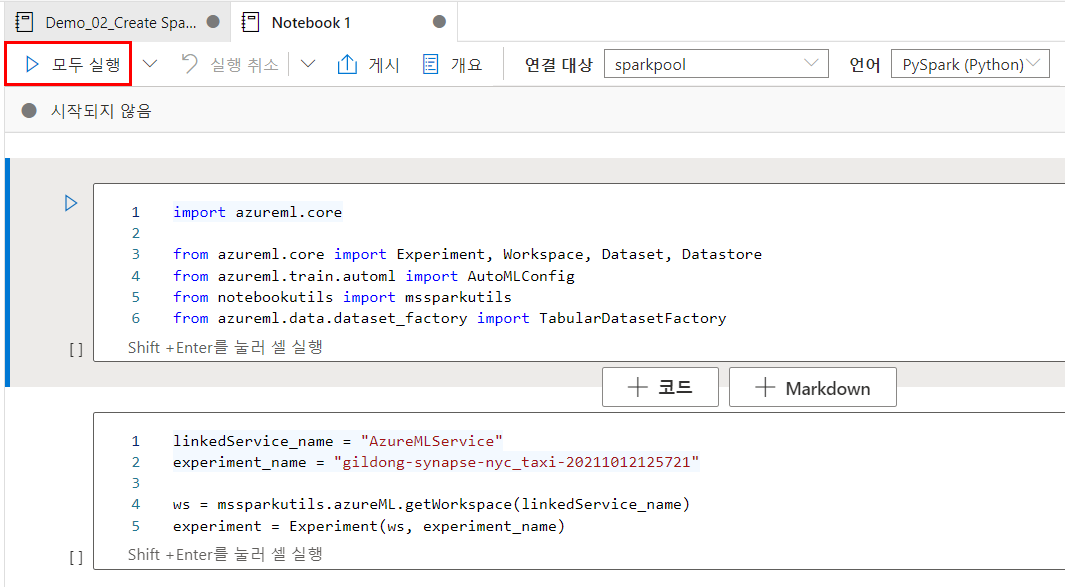
1. 학습 모델의 구성을 설정합니다. **최대 교육기간(시간)을 최솟값인 0.25(15분)로 설정**합니다. Notebook에서 열기를 클릭합니다.



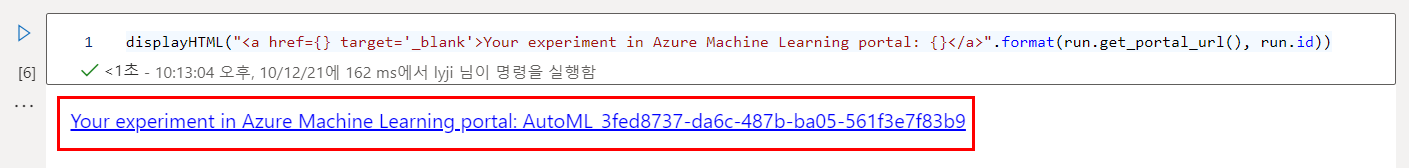
최대 교육 기간(시간) : 0.25

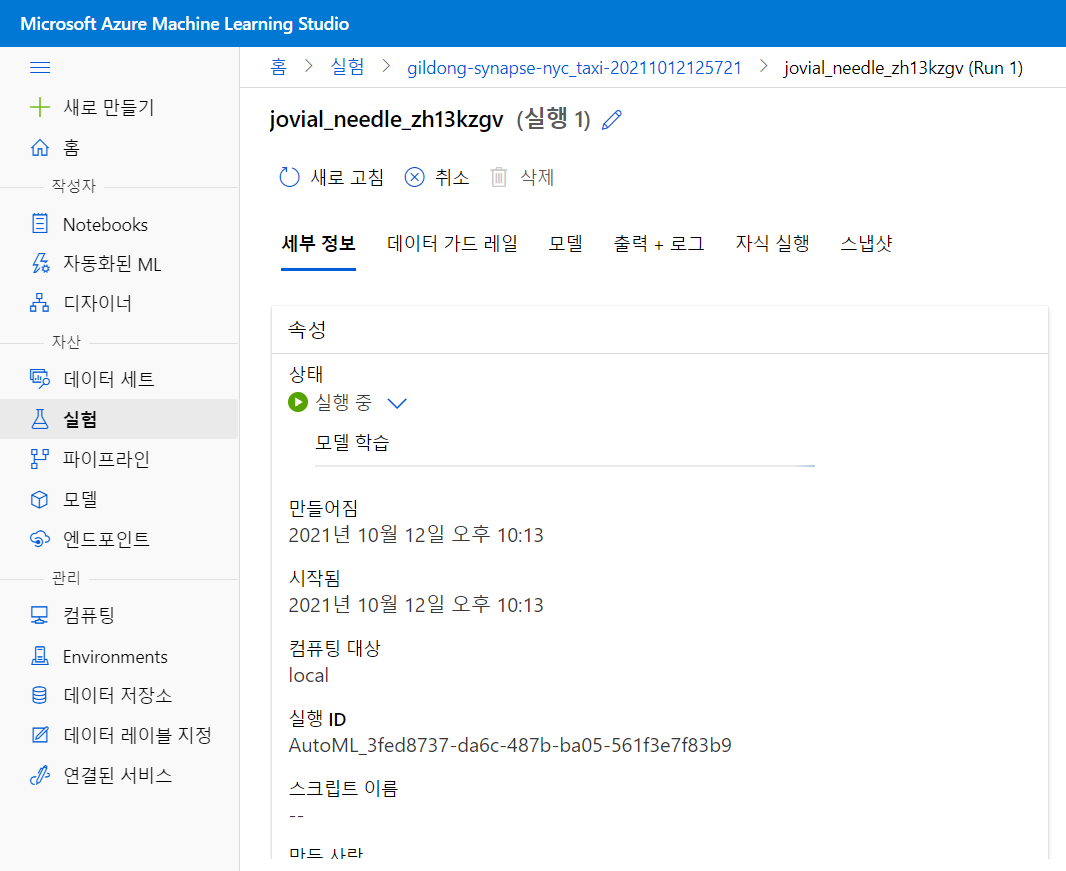
**\*실행 만들기**를 클릭하면 **바로 Machine Learning이 시작**됩니다. 시작이 시작되고 있다는 **알림이 표시**되며 성공을 나타내는 알림이 표시됩니다.

1. **모델을 만드는 Notebook이 생성**되었습니다. **모두 실행**을 클릭하여 모델을 만듭니다. 약**15분 정도 시간이 소요될 수 있습니다.**

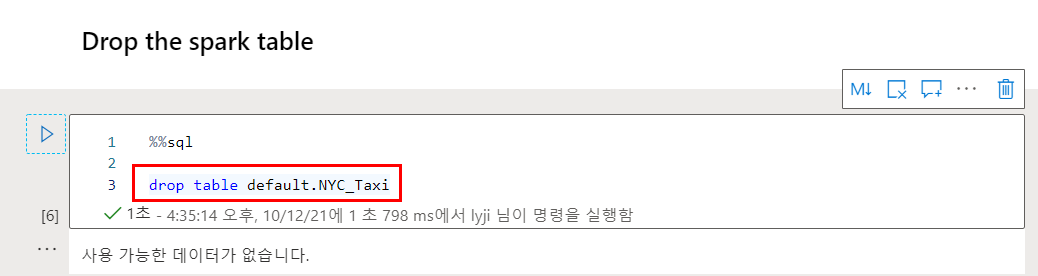


1. **6번 셀**의 링크를 통해 진행과정을 모니터링할 수 있습니다.

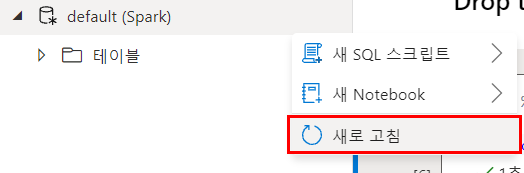


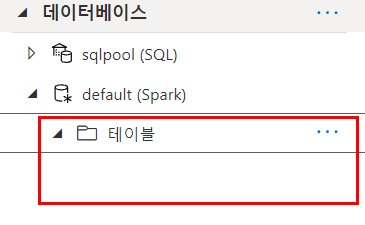


1. **다시 Demo 03번** 파일로 돌아옵니다. **6번 셀**의 주석을 풀고 실행합니다. **Spark table을 Drop** 합니다.

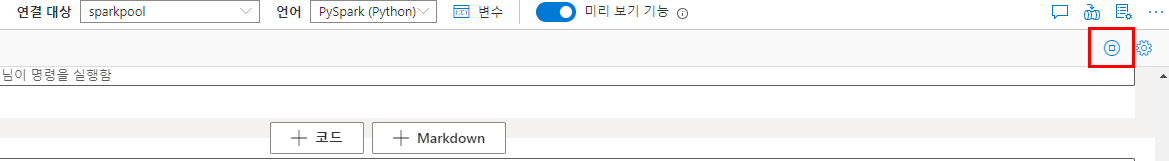


1. **새로고침** 후 데이터베이스를 확인합니다.

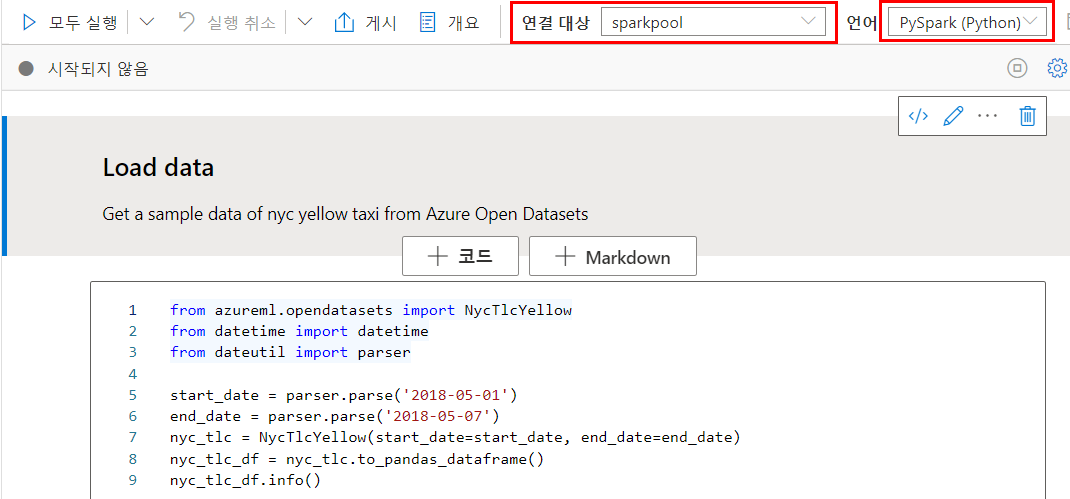


1. 노트북을 닫기 전에 **세션 종료 버튼을 눌러 세션을 종료합니다.**

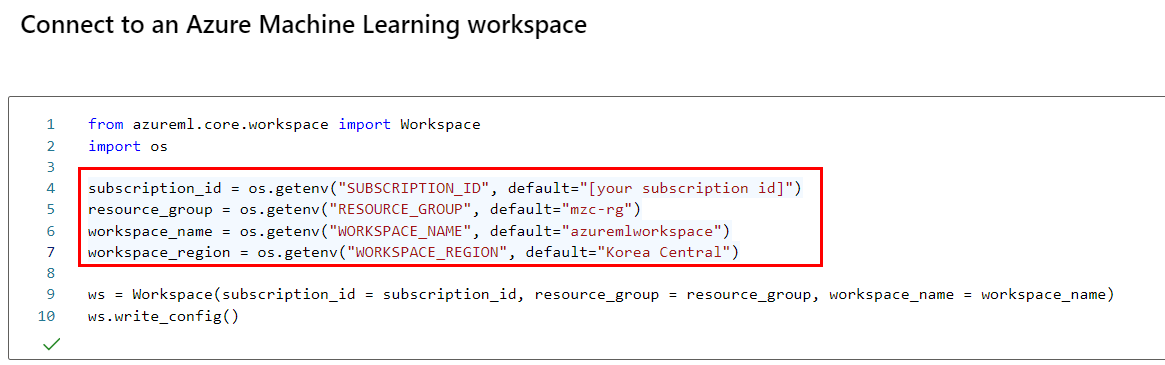


# Task 5 : Demo 04. NYC Taxi Fare AutoML Notebook from scratch

1. **Demo 04번** 파일도 같은 단계를 거쳐서 실행합니다. 연결대상에 spark pool을 할당하고 실행하면서 결과를 조회합니다. 7번 셀까지 실행합니다.



1. **8번셀**은**Azure Machine Learning workspace** 연결하는 부분입니다. 구독 ID, 리소스 그룹, 워크스페이스 이름, 워크스페이스 지역을 알맞게 입력합니다.



Resource Group: mzc-rg

Workspace Name: mzcmlworkspace

Workspace Region: Korea Central

1. 나머지 셀들을 실행하면서 결과를 조회합니다.
2. 노트북을 닫기 전에 **세션 종료 버튼을 눌러 세션을 종료합니다.**

