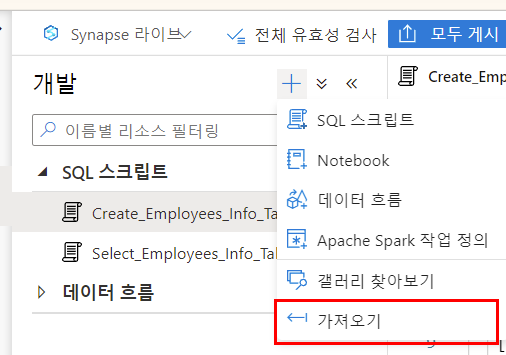
**작업 n : Data Join from two source files**

1. 시나리오



1. Develop 탭에서 +를 누른 후 import를 통해 실습에 필요한 ‘Create\_Employees\_Info\_Table’, ‘Select\_Employees\_Info\_Table’ 스크립트를 가져옵니다.



1. 왼쪽 메뉴의 Develop 메뉴를 선택하고 Create\_Employees\_Info\_Table 스크립트를 열어 실행합니다.
2. Data 탭으로 이동하여 + 버튼을 누르고 Data Set을 생성합니다.
3. 이전과 동일하게 ADLS Gen2를 선택 후 Delimited Text를 선택합니다.
4. 아래와 같이 Property를 입력합니다.

|  |  |
| --- | --- |
| Field | Value |
| Name | hr\_employees\_csv |
| Linked Service | Select <ADLS Gen2 Account Name> |
| File Path – Container | Enter pipeline |
| File Path – Directory | Enter Join |
| File Path – File | Enter hr\_employees.csv |
| First row as header | Checked |
| Import schema | Select From Connection/store |

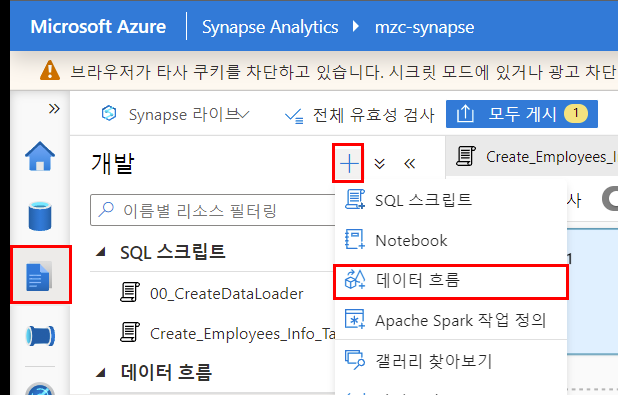
1. 다시 + 를 누르고 Data Set을 생성합니다.
2. 바로 전 작업과 동일하게 ADLS Gen2를 선택 후 Delimited Text를 선택합니다.
3. 아래와 같이 Property를 입력합니다.

|  |  |
| --- | --- |
| Field | Value |
| Name | hr\_departments\_csv |
| Linked Service | Select <ADLS Gen2 Account Name> |
| File Path – Container | Enter pipeline |
| File Path – Directory | Enter Join |
| File Path – File | Enter hr\_departments.csv |
| First row as header | Checked |
| Import schema | Select From Connection/store |

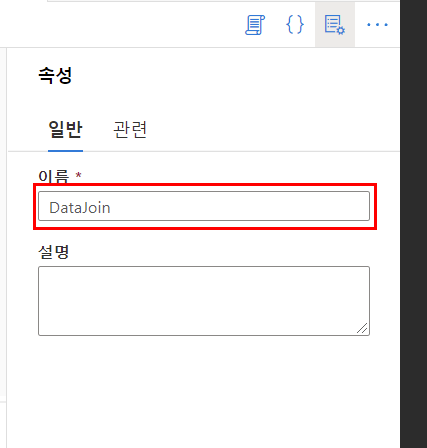
1. 다시 + 를 누르고 Data Set을 생성합니다.
2. 이번엔 Azure Synapse dedicated SQL Pool을 선택합니다.
3. 아래와 같이 Property를 입력합니다.

|  |  |
| --- | --- |
| Field | Value |
| Name | HumanResource\_asa |
| Linked Service | Select WorkspaceDefault |
| Table Name | Select dbo.Employees\_Info |
| Import schema | Select From Connection/store |

1. Publish All 하여 저장합니다.
2. Develop탭으로 돌아와서 +를 누르고 Data Flow를 생성합니다.



1. 이름에 DataJoin을 입력합니다.



1. 소스 추가를 하여 아래와 같이 Property를 입력합니다.



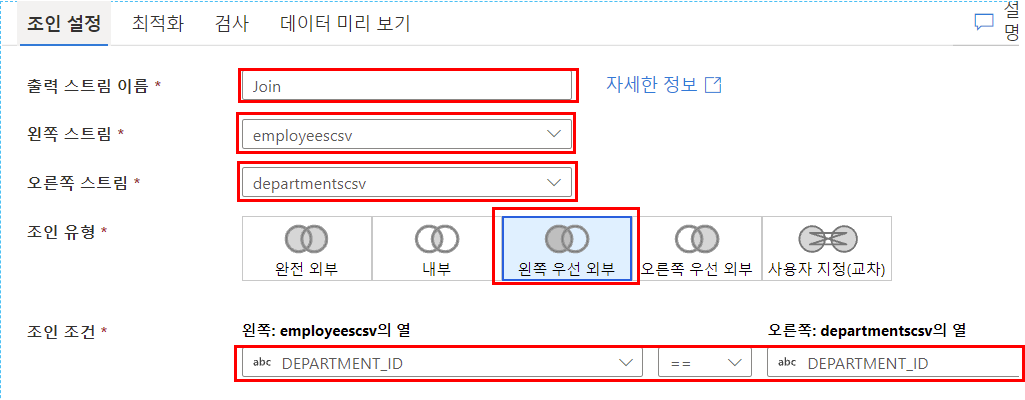
1. 이전 작업과 동일하게 소스를 하나 더 추가하여 Property를 아래와 같이 입력합니다.



1. 아래 +를 클릭하여 Join을 선택합니다.



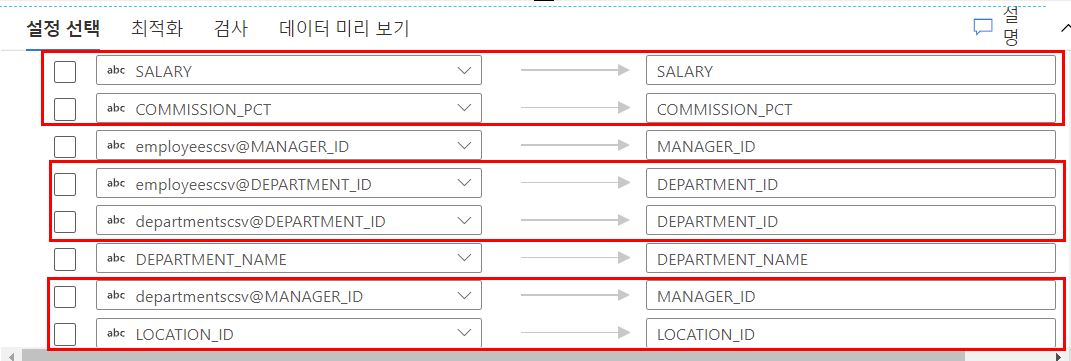
1. Output Stream Name을 Join으로 바꿔주고 Right Stream에 departmentscsv를 선택합니다.
2. employeescsv파일이 기준이 되어 데이터를 조인하기 때문에 조인 유형은 왼쪽 우선 외부를 선택합니다.
3. 조인 조건으로는 DEPARTMENT\_ID를 선택합니다.



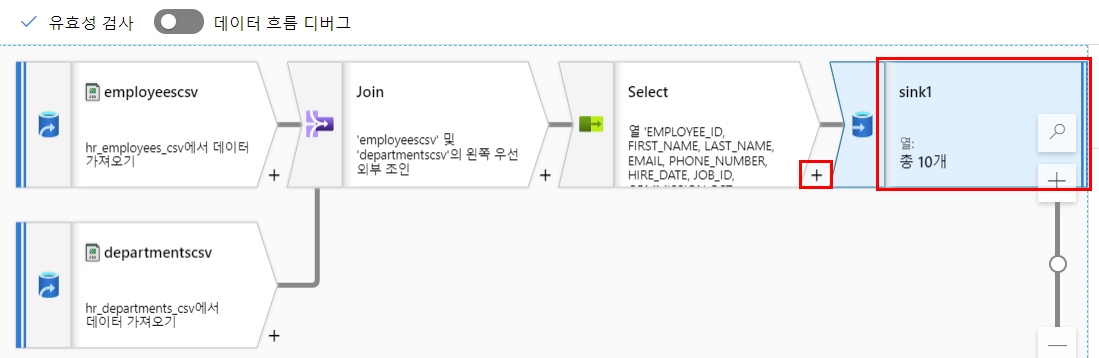
1. 다시 아래 +를 누른 후, Select를 선택합니다.



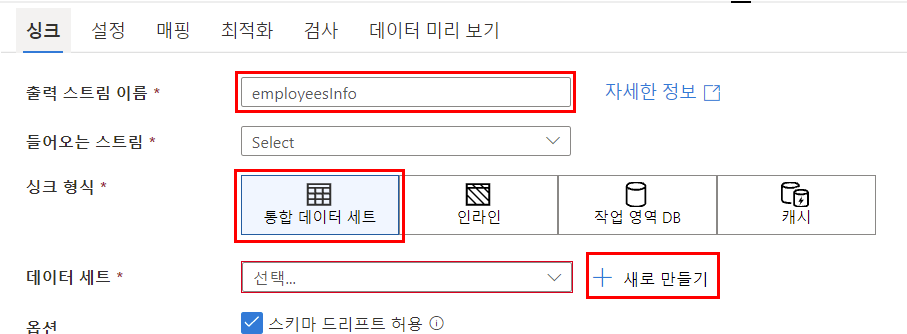
1. Output Stream Name을 Select으로 바꿔주고 설정 선택에서 필요 없는 열을 삭제합니다.



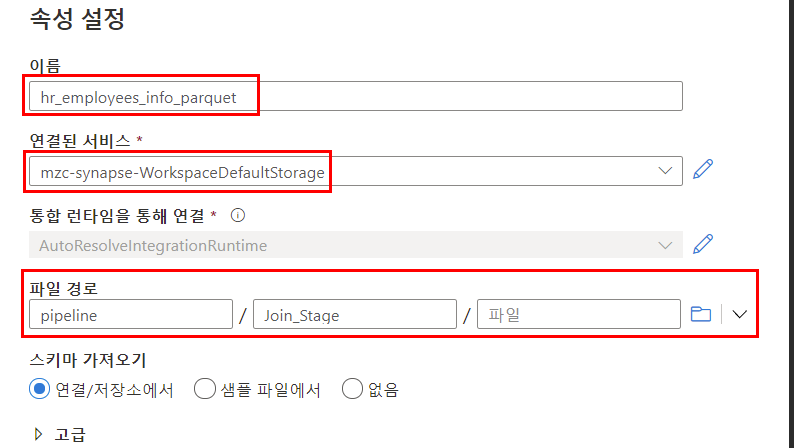
1. 마지막으로 +를 클릭하고 Sink를 선택합니다.



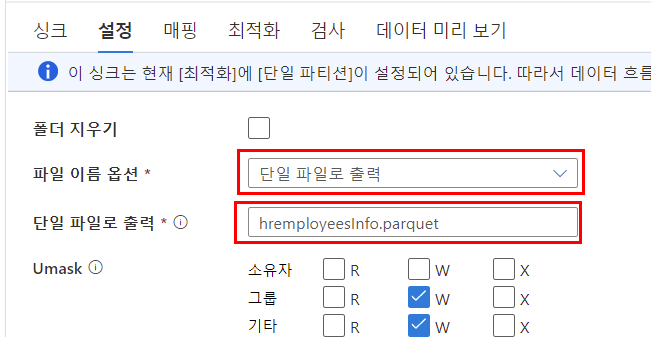
1. Output Stream Name에 employeesInfo를 입력하고 Destination Data Set을 선택할 때 +새로 만들기를 눌러 새로 생성합니다. 이번에는 ADLS Gen2의 Parquet을 선택합니다.



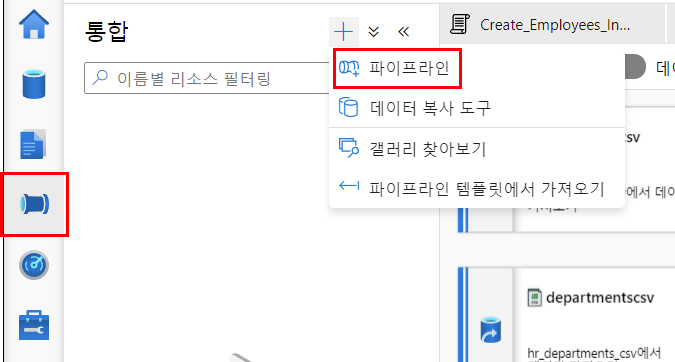
1. 이름에 hr\_employees\_info\_parquet를 입력하고 아래와 같이 Property 값을 넣어줍니다.



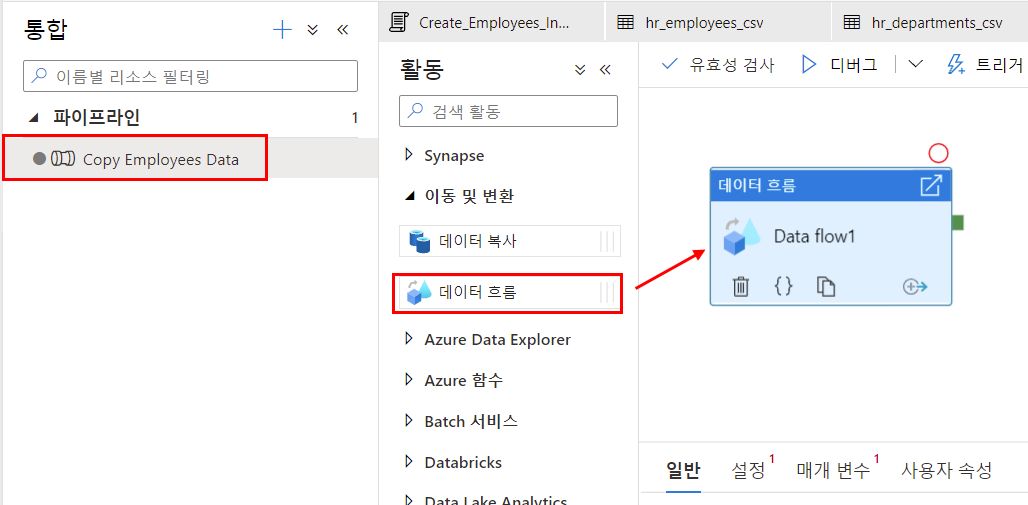
1. 설정 탭으로 가서 File Name Option에서 단일 파일로 출력을 선택합니다. 파일 명에는 hremployeesInfo.parquet를 입력합니다.(매핑 탭에서 단일 파티션 선택)



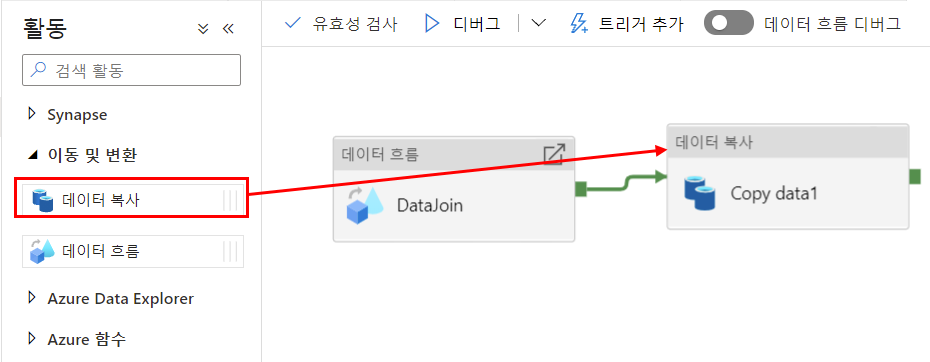
1. Publish All 하여 저장합니다.
2. Integrate탭에서 +를 클릭 후 Pipeline을 생성합니다. 이름을 Copy Employees Data로 입력합니다.



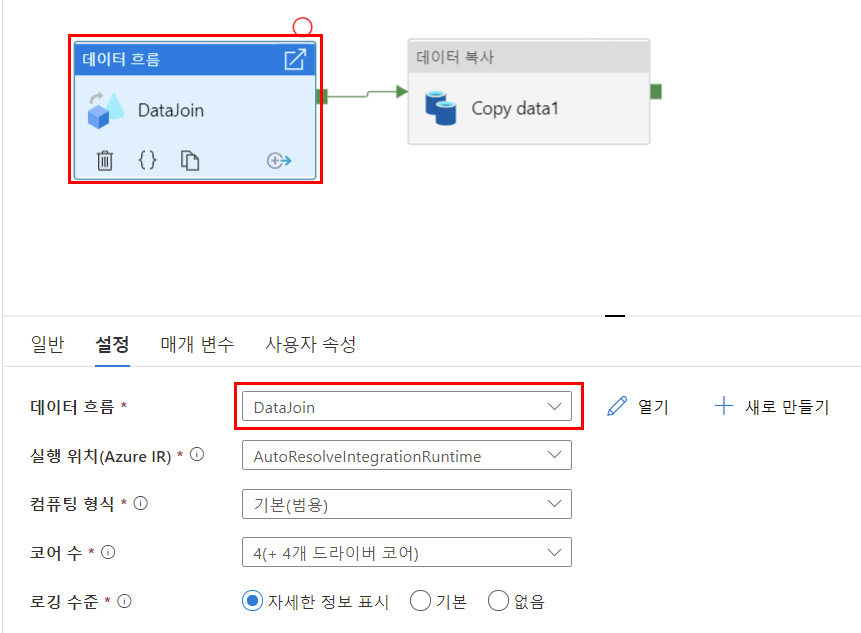
1. Data Flow를 Drag&Drop하고 방금 생성한 Data Flow를 선택합니다.



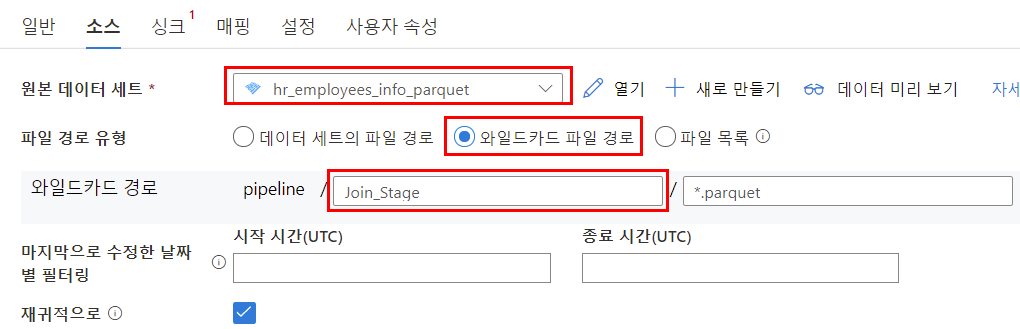
1. Copy Data를 마찬가지로 Drag&Drop하여 아래와 같이 이어줍니다.



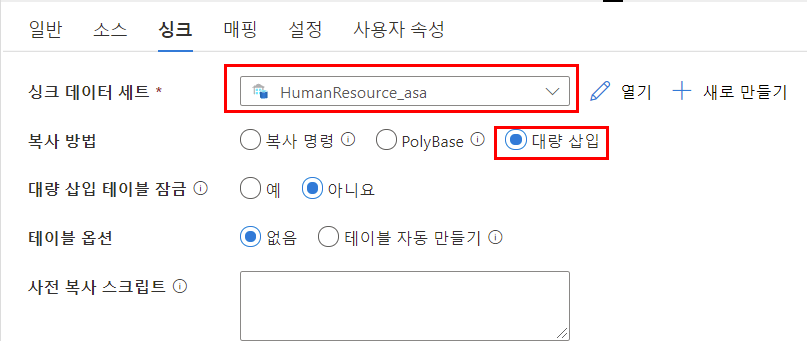
1. Data Flow를 선택하여 방금 전에 만든 DataJoin을 선택합니다.



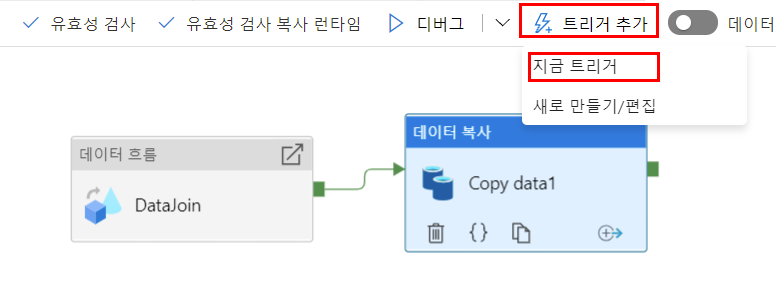
1. Copy Data를 선택하고 Source 탭으로 이동하여, hr\_employees\_info\_parquet를 Source Data Set으로 선택합니다. Wildcard file Path를 선택 후 Join\_Stage를 입력합니다.



1. Sink 탭으로 이동해서 HumanResource\_asa를 선택하고 Bulk Insert를 선택합니다.



1. Publish All하여 저장합니다.
2. Trigger Now하여 실행하고 Monitoring 탭에서 확인합니다.



1. 작업이 완료되면 Select\_Employees\_Info\_Table 스크립트를 실행하여 결과물을 확인합니다.