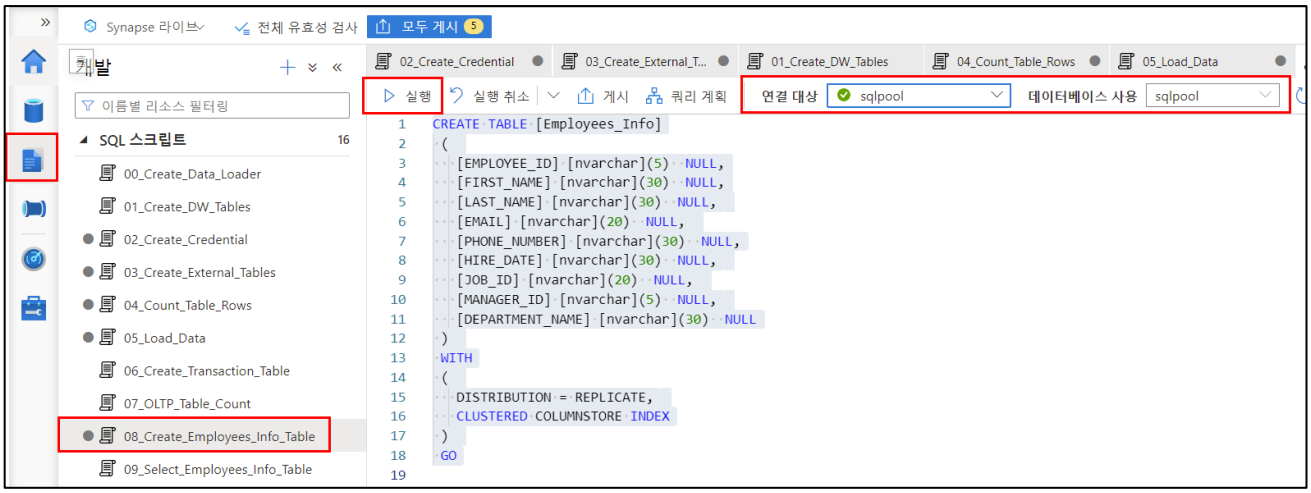
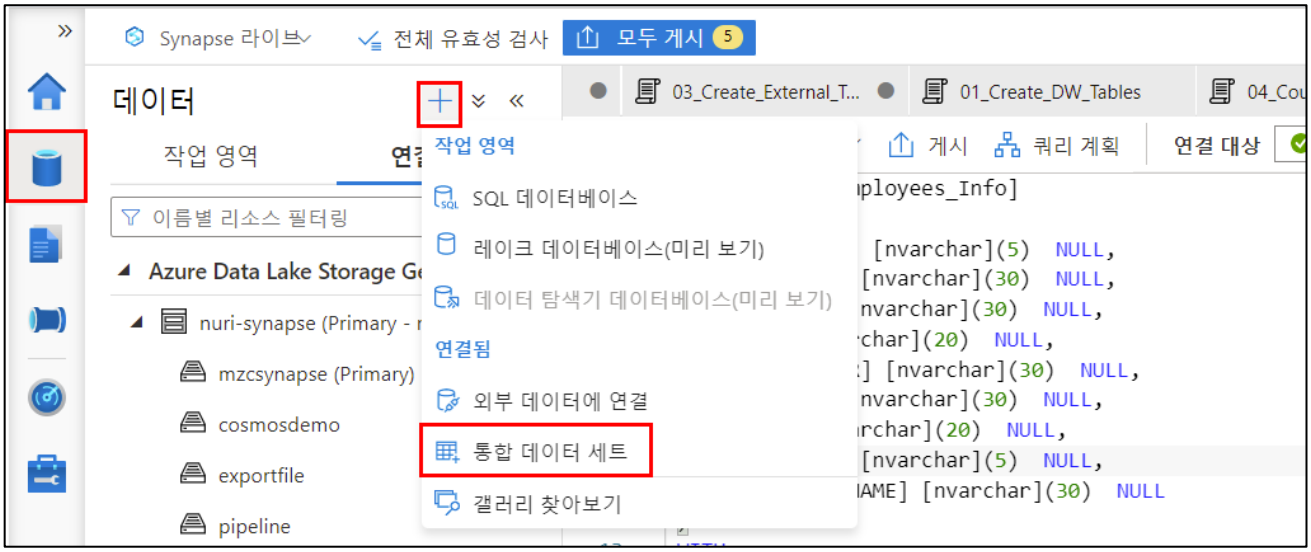
# Lab 4 – Source Files 를 통한 Data Join

## Task 1 : 하나의 테이블에 있는 열의 데이터가 충족되지 않는 경우 다른 테이블에서 필요한 열을 가져오는 데이터 조인 작업

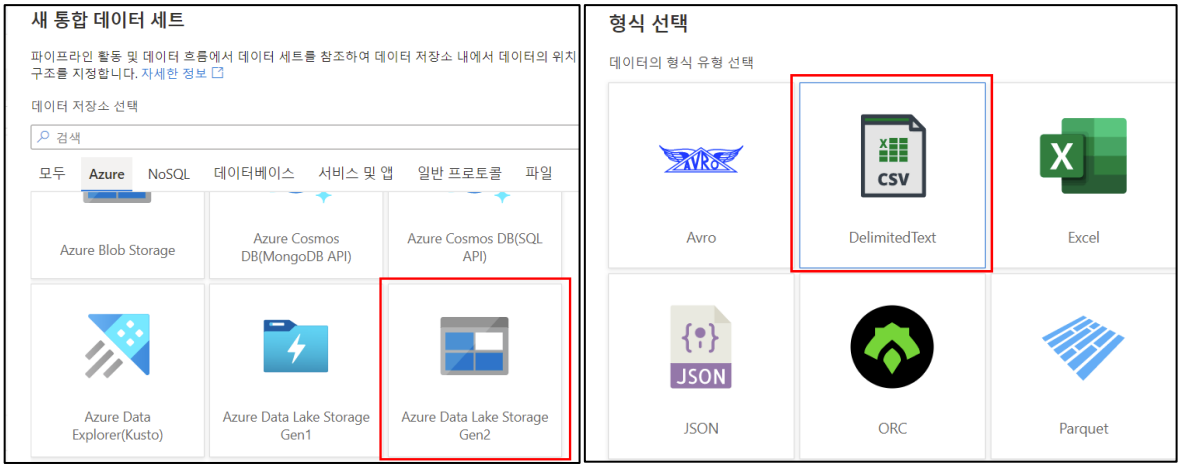
1. 개발 탭에서 **08_Create_Employees_Info_Table** 스크립트를 열어 테이블을 생성합니다.




2. 데이터 탭으로 이동하여 + 버튼을 누르고 **통합 데이터 세트**를 생성합니다.

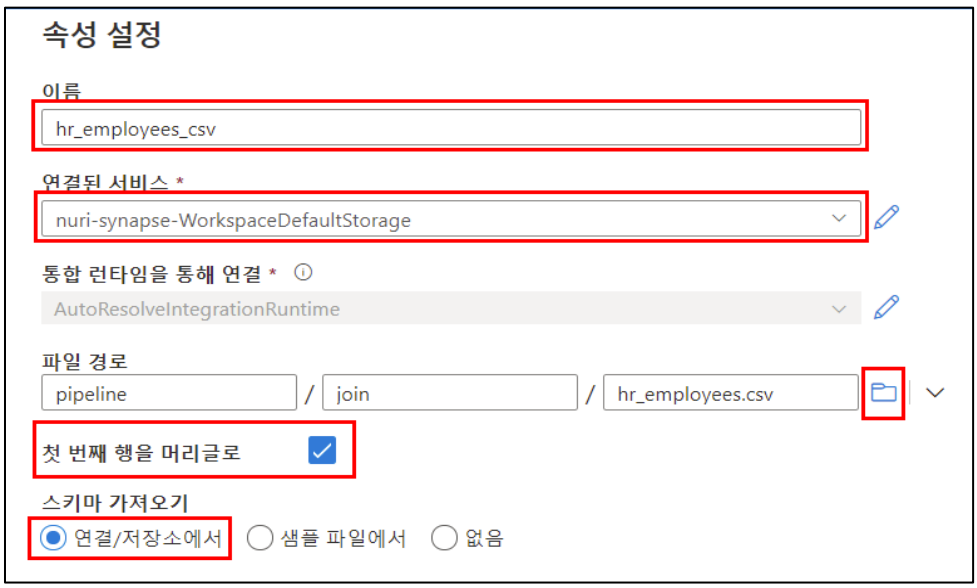


3. **Azure Data Lake Storage Gen2**를 선택 후 **Delimited Text**를 선택합니다.




4. 아래와 같이 **속성 설정**을 입력합니다.

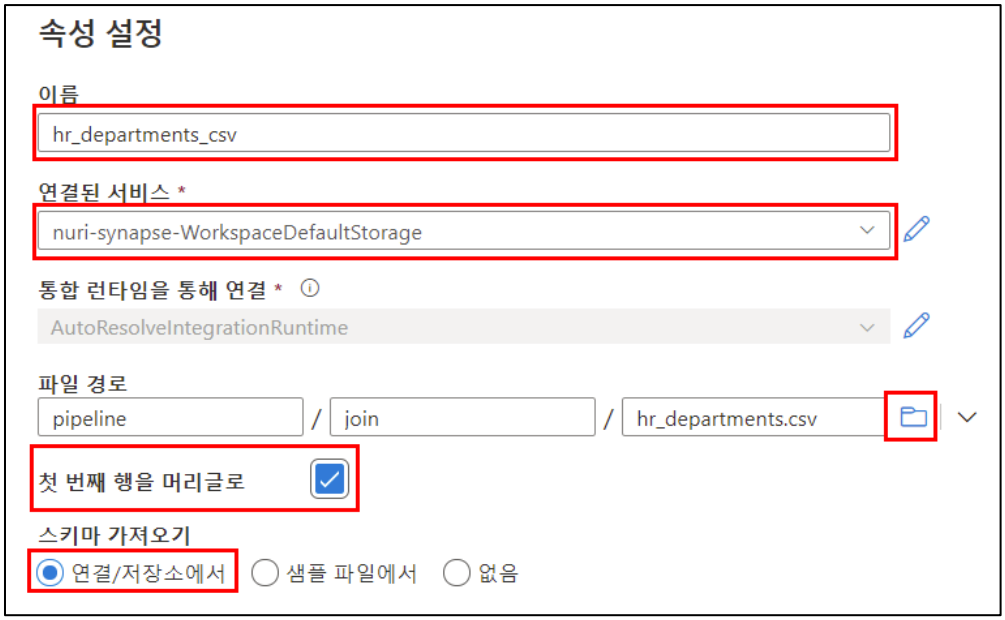
| Field | Value |
|---|---|
| 이름 | hr_employees_csv |
| 연결된 서비스 | <name>-synapse-WorkspaceDefaultStorage 선택 |
| 파일 경로 – 파일 시스템 | pipeline 선택 |
| 파일 경로 – 디렉터리 | join 선택 |
| 파일 경로 – 파일 | hr_employees.csv 선택 |
| 첫번째 행을 머리글로 | 체크박스 선택 |
| 스키마 가져오기 | 연결/저장소에서 선택 |



5. 다시 + 를 누르고 **통합 데이터 세트**를 생성합니다.

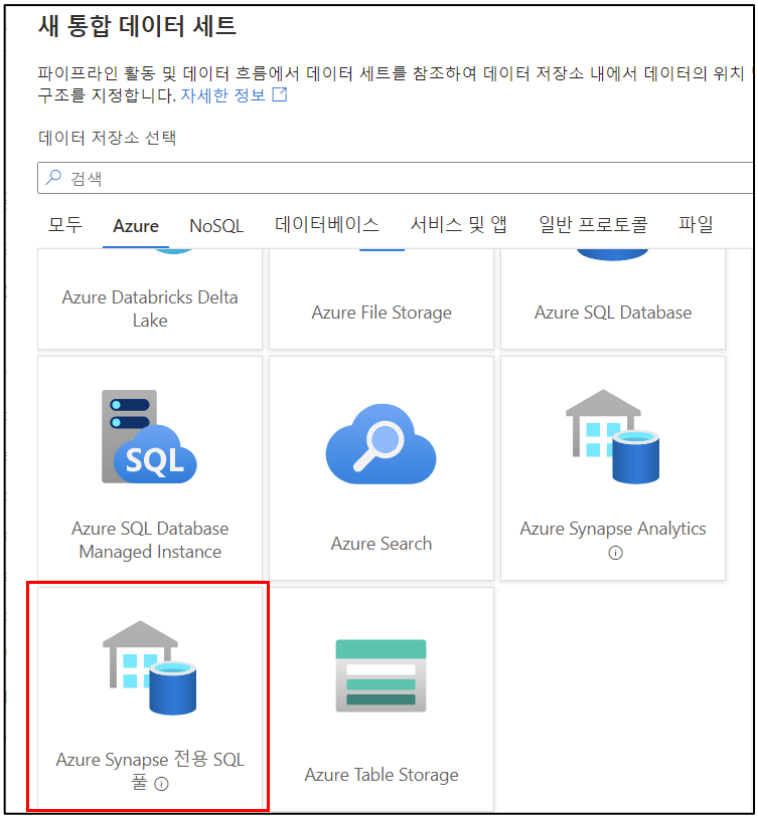
6. 바로 전 작업과 동일하게 **Azure Data Lake Storage Gen2**를 선택 후 **Delimited Text**를 선택합니다. 아래와 같이 **속성 설정**을 입력합니다.

| Field | Value |
|---|---|
| 이름 | hr_departments_csv |
| 연결된 서비스 | <name>-synapse-WorkspaceDefaultStorage 선택 |
| 파일 경로 – 파일 시스템 | pipeline 선택 |
| 파일 경로 – 디렉터리 | join 선택 |
| 파일 경로 – 파일 | hr_departments.csv 선택 |
| 첫번째 행을 머리글로 | 체크박스 선택 |
| 스키마 가져오기 | 연결/저장소에서 선택 |




7. 다시 + 를 누르고 **통합 데이터 세트**를 생성합니다.

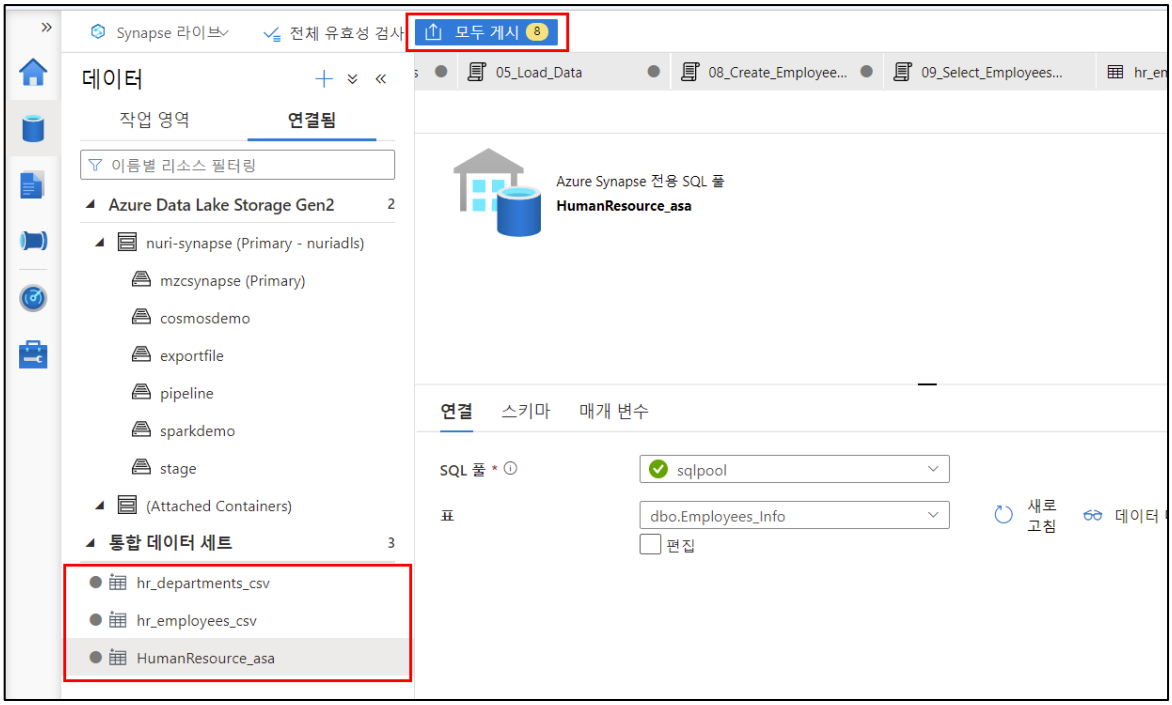
8. 이번엔 **Azure Synapse 전용 SQL 풀**을 선택합니다.



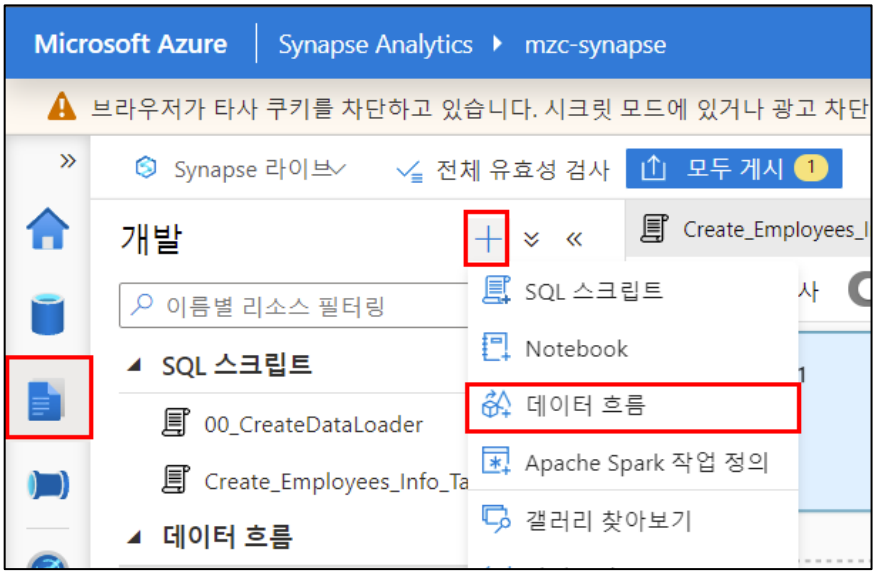

9. 아래와 같이 **속성 설정**을 입력합니다.

| **Field** | **Value** |
|---|---|
| 이름 | HumanResource_asa |
| SQL 풀 | sqlpool 선택 |
| 테이블 이름 | dbo.Employees_Info 선택 |
| 스키마 가져오기 | 연결/저장소에서 선택 |

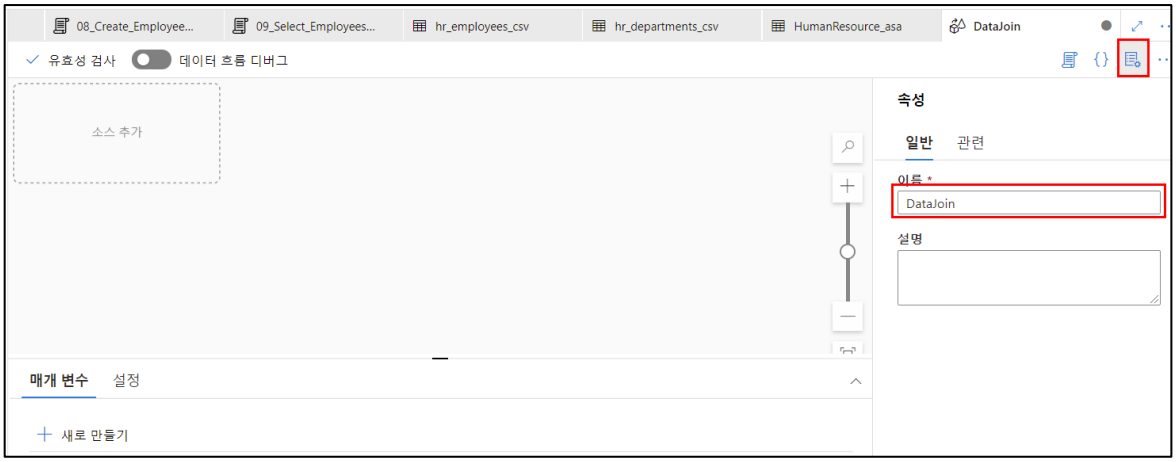
10. **통합 데이터 세트가 3개** 생성됨을 확인하고 **모두 게시**를 선택하여 저장합니다.




11. **개발 탭**으로 돌아와서 +를 누르고 **데이터 흐름**을 생성합니다.




12. 속성 값 이름에 **DataJoin**을 입력합니다.



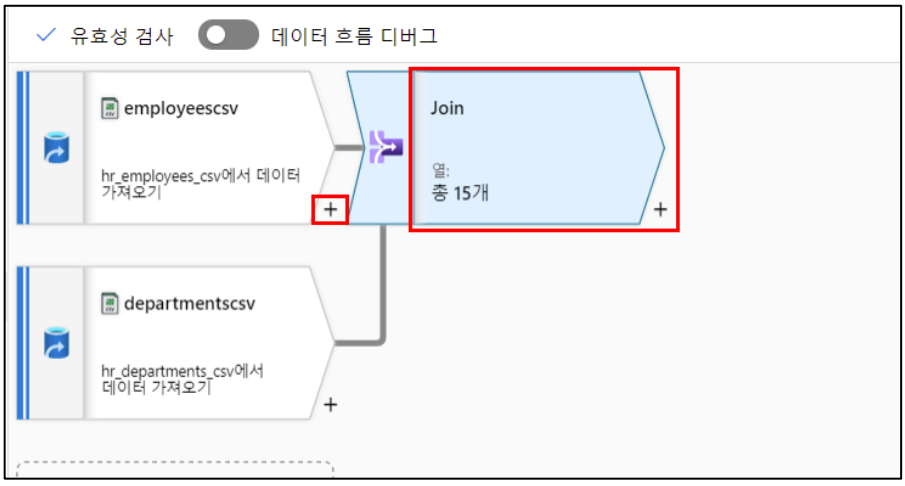

13. **소스 추가**를 선택하여 아래와 같이 **원본 설정**을 입력합니다.

| Field | Value |
|---|---|
| 출력 스트림 이름 | employeescsv |
| 원본 유형 | 통합 데이터 세트 |
| 데이터 세트 | hr_employees_csv 선택 |

14. 이전 작업과 동일하게 소스를 하나 더 추가하여 **원본 설정**을 아래와 같이 입력합니다.

| Field | Value |
|---|---|
| 출력 스트림 이름 | departmentscsv |
| 원본 유형 | 통합 데이터 세트 |
| 데이터 세트 | hr_departments_csv 선택 |

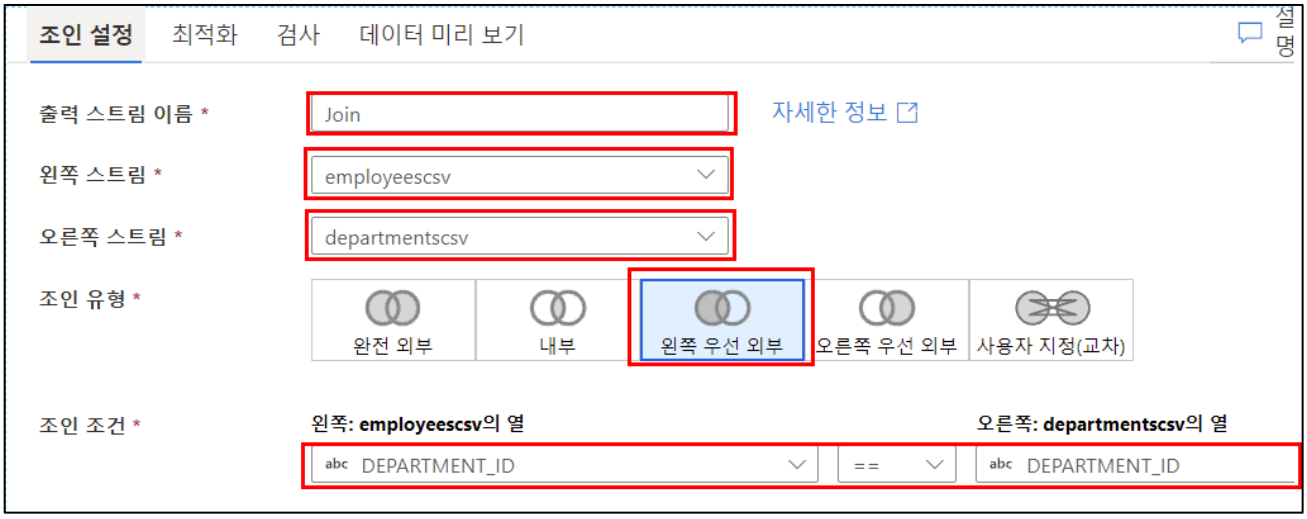

15. 아래 +를 클릭하여 **Join**을 선택합니다.



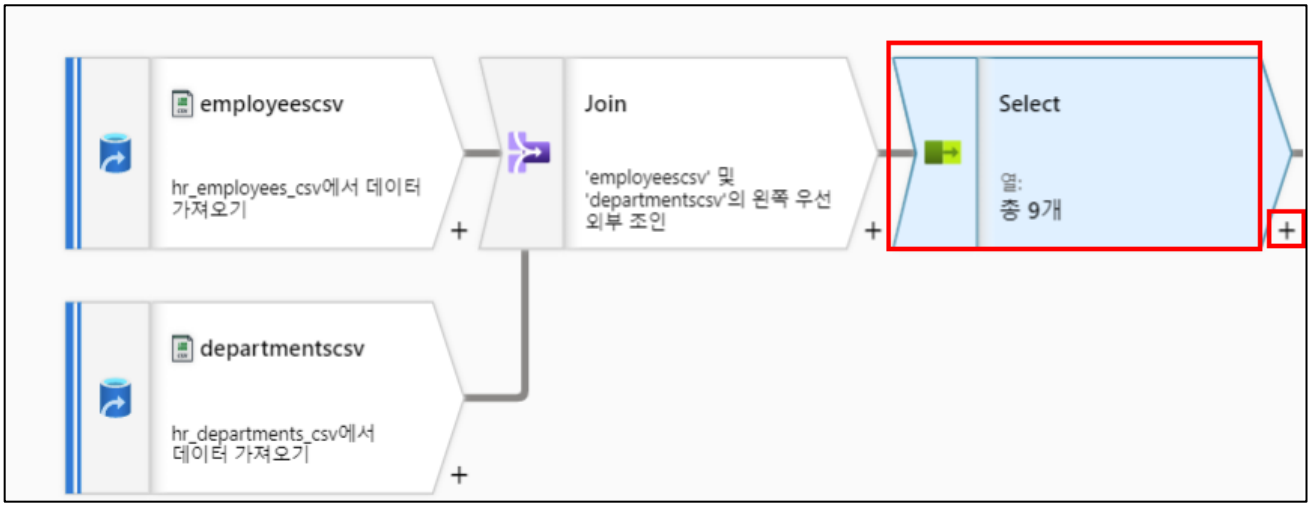
16. **출력 스트림 이름**을 **Join**으로 바꿔주고 **왼쪽 스트림은 employeescsv, 오른쪽 스트림은 departmentscsv**를 선택합니다.

17. employeescsv 파일이 기준이 되어 데이터를 조인하기 때문에 조인 유형은 **왼쪽 우선 외부**를 선택합니다.

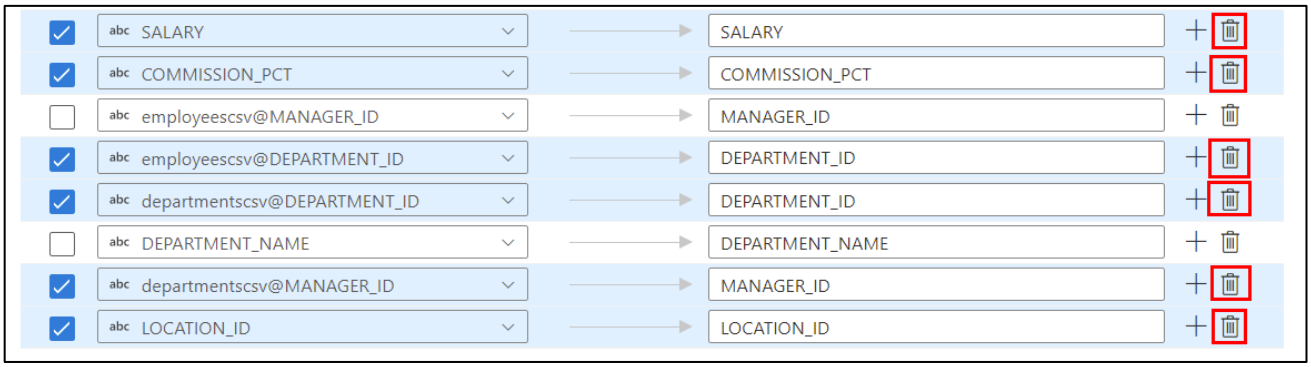
18. 조인 조건으로는 **DEPARTMENT_ID**를 선택합니다.



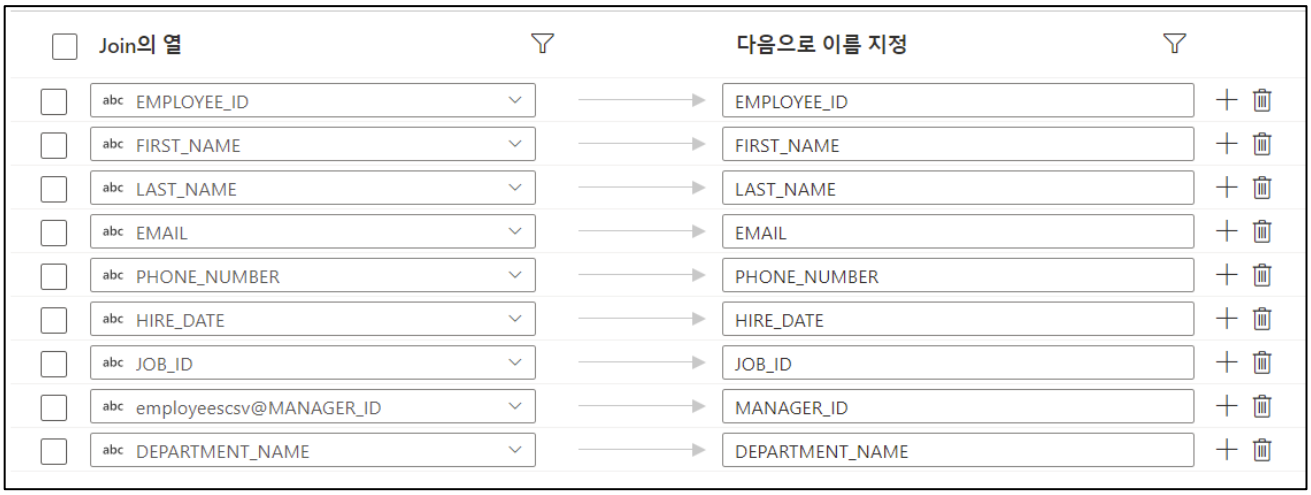

19. 다시 아래 +를 누른 후, **Select**를 선택합니다.



20. 출력 스트림 이름을 **Select**으로 바꿔주고 설정 선택에서 **필요 없는 열을 삭제**합니다.

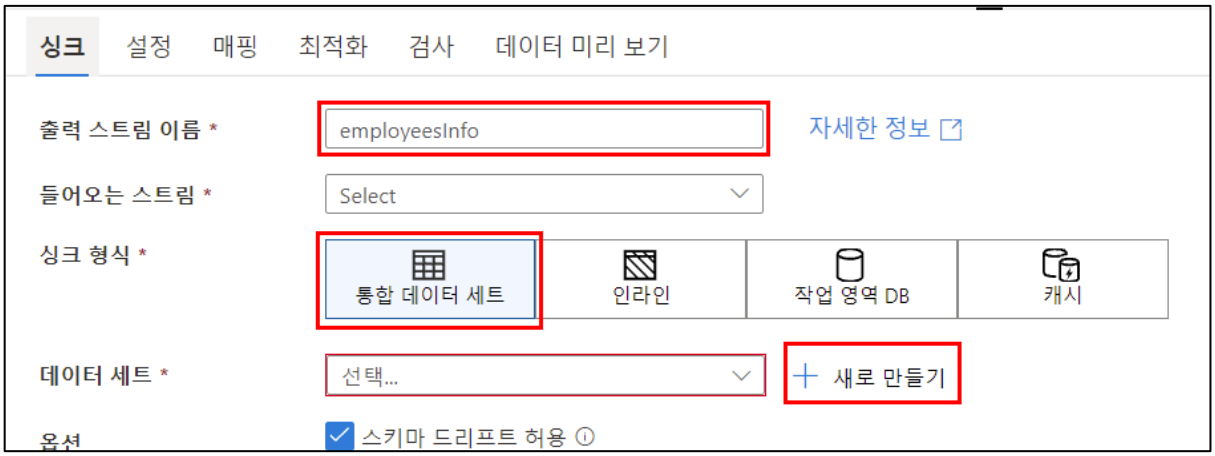



21. 삭제하고 난 뒤에 아래와 같이 보여야 합니다.

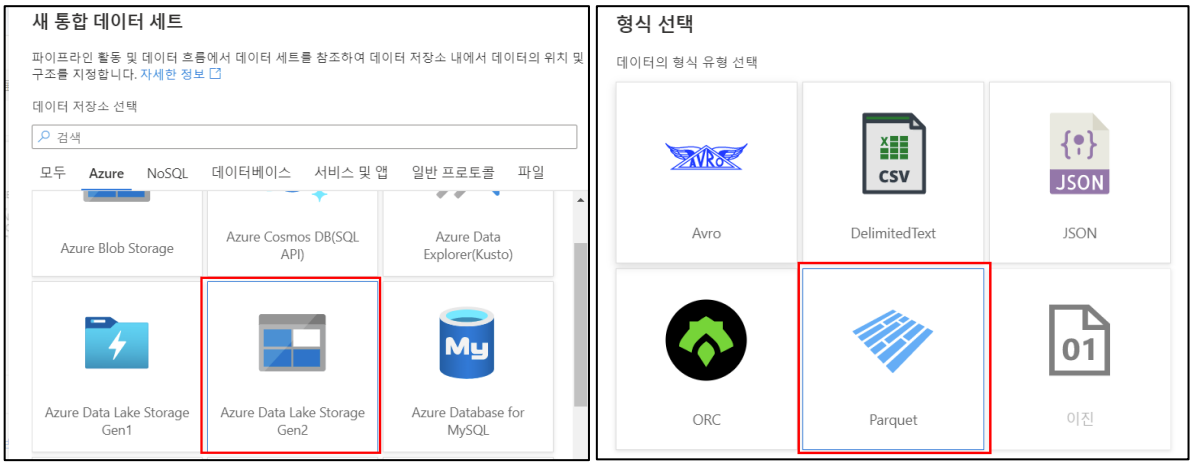



22. 마지막으로 +를 클릭하고 **Sink**를 선택합니다.

23. 출력 스트림 이름에 **employeesInfo**를 입력하고 싱크 형식은 **통합 데이터 세트**를 선택합니다. **데이터 세트**를 선택할 때 **+새로 만들기**를 눌러 새로 생성합니다.

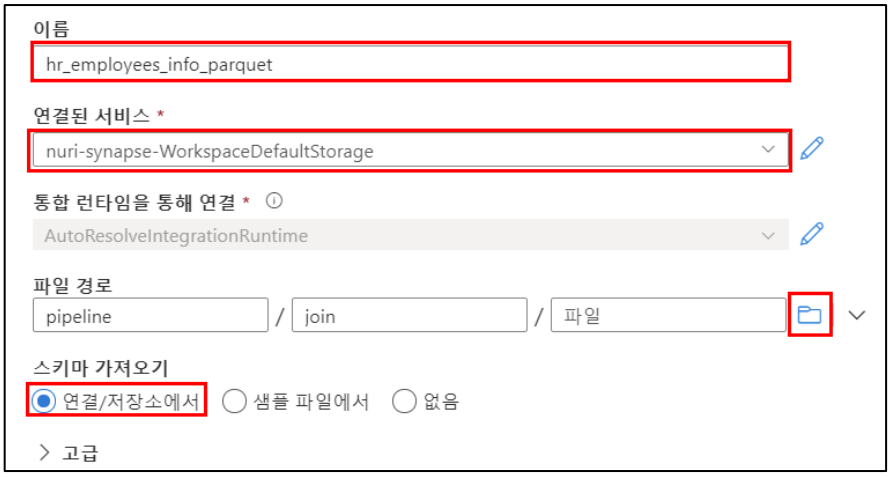



24. **Azure Data Lake Storage Gen2**를 선택하고 파일 형식은 **Parquet**을 선택합니다.

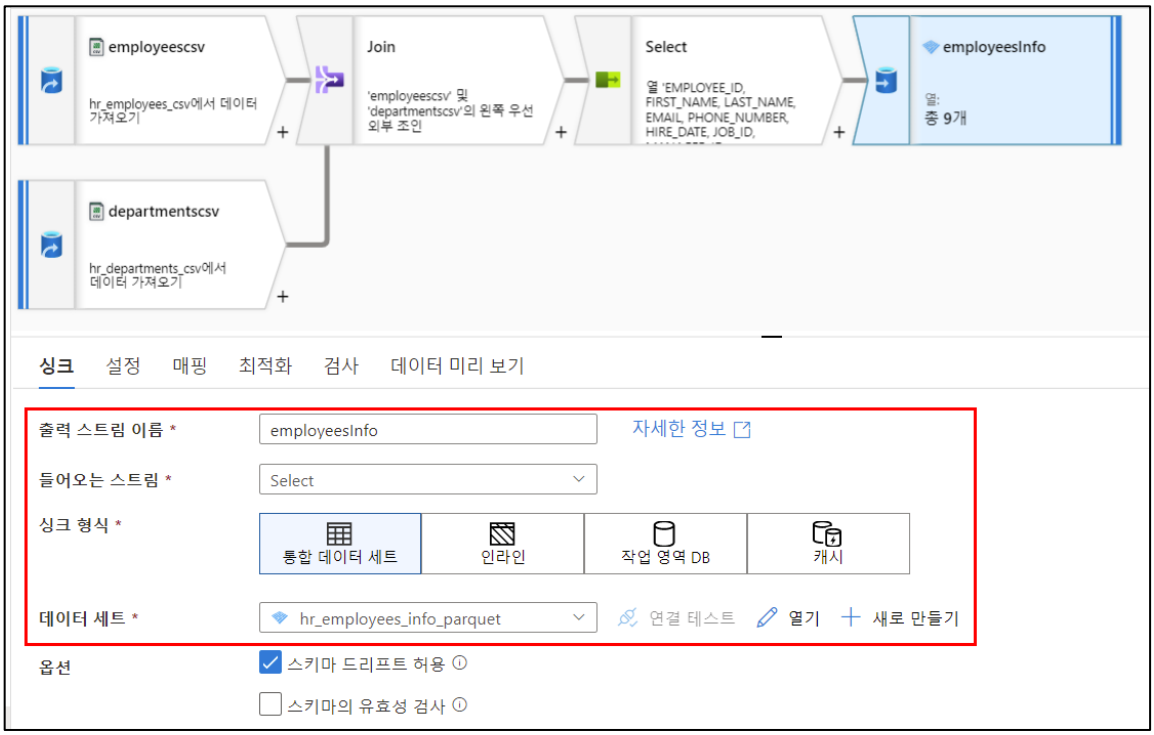



25. **속성 설정** 값은 아래와 같이 입력해줍니다.

| Field | Value |
|---|---|
| 이름 | hr_employees_info_parquet |
| 연결된 서비스 | <name>-synapse-WorkspaceDefaultStorage 선택 |
| 파일 경로 – 파일 시스템 | pipeline 선택 |
| 파일 경로 – 디렉터리 | join 선택 |
| 파일 경로 – 파일 | 공란 |
| 스키마 가져오기 | 연결/저장소에서 선택 |

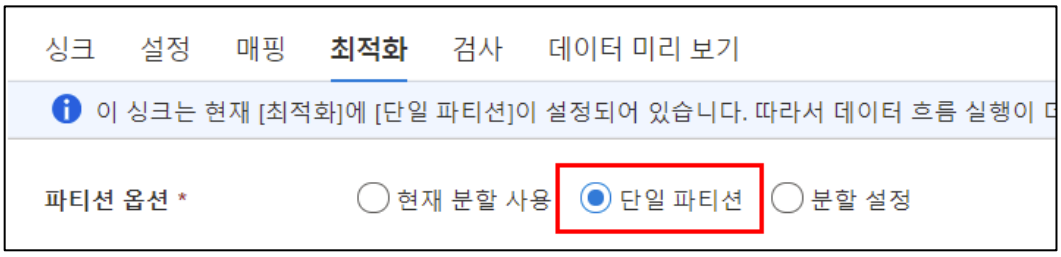



26. 데이터 세트 설정 후에 **sink 정보**는 아래와 같이 보여야 합니다.



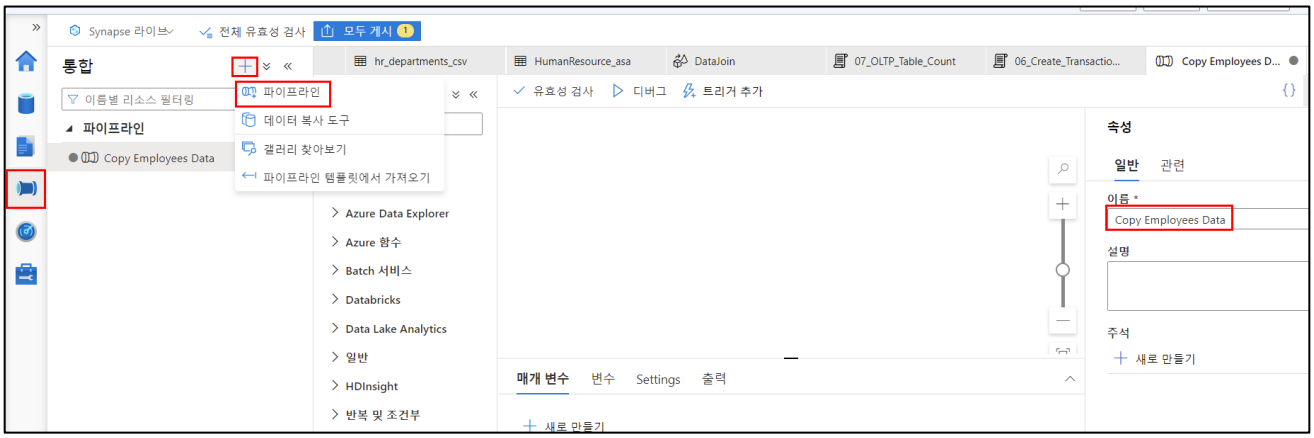

27. **Sink 설정 탭**으로 가서 파일 이름 옵션을 **단일 파일로 출력**을 선택합니다. 파일 명에는 **hremployeesInfo.parquet**를 입력합니다.


28. **Sink 최적화 탭**으로 가서 파티션 옵션을 **단일 파티션**으로 설정해야 합니다.

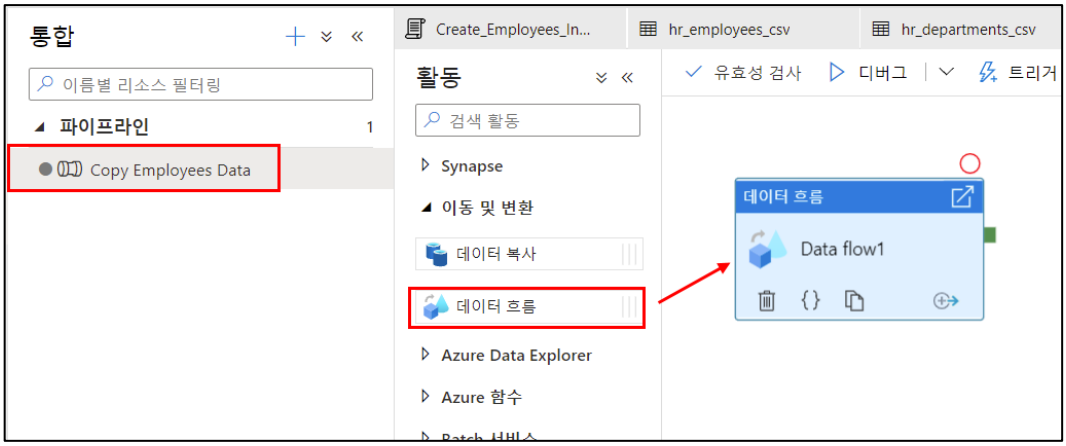



29. **모두 게시** 선택하여 저장합니다.

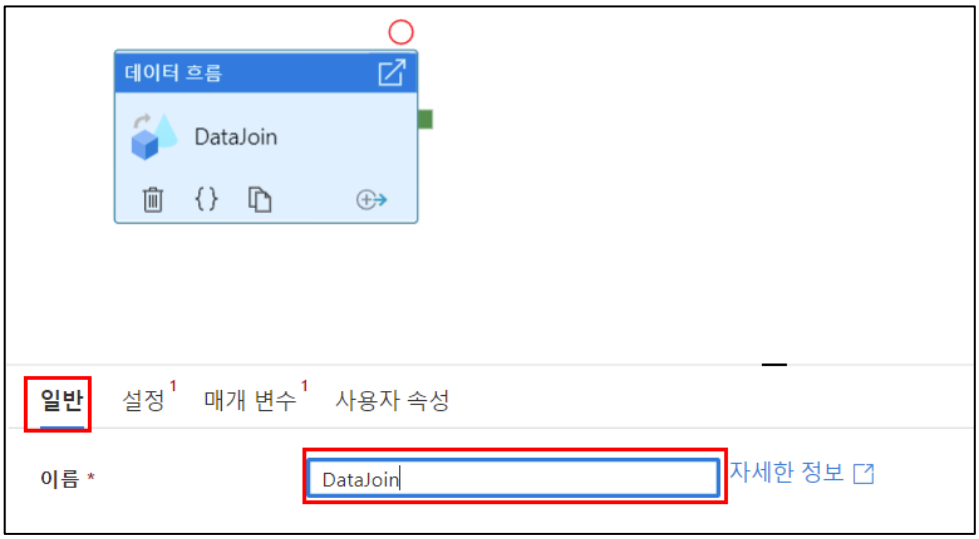
30. **통합** 탭에서 +를 클릭 후 **파이프라인**을 생성합니다. 이름을 **Copy Employees Data**로 입력합니다.



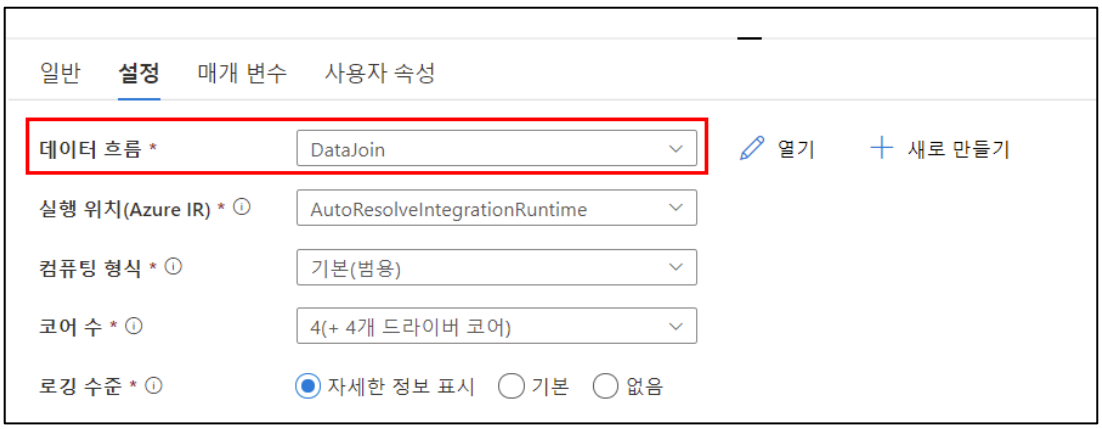
31. 이동 및 변환 메뉴에서 **데이터 흐름**을 Drag&Drop 합니다.



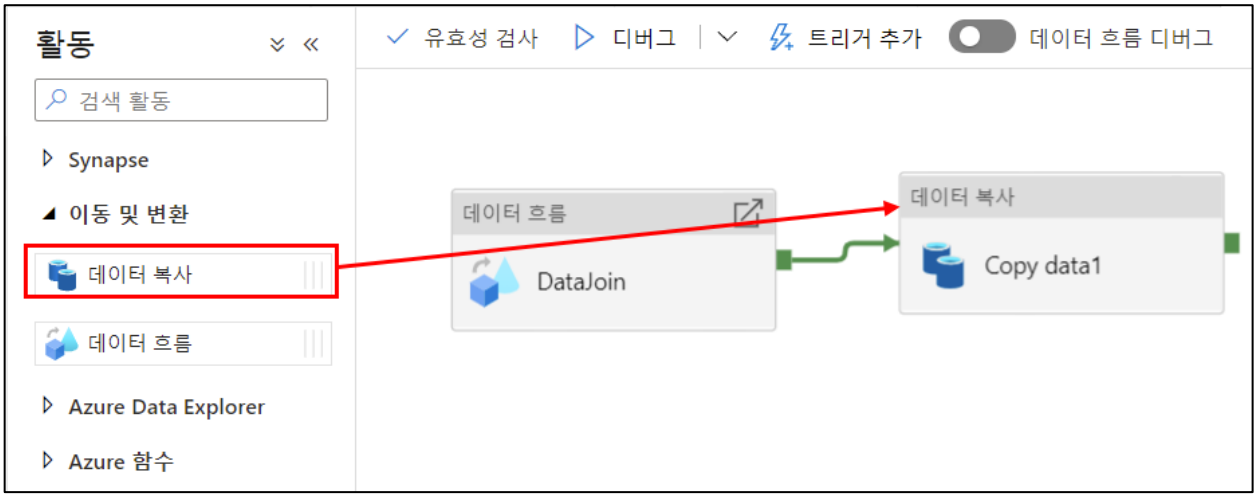

32. 데이터 흐름 일반 탭에서 이름을 **DataJoin**으로 수정합니다.



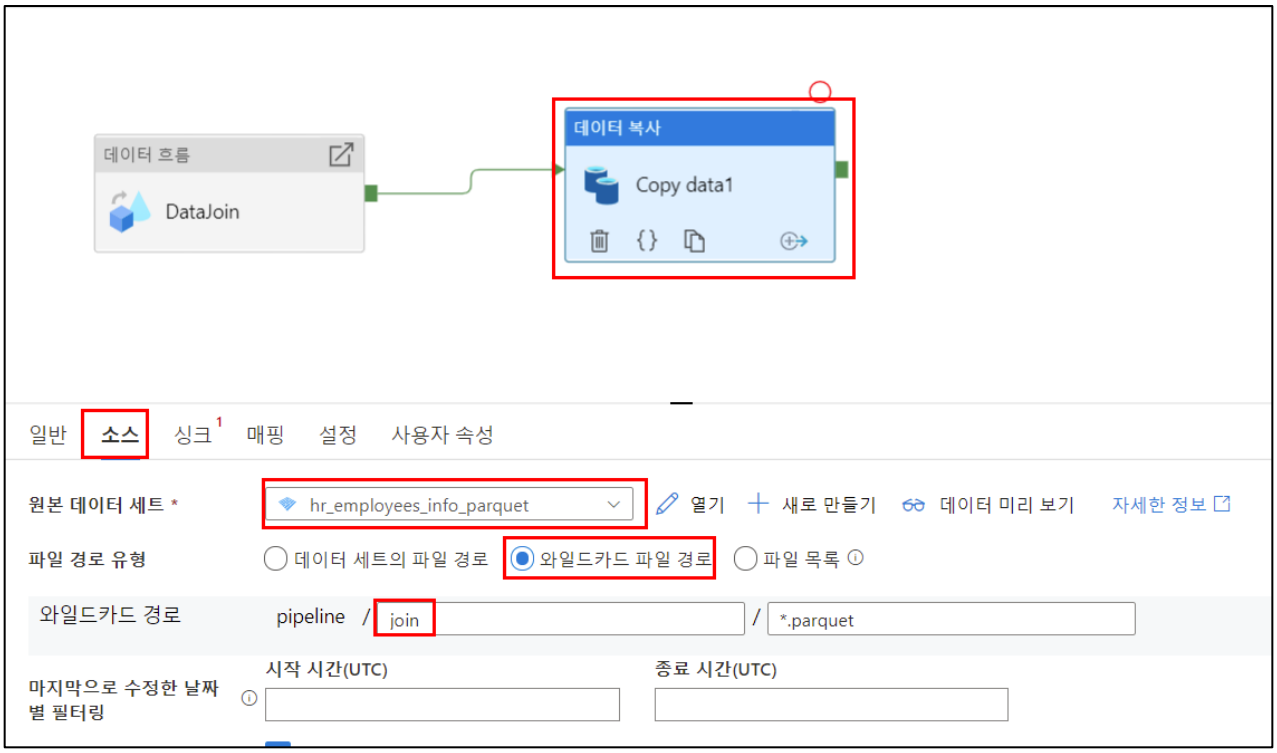

33. **설정 탭**에서 데이터 흐름에 **DataJoin**을 선택해 줍니다.



34. **데이터 복사**를 마찬가지로 Drag&Drop하여 아래와 같이 이어줍니다.



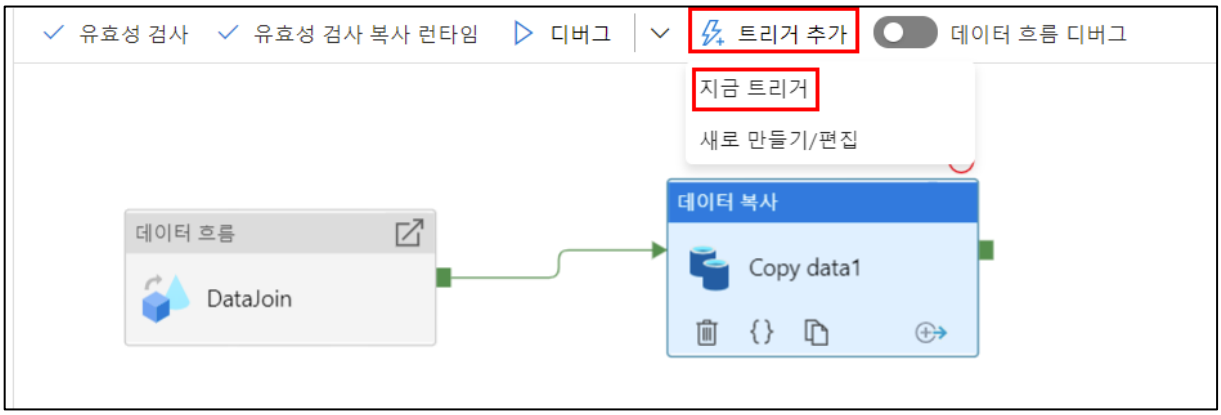

35. 데이터 복사를 선택하고 **소스 탭**으로 이동하여 **hr_employees_info_parquet**를 원본 데이터 세트로 선택합니다. 파일 경로 유형은 **와일드 카드 파일 경로**를 선택 후 **Join**를 입력합니다.



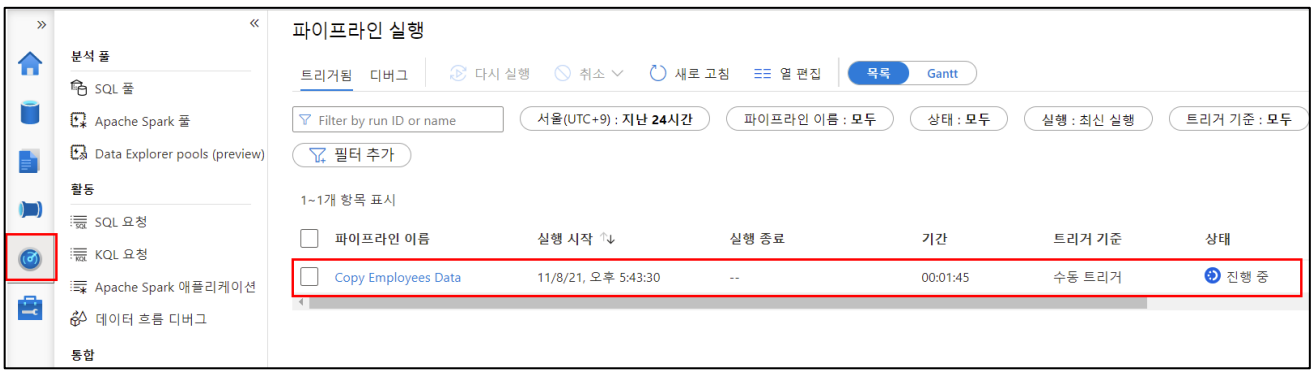
36. **싱크 탭**으로 이동해서 **HumanResource_asa**를 선택하고 **대량 삽입**을 선택합니다.


37. **모두 게시를** 선택하여 저장합니다.

38. 이제 Join된 데이터를 SQL 데이터베이스에 복사하기 위한 **트리거를 실행**합니다.




39. 파이프라인 진행 과정은 **모니터링 탭**에서 확인해볼 수 있습니다.



40. 작업이 완료되면 개발 탭에서 **09_Select_Employees_Info_Table** 스크립트를 실행하여 Join된 데이터가 SQL 데이터베이스에 복사됨을 확인할 수 있습니다.