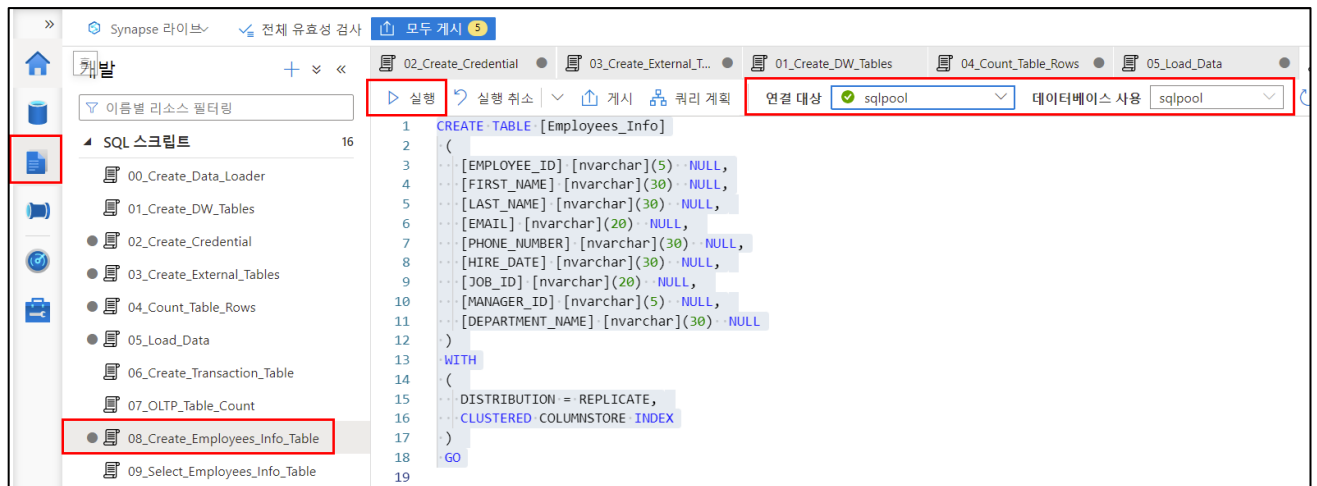


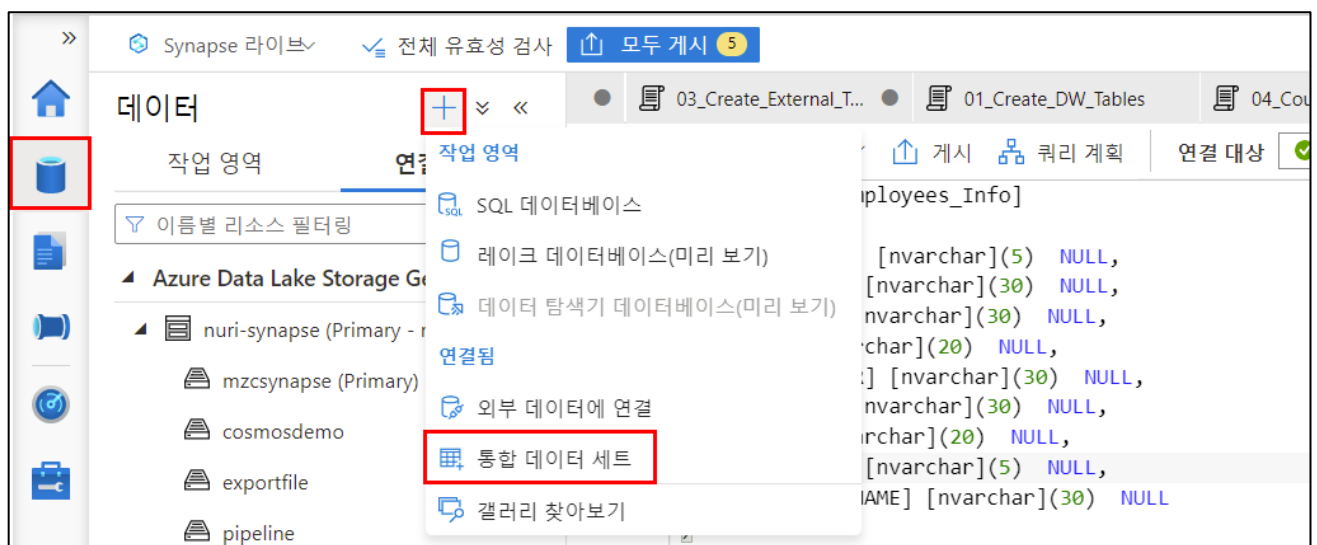
# Lab 4 – Source Files 를 통한 Data Join

**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 Gen2를 선택 후 Delimited Text를 선택합니다.

**새 통합 데이터 세트**

파이프라인 활동 및 데이터 흐름에서 데이터 세트를 참조하여 데이터 저장소 내에서 데이터의 위치 구조를 지정합니다. [자세한 정보](#)

데이터 저장소 선택

검색

모두 Azure NoSQL 데이터베이스 서비스 및 앱 일반 프로토콜 파일

Azure Blob Storage

Azure Cosmos DB(MongoDB API)

Azure Cosmos DB(SQL API)

Azure Data Explorer(Kusto)

Azure Data Lake Storage Gen1

**Azure Data Lake Storage Gen2**

**형식 선택**

데이터의 형식 유형 선택

Avro

**DelimitedText**

Excel

JSON

ORC

Parquet

### 4. 아래와 같이 속성 설정을 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

**속성 설정**

이름

hr\_employees\_csv

연결된 서비스 \*

nuri-synapse-WorkspaceDefaultStorage

통합 런타임을 통해 연결 \* ⓘ

AutoResolveIntegrationRuntime

파일 경로

pipeline / join / hr\_employees.csv

첫 번째 행을 머리글로 ☒

스키마 가져오기

☒ 연결/저장소에서 ☐ 샘플 파일에서 ☐ 없음

- 다시 + 를 누르고 **통합 데이터 세트**를 생성합니다.
- 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

### 속성 설정

이름

연결된 서비스 \*

통합 런타임을 통해 연결 \* ⓘ

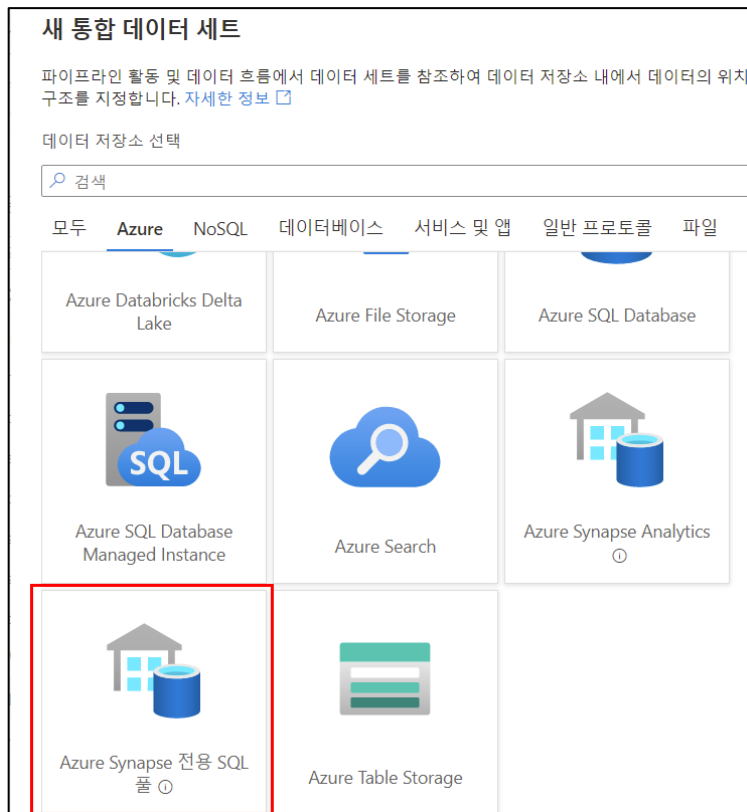
파일 경로  
 /  /

첫 번째 행을 머리글로 ☒

스키마 가져오기  
☒ 연결/저장소에서   
☐ 샘플 파일에서   
☐ 없음

- 다시 + 를 누르고 **통합 데이터 세트**를 생성합니다.

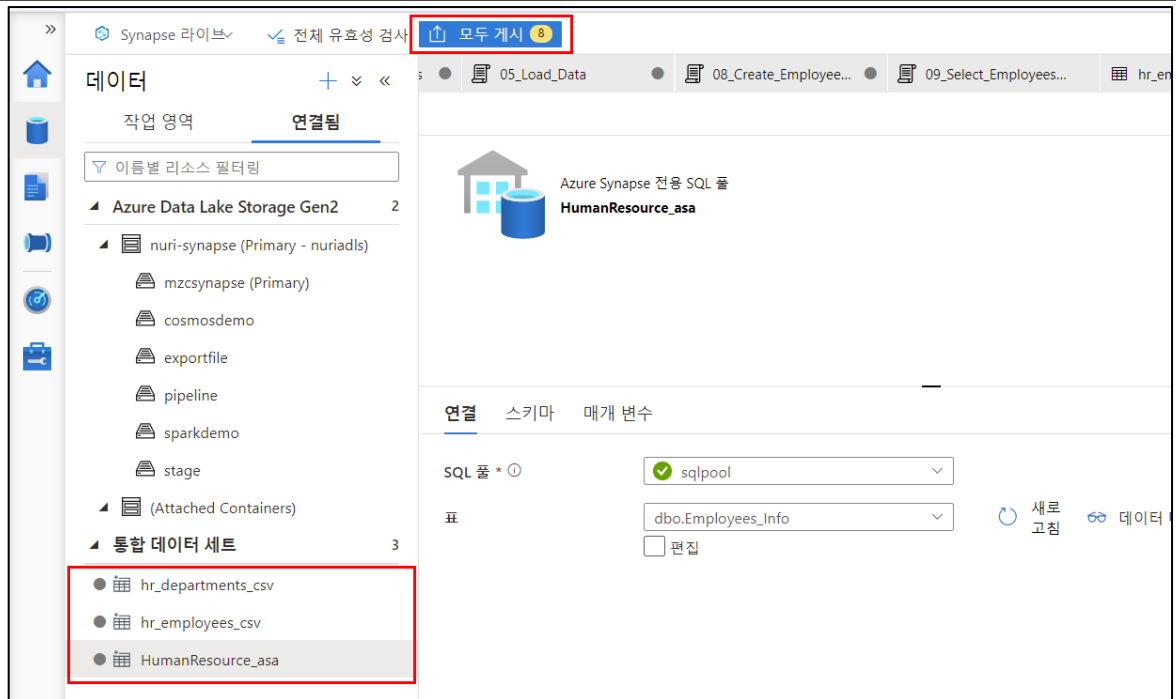
8. 이번엔 **Azure Synapse 전용 SQL 풀**을 선택합니다.



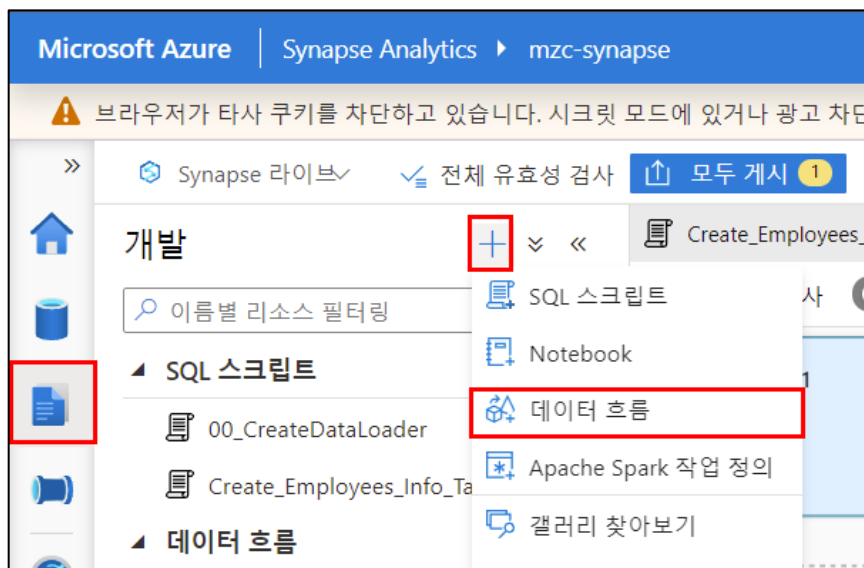
9. 아래와 같이 **속성 설정**을 입력합니다.

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

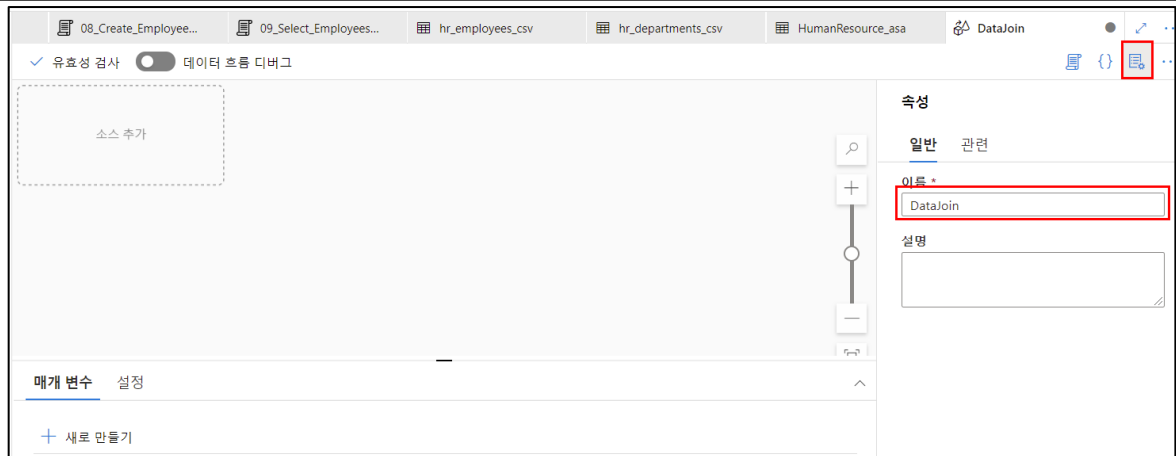
10. **통합 데이터 세트가 3개** 생성됨을 확인하고 **모두 게시**를 선택하여 저장합니다.



11. 개발 탭으로 돌아와서 +를 누르고 데이터 흐름을 생성합니다.



12. 속성 값 이름에 **DataJoin**을 입력합니다.



13. 소스 추가를 선택하여 아래와 같이 원본 설정을 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

✓ 유효성 검사    ☐ 데이터 흐름 디버그

 employeeescv  
hr\_employees\_csv에서 데이터 가져오기

+

 departmentscsv  
열: 총 4개

+

---

원본 설정    원본 옵션    프로젝트    최적화    검사    데이터 미리 보기

출력 스트림 이름 \*

departmentscsv

자세한 정보 [□](#)

원본 유형 \*

통합 데이터 세트

인라인

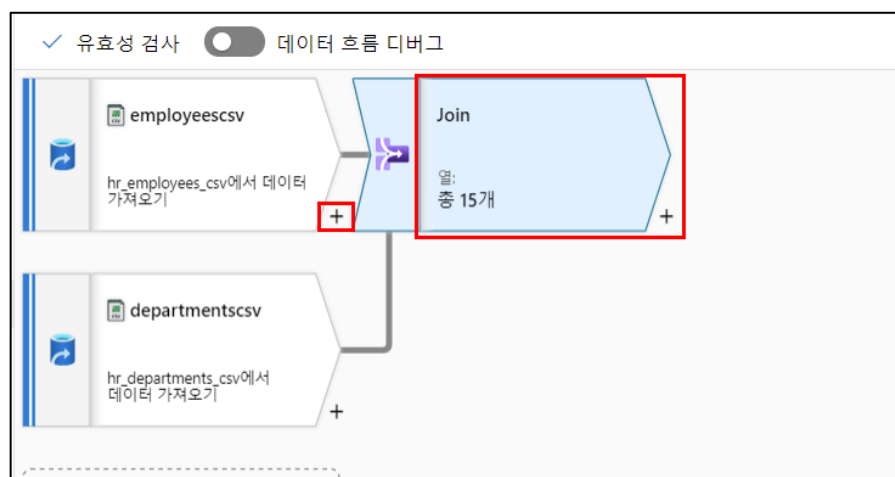
작업 영역 DB

데이터 세트 \*

hr\_departments\_csv

[연결 테스트](#)    [열기](#)    [+ 새로 만들기](#)

15. 아래 +를 클릭하여 **Join**을 선택합니다.



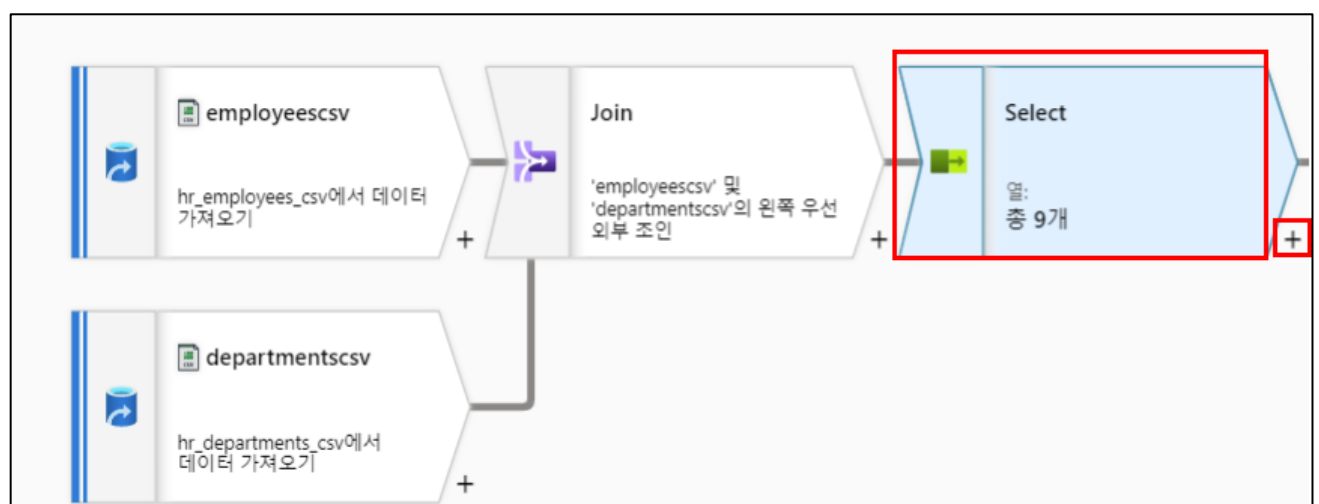
16. 출력 스트림 이름을 **Join**으로 바꿔주고 왼쪽 스트림은 **employeescsv**, 오른쪽 스트림은 **departmentscsv**를 선택합니다.

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 삭제합니다.

<input checked="" type="checkbox"/>	abc SALARY	→	SALARY	+	
<input checked="" type="checkbox"/>	abc COMMISSION_PCT	→	COMMISSION_PCT	+	
<input type="checkbox"/>	abc employeeescsv@MANAGER_ID	→	MANAGER_ID	+	
<input checked="" type="checkbox"/>	abc employeeescsv@DEPARTMENT_ID	→	DEPARTMENT_ID	+	
<input checked="" type="checkbox"/>	abc departmentscsv@DEPARTMENT_ID	→	DEPARTMENT_ID	+	
<input type="checkbox"/>	abc DEPARTMENT_NAME	→	DEPARTMENT_NAME	+	
<input checked="" type="checkbox"/>	abc departmentscsv@MANAGER_ID	→	MANAGER_ID	+	
<input checked="" type="checkbox"/>	abc LOCATION_ID	→	LOCATION_ID	+	

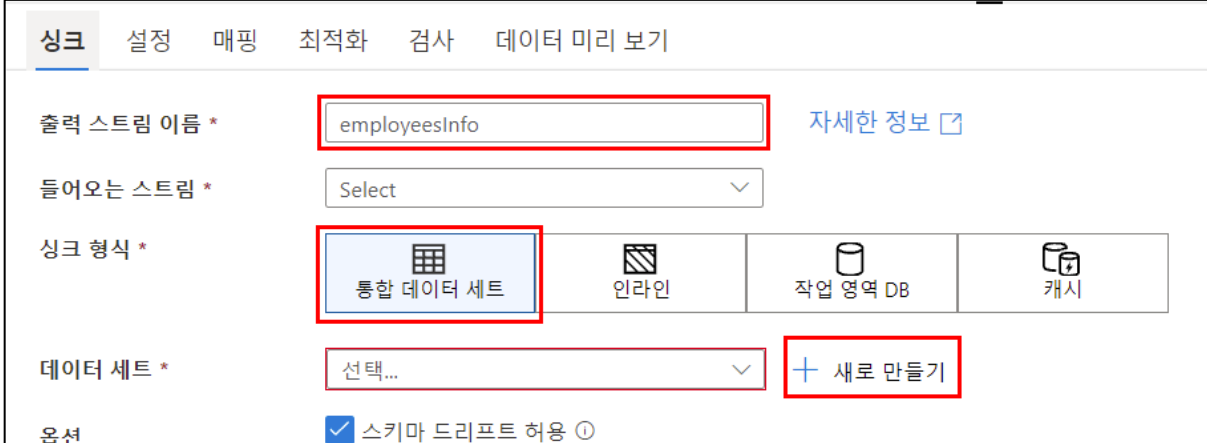
21. 삭제하고 난 뒤에 아래와 같이 보여야 합니다.

<input type="checkbox"/> Join의 열	↓	다음으로 이름 지정	↓
<input type="checkbox"/> abc EMPLOYEE_ID	→	EMPLOYEE_ID	+
<input type="checkbox"/> abc FIRST_NAME	→	FIRST_NAME	+
<input type="checkbox"/> abc LAST_NAME	→	LAST_NAME	+
<input type="checkbox"/> abc EMAIL	→	EMAIL	+
<input type="checkbox"/> abc PHONE_NUMBER	→	PHONE_NUMBER	+
<input type="checkbox"/> abc HIRE_DATE	→	HIRE_DATE	+
<input type="checkbox"/> abc JOB_ID	→	JOB_ID	+
<input type="checkbox"/> abc employeeescsv@MANAGER_ID	→	MANAGER_ID	+
<input type="checkbox"/> abc DEPARTMENT_NAME	→	DEPARTMENT_NAME	+

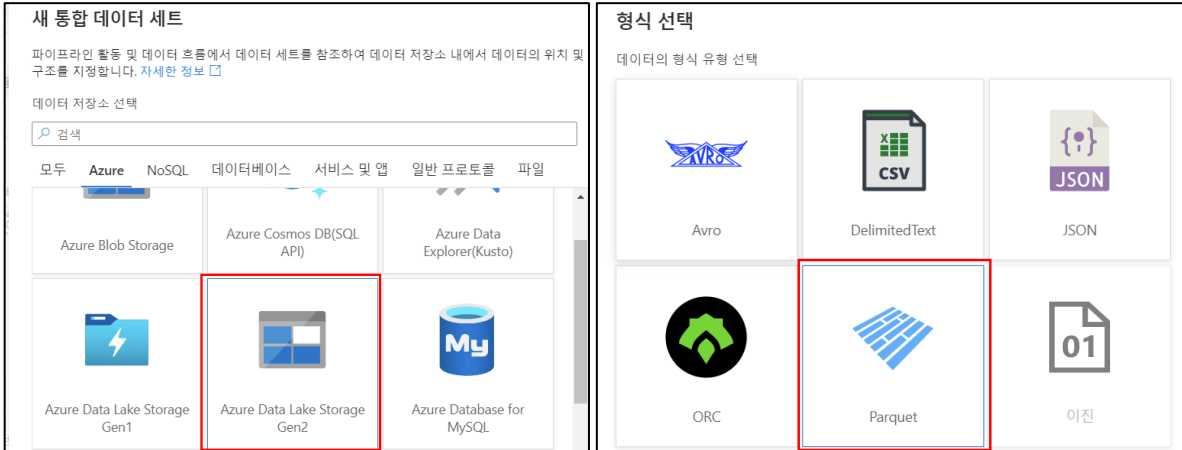
22. 마지막으로 +를 클릭하고 **Sink**를 선택합니다.



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

이름

hr\_employees\_info\_parquet

연결된 서비스 \*

nuri-synapse-WorkspaceDefaultStorage

통합 런타임을 통해 연결 \* ①

AutoResolveIntegrationRuntime

파일 경로

pipeline / join / 파일

스키마 가져오기

☒ 연결/저장소에서
 ☐ 샘플 파일에서
 ☐ 없음

[> 고급](#)

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

employeescsv

hr\_employees\_csv에서 데이터 가져오기

+

Join

'employeescsv' 및 'departmentscsv'의 왼쪽 우선 외부 조인

+

Select

컬 'EMPLOYEE\_ID', FIRST\_NAME, LAST\_NAME, EMAIL, PHONE\_NUMBER, HIRE\_DATE, JOB\_ID, ...

+

employeesInfo

컬: 9개

departmentscsv

hr\_departments\_csv에서 데이터 가져오기

+

싱크

설정

매핑

최적화

검사

데이터 미리 보기

출력 스트림 이름 \*

employeesInfo

자세한 정보

들어오는 스트림 \*

Select

싱크 형식 \*

통합 데이터 세트

인라인

작업 영역 DB

캐시

데이터 세트 \*

hr\_employees\_info\_parquet

연결 테스트

열기

새로 만들기

옵션

☒ 스키마 드리프트 허용 ①
   
☐ 스키마의 유효성 검사 ①

27. Sink 설정 탭으로 가서 파일 이름 옵션을 단일 파일로 출력을 선택합니다. 파일 명에는 hremployeesInfo.parquet를 입력합니다.

싱크 **설정** 매핑 최적화 검사 데이터 미리 보기

**i** 이 싱크는 현재 [최적화]에 [단일 파티션]이 설정되어 있습니다. 따라서 데이터 흐름

폴더 지우기 ☐

파일 이름 옵션 \* 

단일 파일로 출력

단일 파일로 출력 \* ① 

hremployeesInfo.parquet

Umask ①

소유자	<input type="checkbox"/> R	<input type="checkbox"/> W	<input type="checkbox"/> X
그룹	<input type="checkbox"/> R	<input checked="" type="checkbox"/> W	<input type="checkbox"/> X
기타	<input type="checkbox"/> R	<input checked="" type="checkbox"/> W	<input type="checkbox"/> X

28. Sink 최적화 탭으로 가서 파티션 옵션을 단일 파티션으로 설정해야 합니다.

싱크 설정 매핑 **최적화** 검사 데이터 미리 보기

**i** 이 싱크는 현재 [최적화]에 [단일 파티션]이 설정되어 있습니다. 따라서 데이터 흐름 실행이

파티션 옵션 \* ☐ 현재 분할 사용 ☒ 단일 파티션 ☐ 분할 설정

29. 모두 게시 선택하여 저장합니다.

30. 통합 탭에서 +를 클릭 후 파이프라인을 생성합니다. 이름을 Copy Employees Data로 입력합니다.

Synapse 라이브러리 전체 유효성 검사 모두 게시

통합

이름별 리소스 필터링

파이프라인

Copy Employees Data

파이프라인

데이터 복사 도구

쿼리 찾아보기

파이프라인 템플릿에서 가져오기

Azure Data Explorer

Azure 함수

Batch 서비스

Databricks

Data Lake Analytics

일반

HDInsight

반복 및 조건부

유효성 검사 디버그 트리거 추가

속성

일반 관련

이름 \* Copy Employees Data

설명

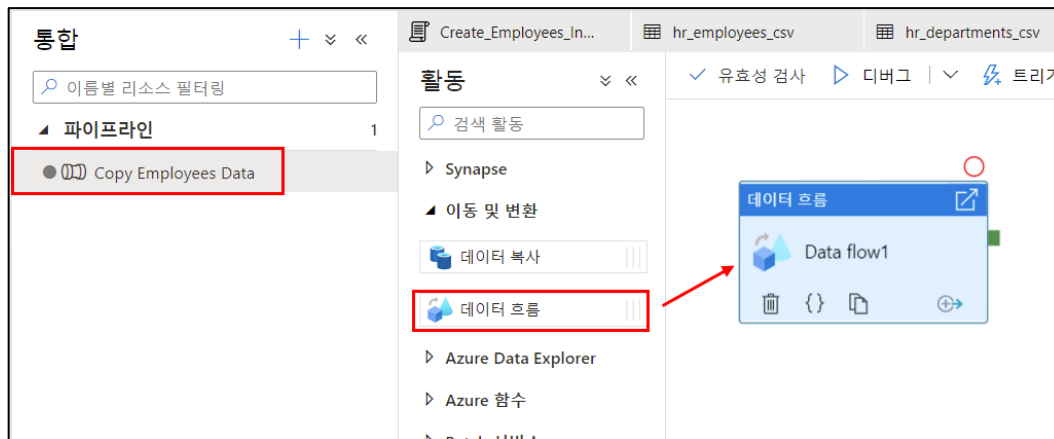
주석

+ 새로 만들기

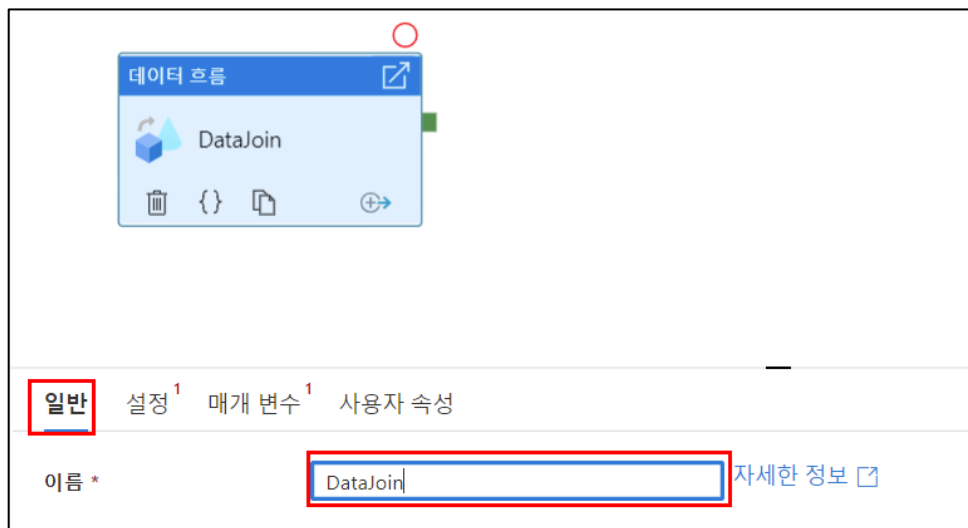
매개 변수 변수 Settings 출력

+ 새로 만들기

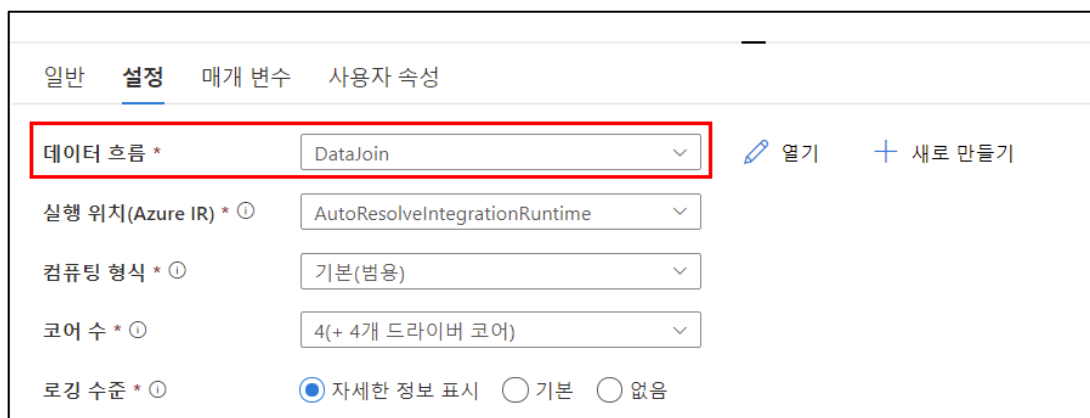
31. 이동 및 변환 메뉴에서 **데이터 흐름**을 Drag&Drop 합니다.



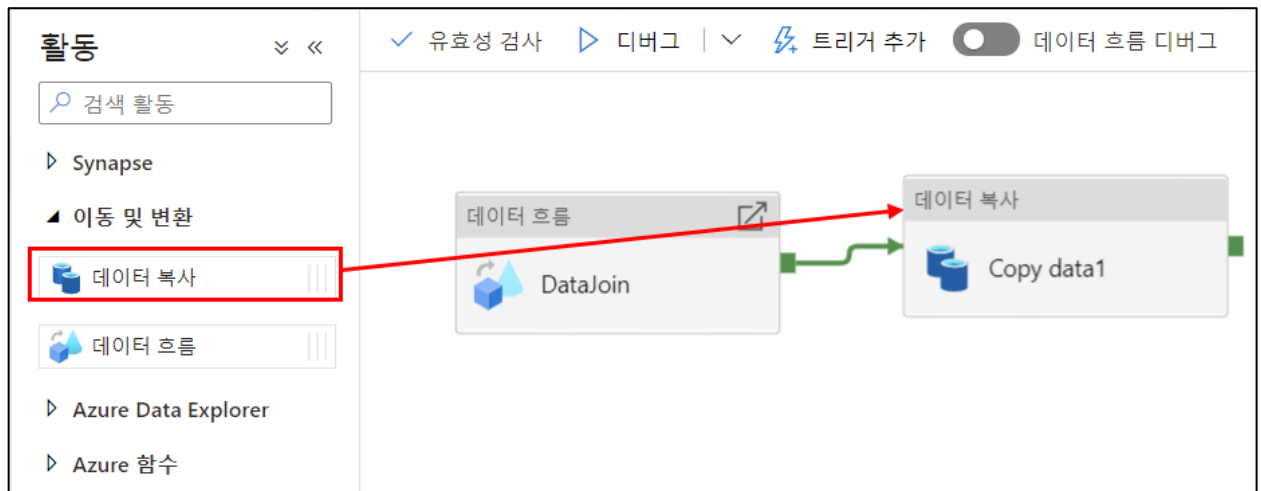
32. 데이터 흐름 일반 탭에서 이름을 **DataJoin**으로 수정합니다.



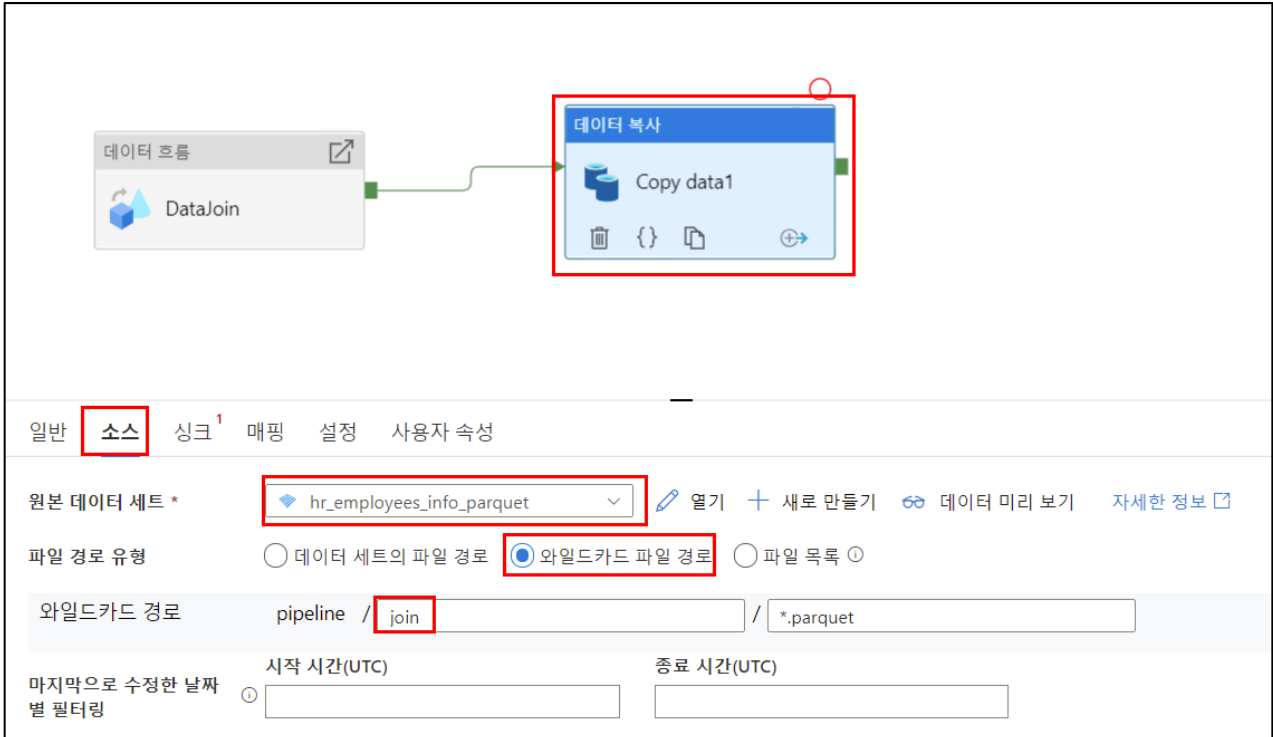
33. **설정** 탭에서 데이터 흐름에 **DataJoin**을 선택해 줍니다.



34. 데이터 복사를 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.



일반 **소스** 싱크<sup>1</sup> 매핑 설정 사용자 속성

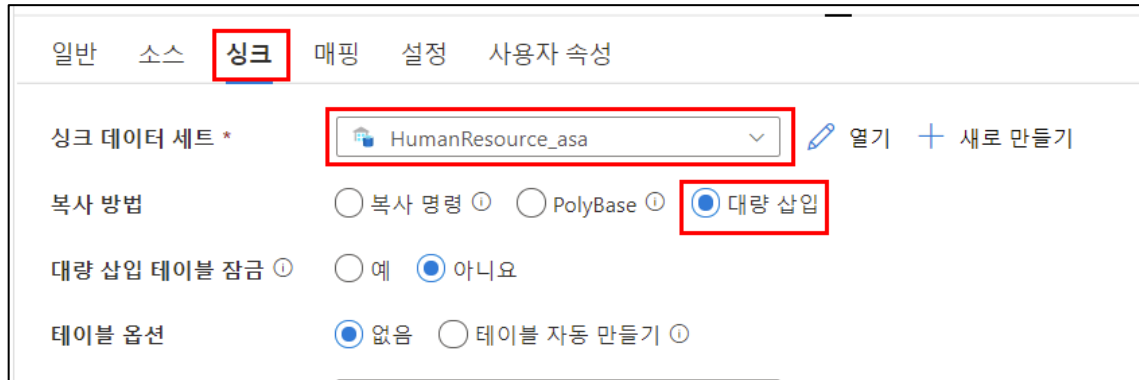
원본 데이터 세트 \* hr\_employees\_info\_parquet 열기 새로 만들기 데이터 미리 보기 자세한 정보

파일 경로 유형 ☐ 데이터 세트의 파일 경로 ☒ 와일드카드 파일 경로 ☐ 파일 목록 ⓘ

와일드카드 경로 pipeline / join / \*.parquet

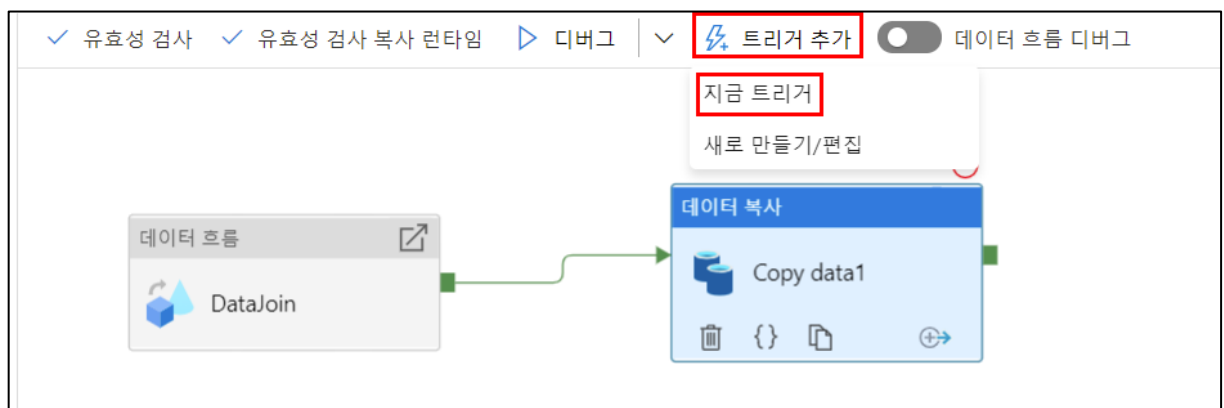
마지막으로 수정한 날짜 별 필터링 시작 시간(UTC) 종료 시간(UTC)

36. 싱크 탭으로 이동해서 **HumanResource\_asa**를 선택하고 **대량 삽입**을 선택합니다.

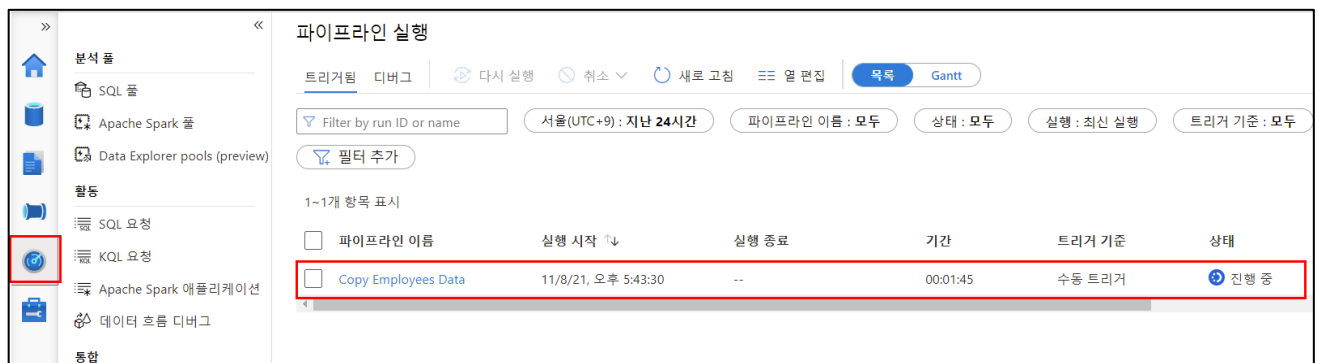


37. 모두 게시를 선택하여 저장합니다.

38. 이제 Join된 데이터를 SQL 데이터베이스에 복사하기 위한 **트리거**를 실행합니다.

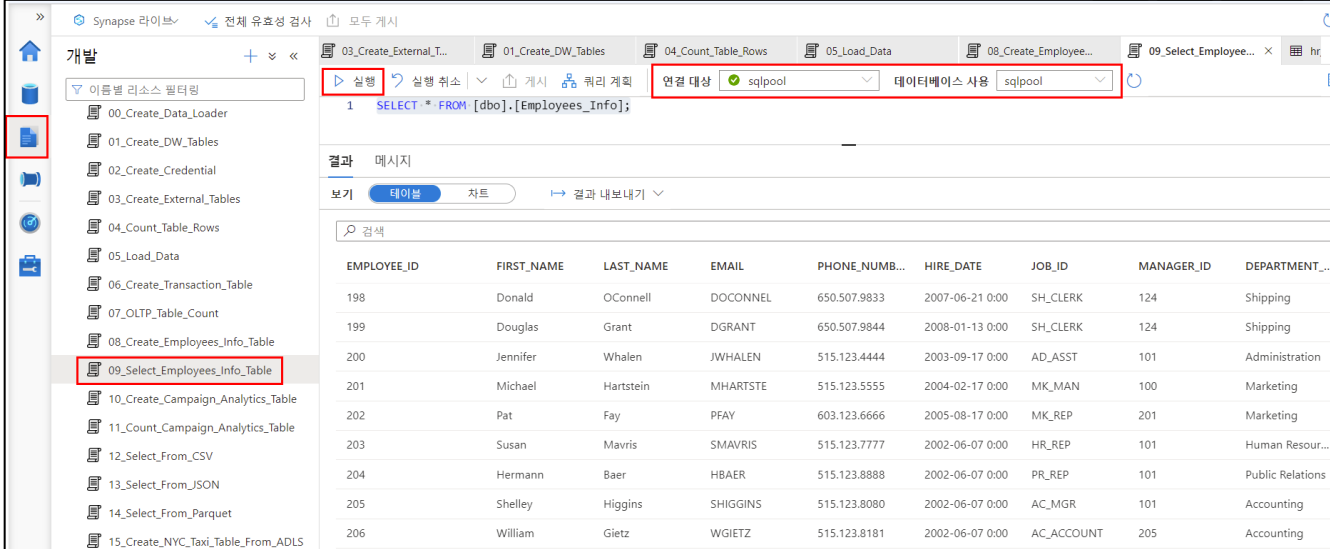


39. 파이프라인 진행 과정은 **모니터링** 탭에서 확인해볼 수 있습니다.



파이프라인 이름	실행 시작	실행 종료	기간	트리거 기준	상태
Copy Employees Data	11/8/21, 오후 5:43:30	--	00:01:45	수동 트리거	진행 중

40. 작업이 완료되면 개발 탭에서 **09\_Select\_Employees\_Info\_Table** 스크립트를 실행하여 Join된 데이터가 SQL 데이터베이스에 복사됨을 확인할 수 있습니다.



The screenshot shows the Synapse IDE interface. On the left, a list of scripts is displayed, with '09\_Select\_Employees\_Info\_Table' selected. The main editor shows the SQL query: `SELECT * FROM [dbo].[Employees_Info];`. The '연결 대상' (Connection Target) is set to 'sqlpool' and '데이터베이스 사용' (Database Use) is also set to 'sqlpool'. The '결과' (Results) tab is active, displaying a table of employee data.

EMPLOYEE_ID	FIRST_NAME	LAST_NAME	EMAIL	PHONE_NUMB...	HIRE_DATE	JOB_ID	MANAGER_ID	DEPARTMENT_...
198	Donald	OConnell	DOCONNEL	650.507.9833	2007-06-21 0:00	SH_CLERK	124	Shipping
199	Douglas	Grant	DGRANT	650.507.9844	2008-01-13 0:00	SH_CLERK	124	Shipping
200	Jennifer	Whalen	JWHALEN	515.123.4444	2003-09-17 0:00	AD_ASST	101	Administration
201	Michael	Hartstein	MHARTSTE	515.123.5555	2004-02-17 0:00	MK_MAN	100	Marketing
202	Pat	Fay	PFAY	603.123.6666	2005-08-17 0:00	MK_REP	201	Marketing
203	Susan	Mavris	SMAVRIS	515.123.7777	2002-06-07 0:00	HR_REP	101	Human Resour...
204	Hermann	Baer	HBAER	515.123.8888	2002-06-07 0:00	PR_REP	101	Public Relations
205	Shelley	Higgins	SHIGGINS	515.123.8080	2002-06-07 0:00	AC_MGR	101	Accounting
206	William	Gietz	WGIEZ	515.123.8181	2002-06-07 0:00	AC_ACCOUNT	205	Accounting