

6.6 탐색 파일럿 실행 4단계 - 주제영역 5

 스마트카 운전자 차량용품 구매이력 정보

과제 IV

다음 페이지 부터 내용들을 실습해서 주제영역 5

“운전자 차량용품 구매이력 정보” 마트를 생성 하세요.

주제 영역 5. 스마트카 운전자 차량용품 구매 이력 정보 - 워크플로 작성 _ 실습

이번 주제 영역 5의 워크플로는 2020년 03월 스마트카 운전자들이 구매한 “스마트카 차량용품 구매 이력”과 “스마트카 마스터 데이터”를 결합한 데이터셋을 만들고, 차량번호별 구매한 상품 리스트를 로컬 파일시스템에 생성하는 것이다. 스마트카 차량용품 구매 이력 데이터는 SmartCar_Item_BuyList 테이블로 구성돼 있고, 스마트카 마스터 데이터는 앞서 여러 번 사용했던 SmartCar_Master_Over18 테이블을 이용할 것이다. 워크플로의 하이브 작업에 사용되는 하이브 QL은 C://예제소스/bigdata2nd-master/CH06/HiveQL/에서 제공되므로 필요 시 해당 파일을 열어 복사 및 붙여넣기 방식으로 이용한다.

01. 휴의 좌측 드롭박스 메뉴에서 [문서]를 선택해 [내 문서]에 생성해 놓은 주제 영역 5의 작업 디렉터리로 이동한다.
 - 휴 내 문서: /workflow/hive_script/subject5
02. 주제 영역 5에서는 하이브 스크립트 파일을 두 개 작성한다. 먼저 내 문서의 /workflow/hive_script/subject5 위치로 이동해서 [새 문서] → [Hive 쿼리]를 선택한다.

03. 하이브 에디터가 나타나면 구매 이력을 관리하기 위한 하이브 테이블 생성 스크립트를 그림 6.113처럼 작성하고 저장한다. 파일 이름은 create_table_smartcar_item_buylist_info.hql로 입력한다.

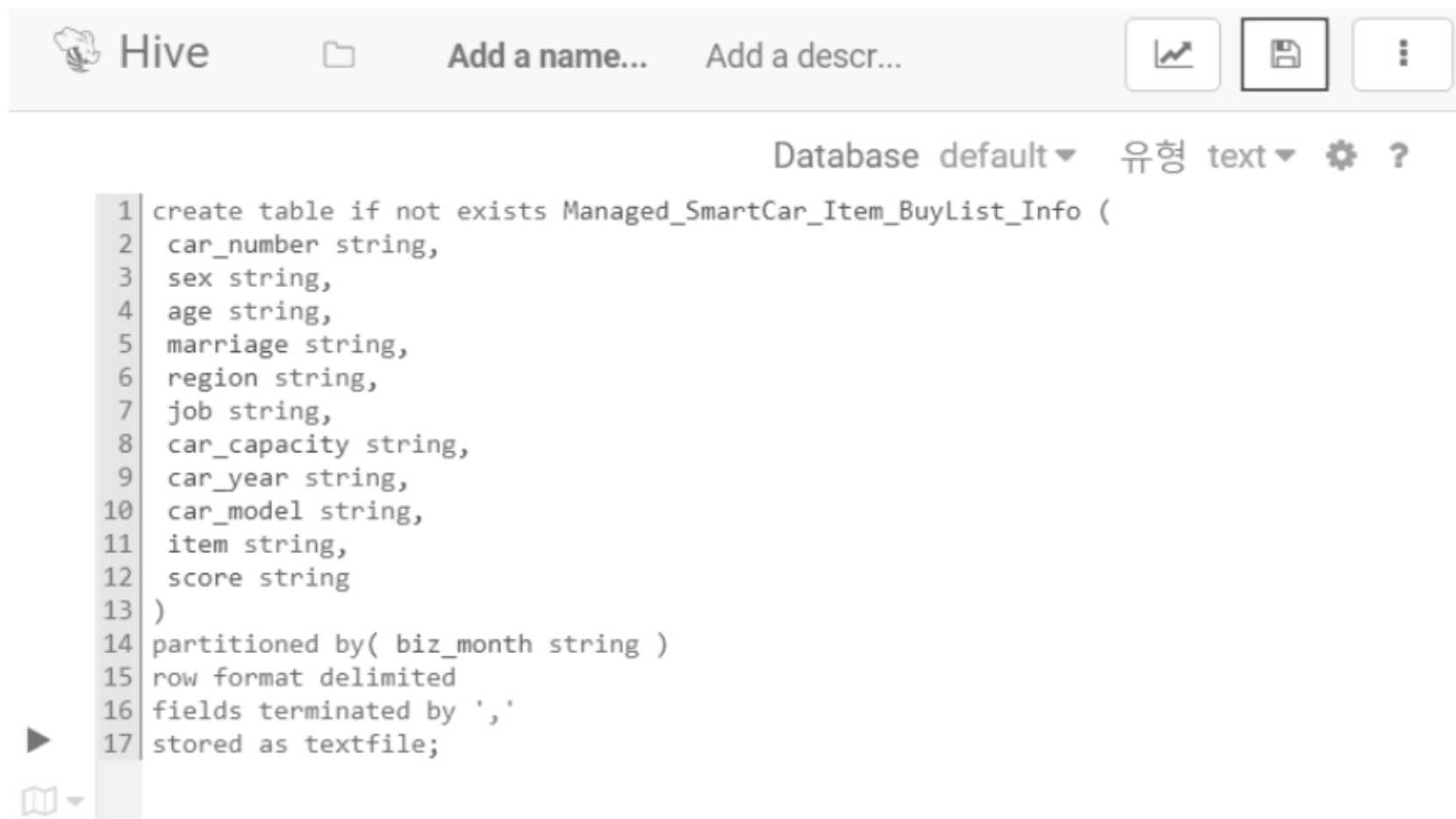




그림 6.113 주제 영역 5의 Managed 테이블을 생성하는 하이브 쿼리

04. 계속해서 내 문서의 /workflow/hive_script/subject5의 위치에 두 번째 하이브 스크립트 파일을 만들어 본다. subject5 디렉터리에서 [새 문서] → [Hive 쿼리]를 선택한다.

05. 하이브 에디트 창이 나타나면 동적 파티션을 설정하고 “차량 물품 구매리스트”와 “스마트카 마스터” 데이터를 조인하는 하이브 스크립트를 작성하고 저장한다. 파일 이름은 insert_table_managed_smartcar_item_buylist_info.hql로 입력한다.


- set hive.exec.dynamic.partition=true;
- set hive.exec.dynamic.partition.mode=nonstrict;


 Hive




Add a name...

Add a description...







Database default▼ 유형 text▼ ⚙️ ?

```
1 set hive.exec.dynamic.partition=true;
2 set hive.exec.dynamic.partition.mode=nonstrict;
3
4 insert overwrite table Managed_SmartCar_Item_BuyList_Info partition(biz_month)
5 select
6     t1.car_number,
7     t1.sex,
8     t1.age,
9     t1.marriage,
10    t1.region,
11    t1.job,
12    t1.car_capacity,
13    t1.car_year,
14    t1.car_model,
15    t2.item,
16    t2.score,
17    t2.month as biz_month
18 from
19     SmartCar_Master_Over18 t1 join SmartCar_Item_Buylist t2
20 on
21     t1.car_number = t2.car_number
22 where
23     t2.month = '202003'
```

그림 6.114 주제 영역 5의 Managed 테이블에 데이터를 생성하는 하이브 쿼리

이번 주제 영역 5의 동적 파티션은 월(Month) 단위다. 앞의 주제 영역 1~4는 일(Day) 단위였다. 그림 6.114의 하이브 QL의 마지막 줄을 보면 “2020년 03월”에 해당하는 물품 구매 데이터를 가져오는 것을 확인할 수 있다.

06. 내 문서의 /workflow/hive_script/subject5에 세 번째 하이브 스크립트 파일을 만들어 본다. subject5 디렉터리에서 [새 문서] → [Hive 쿼리]를 선택한다.

07. 하이브 에디트 창이 나타나면 구매한 상품 리스트를 조회해서 로컬 파일시스템의 특정 위치인 “/home/pilot-pjt/item-buy-list”에 파일을 생성하는 하이브 스크립트를 작성하고 저장한다. 파일 이름은 local_save_managed_smartcar_item_buylist_info.hql로 입력한다.



그림 6.115 주제 영역 5의 차량별 상품 구매 리스트 결과 파일 생성

그림 6.115의 하이브 QL은 SELECT 쿼리를 실행한 결과를 로컬 디렉터리인 /home/pilot-pjt/item-buy-list에 생성하는 하이브 쿼리다. SELECT 구문을 보면 collect_set() 함수를 이용해 차량번호별로 그룹핑한 결과를 하나의 상품 리스트로 구성한다.

08. 이제 워크플로를 만든다. 휴의 상단 쿼리 콤보박스에서 [스케줄러] → [Workflow]를 선택해 우지 편집기를 실행한다.
09. 첫 번째 작업으로 워크플로의 작업 툴박스에서 “Hive 쿼리” 작업을 워크플로의 첫 번째 작업 노드에 드래그 앤드 드롭한다.
10. 사용할 Hive 스크립트 파일을 선택한다. 앞 단계에서 만든 내 문서의 /workflow/hive_script/subject5에 create_table_managed_smartcar_item_buylist_info.hql을 선택한 후 [추가] 버튼을 클릭한다.
11. 두 번째 작업을 위해 워크플로의 작업 툴박스에서 “Hive 쿼리” 작업을 워크플로의 두 번째 작업 노드에 드래그 앤드 드롭한다.
12. 사용할 Hive 스크립트 파일을 선택한다. 앞 단계에서 내 문서의 /workflow/hive_script/subject5 위치에 만들었던 insert_table_managed_smartcar_item_buylist_info.hql을 선택한 후 [추가] 버튼을 누른다.
13. 세번째 작업을 위해 Workflow의 작업 툴박스에서 “Hive 쿼리” 작업을 Workflow의 세 번째 작업 노드에 드래그 앤드 드롭한다.
14. 사용할 Hive 스크립트 파일을 선택한다. 앞 단계에서 내 문서의 /workflow/hive_script/subject5 위치에 만들었던 local_save_managed_smartcar_item_buylist_info.hql을 선택한 후 [추가] 버튼을 누른다

15. 워크플로의 이름을 작성한다. 워크플로 상단의 [My Workflow]를 클릭하고, “Subject 5 – Workflow”로 변경한 후 [확인] 버튼을 클릭한다.
16. 워크플로 작성을 완료한다. 우측 상단의 [저장] 버튼을 누른다.
17. 이제 작성한 워크플로를 작동시키기 위한 예약 작업을 생성한다. 상단의 쿼리 콤보박스에서 [스케줄러] → [예약]을 선택한다.
18. 먼저 예약 작업의 이름을 입력한다. 상단의 [My Scheduler]를 클릭하고 “Subject 5 – 예약”으로 입력한다.
19. 예약 작업이 사용할 워크플로를 선택한다. “예정된 Workflow는 무엇입니까?”라는 메시지 하단에 있는 “Workflow 선택…”을 클릭해 앞서 만든 주제 영역 5의 워크플로 “Subject 5 – Workflow”를 선택한다.
20. 예약 작업이 워크플로를 실행시키기 위한 스케줄 값을 입력한다.
 - 실행 간격: 매일, 05시
 - 시작 일자: 2020년 03월 23일, 00시 00분
 - 종료 일자: 2020년 12월 31일, 23시 59분
 - 시간대: Asia/Seoul
21. 우지의 예약 작업 설정이 모두 끝났다. [저장] 버튼을 누르고 작성을 완료한다.
22. 작성이 완료된 예약 작업을 우측 상단의 [제출] 버튼을 클릭해 실행한다.

23. 제출된 예약 작업을 확인해 본다. 좌측 상단의 드롭박스 메뉴를 클릭해 [Job] → [일정]을 차례로 선택한다. 앞서 등록한 “Subject 5 – 예약”이 “Running” 상태로, 매일 새벽 05시가 되면 등록된 워크플로를 작동시키게 된다. 새벽 5시까지 기다릴 수 없으니 앞서 주제 영역 1에서 소개한 “워크플로 즉시 실행해 보기”를 참고해 곧바로 실행해 본다.

24. “Subject 5 – Workflow”가 정상적으로 작동했는지 확인한다. 상단 쿼리 콤보박스에서 [편집기] → [Hive]를 선택해 하이브 에디터 창에 그림 6.116과 같이 하이브 QL을 “biz_month=202003” 조건으로 작성해서 실행한다.

The image shows the Hive query editor interface. At the top, there's a toolbar with icons for undo, redo, save, and other functions. Below the toolbar, the query text is displayed in a monospace font. The query is a SELECT statement filtering for a specific month. Below the query, there's a play button icon. Underneath the play button, there are tabs for '쿼리 기록' (Query History), '저장된 쿼리' (Saved Queries), and '결과 (100+)' (Results (100+)). The '결과 (100+)' tab is selected, showing a table of results. The table has four columns: 'managed_smartcar_item_buylist_info.car_number', 'managed_smartcar_item_buylist_info.sex', 'managed_smartcar_item_buylist_info.age', and 'managed_smartcar_item_buylist_info.name'. There are five rows of data.

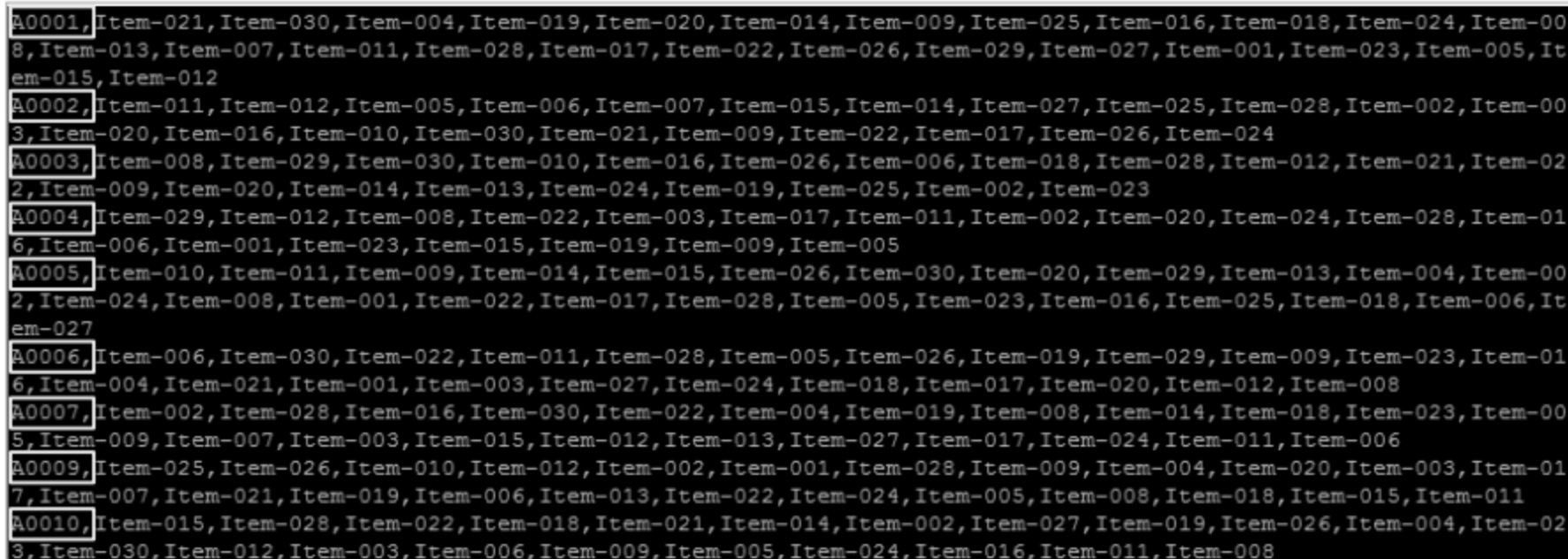
```
1 select * from managed_smartcar_item_buylist_info
2
3 where biz_month = '202003'
```

	managed_smartcar_item_buylist_info.car_number	managed_smartcar_item_buylist_info.sex	managed_smartcar_item_buylist_info.age	managed_smartcar_item_buylist_info.name
1	G0035	♂	65	기혼
2	I0090	♂	40	기혼
3	K0095	♀	48	미혼
4	Y0042	♂	66	기혼
5	W0023	♂	44	미혼

그림 6.116 주제 영역 5 워크플로의 실행 결과 확인 1

차량 번호별로 용품을 구매한 리스트 파일이 Server02의 로컬 파일 시스템에 생성됐는지 확인한다. Server02에 PuTTY로 SSH 접속한 후 다음 명령을 실행한다.

```
$ more /home/pilot-pjt/item-buy-list/000000_0
```



```
A0001,Item-021,Item-030,Item-004,Item-019,Item-020,Item-014,Item-009,Item-025,Item-016,Item-018,Item-024,Item-008,Item-013,Item-007,Item-011,Item-028,Item-017,Item-022,Item-026,Item-029,Item-027,Item-001,Item-023,Item-005,Item-015,Item-012
A0002,Item-011,Item-012,Item-005,Item-006,Item-007,Item-015,Item-014,Item-027,Item-025,Item-028,Item-002,Item-003,Item-020,Item-016,Item-010,Item-030,Item-021,Item-009,Item-022,Item-017,Item-026,Item-024
A0003,Item-008,Item-029,Item-030,Item-010,Item-016,Item-026,Item-006,Item-018,Item-028,Item-012,Item-021,Item-022,Item-009,Item-020,Item-014,Item-013,Item-024,Item-019,Item-025,Item-002,Item-023
A0004,Item-029,Item-012,Item-008,Item-022,Item-003,Item-017,Item-011,Item-002,Item-020,Item-024,Item-028,Item-016,Item-006,Item-001,Item-023,Item-015,Item-019,Item-009,Item-005
A0005,Item-010,Item-011,Item-009,Item-014,Item-015,Item-026,Item-030,Item-020,Item-029,Item-013,Item-004,Item-002,Item-024,Item-008,Item-001,Item-022,Item-017,Item-028,Item-005,Item-023,Item-016,Item-025,Item-018,Item-006,Item-027
A0006,Item-006,Item-030,Item-022,Item-011,Item-028,Item-005,Item-026,Item-019,Item-029,Item-009,Item-023,Item-016,Item-004,Item-021,Item-001,Item-003,Item-027,Item-024,Item-018,Item-017,Item-020,Item-012,Item-008
A0007,Item-002,Item-028,Item-016,Item-030,Item-022,Item-004,Item-019,Item-008,Item-014,Item-018,Item-023,Item-005,Item-009,Item-007,Item-003,Item-015,Item-012,Item-013,Item-027,Item-017,Item-024,Item-011,Item-006
A0009,Item-025,Item-026,Item-010,Item-012,Item-002,Item-001,Item-028,Item-009,Item-004,Item-020,Item-003,Item-017,Item-007,Item-021,Item-019,Item-006,Item-013,Item-022,Item-024,Item-005,Item-008,Item-018,Item-015,Item-011
A0010,Item-015,Item-028,Item-022,Item-018,Item-021,Item-014,Item-002,Item-027,Item-019,Item-026,Item-004,Item-023,Item-030,Item-012,Item-003,Item-006,Item-009,Item-005,Item-024,Item-016,Item-011,Item-008
```

그림 6.117 주제 영역 5 워크플로의 실행 결과 확인 2

그림 6.117을 보면 차량 번호별로 그룹핑된 상품 구매 리스트를 볼 수 있다. 7장에서는 이 구매 리스트 정보를 분석해서 상품 추천 모델을 만들게 된다.

저사양 파일럿 환경: 수집/적재 서비스를 정지한다.

- 플럼 서비스: CM 홈 → [Flume] → [정지]
- 카프카 서비스: CM 홈 → [Kafka] → [정지]
- 스톰 서비스: Server02에 SSH로 접속한 후 다음 명령을 실행
 - \$ service storm-ui stop
 - \$ service storm-supervisor stop
 - \$ service storm-nimbus stop
- 레디스 서비스: Server02에 SSH로 접속한 후 다음 명령을 실행
 - \$ service redis_6379 stop
- HBase 서비스: CM 홈 → [HBase] → [정지]
- 우지 서비스: CM 홈 → [Oozie] → [정지]