6.5 탐색 파일럿 실행 3단계-데이터 탬색&처리 5

스파크를 이용한 추가 데이터셋 탐색

01. 스파크가 설치된 Server02에 SSH를 통해 접속한 후 스파크-셸을 실행한다. 정상적인 스파크-셸 기동이 완료되면 "scala〉" 프롬프트가 나타난다.

```
$ spark-shell
```

```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.4.0-cdh6.3.2
      /_/

Using Scala version 2.11.12 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_181)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

그림 6.62 스파크-셸 시작

6.5 탐색 파일럿 실행 3단계-데이터 탬색&처리 5

스파크를 이용한 추가 데이터셋 탐색

02. 스파크-SQL 컨텍스트를 이용해 하이브에서 생성한 "스마트카 마스터 데이터"인 SmartCar_Master 테이블을 조회할 수 있다. "Age >= 18" 조건으로 스파크-SQL 컨텍스트를 정의해 스파크 DataFrame 변수인 smartcar_master_df에 할당한다.

```
$ scala> val smartcar_master_df = spark.sqlContext.sql("SELECT * from SmartCar_Master where age >= 18")
```

6.5 탐색 파일럿 실행 3단계-데이터 탬색&처리 5

스파크를 이용한 추가 데이터셋 탐색

03. 스파크-SQL로 Age가 18 이상으로 만들어진 DataFrame을 출력한다. 상위 20개의 항목이 표시되고, Age 필드를 보면 18 미만인 데이터는 보이지 않는다.

```
$ scala> smartcar_master_df.show()
```

```
scala> smartcar_master_df.show()
+----------+---+---+--------+------+--------+------------+--------+---------+
|car_number|sex|age|marriage|region|     job|car_capacity|car_year|car_model|
+----------+---+---+--------+------+--------+------------+--------+---------+
|     A0001| 여| 32|    미혼|  서울|프리랜서|        1000|    2009|        F|
|     A0002| 남| 53|    미혼|  충남|    주부|        2500|    2015|        A|
|     A0003| 여| 62|    기혼|  대전|  회사원|        2500|    2012|        B|
|     A0004| 남| 31|    미혼|  광주|  공무원|        2000|    2010|        D|
|     A0005| 남| 67|    미혼|  대구|  공무원|        1700|    2002|        C|
|     A0006| 여| 30|    미혼|  인천|  전문직|        2000|    2016|        D|
|     A0007| 남| 61|    미혼|  전남|개인사업|        1700|    2003|        E|
|     A0008| 여| 20|    미혼|  충북|개인사업|        1500|    2013|        G|
|     A0009| 여| 60|    미혼|  경남|프리랜서|        3500|    2015|        D|
|     A0010| 여| 69|    미혼|  제주|개인사업|        1200|    2003|        A|
|     A0011| 남| 29|    기혼|  충남|    주부|        1000|    2008|        G|
|     A0012| 여| 53|    미혼|  세종|    학생|        2500|    2006|        E|
|     A0013| 여| 43|    미혼|  인천|  전문직|        1500|    2016|        E|
|     A0014| 남| 63|    미혼|  대구|  공무원|        1500|    2006|        C|
|     A0015| 남| 45|    미혼|  충남|프리랜서|        1700|    2012|        F|
|     A0017| 남| 48|    미혼|  충북|프리랜서|        1000|    2010|        B|
|     A0018| 여| 70|    기혼|  경북|개인사업|        2000|    2004|        H|
|     A0019| 남| 32|    기혼|  인천|    주부|        1700|    2004|        C|
|     A0020| 남| 65|    기혼|  대구|    주부|        3500|    2009|        F|
|     A0021| 여| 22|    기혼|  대구|  전문직|        2000|    2001|        A|
+----------+---+---+--------+------+--------+------------+--------+---------+
only showing top 20 rows
```

그림 6.63 스파크-SQL을 이용한 조회 결과 출력

6.5 탐색 파일럿 실행 3단계-데이터 탬색&처리 5

스파크를 이용한 추가 데이터셋 탐색

04. Age가 18세 이상으로 정제된 데이터셋을 하이브의 Managed 테이블인 SmartCar_Master_Over18에 별도로 저장 한다.

```
$ scala> smartcar_master_df.write.saveAsTable("SmartCar_Master_Over18")
```

05. 휴의 [Query 편집기] → [Hive]로 이동해서 스파크-SQL에서 만든 테이블인 SmartCar_Master_Over18가 생성됐는지 확인해 보고 SmartCar_Master_Over18에 "Age 〉 18"에 해당하는 데이터만 존재하는지 하이브 QL로 조회해 본다.

- `Select * from SmartCar_Master_Over18 where Age 〉 18 limit 10`

6.5 탐색 파일럿 실행 3단계-데이터 탬색&처리 5

스파크를 이용한 추가 데이터셋 탐색

06. 표 6.6은 동일 쿼리를 파일럿 환경의 하이브와 스파크에서 각각 실행했을 때 수행 시간을 비교한 결과다. 개인 PC의 가상 환경에서 실행한 결과지만 인메모리 기반의 스파크가 3배 이상 빠른 응답 속도로 측정됐다.

- **실행 쿼리:** `select * from SmartCar_Master_Over18 where age > 30 and sex = '남'`

표 6.6 파일럿 환경에서 하이브 vs. 스파크 성능 비교

	Client	수행시간	환경
Hive	Hue 〉 Hive Editor	62 Sec	▪ OS: 가상화 리눅스 x 3 ▪ CPU: i7 ▪ Mem: 16G ▪ H/D: 256 GB ▪ Hadoop: DataNode x 3
Spark	Spark-Shell 〉 Spark-SQL	20 Sec	

6.5 탐색 파일럿 실행 3단계-데이터 탬색&처리 5

스파크를 이용한 추가 데이터셋 탐색

저사양 파일럿 환경: 스파크 서비스를 정지시킨다.

- **스파크 서비스:** CM 홈 → [Spark] → [정지]

6.5 탐색 파일럿 실행 3단계-데이터 탬색&처리 5

스파크를 이용한 추가 데이터셋 탐색

실습