#하이브적용예제문제\_20190920(Friday)\_Answer Sheet

**#insert overwrite directory '/user/hive/ywjang'**

**select() /from()/group by()/order by()...로하여**

**dwh에서 가져오면됨(by cat())**

1번 : 최대 지연을 일으킨 비행기를 찾아내시오.

[관련 컬럼:arrdelay/ flightnum]

**select year,month,flightnum,max(arrdelay) maxd**

**from airdelay**

**group by year,month,flightnum**

**order by maxd desc**

**limit 10;**

2번 : 가장 많이 취소된 비행일정은 몇 월에 주로 나타나는가?

[관련컬럼 : cancelled/ month]

**hive>select month,count(cancelled)**

**from aiidelay**

**group by month**

**order by month**

3번 : 비행일정이 주로 최소되는 원인은 무엇인가?

[관련컬럼 : cancellationcode/ cancelled]

**select cancellationcode,count(cancelled) countcc**

**from airdelay**

**group by cancellationcode**

**order by canellationcode;**

4번 : 월별 비행거리가 차이가 있는지 살펴보시오?

[관련컬럼 : month/ distance]

**select month, avg(distance) disavg**

**from airdelay**

**group by month**

**order by month;**

5번 : 년도별 취소 비행건수를 비교하시오.

[관련컬럼 : cancelled/ year]

**select year, sum(cancelled)**

**from airdelay**

**group by year**

**order by year;**

6번 : 비행시간(airtime)을 년, 월별로 하여 평균시간으로 나타내시오.

[관련컬럼 : year/month/airtime]

**select year,month,avg(airtime)**

**from aidelay**

**group by year,month**

**order by year asc, month asc;**

7번 : 비행시간(airtime)을 비행기별(flightnum)로 큰순서부터 나열하시오. [10개만 추출]

[관련컬럼 : flightnum/airtime]

**select flightnum, sum(airtime) as m\_at**

**from aidelay**

**group by flightnum**

**order by m\_at desc**

**limit 10;**

8번 : 요일(dayofweek)별로 지연(arrdelay)시간과 비행시간(airtime)의 차이를 보시오.

[관련컬럼 : dayofweek/arrdelay/arrtime]

**select dayofweek,avg(arrdelay), avg(airtime)**

**from aidelay**

**group by dayofweek**

**order by dayofweek;**

9번 : 매월중 어느날(DayofMonth)별로 지연(arrdelay)차이가 있

는지 살피시오.

[관련컬럼 : dayofmonth/ arrdelay]

**select dayofmonth,avg(arrdelay)**

**from aidelay**

**group by dayofmonth**

**order by dayofmonth;**

10번 : 출발지(Origin)와 도착지(Dest)별로 가장 빈번한 지연

(arrdelay)내용들을 비교하시오.[10개만 추출]

[관련컬럼 : origin/dest/arrdelay]

**select origin,dest,avg(arrdelay) as avg\_arrdelay**

**from aidelay**

**group by origin,dest**

**order by avg\_arrdelay desc**

**limit 10;**

11번 : 출발지와 도착지별로 평균적으로 가장 비행시간이 큰 순

서대로 나열하시오.[10개만 추출]

[관련컬럼 : origin/dest/airtime]

**select origin,dest,avg(airtimne) as avg\_airtime**

**from aidelay**

**group by origin,dest**

**order by avg\_airtime desc**

**limit 10;**

12번 : 연도별 요일별로 비행시간 차이가 있는지 살피시오.

[관련컬럼 : year/dayofweek/airtime]

**select year,dayofweek,avg(airtime) as avg\_airtime**

**from aidelay**

**group by year,dayofweek**

**order by year asc, dayofweek asc;**

**###아래내용은 모두 저장타입으로 구성한다음 진행하시기 바람.**

**insert overwrite directory '/user/hive/jaden'**

**select() /from()/group by()/order by()...로하여**

**hive\_data 폴더로 가서 hadoop fs –lsr / 입력**

**위에 생성한 파일 찾은 후 temp 폴더로 이동 dwh에서 가져오면됨(by cat())**

**13. 5년간의 매월 비행시간(airtime) 패턴을 구성시키시오.**

**Select month, sum(airtime)**

**From airdelay**

**Group by month**

**Order by month;**

**14. 5년동안 연별 도착지연(arrdelay)의 패턴을 보이라.**

**Select year, count(arrdelay) as ad**

**From airdelay**

**Group by year**

**Order by year;**

**15. 년도별 비행취소건수를 나타내시오.**

**insert overwrite directory '/user/hive/jaden'**

**Select year, sum(cancelled)**

**From airdelay**

**Group by year**

**Order by year;**

**16. 5년간의 출발지 도착지의 평균적인 비행시간이**

**가장 큰 순서대로 나열하시오.**

**> insert overwrite directory '/user/hive/jaden'**

**> select origin, dest, avg(airtime) as avg\_airtime**

**> from airdelay**

**> group by origin, dest**

**> order by avg\_airtime desc;**

**Limit 10;**

**17. 연도별 가장 많은 비행시간을 가진 비행기와 가장짧은 비행시간을 가진 비행기는?**

**insert overwrite directory '/user/hive/jaden'**

**select year, flightnum, airtime**

**from airdelay**

**join (select year y2, max(airtime) max\_at from airdelay group by year) t2**

**on airdelay.airtime=t2.max\_at**

**where airdelay.year=t2.y2**

**order by year asc;**

**-in case of 년도별 최소 비행시간**