Hive 효율적으로 사용하기

- Hive는 데이터베이스가 아닌 **데이터 처리를 위한 배치 처리 구조**
- 읽어 들이는 데이터의 양을 의식하면서 쿼리를 작성해야 원하는 성능이 나올 수 있음
  - 가능한 의식을 해서 sub query 안에서 fact table을 작게 하도록 해서 중간 데이터를 줄여야 함(`Disk I/O'를 최대한 줄이기 위해서)
    - 그냥 JOIN하게 되면 매우 거대한 중간 데이터를 만들고, 메모리를 낭비할 수 있음
- 데이터의 편향을 피해야함
  - 데이터의 편차(data skew)는 고속화를 방해함
  - 분산 시스템의 성능을 발휘하기 위해서는 데이터의 편차를 최대한 없애고, 모든 노드에 데이터가 균등하게 분산되도록 해야 함
  - 중복을 제거하면 부하를 분산시킬 수 있음