* 필요 개념
- Data Warehouse(DW) : 하나의 통합된 데이터 저장공간으로서 다양한 운영 환경의 시스템들로부터 데이터를 추출, 변환, 통합해서 요약한 데이터베이스
    - 데이터베이스가 데이터는 잘 저장하나, 저장된 데이터들을 제대로 활용하지 못 하는 것에서 착안
    - 기본적으로 관계형 데이터베이스가 있는 상태를 가정하여 DW를 구성하며 동영상이나 음악처럼 DB에 저장할 수 없는 파일도 필요한 부분을 추출하여 보여주어야 함
- ETL(Extract, Tranform, Load) : 데이터를 추출하고, 변형하여, (Data Warehouse에) 적재하는 과정을 일컫는 말
- BI(Business Intelligence) : 데이터 추출/통합/리포팅을 위한 기본도구 집합, DW에서 분석된 데이터를 통해 숨겨진 패턴을 찾아냄
    - ETL을 통해 뽑아낸 데이터를 DW에 적재하고, BI를 이용하여 분석하는 기본 과정을 거침

* Redshift란?(모든 개념 중요!!!)
- PostgreSQL를 기반으로 하는 AWS의 Data Warehouse Service
    - OLAP(Online Analytical Processing)에 사용되는 서비스
- 모든 데이터를 표준 SQL 혹은 BI 도구를 사용하여 효율적으로 분석할 수 있도록 지원
- 대량 병렬처리(Massively Parallel Processing, MPP) 엔진을 통해 복잡한 쿼리라도 빠른 속도로 실행하여 대용량 처리 가능
- 열(Column) 단위 데이터 저장방식
- COPY 명령어를 통해 Amazon EMR, Amazon DynamoDB, S3로부터 데이터를 병렬 로드 가능
- Enhanced VPC Routing을 통해 클러스터와 VPC 외부의 COPY, UNLOAD 트래픽을 모니터링할 수 있음
- WLM(Workload Management)를 통해 사용자가 작업 부하 내 우선 순위를 유연하게 관리하도록 지원

* Redshift의 구성
- 클러스터 : Redshift의 핵심 요소로 하나의 리더 노드와 다수의 컴퓨팅 노드를 가지고 있는 구성 요소
- 리더 노드 : 클라이언트 프로그램과 일어나는 통신을 비롯해 컴퓨팅 노드간의 모든 통신/작업 관리하는 노드로서 쿼리를 계획하고 결과를 집계함
- 컴퓨팅 노드 : 실제 작업을 수행하고 쿼리 결과를 리더 노드에 보내는 노드로 각 노드마다 전용 CPU와 메모리 내장 디스크 스토리지를 따로 보유함
- 클러스터는 Single AZ 모드만을 지원하며 Multi-AZ 모드를 지원하지 않음(Multi-AZ 모드는 향후 가능할 것으로 보임)

* 스냅샷과 백업
- 스냅샷은 Redshift의 특정 시점을 백업한 것으로 S3 내에 저장되며 증분식 저장 방법을 사용함
- 새로운 클러스터에 스냅샷을 복원하여 사용할 수 있음
- 스냅샷은 2가지 모드가 있음
    - 자동 모드 : 8시간 혹은 5GB마다 생성되도록 일정을 예약할 수 있는 모드로 보존 기간은 1일에서 최대 35일
    - 수동 모드 : 언제나 생성할 수 있는 모드로 사용자가 직접 삭제할 때까지 보존됨
- 스냅샷을 다른 리전에 자동으로 복사되도록 구성할 수 있어 다른 리전에 클러스터를 즉시 배포하는 것이 가능

* Redshift vs RDS
- Redshift는 OLAP(보고 및 분석)에 사용되지만, RDS는 OLTP(온라인 트랜잭션 프로세싱) 워크로드에 사용
- Redshfit는 대용량 데이터 세트를 대상을 복합적인 분석 쿼리를 빠르게 실행하는 것에 목표를, RDS는 단일 행 트랜잭션에 목표를 둠

* Athena
- 표준 SQL을 활용해 S3에 있는 데이터를 분석할 수 있는 "Serverless" 대화형 쿼리 서비스
- Athena로 데이터를 로드할 필요 없이 S3에서 데이터를 쿼리하고 분석할 수 있음
- CSV, JSON, ORC, Avro 등의 형식을 지원함
- 데이터 시각화를 위해 Quicksight와 통합되어 사용하며 이를 통해 대시보드와 보고서를 생성함
- Federation Query(연합 쿼리)를 지원하여 S3 이외의 소스에 데이터가 있는 경우에도 Athena를 통해 해당 소스의 위치에서 쿼리할 수 있음

* Athena의 요금 책정과 절감
- Athena는 쿼리당 요금이 부과되며 쿼리로 스캔한 데이터의 양을 기준으로 요금이 청구됨
- 아래 2가지 방법을 통해 쿼리로 스캔한 데이터의 양을 줄이는 것이 요금 절감으로 이어짐
    - S3에 데이터 저장시 압축하여 저장하는 것

- S3에 데이터 저장시 파티셔닝을 통해 데이터를 분류하여 저장
- 위 2가지 방법을 통해 Athena가 쿼리하는 데이터의 양을 줄여 쿼리 속도를 향상시키고 요금을 절감할 수 있음

- 실패한 쿼리에 대해서는 요금이 청구되지 않지만 취소한 쿼리는 취소한 시점까지 스캔한 데이터에 대해 요금을 청구함