ETL(Extract, Transform, and Load) Process

엑스트라이버 데이터 사이언스 지원자
안지용

ETL이란 관계형 데이터베이스 관리 시스템(RDBMS)에서 데이터를 추출(Extract)하고 변형하여(Transform) 데이터 웨어하우스에 적재(Load)하는 절차다. ETL은 주기적으로 이뤄지는 작업이기에 자동화, 문서화, 애자일(Agile)하게 관리해야 한다. ETL처리에 앞서 원본 데이터가 추출되어 데이터베이스에 문제없이 저장되려면 논리적인 map을 구상 해야한다. map은 단계별 데이터가 잘 이행되었는지 검증 작업에 사용된다.

1단계_Extract

데이터 원천(Source)에서 꺼낸 데이터는 준비 영역(Staging area)으로 넘어간다. 준비 영역은 추출된 데이터를 데이터 웨어하우스로 적재하기 전에 검증하는 공간이다. 데이터 추출 방식으로 Full 방식, Partial 방식(알림 유/무)이 있다. 어떤 방법을 사용하더라도 데이터 원천 시스템의 성능과 반응 속도에 영향을 끼치면 안 된다. 추출단계에서 데이터의 key 값이 모두 있는지, 조각나거나 복사된 혹은 불필요한 데이터는 제거한다.

2단계_Transform

추출된 데이터는 raw 상태여서 그대로 사용할 수 없기 때문에 여러 함수를(정해진 map에 맞게 변형시키는 함수) 적용한다. 가치를 더하고 통찰력 있는 비즈니스 정보를 만드는 단계이기도 하다. 어떤 변형도 필요하지 않은 데이터는 direct move data 또는 pass through data로 부른다. 데이터 변형할 때, 공백이나 오기입 여부를 확인해본다. 특히, 문자열 데이터를 눈여겨 봐야한다.

3단계_Load

원하는 데이터 웨어하우스의 데이터베이스에 데이터를 적재하는 단계다. 주로 심야 시간에 데이터가 적재되기 때문에 Load 단계는 최적화 되어야 한다. 적재에 실패할 경우, 리커버 메커니즘이 구성되어야 실패 지점부터 다시 시작하여 데이터 통합에 손실이 없도록 해야 한다. 적재의 방식은 Initial load, Incremental load, Full refresh 방식이 있다. key란에 null값은 없는지 재확인하며 여러 속성을 결합한 테이블이나 시계열 테이블도 불러오는지 검증해본다.

ETL을 다룰 때 모든 데이터를 정제하는 시도는 옳지 않다. 시간이 많이 필요하기 때문에 어떤 데이터를 정제할지 계획을 세운 후 시행한다. 계획에 맞게 데이터를 처리하여 데이터의 신뢰도를 높이는 편이 낫다. 데이터 정제할 때는 소요되는 비용을 계산해야 하고 저장 비용을 줄이기 위해 데이터 크기와 사용의 Trade-off 관계를 고려해야 한다.