**3차 멘토링 피드백**

2024/05/25 2팀 김건호

1. **공통**

공통 장표는 같이 만들어서 PPT를 작성할 것

공통 장표를 만들고 나서 내가 한 부분은 따로 만들어야 함.

→ 얘네들 같은 팀이었구나 라는 느낌을 들게 해야 함

팀 공통적으로 고민하는 것, 노력할 것들도 어필할 것.

WBS에 내가 담당한 것을 하이라이트로 보여주는 것이 필요

전체적으로 백업에 대한 설명이 부족함

발표할 때 이미지는 운영적 영역만 보여주면 됨

현재 코드 파이프라인 자동화만 있지, 데이터 파이프라인 자동화에 대한 내용이 다들 없음

* stepfunction, airflow로 추후에 자동화하는 것을 권장드린다고 말해도 좋음
* glue workflow를 사용해도 좋을 것 같음

1. **개인**

**PPT**

데이터 분석이 필요한 이유는 굳이 PPT에 적용시킬 필요는 없음.

어차피 발표시에 고객사가 요청했다는 가정이기에⇒ presales 영역임.

발표 용어 사용 시에는

* ‘~이렇게 하는 것으로 제안 드립니다’ 식으로 이야기 하기
* sa는 컨설턴트 입장에서 제안하는 직무임

PPT 이미지는 깔끔한 걸로 쓸 것

데이터 파이프라인 아키텍쳐에 나와있는 순서대로 ppt 또한 설문조사, 로그정보, 고객정보 순으로 적기

눈에 보기 쉽게 매핑하는 것도 중요함

PPT 내용은 가장 적합하지만, 개인이 진행한 내용이 장표에 들어가 있어야 함.

본인이 한 작업은 ppt에 캡쳐해서 보여주기

Q. csv경로를 data source를 잡으면 crawler가 왜 csv 데이터를 못 읽는가?

-> crawler는 샘플 데이터도 아마 일정 수준 이상 있어야 할 것임.

Q. 왜 전처리를 대부분 하는가?

-> 데이터 웨어하우스는 분석하기 위한 용도로 영구 저장 DB임. 모든 데이터를 통합해서 만듦

그대로 Quicksight에서 표현하기에는 데이터 웨어하우스는 너무 큼

그래서 용도에 맞게 압축해서 데이터 마트를 구성함. 보고 싶은 것을 볼 수 있게

누군가 한 번만 돌려주면 될 것을 계속 돌리게 되니 SQL은 비용이 많이 발생하게 될 수 밖에 없는 구조

Q. ElasticSearch는 실시간인가? 실시간이면 정말 완전 실시간으로 수집, 정제, 분석, 시각화가 가능한가?

→ 아니다. 모든 데이터 분석에서 실시간은 준실시간이다.

컬럼이 추가 될 경우에 crawler가 효율적

ex)partition

s3://sales/year=2024/month=05/day=24/a.parquet

s3://sales/year=2024/month=05/day=24/b.parquet

이런식으로 파티션 돼 있음

크롤링 할 때 파티션의 메타 데이터를 한 번 확인함. 즉, 매핑의 개념

하지만, 메타 정보를 가지고 조회하기 때문에 메타 정보가 없으면…(파티션 정보가 없으면) 안됨. 장애가 일어남

테이블에 파티셔닝을 수동으로 추가해야함.

→ glue crawler가 이것을 알아서 함

crawler는 데이터를 긁어오는게 아니라 읽는 개념임

파티션이 달라지면 같은 테이블로 합쳐서 못 읽어옴

TS) 테이블 이름에 -가 아닌 \_를 해야 partionts 잘 적용됨

데이터 분석에서는 실시간이라는 개념이 잘 적용 안됨

firehose의 buffer 예시

→ 일례로 10분 뒤에 데이터가 쌓이면 누군가는 실시간, 누군가는 버그냐 이럴 수도 있음

spice 설정시 주기적으로 걸려 있는 SQL을 동작시키게 함

만약, sql 작업과 spice 주기가 안 맞을 시 sql까지 온 시간과 spice 시간이 더해져서 기다려야 하는 상황이 발생

**Kinesis Firehose**

새 줄 구분자 → datafirehose 에서 json 레코드 구분을 위해 사용

만약 한 줄에 다 들어가면 레코드가 하나만 생김. 그래서 사용

동적 파티셔닝

→ 데이터 안에 있는 값을 씀. log나 firehose, s3에 들어가는 시간이 아니라!

→ 데이터가 전송되는데 시간이 걸리기 때문에 사용함

→ 로그에 있는 시간을 꼭 써야 하는 경우에 사용

→ 물론, 로그는 어지간하면 씀

S3 버킷 접두사

→ 예전에 동적 파티셔닝이 없었을 때 사용하였음

→ 하지만 이것도 시간 불일치 문제가 생김

데이터 레코드 압축

→ 압축 파일 종류가 4가지 있는데 다 테스트 해 볼 것

하나가 crawler 안 될 수도 있음