<지원>

* 설문조사할 때 일반적으로 구글 form를 사용.
* **마케터 입장에서는 설문조사를 위해서는 굳이 인프라로 구축하지 않는다.**
  + 쉽게 공짜로 사용할 수 있는 설문조사 도구가 있는데, 비용 효율적이지 못하다.
* 설문이라는 것은 교육같은 것 할 때 교육에 대한 피드백 내용을 구글 form을 사용하고, 광고를 활용한다고 할 때 facebook 등으로 Google Ads같은 광고 노출을 통해 데이터를 수집하기도 한다.

**질문**

* DB가 있으면 굳이 EC2에 접속하지 않고 DB 접근 권한만 받아서 사용. 비용적인 부분 성능적인 부분 사용할 사람의 skillset에 따라 다르다.
* 그들이 쓰는 skillset에 맞춰서 솔루션 제시가 나음

원래는 고객이 관리를 하는게 맞는데, 데이터 관련 조직이 없는 고객이 있으면 다름.

MSP의 운영 같은 경우는 서버의 이슈에 대한 대응을 하는 것인데, 데이터 부분의 운영은 리포트 작성 및 파이프라인의 모니터링, 대시보드 내용 추가 등 일부 대응을 하는 것

인사이트가 나온다는 것은 새로운 뷰를 요청하고 싶은 것이 당연함. 인사이트가 있어야 그 조직이 살아남. 그런 작업이 꾸준히 일어나야 하는데 기술적인 부분이 부족하다면 추가 수정해달라라는 요청이 들어옴

Q. 고객은 인사이트를 더 생각하나? 비용적인 부분을 생각하나?

A: 성능은 당연하고, 비용을 줄일 수 있으면 좋다.

A : QuickSight를 실시간으로 쓰고 싶으면 SPICE를 통해 실시간으로 새로 데이터를 받아오게 할 것인지.(30분 주기로 설정해도 실시간이라 생각할 수 있다.)

<민경>

* Crawler를 돌려야 하는가?에 대한 내용으로 크롤러를 돌리면 카탈로그를 만들 때
  + athena에서 사용할 때 효율적으로 사용하기 위해 파티셔닝
  + 연/월/일 형식으로 날짜로 파티셔닝 이렇게 파티셔닝하면 크롤러는 파티션으로 인식함.
  + Catalog 테이블 안에 파티션 정보가 존재. 하나하나 모두 키 값임.
  + 파티션이 많아질 경우 관리하기 힘들어 수동으로 하나씩 파티셔닝하는 것이 좋음
  + 하나의 버킷에 연/월/일로 파일을 파티셔닝
  + **파티션 키가 잡혀있는 경우 키 기준으로 데이터가 있다는 것을 기준으로 데이터가 없으면 조회 자체를 안함.**
* .csv 파일로 크롤링하는 것은 바람직 하지 않다.
  + orders.csv 파일 파티셔닝은 어떻게?
    - 하나하나 .csv파일도 객체기 때문에 키다.
      * 버킷/shop\_orders.csv/ 까지 키다. 객체는 하나씩만 생성됨. 만약 구매 테이블 정보처럼 새로운 데이터가 추가된다면? 새로운 객체로 생성되어 들어갈 것..
      * 따라서 데이터가 같이 있어야 된다면 버킷에 모아 두어 연/월/일로 나누어 파일로 저장할 것.
      * orders같은 경우에는 파티셔닝이 되어 있어야 됨.
* 아이디어 같은 경우에는 가장 현실적임.
* 사람들은 재확인하고 싶어하는 습성이 있다.
  + 리뷰가 많이 달려있으면 판단하는 것에 대한 확신을 갖게 된다.

<건호>

* 키네시스 firehose 액세스 로그 설정 중 에러 발생
  + **버퍼 사이즈 및 시간쪽 문제** 같다..
* **키네시스 Stream을 쓰는 이유**
  + 카프카는 스트림에도 불구하고 DB다고 말하는 사람이 있다
    - ⇒ 실시간 데이터를 저장한다.
    - 특정 기간동안만 저장.
      * why? 버퍼 역할 및 부하도 분산
      * → **누락되지 않게 실시간 데이터를 저장**하고 보내는 것은 다른 것.
* 단순 액세스 로그 자체는 볼 것이 없다. 따라서 가공해서 클릭 스트림을 분석하기도 한다.
  + 누가 접근했는지를 알 수 없다.
  + 유저 정보의 세션 id 값과 ip 정보를 매핑 테이블로 만들어 저장해두었다가 로그인 안된 상태에서도 세션 id와 ip값을 매핑해서 확인.
* 사용자 요구 및 행동 이해가 구체화되어 있지 않아 어떻게 분석하는지 모르겠다.
* 이 사람이 사이트에서 어디로 이동하는지 분석이 되어있어야한다. 그 다음에 경험 최적화가 가능하다.
* **페이지 네이트 레포트 옵션 → 자동으로 켜져있는 것 조심. 과금 주의.**

엔드포인트를 최소화

**공통 방향성**

리뷰 개수와 판매율 내용이 혜성 + 민경 추가

* 섞기 좋을 것 같음.
* **지원**님 같은 경우에는 비현실적인 경우가 있다. Source로 많이 쓰는게
* exel 작업 로딩에 대한 수요가 많다. 로딩 및 파싱해서 아테나 분석
  + 그 데이터를 S3로 밀어넣는 것이 제일 좋을 것 같다.
* 스트림으로 꽂는 것이 모범사례임.
* DB 데이터도 가져올 필요가 있다. 고객 정보라던지..
  + 배치성이면 된다.(상품 정보 같은 경우에는 웹사이트에서 관리하는데, 긴급성을 가진 것이 아니다. 데일리로 반영되도 충분함.)
* 액세스 로그같이 사용하기 위한 스크립팅 작업
* 배치성 데이터(DB) + 구매 이력()
* GA(Google Analytics)의 스크립트 예시를 통해 화면에서 일어나는 이벤트를 캐치할 수 있다.
  + 실제 일어나는 액션에 대한 데이터를 던지게 되어있음.
* 쉽게 할 수 있는 방법은 javascript를 만들어서 현재 사람의 세션 id 정보라던지 간단하게 넘기는 방식도 있다. 이렇게 실시간성 넘기는 데이터를 저장해서 사용.
* **상품 사진을 클릭했을 때 javascript를 통해 고객 데이터 및 액션에 대한 정보를 넘기는 것.**

**공통 과제 설정**

주제를 버무리면 괜찮을 것 같다.

특정 서비스 기준으로 설정 후 R&R 나눠보기.

정리를 한다면

**데이터 관점에서는 수집, 데이터 마트 구성, 시각화**

수집

* 설문
  + excel로 떨어지면, 수기 or 자동으로 S3에 데이터 저장
  + 파싱하기 위해서는 Lambda를 통해 코드작성 필요.
* DB 데이터

일반 트랜잭션 데이터보다는 고객 정보 동기화 용으로 쓰는게 좋다. → S3

* Snapshot
* 실제 트랜잭션에 대한 데이터
  + Access로그만으로는 정보가 부족함.
  + 로그 수집에 대한 고민
    - javascript 등의 스크립트로 수집
    - 수집 데이터에 대한 수집 서버
    - 수집 서버를 통해 → S3
* S3
  + 모인 데이터들을 통해 Glue ETL 작업
* Glue ETL
  + Crawler
    - meta partitioning
  + Data Catalog
    - 생성된 테이블을 통해 Athena에서 쿼리 가능.
* QuickSight
  + 대시보드를 어떻게 구성할지

**dashboard가 제일 중요**

* 시연을 보여주는 디테일도 있으면 좋을 것

S3 하나의 버킷이든 여러개든 상관 없다.

* 고객 정보에 대한 버킷 데이터는 특정 유저만 관리할 수 있도록 설정할 것
  + 더 나아가면 권한 있는 사용자만 확인할 수 있도록
* 버킷 나누는게 제일 심플하다.

현실적으로 사용하기 위한 구체화

1. 인사이트 도출
2. 대시보드 구체화
3. 태스크 분리

WBS형식으로 쪼갠다.

* 필요한 데이터 정의 및 적용하는 것은 실무쪽
* 어떤 데이터를 사용할지 정의하고 더미 데이터를 넣어두어 테스트
* 태스크별 구체화
* R&R 을 통해 상태 체크

1. 고객이 가지고 있는 데이터 조사.
2. Bucket 구성 및 데이터 어떻게 들어올지 정의
3. 카탈로그화

====================================================오후

NAT 두 개 두는 이유로 하나 AZ가 문제 발생시에도 가용성을 챙기기 위해서..

NAT처럼 아웃바운드 호출할 것 아니면 크게 사용되지 않는다.

* 서버 관리, 패치, 결제 시스템 통신

멘토님:

* 조금 더 개인이 하는 것을 편하게 오픈해서 얘기하는 자리가 필요

**지원**

멘토님:

일반적으로 클라이언트 툴을 사용하게 되는데 session manager를 사용하지 못할 수 있다.

고객에게 가이드를 주라고 할 것인데, 일반적인 가이드를 제시할 것.

나쁘지는 않지만 일반적으로 사용하지 않다.

Q. Session Manager를 사용할 때 NAT를 이용하여 사용하기도 하는데 ..? 실무에서는?

A. 일반적으로는 Private key두고 배스천으로 접근

Q: 스냅샷을 S3 Glacier를 통해 백업하는 것이 좋을지?

A:

스냅샷 자체가 RDS 데이터를 S3로 내리는 것인데.. 큰 의미가 있나?

AWS Backup을 통해 백업하는 것이 좋지 않을까?

Q: EC2 인스턴스 선택할 때 t시리즈의 버스터 인스턴스 중에서 최신 세대 버전을 사용은 어떤가요?

A:

T타입은 실무에서 잘 사용되지 않는다. unlimited 설정이 있어서 크레딧이 없어도 사용할 수 있게 되긴 했다. 예전에는 unlimited 모드가 없을 때는 크레딧이 부족해지면 서버가 다운되는 상황이 발생.

T타입의 시리즈는 저렴하게 사용하는 모드임. 테스트 용이라는 것으로 전제가 있음.

**WEB: IOPS 쪽이긴한데, 네트워크 트래픽인지, 디스크 트래픽인지… 대부분 M 타입 씀**

**WAS : C타입(CPU)**

**DB : R타입(Memory)**

멘토님: 모니터링 지표는?

지원: CPU 사용률, Memory 사용률, 500번 상태코드

멘토님: WAF, Shield 실습은 해봤나요?

지원: 아직 미정. 실습

**S3가 아닌, 웹서버 사용 타당성 검토**

디스크는 느릴 수 밖에 없다. 가장 좋은 것은 캐시에 적재하는 것이 좋겠지만, 비싸다.

* S3는 저렴하기 때문에 사용하게 된다.
* CloudFront는 정적 컨텐츠를 캐시해서 S3의 정적 페이지를 빠르게 접근할 수 있도록
* 리전엣지캐시가 있어서 한국의 고객은 1번만 접근(캐시)

**건호**

* 굳이 codecommit을 사용할 필요는 없지 않을까..
  + 실제로 개발자들은 github를 위주로 사용한다..!

Q 퍼블릭 서브넷에 웹을 올리지 않는 이유?

* 보안사고 가능성이 열린다..!
* 가능하면 private subnet에 인스턴스 동작시킬 것.
* SecurityGroup 소스에 따른 인바운드 포트 설정 필요.

**민경**

코드를 잘 작성하는 것도 중요하다.!

모니터링 CPU, Memory 사용량

Q: CPU 임계치를 넘었을 때 알람이 발생하도록 설정하고 싶은데 CPU 임계치를 어떻게 설정하는 것이 좋을까요?

A: 서비스에 따라 다를 것 같다. 일반적으로 비용 최적화로 책정하는데, 고객사별로 상시 60-70%로 유지하는 회사도 있을 것이다. 임계치는 잡기 나름이다.

멘토님: 버킷안에 폴더를 나눠 관리하는 것이 좋을까요?

민경: 권한적인 부분을 고려했을 때는 버킷을 나눠서 정책 설정하는 것이 좋다..

멘토님: 백업 부분은 AWS Backup, S3 사용할 것이면 Lifecycle을 어떻게 설정할 것인지 기준을 잡을 필요가 있을 것.

* 실제로는 1년 or 2년 데이터만 백업할 것이고 이 기간 전까지는 법률상 데이터를 보존해야 한다라는 등 → 정책 기준으로 ~ 설정 해보았습니다.
* 가이드 경우 고객사에 어떻게 유지보수 할 지에 대한 내용을 가이드.

**혜성**

1. ETL에 대한 자동화 하면 좋긴 하다...
2. DBT, APP Flow, Stepfunction
3. Glue WorkFlow 데이터 파이프라인 자동화하면 좋긴 하다. -> 데이터 파이프라인만 먼저 구축하고, 자동화까지는 이외의 문제. -> 강사님에게 조언을 받아보자..

RAW 데이터에서 대시보드 뜨는 것까지의 코드를 구하고, 데이터가 들어왔을 때 변경이 되는 것도 보여주면 좋다.