데이터 엔지니어 초급 5일차

LGDE.com 서비스 지표 파일럿 프로젝트

Park Suhyuk
Data Ingestion Team Leader


psyoblade


psyoblade

hive 2.3.2
spark 2.4.5
fluentd 3.8.0
sqoop 1.7.4
docker-ce 19.03.13
ubuntu 18.04 LTS

NCSOFT®

목차

지표 서비스 개요

지표 서비스 개요

What to do as a 'Data Engineer'?

가상의 인터넷 쇼핑몰 LGDE.com 을 오픈 예정에 있습니다. "데이터 엔지니어"는 어떤 지표를 추출해 두어야 **데이터에 기반한 의사결정**에 도움이 되는지, 그리고 어떻게 **모델링** 해야 하는지, 혹은 향후 **고객이 늘어났을 때에도 유연하게 대응**할 수 있는 데이터 시스템에 대해 고민해보겠습니다

오픈 첫 날 매출은 얼마나 되는지?

오늘 신규 가입 유저는 얼마나 되는가?

오늘은 몇 명이나 접속 하였는가?

우리 서비스에 고객이 평균 체류 하는 시간은 몇 분이나 되나?

구매 유저는 몇 명이나 되는가?

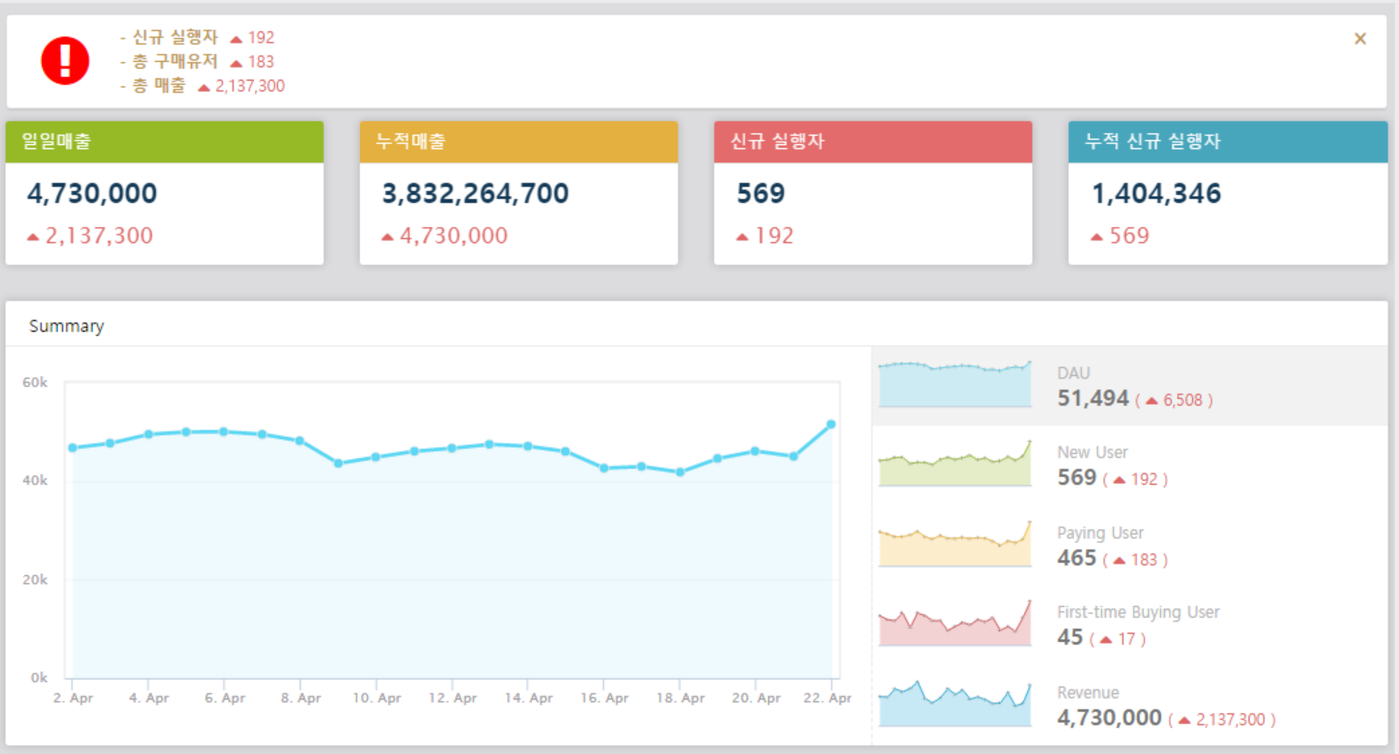
오픈 첫 주의 누적 매출은 얼마나 되는지?

일자 별 최초 구매 유저 수는?

우리 고객의 생애 가치 (LTV)는 얼마인가?

고객 한 명당 평균 매출은 얼마나 되는가?

구매 유저 당 평균 매출은 얼마나 되는가?



http://help.igaworks.com/hc/ko/3_3/Content/Article/adbrix_analysis?share=True

활용 가능한 데이터

활용 가능한 데이터

What 'Dataset' we have?

수집 가능한 데이터는 더 많지만, 보다 명확한 이해를 돕기 위해 가장 핵심 데이터 집합만 활용하기로 합니다. 고객의 정보와 같이 **수시로 변경 가능성**이 있는 데이터의 경우 반드시 동일한 복제본을 일자 별로 유지되고 있어야 향후 **언제든지 해당 시점의 데이터를 분석하거나 비교**가 가능합니다. 또한 다양한 유형의 데이터가 존재 하므로 **일관된 저장소에 언제든지 접근이 가능한 저장소**에 저장될 필요가 있습니다. 그리고 데이터의 크기가 충분히 커지는 경우에도 일정한 **저장 및 조회 성능을 보장**하기 위해서는 분산 저장 및 처리가 가능한 Hadoop Eco System 과 같은 플랫폼의 도입이 필요합니다.

한글명	영문명	유형	설명
고객정보	user	Table	고객 정보는 고객관리 부서에서 관리하며 Maria DB 에 저장되어 관리되고 있습니다 가입 이후에 고객의 행위에 따라 등급 정보는 일 별로 변경되어 해당 시점에만 정확한 정보가 유지됩니다
매출정보	purchase	Table	매출 정보는 Sales 부서에서 관리하고 MySQL 데이터베이스에 저장 되어 있습니다 로그를 통해서도 저장가능하지만 유실 지연 문제가 있으므로 실제 매출 테이블이 가장 정확합니다
접속정보	access	File	접속 정보는 Internet 외부 서비스 제공 부서에서 관리하며 해당 서버의 임의의 경로에 저장되고 있습니다 로컬 저장의 공간 제약이 있기 때문에 일주일의 데이터만 롤링 되며, 과거 데이터는 삭제됩니다 로그인과 로그아웃 시에 해당 이용자의 아이디와 시간을 남깁니다

https://docs.docker.com/compose/reference/overview/

요구사항 분석 및 지표 설계

지표 설계

User Analysis KPI (Key Performance Index)

고객을 분석을 위한 기본 지표를 정의하고, 해당 지표를 계산하는 산식을 설계합니다. 해당 지표를 추출하기 위해 필요한 데이터가 존재하는지, 존재하지 않는 경우는 로그 정의 부터 데이터 수집에 이르기까지 유관부서와 협의 및 관련 인프라의 설계가 필요합니다

한글명	영문명	설명
일일 활성 유저	**DAU** Daily Active User	오늘 접속 로그에 로그인 혹은 로그아웃 정보가 남아있는 유저 수 (access)
일일 신규 유저	**DNU** Daily New User	오늘 처음 가입한 고객의 수 (access, dimension)
일일 매출	**DR** Daily Revenue	오늘 발생한 총 매출의 합 (purchase)
일일 구매 유저	**DPU** Daily Paying User	오늘 한 번이라도 구매한 이력이 있는 고객의 수 (purchase)
일일 최초 구매 유저	**DFPU** Daily First-time Paying User	오늘 처음 구매한 고객의 수 (purchase, dimension)
일일 평균 접속 시간	**DAAT** Daily Average Access Time	로그아웃 시간 - 로그인 시간 = 오늘 접속한 고객의 평균 서비스 이용 시간 (access)
누적매출	**AR** Accumulated Revenue	오픈 이후에 발생한 전체 누적 매출 금액 (purchase, dimension)
유저 당 평균 매출	**ARPU** Average Revenue Per User	총 매출 / 전체 유저 수 = 유저 당 평균 매출 금액 (access, purchase)
구매 유저 당 평균 매출	**ARPPU** Average Revenue Per Paying User	총 매출 / 구매 유저 수 = 구매 유저 당 평균 매출 금액 (access, purchase)

http://help.igaworks.com/hc/ko/3_3/Content/Article/adbrix_analysis?share=True

스냅샷 및 파일 수집 (테이블, 로그)

수집 실습 – 일일 고객 정보 스냅샷

Daily User Snapshot

지표정의 : 지정한 일자의 이용자 정보

지표산식 : 단순 테이블 수집 데이터 제공

입력형태 : 원본 소스의 접속 정보는 jdbc://mysql://mysql:3306/testdb, 접속 계정과 패스워드는 sqoop, sqoop 이며, 테이블 이름은 user 입니다

출력형태 : 저장 타깃의 포맷은 parquet 이며, 서버의 로컬 경로 /tmp/target/user/yyyyMMdd 경로에 저장하며, 아래와 같이 설계합니다

컬럼 명	컬럼 타입	설명
u_id	integer	아이디
u_name	string	이름
u_gender	string	성별
u_signup	integer	가입일자
dt	string	dt=yyyyMMdd 포맷의 문자열 (파티션 키)

수집 실습 – 일일 구매 테이블 스냅샷

Daily Purchase Snapshot

지표정의 : 지정한 일자의 매출 정보

지표산식 : 단순 테이블 수집 데이터 제공

입력형태 : 원본 소스의 접속 정보는 jdbc://mysql://mysql:3306/testdb, 접속 계정과 패스워드는 sqoop, sqoop 이며, 테이블 이름은 purchase 입니다

출력형태 : 저장 타깃의 포맷은 parquet 이며, 서버의 로컬 경로 /tmp/target/purchase/yyyyMMdd 경로에 저장하며, 아래와 같이 설계합니다

컬럼 명	컬럼 타입	설명
p_time	string	구매 시간
p_uid	integer	구매 고객의 아이디
p_id	integer	구매 아이템 아이디
p_name	string	구매 아이템 이름
p_amount	integer	구매 아이템 금액
dt	string	dt=yyyyMMdd 포맷의 문자열 (파티션 키)

수집 실습 – 일일 접속 정보 로그

Daily Access Logs

지표정의 : 지정한 일자의 접속 로그

지표산식 : 단순 파일 수집 데이터 이지만, 중복 로그가 발생할 수 있으므로 최종 테이블 저장시에 중복 제거가 필요합니다

입력형태 : 원본 소스의 포맷은 csv 이며, 접속된 서버의 로컬 경로 /tmp/fluentd/csv/access.csv 파일로 저장됩니다

출력형태 : 수집 데이터 포맷은 json 이며, 서버의 로컬 경로 /tmp/target/access/yyyyMMdd 경로에 저장하며, 아래와 같이 설계합니다

컬럼 명	컬럼 타입	설명
a_id	string	접속로그 유형 (login, logout)
a_tag	string	태그
a_time	string	실제 접속 시간
a_timestamp	string	데이터 수신 시간
a_uid	integer	접속 고객 아이디
dt		dt=yyyyMMdd 포맷의 문자열

기본 지표 (DAU, DR, PU, ARPU, ARPPU)

지표 실습 – 일일 활성 유저

Daily Active User

지표정의 : 지정한 일자의 접속한 유저 수

지표산식 : 지정한 일자의 접속 테이블에 로그(로그인 혹은 로그아웃)가 한 번 이상 발생한 이용자의 빈도수

입력형태 : access 테이블

출력형태 : number

지표 실습 – 일일 구매 유저

Daily Paying User

지표정의 : 지정한 일자의 구매 유저 수

지표산식 : 지정한 일자의 구매 테이블에 한 번이라도 구매가 발생한 이용자의 빈도수

입력형태 : purchase 테이블

출력형태 : number

지표 실습 – 일일 매출

Daily Revenue

지표정의 : 지정한 일자에 발생한 총 매출 금액

지표산식 : 지정한 일자의 구매 테이블에 저장된 전체 매출 금액의 합

입력형태 : access 테이블

출력형태 : number

지표 실습 – 유저 당 평균 매출

Average Revenue Per User

지표정의 : 유저 당 평균 발생 매출 금액

지표산식 : 총 매출 / 전체 유저 수 = DR / DAU

입력형태 : Daily Revenue, Daily Active User

출력형태 : number

지표 실습 – 구매 유저 당 평균 매출

Average Revenue Per Paying User

지표정의 : 유저 당 평균 발생 매출 금액

지표산식 : 총 매출 / 전체 유저 수 = DR / DPU

입력형태 : Daily Revenue, Daily Paying User

출력형태 : number

디멘젼 테이블 (유저)

지표 실습 – 유저 디멘젼

Daily User Dimension

지표정의 : 이용자 누적 상태 정보

지표산식 : 오늘까지 접속한 모든 유저의 정보를 저장하는 테이블

입력형태 : user, purchase, access

출력형태 : 아래와 같이 설계합니다

컬럼 명	컬럼 타입	설명
d_uid	integer	유저 아이디
d_name	string	유저 이름
d_pamount	integer	누적 구매 금액
d_pcount	integer	누적 구매 횟수
d_acount	integer	누적 접속 횟수
dt	string	dt=yyyyMMdd 포맷의 문자열

고급 지표 (NU, FPU)

지표 실습 – 신규 유저

Daily New User

지표정의 : 처음 가입한 고객의 수

지표산식 : 오늘 접속한 유저 중에, 어제 까지 한 번도 접속한 이력이 없는 이용자 빈도 (가입 로그가 별도로 없다는 가정)

입력형태 : Daily Dimension, access

출력형태 : number

지표 실습 – 최초 구매 유저

Daily First-time Paying User

지표정의 : 처음 구매한 고객의 수

지표산식 : 오늘 구매한 유저 중에, 어제 까지 한 번도 구매한 이력이 없는 이용자 빈도

입력형태 : Daily Dimension, purchase

출력형태 : number

지표 실습 – 누적 매출

Accumulated Revenue

지표정의 : 누적 매출 금액

지표산식 : 구매한 이력이 있는 이용자의 누적 구매 금액의 누적 합 (혹은 횟수)

입력형태 : Daily Dimension

출력형태 : number

지표 실습 – 평균 접속 시간

Daily Average Access Time

지표정의 : 유저의 평균 접속 시간

지표산식 : 로그아웃 시간 - 로그인 시간 / 오늘 접속 횟수

입력형태 : access

출력형태 : number

https://github.com/psyoblade/data-engineer-basic-training/tree/master/day5