

명품 브랜드의 인스타그램 화제성 분석 프로젝트

2팀 1조

이성희
이하운
임형우

프로젝트 개요







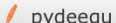











- 명품 브랜드의 인스타그램 계정과 관련 해시태그 정보를 수집하여 이를 분석하여 각 브랜드의 화제성을 평가합니다.

주제 선정 이유

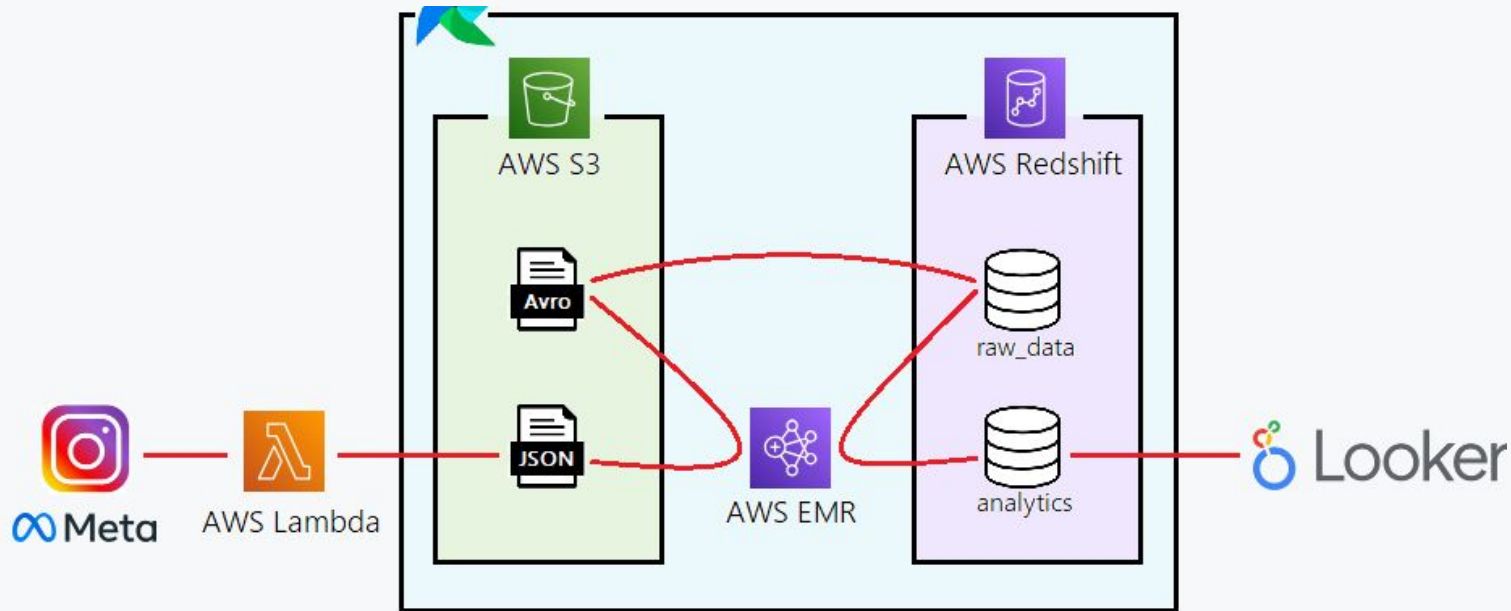
- 명품 브랜드의 인스타그램 계정과 관련 해시태그로 미디어를 수집하여 브랜드의 활동성과 인지도를 측정하고, 각 브랜드의 고유한 온라인 마케팅 전략을 파악합니다.
- 이를 통해 브랜드 간의 차별화를 확인하고 경쟁 우위를 평가하며, 소비자 동향을 파악하는 중요한 정보를 얻을 수 있습니다.

참여자 및 각 역할

업무	이성희	이하윤	임형우
기획	데이터 모델링	- 데이터 모델링 - 분석지표 정의	데이터 모델링
Infra	- AWS 네트워크 관리 - 데이터/컨테이너 환경 구축 - Airflow CI/CD	-	- Lambda 환경 관리 - 스크래퍼 CI/CD
Scrapping	-	-	- Instagram API 스크래퍼 - Lambda Event 스케줄링
ETL	- Airflow ETL 스케줄링 - EMR 프로세스 개발 - Slack 모듈 개발	-	-
ELT	-	- 데이터 마트 모델링 - 마트 쿼리 작성 - Airflow ELT 스케줄링 - EMR 프로세스 개발	- 데이터 마트 모델링 - 마트 쿼리 작성
Visualization	-	Looker 대시보드	Looker 대시보드

Field	Stack
Infra	 AWS Secrets Manager  AWS Cloudwatch  AWS EC2
Scrapping	 AWS Lambda
ETL & ELT	 Airflow  Spark  pydeequ
Data Storage	 AWS S3  AWS Redshift  Snowflake  Amazon RDS
BI tool	 Looker
CI/CD	 Docker  Github Actions  AWS ECR  AWS ECS
ETC	 Slack  Notion

데이터 분석 아키텍처



<RAW_DATA Schema>

Overwrite -> Blue
Append -> Red
Upsert -> Purple

media				
media_id	VARCHAR(30)	NOT NULL	Default value	미디어 아이디 코드
user_id	VARCHAR(30)	NOT NULL	Default value	유저 아이디 코드
like_count	INTEGER	NOT NULL	0	좋아요 수
comments_count	INTEGER	NOT NULL	0	댓글 수
media_product_type	VARCHAR(10)	NOT NULL	Default value	미디어가 게시된 위치
media_type	VARCHAR(20)	NOT NULL	Default value	미디어의 유형
caption	VARCHAR(max)	NULL	Default value	미디어의 캡션 텍스트
permalink	VARCHAR(100)	NOT NULL	Default value	미디어의 영구 URL
media_url	VARCHAR(1000)	NOT NULL	Default value	미디어의 url
ts	TIMESTAMP	NOT NULL	Default value	미디어 게시 날짜
del_yn	CHAR(1)	NOT NULL	'Y'	미디어 삭제 여부
created_at	TIMESTAMP	NOT NULL	SYSDATE	적재된 시각
updated_at	TIMESTAMP	NOT NULL	Default value	update된 시각

media_hashtag				
user_id	VARCHAR(30)	NOT NULL	Default value	유저 아이디 코드
media_id	VARCHAR(30)	NOT NULL	Default value	미디어 아이디
media_type	VARCHAR(20)	NOT NULL	Default value	미디어 타입
caption	VARCHAR(3000)	NULL	Default value	미디어 글
comments_count	INTEGER	NOT NULL	0	댓글 수
like_count	INTEGER	NOT NULL	0	좋아요 수
ts	TIMESTAMP	NOT NULL	Default value	미디어 생성시간
created_at	TIMESTAMP	NOT NULL	SYSDATE	DB적재시간

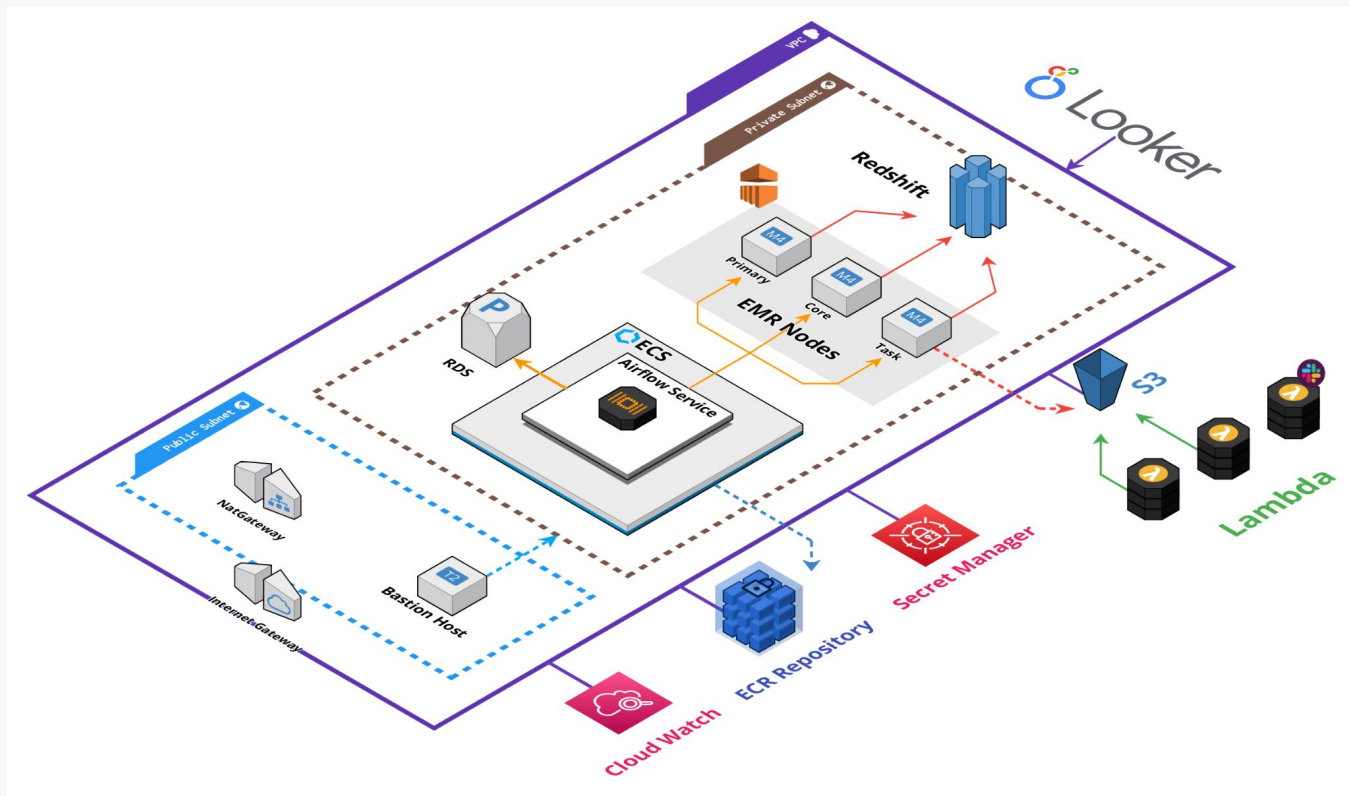
brand				
user_id	VARCHAR(30)	NOT NULL	Default value	유저 아이디 코드
username	VARCHAR(50)	NOT NULL	Default value	유저 이름 (검색id)
tagname	VARCHAR(20)	NOT NULL	Default value	태그 이름
name	VARCHAR(100)	NULL	Default value	유저 이름
profile_picture_url	VARCHAR(1000)	NULL	Default value	프로필 사진 이미지 url
followers_count	INTEGER	NOT NULL	0	팔로워 수
media_count	INTEGER	NOT NULL	0	미디어 수
del_yn	CHAR(1)	NOT NULL	'Y'	계정삭제여부
updated_at	TIMESTAMP	NOT NULL	Default value	update된 시각

media_log				
user_id	VARCHAR(30)	NOT NULL	Default value	유저 아이디 코드
like_count	INTEGER	NOT NULL	0	좋아요 수
comments_count	INTEGER	NOT NULL	0	댓글 수
created_at	TIMESTAMP	NOT NULL	SYSDATE	적재된 시각

brand_log				
user_id	VARCHAR(30)	NOT NULL	Default value	유저 아이디 코드
followers_count	INTEGER	NOT NULL	0	팔로워 수
media_count	INTEGER	NOT NULL	0	미디어 수
created_at	TIMESTAMP	NOT NULL	SYSDATE	적재된 시각

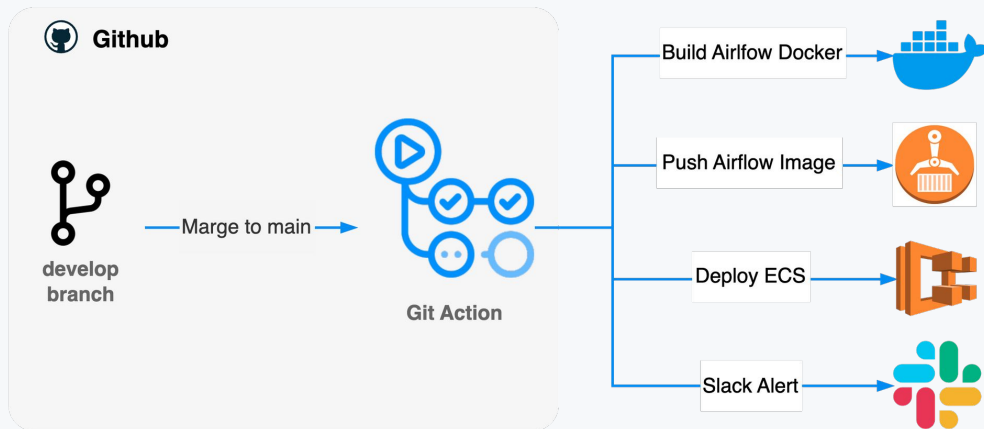
AWS 인프라 아키텍처

S3	환경변수 및 EMR 저장소, 데이터 레이크
Lambda	데이터 수집 프로세스 실행 및 스케줄링
Secret Manger	토큰, Connection 등 민감 정보 관리
EMR	ETL, ELT 작업을 처리하는 클러스터 - m4.large Spot, 최소 노드 2 (Primary, Core)
Redshift	Raw Data, Mart Data가 적재된 데이터 웨어하우스 - dc2.large, 2노드
ECR	Docker 이미지 저장소
ECS	Airflow 컨테이너를 관리하는 Serverless 오케스트레이션 - fargate Spot, 4개 컨테이너
RDS	Airflow 서비스의 메타 데이터 저장소

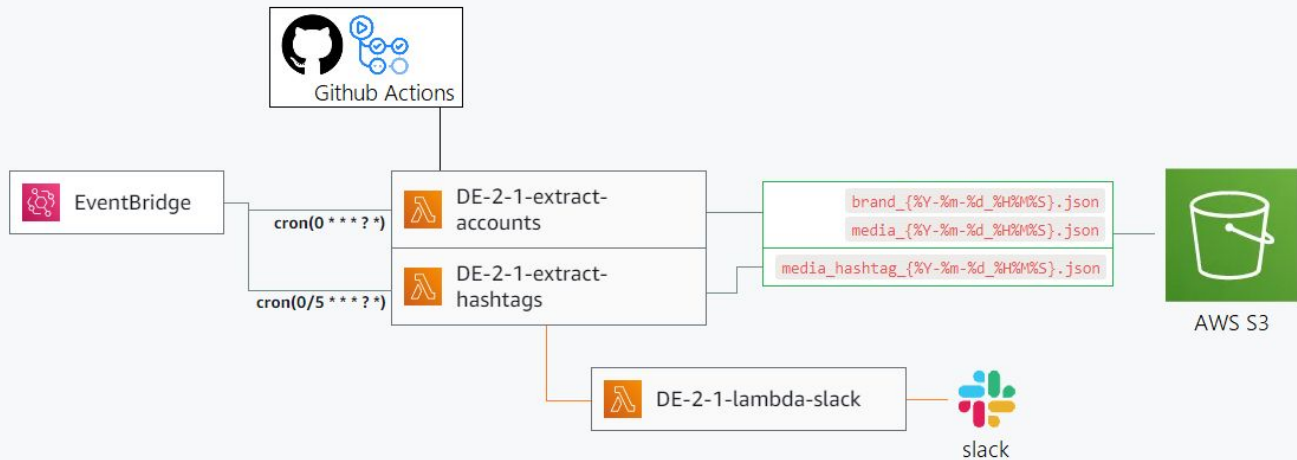


Airflow 배포 프로세스

- Github Action을 이용하여 Docker 이미지를 생성해 ECR에 저장
- Airflow 빌드 정보가 담긴 task-definition의 이미지 버전정보 업데이트
- ECS를 통해 Airflow 서비스 배포 (롤링 업데이트)
- 컨테이너별 Health Check 후 정상이 아닐 경우 재배포
- 배포 결과 Slack으로 알람



데이터 수집



데이터 수집 및 저장

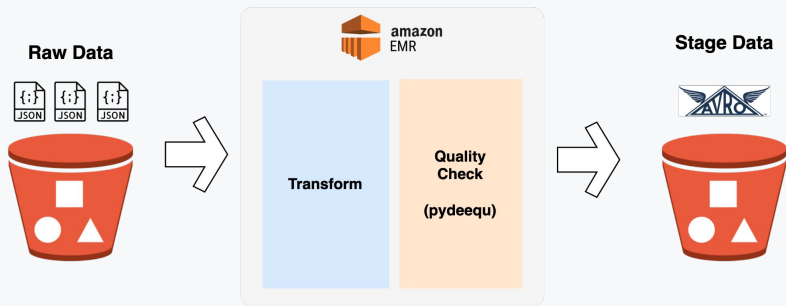
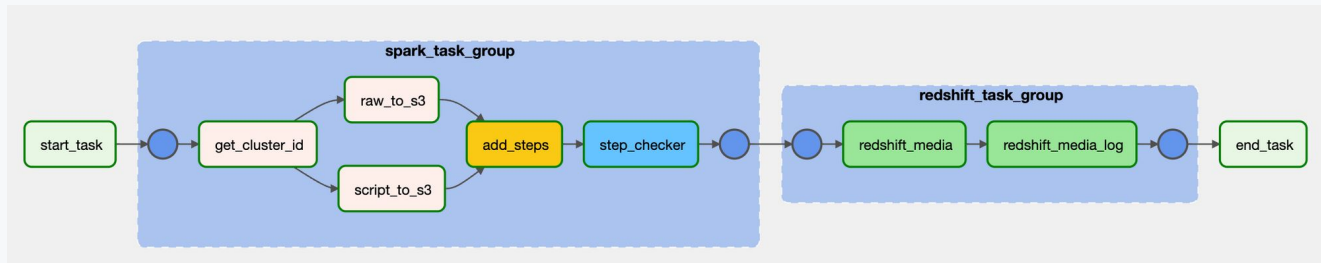
- Instagram Graph API를 통한 데이터 수집기 개발
- API 응답 데이터 검증 및 필드 전처리 후 JSON 포맷으로 S3 적재
- Github Actions를 이용하여 CI/CD 및 AWS Lambda 배포

데이터 품질 점검 및 로깅

- 수집 대상, 맵핑 필드, 저장 포맷 등의 메타데이터를 **configure.ini** 파일로 분리하여 관리
- 일부 데이터 문제로 프로세스 자체가 종료되지 않도록 느슨한 에러 처리
- 알람의 경우 성공/실패 여부보단 수집된 데이터에 대한 품질사항에 초점

ETL 프로세스

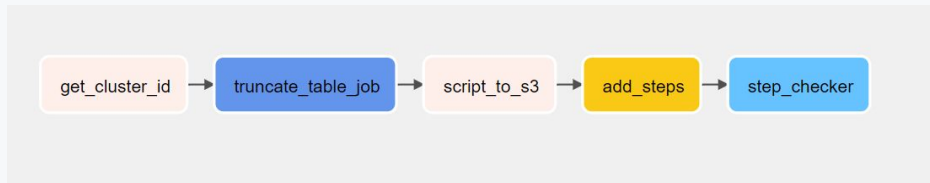
→ AWS EMR을 이용한 Raw Data 가공, 검증 및 Redshift 적재



- Raw Data와 EMR 클러스터가 수행할 스크립트 정보를 S3에 적재
- **pydeequ** 라이브러리를 이용해 데이터 품질체크 수행
- 처리 결과 S3 /Stage 버킷에 Avro 타입으로 저장 (Avro : 자동매핑 편의성 제공)
- Sensor를 통해 EMR 작업의 결과 확인

ELT 프로세스

→ AWS EMR을 이용한 Raw Data 가공, 검증 및 Redshift 적재



- 품질체크 쿼리를 통해 Raw 데이터를 검증 후 로깅 및 실패시 Dag 종료
- 마트 테이블 모두 Overwrite 적재 수행
- Cross Filtering 차트를 위한 Fact 테이블에 데이터 적재
- 데이터 집계 연산 후 집계(Agg) 테이블에 적재
- 분석 지표에 맞춰 데이터 가공 후 Fact 테이블에 각각 적재
 - a. 인스타그램 미디어 Caption 데이터에서 해시태그 키워드 추출
 - b. Timestamp 데이터를 시간, 요일 데이터로 각각 변환
 - c. 인기도 지표에 필요한 3가지 지수를 계산 후 랭킹 처리

데이터 마트 모델링

Fact -> Green
Agg -> Purple

brand_information

tag_name	VARCHAR(20)	NOT NULL	Default value	태그 이름 (브랜딩 고유명사)
user_id	VARCHAR(30)	NOT NULL	Default value	유저 ID 코드(브랜딩 ID 코드)
media_id	VARCHAR(30)	NOT NULL	Default value	미디어 아이디 코드
like_count	INTEGER	NOT NULL	0	좋아요 수
comments_count	INTEGER	NOT NULL	0	댓글 수
ts	TIMESTAMP	NOT NULL	Default value	미디어 게시 시간
media_count	INTEGER	NOT NULL	0	미디어 수
followers_count	INTEGER	NOT NULL	0	팔로워 수
media_type	VARCHAR(20)	NOT NULL	Default value	미디어의 유형
media_product_type	VARCHAR(10)	NOT NULL	Default value	미디어가 게시된 위치
updated_at	TIMESTAMP	NOT NULL	Default value	업데이트 시간
engagement	INTEGER	NOT NULL	Default value	좋아요 수 + 댓글 수
engagement_rate	FLOAT	NOT NULL	Default value	참여율

followers_growth

tag_name	VARCHAR(20)	NOT NULL	Default value	태그 이름 (브랜딩 고유명사)
created_at	TIMESTAMP	NOT NULL	Default value	DB 적재된 시간
followers_count	INTEGER	NOT NULL	0	팔로워 수

hashtag_search

created_at	TIMESTAMP	NOT NULL	Default value	DB 적재된 시간
tag_name	VARCHAR(20)	NOT NULL	Default value	태그 이름 (브랜딩 고유명사)
user_id	VARCHAR(30)	NOT NULL	Default value	유저 ID 코드(브랜딩 ID 코드)
ts	TIMESTAMP	NOT NULL	Default value	미디어 생성시간
media_id	VARCHAR(30)	NOT NULL	Default value	미디어 아이디 코드
caption	VARCHAR(5000)	NOT NULL	**	미디어의 캡션 텍스트

trending_topics

created_at	TIMESTAMP	NOT NULL	Default value	DB 적재된 시간
tag_name	VARCHAR(20)	NOT NULL	Default value	태그 이름 (브랜딩 고유명사)
ts	TIMESTAMP	NOT NULL	Default value	미디어 생성시간
related_hashtag	VARCHAR(100)	NOT NULL	Default value	연관 해시태그
pos	INTEGER	NOT NULL	Default value	합수 사용시 필수인자

brand_basic_info

tag_name	VARCHAR(20)	NOT NULL	Default value	태그 이름 (브랜딩 고유명사)
user_id	VARCHAR(30)	NOT NULL	Default value	유저 ID 코드(브랜딩 ID 코드)
profile_picture_url	VARCHAR(1000)	NULL	**	프로필 사진 이미지 url
followers_count	INTEGER	NOT NULL	0	팔로워 수

aggregated_brand_information

tag_name	VARCHAR(20)	NOT NULL	Default value	태그 이름 (브랜딩 고유명사)
avg_engagement_rate	FLOAT	NOT NULL	Default value	평균 참여율
brand_media_cnt	INTEGER	NOT NULL	Default value	업로드한 미디어 총 개수

brand_media_post_time

tag_name	VARCHAR(20)	NOT NULL	Default value	태그 이름 (브랜딩 고유명사)
ts	TIMESTAMP	NOT NULL	Default value	미디어 게시 시간
post_time	INTEGER	NOT NULL	Default value	미디어 게시 시간
day_of_week	INTEGER	NOT NULL	Default value	미디어 게시 요일 정수값(1~7)
post_day_of_week	VARCHAR(15)	NOT NULL	Default value	미디어 게시 요일

aggregated_hashtag_search

tag_name	VARCHAR(20)	NOT NULL	Default value	태그 이름 (브랜딩 고유명사)
hashtagged_media_cnt	INT	NOT NULL	Default value	해시태그 검색에서의 미디어 총 개수

popularity_factor

tag_name	VARCHAR(20)	NOT NULL	Default value	태그 이름 (브랜딩 고유명사)
avg_engagement_rate	FLOAT	NOT NULL	Default value	평균 참여율
hashtagged_media_cnt	INTEGER	NOT NULL	Default value	해시태그 검색에서의 미디어 총 개수
followers_count	INTEGER	NOT NULL	Default value	팔로워 수
rank_avg_engagement_rate	INTEGER	NOT NULL	Default value	참여율 지표 행렬 지수
rank_followers_count	INTEGER	NOT NULL	Default value	미디어 지표 행렬 지수
rank_hashtagged_media_cnt	INTEGER	NOT NULL	Default value	커뮤니티 지표 행렬 지수

popularity_calculation

tag_name	VARCHAR(20)	NOT NULL	Default value	태그 이름 (브랜딩 고유명사)
sum_rank	INTEGER	NOT NULL	Default value	행렬지수 합산 - 넣을수록 좋음
popularity_rank	INTEGER	NOT NULL	Default value	인기도 (랭크 최음)

hashtag_count

related_hashtag	VARCHAR(100)	NOT NULL	Default value	연관 해시태그
hashtag_frequency	INTEGER	NOT NULL	Default value	빈도

popularity_factor_early_stage

tag_name	VARCHAR(20)	NOT NULL	Default value	태그 이름 (브랜딩 고유명사)
avg_engagement_rate	FLOAT	NOT NULL	Default value	평균 참여율
hashtagged_media_cnt	INTEGER	NOT NULL	Default value	해시태그 검색에서의 미디어 총 개수

<ANALYTICS
Schema>

대시보드

분석 지표 설정

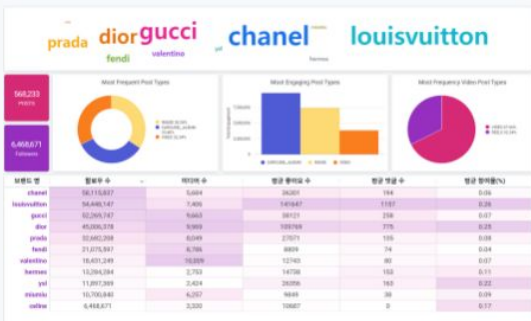
- 참여율(%) = (댓글 수 + 좋아요 수) / (팔로워 수) * 100
- 인기도 = (미디어 지수 + 팔로워 지수 + 커뮤니티 지수) 지표 합산 랭킹
 - 인기도 지표를 정의해서 명품브랜드 11개의 실시간 변동 순위 제시
 - 미디어 지수 : 브랜드 인스타그램 미디어의 평균 참여율에 따라 랭킹 부여
 - 팔로워 지수 : 브랜드 인스타그램 팔로워수에 따라 랭킹 부여
 - 커뮤니티 지수 : 브랜드 이름이 해시태그된 미디어의 개수에 따라 랭킹 부여

대시보드 생성

- 종합 대시보드
- 반응형 대시보드

Luxury Brand Analysis Project on Instagram

Instagram Graph API를 통해 데이터를 수집하여 분석한다



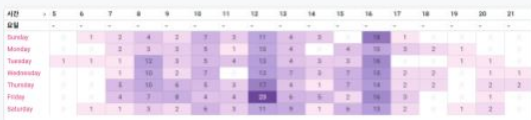
Post Engagement Average Rate



Brand Post Upload Activity

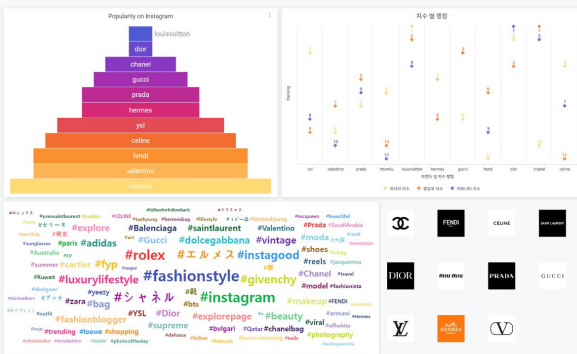


Instagram Post Upload Time Distribution



Luxury Brand Analysis Project on Instagram

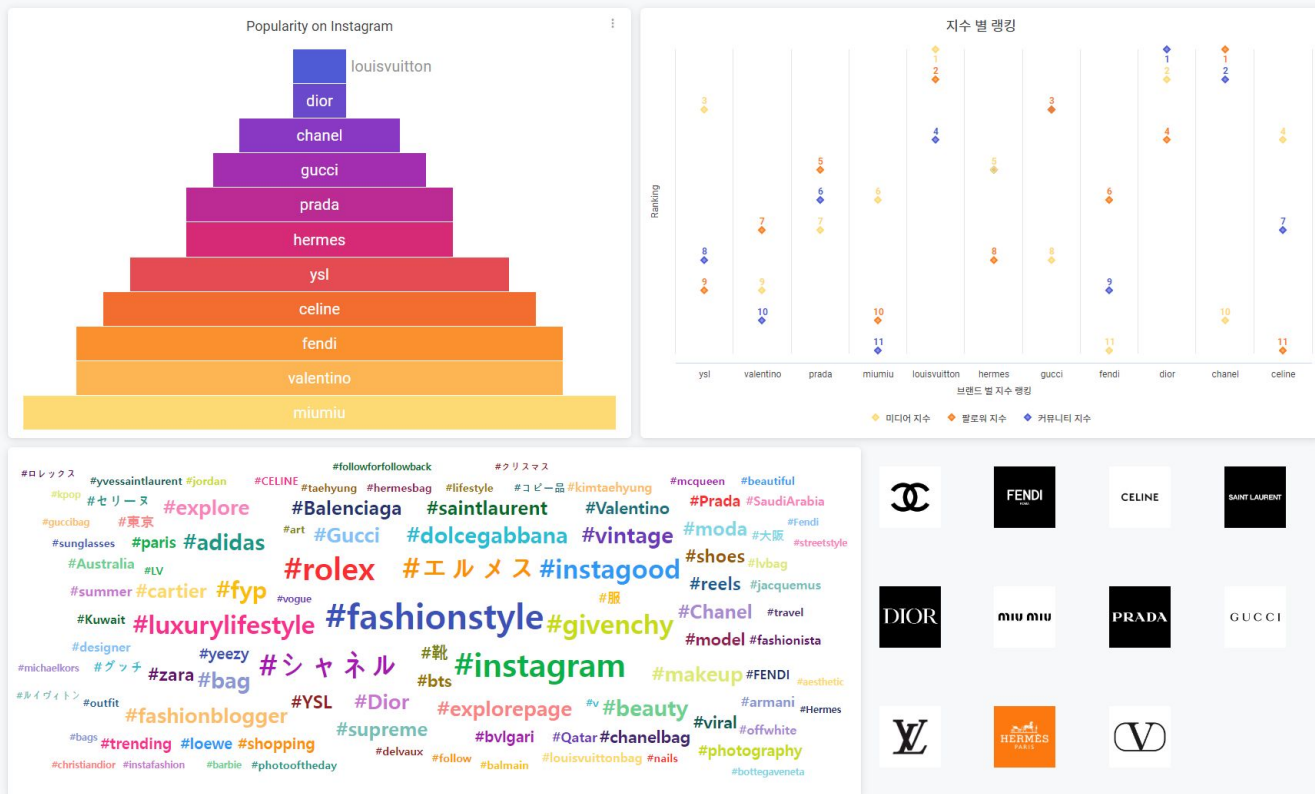
이해 군 프로젝트를 위하여 작성한 보고서. 첫 번째 장은 '인기도' 지표에 대한 설명이다. Instagram Graph API를 통해 인스타그램의 브랜드를 분석하고 있다.



종합 대시보드

Luxury Brand Analysis Project on Instagram

이제 곧 취준생을 벗어나 직장인이 될텐데, 첫 명품 입문을 어느 브랜드로 하면 좋을까? Instagram Graph API를 통해 인스타그램 내 명품 브랜드의 화제성 분석 프로젝트!



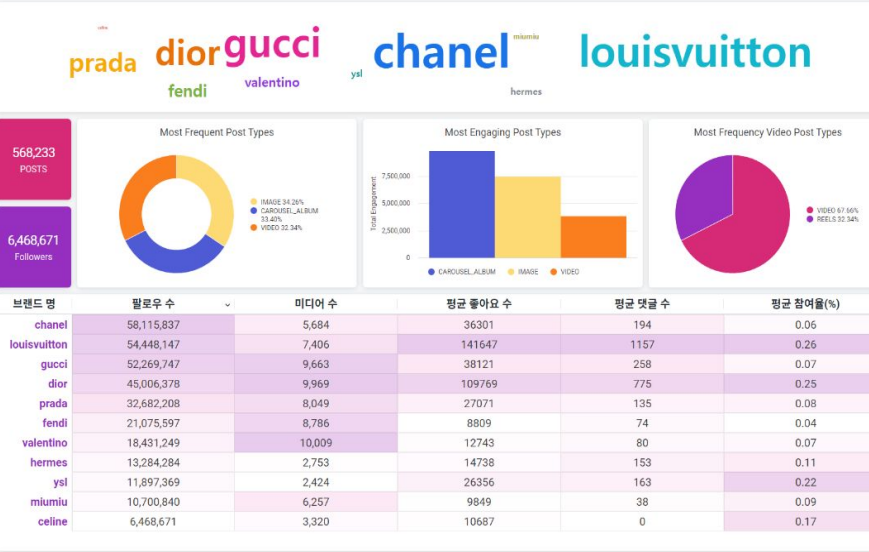
반응형 대시보드

Detailed Analysis by Luxury Brands

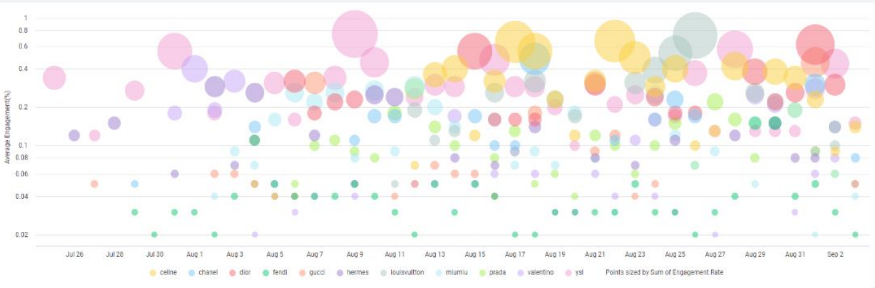
100% 완료도 남음 0 100% 완료도 선택 100% 완료도 선택 Today 8 100% 완료도 선택 100% 완료도 선택

Luxury Brand Analysis Project on Instagram

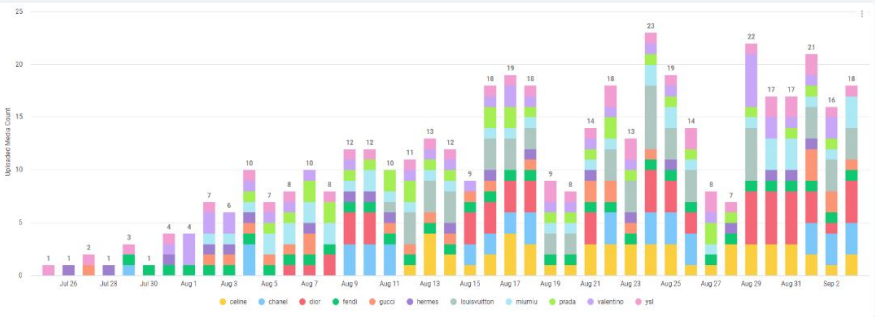
Instagram Graph API를 통해 영문 브랜드의 상세 분석



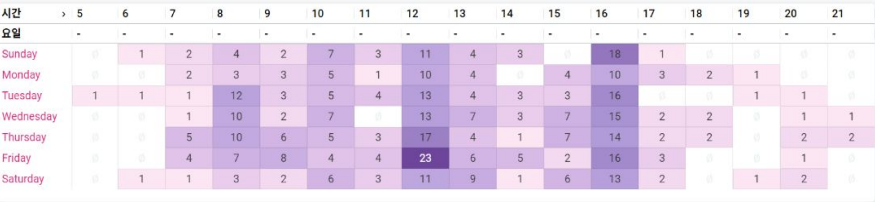
Post Engagement Average Rate



Brand Post Upload Activity



Instagram Post Upload Time Distribution



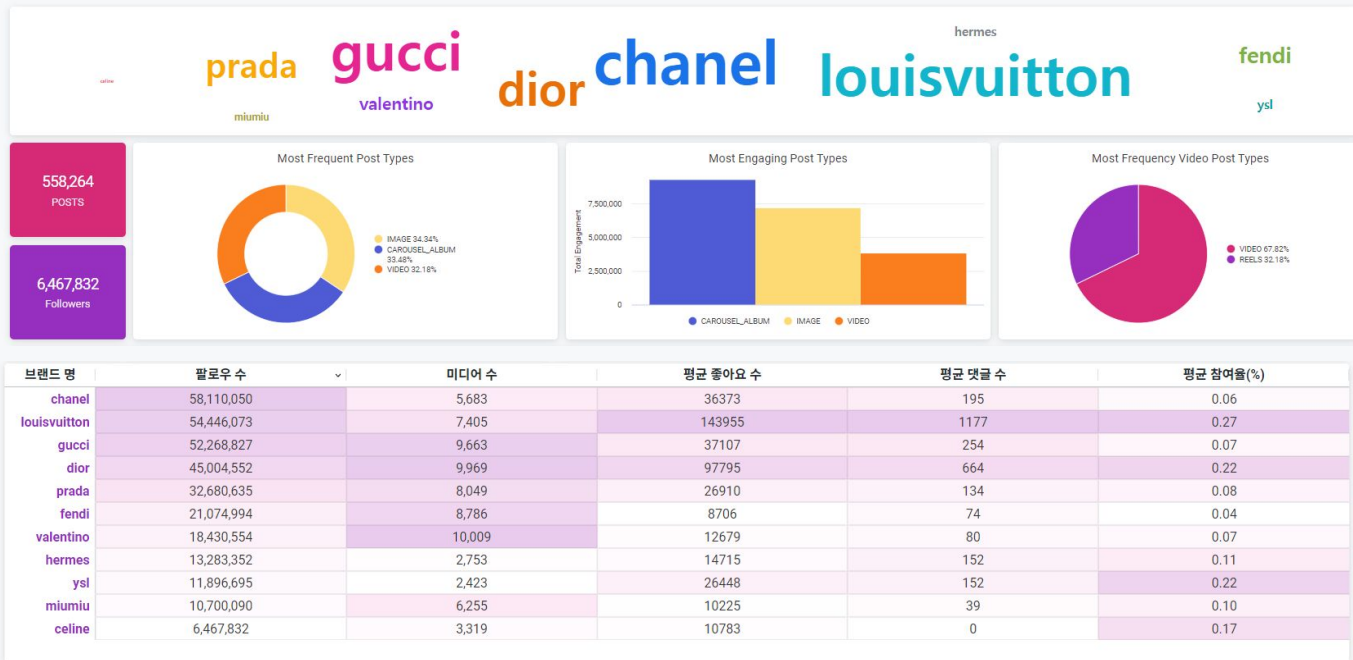
반응형 대시보드

Detailed Analysis by Luxury Brands ♥

미디어 업로드 날짜 + 미디어 업로드 시간대 + 미디어 업로드 요일
Last 30 Days 0 24 is Monday or Friday or Tuesday or Wednesday...

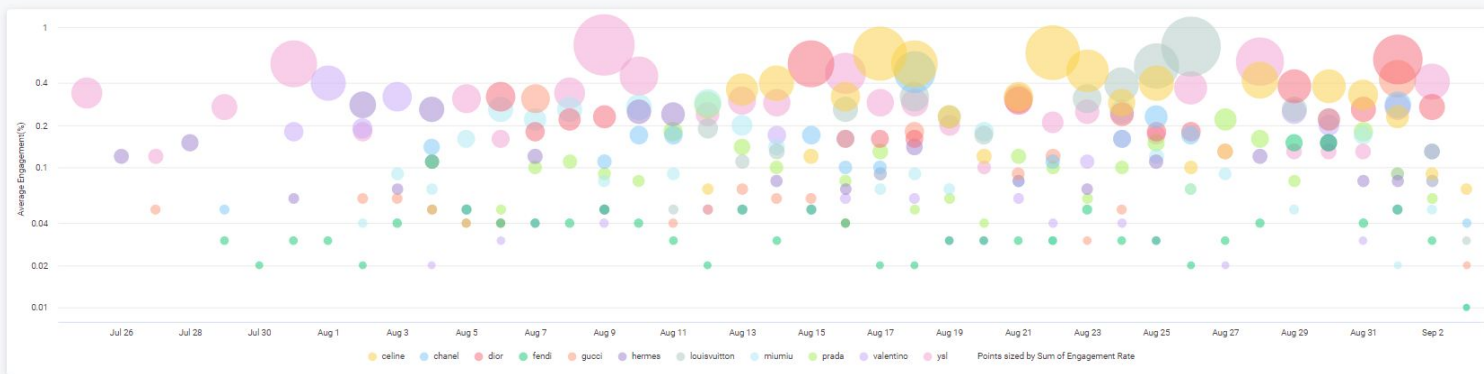
Luxury Brand Analysis Project on Instagram

Instagram Graph API를 통해 명품 브랜드의 상세 분석

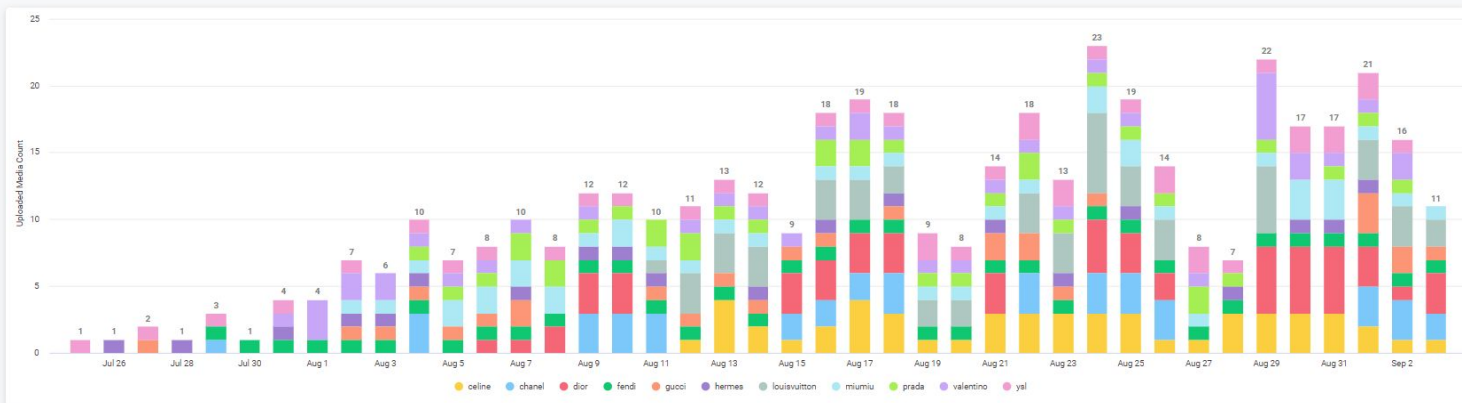


반응형 대시보드

Post Engagement Average Rate



Brand Post Upload Activity



반응형 대시보드

Instagram Post Upload Time Distribution

시간	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
요일	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sunday	0	1	2	4	2	7	3	11	3	2	0	14	0	0	0	0	0
Monday	0	0	2	3	3	5	1	10	4	0	4	10	3	2	1	0	0
Tuesday	1	1	1	12	3	5	4	13	4	3	3	16	0	0	1	1	0
Wednesday	0	0	1	10	2	7	0	13	7	3	7	15	2	2	0	1	1
Thursday	0	0	5	10	6	5	3	17	4	1	7	14	2	2	0	2	2
Friday	0	0	4	7	8	4	4	23	6	5	2	16	3	0	0	1	0
Saturday	0	1	1	3	2	6	3	11	9	1	6	13	2	0	1	2	0

<인터랙티브 차트 시연 영상>

