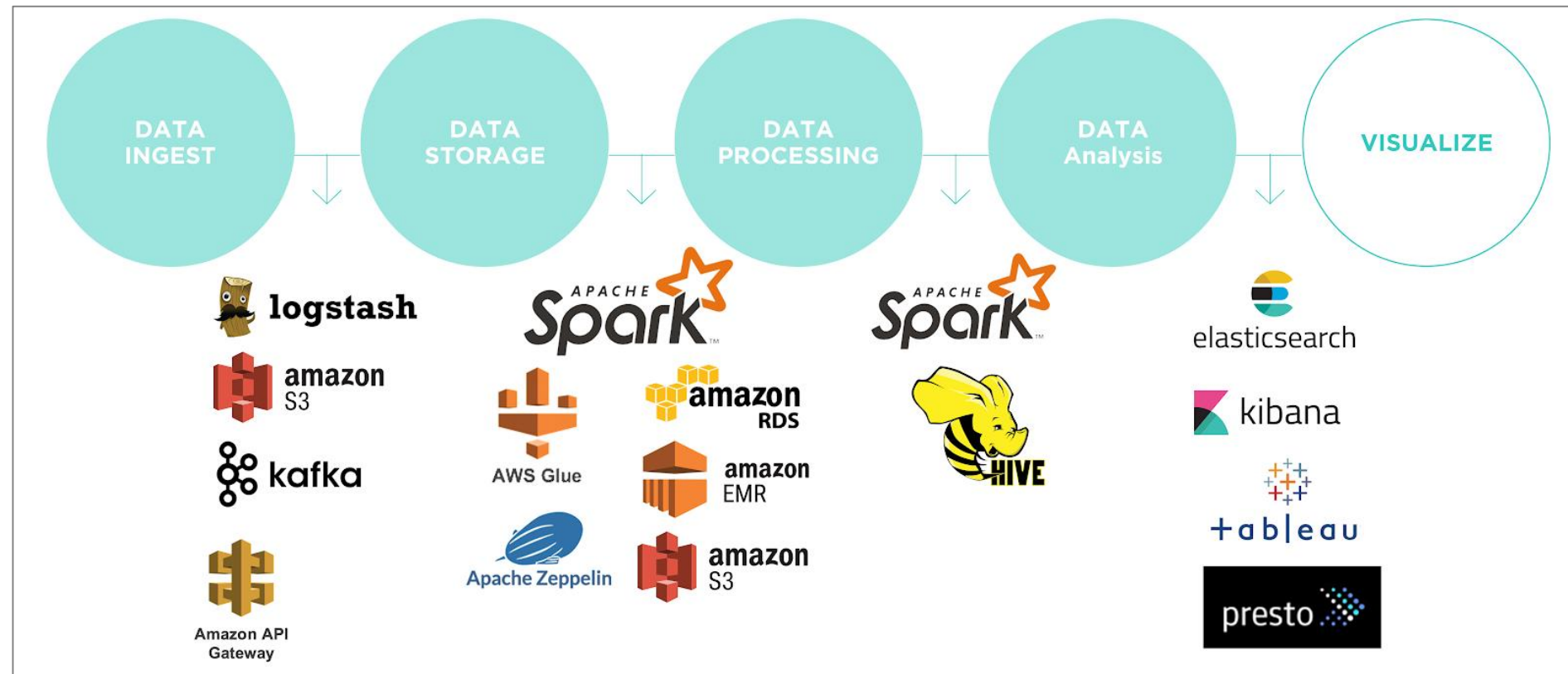


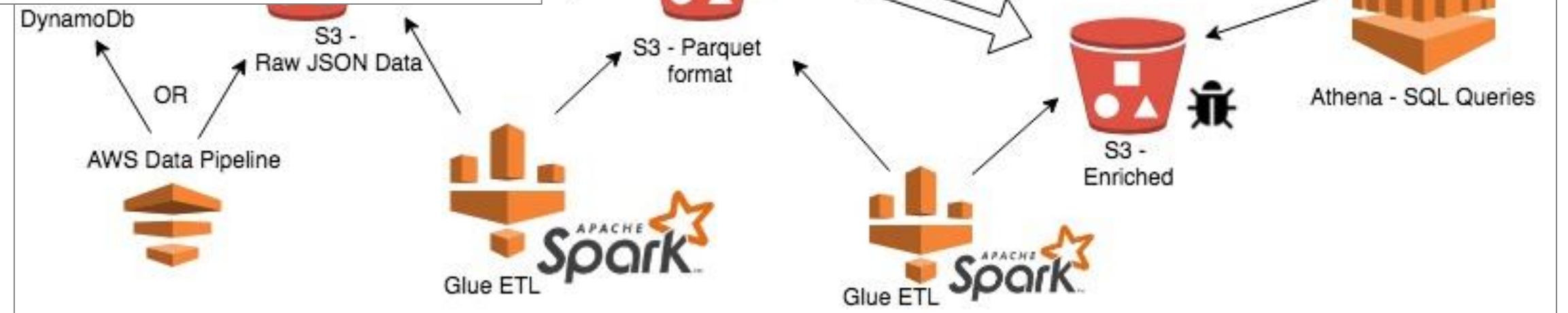
그로스해킹 101

5-1. 분석 환경 구축

데이터 파이프라인



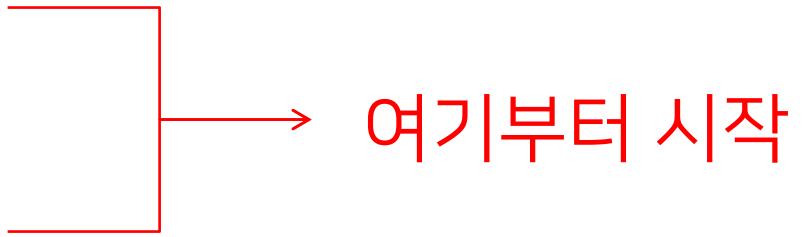
데이터 파이프라인이요? 이런 거?



데이터 파이프라인 만들기

= 데이터 엔지니어 + 데이터 사용자의 협업이 필요

기술 stack 선택은 개발자의 몫이지만, 데이터 사용자의 의견이 반영되어야 하는 것들이 훨씬 많음

- 어떤 데이터를 쌓을 것인가?
 - 어떤 형태로 쌓을 것인가?
 - 어디에 쌓을 것인가?
 - 어떻게 꺼내서 볼 것인가?
- 
- 여기부터 시작

데이터 파이프라인의 출발점

- 어떤 데이터를 쌓을 것인가?
 - 어떤 형태로 쌓을 것인가?
- 의외로(?) 많은 회사들이 여기에서부터 막힌다...



들어오시면 데이터부터
다시 쌓으셔야 해요.

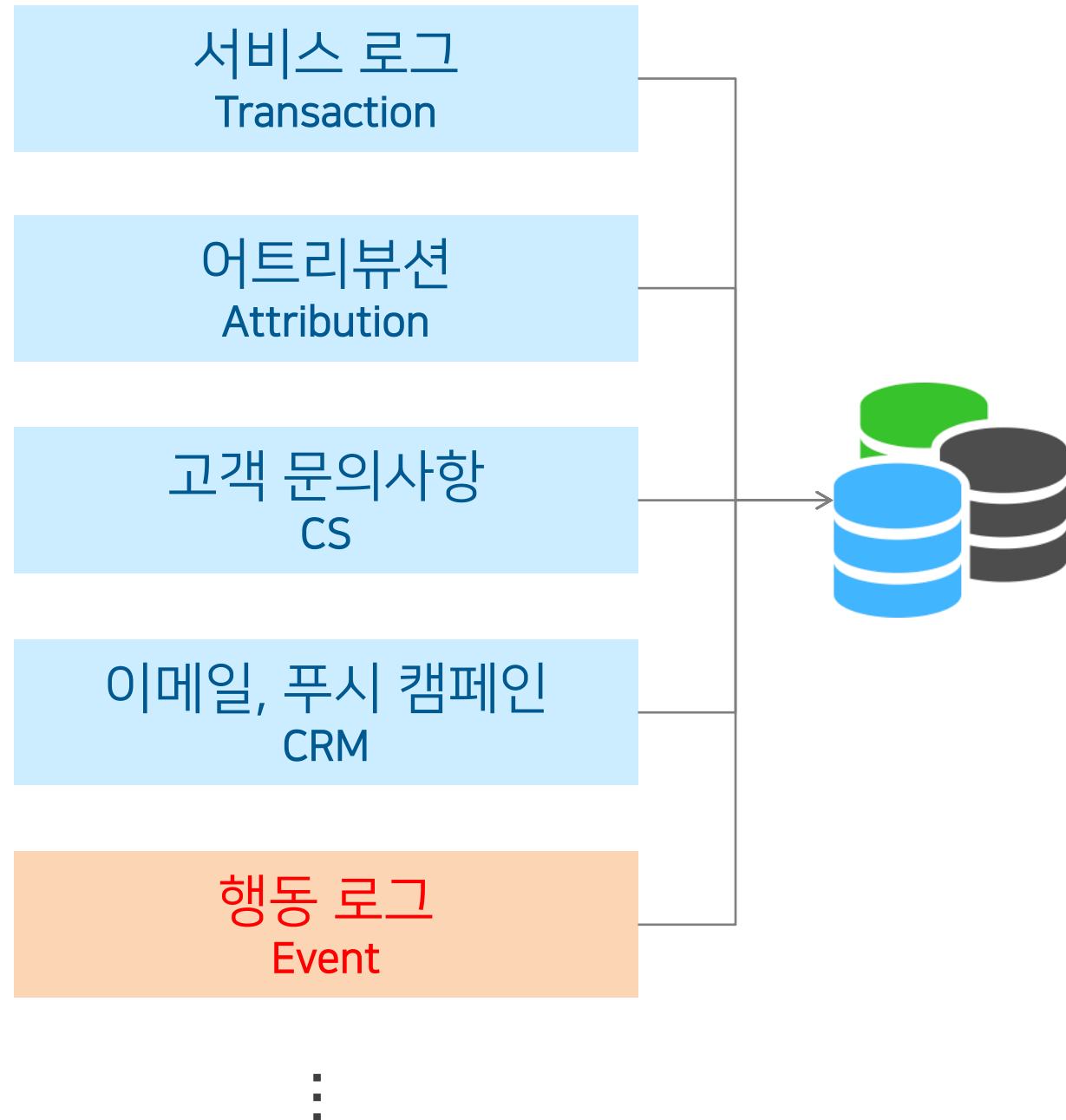
데이터분석가 면접에서 많이 듣는 말 1위



분석하려고 봤더니 쓸 수
있는 데이터가 없어요.

데이터분석가가 입사하고 많이 하는 말 1위

데이터 소스



- 행동 로그 (Event log)

- 사용자들이 서비스 내에서 어떤 action을 했는지 기록한 데이터
- 클릭했다, 검색했다, 스와이프했다...
- 다른 데이터와 연계해서 분석하는 경우 굉장히 풍부한 인사이트를 얻을 수 있음
- 하지만 잘 관리되지 않는 곳이 많음
 - 일반적으로 데이터 구조가 복잡하고 양이 많은 편
 - 없다고 해서 서비스에 큰 문제가 생기는 건 아님;;;

- 행동 로그 설계하기

- 어디에 쌓을 것인가... 에 따라서 설계 방식이 조금씩 다르지만, 큰 틀은 동일함



속도
안정성
유지보수 효율
가격
연동편의성



Google BigQuery

참고 문서

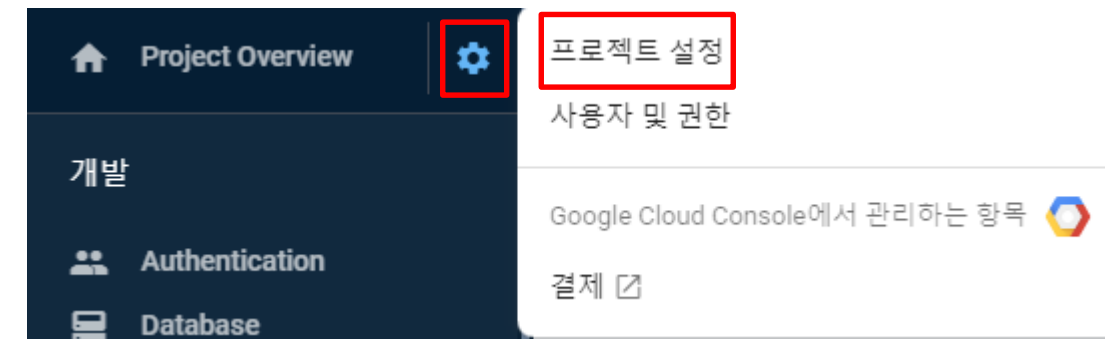
- [구글 BigQuery 공식 문서](#)
- 변성운 님의 [BigQuery 튜토리얼](#)
- 이민우 님의 [BigQuery 시작하기](#)

BigQuery 기본 설정

1. Blaze 요금제 사용

Spark 요금제 아마추어 개발자에게 넉넉한 한도 무료	Flame 요금제 성장하는 앱을 위한 고정 가격 \$25/월	Blaze 요금제 규모별 앱 가격 계산 종량제 <small>체크표시 Spark 요금제의 무료 사용 혜택 포함*</small>
--	---	--

2. Firebase 콘솔 > 설정 > 프로젝트 설정



3. 설정 > 통합 > BigQuery 관리

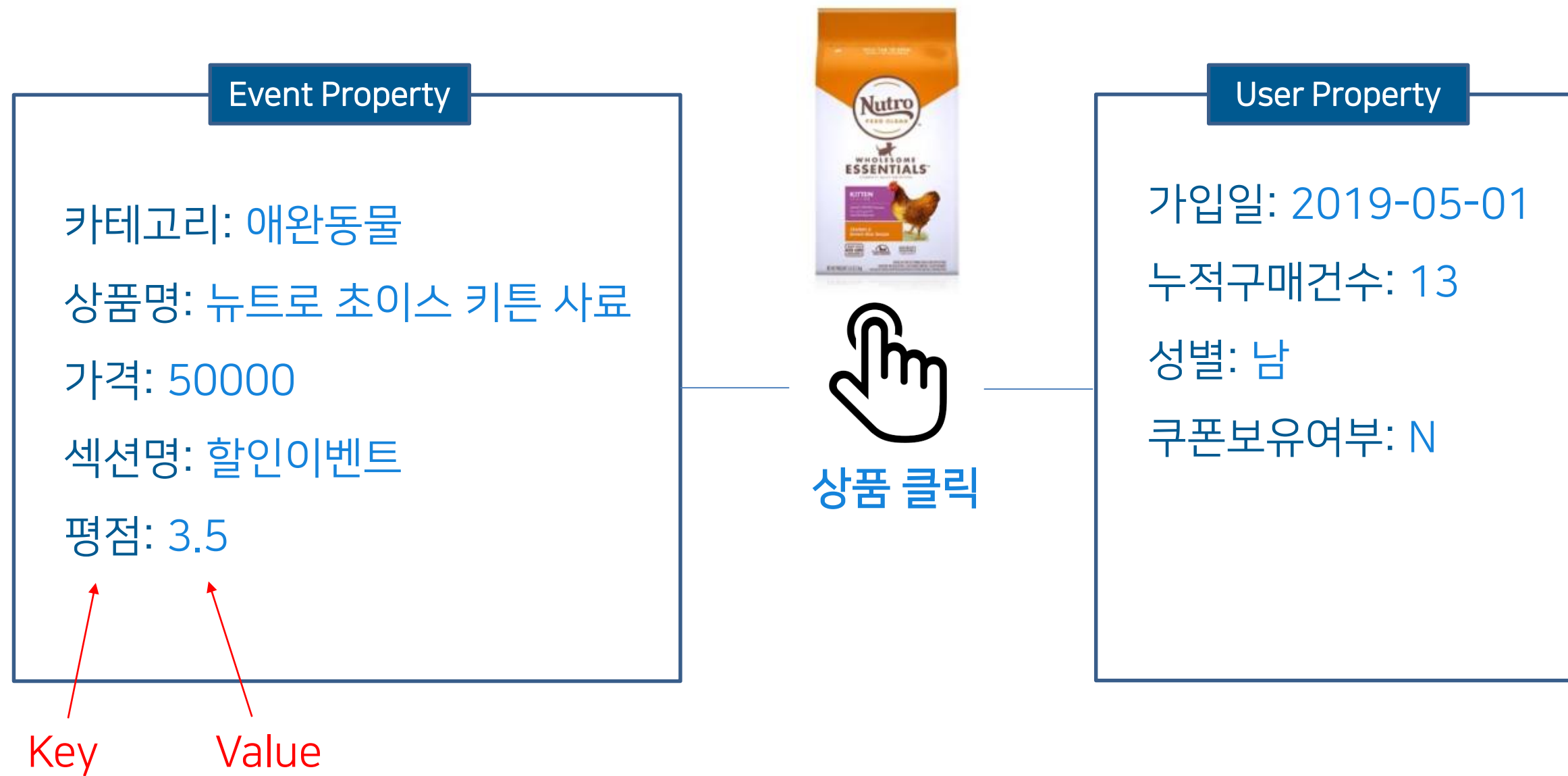
4. 앱 선택

5. 관련 링크 / 기술 문서

- 자동으로 수집하는 이벤트 <https://support.google.com/firebase/answer/6317485?hl=ko>
- 자동으로 수집하는 사용자 속성 <https://support.google.com/firebase/answer/6317486?hl=ko>
- 수집 및 구성 한도 <https://support.google.com/firebase/answer/9237506?hl=ko>
- 이벤트 로깅 <https://firebase.google.com/docs/analytics/cpp/events?hl=ko>

이벤트 설계

- 행동 로그 설계의 핵심은, Property 설계



이벤트 설계

	요약	예시
1단계	단순 이벤트 집계	오늘 하루동안 상품 상세보기 버튼 클릭수는 100입니다.
2단계	이벤트 Property 집계	상품 상세보기 클릭이 100회 발생했는데, 이 중 60번은 애완동물 카테고리에 있는 상품을 클릭했습니다. 할인이벤트 섹션에 있는 상품 클릭이 가장 많았으며, 클릭한 상품의 평균 가격은 7만 5천원입니다. 상품별로는 뉴트로 초이스 키튼 사료 상품의 클릭이 25회로 가장 많았고, 이 상품의 평점은 3.5점 입니다.
3단계	이벤트 Property + 사용자 Property 집계	상품 상세보기 클릭이 100회 발생했는데, 이 중 60번은 애완동물 카테고리에 있는 상품을 클릭했습니다. 할인이벤트 섹션에 있는 상품 클릭이 가장 많았으며, 클릭한 상품의 평균 가격은 7만 5천원입니다. 상품별로는 뉴트로 초이스 키튼 사료 상품의 클릭이 25회로 가장 많았고, 이 상품의 평점은 3.5점 입니다. 이 상품을 클릭한 사람 중 25%는 당일에 가입했으며, 여성 비율이 70% 입니다. 상품 상세보기를 클릭한 사람의 40%는 쿠폰을 가지고 있고, 일주일 이내 쿠폰이 만료되는 사람이 그 중 75% 입니다.

이벤트 설계 - Firebase

Event Property

카테고리: 애완동물

상품명: 뉴트로 초이스 키튼 사료

가격: 50000

섹션명: 할인이벤트

평점: 3.5

이벤트 발생 시점	event_name	event_dim.params			
		key: string_value	key: int_value	key: double_value	key: float_value
아이템 상세화면 보기 클릭	click_item	category : {{item_category}} title : {{item_name}} section : {{section_name}}	price : {{item_price}}		review : {{review_score}}

이벤트 설계 - BigQuery 적재구조

event_dim	RECORD
event_dim.name	STRING
event_dim.params	RECORD
event_dim.params.key	STRING
event_dim.params.value	RECORD
event_dim.params.value.string_value	STRING
event_dim.params.value.int_value	INTEGER
event_dim.params.value.float_value	FLOAT
event_dim.params.value.double_value	FLOAT
event_dim.timestamp_micros	INTEGER
event_dim.previous_timestamp_micros	INTEGER
event_dim.date	STRING

Schema	Details	Preview	
event_dim.date	event_dim.name	event_dim.params.key	event_dim.params.value.string_value
20161216	상세보기 클릭	카테고리	애완동물
		상품명	뉴트로 초이스 키튼 사료
20161216	first_open	가격	50000
		previous_first_open_count	null
		system_app	null
		update_with_analytics	null
		firebase_conversion	null
20161216	user_engagement	engagement_time_msec	null
		firebase_event_origin	auto
20161216	session_start	firebase_event_origin	auto

Firestore → BigQuery 로 적재할 때의 기본 구조입니다.

이벤트 설계 - Google Analytics

Implementation

Event hits can be sent using the `send` command and specifying a `hitType` of `event`. The `send` command has the following signature for the `event` hit type:

```
ga('send', 'event', [eventCategory], [eventAction], [eventLabel], [eventValue], [fieldsObject]);
```



Google Analytics

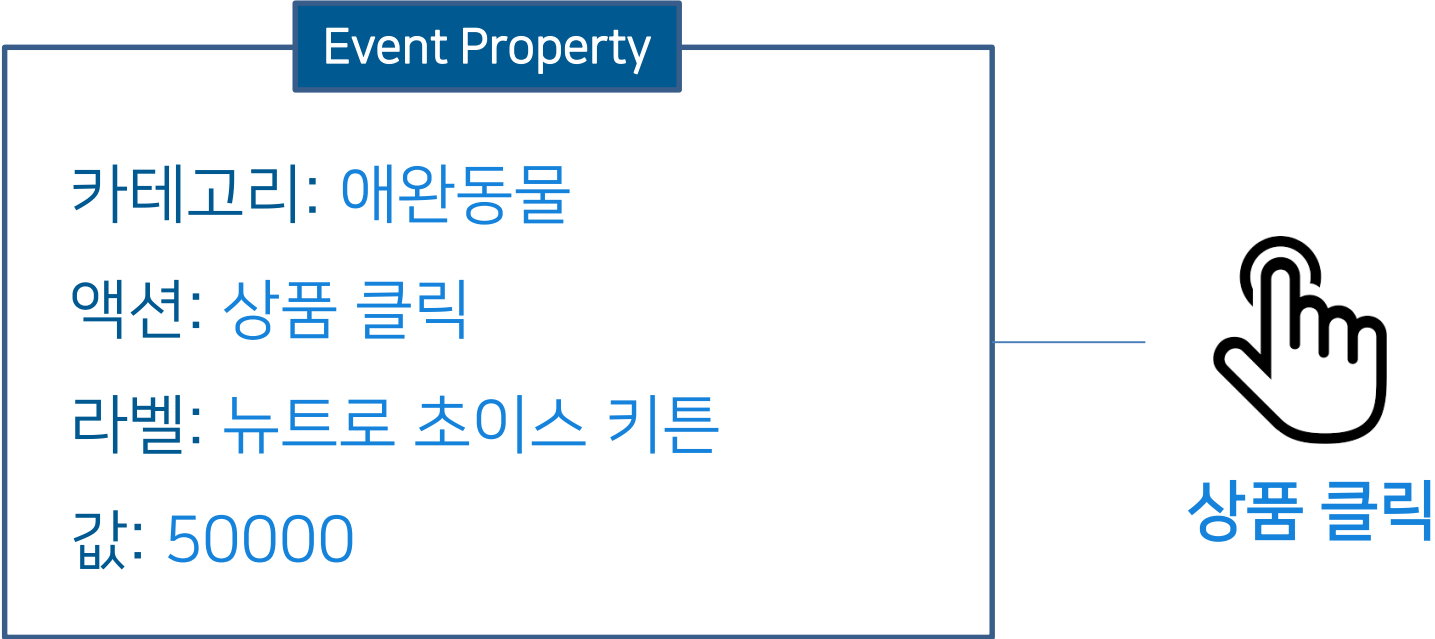
Event fields

The following table summarizes the event fields:

Field Name	Value Type	Required	Description
<code>eventCategory</code>	text	yes	Typically the object that was interacted with (e.g. 'Video')
<code>eventAction</code>	text	yes	The type of interaction (e.g. 'play')
<code>eventLabel</code>	text	no	Useful for categorizing events (e.g. 'Fall Campaign')
<code>eventValue</code>	integer	no	A numeric value associated with the event (e.g. 42)

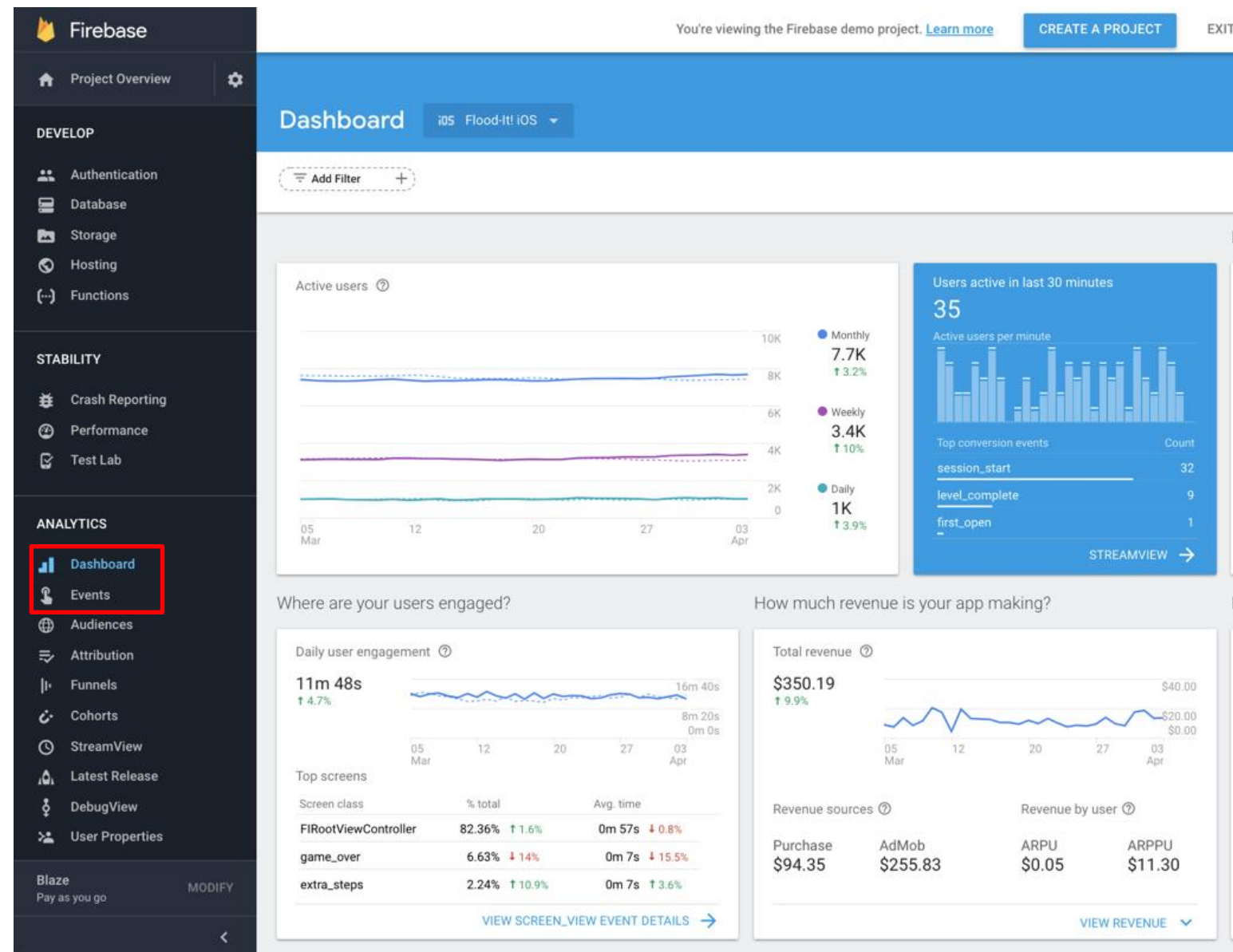
<https://developers.google.com/analytics/devguides/collection/analyticsjs/events>

이벤트 설계 - Google Analytics



이벤트 발생 시점	이벤트 카테고리	이벤트 액션	이벤트 라벨	이벤트 값
상품 클릭	{{item_category}}	상품 클릭	{{item_title}}	{{item_price}}

이벤트 대시보드 - Firebase



Analytics						
com.labpixies.flood						
DASHBOARD EVENTS AUDIENCES ATTRIBUTION FUNNELS COHORTS USER PROPERTIES						
Add Filter						
Last 30 days						
Compared to Aug 26, 2016 - Sep 24, 2016						
Event name ↑	Count		Value		Users	Mark as conversion
ad_click	659	+9.1%	-		279	-18.2%
app_clear_data	15	-55.9%	-		13	-61.8%
app_exception	58	-22.7%	-		39	-40.9%
app_remove	1,026	-1%	-		1,026	-1%
app_update	1	-	-		1	-
completed_5_levels	835	+61.8%	17,518	+62.7%	743	+73.2%
first_open	1,977	-7.9%	-		1,977	-7.9%
in_app_purchase	13	-60.6%	\$17.07	-79.4%	10	-44.4%
level_complete	16,298	-7.7%	306,558	-5.8%	1,981	-8.7%
level_complete_quickplay	58,113	+16.9%	1,273,062	+16.5%	1,927	-4.2%

Source: [Firebase](#)

이벤트 raw data 조회 - BigQuery

The screenshot displays the Google Cloud Platform BigQuery interface. The top navigation bar shows 'Google Cloud Platform' and the project 'gcpidentity-asset-export-1'. The left sidebar contains navigation links for 'Query history', 'Saved queries', 'Job history', 'Transfers', and 'Resources'. The 'Resources' section is expanded, showing a search bar and a list of datasets under 'gcpidentity-asset-export-1', including 'asset_inventory' and various Google App Engine and Cloud services.

The main area is the 'Query editor', which contains a SQL query:

```
1 SELECT instance.timestamp, REGEXP_EXTRACT(instance.resource.data.machineType, '.*/(.*)') AS machine_type, REGEXP_E
2 FROM `asset_inventory.google_compute_Instance` AS instance
3 JOIN UNNEST(instance.resource.data.disk) AS instance_disk
4 JOIN `asset_inventory.google_compute_Disk` AS disk
5 ON instance_disk.source = disk.resource.data.selfLink and instance.timestamp = disk.timestamp
6 where instance.resource.data.status = 'RUNNING'
7 group by timestamp, machine_type, source_image
8
```

Below the query editor, a green checkmark indicates the query is 'Valid'. The 'Processing location' is set to 'US'. Buttons for 'Run', 'Save query', 'Save view', and 'More' are visible.

The 'Query results' section shows the query is complete (0.020 sec elapsed, cached). The 'Job information' tab is selected, displaying the results in a table:

Row	timestamp	machine_type	source_image	num_disks
1	2019-01-17 16:25:11.805264 UTC	n1-standard-1	debian-9-stretch-v20190116	1
2	2019-01-17 04:08:52.968214 UTC	n1-standard-1	debian-9-stretch-v20190116	1
3	2019-01-17 00:11:39.614887 UTC	n1-standard-1	debian-9-stretch-v20190116	1

이벤트 대시보드 - Google Analytics

▼ 행동

개요

행동 흐름

▶ 사이트 콘텐츠

▶ 사이트 속도

▶ 사이트 검색

▼ 이벤트

개요

인기 이벤트

페이지

기본 측정기준: 이벤트 카테고리 이벤트 액션 이벤트 라벨

선택 행 도표 만들기 보조 측정기준 정렬 유형: 기본

이벤트 액션 ?

총 이벤트 수 ?

순 이벤트 수 ?

이벤트 값 ?

		22,553 전체 대비 비율(%): 100.00% (22,553)	10,262 전체 대비 비율(%): 100.00% (10,262)	0 전체 대비 비율(%): 0.00% (0)
<input type="checkbox"/>	1. Quickview Click	18,766 (83.21%)	7,937 (77.34%)	0 (0.00%)
<input type="checkbox"/>	2. Product Click	1,979 (8.77%)	1,286 (12.53%)	0 (0.00%)
<input type="checkbox"/>	3. Add to Cart	1,301 (5.77%)	696 (6.78%)	0 (0.00%)
<input type="checkbox"/>	4. Remove from Cart	404 (1.79%)	272 (2.65%)	0 (0.00%)
<input type="checkbox"/>	5. Onsite Click	103 (0.46%)	71 (0.69%)	0 (0.00%)

▼ 행동

개요

행동 흐름

▶ 사이트 콘텐츠

▶ 사이트 속도

▶ 사이트 검색

▼ 이벤트

개요

인기 이벤트

페이지

기본 측정기준: 이벤트 카테고리 이벤트 액션 이벤트 라벨

선택 행 도표 만들기 보조 측정기준: 이벤트 액션 정렬 유형: 기본

이벤트 카테고리 ?

이벤트 액션 ?

총 이벤트 수 ?

순 이벤트 수 ?

이벤트 값 ?

			22,553 전체 대비 비율(%): 100.00% (22,553)	10,262 전체 대비 비율(%): 100.00% (10,262)	0 전체 대비 비율(%): 0.00% (0)
<input type="checkbox"/>	1. Enhanced Ecommerce	Quickview Click	18,766 (83.21%)	7,937 (77.34%)	0 (0.00%)
<input type="checkbox"/>	2. Enhanced Ecommerce	Product Click	1,979 (8.77%)	1,286 (12.53%)	0 (0.00%)
<input type="checkbox"/>	3. Enhanced Ecommerce	Add to Cart	1,301 (5.77%)	696 (6.78%)	0 (0.00%)
<input type="checkbox"/>	4. Enhanced Ecommerce	Remove from Cart	404 (1.79%)	272 (2.65%)	0 (0.00%)
<input type="checkbox"/>	5. Contact Us	Onsite Click	103 (0.46%)	71 (0.69%)	0 (0.00%)

- 마이리얼트립의 앱 로그분석 환경 구축기



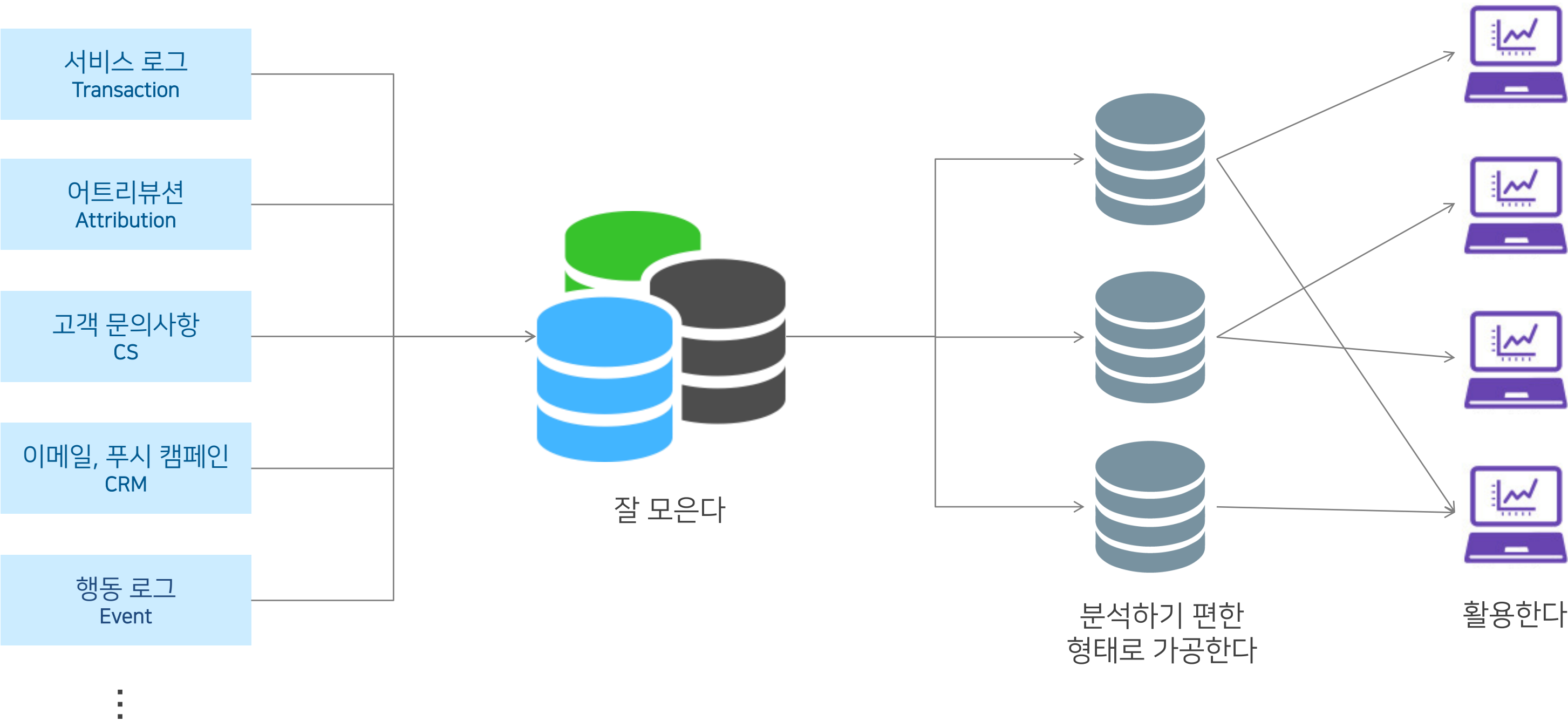
모바일 앱 로그분석, 어떻게 시작해야 할까?

Firebase와 BigQuery를 이용한 로그분석 시스템 구축하기

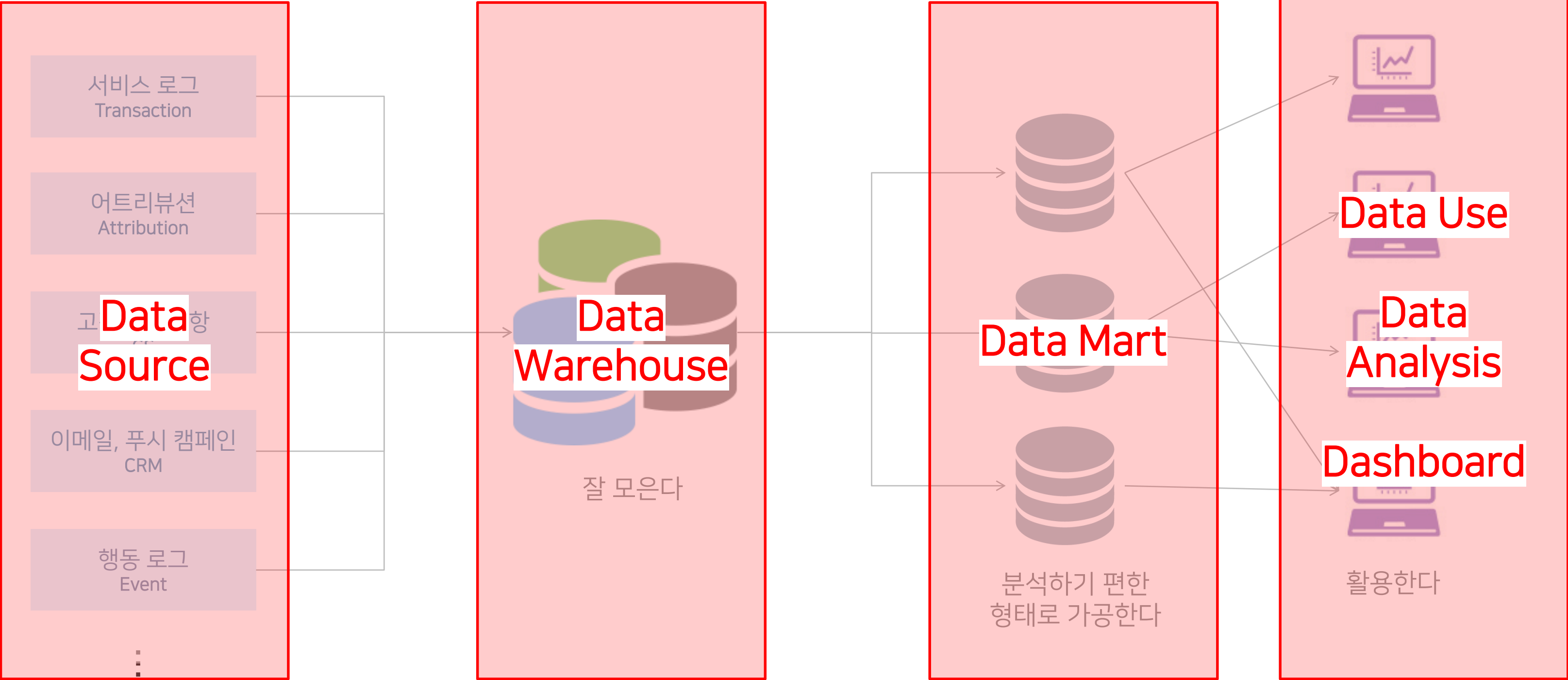
이벤트 로그 설계

- Labeling을 1시간 더 고민하면, 나중에 100시간이 편해집니다.
- QA의 퀄리티가 데이터셋의 퀄리티를 결정합니다.
- 전체를 갈아엎고 새로 싹 만드는 일의 난이도가 10이라면,
이후 서비스가 업데이트 될 때 꾸준히 잘 챙기는 일의 난이도는 100입니다.
- 로깅 작업을 잘 도와주는 개발자는 있지만, 좋아하는 개발자는 없습니다.
- ~~애초에 사람이 할 일이 아니거든요...~~

행동 로그를 예로 들었지만,
잘 쌓아둬야 하는 데이터도 매우 많습니다.



데이터가 흐르는 Flow 만들기 → 데이터 파이프라인



데이터가 흐르는 Flow 만들기 → 데이터 파이프라인

- 데이터 파이프라인도 기획과 설계가 필요합니다.
- 데이터 엔지니어와 데이터 분석가의 긴밀한 협업이 필요!
 - 어떤 데이터가 필요한지
 - 그 데이터를 어떻게 수집할지
 - 배치와 스트리밍을 어떻게 활용할지 (실시간 처리가 필요한지?)
 - 분석하기 편한 형태로 만들려면 어떤 전처리를 해야할지
 - 대시보드 등 데이터 사용자를 위한 Visualization은 어떻게 할지
 - 파이프라인 설계/구축을 위해 어느 정도의 리소스와 비용을 투자할지