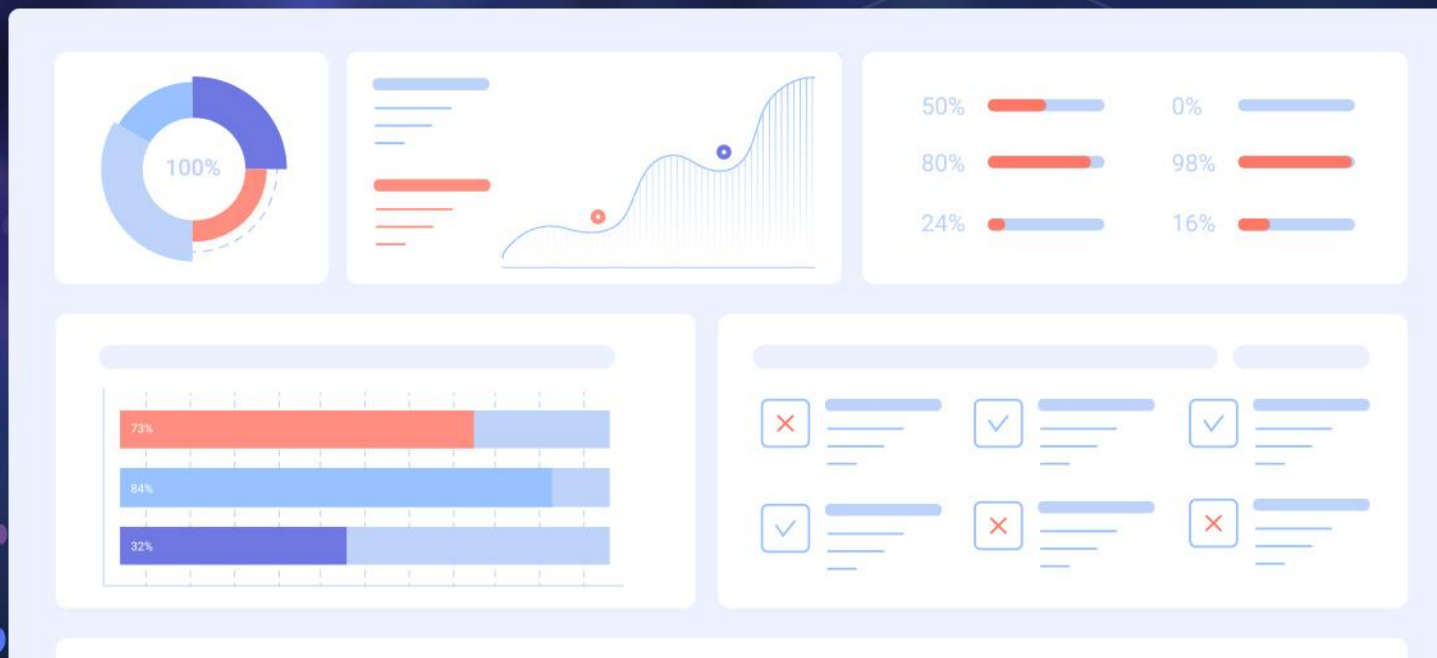


#빅쿼리 #데이터웨어하우스 #데이터시각화 #데이터스튜디오

빅데이터 한눈에 보기







3월 24일 수요일 오후 2시  YouTube 라이브

베스핀글로벌 소개

베스핀글로벌

Google Cloud Managed Service Provider Partner

전세계 28개* MSP 파트너 중 APAC에 MSP 파트너사이고 Infrastructure - Training 및 Service 부분의 Specialization Partner입니다.

 <p>Google Cloud Partner</p> <p>Managed Service Provider</p>	<p>Premier Partner</p>  <p>Infrastructure</p>
 <p>Specialization</p> <p>Infrastructure - Training, Service</p>	 <p>Premier Partner</p>

BESPIN GLOBAL



accenture Doif SUTHERLAND rackspace
 oXya SADA devoteam G Cloud Deloitte. wipro
 Cloud Ace claranet Nordcloud
 REPLY CLOUDREACH DATAPIPE Infosys HCL
 claranet Smartronix Cloud ASSURED taos oXya IBM redapt
 sopra steria REPLY BOLD

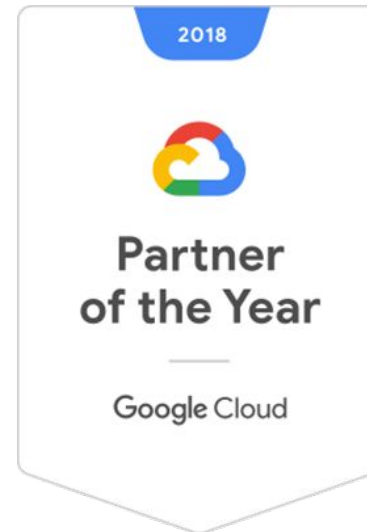
↑ Bospin Global holds GCP / G Suite resale rights for Regional Expansion (APAC, JAPAN, EMEA, NA)

* '20년 3월 기준

베스핀글로벌

Google Cloud 2018년, 2019년 2년 연속 Google Cloud Partner Award 수상

- 2018년 아시아 태평양지역 BreakThrough Partner of the Year ‘혁신적인 파트너상’
- 2019년 일본 및 아시아 태평양지역 Reseller Partner of the Year ‘리셀러 파트너상’



국내에서 가장 강력한 Google Cloud Partner.
 유능한 Google Cloud Specialist를 보유하고 있으며,
 다수의 자격증을 획득하고 있습니다.

148+

Google Cloud Platform Specialist

- ↑ 국내 다수 GCP Specialist 보유
- ↑ 레거시와 Cloud에 대한 IT 베테랑들의 전문가 집단
- ↑ 국내 최고 Anthos(Hybrid Cloud) 역량 보유



120+

Google Workspace

3 +



Google Certified Professional Cloud Architect

135 +



Google Certified Professional Data Engineer

4 +



Google Certified Professional Network Engineer

6 +



Google Certified Professional Developer

2 +

Deployment Services

3 +

- ↑ Google Cloud Managed Service Provider 획득
- ↑ APAC 다수의 Google Cloud 관련 인증 자격 보유
- ↑ Google Cloud Partner of the Year 2년 연속 수상 ('18년 Breakthrough, '19년 ReSeller)

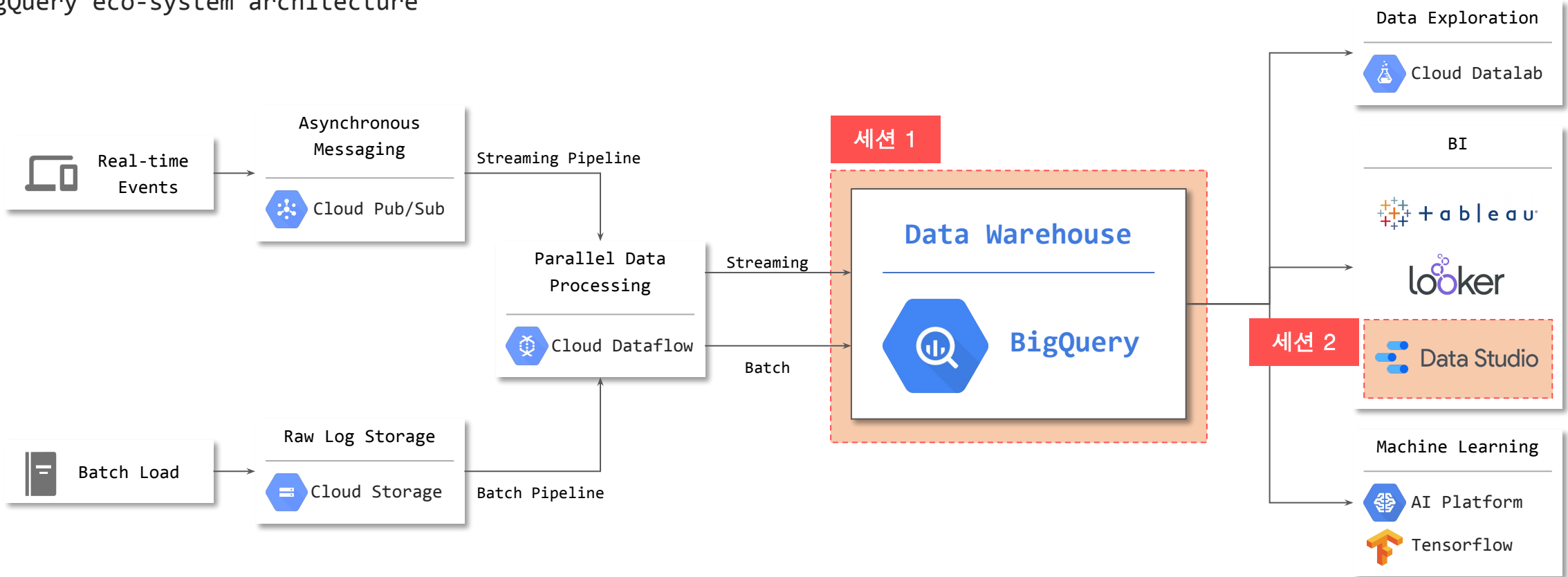
| Agenda

세션 구성

세션 1 BigQuery Data Warehouse 서비스 소개 (약 40분)

세션 2 DataStudio를 이용한 데이터 시각화 구현 (약 40분)

BigQuery eco-system architecture



BigQuery Data Warehouse 서비스 소개

베스핀글로벌 이지영

Contents

1. 서비스 소개

2. 사용 관리



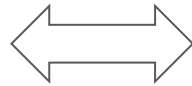
- BigQuery: 빅데이터 분석 및 활용 목적의 완전 관리형 클라우드 데이터 웨어하우스



데이터 소스



BigQuery



분석 활용

- Serverless
- Petabyte Scale
- Data Encryption and Security
- Programmatic Interaction
- Standard SQL
- ...

리소스 운영 및 관리 불필요 → 데이터와 분석에 집중 가능한 환경 제공.

페타바이트 이상의 데이터를 빠르고 쉽게 저장 및 분석 가능.

액세스 관리로 보안 유지. 데이터는 미사용 및 전송 상태에서 항상 암호화.

REST API 및 Java, Python, Node.js 등의 클라이언트 라이브러리 제공.

표준 SQL을 이용한 데이터 조회로 데이터 분석을 위한 손쉬운 접근 가능.

기능 구성

- BigQuery 서비스는 빅데이터 저장 및 조회 + 주요 기능으로 구성

BigQuery

Data Warehousing

: Serverless, full managed and Petabyte Scale

Query as a Service

: SQL 인터페이스를 통한 쉽고 빠른 데이터 접근 조회

Data Transfer Service

: Google Cloud Storage 및 AWS S3 등의 저장소의 데이터를 BigQuery로 전송하는 서비스

Scheduled Query Service

: 반복 실행하는 query에 대한 스케줄링 서비스

Key Features

BigQuery ML

: BigQuery 데이터 + SQL를 사용하여 머신러닝 모델 구현 및 예측

```

Step 1: Model creation
CREATE MODEL `<PROJECT>.<DATASET>.<MODEL>`
OPTIONS(
  model_type='<Model_Type>'
  , input_label_cols=['label_column']
  , option_name='option_value') AS
SELECT
  feature_1
  , feature_2
  , feature_3
    
```

```

Step 2: Prediction
SELECT
  feature_1, feature_2, feature_3
  , ...
  , label_column
  , predicted_label
FROM
  ML.PREDICT(
    MODEL `<PROJECT>.<DATASET>.<MODEL>`, (
      SELECT
        feature_1, feature_2, feature_3
    )
    )
    
```

BigQuery BI Engine

: BI 프로세스를 위한 인메모리 연산 (computing)엔진 활성화 기능.

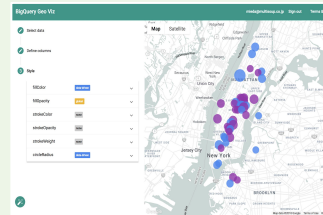


BigQuery GIS

: BigQuery 내 지리정보 분석의 네이티브 지원 기능. 공간 데이터에 대한 분석 기능.

```

BigQuery Geo Viz
1 #standardSQL
2 -- Finds Citi Bike stations with > 30 bikes
3 SELECT
4   ST_GeogPoint(longitude, latitude) AS WKT,
5   num_bikes_available
6 FROM
7   `bigquery-public-data.new_york.citibike_stations`
8 WHERE num_bikes_available > 30
9
    
```

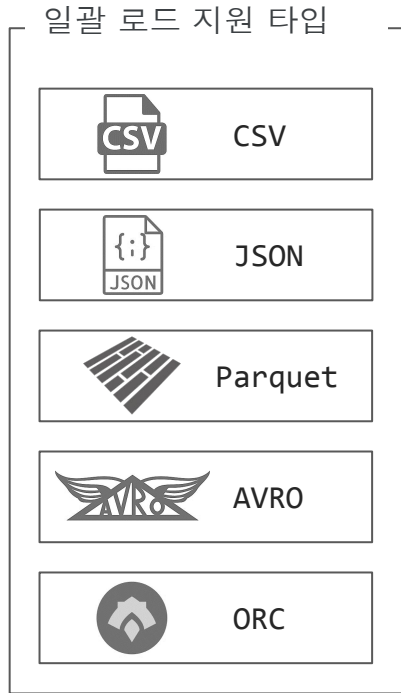


- 퍼블릭 클라우드 간 비교

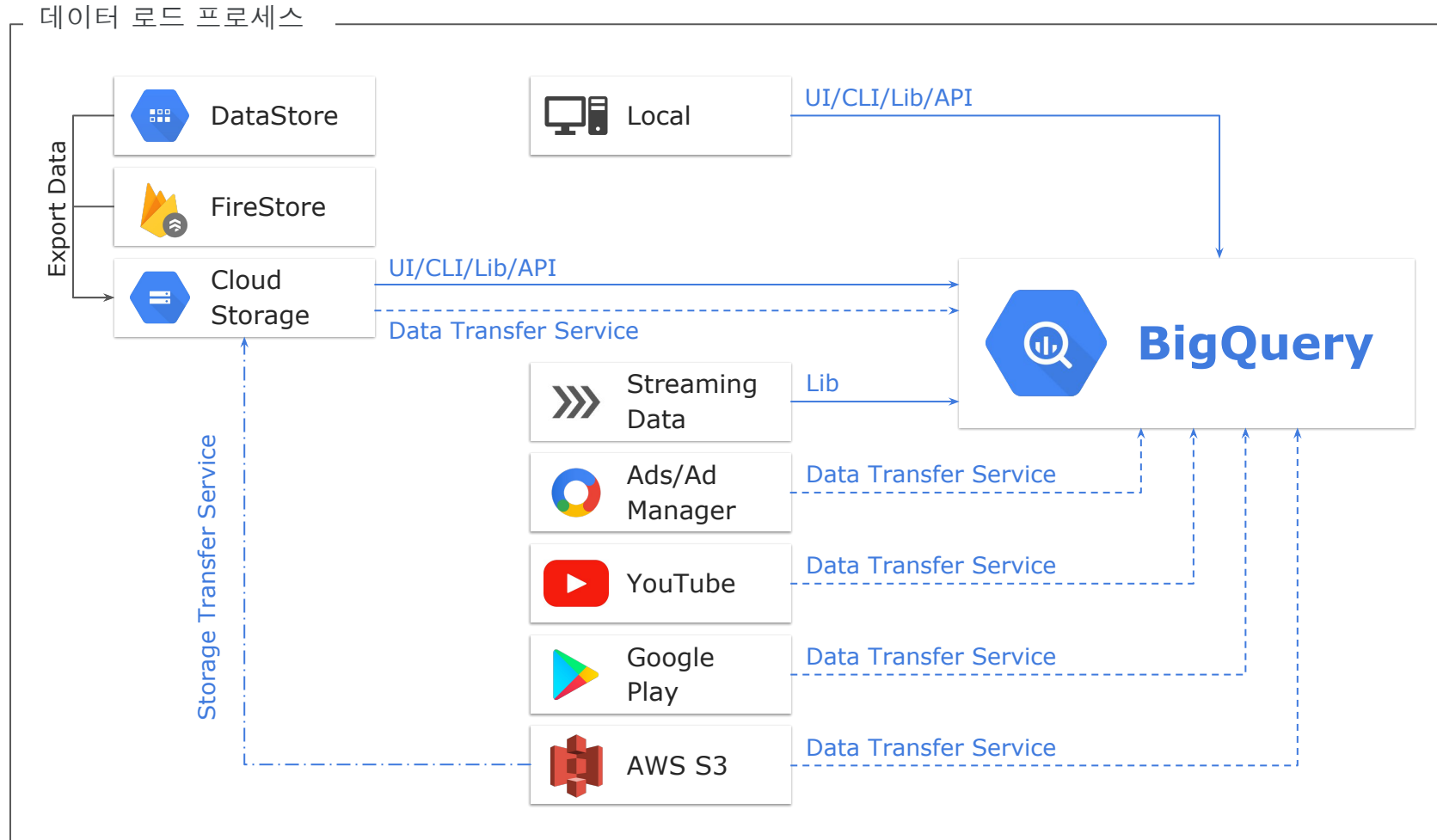
Service	Data Warehousing	Big Data Query as a Service

데이터 로드

▪ BigQuery 지원 파일 형식 및 데이터 로드

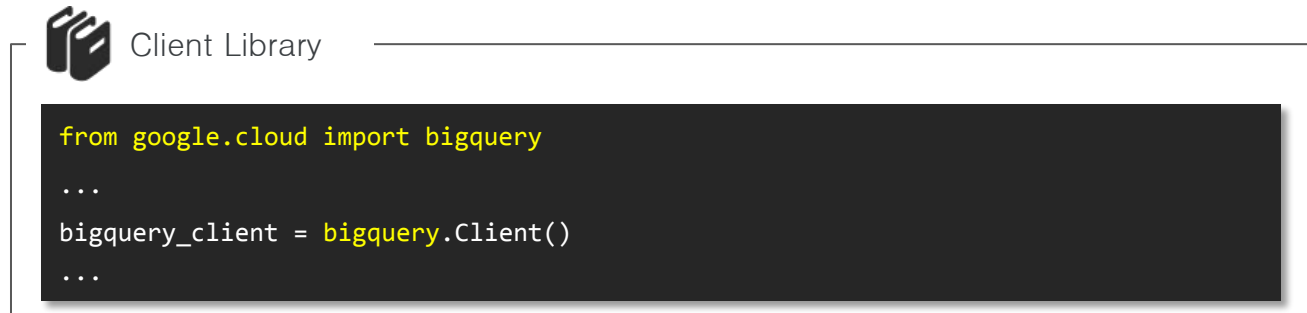
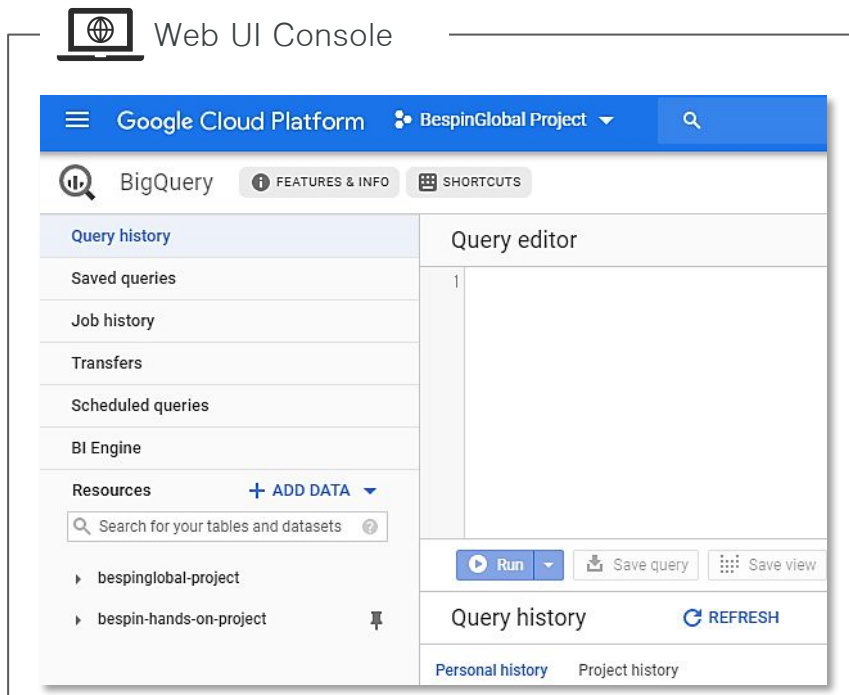
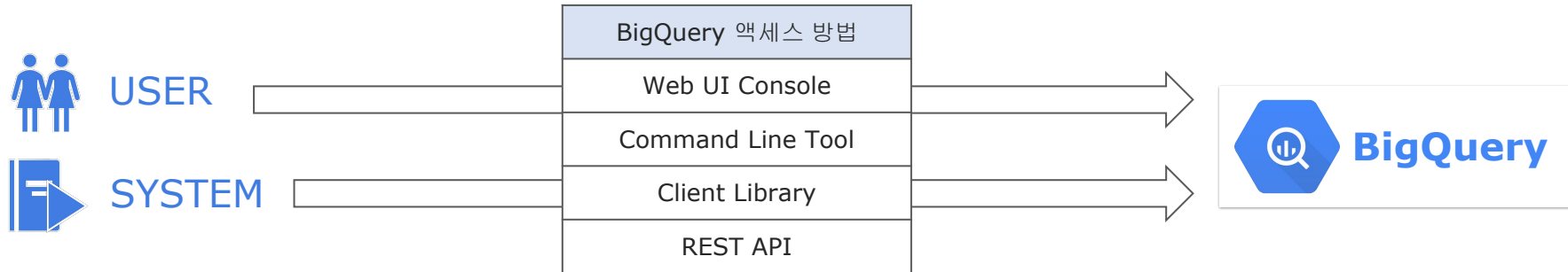


* CSV : 쉼표(,)로 구분
 * JSON : 줄바꿈으로 구분

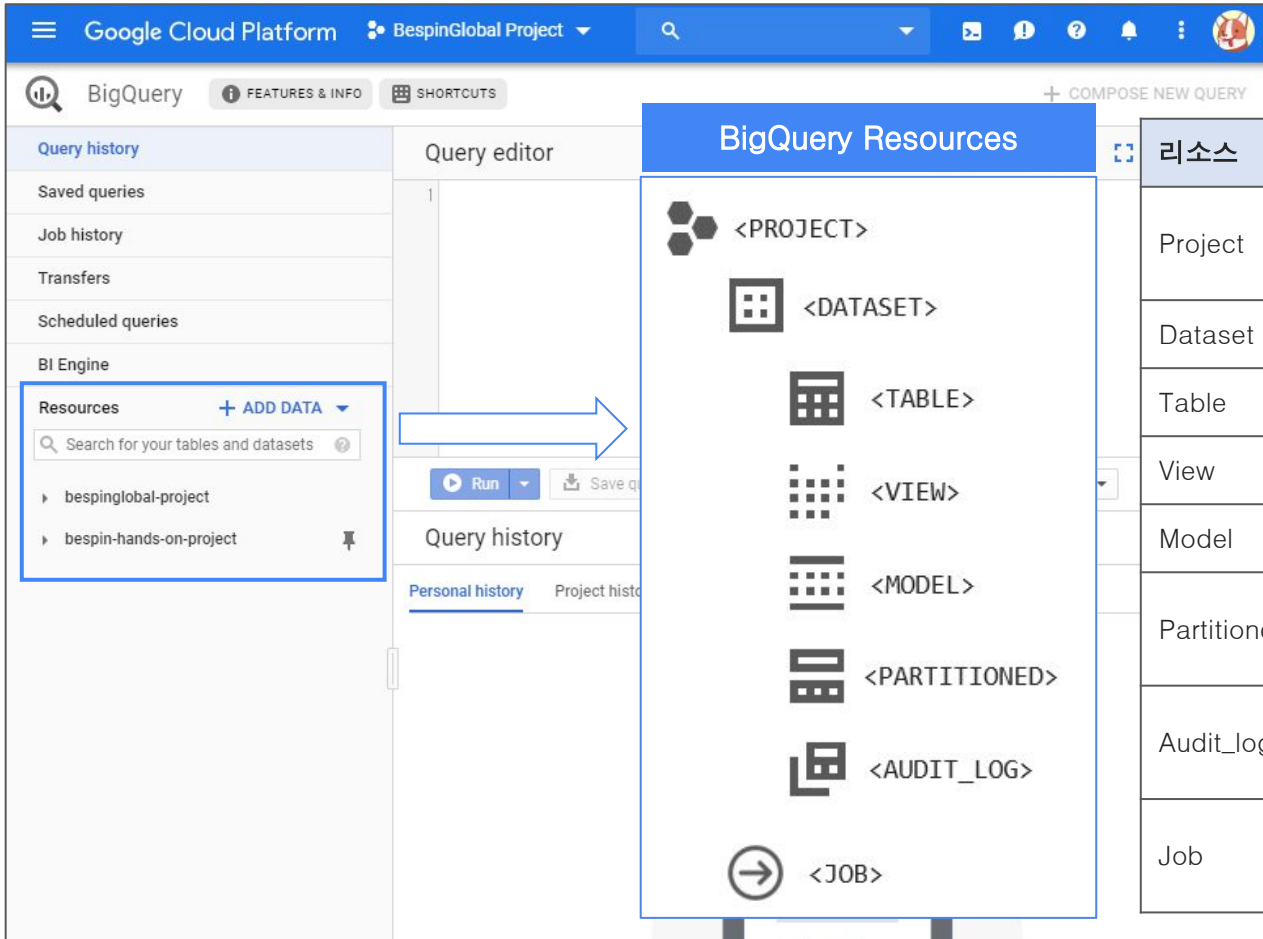


데이터 액세스

- BigQuery 데이터에 대한 액세스 방법



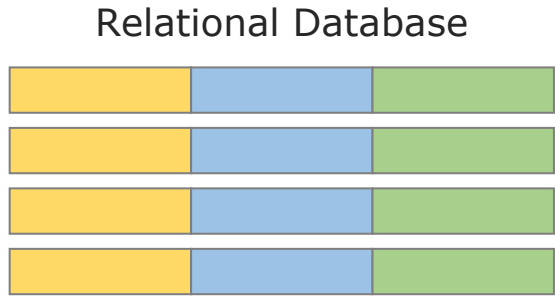
- 데이터가 저장되는 <테이블>은 <데이터세트> 하위의 리소스로 구성
- GCP Web UI 콘솔 화면 기준



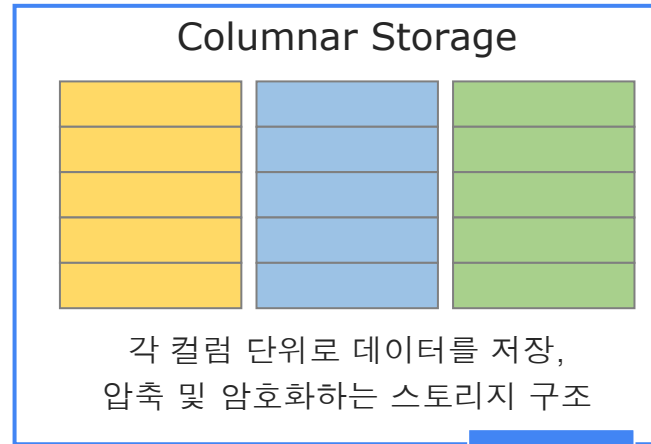
리소스	설명
Project	결재를 사용 설정하고, GCP 리소스에 대한 접근 권한 및 서비스 사용에 대한 관리를 위한 리소스 단위
Dataset	테이블과 뷰에 대한 액세스 구성 및 제어에 사용되는 최상위 컨테이너
Table	데이터를 저장하는 리소스 단위
View	SQL 쿼리로 정의한 가상의 테이블이며 실제 데이터를 저장하지 않음.
Model	BigQuery ML 기능으로 생성된 머신러닝 모델
Partitioned	쿼리 성능 및 비용 관리를 목적으로 데이터의 ingestion 시간이나 테이블 내 TIMESTAMP 및 DATE 컬럼으로 데이터를 파티션한 테이블
Audit_log	데이터 액세스, 시스템 이벤트, 관리자 액션과 같은 GCP 시스템에서 제공 가능한 서비스 사용 및 운영 관련 로그 데이터
Job	데이터 조회와 같은 단기 실행 작업 외 BigQuery가 데이터 로드/내보내기 /복사 등의 장기 실행하는 비동기 작업

아키텍처

- Storage는 column 형식으로 데이터를 저장

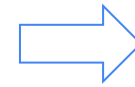


레코드 기반의 스토리지 구조
트랜잭션 업데이트를 지원



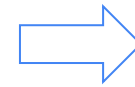
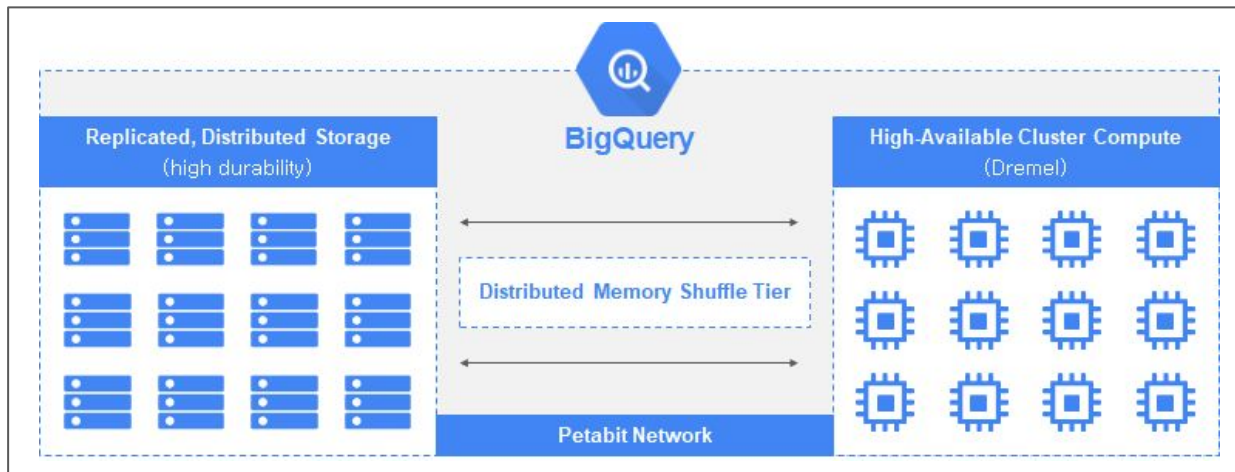
각 컬럼 단위로 데이터를 저장,
압축 및 암호화하는 스토리지 구조

BigQuery



유사한 column 데이터를 효율적으로 압축하고 쿼리 응답에 필요한 데이터만 READ하여 디스크 스캔범위 축소
→ 빅데이터 조회 성능 향상, 복잡한 쿼리에 대한 처리에 적합

- Computing과 Storage가 분리된 아키텍처



- 컴퓨팅과 스토리지가 분리되어 있음에도 페타비트 네트워크로 연결되어 빠른 데이터 처리 가능
- 컴퓨팅과 분리된 스토리지 구조로 탄력적인 용량 확장 가능

객체 기반 스키마 지원

- columnar 형식으로 repeated, nested 필드 저장 가능

"addresses" nested fields

groupId	accessDate	addresses.zip	addresses.city	addresses.state	addresses.address
1501583974	20170801	05626	Seoul	Songpa-gu	Songpa-dong
		05643	Seoul	Songpa-gu	Bangei-dong
		16585	Suwon	Kwonseon-gu	Kwonseon-dong
		16475	Suwon	Paldal-gu	Ingye-dong
		16543	Suwon	Yeongtong-gu	Maetan-dong
1675431231	20170803	05626	Seoul	Songpa-gu	Songpa-dong
		05643	Seoul	Songpa-gu	Bangei-dong
		13311	Seongnam	Sujeong-gu	Taepyeong-dong
		13119	Seongnam	Sujeong-gu	Bongjeong-dong
		13366	Seongnam	Jungwon-gu	Seongnam-dong
		16475	Suwon	Paldal-gu	Ingye-dong
		16543	Suwon	Yeongtong-gu	Maetan-dong
1684312458	20170804	05626	Seoul	Songpa-gu	Songpa-dong
		16475	Suwon	Paldal-gu	Ingye-dong
...

repeated fields

JSON 파일, Avro 파일, Cloud Firestore 내보내기 파일과 같은 객체 기반의 스키마를 지원하는 소스 형식의 중첩(nested) 및 반복(repeated) 데이터를 BigQuery로 직접 로드 가능

Schema

Name	Type	Mode	
groupId	STRING	REQUIRED	X
accessDate	STRING	NULLABLE	X
addresses	RECORD	REPEATED	X
addresses.zip	STRING	NULLABLE	X
addresses.city	STRING	NULLABLE	X
addresses.state	STRING	NULLABLE	X
addresses.address	STRING	NULLABLE	X

+ Add field

할당량 및 한도

▪ BigQuery로의 요청에 대한 주요 할당량 및 한도



* 각 할당량 및 한도에 대한 상향 조정은 구글 문의 필요

구분	항목	수량	비고
데이터 로드 JOB	'Project'당 일일 데이터 로드 JOB	100,000 개	- 실행 실패한 JOB도 포함.
	'Table'당 일일 데이터 로드 JOB	1,500 개	- 실행 실패한 JOB도 포함.
	'Table'당 최대 열(Column) 수	10,000 개	
	데이터 로드 JOB당 처리되는 최대 파일 수	10,000,000 개	- 데이터 로드 설정에서 모든 와일드카드(*) URI와 일치하는 모든 파일을 포함.
	데이터 로드 JOB 실행시간 제한	6 시간	
데이터 쿼리 JOB	'Project'별 동시에 처리되는 쿼리 수 제한	100 개	- 쿼리 캐시로 리턴되는 결과 및 테스트 실행 쿼리는 적용범위에서 제외.
	쿼리 실행 시 참조되는 테이블 수 제한	1,000 개	- 데이터 조회 시, JOIN 등으로 참조되는 모든 테이블.
	'Project'별 동시에 사용가능한 최대 Slot 수	2,000 개	- 주문형(On-demand) 요금제에서는 사용되는 Slot 수는 'Project' 단위로 적용 - 2,000 개가 초과되는 Slot 리소스 필요 시, Flat-rate 요금제 사용
...

* JOB : BigQuery가 데이터 로드, 데이터 내보내기, 데이터 쿼리 또는 데이터 복사를 위해 사용자를 대신하여 실행하는 행위

Contents

1. 서비스 소개

2. 사용 관리



가격 정책

▪ BigQuery 과금은 데이터 수집 / 저장 / 조회 비용으로 구성

* 서울(asia-northeast3) 기준의 가격 정보이며, 지역별로 비용이 다름

데이터 수집 (로드)

구분	과금	세부정보
일괄 로드 (Batch Loading)	공유 슬롯 풀 사용시 무료	데이터가 BigQuery에 로드되면 스토리지 비용이 청구됩니다.
스트리밍 (Streaming Inserts)	\$ 0.012 / 200 MB	- 무료제공 범위 없음. Insert를 성공한 row만 과금. 최소 크기는 1KB로 적용하여 개별 row를 계산

데이터 저장 (스토리지)

구분	과금	세부정보
활성 스토리지 (Active Storage)	\$ 0.023 / GB	- 매월 10 GB 까지 무료범위 제공
장기 스토리지 (Long-term Storage)	\$ 0.016 / GB	- 매월 10 GB 까지 무료범위 제공. 90일동안 Table 변경이 없는 경우, Active Storage에서 자동으로 변경

데이터 조회 (분석)

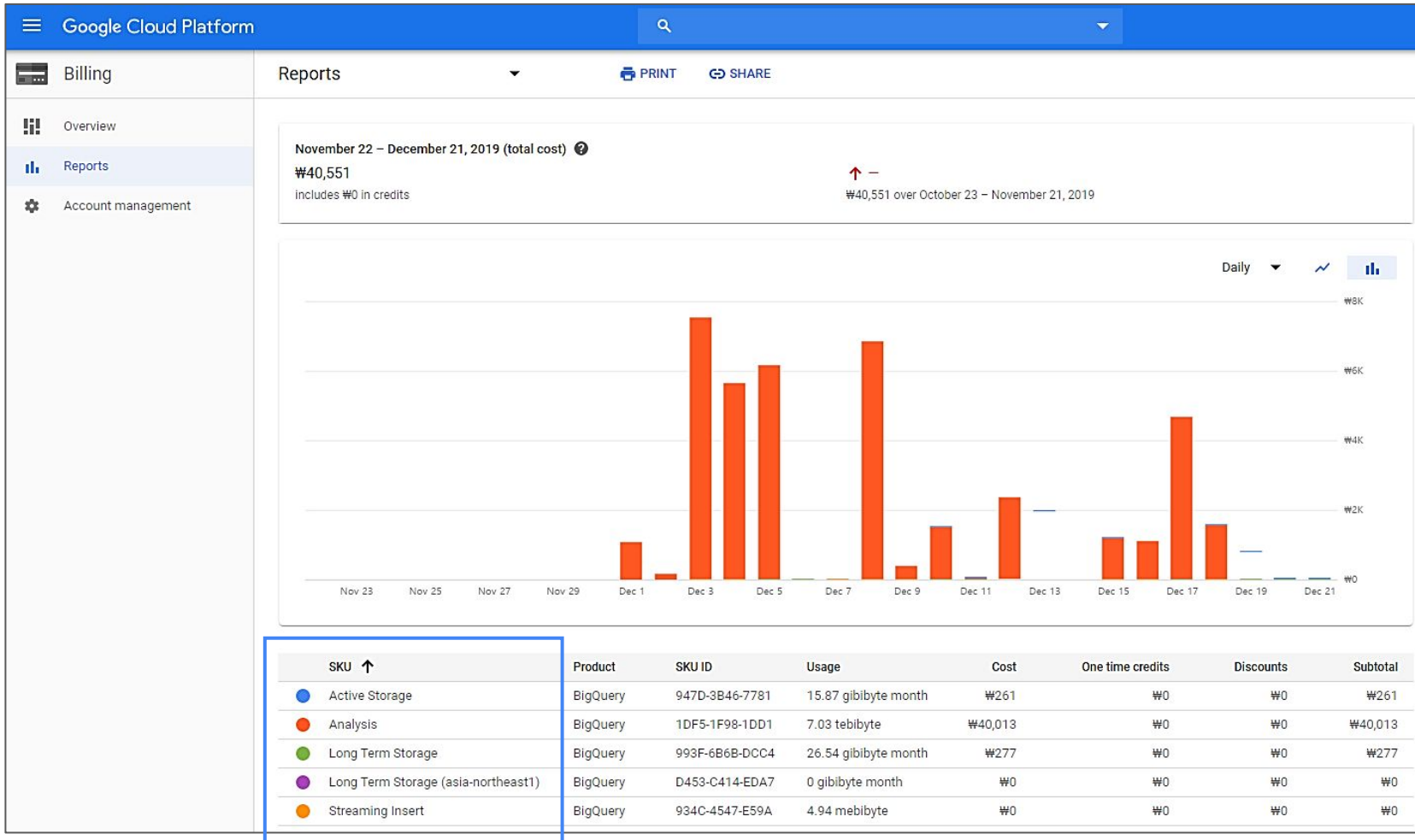
- 주문형(On-Demand) 요금제 과금

Operation	Pricing	Description
Query	\$ 6.00 / TB	- 매월 최초 1 TB 무료 - SQL 명령 실행 시 처리되는 데이터에 따라 과금 - MB 단위로 과금, 최소 데이터 처리량 = 10 MB - 오류 발생 및 캐시된 쿼리는 과금 제외

- 정액제(Flat-rate) 요금제 과금

Monthly cost	Number of Slot	Description
\$ 2,400	100	- Flat-rate 요금제는 처리되는 데이터 크기와는 무관하게 Slot 크기에 따라 과금 - 연 단위 계약 시 매월 \$ 2,040 / 100 Slots * Slot: SQL 쿼리 실행에 사용하는 가상 CPU 단위

- 웹 콘솔 > Billing > Report > "group by SKU"



전송 비용 SKU
- Streaming Insert

분석 비용 SKU
- Analysis

저장 비용 SKU
- Active Storage
- Long Term Storage
- Long Term Storage (region)

* SKU : Stock-Keeping Unit

READ 데이터 사전 체크

- 쿼리 실행 전에 미리 예상되는 READ 데이터 크기를 체크

- GCP 웹UI 콘솔: Query Validator

```
1 SELECT
2   visitorId, visitNumber, visitId, visitStartTime, date
3 FROM
4   `bigquery-public-data.google_analytics_sample.ga_sessions_*`
```

Run Save query Save view Schedule query More

This query will process 29.3 MB when run. ✓

- CLI: '--dry_run'

```
bq --location=[LOCATION] query --use_legacy_sql=false --dry_run 'SELECT [FIELD1], [FIELD2], [FIELD3] FROM `[PROJECT].[DATASET].[TABLE]`'
```

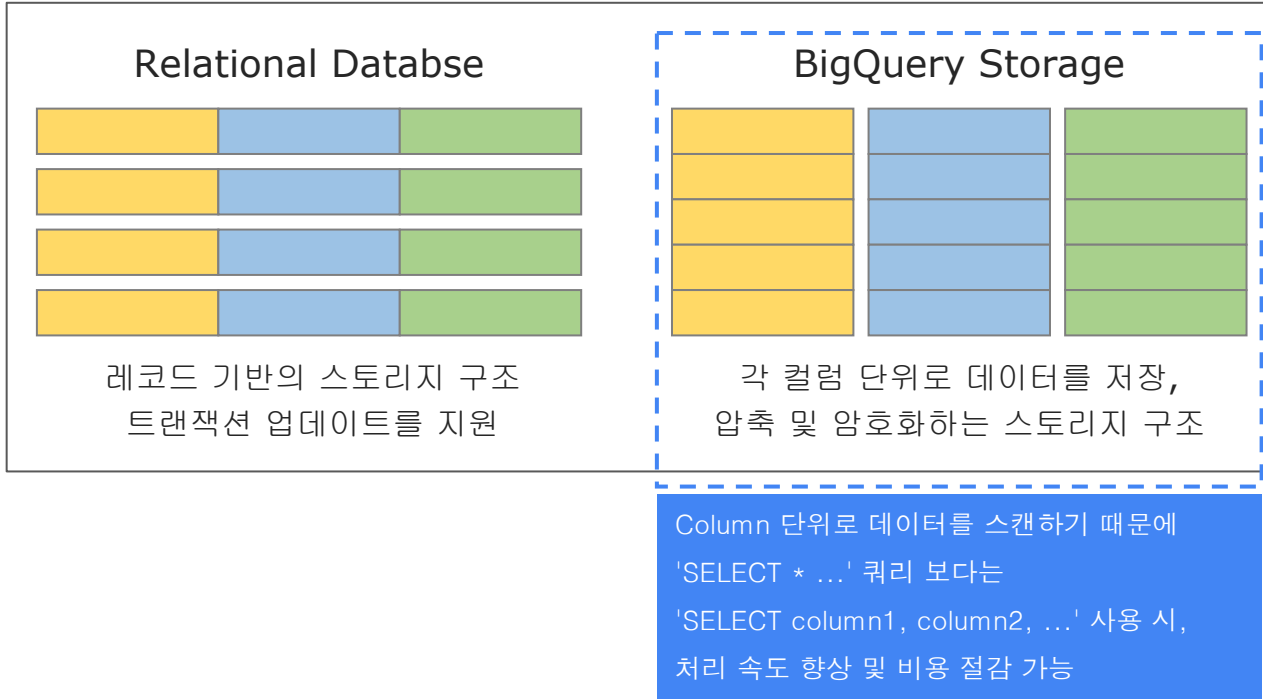
- Client Library: QueryJobConfig.dry_run 속성을 True로 설정

```
from google.cloud import bigquery
client = bigquery.Client()
job_config = bigquery.QueryJobConfig()
job_config.dry_run = True
job_config.use_query_cache = False
query_job = client.query(
    ...
)
assert query_job.state == "DONE"
assert query_job.dry_run
print("This query will process {} bytes.".format(query_job.total_bytes_processed))
```

- API: 쿼리 request를 submit할 때 'dryRun' 파라미터 값을 'true'로 설정

전체 column 조회 제한

- 'SELECT * ...' 조회 쿼리에 대한 사용 제한



```

1 SELECT
2 *
3 FROM
4 `bigquery-public-data.google_analytics_sample.ga_sessions_*`
5

```

This query will process 5.4 GB when run. ✓

→ 전체 column 조회 시, 5.4 GB 데이터 처리

```

1 SELECT
2 visitorId, visitNumber, visitId, visitStartTime, date
3 FROM
4 `bigquery-public-data.google_analytics_sample.ga_sessions_*`
5

```

This query will process 29.3 MB when run. ✓

→ 일부 column 조회 시, 29.3 MB 데이터 처리

- 'LIMIT ...' 사용은 비용과 무관 → 'LIMIT'를 적용해도 스캔되는 데이터

```

1 SELECT visitorId, visitNumber, visitId, visitStartTime, date
2 FROM `bigquery-public-data.google_analytics_sample.ga_sessions_*`
3 LIMIT 1000

```

This query will process 29.3 MB when run. ✓

데이터 분할 - 테이블 파티셔닝

- 테이블을 파티션이라고 하는 세그먼트로 분할하여 데이터 저장

- EVENT_TABL

EVENT_DT [DATE]	EVENT_NM [STRING]	EVENT_TYPE [INTEGER]	...	EVENT_TS [TIMESTAMP]
2019-09-17	VIEW	1208	...	1568916613
2019-09-17	PURCHASE	2134	...	1568752216
2019-09-18	LIKE	1101	...	1568848259
2019-09-19	PURCHASE	2134	...	1568922160
2019-09-19	PLAY	2201	...	1568933741



- Ingestion-time(수집 시간) partitioned table

_PARTITION DATE	_PARTITION TIME	EVENT_DT [DATE]	EVENT_NM [STRING]	EVENT_TYPE [INTEGER]	...	EVENT_TS [TIMESTAMP]
2020-08-16	1597622397	2019-09-17	VIEW	1208	...	1568916613
2020-08-16	1597622398	2019-09-17	PURCHASE	2134	...	1568752216
2020-08-16	1597622399	2019-09-18	LIKE	1101	...	1568848259
2020-08-17	1597622400	2019-09-19	PURCHASE	2134	...	1568922160
2020-08-17	1597622401	2019-09-19	PLAY	2201	...	1568933741

pseudo columns

데이터 수집 날짜 및 시간을
기준으로 파티션 분할

- Date/Timestamp(시간 단위) partitioned table

EVENT_DT [DATE]	EVENT_NM [STRING]	EVENT_TYPE [INTEGER]	...	EVENT_TS [TIMESTAMP]
2019-09-17	VIEW	1208	...	1568916613
2019-09-17	PURCHASE	2134	...	1568752216
2019-09-18	LIKE	1101	...	1568848259
2019-09-19	PURCHASE	2134	...	1568922160
2019-09-19	PLAY	2201	...	1568933741

테이블 내 DATE 또는
TIMESTAMP 타입의 필드 데이터를
기준으로 파티션 분할

- Integer range(정수 범위) partitioned table

EVENT_DT [DATE]	EVENT_NM [STRING]	EVENT_TYPE [INTEGER]	...	EVENT_TS [TIMESTAMP]
2019-09-17	VIEW	1208	...	1568916613
2019-09-17	PURCHASE	2134	...	1568752216
2019-09-18	LIKE	1101	...	1568848259
2019-09-19	PURCHASE	2134	...	1568922160
2019-09-19	PLAY	2201	...	1568933741

테이블 내 정수 필드 데이터를
기준으로 파티션 분할

시작(start), 종료(end), 범위(interval)
설정으로 파티션 범위 지정

90일간 테이블에 대한 업데이트가 발생하지 않는 경우,
storage 저장 비용 50% 절감

파티션 범위조건으로 조회 쿼리 실행 시,
데이터 스캔범위가 축소되어 쿼리 비용 절감

데이터 분할 - 테이블 샤딩

- 테이블 명을 "[PREFIX] + YYYYMMDD" 으로 구성하여 테이블 데이터를 분할 저장

• EVENT_TABL

EVENT_DT [DATE]	EVENT_NM [STRING]	EVENT_TYPE [INTEGER]	...	EVENT_TS [TIMESTAMP]
2019-09-17	VIEW	1208	...	1568916613
2019-09-17	PURCHASE	2134	...	1568752216
2019-09-18	LIKE	1101	...	1568848259
2019-09-19	PURCHASE	2134	...	1568922160
2019-09-19	PLAY	2201	...	1568933741
2019-09-19	LIKE	1101	...	1568935516



• EVENT_TABLE_201909

EVENT_DT [DATE]	EVENT_NM [STRING]	EVENT_TYPE [INTEGER]	...	EVENT_TS [TIMESTAMP]
2019-09-17	VIEW	1208	...	1568916613
2019-09-17	PURCHASE	2134	...	1568752216

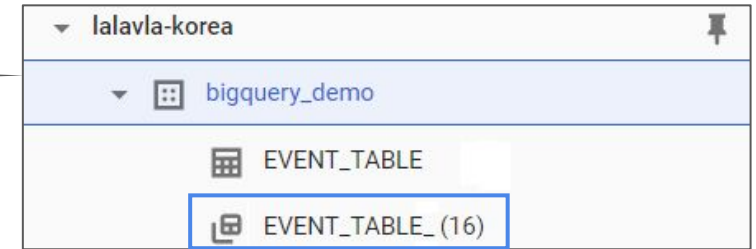
• EVENT_TABLE_201909

EVENT_DT [DATE]	EVENT_NM [STRING]	EVENT_TYPE [INTEGER]	...	EVENT_TS [TIMESTAMP]
2019-09-18	LIKE	1101	...	1568848259

• EVENT_TABLE_201909

EVENT_DT [DATE]	EVENT_NM [STRING]	EVENT_TYPE [INTEGER]	...	EVENT_TS [TIMESTAMP]
2019-09-19	PURCHASE	2134	...	1568922160
2019-09-19	PLAY	2201	...	1568933741
2019-09-19	LIKE	1101	...	1568935516

샤딩된 테이블은 웹콘솔에서 그룹화되어 화면 출력



Field name	Type	Mode
2019-09-19		
2019-09-18		
2019-09-17		

90일간 테이블에 대한 업데이트가 발생하지 않는 경우, storage 저장 비용 50% 절감

여러 테이블 동시 참조 시, 와일드카드(*) 또는 _TABLE_SUFFIX 조건을 이용하여 데이터 조회

```
FROM `projectid.dataset.EVENT_TABLE_201909*`
```

```
WHERE _TABLE_SUFFIX BETWEEN '17' AND '19'
```

데이터 스캔 범위를 축소하여 조회비용 절감

데이터 분할 - 테이블 클러스터링

- 테이블 내 1개 이상의 필드를 자동 정렬하여 데이터 배치

- EVENT_TABL

EVENT_DT [DATE]	EVENT_NM [STRING]	EVENT_TYPE [INTEGER]	...	EVENT_TS [TIMESTAMP]
2019-09-17	VIEW	1208	...	1568916613
2019-09-17	PURCHASE	2134	...	1568752216
2019-09-18	LIKE	1101	...	1568848259
2019-09-19	PURCHASE	2134	...	1568922160
2019-09-19	PLAY	2201	...	1568933741



- "EVENT_NAME" clustered

EVENT_DT [DATE]	EVENT_NM [STRING]	EVENT_TYPE [INTEGER]	...	EVENT_TS [TIMESTAMP]
2019-09-18	LIKE	1101	...	1568848259
2019-09-17	VIEW	1208	...	1568916613
2019-09-19	PURCHASE	2134	...	1568922160
2019-09-17	PURCHASE	2134	...	1568752216
2019-09-19	PLAY	2201	...	1568933741

클러스터링 필드를 적용 가능한 필드 타입
 - DATE, BOOL, GEOGRAPHY, INT64, NUMERIC, STRING, TIMESTAMP

최대 4개 필드까지 클러스터링 필드로 지정 가능

데이터 조회 실행 시, 정렬된 블록을 스캔하여 불필요한 데이터 검색을 축소하여 쿼리 성능 향상

- 테이블 파티셔닝 + 클러스터링 예시

```
SELECT
    EVENT_DT
    , EVENT_NM
    , EVENT_TYPE
    , ...
    . EVENT_TS
FROM
    `projectid.dataset.TABLE_NAME`
WHERE
    _PARTITIONTIME = '2020-08-16'
    AND EVENT_NM = 'PLAY';
```

	_PARTITION DATE	EVENT_DT [DATE]	EVENT_NM [STRING]	...
2020-08-15	2020-08-15	2019-09-09	WISH	...
	2020-08-15	2019-09-09	WISH	...
2020-08-16	2020-08-16	2019-09-18	LIKE	...
	2020-08-16	2019-09-19	PLAY	...
	2020-08-16	2019-09-19	PLAY	...
	2020-08-16	2019-09-17	PURCHASE	...
	2020-08-16	2019-09-19	PURCHASE	...
	2020-08-16	2019-09-17	VIEW	...
2020-08-17	2020-08-17	2019-09-21	LIKE	...
	2020-08-17	2019-09-21	LIKE	...

데이터 조회 실행 시,
 - 날짜 분할된 해당 테이블만 스캔
 - 정렬된 블록을 스캔하여 조건에 맞는 데이터 취합
 → 불필요한 데이터 검색을 축소하여 쿼리 성능 향상

테이블 분할 비교

테이블 데이터 분할 처리 비교

미분할

테이블 샤딩

'date' 필드 파티셔닝

'date' 필드 파티셔닝 +
'country_name' 필드 클러스터링

리소스
화면

쿼리
조회
결과

```

1 SELECT
2 *
3 FROM
4 `lalavla-korea.bigquery_demo.covid_19_open_data`
5 WHERE
6 date BETWEEN '2020-04-01' AND '2020-06-30'
7 AND country_name = 'South Korea'
    
```

Query complete (3.2 sec elapsed, 1,016.4 MB processed)

1,016.4 MB 데이터 처리

3.2 seconds 소요

```

1 SELECT
2 *
3 FROM
4 `lalavla-korea.bigquery_demo.covid_19_open_data_p`
5 WHERE
6 date BETWEEN '2020-04-01' AND '2020-06-30'
7 AND country_name = 'South Korea'
    
```

Query complete (1.4 sec elapsed, 403.5 MB processed)

403.5 MB 데이터 처리

2.8 seconds 소요

```

1 SELECT
2 *
3 FROM
4 `lalavla-korea.bigquery_demo.covid_19_open_data_2020*`
5 WHERE
6 _TABLE_SUFFIX BETWEEN '0401' AND '0630'
7 AND country_name = 'South Korea'
    
```

Query complete (2.8 sec elapsed, 403.5 MB processed)

403.5 MB 데이터 처리

1.4 seconds 소요

```

1 SELECT
2 *
3 FROM
4 `lalavla-korea.bigquery_demo.covid_19_open_data_pc`
5 WHERE
6 date BETWEEN '2020-04-01' AND '2020-06-30'
7 AND country_name = 'South Korea'
    
```

Query complete (1.2 sec elapsed, 402.6 MB processed)

402.6 MB 데이터 처리

1.2 seconds 소요

쿼리 사용량 제한 설정

- 프로젝트 및 사용자 계정 레벨의 쿼리 사용량 제한 설정 가능

Google Cloud Platform IAM & admin Quotas

Quota type: All quotas | Service: BigQuery API | Metric: All metrics | Location: All locations | Clear

Quota type	Service	Metric	Location	Limit	
<input type="checkbox"/> BigQuery API Query usage per day	Global	0 MIB	9,823,029 MIB (9.368 TiB)	Unlimited	View hierarchy
<input type="checkbox"/> BigQuery API Streaming insert rows per second	Global	0	0	100,000	View hierarchy
<input type="checkbox"/> BigQuery API Query usage per day per user	Global	-	-	Unlimited	View hierarchy
<input type="checkbox"/> BigQuery API Google internal requests per day	Global			0	

Daily quotas reset at midnight Pacific Time (PT).

Extract bytes

Quota Name	Limit
Extract bytes per day	10,995,116,277,760

쿼리 사용 크기에 대한 제한 설정으로 예산범위 이상의 과다 청구 발생 방지

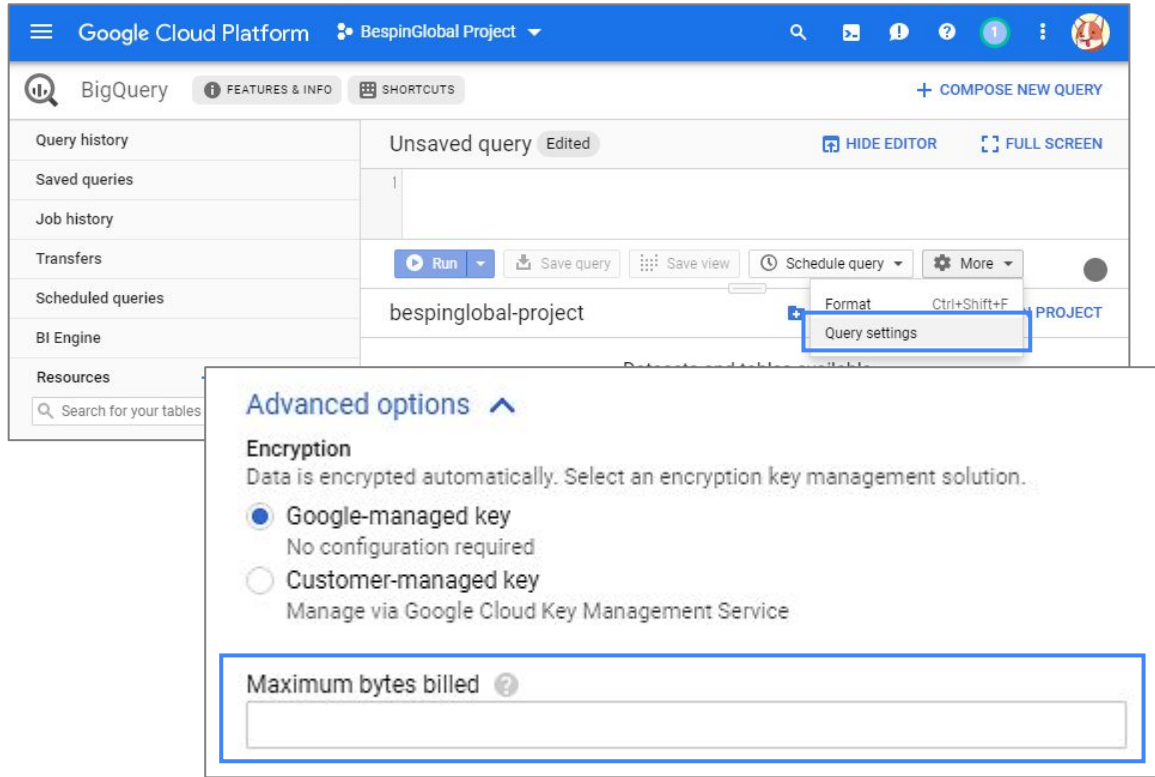
- 제한범위 이상으로 사용 시 발생 오류

Custom quota exceeded: Your usage exceeded the custom quota for **QueryUsagePerDay**, which is set by your administrator. For more information, see <https://cloud.google.com/bigquery/cost-controls>

Custom quota exceeded: Your usage exceeded the custom quota for **QueryUsagePerUserPerDay**, which is set by your administrator. For more information, see <https://cloud.google.com/bigquery/cost-controls>

최대 과금 범위 및 테이블 만료기간 설정

- 비용 청구되는 최대 바이트 수를 설정

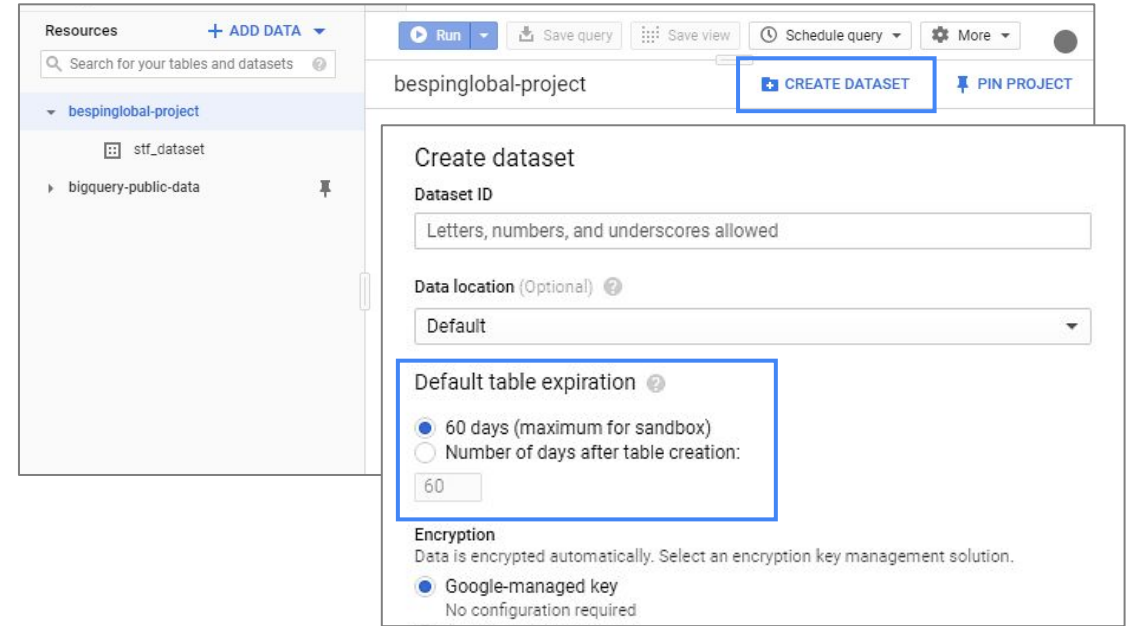


청구되는 최대 바이트 설정 가능. 초과 시 쿼리실행은 실패하며 오류 출력

- 청구되는 최대 바이트 초과 시, 출력되는 오류 메시지

```
Error: Query exceeded limit for bytes billed: 1000000.
10485760 or higher required.
```

- 데이터세트 생성 시점에 테이블 만료기간 설정



데이터세트 하위의 테이블은 기간이 만료되면 자동 삭제

- 파티션 만료일자도 설정 가능: DML 적용

```
ALTER TABLE mydataset.mytable
SET OPTIONS (
  -- Sets partition expiration to 5 days
  partition_expiration_days=5
)
```

Data Studio를 이용한 데이터 시각화 구현

베스핀글로벌 윤영훈

Contents

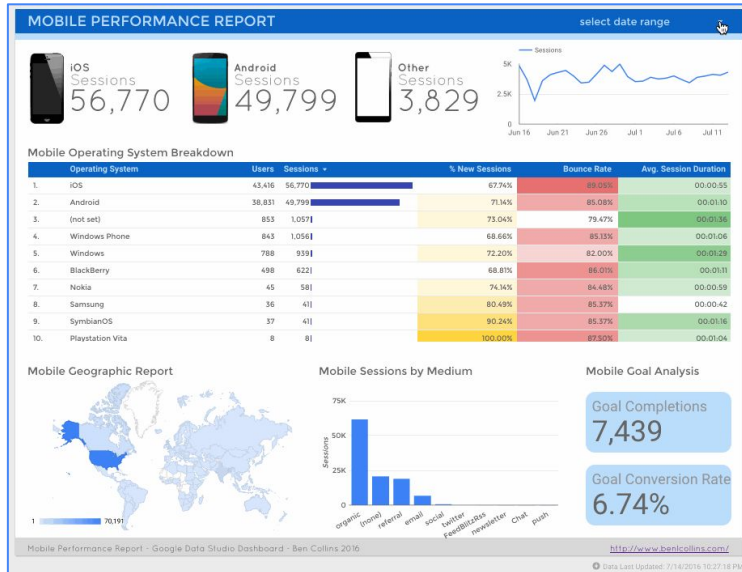
1. 서비스 소개

2. 데모 시연



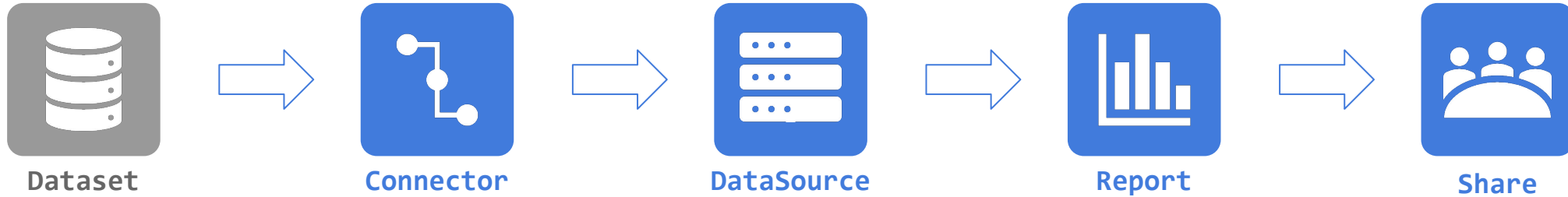
Data Studio

- 다양한 데이터 소스를 기반으로 대시보드 및 보고서를 만들고 공유할 수 있는 데이터 시각화 도구



Service concept

- 데이터 저장소의 데이터를 대시보드 형태로 시각화 처리하여 공유



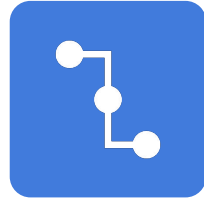
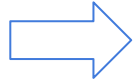
Dataset	데이터를 저장하는 물리적인 레이어로 BigQuery, MySQL 등의 데이터 저장소
Connector	Dataset 레이어와 DataSource 레이어를 연결하는 데이터 커넥터로 구글, 파트너사, 오픈소스 커넥터들을 사용 가능
DataSource	데이터 필드와 해당 필드들을 활용한 추가 속성들을 포함하는 논리적인 레이어로 데이터를 실제로 저장하지 않음
Report	데이터를 시각화하여 화면으로 보여주는 프레젠테이션 레이어로 scorecard, column bar, line graph 등의 chart를 구성
Share	생성된 report를 public 공개 또는 특정 사용자와 공유가 가능하며, 타 사이트로의 embed 및 PDF 다운로드 가능

Dataset and Connector

- 구글/파트너/오픈소스 Connector를 이용하여 물리적인 데이터 저장소와 연동



Dataset



Connector



The screenshot displays the 'Connectors' page in Data Studio, categorized into three sections:

- Google Connectors (18):** Connectors built and supported by Data Studio. Includes:
 - Google Analytics: Connect to Google Analytics reporting views.
 - Google Ads: Connect to Google Ads performance report data.
 - Google Sheets: Connect to Google Sheets.
 - BigQuery: Connect to BigQuery tables and custom queries.
 - File Upload: Connect to CSV (comma-separated values) files.
 - Campaign Manager: Connect to Campaign Manager data.
 - Cloud Spanner: Connect to Google Cloud Spanner databases.
 - Cloud SQL for MySQL: Connect to Google Cloud SQL for MySQL databases.
- Partner Connectors (156):** Connectors built and supported by Data Studio partners. Includes:
 - Custom JSON/CSV/XML: Fetch data from any JSON, CSV or XML with a URL or from your Google Drive.
 - 400+ Marketing Data Sources: Fetch data from over 400 marketing sources through a single connector.
 - Ad Data + All Other Sources: DataTap provides hundreds of native API connector alongside a scalable and fully fetched ETL engine.
- Open Source Connectors (3):** Connectors built by the Data Studio open source community. Includes:
 - Chrome UX Report: Explore user experience stats about an origin in the Chrome UX Report.
 - GKE Usage Metering: Connect to GKE usage metering data.
 - Kaggle: Connect to and visualize Kaggle datasets. This connector supports a single CSV file up to 20MB size for each data ...

Google Connector
 - Google에서 개발하여 제공하는 Connector
 - Google 서비스 제품과 연동

Partner Connector
 - Supermetric, TapClicks 등의 파트너사에서 개발하여 제공하는 Connector

Open Source Connector

DataSource

- 물리적인 데이터 저장소와 Connector로 연결된 논리적인 데이터 리소스 객체



DataSource

ecommerce_all_sessions Data credentials: Owner Data freshness: 12 hours

EDIT CONNECTION Field editing in reports: On CANCEL ADD TO REPORT

Index	Field	Type	Aggregation
1	Record Count	123 Number	Auto
2	fullVisitorId	ABC Text	None
3	channelGrouping	ABC Text	None
4	time	123 Number	None
5	country	Country	None
6	city	ABC Text	None
7	totalTransactionRevenue	123 Number	None
8	transactions	123 Number	None
9	timeOnSite	123 Number	None
10	pageviews	123 Number	None
11	sessionQualityDim	123 Number	None
12	date	Date (YYYYMMDD)	None
13	visitId	Date (YYYYMMDD)	None
14	type	ABC Text	None
15	productRefundAmount	123 Number	None
16	productQuantity	123 Number	None
17	productPrice	123 Number	None
18	product		
19	product		

DataSource 생성 시, 필드명 및 필드타입 등에 대한 변경 가능

REFRESH FIELDS



You are about to add data to this report

ecommerce_all_sessions

Note that **Report Editors** can create charts using the new data source(s), and can add dimensions and metrics not currently included in the report.

CANCEL ADD TO REPORT

DataSource

DataSource 및 Field에 대한 변환처리 및 신규생성 가능

- Calculated Field

데이터소스의 필드들을 연산하여 새로운 필드를 만들 수 있다.
 COUNT, LENGTH, REGEXP_MATCH 등의 function 적용 가능
support.google.com/datastudio/table/6379764?hl=en&ref_topic=7570421

Field Name: channelGroupingKr
 Field ID: calc_q7zuidt1b

```

    1 CASE
    2   WHEN channelGrouping = "Direct" THEN "직접 접속"
    3   WHEN channelGrouping = "Organic Search" THEN "자연 검색"
    4   WHEN channelGrouping = "Social" THEN "소셜"
    5   WHEN channelGrouping = "Email" THEN "이메일"
    6   WHEN channelGrouping = "Affiliates" THEN " 제휴사"
    7   WHEN channelGrouping = "Referral" THEN "추천"
    8   WHEN channelGrouping = "Paid Search" THEN "유료 검색"
    9   WHEN channelGrouping = "Display" THEN "광고"
    10  ELSE "Other"
    11  END
    
```

기존 필드들의 값을 계산 및 조합하여 데이터소스 내 새로운 필드 생성 가능

- Data Blending

각 DataSource의 필드를 조합하여 새로운 DataSource를 생성할 수 있다.
 생성한 DataSource는 report에 바로 활용할 수 있으며 Blended DataSource는 left outer join 의 결과물이다.

Data Source A, Data source B, Blended data source

Data Source, Data Source, Data Source → BLEND → Data View

Join keys: Unique ID (selected)

Available Fields: Customer Name, Date of Birth, Favourite Colour, Mode of Transport, Unique ID

Blended Data: Unique ID, Customer Name, Date of Birth, Favourite Colour, Mode of Transport, Unique ID, Customer Name, Date of Purchase, Unique ID, What Customer bought

Report and Share

- Chart, Graph, Filter 등으로 데이터를 시각화하고 여러 사람과 공유



Report



Share

+ Add a chart ▾ | ⌵ ▾ | 📅 | ≡ | 🔄 | <> | 🖨️ | 🔍 | 🔗 | 🔍 | 🔍

Table

Scorecard

Total: 1,168 | Sessions: 69.3K

Time series

Bar

Pie

Geo map

Line

Area

Scatter

Pivot table

Bullet

Treemap

🔍
📄
👤+ Share ▾
🗪
🐱

- 👤+ Invite people
- 🕒 Schedule email delivery
- 🔗 Get report link
- 🔗 Embed report
- 📄 Download report

Sharing settings

Link to share (only accessible by collaborators)

https://datastudio.google.com/open/1U3lqPE77EQaG_trmKzx0opjFJBo2WpFk

Who has access

👤 Specific people can access Change...

🐱	younghun yun (you) younghun29.yun@gmail.com	✎ ✕
👤	윤영준YoungHun Yun younghun.yun@bespinglobal.com	✔ Can edit Can view

Invite people:

Enter names or email addresses... ✎

Done

Get a link to this report ✕

Only specific people can access this report. [Change sharing settings](#)

Link to share

<https://datastudio.google.com/s/gKwlgqk9CPA> COPY LINK

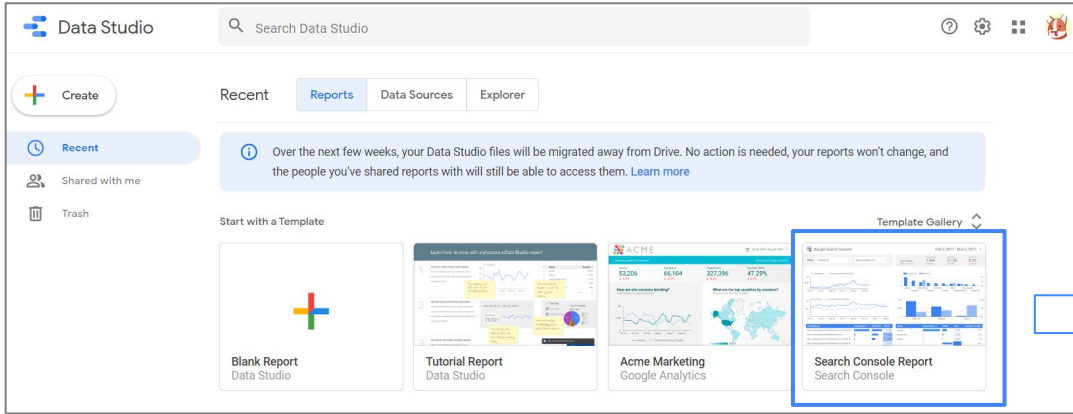
Link to your current report view

Your current report view includes your selections in date range controls, filter controls, and more. [Learn more](#)

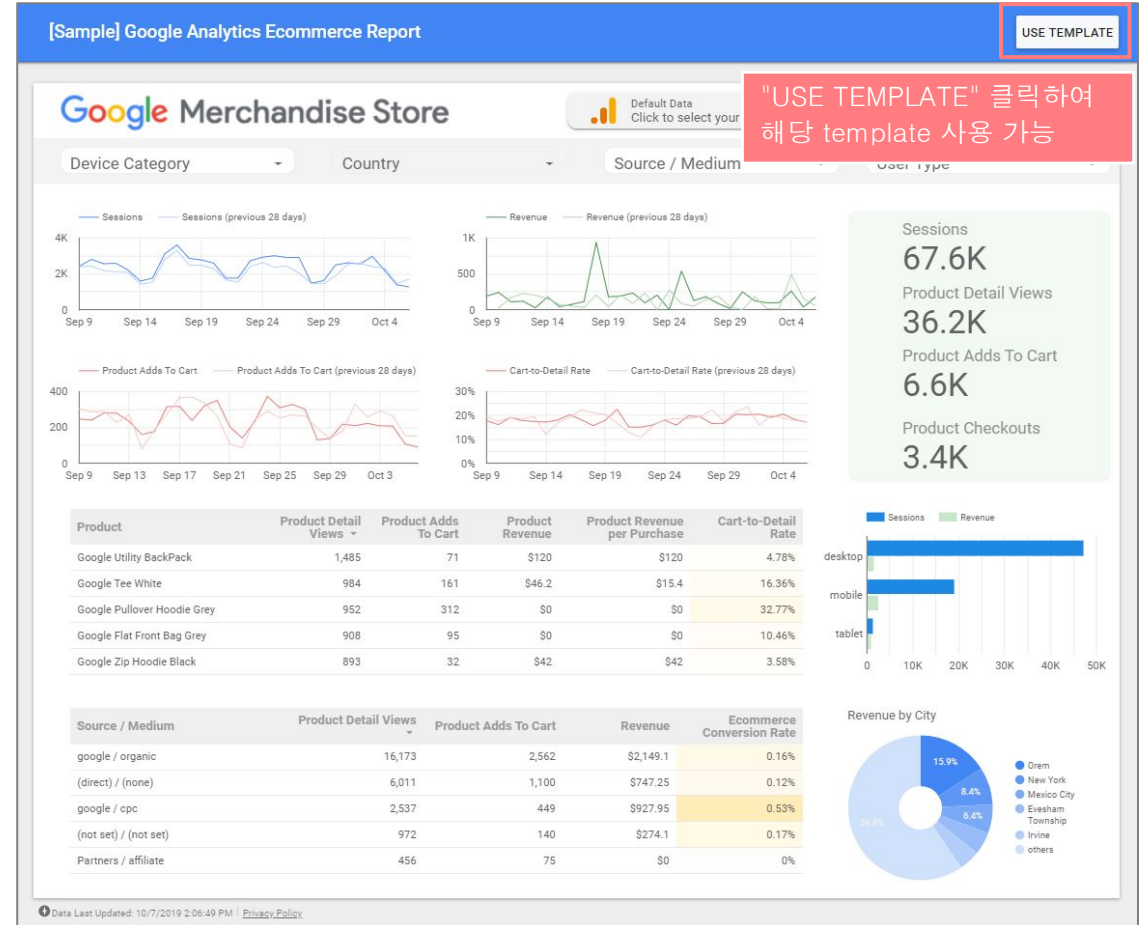
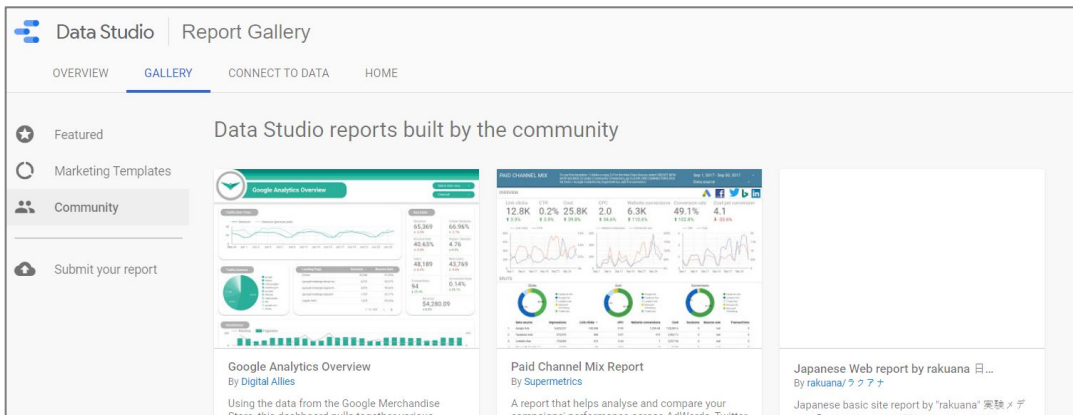
Template

- Template를 사용하여 Report 생성 가능

- DataStudio 목록 페이지



- <https://datastudio.google.com/gallery?category=community>



Contents

1. 서비스 소개

2. 데모 시연



- 구글 클라우드에서 제공하는 public data를 dataset로 활용합니다.

The image shows a composite screenshot. On the left is the Google Merchandise Store website, displaying various merchandise items like water bottles and caps. On the right is the Google Cloud BigQuery console interface, showing details for a dataset named 'ga_sessions_'. A blue arrow points from the store URL to the BigQuery dataset URL. A grey box at the bottom right contains summary statistics for the dataset.

Google Merchandise Store <https://shop.googlemerchandisestore.com/>

Google Merchandise Store의 샘플 데이터

<https://console.cloud.google.com/bigquery?project=bespinn-webinar-0324>

Resources + ADD DATA

Search for your tables and datasets

- bespin-gcp-lab
- bigquery-public-data
 - google_analytics_sample
 - ga_sessions_ (366)

Run Save query Save view Schedule query More

ga_sessions_ 2017-08-01 QUERY TABLE COPY TABLE

Schema Details Preview

Description Obfuscated Google Analytics 360 data. It's a great way to look at business data and experiment and learn the benefits of analyzing Google Analytics 360 data in BigQuery. Learn more at: <https://support.google.com/analytics/answer/7586738>

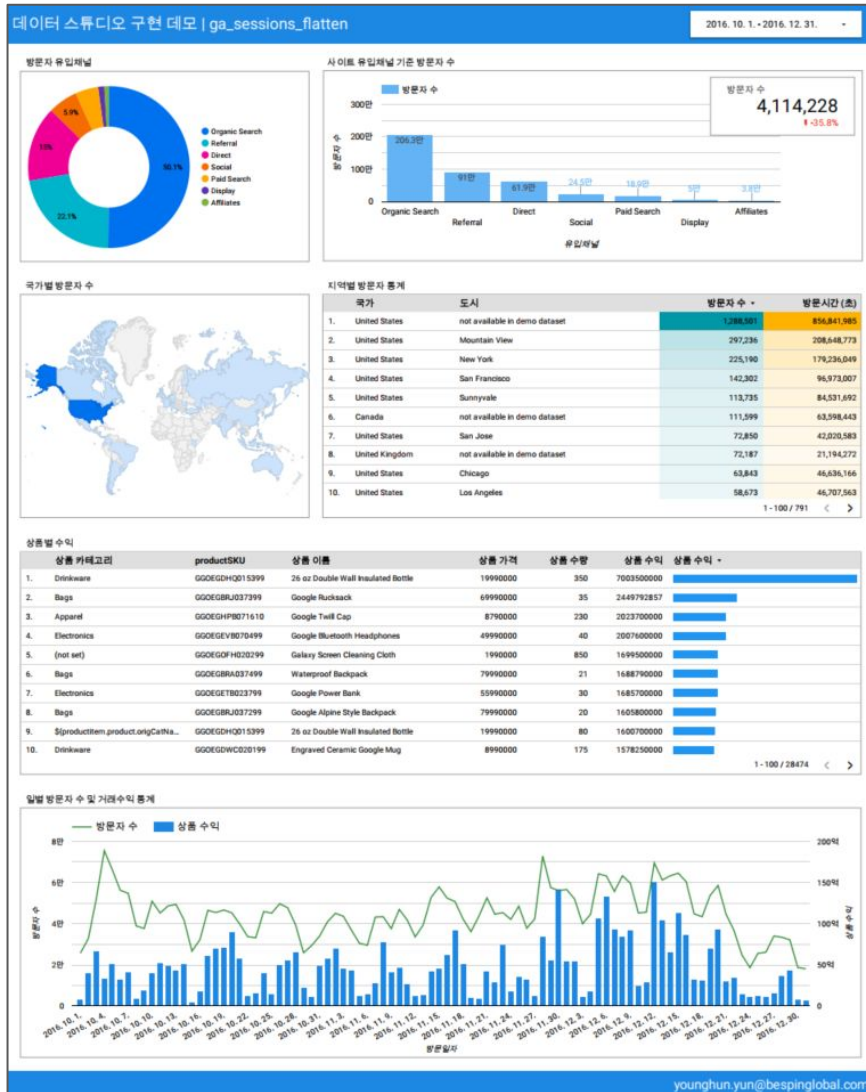
Labels None

Table info

Table ID	bigque
Table size	15.86 t
Long-term storage size	15.86 t
Number of rows	2,556

– 일자 기준으로 테이블이 분리되어 저장
– 테이블 수: 366개
– 수집 기간: 2016년 8월 1일 ~ 2017년 8월 1일
– 전체 데이터 크기: 5.38 GB

- 데이터 스튜디오를 이용하여 시각화된 대시보드를 구현합니다.




Google merchandise store
샘플 데이터를 이용하여
대시보드 구현 데모

<https://datastudio.google.com>

끝, 감사합니다:)

BESPIN GLOBAL  Google Cloud