빅쿼리를 활용한 빅데이터 분석 (10)

서진호

제 10 강 목표

1. 빅쿼리란 무엇인가?
2. 빅쿼리의 간단한 구조
3. 빅쿼리 SQL
4. 빅쿼리 데이터 타입

Cloud Innovator
MEGAZONE

빅쿼리(BigQuery) 주요 특징(1)

- 확장성이 뛰어난 구글 클라우드의 서버리스 기반의 데이터 분석 도구
- 관리할 인프라가 없기 때문에 데이터 분석에 집중할 수 있으며, 인프라와 데이터를 관리할 관리자도 필요하지 않다.
- 빅쿼리는 ANSI:2011을 준수하는 표준 SQL을 지원하기 때문에 기존에 SQL 을 알고 있는 사용자도 손쉽게 이용할 수 있다.
- ODBC 및 JDBC 드라이버를 제공하여 데이터를 쉽고 빠르게 통합할 수 있다.
- 몇 초 만에 기가바이트급에서 패타바이트급에 이르는 데이터를 대상으로 초고속으로 SQL 쿼리를 실행함..
- 매월 무료로 최대 1TB 상당의 데이터를 분석하고 10GB의 데이터를 저장할 수 있음.

빅쿼리(BigQuery) 주요 특징(2)

- 스토리지와 컴퓨팅(연산)이 분리되어 있기 때문에 데이터 웨어하우스의 용량을 원하는 대로 계획할 수 있는 탄력적인 확장성을 가짐.
- 자동확장과 고성능 스트리밍 수집 방식을 지원해서 실시간분석의 어려움도 간편하게 해결할 수 있음.
- 내부적으로 관리형 열 형식 스토리지, 대량 동시 실행, 자동 성능 최적화 기능을 제공하고 있어서 데이터 크기에 관계없는 클라우드 데이터 레이크를 구축 및 빠르게 분석함.
- 구글 클라우드 스토리지와 구글 시트, 구글 드라이브 등으로 손쉽게 데이터를 읽을 수 있으며 인포매티카와 탈랜드 같은 기존 ETL 도구와의 연동도 지원함.
- 태블로, 마이크로스트레티지, 루커, 데이터 스튜디오와 같은 BI 도구와 자체적으로도 BI 엔진을 지원하여 누구나 손쉽게 보고서와 대시보드를 만들 수 있음.

빅쿼리(BigQuery) 주요 특징(3)

- 모든 배치와 스트리밍 데이터를 분석함
- 강력한 스트리밍 수집 기능은 실시간으로 데이터를 캡처하고 분석해, 통계를 항상 최신 상태로 유지함.
- 데이터세트, 쿼리, 스프레드시트, 보고서로 조직 안팎에서 유용한 정보를 안전하게 공유할 수 있음.
- 최근 릴리즈된 빅쿼리ML을 이용하면 SQL 쿼리를 통해 ML 모델을 학습시키는 것이 가능하며, 클라우드 ML 엔진 및 텐서플로와도 통합이 가능함.
- 빅쿼리 GIS를 이용하면 일반적으로 GIS 함수에 대한 SQL지원을 빅쿼리 내에서 이용할 수 있음.

빅쿼리(BigQuery) 구조

- 프로젝트(Project):
  - 프로젝트에는 결제 및 승인된 사용자에 대한 정보가 저장됨.
  - 각 프로젝트에는 프로젝트명과 프로젝트ID가 존재함.
  - 하나의 프로젝트에는 여러 개의 데이터셋을 포함할 수 있음.

- 데이터셋(Dataset):
  - 관계형 데이터베이스 시스템의 테이터베이스(Database)와 같은 개념
  - 데이터셋은 특정 프로젝트에 포함되며, 테이블과 뷰에 대한 액세스를 구성하고 제어하는 데 사용함.
  - 하나의 데이터셋에는 여러 개의 테이블을 가질 수 있음.

- 뷰(View):SQL 쿼리로 정의된 가상 테이블
- 잡(Job): 쿼리, 데이터로딩, 생성,삭제 등 작업에 대한 단위

빅쿼리(BigQuery) 구조

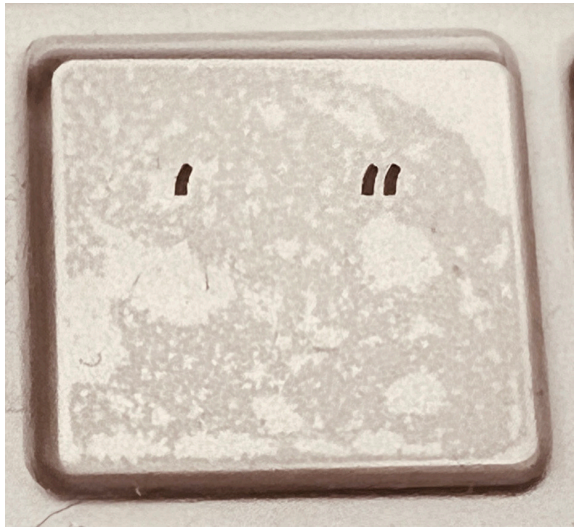
- Table:
  - 관계형 데이터베이스 시스템의 Table과 같은 개념
  - 행으로 구성된 개별 레코드가 포함함.
  - 각 레코드는 컬럼으로 구성되며, 모든 테이블은 컬럼명, 데이터 유형, 기타 정보를 설명하는 스키마로 정의함.

  - 빅쿼리에서 지원되는 테이블 유형
    - 기본테이블: 기본 빅쿼리 리포지토리에서 지원되는 테이블
    - 외부테이블: 빅쿼리 외부 리포지토리에서 지원되는 테이블

빅쿼리(BigQuery) SQL

Standard SQL	Legacy SQL
‘프로젝트명,데이터셋명.테이블명’ Select * from ‘my-project.my-dataset.my-table’	[프로젝트명:데이터셋.테이블명] Select * from [my-project:my-dataset.my-table]
‘with’절 사용 가능	
DML(Insert, Update, Delete) 사용	
Array 및 Struct 데이터 타입 사용 가능	
더 엄격한 TIMESTAMP 값의 범위	
모든 위치에서 서브 쿼리(SubQuery) 지원	

프로젝트 사용시 주의점



홀따옴표, Single Quotation (X)

백틱, Backtick(O)

참고: https://cloud.google.com/bigquery/docs/reference/standard-sql/lexical?hl=ko

표준 SQL 데이터 타입

타입	데이터 타입	설명
숫자	INT64	정수 (범위: -9,223,372,036,854,775,808~9,223,372,036,854,775,807)
	NUMERIC	좀 더 정밀한 숫자
	FLOAT64	부동 소수점, 배정밀도 십진수 값
부울	BOOL	TRUE or FALSE (대소문자 구분하지 않음)
문자	STRING	유니코드 데이터
	BYTES	가변 길이 바이너리 데이터
	DATE	날짜 (범위: 0001-01-01, 9999-12-31)
	DATETIME	날짜+시간 (범위: 0001-01-01 00:00:00, 9999-12-31 23:59:59.999999)

참고: https://cloud.google.com/bigquery/docs/reference/standard-sql/data-types/

표준 SQL 데이터 타입

타입	데이터 타입	설명
	TIME	시간
	TIMESTAMP	MS 단위의 절대 시점 값. 웹로그 또는 시계열 데이터 사용
	ARRAY	동일한 자료형의 리스트
	STRUCT	서로 다른 자료형의 리스트

참고: https://cloud.google.com/bigquery/docs/reference/standard-sql/data-types/

표준 SQL 데이터 타입 유의점

속성	설명	적용 대상
Nullable	Null 허용	ARRAY는 NULL 불가
Orderable	ORDER BY에서 사용	ARRAY, STRUCT 사용 불가
Groupable	GROUP BY, DISTINCT, PARTITION By 에서 사용	ARRAY, STRUCT 사용 불가
Comparable	동일한 유형의 값을 서로 비교	ARRAY 사용 불가 STRUCT '=' 만 필드 순서로 지원 가능