빅쿼리를 활용한 빅데이터 분석 (8)

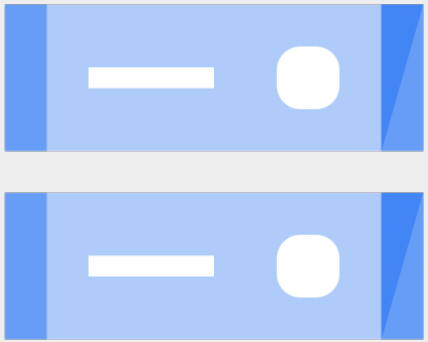
서진호

제 8 강 목표

1.구글 클라우드 스토리지 개념
2. 구글 클라우드 스토리지 특징
3. 리포지토리 등급 차이 비교
4. gsutil 명령어 사용법

클라우드 스토리지(GCS)란?

- 구글 대표적인 객체 리포지토리
- 빅데이터 분석에서 빈번이 사용함
- 데이터량에 관계없이 언제, 어디서나 데이터를 저장하고 가져 올 수 있음(불러 올 수 있음)
- 용도: 컨텐츠를 제공하거나 백업 테이프를 저장하거나 사용자에게 대량의 데이터 객체를 배포
- AWS S3와 유사한 서비스



GCS 주요 개념 (1)

개념	설명
프로젝트	GCP의 프로젝트와 동일한 개념으로 GCS의 모든 데이터는 프로젝트에 속함..
버킷	• 버킷은 GCS에서 저장하는 모든 데이터들은 버킷에 포함. • 중복 버킷 생성 불가(버킷 안에 또 다른 버킷 만들 수 없음) • 버킷명은 전역, 고유하기 때문에 GCS 전체 버킷명을 고려 • 프로젝트마다 2초당 작업 1개라는 제한이 버킷 생성 및 삭제에 적용 • 버킷 수는 적고 객체가 많도록 설계 추천 • 버킷과 컨텐츠가 저장되는 지리적 위치 및 리포지토리 등급을 지정 • 레이블을 달 수 있으면 라벨의 최대 갯수는 버킷당 64개 • 키값(Key-Value) 형식으로 GCP의 다른 리소스와 그룹화 할 수 있음.
리포지토리 등급	• 버킷을 생성할 때 지정할 데이터의 특징에 따라 리포지토리 등급이 나눔. • 멀티 리전, 리전, 니어라인(Nearline), 콜드라인(Coldline)

GCS 주요 개념 (2)

개념	설명
객체	• 객체는 버킷에 저장하는 파일을 말함. • 객체는 객체 데이터와 객체 메타데이터로 나눔. • 객체 데이터는 일반적으로 GCS에 저장되는 파일을 말함. • 객체 메타 데이터는 키-값 형태로 구성이 되며 다양한 객체의 퀄리티 설명 담당. • 버킷에서 만들 수 있는 객체 수에는 제한이 없음. • 객체 이름은 유니코드 문자 조합(UTF-8 인코딩)을 포함함. • 길이는 1,024 바이트를 초과할 수 없음. • 객체 이름에 포함되는 일반 문자는 슬래시(/)를 사용하면 디렉토리 구조가 없는 GCS에서 디렉토리
지리적 중복	• 지리적 중복 데이터는 최소 100마일 이상 떨어진 두 곳 이상의 중복 저장 됨. • 자연 재해와 같은 대규모 장애 발생시에도 최대한의 데이터 가용성을 보장함. • 다중 지역 위치에 저장된 객체는 리포지토리 등급에 관계없이 지리적으로 중복 가능됨. • 지리적 중복성은 비동기적으로 발생하지만, 모든 GCS는 사용자가 업로드하는 즉시 최소 한 곳 이상 지리적 장소 내에 중복됨
객체 불변성	• 버킷을 생성할 때 지정할 데이터의 특징에 따라 리포지토리 등급이 나눔. • 멀티 리전, 리전, 니어라인(Nearline), 콜드라인(Coldline)

리포지토리 등급 차이 비교 (1)

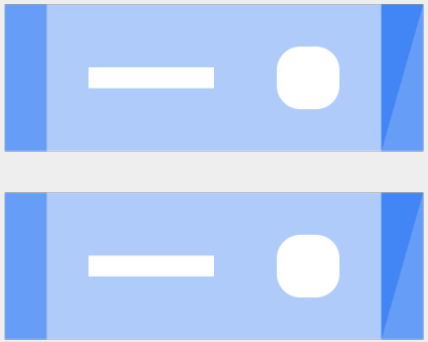
리포지토리 등급	특성	사용 사례	비용(1GB당 월별 과금)
멀티리저널 스토리지	• >99.99% 월간 평균 가용성 • 99.95% 가용성 SLA • 지리적 중복	웹 사이트 콘텐츠, 비디오 스트리밍 또는 게임 및 모바일 앱 등 전세계적으로 자주 데이터 접근함.	$0.026
리저널 스토리지	• 99.99% 월간 평균 가용성 • 99.95% 가용성 SLA • 저장된 GB 대비 낮은 비용 • 좁은 지역에 데이터 저장 • 가용성 영역 내에서 중복	• 데이터 분석과 같이 자주 사용하는 구글 클라우드 DataProc • 구글 컴퓨트 엔진 인스턴스와 동일한 지역에 접근해서 저장해 데이터 집약적인 컴퓨트 수행 시 높은 성능의 장점을 가질 수 있음	$0.02

리포지토리 등급 차이 비교 (2)

리포지토리 등급	특성	사용 사례	비용(1GB당 월별 과금)
니어라인 스토리지	• 99.9% 월간 평균 가용성 • 99.0% 가용성 SLA • 저장된 GB 대비 낮은 비용 • 데이터 검색 비용 • 높은 운영 비용 • 30일 최소 저장 기간	• 자주 액세스하지 않을 데이터, 예) 한 달에 한번. • 백업 및 지연 시간이 긴 멀티미디어 콘텐츠에 적합	$0.026
리전 스토리지	• 99.9% 월간 평균 가용성 • 99.0% 가용성 SLA • 저장된 GB 대비 낮은 비용 • 데이터 검색 비용 • 높은 운영 비용 • 90일 최소 저장 기간	• 데이터 분석과 같이 자주 사용하는 구글 클라우드 DataProc • 구글 컴퓨트 엔진 인스턴스와 동일한 지역에 접근해서 저장해 데이터 집약적인 컴퓨트 수행 시 높은 성능의 장점을 가질 수 있음	$0.02

스토리지 클래스

- Standard: 정기적 액세스, 최소 보관 기간 없음, '핫' 데이터
- Nearline: 한 달에 한 번 미만 액세스, 한 달 이상 보관
- Coldline: 한 분기에 한 번 미만 액세스, 한 분기 이상 보관
- Archive: 일 년에 한 번 미만 액세스, 일 년 이상 보관



gsutil에 대하여

synabreu — -zsh — 138×54

~ — -zsh

```
Last login: Wed Jul 27 17:18:35 on ttys000
(base) synabreu@synabreu ~ % gsutil --help
Usage: gsutil [-D] [-DD] [-h header]... [-i service_account] [-m] [-o section:flag=value]... [-q] [-u user_project] [command [opts...] arg
s...]
Available commands:
  acl              Get, set, or change bucket and/or object ACLs
  autoclass        Configure autoclass feature
  bucketpolicyonly Configure uniform bucket-level access
  cat              Concatenate object content to stdout.
  compose          Concatenate a sequence of objects into a new composite object.
  config           Obtain credentials and create configuration file
  cors             Get or set a CORS JSON document for one or more buckets
  cp               Copy files and objects
  defacl           Get, set, or change default ACL on buckets
  defstorageclass  Get or set the default storage class on buckets
  du               Display object size usage
  hash             Calculate file hashes
  help             Get help about commands and topics
  hmac             CRUD operations on service account HMAC keys.
  iam              Get, set, or change bucket and/or object IAM permissions.
  kms              Configure Cloud KMS encryption
  label            Get, set, or change the label configuration of a bucket.
  lifecycle        Get or set lifecycle configuration for a bucket
  logging          Configure or retrieve logging on buckets
  ls               List providers, buckets, or objects
  mb               Make buckets
  mv               Move/rename objects
  notification     Configure object change notification
  pap              Configure public access prevention
  perfdiag         Run performance diagnostic
  rb               Remove buckets
  requesterpays    Enable or disable requester pays for one or more buckets
  retention        Provides utilities to interact with Retention Policy feature.
  rewrite          Rewrite objects
  rm               Remove objects
  rpo              Configure replication
  rsync            Synchronize content of two buckets/directories
  setmeta          Set metadata on already uploaded objects
  signurl          Create a signed URL
  stat             Display object status
  test             Run gsutil unit/integration tests (for developers)
  ubla             Configure Uniform bucket-level access
  update           Update to the latest gsutil release
  version          Print version info about gsutil
  versioning       Enable or suspend versioning for one or more buckets
  web              Set a website configuration for a bucket

Additional help topics:
  acls             Working With Access Control Lists
  crc32c           CRC32C and Installing crcmod
  creds            Credential Types Supporting Various Use Cases
  dev              Contributing Code to gsutil
  encoding         Filename encoding and interoperability problems
  metadata         Working With Object Metadata
```

gs://[버킷 이름]/[객체 이름]