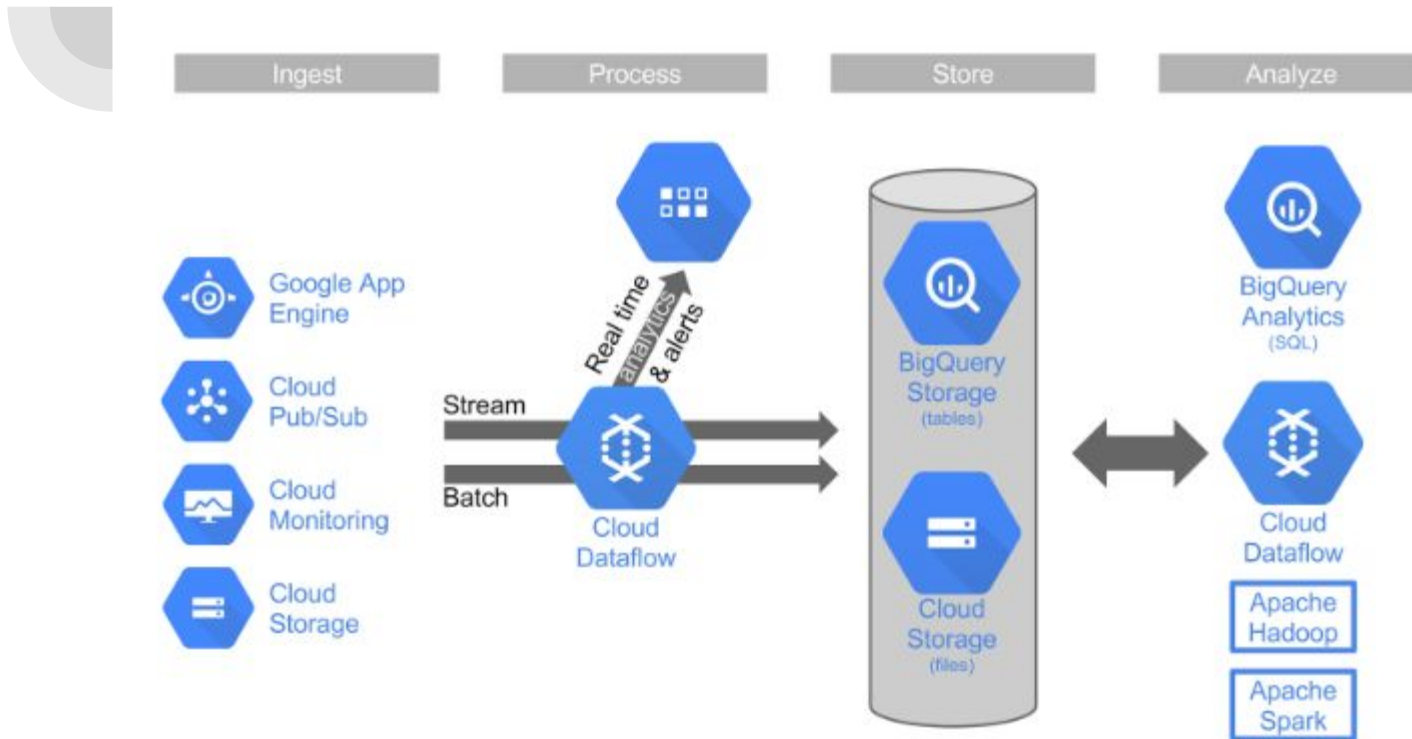


구글 클라우드 빅데이터 플랫폼

데이터 처리 Cycle



1. Google Cloud 빅데이터 플랫폼

Cloud Dataflow

데이터 파이프 라인

클러스터는 GCE 가상머신에 생성됨 , 오토스케일링 가능

크기를 알고 있는 데이터셋이 있을때 사용가능

매니지드 방식으로 리소스 관리 용이 (필요한 리소스 자동 배치)

Write Code -> Batch -> Streaming

복합적 프로그래밍 모델 (ETL, 배치연산, 지속적연산)

Cloud DataProc

Cloud Dataproc은 클라우드 네이티브 아파치 하둡 및 아파치 스파크 서비스 (완전 관리형 클라우드 서비스)

90초 이내에 클러스터 생성, 클러스터는 GCE 가상 머신에 생성됨(GKE 와 같은 방식)

클러스터는 Job이 구동 중에도 스케일링 가능

스택드라이버로 모니터링

Cloud Storage 등에 적재된 (로그성) 데이터의 빠른 분석 가능

데이터 마이닝과 분석에 Spark SQL 사용

MLib(Spark 머신러닝 라이브러리)를 분류 알고리즘에 사용

Dataflow	Dataproc
Apache Beam 기반	Apache Hadoop / Spark 기반
Serverless	DevOps
기존에 레거시 없이 새로 접근할 때 적합	Apache 빅데이터 생태계(=Hadoop Eco System)에 적합

[표 15-3] Dataflow vs Dataproc

1. Google Cloud 빅데이터 플랫폼



BigQuery

매니지드 데이터웨어하우스(DW)

실시간의 대량의 데이터 분석

SQL 2011 문법 사용

클러스터 관리가 필요없음

데이터 ingestion

데이터 소스의 위치와 관계없이 글로벌 가용성 제공

타 구글 제품과 연동 가능

분석, OLAP에 특화되어있기 때문에 OLTP에는 부적합.

적은 양의 데이터를 계속해서 입력하고 업데이트 해야하는 작업에서는 오히려 성능이 떨어지는 모습을 보임.

최근 릴리즈된 BigQuery ML을 이용하면 SQL 쿼리를 통해 ML 모델을 학습시키는 것이 가능 (간단한 회귀분류)

Cloud Datalab

대량의 데이터 조회, 변환, 분석 시각화를 위한 인터랙티브 도구

Jupyter Notebooks 기반으로 사용

GCE 가상머신에서 작동, 머신타입과 생성 리전 선택필요

메가바이트 또는 테라바이트 단위의 분석 처리 가능

Big Query에서 테라바이트 단위의 데이터를 쿼리하고 샘플데이터의 로컬분석을 실행,

AI Platform에서 테라바이트 단위의 데이터에 관한 학습 가능

현재는 Datalab보다 AI Platform Notebooks를 주요 서비스로 제공하고 있음.

Google Cloud AI Platform Notebooks 는 사실상 Google Cloud Datalab 의 업그레이드 된 버전이며 ssh 터널을 설정하지 않고도 브라우저에서 직접 노트북 을 사용할 수 있다는 이점을 제공

2. Cloud AI Platform

시각



비전

클라우드나 에지의 이미지에서 유용한 정보를 도출합니다.



동영상

강력한 콘텐츠 탐색 기능과 매력적인 동영상 환경을 지원합니다.

대화



Dialogflow

가상 에이전트 및 기타 대화형 환경을 빌드합니다.



Cloud Text-to-Speech API

WaveNet 음성을 사용해 텍스트를 자연스러운 음성으로 변환합니다.



Cloud Speech-to-Text API

자동으로 음성을 텍스트로 정확하게 변환합니다.

언어



Translation

언어를 동적으로 감지하고 각 언어로 번역합니다.



Natural Language

머신러닝을 통해 텍스트의 구조와 의미를 드러냅니다.

구조화된 데이터



AutoML Tables

구조화된 데이터를 사용하는 최신 머신러닝 모델을 자동으로 빌드하고 배포합니다.



Recommendations AI

고도로 맞춤화된 제품을 대규모로 추천합니다.



Cloud Inference API

입력된 시계열 데이터셋에 대한 대규모 상관관계 분석을 신속하게 수행할 수 있습니다.

2. Cloud AI Platform



- 신경망 모델을 빌드 및 실행하는 오픈 소스 도구
 - 폭넓은 플랫폼 지원 : CPU 또는 GPU, 모바일, 서버, 클라우드
- 완전 관리형 머신러닝 서비스
 - 익숙한 메모장 기반 개발자 환경
 - Google 인프라에 최적화, BigQuery 및 Cloud Storage와 통합
- Google에서 빌드한 선행 학습된 머신러닝 모델
 - 음성: 결과를 실시간으로 스트리밍 하고, 80개의 언어를 이해함
 - 비전: 객체, 랜드마크, 텍스트, 콘텐츠 식별
 - 번역: 언어 감지 및 번역
 - 자연어: 텍스트의 의미, 구조
- 구조화된 데이터 : 분류 및 회귀 / 추천 / 이상감지
- 구조화되지 않은 데이터 : 이미지 및 동영상 분석 / 텍스트 분석



Vision AI

AutoML Vision을 사용하여 클라우드나 에지 이미지에서 유용한 정보를 도출하거나 선행 학습된 Vision API 모델을 사용하여 이모티콘 인식, 텍스트 이해 등을 수행하며 객체 자동인식, 데이터 라벨링 서비스 등을 제공하고 있다.

커스텀 라벨 생성을 필요로 하는 경우 AutoML Vision을 사용하면 된다.

해당 API를 사용하려면 현재 프로젝트 리소스가 **us-central1** 리전에 있어야 함.

AWS의 경우 비슷한 서비스를 제공하고 있고, MS Azure는 데이터라벨링서비스나 객체 자동인식 서비스를 제공하지 않고 있다.

AutoML Vision

자체 커스텀 머신러닝 모델의 학습을 자동화하세요. AutoML Vision의 사용하기 쉬운 그래픽 인터페이스로 간단하게 이미지를 업로드해 커스텀 이미지 모델을 학습시킬 수 있습니다. 정확성, 지연 시간, 크기에 맞춰 모델을 최적화한 후 클라우드의 애플리케이션 또는 에지의 다양한 기기로 배포할 수도 있습니다.

Vision API

Google Cloud의 Vision API는 REST 및 RPC API를 통해 선행 학습된 강력한 머신러닝 모델을 제공합니다. 이미지에 라벨을 할당하고 사전 정의된 수백만 개의 카테고리로 빠르게 분류할 수 있습니다. 객체와 얼굴을 인식하고 인쇄 및 필기 텍스트를 읽으며 이미지 카탈로그에 유용한 메타 데이터를 구축합니다.

AutoML Vision vs Vision API

	AutoML Vision	Vision API
사용자 인터페이스		
API 사용 REST 및 RPC API를 사용합니다.	✓	✓
그래픽 UI 사용 그래픽 사용자 인터페이스를 사용합니다.	✓	
사전 정의된 라벨 또는 커스텀 라벨		
사전 정의된 라벨을 사용한 이미지 분류 선행 학습된 모델에서는 사전 정의된 라벨이 포함된 방대한 라이브러리를 활용합니다.		✓
커스텀 라벨을 사용한 이미지 분류 선택한 라벨을 통해 이미지를 분류하도록 모델을 학습시킵니다.	✓	
Google의 데이터 라벨링 서비스 사용 Google팀에서 이미지, 동영상, 텍스트에 주석을 추가하도록 도와드립니다.	✓	✓
예지에 배포		
예지에 머신러닝 모델 배포 예지 기기에 최적화된 지면 시간이 짧고 정확성이 높은 모델을 배포합니다.	✓	ML Kit과 통합

추가 기능		
객체 인식 객체와 객체의 위치 및 개수를 인식합니다.	✓	✓
Vision 제품 검색 사용 설정 사진이 제품 카탈로그의 이미지와 일치하는지 비교하고 유사한 상품의 순위 목록을 반환합니다.		✓
인쇄 및 필기 입력 텍스트 인식 OCR을 사용해 언어를 자동으로 식별합니다.		✓
얼굴 감지 얼굴과 얼굴 속성을 감지합니다. (얼굴 인식은 지원되지 않습니다.)		✓
명소 및 제품 로고 식별 잘 알려진 명소와 제품 로고를 자동으로 식별합니다.		✓
일반 이미지 속성 할당 일반적인 속성과 적절한 자르기 힌트 를 인식합니다.		✓
웹 항목 및 페이지 인식 웹에서 유사 이미지, 뉴스 이벤트, 로고를 찾습니다.		✓
콘텐츠 검토 이미지 내에서 성인용 콘텐츠, 폭력적인 콘텐츠와 같은 유해성 콘텐츠 를 감지합니다.		✓
유명인 인식 이미지에서 유명인의 얼굴을 인식합니다.(제한된 액세스, 문서 참조).		✓



음성 및 텍스트 변환 서비스

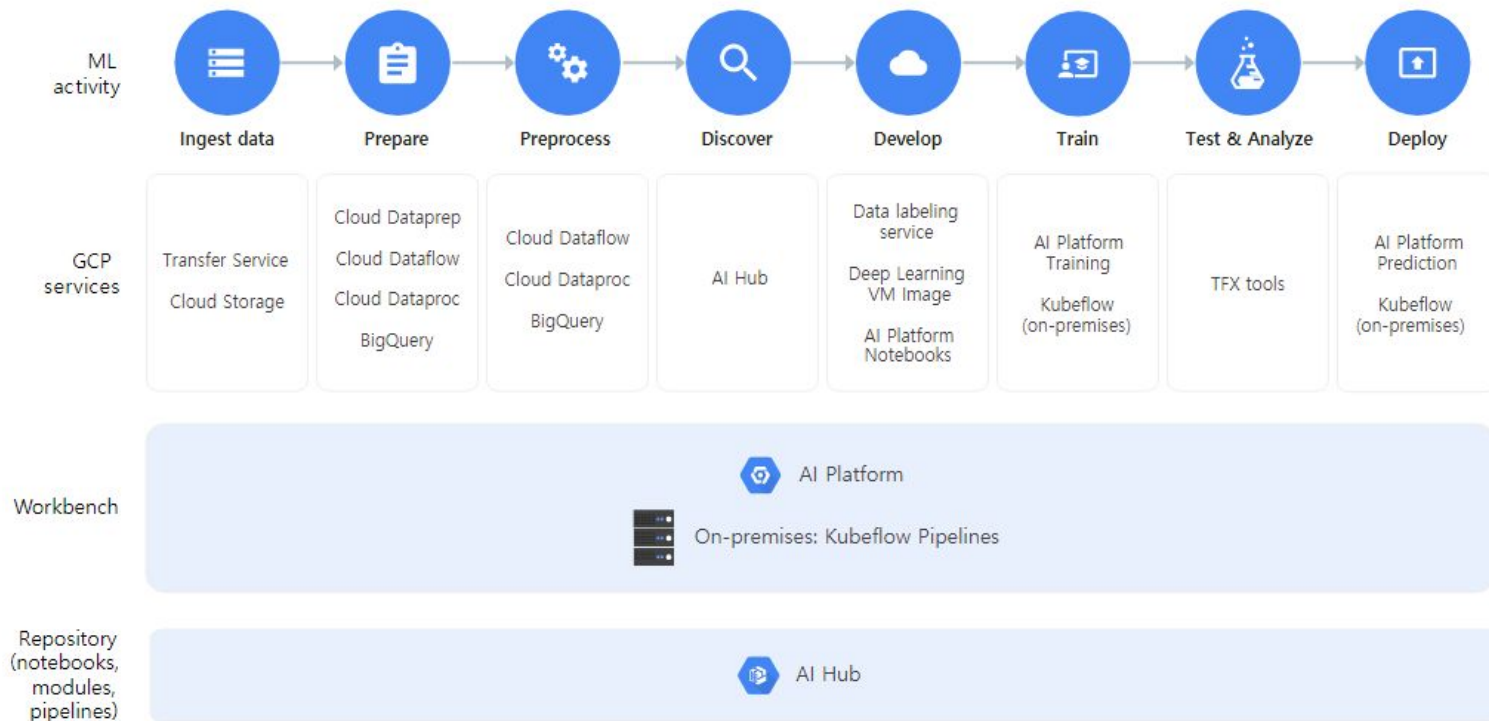
- 분야별 품질 요구사항에 따라 최적화된 음성 제어, 전화 통화, 동영상 텍스트 변환에 맞게 학습 모델 옵션을 선택적으로 사용
- **Dialogflow**
구글의 챗봇 플랫폼
- **Cloud Text-to-Speech API**
WaveNet 음성을 사용해 텍스트를 음성으로 변환.
- **Cloud Speech-to-Text API**
자동으로 음성을 텍스트로 정확하게 변환.



Cloud Inference API (시계열데이터 상관관계 분석 Alpha 버전)

- 실시간 분석 : 실시간 데이터 스트리밍으로 상관관계 컴퓨팅 가능
- 다른 **Google Cloud Storage** 서비스와 통합, 일관된 액세스 방법 제공
- 이상치 및 트렌드 감지
- 최대 수백억개의 이벤트로 구성된 데이터 세트처리
- 시계열 데이터 세트 처리 :
 - **JSON**에서 쿼리 효율적인 내부 형식으로 수집
 - 시스템에서 이전에 제출 한 데이터 세트 제거
 - 프로젝트에서 제출 한 시스템의 활성 데이터 세트 나열
- 로드 된 데이터 세트에 대한 추론 쿼리를 실행
 - 서로 다른 유형의 값은 어떻게 연관됩니까? 예를 들어 레이블이 지정된 뉴스 기사의 데이터 세트에서 휴가 관련 기사와 어떤 레이블이 상호 연관되어 있습니까?
 - 이벤트 빈도는 시간에 따라 어떻게 달라 집니까? 예 : 특정 주제와 관련된 이벤트가 비정상적으로 많은 날이있는 날은 언제입니까?
 - 시스템의 이벤트에 대한 배경 확률은 얼마입니까? 예 : 다양한 스포츠 이미지가 기사에 얼마나 자주 등장합니까?

머신러닝 개발: 엔드 투 엔드 주기



클라우드 서비스별 비교표

머신러닝서비스(분류, 회귀, 군집)

CLOUD MACHINE LEARNING SERVICES COMPARISON

	Amazon	Microsoft	Google	IBM
Automated and semi-automated ML services				
	Amazon ML	Microsoft Azure ML Studio	Cloud AutoML	IBM Watson ML Model Builder
Classification	✓	✓	✓	✓
Regression	✓	✓	✓	✓
Clustering	✓	✓	✗	✗
Anomaly detection	✗	✓	✗	✗
Recommendation	✗	✓	✓	✗
Ranking	✗	✓	✗	✗
Platforms for custom modeling				
	Amazon SageMaker	Azure ML Services	Google ML Engine	IBM Watson ML Studio
Built-in algorithms	✓	✗	✓	✓
Supported frameworks	TensorFlow, MXNet, Keras, Gluon, Pytorch, Caffe2, Chainer, Torch	TensorFlow, scikit-learn, Microsoft Cognitive Toolkit, Spark ML	TensorFlow, scikit-learn, Microsoft XGBoost, Keras	TensorFlow, Spark MLlib, scikit-learn, XGBoost, PyTorch, IBM SPSS, PMML

이미지분석

IMAGE ANALYSIS APIs COMPARISON

	Amazon	Microsoft	Google	IBM
Object Detection	✓	✓	✓	✓
Scene Detection	✓	✓	✓	✗
Face Detection	✓	✓	✓	✓
Face Recognition (person face identification)	✓	✓	✗	✗
Facial Analysis	✓	✓	✓	✓
Inappropriate Content Detection	✓	✓	✓	✓
Celebrity Recognition	✓	✓	✓	✗
Text Recognition	✓	✓	✓	✓
Written Text Recognition	✗	✓	✓	✗
Search for Similar Images on Web	✗	✗	✓	✗
Logo Detection	✗	✗	✓	✗
Landmark Detection	✗	✓	✓	✗
Food Recognition	✗	✗	✗	✓
Dominant Colors Detection	✗	✓	✓	✗

영상분석

VIDEO ANALYSIS APIs COMPARISON

	Amazon	Microsoft	Google
Object Detection	✓	✓	✓
Scene Detection	✓	✓	✓
Activity Detection	✓	✗	✗
Facial Recognition	✓	✓	✗
Facial and Sentiment Analysis	✓	✓	✗
Inappropriate Content Detection	✓	✓	✓
Celebrity Recognition	✓	✓	✗
Text Recognition	✓	✓	✗
Person Tracking on Videos	✓	✓	✗
Audio Transcription	✗	✓	✓
Speaker Indexing	✗	✓	✗
Keyframe Extraction	✗	✓	✗
Video Translation	✗	9 languages	✗
Keywords Extraction	✗	✓	✗
Brand Recognition	✗	✓	✗
Annotation	✗	✓	✗
Dominant Colors Detection	✗	✗	✗
Real-Time Analysis	✓	✗	✗

클라우드 서비스별 비교표



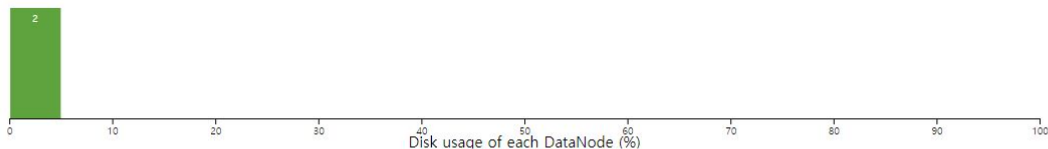
Service	GCP	AWS	Azure
Machine Learning	Cloud ML Engine	SageMaker	Machine Learning
	Cloud AutoML	Ground Truth	Azure Databricks
	Cloud TPU	AML	
		Apache MXNet on AWS	
		Tensorflow on AWS	
		Personalize	
		Forecast	
		Elastic Inference	
		DeepRacer	
Cognitive Services	Cloud Natural Language	Comprehend	Cognitive Services
	Cloud Speech-toText	Lex	
	Cloud Text-toSpeech	Polly	
	Cloud Vision	Rekognition	
	Cloud Translation	Translate	
	Cloud Video Intelligence	Transcribe	
		DeepLens	
		Textract	
Big Data Analytics	Cloud Dataflow	Redshift	HDInsight
	Cloud Dataproc	Athena	Stream Analytics
	Cloud Dataprep	EMR	Data Lake Analytics
	BigQuery	Kinesis Data Stream	Analysis Services
		Kinesis Firehose	Azure Data Explorer
		Kinesis Analytics	
		Glue	
		Data pipeline	
		Lake Formation	
		QuickSight	

Dataproц 실습 (hadoop)

Datanode Information

✓ In service ⬇ Down ⚡ Decommissioned ⚡ Decommissioned & dead ⚡ In Maintenance & dead

Datanode usage histogram



In operation

Show 25 entries

Search:

Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool used	Version
✓ test-cluster-w-0.c.sonorous-nomad-287601.internal:9866 (10.182.0.3:9866)	http://test-cluster-w-0.c.sonorous-nomad-287601.internal:9864	0s	113m	491.97 GB <div><div></div></div>	3	132.68 KB (0%)	2.9.2
✓ test-cluster-w-1.c.sonorous-nomad-287601.internal:9866 (10.182.0.4:9866)	http://test-cluster-w-1.c.sonorous-nomad-287601.internal:9864	0s	113m	491.97 GB <div><div></div></div>	3	140 KB (0%)	2.9.2

Showing 1 to 2 of 2 entries

Previous 1 Next


Dataproц 실습 (hadoop)

☰

Google Cloud Platform

● My First Project ▼

🔍 제품 및 리소스

 Dataproц

클러스터

+

 클러스터 만들기

↺

 새로고침

🗑

 삭제

리전 ▼

클러스터

작업

워크플로

자동 확장 정책

구성요소 교환

메모장

☰

클러스터를 검색하려면 Enter 키를 누르세요.

<input type="checkbox"/>	이름 ↑	리전	영역	총 워커 노드 수	예약된 삭제
<input type="checkbox"/>	✔ test-cluster	global	us-west4-b	2	사용 안함