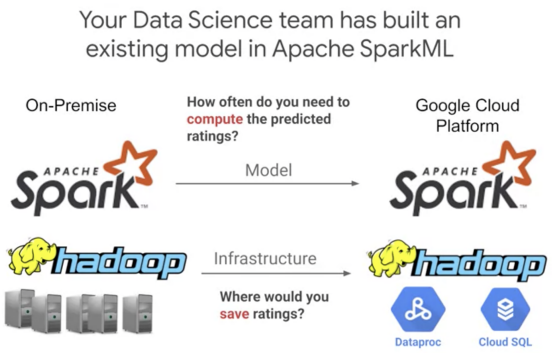
Dataproc for Spark ML



- Spark 컴퓨팅 작업은 Dataproc으로, 작업결과는 Cloud SQL RDBMS에 저장
- Dataproc은 일괄 처리, 쿼리, 스트리밍, 머신 러닝에 오픈소스 데이터 도구를 활용할 수 있는 관리형 Spark 및 Hadoop 서비스입니다. Dataproc 자동화를 통해 신속하게 클러스터를 만들고 손쉽게 관리하며 불필요한 클러스터를 사용 중지하여 비용을 절감할 수 있습니다. 관리 시간과 비용이 절감되므로 작업과 데이터에 집중할 수 있습니다.
  - Created a fully-managed Cloud SQL instance for rentals
  - Created tables and explored the schema with SQL
  - Ingested data from CSVs
  - Edited and ran a Spark ML job on Dataproc
  - Viewed prediction results

1. Hadoop에 대한 지원이 기본 제공.
2. GCP의 완전 관리형 서비스입니다.
   - 즉, 하드웨어 또는 소프트웨어 업데이트 및 설치에 대해 걱정할 필요가 없습니다. 모든 작업이 완료되고 관리됩니다. 또한 더 큰 클러스터가 필요한 경우 더 많은 기계를 기다리거나 주문할 필요가 없습니다. 간단히 추가하거나 단 몇 분 만에 Cloud Dataproc UI를 통해 클러스터의 노드를 제거할 수 있습니다.
3. 버전관리 시스템

4. 일반적인 온프레미스 Hadoop 설정은 다양한 용도로 사용되는 단일 클러스터를 사용합니다. GCP로 이동하면 필요한 만큼 클러스터를 생성하여 개별 작업에 집중할 수 있습니다.

Using Cloud Dataproc

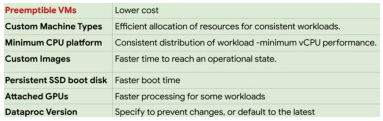
Setup

- 기존 클러스에서 YAML 파일을 내보낸다.
- Deployment Manager
- Rest API

Configuration

- Cluster Options
- Master node
- Worker nodes
- Preemptible nodes
  : Yarn node manger 는 있으나 HDFS를 실행하지 않음

Optimization



Preemptible VMs	Lower cost
Custom Machine Types	Efficient allocation of resources for consistent workloads.
Minimum CPU platform	Consistent distribution of workload -minimum vCPU performance.
Custom Images	Faster time to reach an operational state.
Persistent SSD boot disk	Faster boot time
Attached GPUs	Faster processing for some workloads
Dataproc Version	Specify to prevent changes, or default to the latest

Utilize

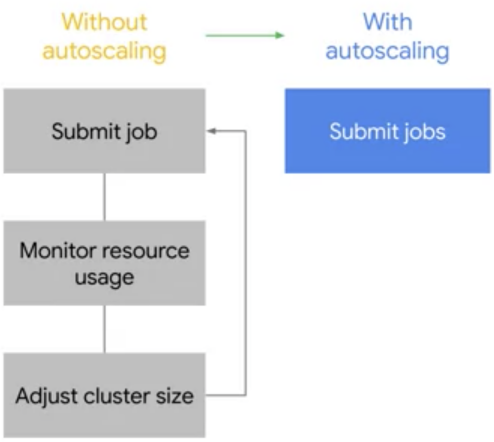
- Submit a job

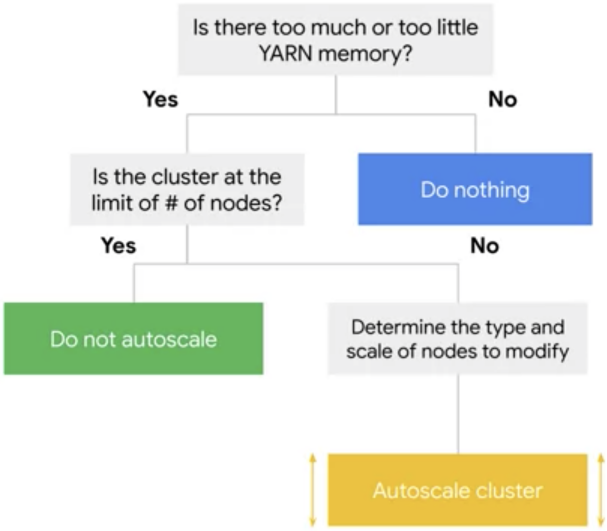
Monitor

- Stackdriver Monitoring
- CPU
- Disk
- HDFS/YARN

Autoscaling

Cloud Dataproc autoscaling workflow





- 클러스터 자동 종료
- Flexible