Google Cloud Dataproc

Google Cloud Dataproc is a managed Spark and Hadoop service that lets you take advantage of open source data tools for batch processing, querying, streaming, and machine learning.

Credit to reddit user lpetrazickis

Features

- Fast, easy, managed way to run Hadoop and Spark/Hive/Pig.
- Benefit from cloud integration (Storage, Stackdriver, etc).
- Customize and configure clusters with initialization actions.
- Create clusters in **90 sec or less**.
- Pay-per-minute billing.
- Scale clusters up and down even when jobs are running.
- Tools including RESTful API and GCP SDK integration.
- Can be managed using Hadoop native tools such as:
  - YARN Web UI.
  - HDFS Web UI.
  - SSH.
  - SOCKS.

When to Use

- Migrate on-prem Hadoop jobs to the cloud.
- Analyze data stored in Cloud Storage.
- Use Spark/Spark SQL to quickly perform data mining and analysis.
- Use Spark Machine Learning Libraries for classification models.

ELI5

Hadoop is an open-source implementation of MapReduce.

What is MapReduce?

> MapReduce is a software approach for cutting problems like this one into smaller problems, mapping each sub-problem to a different processor (usually different machines on a network) and then reducing each intermediate answer to the single final answer you're looking for. Not all problems can be distributed out to separate solvers, but for the ones that can be, you can gain a speedup of several orders of magnitude when compared to the classic, single-processor approach most of us are used to.

Credit to reddit user alk509

What is Spark?

> Spark is a framework for efficiently processing large amounts of data in parallel. It has built-in libraries for machine learning and other statistical analysis. It can be applied for data journalism, business analysis, or any other data science field. In my experience, Spark is both much easier to work with and more efficient than traditional MapReduce or Hadoo, so Spark is displacing Hadoop within the Hadoop family of tools. Spark is also language agnostic, so you can apply your existing Python, R, SQL, or Scala skills with equal efficiency.

What is Hive?

> Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data summarization, query and analysis. Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop. Traditional SQL queries must be implemented in the MapReduce Java API to execute SQL applications and queries over distributed data. Hive provides

the necessary SQL abstraction to integrate SQL-like queries (HiveQL) into the underlying Java without the need to implement queries in the low-level Java API.

Credit Wikipedia

What is Pig?

Apache Pig is a high-level platform for creating programs that run on Apache Hadoop. The language for this platform is called Pig Latin. Pig can execute its Hadoop jobs in MapReduce, Apache Tez, or Apache Spark. Pig Latin abstracts the programming from the Java MapReduce idiom into a notation which makes MapReduce programming high level, similar to that of SQL for relational database management systems. Pig Latin can be extended using user-defined functions (UDFs) which the user can write in Java, Python, JavaScript, Ruby or Groovy and then call directly from the language.

Credit Wikipedia