

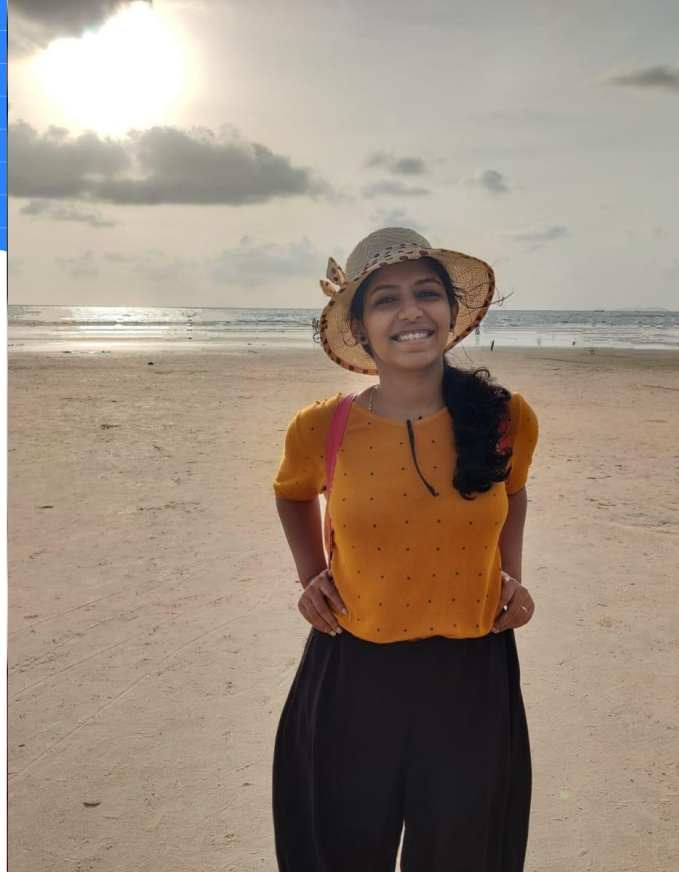
devfest 2022

```
book-nav-toggle'  
dden="" fixed="" aria-label='Hide  
Hide side navigation'
```

Make your Data pipelines easy with Dataproc and Composer



Google Developer Groups
Vancouver



Julian Sara Joseph
Data Analytics DRE, Google Cloud

In this session



Dataproc

- Personal Story
- Serverless
- Demo on Serverless Spark



Composer

- DAGs and Operators
- Demo - Composer for Dataproc Serverless Workloads



Q & A

A walk down the memory lane

2016

- Data Engineer in a **small** product team - TrialRun
- 1st Client's **Data was in TBs** and it was sent **weekly**
- ETL - It was our job to create deltas and make aggregates before storing the data for our DS team.

2018

- 2nd client - **Data was in PBs**
- Our Spark jobs needed varying levels of compute power
- Still manually triggering cron jobs weekly to start the Spark jobs 😓

That would be me every Monday
#TrueStory



We needed :

01 | **Auto-scaling clusters/nodes**
Cost-effective (Small budget team)

02 | **Automation of our cron jobs**

devfest
2022



Enter Dataproc

Fully managed Spark and Hadoop service

Seamless AI/ML



Open source - no lock in

Flexibility - GCE, GKE, Serverless



Higher Scale & Productivity

Low Cost

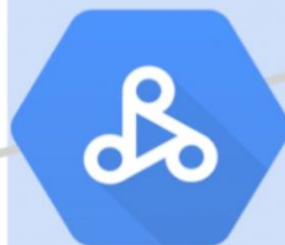
Low Rate & Billed By The Minute

Less Waiting

Clusters Start & Stop in < 90 Sec

Minimal Ops

We Manage Spark and Hadoop



Google Cloud Dataproc

Easier To Use

Manage Jobs With Easy Tools

Up To Date

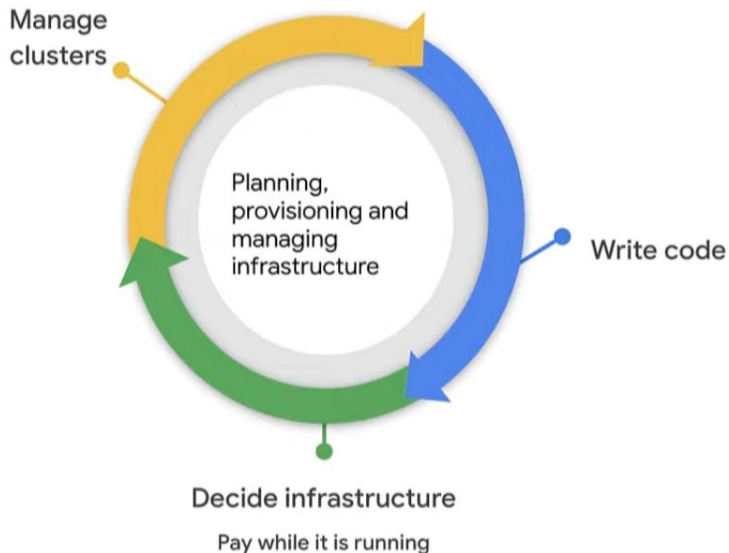
Recent Spark and Hadoop Releases

Cloud Native

Integrated w/ Google Cloud Products

Lower Cost & Complexity

What's better than Dataproc ? Serverless Spark!



- ✓ Job auto-scales
- ✓ No infrastructure to tune
- No clusters to manage
- Only pay for the job duration

Focus on Spark, not infrastructure

Demo running a **Serverless Spark job** from GUI for a simple **Spark Transformation job** (3 min)

Dive deeper with codelabs



Head over to g.co/codelabs/dataproc-serverless to gain hands-on experience

Dataproc Serverless

🕒 37 mins remaining 🌐 English ▼

1 Overview - Google Dataproc

2 Set up

3 Run Serverless Spark jobs with Dataproc Batches

4 Dataproc Metrics and Observability

5 Dataproc Templates: CSV -> GCS

6 Dataproc Templates: CSV to parquet

7 Clean up resources

8 What's next

Dataproc Serverless

About this codelab

📅 Last updated Oct 12, 2022

👤 Written by Brad Miro

1. Overview - Google Dataproc

Dataproc is a fully managed and highly scalable service for running Apache Spark, Apache Flink, Presto, and many other open source tools and frameworks. Use Dataproc for data lake modernization, ETL / ELT, and secure data science, at planet scale. Dataproc is also fully integrated with several Google Cloud services including [BigQuery](#), [Cloud Storage](#), [Vertex AI](#), and [Dataplex](#).

Dataproc is available in three flavors:

- [Dataproc Serverless](#) allows you to run PySpark jobs without needing to configure infrastructure and autoscaling. Dataproc Serverless supports PySpark batch workloads and sessions / notebooks.
- [Dataproc on Google Compute Engine](#) allows you to manage a Hadoop YARN cluster for YARN-based Spark workloads in addition to open source tools such as Flink and Presto. You can tailor your cloud-based clusters with as much vertical or horizontal scaling as you'd like, including autoscaling.
- [Dataproc on Google Kubernetes Engine](#) allows you to configure Dataproc virtual clusters in your GKE infrastructure for submitting Spark, PySpark, SparkR or Spark SQL jobs.

Dataproc Serverless

37 minutes Updated October 12, 2022

In this codelab, you'll learn all about Dataproc Serverless, including how to get started and how to access its rich featureset.

</>

Start

Dataproc on Google Compute Engine

16 minutes Updated October 7, 2022

In this codelab, you will learn about using Dataproc on Google Compute Engine (GCE).

</>

Start

Preprocessing BigQuery Data with PySpark on Dataproc

42 minutes Updated January 24, 2022

This lab shows you how to use PySpark on Dataproc to load data from BigQuery and save it to Google Cloud Storage.

Start

Apache Spark and Jupyter Notebooks on Cloud Dataproc

52 minutes Updated June 25, 2021

Create Spark ML models with Google Dataproc

31 minutes Updated October 12, 2022

In this codelab, you'll submit Spark ML jobs to Google's Dataproc service.

</>

Start

Provisioning and Using a Managed Hadoop/Spark Cluster with Cloud Dataproc (Command Line)

20 minutes Updated May 2, 2022

In this codelab, you will learn how to start a managed Spark/Hadoop cluster using Dataproc, submit a sample Spark job, and shut down your cluster using the command line.

Start

PySpark for Natural Language Processing on Dataproc

25 minutes Updated June 25, 2021

This lab shows you how to use Spark MLlib and spark-nlp for performing machine learning and NLP on large quantities of data.

Start

AI Speech Recognition with TensorFlow Lite for Microcontrollers and SparkFun Edge

15 minutes Updated October 11, 2020

We needed :

01 |

**Auto-scaling clusters/nodes
Cost-effective (Small budget team)**

Dataproc Serverless

02 |

Automation of our cron jobs

?

“

**You do not rise to the
level of your goals.
You fall to the level of
your systems.**

JAMES CLEAR
Atomic Habits

Enter Airflow



DAGs Security Browse Admin Docs

21:11 UTC RH

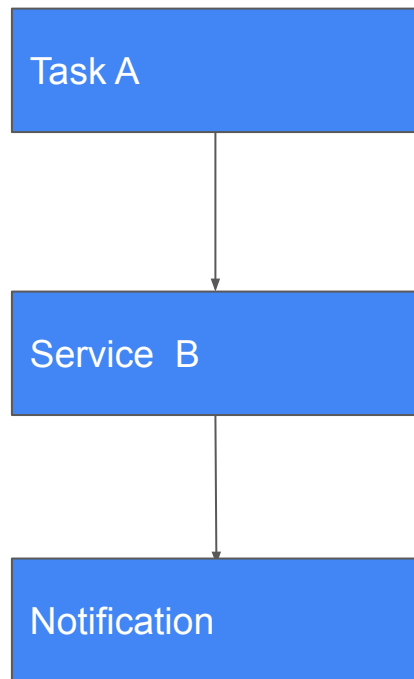
DAGs

All 26		Active 10		Paused 16		Filter DAGs by tag		Search DAGs	
DAG	Owner	Runs	Schedule	Last Run	Recent Tasks	Actions	Links		
<div><div><div><div><div></div><div>example_bash_operator</div><div>exampleexample2</div></div></div></div></div> <div>airflow</div> <div>2</div> <div>0 0 * * *</div> <div>2020-10-26, 21:08:11</div> <div>6</div> <div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>									

I wish I knew of this in 2016

Composer

- a. Why ?
- b. Under the hood
- c. DAGs
- d. Dataproc Operators Demo



Why Composer ?

- **Managed Airflow**
- Focus on your **workflows** and not your **infrastructure**
- Since its based on **open source** framework your workflows can be **multi-cloud** and/or **on-prem**.
- Uses **Python** - so no new lang.
- Use **directed acyclic graphs (DAGs)** - Simple to decipher
- Widely used - several operators and connectors available for almost every tool



“Did someone say Multi-cloud?”



Composer

Environments

[+ CREATE](#)[DELETE](#)[Filter](#) Filter environments

<input type="checkbox"/>	<input type="radio"/>	Name ↑	Location	Composer version	Airflow version	Creation time	Update time	Airflow webserver	DAG list NEW	Logs	DAGs folder	Labels
<input type="checkbox"/>	<input type="radio"/>	testing-bq	us-central1	2.0.23	2.2.5	8/18/22, 12:12 PM	8/18/22, 12:12 PM	None	DAGs	Logs	None	None



Airflow webserver	DAG list NEW	Logs	DAGs folder	Labels
Airflow	DAGs	Logs	DAGs	None

Go here to see the airflow dashboard

Add your DAG python file in this folder

Under the hood

- Airflow **workers** are executed on a GKE clusters
- Airflow **metadata** is stored on a Cloud SQL
- Airflow **Webserver** runs as an App Engine Flex which is protected by an Identity Aware Proxy.



DAGs and Operators

```
with DAG(  
    dag_id="GCS_BQ_Bash_operator",  
    schedule_interval='0 0 * * *',  
    start_date=datetime.datetime(2022, 11, 15)  
  
    ) as dag:  
  
    create_dataset = BashOperator(  
        task_id="dataset_creation",  
        bash_command="bq mk mydataset",  
    )
```

devfest
2022



Automation to its fullest for Dataproc Serverless

DataProcCreateBatchOperator,
DataProcGetBatchOperator,
DataProcListBatchOperator,
DataProcDeleteBatchOperator



Demo in console (5min)

We needed :

01 |

**Auto-scaling clusters/nodes
Cost-effective (Small budget team)**

Dataproc Serverless

02 |

Automation of our cron jobs

Cloud Composer

Further reading



1. [Run Dataproc serverless workloads with Cloud composer](#)
2. [Cloud Composer Tag in GCP Medium Publication](#)
3. [Composer codelabs](#) by Leah Cole

Build cool things with Google Cloud



/in/julian-s-joseph



medium.com/@juliansarajoseph

bit.ly/julian-feedback