devfest 2022

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

Make your Data pipelines easy with Dataproc and Composer

Julian Sara Joseph
Data Analytics DRE, Google Cloud

Google Developer Groups
Vancouver

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

Date: 19 November 2022

A walk down the memory lane

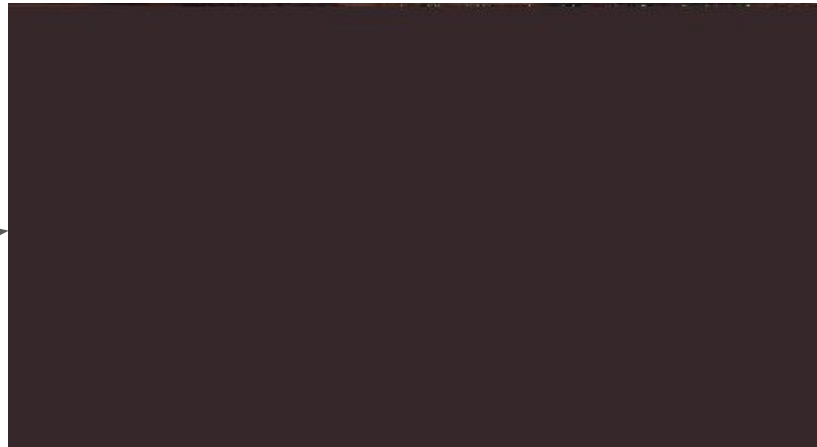
2016

- 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



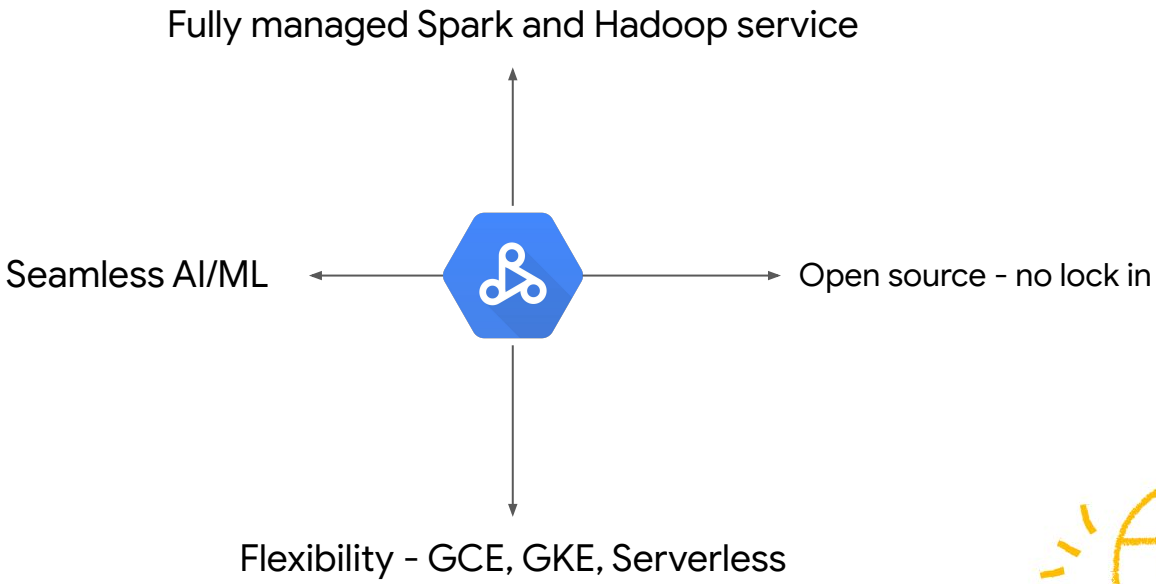
We needed :

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

02 | Automation of our cron jobs

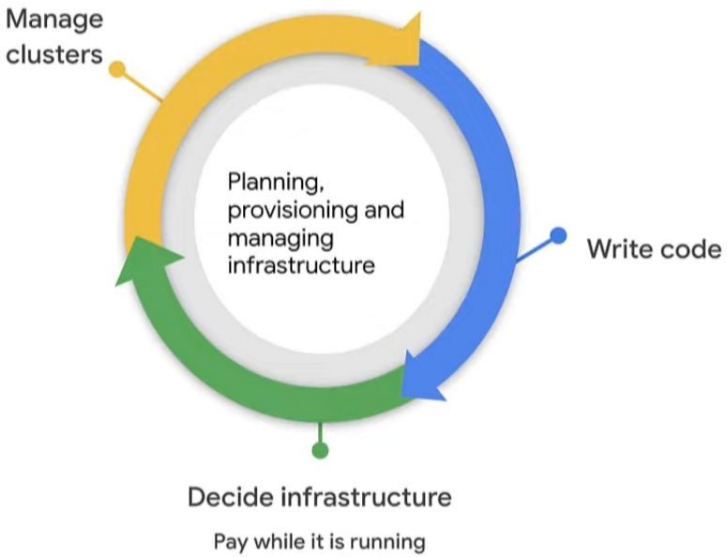
devfest
2022

Enter Dataproc




Google Cloud Platform
Low Cost
Low Rate & Billed By The Minute
Less Waiting
Clusters Start & Stop in < 90 Sec
Minimal Ops
We Manage Spark and Hadoop
Higher Scale & Productivity
Lower Cost & Complexity
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

What's better than Dataproc ? Serverless Spark!



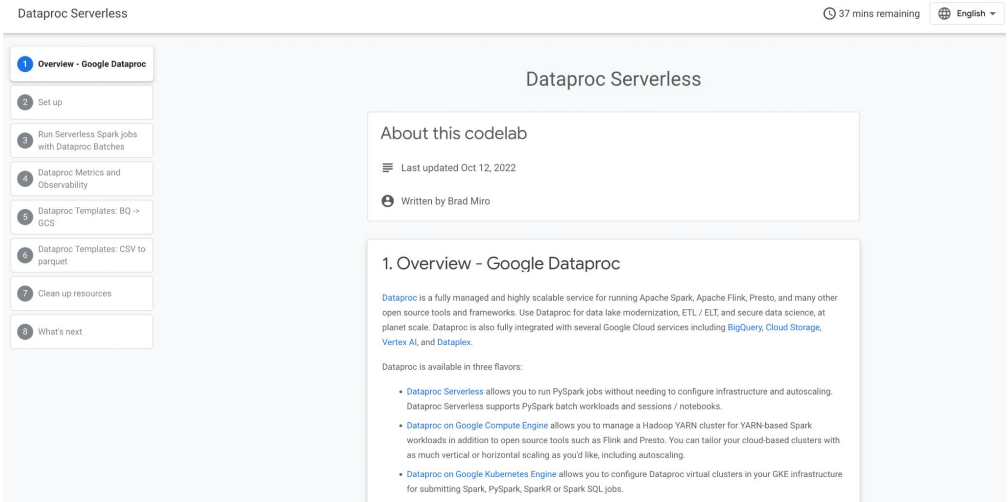
- ✓ Job auto-scales
- ✓ No infrastructure to tune
- No clusters to manage
- Only pay for the job duration

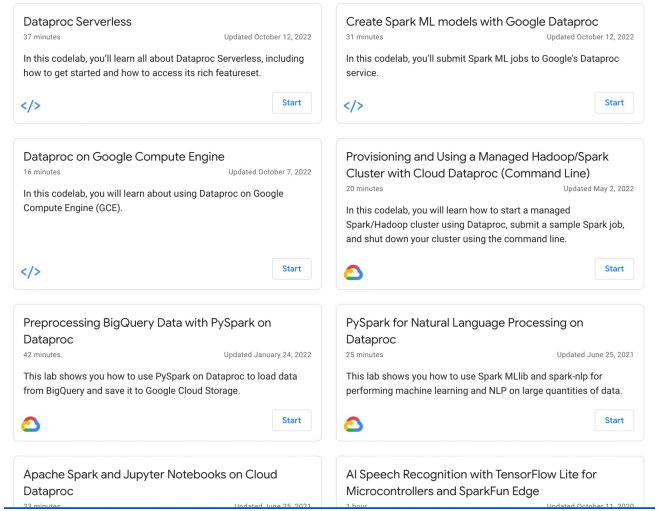
Focus on Spark, not infrastructure

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





We needed :

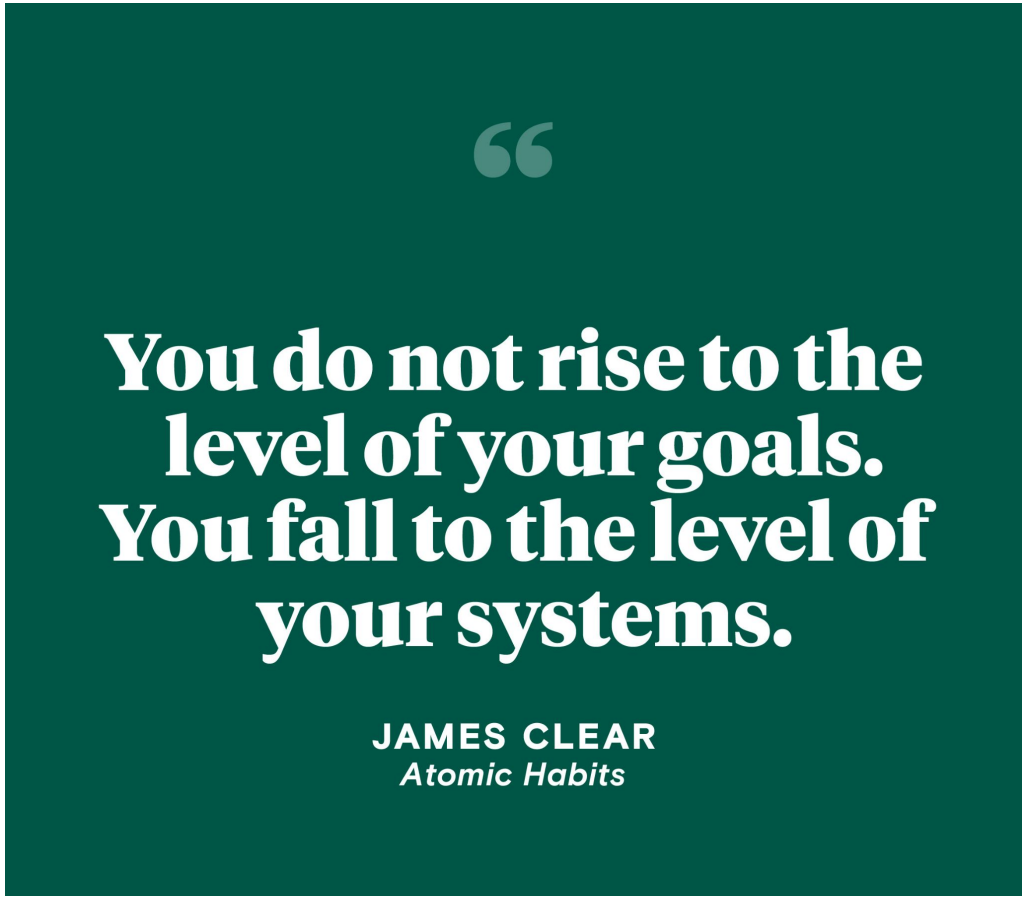
01 | Auto-scaling clusters/nodes
Cost-effective (Small budget team)

Dataproc Serverless

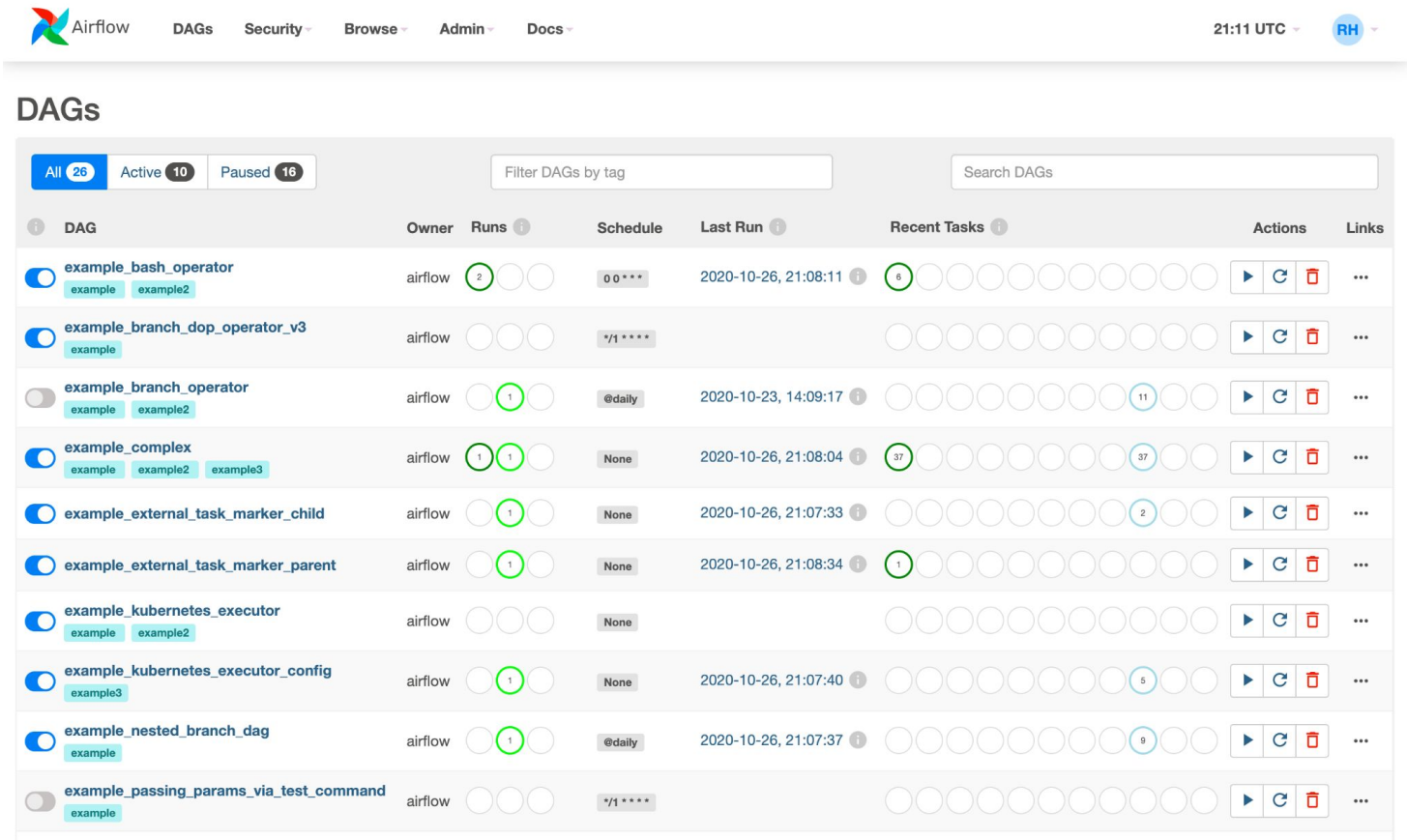
02 | Automation of our cron jobs

?

Automation


“
You do not rise to the level of your goals. You fall to the level of your systems.
JAMES CLEAR
Atomic Habits

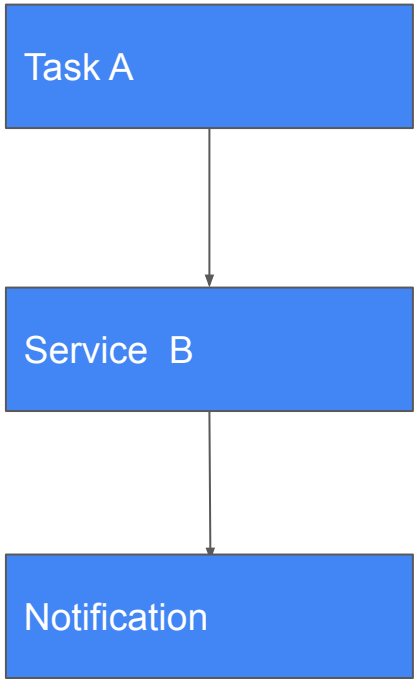
Enter Airflow



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

Name	Location	Composer version	Airflow version	Creation time	Update time	Airflow webserver	DAG list NEW	Logs	DAGs folder	Labels
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.

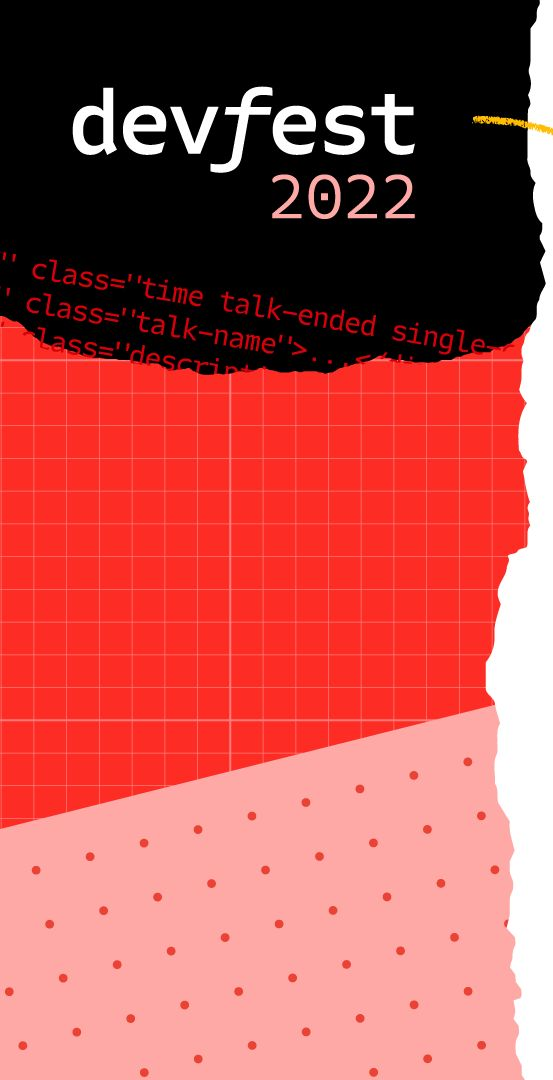
DAGs and Operators

```
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