Open source serverless spark data pipelines

Ajay Yadav
Data Platform Practice Lead
Google Cloud
linkedin.com/in/ajayydv

Shashank Agarwal
Staff Strategic Cloud Eng.
Google PSO
linkedin.com/in/shashank181

Agenda

01 Serverless computing

02 What is Serverless Spark ?

03 Serverless Spark benefits

04 Open source spark templates

05 Demo

06 Questions

What is serverless computing?

Running applications and services without having to manage the infrastructure itself.

“ Serverless architectures enable developers to focus on what they should be doing — writing code and optimizing application design — making way for business agility”

Gartner

Serverless on the rise

Developer Speed
& Efficiency

Security & Flexibility

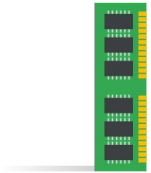
Data Governance
& Compliance

Cost Savings

What is Serverless Spark ?

Offering by GCP Dataproc which lets you run Spark workloads in Serverless mode.

Autoscaling



Auto provisioning and cleanup

Auto tuning

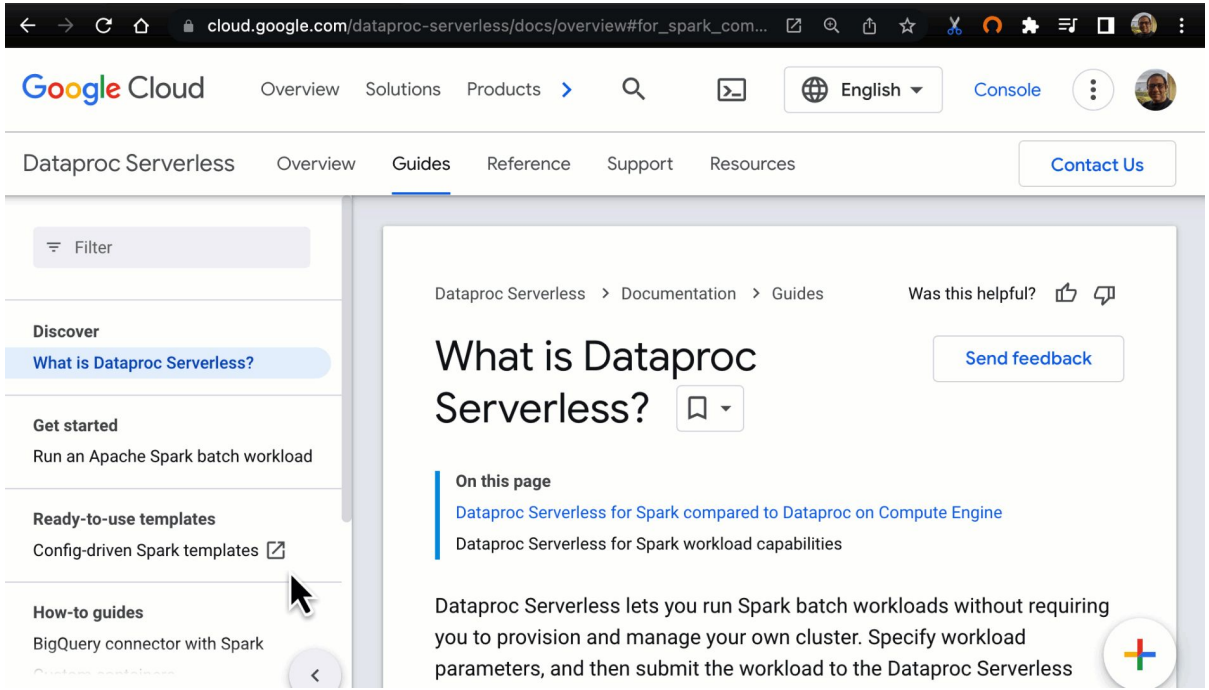
Benefits of Serverless Spark?

- Accelerate spark development and deployment
- Use ready to use Dataproc templates to bootstrap your project.
- Integration with Vertex AI, Jupyter notebooks and Dataplex.
- Submit jobs in Pyspark, Spark SQL, Spark R or Spark Java/Scala

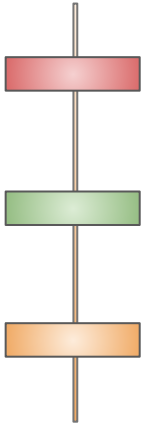
- Reduced Operational overhead
- Simplified Governance & Security
- Accelerate Data lake modernization & migration efforts.
- Utilize existing skill set with a cloud native, serverless environment.

Serverless Spark Templates

Ready to use, open sourced (github), customizable templates and notebooks.



How to use Serverless spark templates?



Step 1: Clone github repo in cloud shell (or preferred location).
https://github.com/GoogleCloudPlatform/dataproc-templates

Step 2: Authenticate with your GCP identity or use service account.
gcloud auth application-default login

Step 3: Launch respective template from the shell.
It will automatically build, deploy and run the spark code.

Example pipeline execution

What is Dataproc Ser | Cloud Shell | Batches – Dataproc – | BigQuery – apchecor

cloud.google.com/dataproc-serverless/docs/overview

Google | Gmail | JDK | mvn | DataEng | PSO Lab | FT | Care.com | Keep | Other Bookmarks

Google Cloud | Overview | Solutions | Products | Pric > | English | Console

Dataproc Serverless | Overview | Guides | Reference | Support | Resources | Contact Us

Filter

Discover

What is Dataproc Serverless?

Get started

Run an Apache Spark batch workload

Ready-to-use templates

Config-driven Spark templates

What is Dataproc Serverless?

Send feedback

On this page

Dataproc Serverless lets you run Spark batch workloads without requiring you to provision and manage your own cluster. Specify workload parameters, and then submit the workload to the Dataproc Serverless service. The service will run the workload on a managed compute infrastructure, autoscaling resources as needed. Dataproc Serverless charges apply only to the time when the workload is executing.

Demo

Java	Python	Notebooks
HiveToBigQuery	BigQueryToGCS (blogpost)	Hive to Bigquery Notebook
HiveToGCS	GCSToBigTable	MySQL to Cloud Spanner Notebook
GCSToBigQuery	GCSToJDBC	SQL Server to Postgresql Notebook (on roadmap)
GCSToGCS	GCSToMongo (blogpost)	SQL Server to BigQuery Notebook (on roadmap)
GCSToSpanner (blogpost)	GCSToGCS	Oracle to Spanner Notebook (on roadmap)
HBaseToGCS	HiveToBigQuery (blogpost)	Oracle to BigQuery Notebook (on roadmap)
SpannerToGCS (blogpost)	HiveToGCS (blogpost)	
JDBCToBigQuery	HbaseToGCS	
JDBCToGCS (blogpost)	MongoToGCS (blogpost)	
PubSubToGCS (blogpost)	SnowflakeToGCS	
GCSToJDBC (blogpost)	JDBCToGCS	
SnowflakeToGCS	JDBCToBigQuery	
KafkaToBQ (blogpost)	RedshiftToGCS	

Community engagement and learning resources

Running Spark Jobs in a Serverless Environment (Link)

How to use Kubernetes with Spark (Link)

How to use Notebooks with Spark (Link)

How to use Iceberg and Delta with Serverless Spark (Link)

Contact Us

Feedback, ideas, thoughts feedback-form

Questions, issues, and comments - dataproc-templates-support-external@googlegroups.com

Thank you.

Google Cloud