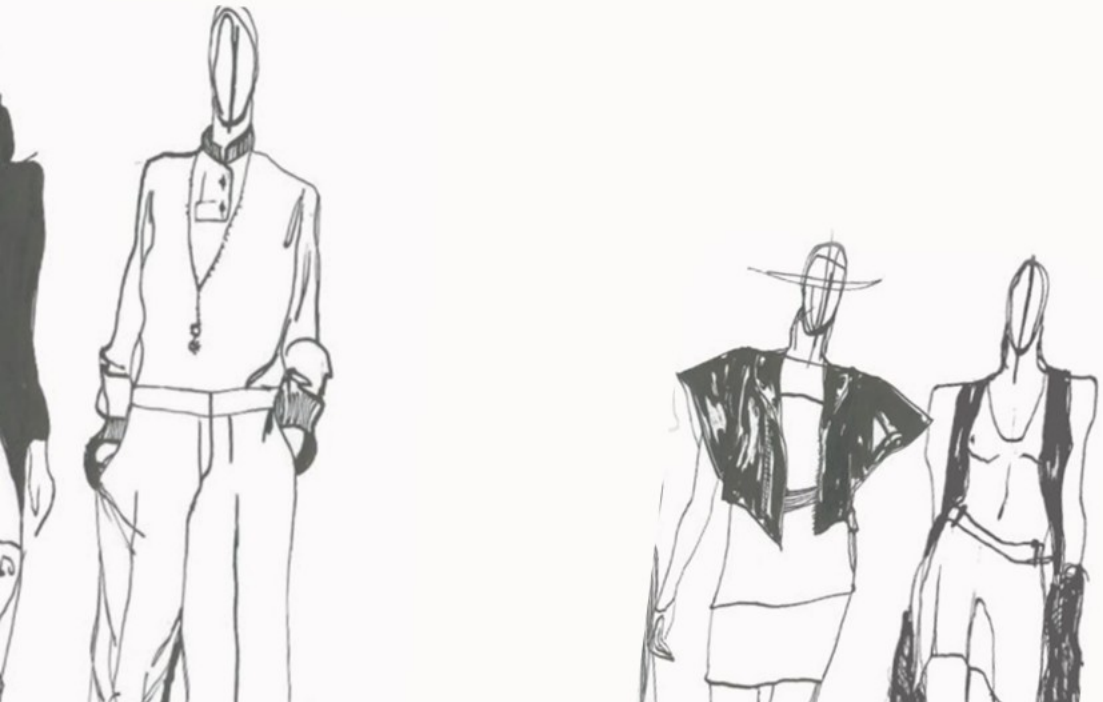
Infrastruktura analityczna w LPP e-commerce



Szymon Chojnacki
Enterprise Architect

LPP

5 MAREK DETALICZNYCH

... Z FLAGOWĄ MARKĄ RESERVED

Stworzyliśmy pięć rozpoznawalnych marek:
Reserved, **Cropp**, **House**, **Mohito** i **Sinsay**.
Każda z nich kierowana jest do innej grupy klientów reprezentujących odmienny styl życia, mający inny sposób na wyrażenie siebie i różne potrzeby.

RESERVED CROPP house MOHITO sinsay

Zespół

Łukasz

Mateusz

Marek

Szymon

Doświadczenie z: **Php, JavaScript, Java, GA360, Excel, Python, SQL, AWS, ML**

Używamy: **GCP, Apache Kafka, App Engine, Dataflow, Airflow**

Uczymy się: **Apache Beam, Cloud ML, Plotly Dash**

LPP RESERVED CROPP house MOHITO sinsay

Infrastruktura analityczna

Steps

1 Source

2 Ingress

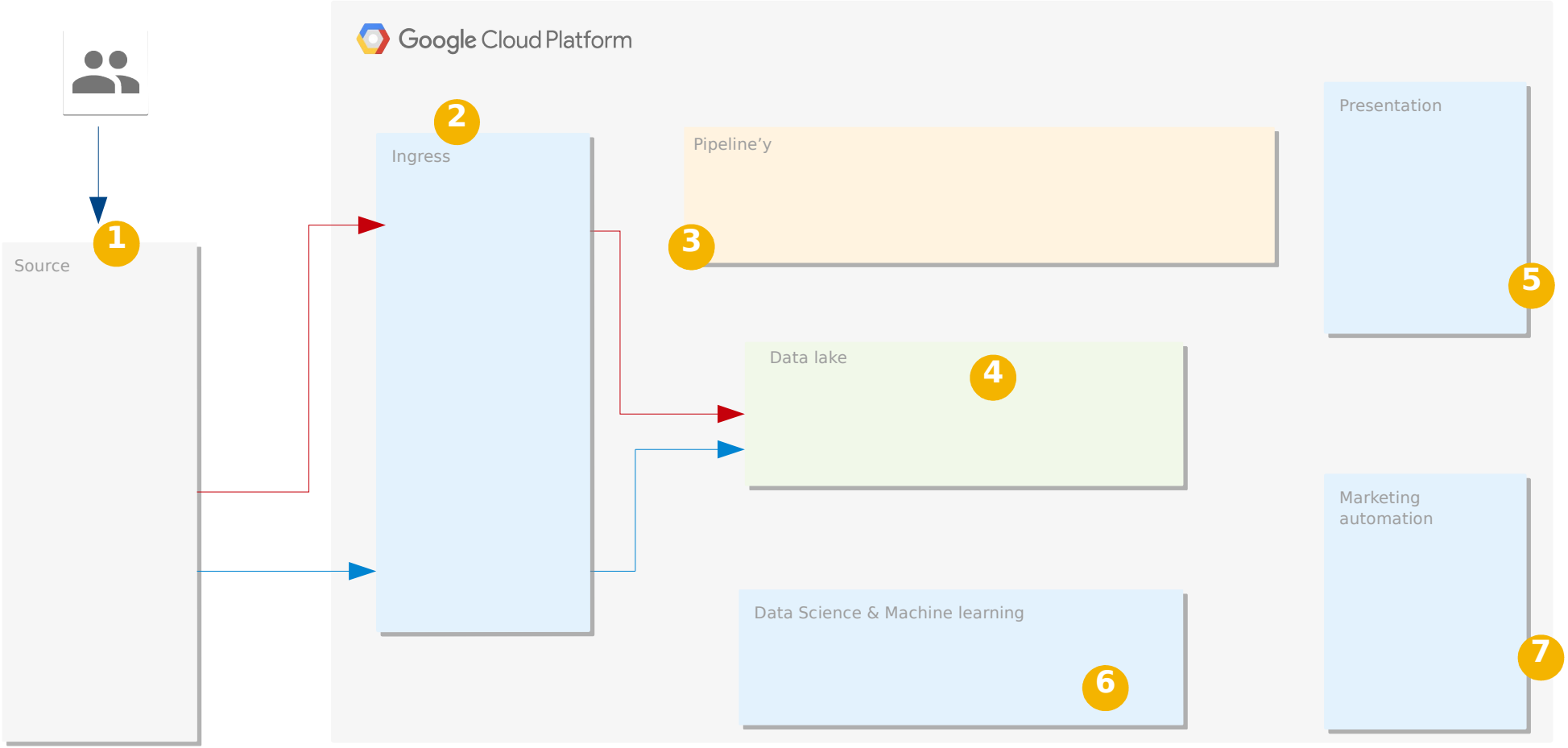
3 Pipelines

4 Data lake

5 Visualization

6 Data science & ML

7 Action



LPP RESERVED CROPP house MOHITO sinsay

Infrastruktura analityczna

Steps

1 Source

2 Ingress

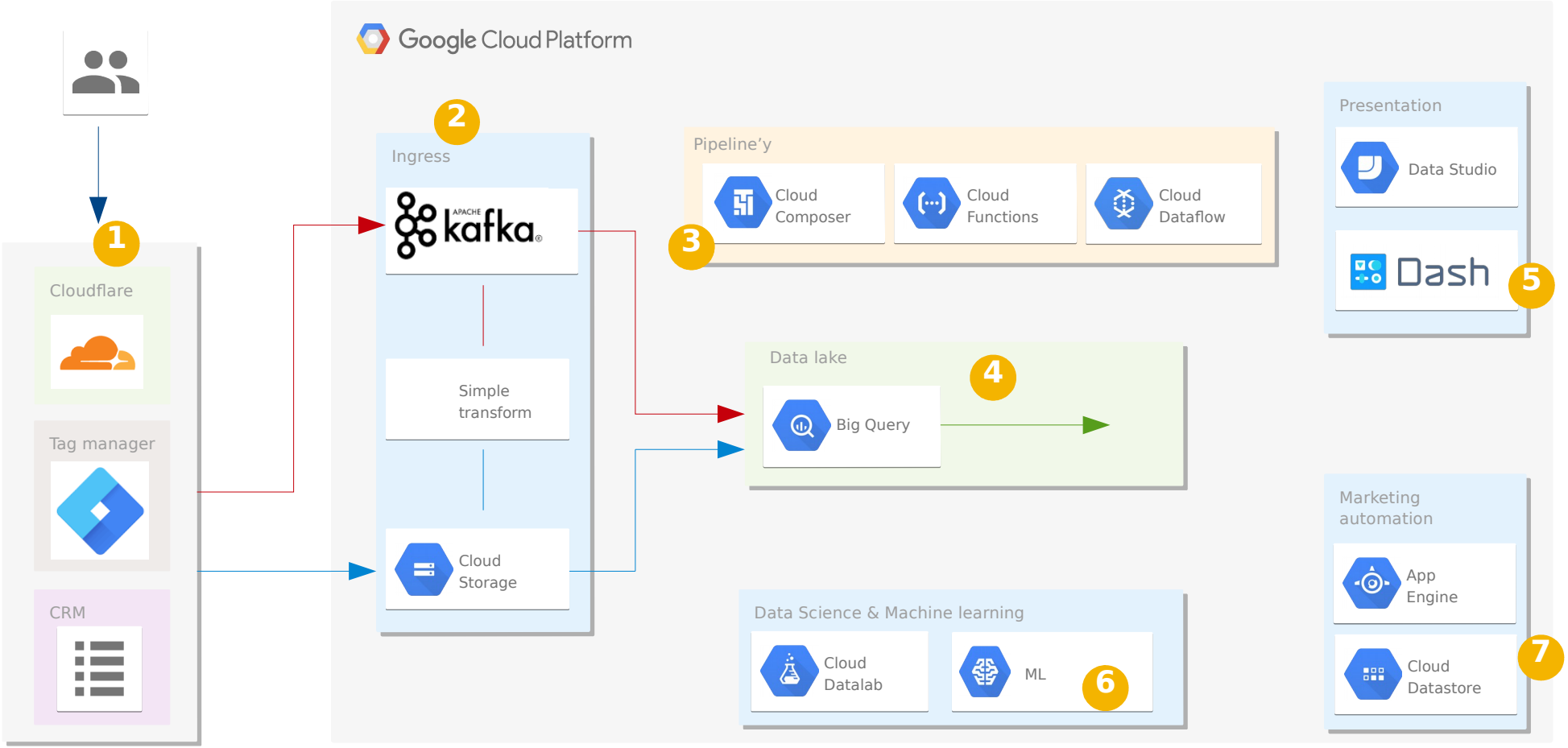
3 Pipelines

4 Data lake

5 Visualization

6 Data science & ML

7 Action



LPP RESERVED CROPP house MOHITO sinsay

Partnerzy

Google Cloud Platform

1. Google Cloud Platform

- metodologia SRE
- Google Kickstart
- Qwiklabs

deepsense.ai BIG DATA SCIENCE

2. Deepsense

- machine learning
- feature engineering
- success stories

SOFTWAREMILL

3. SoftwareMill

- IaaC
- Streaming Pipelines

LPP RESERVED CROPP house MOHITO sinsay

Pytanie: **W jaki gwiazdozbiór układają się kroki** 1 2 3 4 5 6 7

a) Orion

b) Wielki wóz

c) Bliźnięta

Infrastruktura analityczna

Steps

1 Source

2 Ingress

3 Pipelines

4 Data lake

5 Visualization

6 Data science & ML

7 Action

Source 1

Ingress 2

Pipeline’y 3

Data lake 4

Presentation 5

Data Science & Machine learning 6

Marketing automation 7

a) Orion

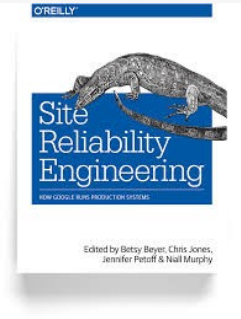
b) Wielki wóz

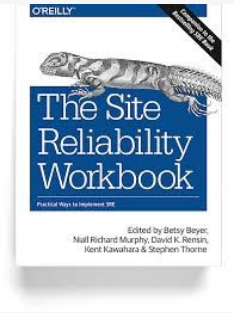
c) Bliźnięta

LPP RESERVED CROPP house MOHITO sinsay

Google Cloud Platform Team - wsparcie

1. **Metodologia SRE (site reliability engineering)**
   - czy big data pasuje do Scrum'a
   - sprinty i 'person on call'
   - SLO (service level objective)





2. **Google Kickstart na Coursera**
   - szkolenia
   - certyfikaty

3. **Qwiklabs**
   - laboratoria

GCP – różne doświadczenia

1. **Help desk**
2. **Skalowalność**
3. **Serwerless**

Support options

	BASIC	ROLE-BASED		ENTERPRISE
Cost per month	Free	$100/user	$250/user	Greater of $15,000

Google Cloud Platform

Support

- Overview
- Cases
- Chat support
- Phone support
- Community support
- Settings

Case ... REOPEN

Hello Szymon,

As I explained yesterday, I have continued working on my reproductions. Mainly, I have built 2 scenarios:
- Migration of 366 small tables: I have created a DAG using the code in the blog post, in order to transfer 366 small tables (<15MB) to a dataset in my project. This DAG has 1100 tasks (= 366 days * 3 tasks/day + 1 init + 1 end), and by now (~2h later) it has completed successfully all the tasks, i.e. the DAG run completely. No "timeout" has shown in the logs, and all tasks are marked as "successful".
- Migration of 1 very big table: I also have another DAT that transfers table [2] (6.76 TB) to a dataset in my project. This DAG has only 5 tasks and as of now (1 hour after its creation) it has completed 2 of the tasks and is running the third one, "GCS_to_GCS", which is over 75 minutes of execution right now. I have not seen any timeout until now either.

GCP – Skalowalność

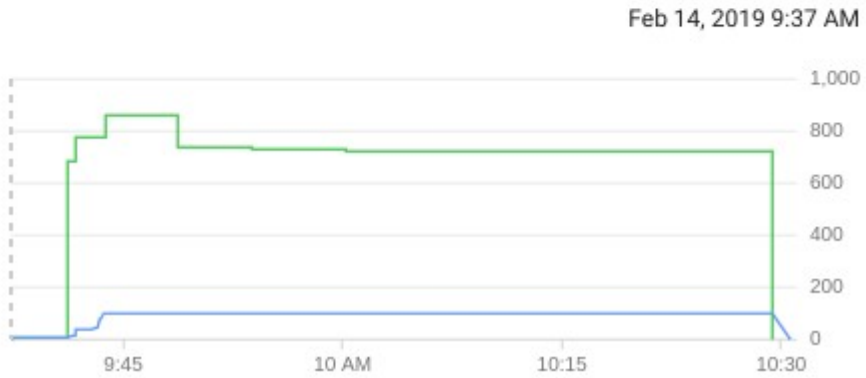
Dataflow

Job summary

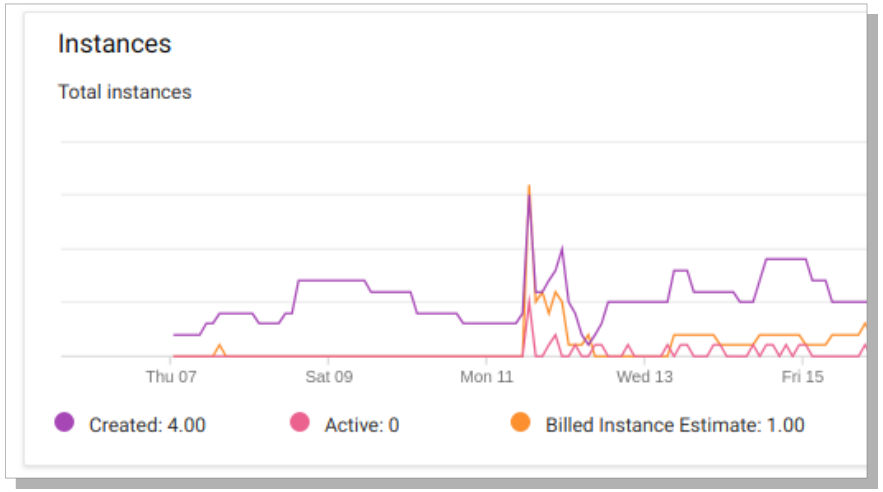
Job name	
Job ID	2019-02-14_00_37_14-12518667443414548076
Region	europe-west1
Job status	Succeeded
SDK version	Apache Beam SDK for Python 2.10.0
Job type	Batch
Start time	Feb 14, 2019, 9:37:15 AM
Elapsed time	53 min 30 sec

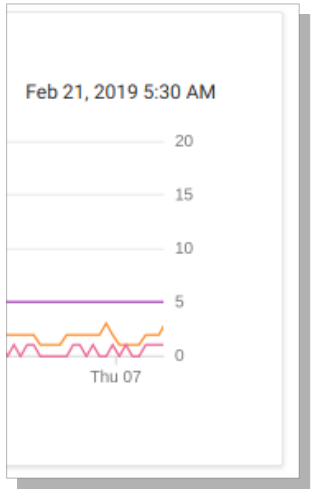
Autoscaling

Workers	0
Current state	Worker pool stopped.



App Engine

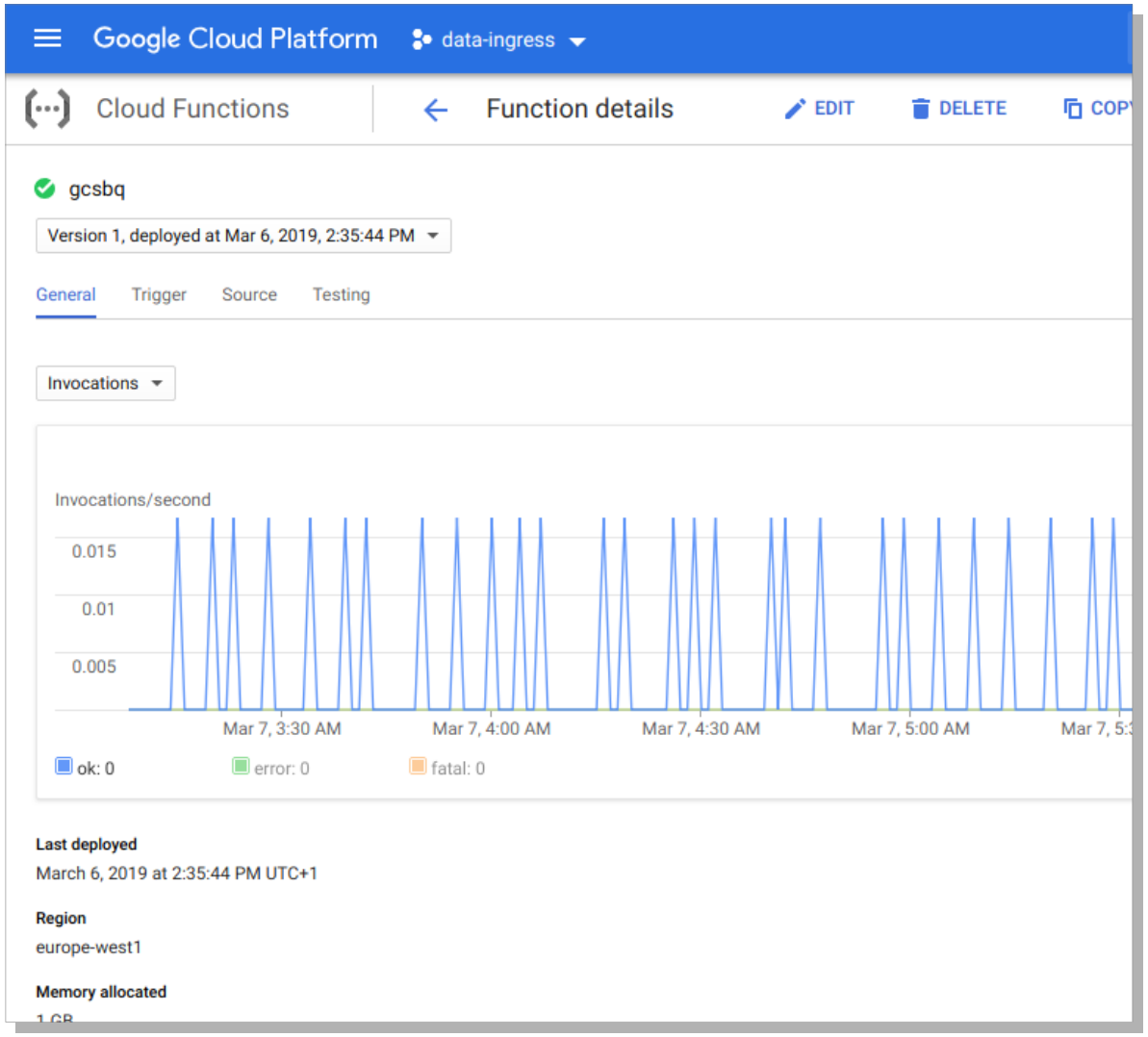




Big Query

Feb 1 – 28, 2019	BigQuery Analysis: 591.399 Tebibytes (Source: ...])	$2,951.99

GCP – Serwerless



index.js

```
'use strict'

const { BigQuery } = require('@google-cloud/bigquery')
const { Storage } = require('@google-cloud/storage')

async function gcsbq (file, context) {
    const _schema = require(process.env.SCHEMA)

    const datasetId = process.env.DATASET
    const tableId = process.env.TABLE

    const bigquery = new BigQuery()

    const storage = new Storage()

    console.log(`Starting job for ${file.name}`)

    const filename = storage.bucket(file.bucket).file(file.name)

    /* Configure the load job and ignore values undefined in schema */
    const metadata = {
        sourceFormat: 'NEWLINE_DELIMITED_JSON',
        schema: {
            fields: _schema
        },
        ignoreUnknownValues: true
    }

    const dataset = bigquery.dataset(datasetId)

    await dataset.get({ autoCreate: true }, (e, dataset, res) => {
        if (e) console.log(e)
        dataset.table(tableId).get({ autoCreate: true }, (e, table, res) => {
            table.load(filename, metadata)
        })
    })
}

exports.gcsbq = gcsbq
```

RESERVED CROPP house MOHITO sinsay

Zapraszamy na staż

Data Engineer Intern

Data Scientist Intern

/lppkariera

@discoverlpp

LPP S.A.

LPP

DZIĘKUJĘ