Google Cloud Platform

ENGINEERING SCHOOL
Creating the future together

CASTLE BEE

Mini Project

Context and problem

You have been hired as a Data Engineer by a police department of analytics and have been given a dataset of all crimes happening in your area.
Here are the question your supervisor what's you to answer.

0- Before starting, add 3 years to all dates in the dataset

1 – What is the total number of crime per month for the past 5 years (sum of all crimes that happened during that month over the past 5 years)

2 – What are the top 10 location of "THEFT" crimes, for each of the past 3 years ?

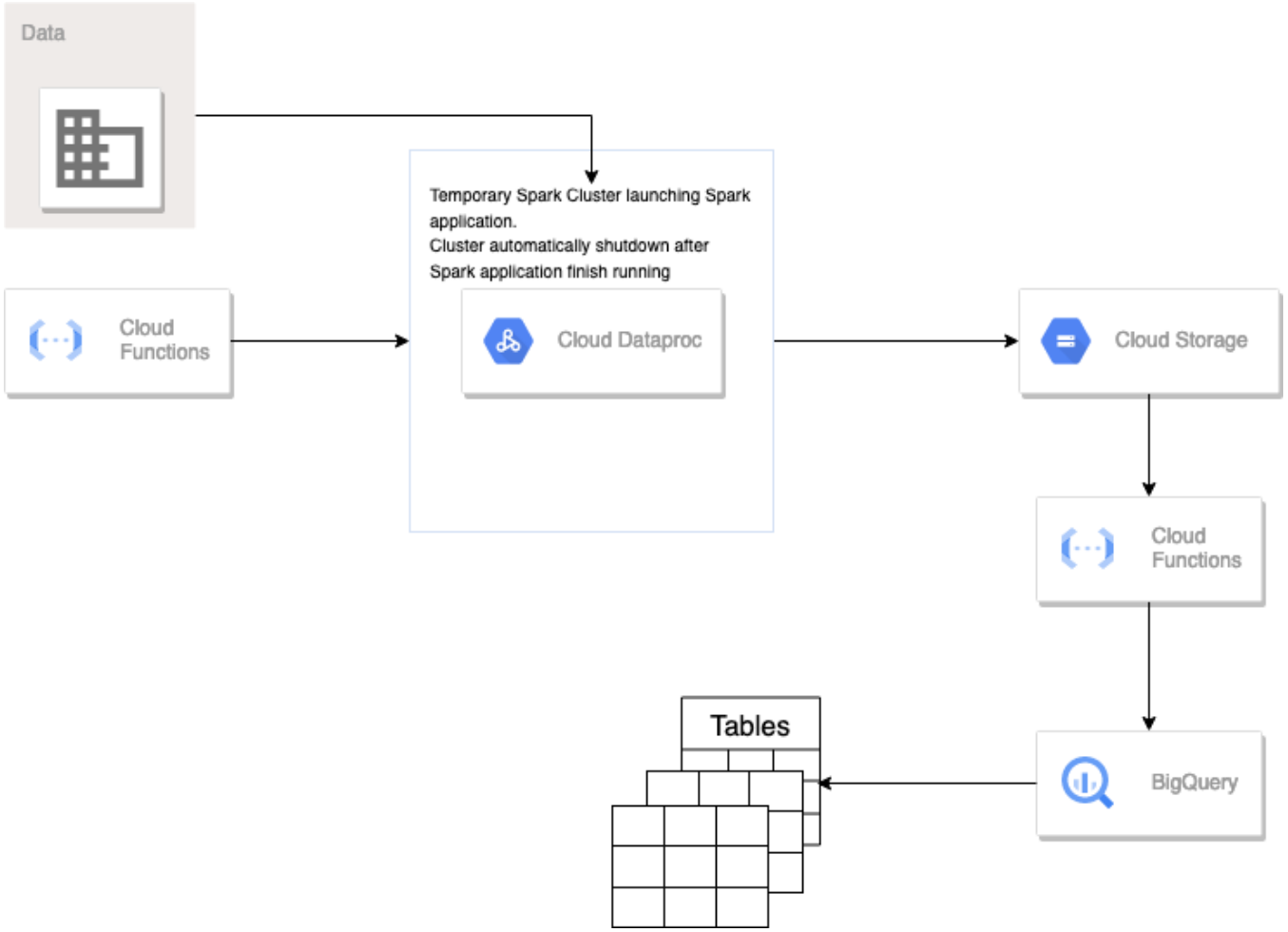
3 – Is crime declining or increasing overall ?

4 – Which areas are the safest between 10pm and 4 am ?

5 – Which types of crimes are suspects most arrested for between 2016 and 2019 ?

Given architectures contains services that are mandatory to use, but you are free to add or duplicate any service you think may be needed to accomplish given goals.

Architecture 1



Download data from this link

- Spawn a Cloud function that will create a temporary spark cluster. Your spark Cluster will transform data to obtain data small enough to be read with a cloud function and respond to the questions your supervisor.

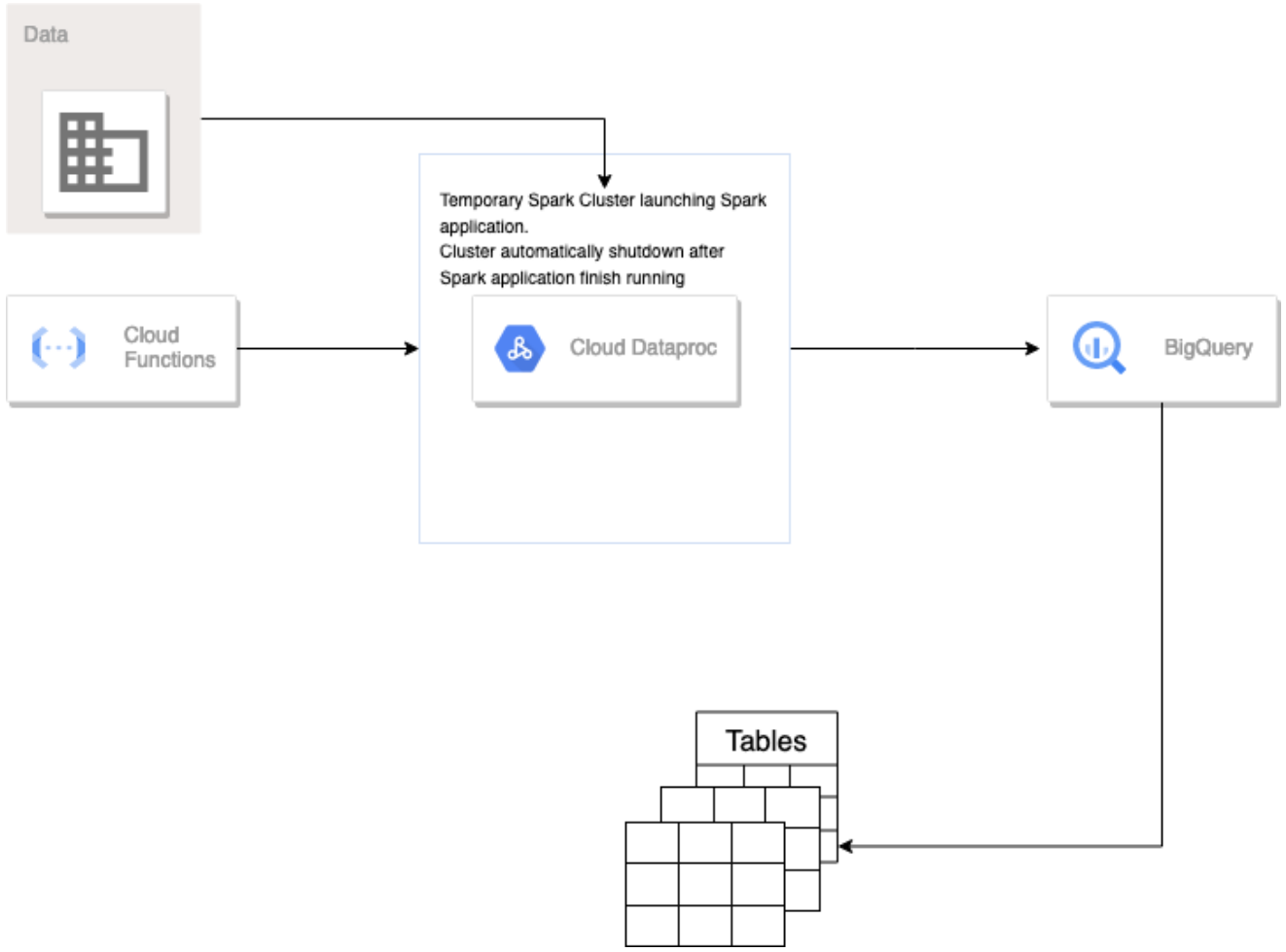
- Store those datasets (saved as parquet) to a given bucket

- Using a cloud function, create several bigquery tables for each questions

- Bonus : using a cloud function, create graphics that answers each questions and those graphics in a given bucket.

- Orchestrate your workflow using Cloud Scheduler (and Cloud Pub/Sub if necessary).

Architecture 2



Download data from this link

- Spawn a Cloud function that will create a temporary spark cluster. Your spark Cluster will directly transfer the raw data into a bigquery table.

- Using SQL, create several bigquery tables for each questions

- Orchestrate your workflow using Cloud Scheduler (and Cloud Pub/Sub if necessary).

Notes

Notes Architecture 2 :
- Bigquery : create 2 datasets. The first one will contain the raw table. The second one will contain the transformed tables

Notes Architecture 1 :
- Bigquery : only create one dataset containing transformed tables

Evaluation metrics :
- The student will provide a schema of their final architecture and a small paragraph explaining how to launch their data pipelines and where to view the different results

3 evaluation metrics :
- Orchestration
- Use of Cloud services
- Cleanness of code

Contact us

- georges.awono@castlebee.fr / 06 67 15 37 06
- ahmed.chreif@castlebee.fr / 06 45 81 20 79
- antoine.gademer@epf.fr
- Follow us on :
  - Our website : CASTLE BEE
  - Linkedin : CASTLE BEE
  - Welcome to the jungle : CASTLE BEE