

Execution Steps



Von

Roland Bauer
Dominik Prüll
Sebastian Weidele
Oliver Tomondy

in

Big Data Infrastructure

am

FH Technikum Wien

Sommersemester 2021

16.6.2021

Inhaltsverzeichnis

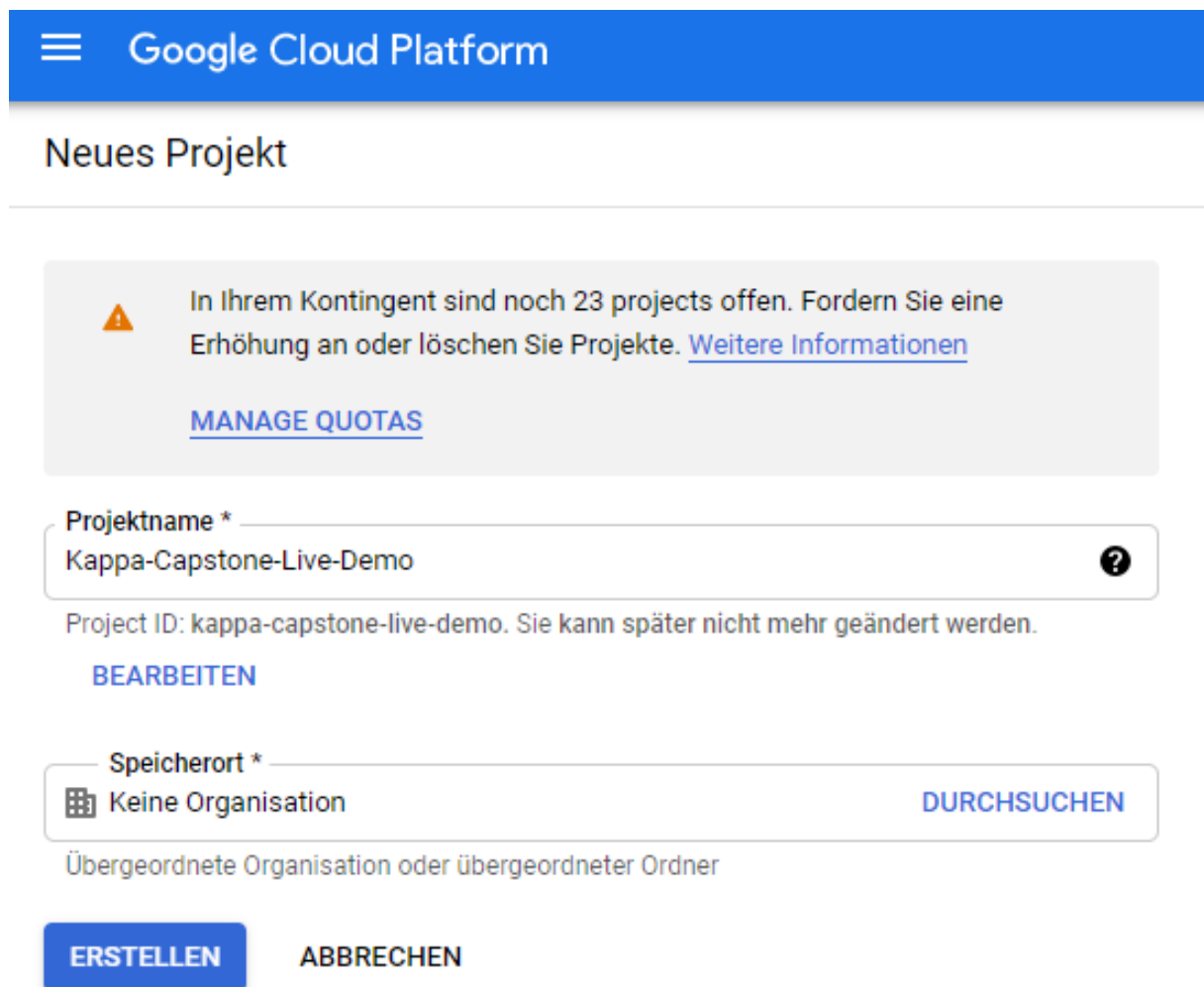
Set up Umgebung.....	3
<i>Projekt erstellen</i>	<i>3</i>
<i>Projektübersicht</i>	<i>4</i>
<i>Erstellen einer VM-Instanz</i>	<i>4</i>
<i>Instanz Einstellungen</i>	<i>5</i>
<i>Bootlaufwerk Einstellungen</i>	<i>5</i>
<i>VM-Instanz Übersicht.....</i>	<i>6</i>
<i>Docker installieren</i>	<i>7</i>
<i>Firewall.....</i>	<i>7</i>
Set up Apache Kafka	8
Set up Apache Druid	8
Set up Apache Superset	10
<i>Superset Database anbinden</i>	<i>12</i>
<i>Dataset und Dashboard Import</i>	<i>13</i>
<i>Superset Alerts + Reports</i>	<i>16</i>
Set up Datasource	17
Abbildungsverzeichnis.....	18

Set up Umgebung

Grundsätzlich wäre eine lokale Installation auch möglich. Aufgrund der Tatsache, dass Druid sehr ressourcenintensiv ist und viel Arbeitsspeicher benötigt, ist eine Cloud Instanz zu bevorzugen.


Diese Prototypen-Instanz wird in einer Ubuntu VM in der Google Cloud Platform aufgesetzt.

Projekt erstellen




Google Cloud Platform

Neues Projekt

 In Ihrem Kontingent sind noch 23 projects offen. Fordern Sie eine Erhöhung an oder löschen Sie Projekte. [Weitere Informationen](#)

[MANAGE QUOTAS](#)

Projektname * 

Project ID: kappa-capstone-live-demo. Sie kann später nicht mehr geändert werden.

[BEARBEITEN](#)

Speicherort * [DURCHSUCHEN](#)

Übergeordnete Organisation oder übergeordneter Ordner

[ERSTELLEN](#) [ABBRECHEN](#)

Abbildung 1 Projekt erstellen

Projektübersicht

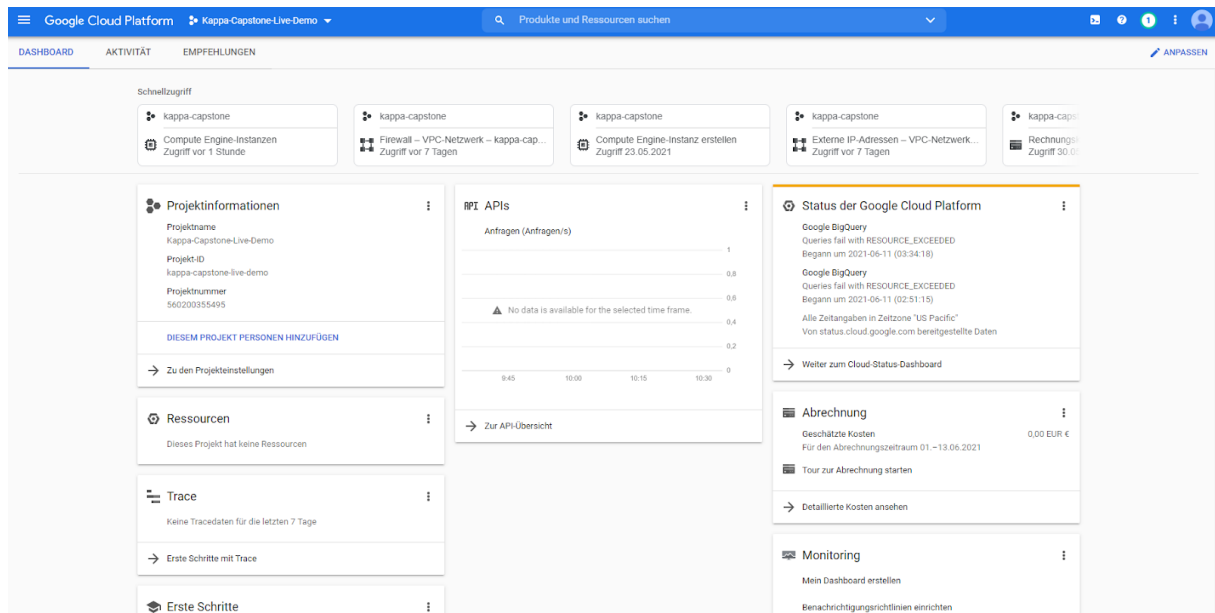


Abbildung 2 Projektübersicht

Erstellen einer VM-Instanz

Google Cloud Platform Menü → Compute Engine → VM-Instanzen
 Compute Engine API → Aktivieren

Nun kann eine Instanz erstellt werden.

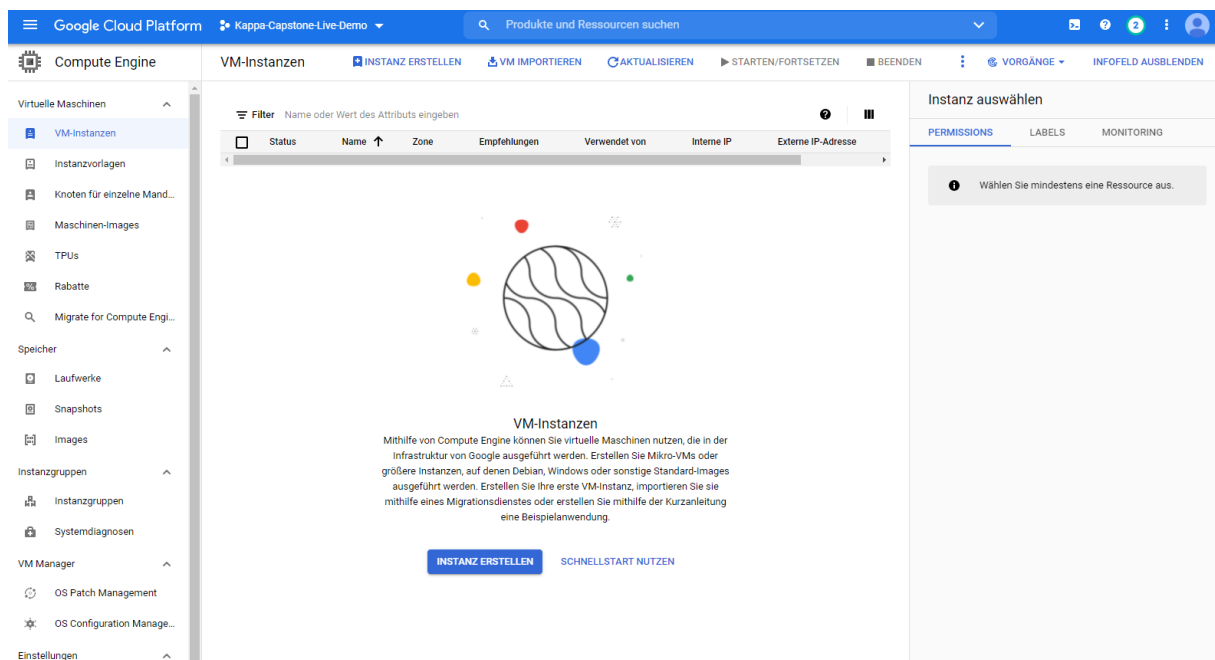


Abbildung 3 VM Erstellung

Instanz Einstellungen

Google Cloud Platform

Kappa-Capstone-Live-Demo

Produkte und Ressourcen suchen

← Instanz erstellen

Wählen Sie eine der Optionen, um eine VM-Instanz zu erstellen:

Neue VM-Instanz

Einzelne VM-Instanz neu erstellen

Neue VM-Instanz aus Vorlage erstellen

Einzelne VM-Instanz aus einer vorhandenen Vorlage erstellen

Neue VM-Instanz aus Maschinen-Image

Einzelne VM-Instanz aus einem vorhandenen Maschinenimage erstellen

Marketplace

Eine sofort einsatzbereite Lösung auf VM-Instanz bereitstellen

Name

Der Name kann später nicht mehr geändert werden

instance-1

Labels

(Optional)

+ Label hinzufügen

Region

Die Region kann später nicht mehr geändert werden

europa-west3 (Frankfurt)

Zone

Die Zone kann später nicht mehr geändert werden

europa-west3-c

Maschinenkonfiguration

Maschinenfamilie

Allgemeiner Zweck | Computing-optimiert | Arbeitsspeicheroptimiert

Maschinentypen für gängige Arbeitslasten, optimiert für Kosten und hohe Flexibilität

Reihe

N1

Powered by Intel Skylake-CPU-Plattform oder einem Vorgänger

Maschinentyp

Benutzerdefiniert

Kerne

6 vCPU

1 - 96

Arbeitsspeicher

32 GB

5,5 - 39

Speicher erweitern

CPU-Plattform und GPU

Vertraulicher VM-Dienst

Confidential Computing-Dienst auf dieser VM-Instanz aktivieren.

Container

Container-Image für diese VM-Instanz bereitstellen. Weitere Informationen

Bootlaufwerk

Neuer gleichmäßig ausgelasteter nichtflüchtiger Speicher mit 50 GB

Image

Ubuntu 20.04 LTS

Ändern

Sie haben noch ein Guthaben von 254,22745 € für die kostenlose Testversion

Pro Monat etwa 219,93 \$

Das sind etwa 0,301 \$ pro Stunde

Sie zahlen nur für die tatsächliche Nutzung: keine Vorauszahlungen und sekundengenaue Abrechnung

Details

Abbildung 4 VM Einstellungen

Bootlaufwerk Einstellungen

Bootlaufwerk

Wählen Sie ein Image oder einen Snapshot aus, um ein Bootlaufwerk zu erstellen, oder fügen Sie ein bestehendes Laufwerk hinzu. Sie finden nicht das, wonach Sie gesucht haben? Sehen Sie sich Hunderte VM-Lösungen im [Marketplace](#) an.

Öffentliche Images | Benutzerdefinierte Images | Snapshots

Vorhandene Laufwerke

Betriebssystem

Ubuntu

Version

Ubuntu 20.04 LTS

amd64 focal image built on 2021-06-10, gVNIC is required to support higher network bandwidths for distributed workloads on VMs that have attached GPUs. This flag indicates that the gVNIC driver is installed in on image - it does not add the NIC; this is done during instance creation. Both flags must be set for gVNIC to work successfully.

Startdatenträgertyp

Gleichmäßig ausgelasteter nichtflücht...

Größe (GB)

50

Abbildung 5 Bootlaufwerk Einstellungen

Identität und API-Zugriff ?

Dienstkonto ?

Compute Engine default service account

Zugriffsbereiche ?

☒ Standardzugriff zulassen

☐ Uneingeschränkten Zugriff auf alle Cloud-APIs zulassen

☐ Zugriff für jede API festlegen

Firewall ?

Sie können Tags und Firewallregeln hinzufügen, um bestimmten Netzwerktraffic aus dem Internet zuzulassen.

☐ HTTP-Traffic zulassen

☐ HTTPS-Traffic zulassen

↕ Verwaltung, Sicherheit, Laufwerke, Netzwerke, einzelne Mandanten

Ihr kostenloses Testguthaben wird für diese VM-Instanz verwendet.
[Kostenlose GCP-Stufe](#)

Erstellen **Abbrechen**

Entsprechende [REST-Anfrage/-Antwort](#) oder [Befehlszeile](#)

Abbildung 6 Identität und API Einstellungen

Wurden alle Einstellungen getätigt, kann die Instanz erstellt werden.

VM-Instanz Übersicht

Anschließend wird die Übersicht der VM-Instanz angezeigt:

The screenshot displays the Google Cloud Platform interface for VM Instances. The main content area shows a table of instances. The first instance, 'instance-1', is in the 'europa-west3-c' zone and has an internal IP of '10.156.0.2 (nic0)'. Its external IP address is '35.198.184.31', which is circled in yellow. In the 'Verbinden' column, there is an 'SSH' button, also circled in yellow. Below the table, there are several 'Weiterführende Aktionen' (Further actions) such as 'Abrechnungsbericht ansehen', 'VMs überwachen', 'VM-Logs prüfen', 'Firewallregeln einrichten', and 'Patchverwaltung'. The left sidebar shows the 'Compute Engine' menu with 'VM-Instanzen' selected. The top navigation bar includes buttons for 'INSTANZ ERSTELLEN', 'VM IMPORTIEREN', 'AKTUALISIEREN', 'STARTEN/FORTSETZEN', and 'BEENDEN'.

Abbildung 7 VM Übersicht

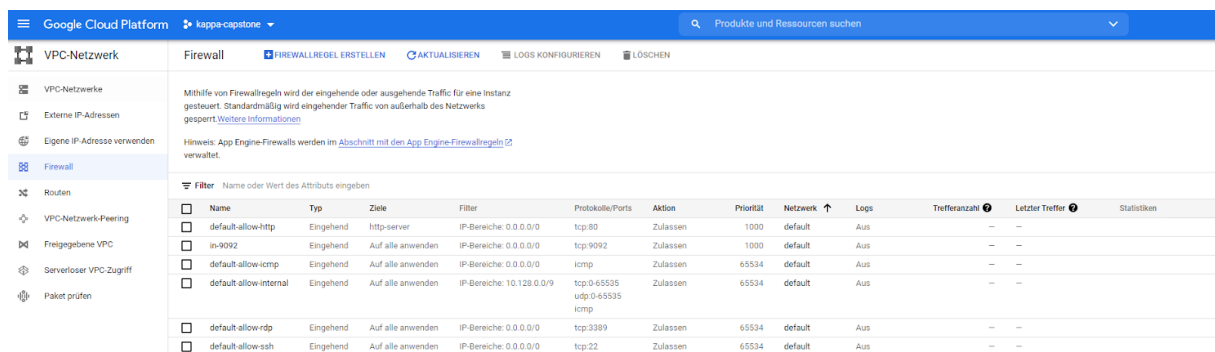
Hier kann die externe IP-Adresse abgelesen werden. Diese wird im weiteren Verlauf benötigt.

Docker installieren

Mittels Klick auf “SSH” kann die Browser Konsole geöffnet und Docker installiert werden:
<https://docs.docker.com/engine/install/ubuntu/>

Firewall

Um Daten von der Maschine in das Kafka-Topic schreiben zu können, musste eine eingehende Port-Weiterleitung (Port 9092) eingerichtet werden:



The screenshot shows the Google Cloud Platform console for a project named 'kappa-egstone'. The 'Firewall' section is selected in the left sidebar. The main area displays a list of firewall rules. The rules are as follows:

Name	Typ	Ziele	Filter	Protokolle/Ports	Aktion	Priorität	Netzwerk	Logs	Trefferanzahl	Letzter Treffer	Statistiken
default-allow-http	Eingehend	http-server	IP-Bereiche: 0.0.0.0/0	tcp:80	Zulassen	1000	default	Aus	--	--	
in-9092	Eingehend	Auf alle anwenden	IP-Bereiche: 0.0.0.0/0	tcp:9092	Zulassen	1000	default	Aus	--	--	
default-allow-icmp	Eingehend	Auf alle anwenden	IP-Bereiche: 0.0.0.0/0	icmp	Zulassen	65534	default	Aus	--	--	
default-allow-internal	Eingehend	Auf alle anwenden	IP-Bereiche: 10.128.0.0/9	tcp:0-65535 udp:0-65535 icmp	Zulassen	65534	default	Aus	--	--	
default-allow-rdp	Eingehend	Auf alle anwenden	IP-Bereiche: 0.0.0.0/0	tcp:3389	Zulassen	65534	default	Aus	--	--	
default-allow-ssh	Eingehend	Auf alle anwenden	IP-Bereiche: 0.0.0.0/0	tcp:22	Zulassen	65534	default	Aus	--	--	

Abbildung 8 Port-Forwarding Einstellungen

Des weiteren müssen folgende Ports weitergeleitet werden:

Druid: 8888

Superset: 8088

Set up Apache Kafka

Zunächst das Kafka Docker-Image von Wurstmeister aus git klonen:

```
"git clone https://github.com/wurstmeister/kafka-docker.git"
```

Im neuen Verzeichnis kafka die datei docker-compose.yml mit der selben Datei aus dem Abgabe-Zip ersetzen. Anschließend in der Datei docker-compose.yml <actualIPorlocalhost> ersetzen mit "localhost" oder der externen VM IP-Adresse.

Nach der Konfiguration kann Kafka mit "sudo docker-compose up -d" gestartet werden.

Set up Apache Druid

Als erstes wird ein neuer Ordner druid erstellt und die Dateien docker-compose und environment aus dem Abgabe-Zip eingefügt.

Das Druid-Cluster wird mit "sudo docker-compose up -d" gestartet.

Sobald das Cluster komplett hochgefahren ist, kann die Druid-Konsole über <actualIPorlocalhost>:8888 geöffnet werden.

Um einen neuen Ingestion-Job mit Kafka zu erstellen, zunächst auf "Load data" klicken. Sollten bereits Jobs vorhanden sein, muss zusätzlich "Start a new spec" ausgewählt werden. Nun Apache Kafka auswählen und auf "Connect data" klicken:

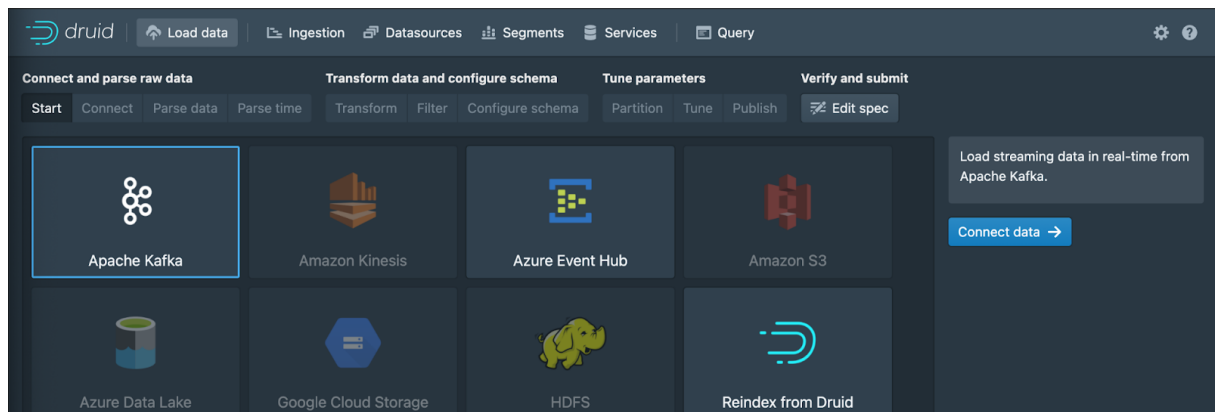


Abbildung 9 Druid Load Data

Zur Konfiguration auf den Button "Edit spec" unter "Verify and submit" klicken. Dort nun der Inhalt der Datei druid.config einfügen und unter "consumer properties" in "bootstrap.servers" die IP-Adresse des Kafka-Clusters mit Port eingeben:

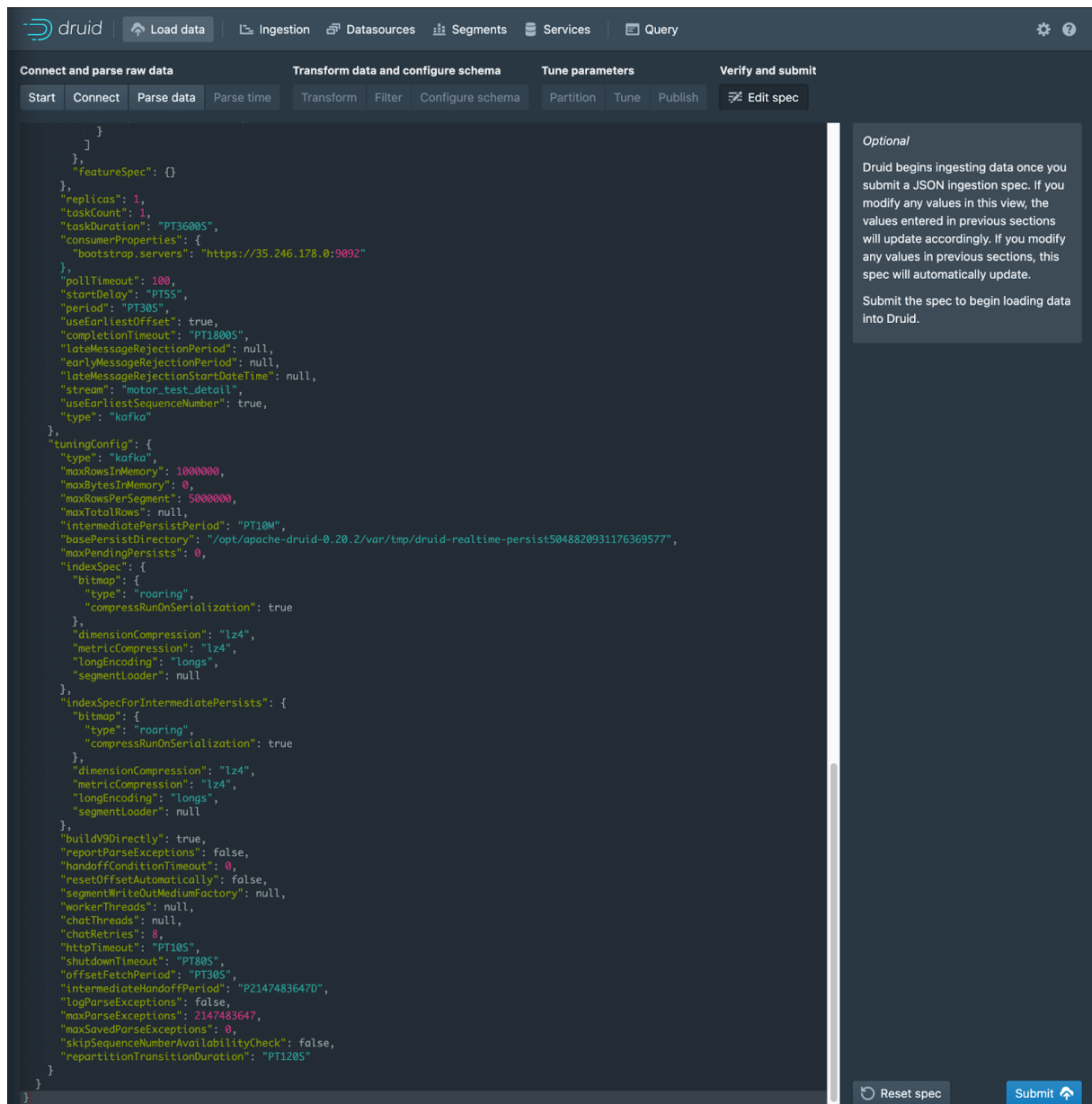


Abbildung 10 Druid Ingestion Specs

Zum Abschluss auf "Submit klicken".

Set up Apache Superset

Eine erfolgreiche Docker und docker-compose Installation wird vorausgesetzt:

<https://docs.docker.com/engine/install/>

<https://docs.docker.com/compose/install/>

Zuerst wird das Superset Repository mit folgendem Command geklont:

“git clone <https://github.com/apache/superset.git>”

Anschließend wechselt man in den Superset-Folder:

“cd superset”

Und führt folgenden Befehl aus:

docker-compose -f docker-compose-non-dev.yml up

```
dominik_pruell@instance-1:~/superset$ sudo docker-compose -f docker-compose-non-dev.yml up
Starting superset_db ... done
Starting superset_cache ... done
Creating superset_worker_beat ... done
Creating superset_app ... done
Creating superset_init ... done
Creating superset_worker ... done
Attaching to superset_db, superset_cache, superset_worker, superset_worker_beat, superset_app, superset_init
```

Abbildung 11 Superset docker-compose

Für das Setup in der VM der Google Cloud war es darüber hinaus auch notwendig eine eingehende Firewall-Regel für Port 8088 zu machen:

Richtung	
Eingehend	
Aktion bei Übereinstimmung	
Zulassen	
Quellfilter	
IP-Bereiche	0.0.0.0/0
Protokolle und Ports	
tcp:8088	
Erzwingung	
Aktiviert	
Statistiken	

Abbildung 12 Port-Forwarding Superset

Superset kann nun lokal unter <http://localhost:8088> oder über die externe IP der VM aufgerufen werden. Die Default Credentials sind:

Username: admin

PW: admin

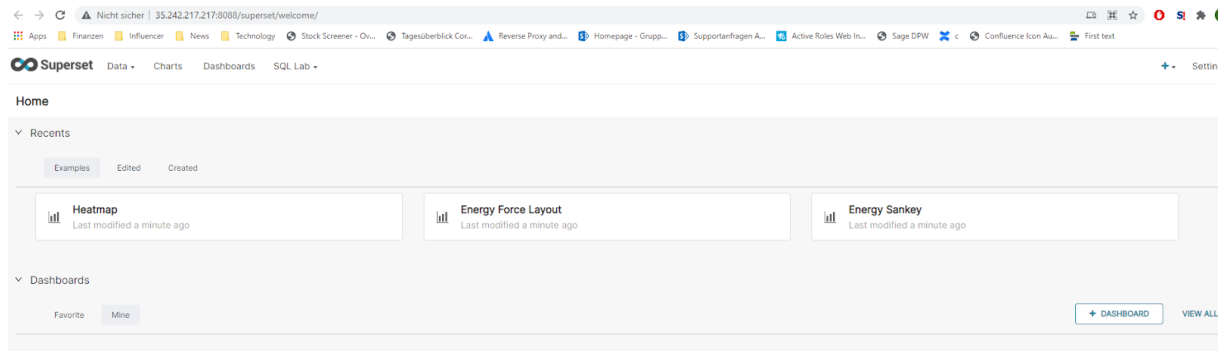


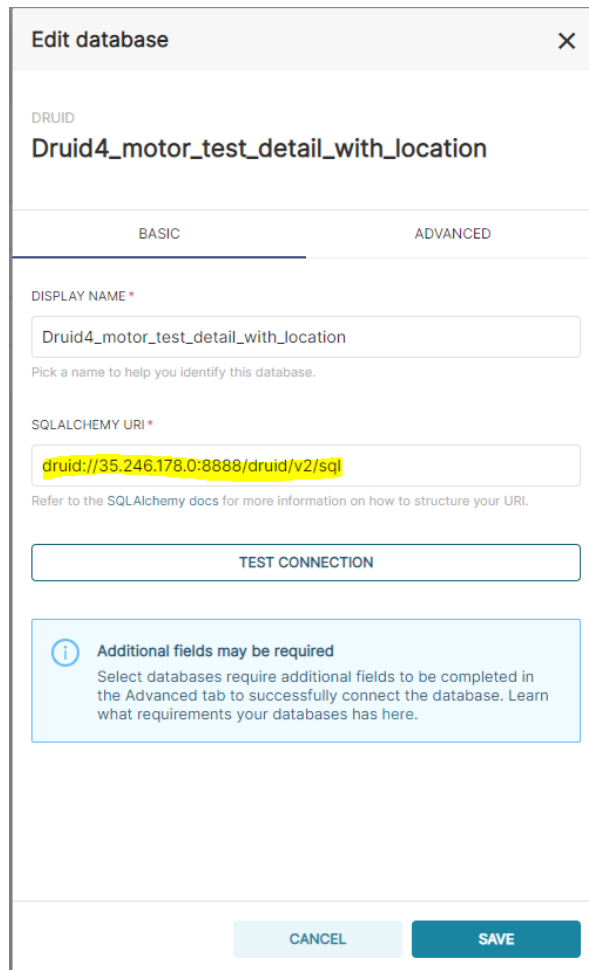
Abbildung 13 Superset Konsole

Superset Database anbinden

Unter "Data > Databases" kann die Druid Data-Source eingebunden werden:

<https://superset.apache.org/docs/databases/druid>

Connection-String: "druid://IP-ADRESSE:8888/druid/v2/sql"



The screenshot shows the 'Edit database' window in Superset. The database type is 'DRUID' and the name is 'Druid4_motor_test_detail_with_location'. The 'BASIC' tab is selected. The 'DISPLAY NAME' field contains the same name. The 'SQLALCHEMY URI' field contains 'druid://35.246.178.0:8888/druid/v2/sql'. Below the URI field is a 'TEST CONNECTION' button. A blue information box states: 'Additional fields may be required. Select databases require additional fields to be completed in the Advanced tab to successfully connect the database. Learn what requirements your databases has here.' At the bottom are 'CANCEL' and 'SAVE' buttons.

Abbildung 14 Superset Database

Achtung: Bei lokalem Setup und eigenen Druid Container, kann es sein dass die Verbindung nicht via "localhost" sondern lokaler IP Adresse hergestellt werden muss.

(In Linux via "ifconfig" einsehbar)

```
dominik@dominik-PC:~$ ifconfig
eth0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST> mtu 1500
    inet 192.168.0.157 netmask 255.255.255.0 broadcast 192.168.0.255
    inet6 2a02:8388:4201:b100:4ddf:696e:9360:6d49 prefixlen 64 scopeid 0x0<global>
    inet6 2a02:8388:4201:b100:850e:b142:6376:3fae prefixlen 128 scopeid 0x0<global>
    inet6 fe80::4ddf:696e:9360:6d49 prefixlen 64 scopeid 0xfd<compat,link,site,host>
    ether bc:5f:f4:cf:67:34 (Ethernet)
    RX packets 0 bytes 0 (0.0 B)
    RX errors 0 dropped 0 overruns 0 frame 0
    TX packets 0 bytes 0 (0.0 B)
    TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0
```

Abbildung 15 ifconfig Beispiel

Dataset und Dashboard Import

Dataset:

<https://superset.apache.org/docs/miscellaneous/importing-exporting-datasources>

Unter Data > Datasets legt man ein neues Dataset mit dem Namen “motor_test_detail” an.

The screenshot shows the 'Edit Dataset' interface in Superset for a dataset named 'motor_test_detail'. The interface has a header bar with the title 'Edit Dataset motor_test_detail' and a close button. Below the header is a yellow warning box that says 'Be careful. Changing these settings will affect all charts using this dataset, including charts owned by other people.' Below the warning box is a tabbed interface with five tabs: 'SOURCE', 'METRICS' (with a blue circle containing the number 1), 'COLUMNS' (with a blue circle containing the number 11), 'CALCULATED COLUMNS' (with a blue circle containing the number 2), and 'SETTINGS'. The 'SOURCE' tab is selected. Under the 'SOURCE' tab, there are two radio buttons: 'Physical (table or view)' (which is selected) and 'Virtual (SQL)'. Below the radio buttons is a section titled 'PHYSICAL' with a help icon. This section contains three dropdown menus: 'Database: druid' (with a sub-menu showing 'Druid4_motor_test_detail_with_loc'), 'Schema: druid', and a table view dropdown showing 'motor_test_detail'. Below these dropdowns is a lock icon and the text 'Click the lock to make changes.' At the bottom right of the interface are three buttons: 'USE LEGACY DATASOURCE EDITOR', 'CANCEL', and 'SAVE'.

Abbildung 16 Superset Dataset

Zusätzlich werden zwei "Calculated Columns definiert":

state_message:

```
CASE WHEN  
"fields.STATE" = 131202  
THEN 'running'  
ELSE 'stopped'  
END
```

state_message STRING ☐ ☒ ☒

SQL EXPRESSION

1	CASE WHEN	
2	"fields.STATE" = 131202	
3	THEN 'running'	
4	ELSE 'stopped'	
5	END	

Abbildung 17 Calculated Column state

temp_upper_limit:

```
SELECT 100 as lim_max  
FROM "druid"."motor_test_detail"  
limit 1
```

temp_upper_limit NUMERIC ☐ ☒ ☒ 1

SQL EXPRESSION

1	SELECT 100 as lim_max	
2	FROM "druid"."motor_test_detail"	
3	limit 1	

LABEL

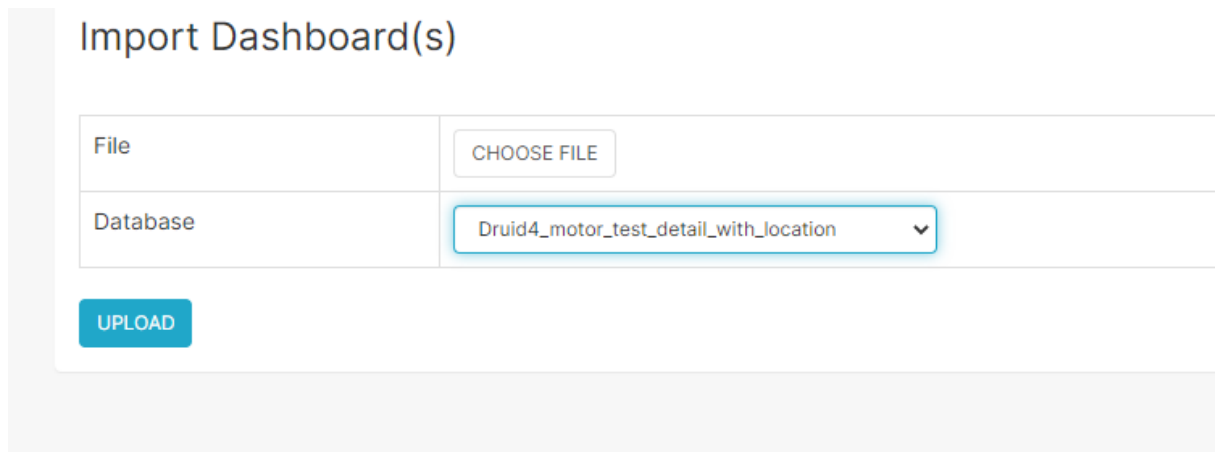
Label

DESCRIPTION

Abbildung 18 Calculated Column temp upper limit

Das Dashboard kann über das mitgelieferte dashboard.json file aus dem Ordner superset importiert werden.

Den Dashboard Import erreicht man rechts oben über "Settings > Import Dashboards". Dort wählt man die Druid Datenbank und importiert das yaml File:



Import Dashboard(s)

File	CHOOSE FILE
Database	Druid4_motor_test_detail_with_location ▼

UPLOAD

Abbildung 19 Superset Dashboard Import

Anschließend findet man das "Motor Dashboard Präsentation" in Dashboards:

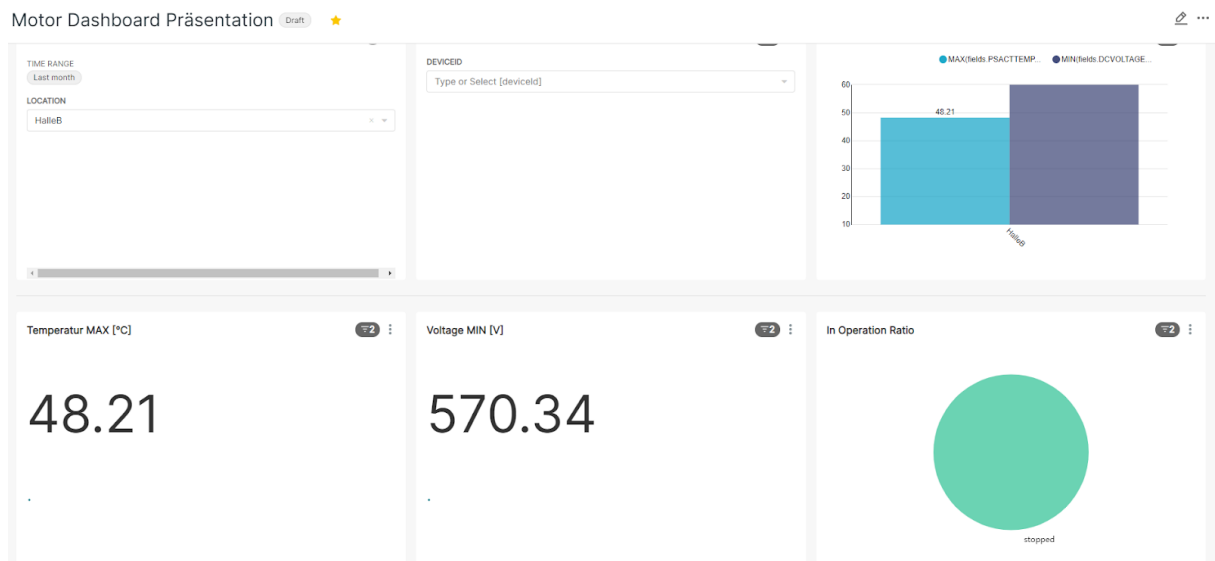


Abbildung 20 Superset Dashboard

Superset Alerts + Reports

Dokumentation: <https://superset.apache.org/docs/installation/alerts-reports>

Das “superset_config.py” File für die Konfiguration von SMTP und Report Settings findet man im Folder “superset/docker/pythonpath_dev/”:

```
roland6bauer@instance-1:/home/dominik_pruell/superset/docker/pythonpath_dev$ sudo vi superset_config.py

ssh.cloud.google.com/projects/boxwood-office-310413/zones/europe-west3-c/instances/instance-1?authuser=0&hl=de&projectNumber=1063257030592&useAdminProxy=true

CELERYD_PREFETCH_MULTIPLIER = 1
CELERY_ACKS_LATE = False
CELERYBEAT_SCHEDULE = {
    "reports.scheduler": {
        "task": "reports.scheduler",
        "schedule": crontab(minute="*", hour="*"),
    },
    "reports.prune_log": {
        "task": "reports.prune_log",
        "schedule": crontab(minute=10, hour=0),
    },
}

CELERY_CONFIG = CeleryConfig

FEATURE_FLAGS = {"ALERT_REPORTS": True}
ALERT_REPORTS_NOTIFICATION_DRY_RUN = False

SMTP_HOST = "smtp.gmail.com"
SMTP_STARTTLS = True
SMTP_SSL = False
SMTP_USER = "kappacapstone@gmail.com"
SMTP_PORT = 587
SMTP_PASSWORD = " "
SMTP_MAIL_FROM = "kappacapstone@gmail.com"

WEBDRIVER_BASEURL = "http://superset:8080/"
# The base URL for the email report hyperlinks.
WEBDRIVER_BASEURL_USER_FRIENDLY = WEBDRIVER_BASEURL

SQLLAB_CTAS_NO_LIMIT = True

#
# Optionally import superset_config_docker.py (which will have been included on
# the PYTHONPATH) in order to allow for local settings to be overridden
#
try:
    import superset_config_docker
    from superset_config_docker import * # noqa
```

Abbildung 21 Superset Config für Alerts

Set up Datasource

Da Sie keinen direkten Zugriff auf die Echtzeitdaten haben, können Sie den Data Stream mit dem folgenden Python Script simulieren. Der Python Script schiebt diese Daten in eine Kafka Topic:

```
from kafka import KafkaProducer
import time
import json
import os

producer = KafkaProducer(bootstrap_servers=["localhost:9092"])
#Replace the ip with your own if needed.

demo_data_file = "demo_data.txt"
file = open(os.path.join(demo_data_file))
topic_name = "motor_data"

while True:
    line = file.readline().rstrip()
    if not line:
        break
    else:
        print("send ... " + line + " ... to topic " + topic_name)
        producer.send(topic_name, bytes(line, 'utf-8'))
        time.sleep(1)

file.close()
print("End of file")
```

Die Demo Daten sind in *demo_data.txt* auffindbar.

Um zu überprüfen, ob die Daten auch tatsächlich in die Kafka Topic geschoben werden, können Sie mit folgendem Command einen Kafka Console Consumer starten, um dies zu überprüfen:

```
sudo docker exec kafka_kafka_1 kafka-console-consumer.sh --topic motor_kafka --from-
beginning --bootstrap-server localhost:9092
```

Abbildungsverzeichnis

Abbildung 1 Projekt erstellen	3
Abbildung 2 Projektübersicht	4
Abbildung 3 VM Erstellung	4
Abbildung 4 VM Einstellungen	5
Abbildung 5 Bootlaufwerk Einstellungen	5
Abbildung 6 Identität und API Einstellungen	6
Abbildung 7 VM Übersicht	6
Abbildung 8 Port-Forwarding Einstellungen	7
Abbildung 9 Druid Load Data	8
Abbildung 10 Druid Ingestion Specs	9
Abbildung 11 Superset docker-compose	10
Abbildung 12 Port-Forwarding Superset	10
Abbildung 13 Superset Konsole	11
Abbildung 14 Superset Database	12
Abbildung 15 ifconfig Beispiel	12
Abbildung 16 Superset Dataset	13
Abbildung 17 Calculated Column state	14
Abbildung 18 Calculated Column temp upper limit	14
Abbildung 19 Superset Dashboard Import	15
Abbildung 20 Superset Dashboard	15
Abbildung 21 Superset Config für Alerts	16