

Data Architecture $\begin{smallmatrix} + \\ \circ \end{smallmatrix} \bullet$

1. Configuración ES-Hadoop

1. Creación de cluster Hadoop en Dataproc

The screenshot shows the Google Cloud Dataproc console interface. The top navigation bar includes the Google Cloud logo, the project name 'kk-data-architecture', a search bar, and various utility icons. The breadcrumb trail indicates the current location: 'Dataproc / Clusters / Cluster: hadoop-elastic-pr / VM instances'.

The main content area is titled 'Cluster details' and includes several action buttons: 'SUBMIT JOB', 'REFRESH', 'START', 'STOP', 'DELETE', and 'VIEW LOGS'. Below these buttons is a URL for the cluster's overview page.

A 'MORE' link is visible above a table of cluster details:

| | |
|--------------|--------------------------------------|
| Name | hadoop-elastic-pr |
| Cluster UUID | f2a49c92-e2d6-4cc6-9e75-37adfd246184 |
| Type | Dataproc Cluster |
| Status | Running |

Below the details table are tabs for 'MONITORING', 'JOBS', 'VM INSTANCES' (which is selected), 'CONFIGURATION', and 'WEB INTERFACES'.

Under the 'VM INSTANCES' tab, there is a 'Filter' section and a table of VM instances:

| | Name | Role | Machine type |
|---|---------------------------------------|--------|---------------|
| ✓ | hadoop-elastic-pr-m | Master | e2-standard-4 |
| ✓ | hadoop-elastic-pr-w-0 | Worker | e2-standard-2 |
| ✓ | hadoop-elastic-pr-w-1 | Worker | e2-standard-2 |

Hadoop interface



Logged in as: dr.who

Nodes of the cluster

Cluster

[About](#)
[Nodes](#)
[Node Labels](#)
[Applications](#)

[NEW](#)
[NEW SAVING](#)
[SUBMITTED](#)
[ACCEPTED](#)
[RUNNING](#)
[FINISHED](#)
[FAILED](#)
[KILLED](#)

[Scheduler](#)

Tools

Cluster Metrics

| Apps Submitted | Apps Pending | Apps Running | Apps Completed | Containers Running | Used Resources | Total Resources | Reserved Resources | Physical Mem Used % | Physical VCores Used % |
|----------------|--------------|--------------|----------------|--------------------|------------------------|-----------------------------|------------------------|---------------------|------------------------|
| 0 | 0 | 0 | 0 | 0 | <memory:0 B, vCores:0> | <memory:12.80 GB, vCores:4> | <memory:0 B, vCores:0> | 17 | 0 |

Cluster Nodes Metrics

| Active Nodes | Decommissioning Nodes | Decommissioned Nodes | Lost Nodes | Unhealthy Nodes | Rebooted Nodes | Shutdown Nodes |
|--------------|-----------------------|----------------------|------------|-----------------|----------------|----------------|
| 2 | 0 | 0 | 0 | 0 | 0 | 0 |

Scheduler Metrics

| Scheduler Type | Scheduling Resource Type | Minimum Allocation | Maximum Allocation | Maximum Cluster Application Priority | Scheduler Busy % |
|--------------------|-------------------------------|----------------------|-------------------------|--------------------------------------|------------------|
| Capacity Scheduler | [memory-mb (unit=Mi), vcores] | <memory:1, vCores:1> | <memory:8554, vCores:2> | 0 | 0 |

Show 20 entries

Search:

| Node Labels | Rack | Node State | Node Address | Node HTTP Address | Last health-update | Health-report | Containers | Allocation Tags | Mem Used | Mem Avail | Phys Mem Used % | VCores Used | VCores Avail | Phys VCores Used % | Version |
|-------------|---------------|------------|---|---|--------------------------------|---------------|------------|-----------------|----------|-----------|-----------------|-------------|--------------|--------------------|---------|
| | /default-rack | RUNNING | hadoop-elastic-pr-w-1.us-west1-a.c.kk-data-architecture.internal:8026 | hadoop-elastic-pr-w-1.us-west1-a.c.kk-data-architecture.internal:8042 | Sun Mar 23 02:42:22 +0000 2025 | | 0 | | 0 B | 6.40 GB | 17 | 0 | 2 | 0 | 3.3.6 |
| | /default-rack | RUNNING | hadoop-elastic-pr-w-0.us-west1-a.c.kk-data-architecture.internal:8026 | hadoop-elastic-pr-w-0.us-west1-a.c.kk-data-architecture.internal:8042 | Sun Mar 23 02:42:18 +0000 2025 | | 0 | | 0 B | 6.40 GB | 17 | 0 | 2 | 0 | 3.3.6 |

Showing 1 to 2 of 2 entries

First Previous 1 Next Last

2. Creación de bucket y carga de archivos jar de configuración

Google Cloud

kk-data-arquitectura

Search (/) for resources, docs, products, and more

Search

49

S

←

Bucket details

GO TO PATH

REFRESH

LEARN

📁

📁

📊

📈

⚙️

🛒

📁

bucket-elastic-pr

Location

Storage class

Public access

Protection

europa-west2 (London)

Standard

Not public

Soft Delete

OBJECTS

CONFIGURATION

PERMISSIONS

PROTECTION

LIFECYCLE

OBSERVABILITY

NEW

INVENTORY REPORTS

OPERATIONS

Folder browser

⏪

📁

bucket-elastic-pr

Buckets > bucket-elastic-pr

CREATE FOLDER

UPLOAD

TRANSFER DATA

OTHER SERVICES



Filter by name prefix only

Filter

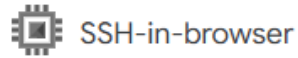
Filter objects and folders

Show

Live objects only

| <input type="checkbox"/> | Name | Size | Type | Created | Storage class | L |
|--------------------------|---|--------|--------------------------|--------------------------|---------------|--------|
| <input type="checkbox"/> |  commons-httpclient-3.1.jar | 305 KB | application/octet-stream | Mar 20, 2025, 4:46:46 AM | Standard | N ⬇️ ⋮ |
| <input type="checkbox"/> |  elasticsearch-hadoop-8.14.1.jar | 2.2 MB | application/octet-stream | Mar 20, 2025, 4:48:02 AM | Standard | N ⬇️ ⋮ |

3. Carga de los archivos jar desde el bucket al cluster desde la consola



```
* Support:      https://ubuntu.com/pro

System information as of Sat Mar 22 04:49:34 UTC 2025

System load:  0.25          Processes:            153
Usage of /:   31.1% of 48.27GB Users logged in:       0
Memory usage: 36%          IPv4 address for ens4: 10.138.0.5
Swap usage:   0%

* Strictly confined Kubernetes makes edge and IoT secure. Learn how MicroK8s
  just raised the bar for easy, resilient and secure K8s cluster deployment.

https://ubuntu.com/engage/secure-kubernetes-at-the-edge

Expanded Security Maintenance for Applications is not enabled.

0 updates can be applied immediately.

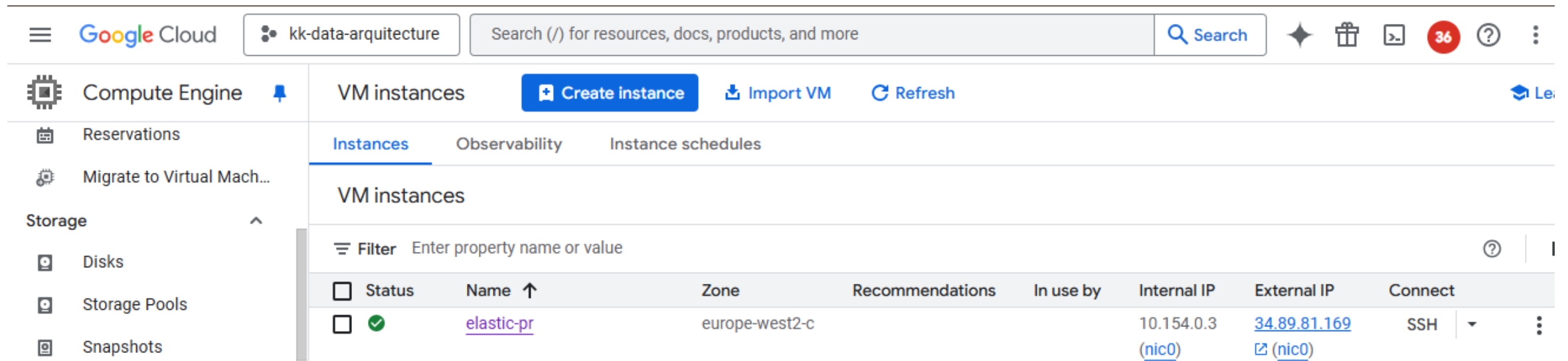
Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status

New release '24.04.2 LTS' available.
Run 'do-release-upgrade' to upgrade to it.

Last login: Sat Mar 22 02:32:38 2025 from 35.235.240.144
keewcripto@hadoop-elastic-pr-m:~$ gsutil cp gs://bucket-elastic-pr/commons-httpclient-3.1.jar .
Copying gs://bucket-elastic-pr/commons-httpclient-3.1.jar...
- [1 files][297.8 KiB/297.8 KiB]
Operation completed over 1 objects/297.8 KiB.
keewcripto@hadoop-elastic-pr-m:~$ gsutil cp gs://bucket-elastic-pr/elasticsearch-hadoop-8.14.1.jar .
Copying gs://bucket-elastic-pr/elasticsearch-hadoop-8.14.1.jar...
- [1 files][ 2.1 MiB/ 2.1 MiB]
Operation completed over 1 objects/2.1 MiB.
keewcripto@hadoop-elastic-pr-m:~$
```

PARTE 2 - Configuración server Elasticsearch


1. Creación VM y configuración de la instancia Elastic-pr



The screenshot shows the Google Cloud Platform console interface. The top navigation bar includes the Google Cloud logo, the project name 'kk-data-architecture', a search bar, and various utility icons. The left sidebar contains navigation links for Compute Engine, Reservations, Migrate to Virtual Mach..., Storage, Disks, Storage Pools, and Snapshots. The main content area is titled 'VM instances' and includes buttons for 'Create instance', 'Import VM', and 'Refresh'. Below this, there are tabs for 'Instances', 'Observability', and 'Instance schedules'. The 'Instances' tab is active, showing a table of VM instances. A filter bar is present above the table. The table has columns for Status, Name, Zone, Recommendations, In use by, Internal IP, External IP, and Connect. One instance, 'elastic-pr', is listed with a status of 'Running' (indicated by a green checkmark), located in the 'europe-west2-c' zone, with an internal IP of '10.154.0.3' and an external IP of '34.89.81.169'. The 'Connect' column shows 'SSH' and a dropdown menu.

| Status | Name | Zone | Recommendations | In use by | Internal IP | External IP | Connect |
|-------------------------------------|----------------------------|----------------|-----------------|-----------|--|--|---------|
| <input checked="" type="checkbox"/> | elastic-pr | europe-west2-c | | | 10.154.0.3 (nic0) | 34.89.81.169 (nic0) | SSH |

2. Descarga e instalación de Elasticsearch y Kibana

 SSH-in-browser

↑ UPLOAD FILE ↓ DOWNLOAD FILE ! ⌨ ⚙

GNU nano 7.2 /etc/elasticsearch/elasticsearch.yml

```
# Enable security features
xpack.security.enabled: false

xpack.security.enrollment.enabled: true

# Enable encryption for HTTP API client connections, such as Kibana, Logstash, and Agents
xpack.security.http.ssl:
  enabled: false
  keystore.path: certs/http.p12

# Enable encryption and mutual authentication between cluster nodes
xpack.security.transport.ssl:
  enabled: true
  verification_mode: certificate
  keystore.path: certs/transport.p12
  truststore.path: certs/transport.p12
# Create a new cluster with the current node only
# Additional nodes can still join the cluster later
cluster.initial_master_nodes: ["elastic-pr"]

# Allow HTTP API connections from anywhere
# Connections are encrypted and require user authentication
http.host: 0.0.0.0

# Allow other nodes to join the cluster from anywhere
# Connections are encrypted and mutually authenticated
#transport.host: 0.0.0.0

#----- END SECURITY AUTO CONFIGURATION -----
|
```

| | | | | | | | | |
|----------------|---------------------|--------------------|-----------------|-------------------|-----------------------|-----------------|---------------------|-----------------------|
| ^G Help | ^O Write Out | ^W Where Is | ^K Cut | ^T Execute | ^C Location | M-U Undo | M-A Set Mark | M-] To Bracket |
| ^X Exit | ^R Read File | ^\\ Replace | ^U Paste | ^J Justify | ^/_ Go To Line | M-E Redo | M-6 Copy | ^Q Where Was |

3. Configuración en Cluster Hadoop de Conexión a ES

1. Configuración de Hive

SSH-in-browser

UPLOAD FILE

DOWNLOAD FILE

hive-server2.service - LSB: Hive Server2

Loaded: loaded (/etc/init.d/hive-server2; generated)

Drop-In: /etc/systemd/system/hive-server2.service.d

└─hive-hbase.conf, restart.conf

Active: active (running) since Sat 2025-03-22 02:54:23 UTC; 2s ago

Docs: man:systemd-sysv-generator(8)

Process: 49899 ExecStart=/etc/init.d/hive-server2 start (code=exited, status=0/SUCCESS)

Main PID: 49910 (java)

Tasks: 0 (limit: 19180)

Memory: 8.0K

CGroup: /system.slice/hive-server2.service

└─ 49910 /usr/lib/jvm/temurin-11-jdk-amd64/bin/java -Dproc_jar -Dhive.log.dir=/var/log/hive

Mar 22 02:54:20 hadoop-elastic-pr-m systemd[1]: Starting LSB: Hive Server2...

Mar 22 02:54:20 hadoop-elastic-pr-m su[49908]: (to hive) root on none

Mar 22 02:54:20 hadoop-elastic-pr-m su[49908]: pam_unix(su:session): session opened for user hive(uid=)

Mar 22 02:54:20 hadoop-elastic-pr-m su[49908]: pam_unix(su:session): session closed for user hive

Mar 22 02:54:23 hadoop-elastic-pr-m hive-server2[49899]: * Started Hive Server2 (hive-server2):

Mar 22 02:54:23 hadoop-elastic-pr-m systemd[1]: Started LSB: Hive Server2.

lines 1-19/19 (END)

keewcripto@hadoop-elastic-pr-m:~\$ SHOW TABLES;

SHOW: command not found

keewcripto@hadoop-elastic-pr-m:~\$ sudo sed -i 's'd' /etc/hive/conf.dist/hive-site.xml

keewcripto@hadoop-elastic-pr-m:~\$ sudo sed -i 's\$a \<property>\n<name>es.nodes</name>\n<value>35.234.149.191</value>\n</property>\n' /etc/hive/conf.dist/hive-site.xml

sudo sed -i 's\$a \<property>\n<name>es.port</name>\n<value>9200</value>\n</property>\n' /etc/hive/conf.dist/hive-site.xml

keewcripto@hadoop-elastic-pr-m:~\$ sudo sed -i 's\$a \<property>\n<name>es.nodes.wan.only</name>\n<value>true</value>\n</property>\n' /etc/hive/conf.dist/hive-site.xml

keewcripto@hadoop-elastic-pr-m:~\$ sudo sed -i 's\$a \<property>\n<name>hive.aux.jars.path</name>\n<value>/usr/lib/hive/lib/elasticsearch-hadoop-8.14.1.jar,/usr/lib/hive/lib/commons-httpclient-3.1.jar</value>\n</property>\n</configuration>' /etc/hive/conf.dist/hive-site.xml

keewcripto@hadoop-elastic-pr-m:~\$ sudo cp elasticsearch-hadoop-8.14.1.jar /usr/lib/hive/lib/

keewcripto@hadoop-elastic-pr-m:~\$ sudo cp commons-httpclient-3.1.jar /usr/lib/hive/lib/

keewcripto@hadoop-elastic-pr-m:~\$ sudo service hive-server2 restart

keewcripto@hadoop-elastic-pr-m:~\$

4. A conectar datos

1. Creación de un índice en Elasticsearch



SSH-in-browser

↑ UPLOAD FILE

↓ DOWNLOAD FILE



```
Linux elastic-pr 6.1.0-32-cloud-amd64 #1 SMP PREEMPT_DYNAMIC Debian 6.1.129-1 (2025-03-06) x86_64
```

```
The programs included with the Debian GNU/Linux system are free software;  
the exact distribution terms for each program are described in the  
individual files in /usr/share/doc/*/copyright.
```

```
Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent  
permitted by applicable law.
```

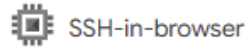
```
Last login: Sat Mar 22 04:57:29 2025 from 35.235.244.114
```

```
keewcripto@elastic-pr:~$ curl -X POST "localhost:9200/alumnos/_doc/6" -H 'Content-Type: application/json' -d'
```

```
{  
  "title": "New Document",  
  "content": "This is a new document for the master class",  
  "tag": ["general", "testing"]  
}
```

```
{ "_index": "alumnos", "_id": "6", "_version": 1, "result": "created", "_shards": { "total": 2, "successful": 1, "failed": 0 }, "_seq_no": 0, "_primary_term": 1 } keewcripto@elastic-pr:~$
```

2. Agregación de datos al índice



```
Welcome to Ubuntu 22.04.5 LTS (GNU/Linux 6.8.0-1025-gcp x86_64)

* Documentation:  https://help.ubuntu.com
* Management:    https://landscape.canonical.com
* Support:       https://ubuntu.com/pro

System information as of Sun Mar 23 00:16:48 UTC 2025

System load: 0.08          Processes:            148
Usage of /:  31.1% of 48.27GB Users logged in:        1
Memory usage: 37%          IPv4 address for ens4: 10.138.0.5
Swap usage:  0%

=> There is 1 zombie process.

* Strictly confined Kubernetes makes edge and IoT secure. Learn how MicroK8s
  just raised the bar for easy, resilient and secure K8s cluster deployment.

https://ubuntu.com/engage/secure-kubernetes-at-the-edge

Expanded Security Maintenance for Applications is not enabled.

0 updates can be applied immediately.

Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status

New release '24.04.2 LTS' available.
Run 'do-release-upgrade' to upgrade to it.

Last login: Sun Mar 23 00:11:11 2025 from 35.235.248.1
keewcripto@hadoop-elastic-pr-m:~$ curl -X POST "http://34.147.207.133:9200/_bulk" -H 'Content-Type: application/json' -d'
{ "index": { "_index": "alumnos", "_id": "3" } }
{ "id": 3, "name": "Carlos", "last_name": "González" }
{ "index": { "_index": "alumnos", "_id": "4" } }
{ "id": 4, "name": "Maria", "last_name": "López" }
{ "index": { "_index": "alumnos", "_id": "5" } }
{ "id": 5, "name": "Luis", "last_name": "Martínez" }
{ "index": { "_index": "alumnos", "_id": "7" } }
{ "id": 7, "name": "Sofia", "last_name": "Ramírez" }
{ "index": { "_index": "alumnos", "_id": "8" } }
{ "id": 8, "name": "Pedro", "last_name": "Hernández" }
'
```

```
{
  "errors": false,
  "took": 110,
  "items": [
    {
      "index": {
        "_index": "alumnos",
        "_id": "3",
        "_version": 1,
        "result": "created",
        "_shards": {
          "total": 2,
          "successful": 1,
          "failed": 0,
          "_seq_no": 1,
          "_primary_term": 4,
          "status": 201
        }
      },
      "index": {
        "_index": "alumnos",
        "_id": "4",
        "_version": 1,
        "result": "created",
        "_shards": {
          "total": 2,
          "successful": 1,
          "failed": 0,
          "_seq_no": 2,
          "_primary_term": 4,
          "status": 201
        }
      },
      "index": {
        "_index": "alumnos",
        "_id": "5",
        "_version": 1,
        "result": "created",
        "_shards": {
          "total": 2,
          "successful": 1,
          "failed": 0,
          "_seq_no": 3,
          "_primary_term": 4,
          "status": 201
        }
      },
      "index": {
        "_index": "alumnos",
        "_id": "7",
        "_version": 1,
        "result": "created",
        "_shards": {
          "total": 2,
          "successful": 1,
          "failed": 0,
          "_seq_no": 4,
          "_primary_term": 4,
          "status": 201
        }
      },
      "index": {
        "_index": "alumnos",
        "_id": "8",
        "_version": 1,
        "result": "created",
        "_shards": {
          "total": 2,
          "successful": 1,
          "failed": 0,
          "_seq_no": 5,
          "_primary_term": 4,
          "status": 201
        }
      }
    ]
  }
}

keewcripto@hadoop-elastic-pr-m:~$ curl -X GET "http://34.147.207.133:9200/alumnos/_search?pretty"
curl -X GET "http://34.147.207.133:9200/alumnos/_search?pretty"
```

```
{
  "took" : 4,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 6,
      "relation" : "eq"
    },
    "max_score" : 1.0,
    "hits" : [
      {
        "_index" : "alumnos",
        "_id" : "6",
        "_score" : 1.0,
        "_source" : {
          "title" : "New Document",
          "content" : "This is a new document for the master class",
          "tag" : [
            "general",
            "testing"
          ]
        }
      },
      {
        "_index" : "alumnos",
        "_id" : "3",
        "_score" : 1.0,
        "_source" : {
          "id" : 3,
          "name" : "Carlos",
          "last_name" : "González"
        }
      }
    ]
  }
}
```

5. KIBANA

