

Big Data Architecture

Final practical deliverable

by Ariane (Ariadna) Heinz Vallribera

PART 1 SCREENSHOTS:

SSH console for the Hadoop cluster once the loading of the configuration files is complete:

ssh.cloud.google.com/v2/ssh/projects/big-data-edition-xvi-ahv/zones/europe-west4-b/instances/es-hadoop-cluster-ahv-m?authu...

SSH en el navegador

SUBIR ARCHIVODESCARGAR ARCHIVO

Linux es-hadoop-cluster-ahv-m 6.1.0-40-cloud-amd64 #1 SMP PREEMPT_DYNAMIC Debian 6.1.153-1 (2025-09-20) 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 Oct 18 23:37:00 2025 from 35.235.245.146

enairaznieh@es-hadoop-cluster-ahv-m:~\$ gsutil cp gs://es-hadoop-bucket-ahv1/jars/elastic/elasticsearch-hadoop-8.14.1.jar .

Copying gs://es-hadoop-bucket-ahv1/jars/elastic/elasticsearch-hadoop-8.14.1.jar...

/ [1 files][2.1 MiB/ 2.1 MiB]

Operation completed over 1 objects/2.1 MiB.

enairaznieh@es-hadoop-cluster-ahv-m:~\$ gsutil cp gs://es-hadoop-bucket-ahv1/jars/elastic/commons-httpclient-3.1.jar .

Copying gs://es-hadoop-bucket-ahv1/jars/elastic/commons-httpclient-3.1.jar...

/ [1 files][297.8 KiB/297.8 KiB]

Operation completed over 1 objects/297.8 KiB.

enairaznieh@es-hadoop-cluster-ahv-m:~\$

Cluster with 2 worker nodes:

Clústeres

Create cluster

Actualizar

Iniciar

Detener

Borrar

Regiones

+ 5 alertas recomendadas

Filtro

Busca el clúster por propiedades y presiona Intro

El servidor no pudo completar tu solicitud.

	Nombre	Estado	Región	Zona	Versión de la imagen base	Total de nodos trabajadores	¿Tiene VMs
<input type="checkbox"/>	es-hadoop-cluster-ahv	En ejecución	europe-west4	europe-west4-b	2.2.68-debian12	2	No

Bucket that contains configuration files:

es-hadoop-bucket-ahv1

Ubicación

Clase de almacenamiento

Acceso público

Protección

europe-west4 (Países Bajos)

Standard

No público

Borrar de forma no definitiva

Objetos

Configuración

Permisos

Protección

Ciclo de vida

Observabilidad

Nuevo

Informes de inventario

Operaciones

Navegador de carpetas

es-hadoop-bucket-ahv1

jars/elastic/

Depósitos

es-hadoop-bucket-ahv1

jars

elastic

Crear carpeta

Subir

Transferir los datos

Otros servicios

Filtrar solo por prefijo de nombre

Filtro

Filtrar objetos y carpetas

	Nombre	Tamaño	Tipo	Fecha de creación
<input type="checkbox"/>	commons-httpclient-3.1.jar	305 KB	application/java-archive	19 oct 2025 01:32:53
<input type="checkbox"/>	elasticsearch-hadoop-8.14.1.jar	2.2 MB	application/java-archive	19 oct 2025 01:29:18

PART 2 SCREENSHOTS:

Configuration changes in the elasticsearch.yml file:

ssh.cloud.google.com/v2/ssh/projects/big-data-edition-xvi-ahv/zones/europe-west4-b/instances/elastic-instance-ah

SSH en el navegador

GNU nano 7.2 /etc/elasticsearch/elasticsearch.yml

```
# The following settings, TLS certificates, and keys have been automatically
# generated to configure Elasticsearch security features on 19-10-2025 00:06:39
#
# -----

# 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-instance-ahv"]

# 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

^X Exit

^R Read File

^N Replace

^U Paste

^J Justify

^_ Go To Line

Verification that the cluster can indeed access the Elastic machine:

```
enairaznieh@es-hadoop-cluster-ahv-m:~$ ping 34.7.74.62
PING 34.7.74.62 (34.7.74.62) 56(84) bytes of data.
64 bytes from 34.7.74.62: icmp_seq=1 ttl=61 time=1.12 ms
64 bytes from 34.7.74.62: icmp_seq=2 ttl=61 time=0.738 ms
64 bytes from 34.7.74.62: icmp_seq=3 ttl=61 time=0.439 ms
64 bytes from 34.7.74.62: icmp_seq=4 ttl=61 time=0.645 ms
64 bytes from 34.7.74.62: icmp_seq=5 ttl=61 time=0.539 ms
```

Additional firewalls I created manually:

<input type="checkbox"/>	elastic-kibana-firewall-1	Entrada	Aplicar a	Rangos de IP:	tcp:5601, 9200	Permitir	1000
<input type="checkbox"/>	hadoop-firewall-1	Entrada	Aplicar a	Rangos de IP:	tcp:8088, 9870	Permitir	1000

PART 3 SCREENSHOTS:

Configuration process of connection with ES in Hadoop Cluster:

```
enairaznieh@es-hadoop-cluster-ahv-m:~$ sudo sed -i 's$' /etc/hive/conf.dist/hive-site.xml
enairaznieh@es-hadoop-cluster-ahv-m:~$ sudo sed -i 's$ \ <property>\n      <name>es.nodes</name>\n      <value>AQUÍ LA IP DE ELASTIC</valu
e>\n    </property>\n' /etc/hive/conf.dist/hive-site.xml
enairaznieh@es-hadoop-cluster-ahv-m:~$ sudo sed -i 's$ \ <property>\n      <name>es.port</name>\n      <value>9200</value>\n    </property>\n' /etc/hive/conf.dist/hive-site.xml
enairaznieh@es-hadoop-cluster-ahv-m:~$ sudo sed -i 's$ \ <property>\n      <name>es.nodes.wan.only</name>\n      <value>true</value>\n    </
property>\n' /etc/hive/conf.dist/hive-site.xml
enairaznieh@es-hadoop-cluster-ahv-m:~$ ^C
enairaznieh@es-hadoop-cluster-ahv-m:~$ sudo sed -i 's$ \ <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.d
ist/hive-site.xml
enairaznieh@es-hadoop-cluster-ahv-m:~$ sudo cp elasticsearch-hadoop-8.14.1.jar /usr/lib/hive/lib/
enairaznieh@es-hadoop-cluster-ahv-m:~$ sudo cp commons-httpclient-3.1.jar /usr/lib/hive/lib/
enairaznieh@es-hadoop-cluster-ahv-m:~$ sudo systemctl restart hive-server2
enairaznieh@es-hadoop-cluster-ahv-m:~$ sudo systemctl restart hive-metastore
enairaznieh@es-hadoop-cluster-ahv-m:~$ jps | grep Hive
enairaznieh@es-hadoop-cluster-ahv-m:~$ which hive
/usr/bin/hive
enairaznieh@es-hadoop-cluster-ahv-m:~$ ls /etc/hive/conf.dist/
beeline-log4j2.properties.template  hive-exec-log4j2.properties  ivysettings.xml  parquet-logging.properties
hive-default.xml.template           hive-log4j2.properties      llap-cli-log4j2.properties.template
hive-env.sh                         hive-site.xml               llap-daemon-log4j2.properties.template
enairaznieh@es-hadoop-cluster-ahv-m:~$
```

The screenshot also shows the three commands I ran to ensure that Hive is indeed in the es-hadoop cluster:

- The command `jps | grep Hive` resulted in no output, signifying that no hive-related java processes were running at the moment.
- The command `which hive` returned the location where the *hive* executable is located.
- The command `ls /etc/hive/conf.dist/` returned the list of files in the configuration directory of *hive*, indicating that they are indeed there.

PART 4 SCREENSHOTS:

Result of query from Hadoop cluster:

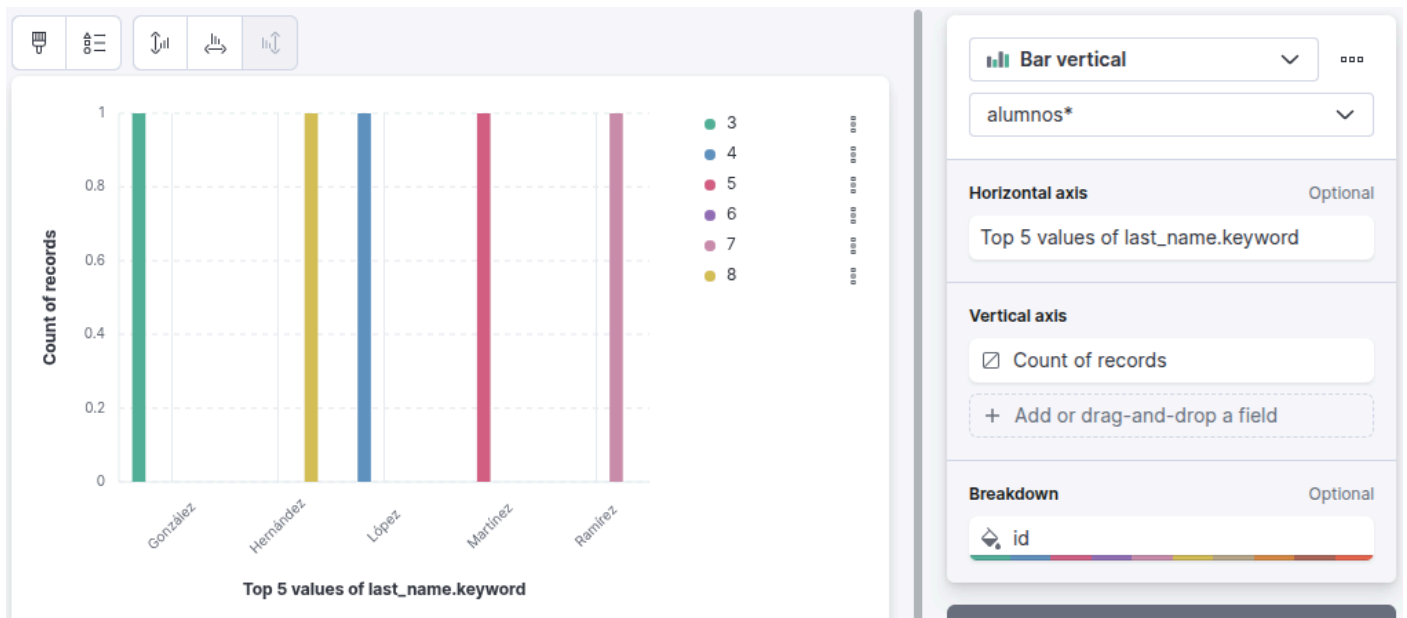
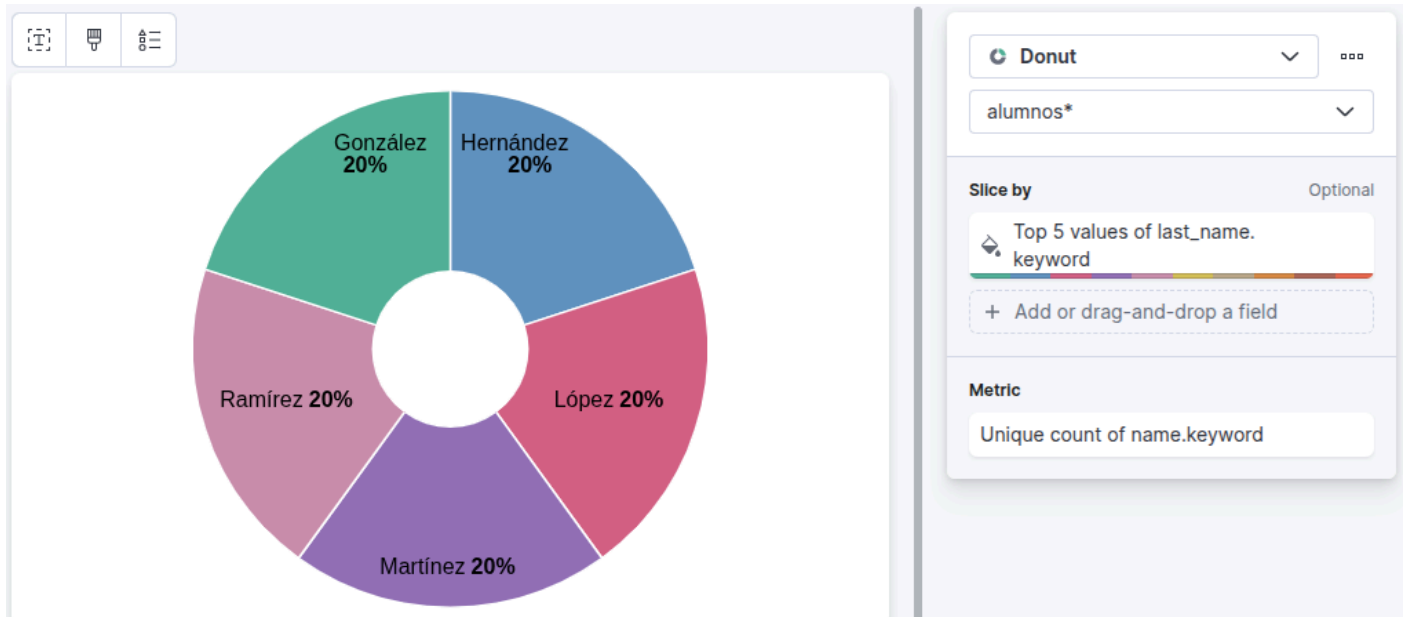
```
enairaznieh@es-hadoop-cluster-ahv-m:~$ curl -X GET "http://34.7.74.62:9200/alumnos/_search?pretty"
{
  "took" : 11,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 6,
      "relation" : "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"
        }
      },
      {
        "_index" : "alumnos",
        "_id" : "4",
        "_score" : 1.0,
        "_source" : {
          "id" : 4,
          "name" : "María",
          "last_name" : "López"
        }
      }
    ]
  }
}
```

(continues on the next page...)

```
      "name" : "Maria",  
      "last_name" : "López"  
    },  
    {  
      "_index" : "alumnos",  
      "_id" : "5",  
      "_score" : 1.0,  
      "_source" : {  
        "id" : 5,  
        "name" : "Luis",  
        "last_name" : "Martínez"  
      }  
    },  
    {  
      "_index" : "alumnos",  
      "_id" : "7",  
      "_score" : 1.0,  
      "_source" : {  
        "id" : 7,  
        "name" : "Sofía",  
        "last_name" : "Ramírez"  
      }  
    },  
    {  
      "_index" : "alumnos",  
      "_id" : "8",  
      "_score" : 1.0,  
      "_source" : {  
        "id" : 8,  
        "name" : "Pedro",  
        "last_name" : "Hernández"  
      }  
    }  
  ]  
}  
}
```

enairaznieh@es-hadoop-cluster-ahv-m:~\$

PART 5 SCREENSHOTS:



Firewalls involved in this project (in magenta, the ones I set up manually):

Big Data Edition XVI AHV

Buscar (/) recursos, documentos, productos y más

Buscar

11

Políticas de firewall

Crear política de firewall

Crear regla de firewall

Más información

<div>Filtro</div> <div>Escribir el nombre o valor de la propiedad</div>							
<input type="checkbox"/>	Nombre	Tipo	Destinos	Filtros	Protocolos/puertos	Acción	
<input type="checkbox"/>	default-allow-http	Entrada	http-server	Rangos de IP:	tcp:80	Permitir	▼
<input type="checkbox"/>	default-allow-https	Entrada	https-server	Rangos de IP:	tcp:443	Permitir	▼
<input type="checkbox"/>	elastic-kibana-firewall-1	Entrada	Aplicar a	Rangos de IP:	tcp:5601, 9200, 10000	Permitir	▼
<input type="checkbox"/>	hadoop-firewall-1	Entrada	Aplicar a	Rangos de IP:	tcp:8088, 9870	Permitir	▼
<input type="checkbox"/>	default-allow-icmp	Entrada	Aplicar a	Rangos de IP:	icmp	Permitir	▼
<input type="checkbox"/>	default-allow-internal	Entrada	Aplicar a	Rangos de IP:	tcp:0-65535 udp:0-65535 icmp	Permitir	▼
<input type="checkbox"/>	default-allow-rdp	Entrada	Aplicar a	Rangos de IP:	tcp:3389	Permitir	▼
<input type="checkbox"/>	default-allow-ssh	Entrada	Aplicar a	Rangos de IP:	tcp:22	Permitir	▼