Entrega Big Data Architecture

Entregable parte 1:

```
ruijiagu@cluster-hadoop-practica-m:~$ gsutil cp gs://bucket-para-elastic-practica/jars/ela
stic/elasticsearch-hadoop-8.14.1.jar .
Copying gs://bucket-para-elastic-practica/jars/elastic/elasticsearch-hadoop-8.14.1.jar...
/ [1 files][  2.1 MiB/  2.1 MiB]
Operation completed over 1 objects/2.1 MiB.
ruijiagu@cluster-hadoop-practica-m:~$ gsutil cp gs://bucket-para-elastic-practica/jars/ela
stic/commons-httpclient-3.1.jar .
Copying gs://bucket-para-elastic-practica/jars/elastic/commons-httpclient-3.1.jar...
/ [1 files][297.8 KiB/297.8 KiB]
Operation completed over 1 objects/297.8 KiB.
ruijiagu@cluster-hadoop-practica-m:~$ ls -ltr
total 2472
-rw-r--r-- 1 ruijiagu ruijiagu 2224041 Oct 12 12:19 elasticsearch-hadoop-8.14.1.jar
-rw-r--r-- 1 ruijiagu ruijiagu  305001 Oct 12 12:20 commons-httpclient-3.1.jar
```

Entregable Parte 2:

```
# ---------------------------------------------------------------------------

# 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: ["vm-para-elastic-practica"]

# 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 -------------------------
```

Entregable parte 3:

```
ruijiagu@cluster-hadoop-practica-m:~$ sudo sed -i '$d' /etc/hive/conf.dist/hive-site.xml
ruijiagu@cluster-hadoop-practica-m:~$ sudo cat /etc/hive/conf.dist/hive-site.xml
```

Aquí he ejecutado un sudo cat para ver el contenido antes de realizar los cambios (más precisamente, después de eliminar la última línea del archivo de configuración y antes de aplicar el resto de las modificaciones).

```
ruijiagu@cluster-hadoop-practica-m:~$ sudo sed -i '$a \  <property>\n    <name>es.nodes</name>\n    <value>AQUÍ LA IP DE ELASTIC</value>\n  </property>\n' /etc/hive/conf.dist/hive-site.xml
ruijiagu@cluster-hadoop-practica-m:~$ sudo cat /etc/hive/conf.dist/hive-site.xml
```

```
<property>
  <name>es.nodes</name>
  <value>AQUÍ LA IP DE ELASTIC</value>
</property>
```

Aquí, al ver el resultado del sudo cat, me di cuenta de que no había puesto la IP de ELASTIC, por lo que después utilicé sudo nano para abrir el archivo y añadirla.

```
ruijiagu@cluster-hadoop-practica-m:~$ sudo nano /etc/hive/conf.dist/hive-site.xml
ruijiagu@cluster-hadoop-practica-m:~$ sudo cat /etc/hive/conf.dist/hive-site.xml
```

Después de editar la IP con sudo nano, en el resultado de sudo cat ya puedo comprobar que la IP de ELASTIC está correctamente colocada.

```
<property>
  <name>es.nodes</name>
  <value>34.175.246.18</value>
</property>
```

Continué con el resto de la configuración.

```
ruijiagu@cluster-hadoop-practica-m:~$ sudo sed -i '$a \  <property>\n    <name>es.port</name>\n    <value>9200</value>\n  </property>\n' /etc/hive/conf.dist/hive-site.xml
ruijiagu@cluster-hadoop-practica-m:~$ sudo sed -i '$a \  <property>\n    <name>es.nodes.wan.only</name>\n    <value>true</value>\n  </property>\n' /etc/hive/conf.dist/hive-site.xml
ruijiagu@cluster-hadoop-practica-m:~$ sudo sed -i '$a \  <property>\n    <name>hive.aux.jars.path</name>\n    <value>/usr/lib/hive/lib/elasticsearch-hadoop-8.14.1.jar,/usr/lib/hive/lib/commo
ns-httpclient-3.1.jar</value>\n  </property>\n</configuration>' /etc/hive/conf.dist/hive-site.xml
ruijiagu@cluster-hadoop-practica-m:~$ sudo cp elasticsearch-hadoop-8.14.1.jar /usr/lib/hive/lib/
ruijiagu@cluster-hadoop-practica-m:~$ ls
commons-httpclient-3.1.jar  elasticsearch-hadoop-8.14.1.jar
ruijiagu@cluster-hadoop-practica-m:~$ sudo cp commons-httpclient-3.1.jar /usr/lib/hive/lib/
ruijiagu@cluster-hadoop-practica-m:~$ ls
commons-httpclient-3.1.jar  elasticsearch-hadoop-8.14.1.jar
ruijiagu@cluster-hadoop-practica-m:~$ sudo systemctl restart hi
hibernate.target        hive-metastore.service  hive-server2.service
ruijiagu@cluster-hadoop-practica-m:~$ sudo systemctl restart hive-server2.service
ruijiagu@cluster-hadoop-practica-m:~$ sudo systemctl status hive-server2
● 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 Sun 2025-10-12 13:11:17 UTC; 1min 2s ago
       Docs: man:systemd-sysv-generator(8)
    Process: 21873 ExecStart=/etc/init.d/hive-server2 start (code=exited, status=0/SUCCES>
   Main PID: 21885 (java)
      Tasks: 0 (limit: 9513)
     Memory: 52.0K
     CGroup: /system.slice/hive-server2.service
             ‣ 21885 /usr/lib/jvm/temurin-11-jdk-amd64/bin/java -Dproc_jar -Dhive.log.dir>

Oct 12 13:11:14 cluster-hadoop-practica-m systemd[1]: Starting hive-server2.service - LSB>
Oct 12 13:11:14 cluster-hadoop-practica-m su[21883]: (to hive) root on none
Oct 12 13:11:14 cluster-hadoop-practica-m su[21883]: pam_unix(su:session): session opened>
Oct 12 13:11:14 cluster-hadoop-practica-m su[21883]: pam_unix(su:session): session closed>
Oct 12 13:11:17 cluster-hadoop-practica-m hive-server2[21873]: Started Hive Server2 (hive>
Oct 12 13:11:17 cluster-hadoop-practica-m systemd[1]: Started hive-server2.service - LSB:>
```

Al final, después de reiniciar Hive, ejecuté un systemctl para comprobar su estado.

Además de eso, ejecuté un sudo cat al finalizar toda la configuración, y así es como se ve el archivo de configuración después de realizar todos los cambios. (Solo muestro la parte final del archivo, que corresponde a los cambios que realizamos.).

```
  </property>
  <property>
    <name>hive.exec.input.listing.max.threads</name>
    <value>20</value>
  </property>
  <property>
    <name>hive.fetch.task.conversion</name>
    <value>minimal</value>
  </property>
  <property>
    <name>es.nodes</name>
    <value>34.175.246.18</value>
  </property>

  <property>
    <name>es.port</name>
    <value>9200</value>
  </property>

  <property>
    <name>es.nodes.wan.only</name>
    <value>true</value>
  </property>

  <property>
    <name>hive.aux.jars.path</name>
   <value>/usr/lib/hive/lib/elasticsearch-hadoop-8.14.1.jar,/usr/lib/hive/lib/commons-httpclient-3.1.jar</value>
  </property>
</configuration>
```

Entregable parte 4:

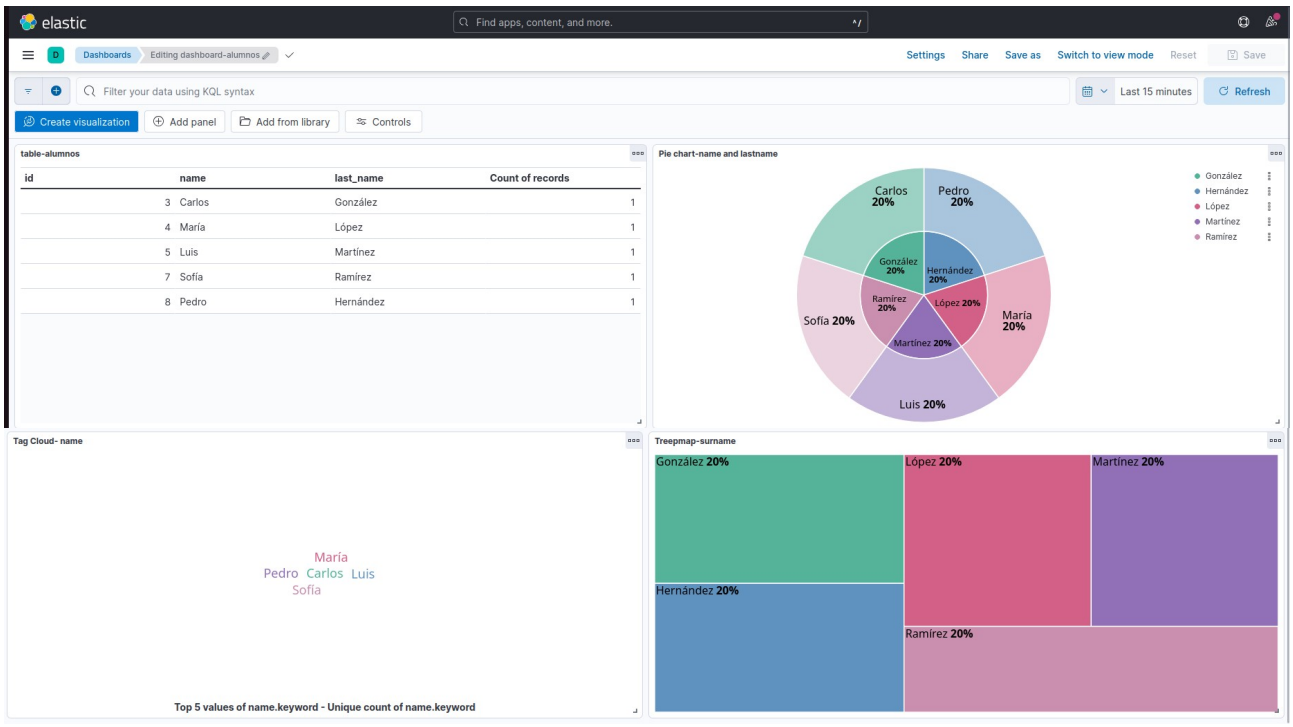
```
ruijiagu@cluster-hadoop-practica-m:~$ curl -X GET "http://34.175.246.18:9200/alumnos/_sear
ch?pretty"
{
  "took" : 8,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 6,
      "relation" : "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",
```

Entregable parte 5:

Al observar los datos, noté que solo contienen el nombre, el apellido y el ID de cada alumno. Por ello, he creado una tabla que muestra los nombres y apellidos de todos ellos.
Cuando se arrastra un campo de texto (como last_name keyword) a “Rows”, Kibana aplica automáticamente una agregación de tipo “Top values”, es decir, solo muestra los *N* valores más frecuentes (por defecto, los cinco primeros). En este caso, para visualizar todas las filas, seleccioné “Top 5 values” (ya que tenemos cinco alumnos en total) y luego cambié el nombre de las columnas a id, name y last_name para que sean legibles.

Para experimentar un poco con los gráficos, también creé un gráfico circular (Pie chart) para visualizar el porcentaje de nombres y apellidos (cada uno representa un 20 %, ya que no se repite ningún nombre ni apellido).
Además, elaboré una nube de etiquetas (Tag cloud) con los nombres y un Treemap con los apellidos.

Finalmente, el dashboard se ve así:



También importé el conjunto de datos *shakespeare.json* en Kibana para analizar y experimentar con distintas visualizaciones.