Implementación de un Clúster Hadoop para Analítica Distribuida y Modelamiento Predictivo

Examen Parcial — Big Data y Computación Distribuida

Franklin Espinoza
Chavez Chico Joel

17 de octubre de 2025

Índice

Resumen Ejecutivo

Este informe documenta el **diseño, levantamiento y validación** de un clúster Hadoop (HDFS+YARN) de tres nodos (1 maestro, 2 trabajadores) sobre máquinas virtuales, y la **ejecución de consultas MapReduce y modelos de aprendizaje automático** (regresión y clasificación) sobre un dataset real de aceleraciones sísmicas. Se presentan evidencias de instalación, configuración de red *host-only*, SSH sin contraseña, despliegue de servicios, **monitoreo de recursos con `htop`**, y **comparativas de rendimiento** entre ejecución distribuida (YARN) y de un solo nodo (framework local).

Los resultados incluyen: (i) tres consultas de estadística descriptiva (promedio, mediana, desviación estándar), (ii) dos consultas con *pipelines* de **tres MapReduce encadenados**, y (iii) dos consultas de **ML** (una de regresión y otra de clasificación). Finalmente, se presentan **tablas de tiempos/speedup** y **gráficos en Power BI**.

1. Introducción

El objetivo es **construir y demostrar** un flujo de analítica distribuida con Apache Hadoop, desde cero: aprovisionamiento de VMs, red privada *host-only*, instalación de Java/Hadoop, configuración de HDFS/YARN y ejecución de trabajos MapReduce y ML. La motivación es **medir y argumentar** el beneficio del paralelismo frente a la ejecución de un solo nodo, cuantificando *speedup* y uso de recursos.

1.1. Objetivos específicos

- Desplegar un clúster Hadoop funcional (1 master, 2 slaves) en VirtualBox.
- Ingerir y preparar el dataset; diseñar consultas y *pipelines* MapReduce.
- Entrenar y evaluar un modelo de **regresión** y otro de **clasificación**.
- Comparar tiempos y recursos entre ejecución distribuida y local.
- Documentar reproducibilidad (comandos, archivos de configuración y código).

2. Marco teórico

2.1. Hadoop: HDFS, YARN y MapReduce

HDFS provee almacenamiento distribuido y tolerante a fallos; **YARN** administra recursos y la cola de trabajos; **MapReduce** es el modelo de cómputo paralelo por lotes.

2.2. Estadística descriptiva

Promedio, mediana y desviación estándar como métricas base para caracterizar la distribución de variables continuas.

2.3. Aprendizaje automático

Regresión lineal para variables continuas; regresión logística binaria para clasificación.

3. Metodología

3.1. Entorno de trabajo

Topología del clúster

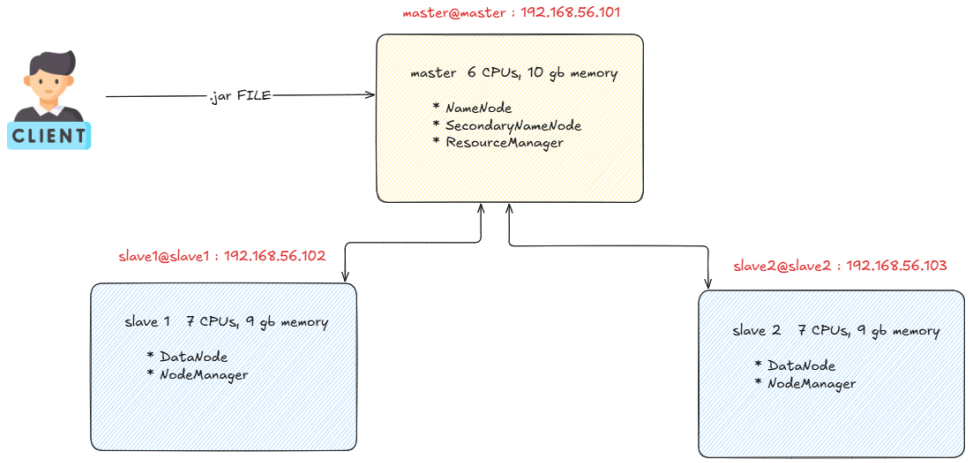

Figura 1: Arquitectura lógica: `master` (NameNode, SecondaryNameNode, ResourceManager) y `slave1`/`slave2` (DataNode, NodeManager).

Red *host-only* y direccionamiento

Se configuró un **adaptador host-only** en VirtualBox, deshabilitando DHCP y asignando IPs estáticas a las VMs mediante `netplan`. Esto permite realizar conexiones por SSH desde el host y entre nodos.

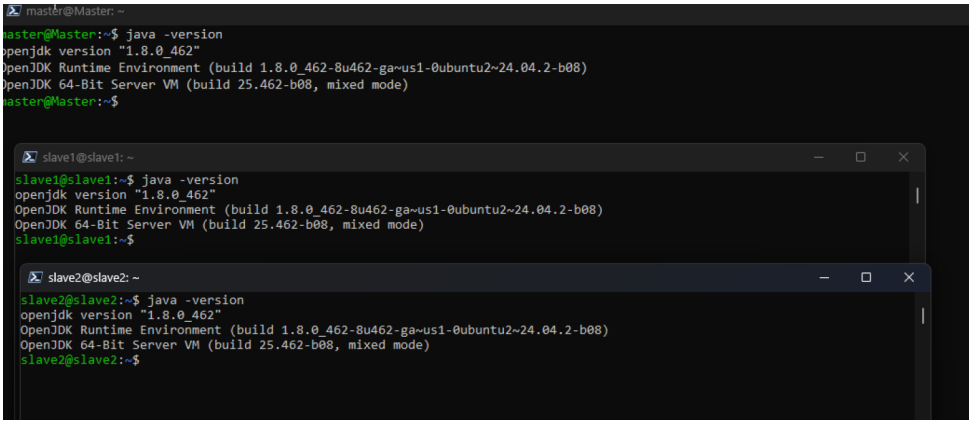

Figura 2: Conexiones SSH a los nodos maestro y esclavos.

3.2. Herramientas utilizadas

- **VirtualBox** (VMs Ubuntu Server 24.04 LTS).
- **OpenJDK 8/11** y **Apache Hadoop** (HDFS + YARN).
- **htop** para monitoreo de CPU/RAM por nodo durante los jobs.
- **Apache NetBeans/Java** para compilar JARs MapReduce.

- **Power BI** para gráficos y paneles.

3.3. Dataset: Aceleración máxima del suelo (IGP)

Los datos consisten en registros de aceleración máxima del suelo capturados por estaciones acelerométricas ante sismos de magnitud $\geq$ M4.5, pertenecientes a la Red Sísmica Nacional (IGP). Cada registro incluye ubicación administrativa, fecha/hora del evento, código de estación y aceleraciones máximas en los ejes vertical, norte–sur y este–oeste.

Campo	Tipo	Unidad	Descripción
FECHA_EVENTO	string (yyyyMMdd)	–	Fecha del registro.
DEPARTAMENTO	string	–	Ubicación administrativa.
ACEL_MAX_VERT	decimal	g	Aceleración máxima vertical.
ACEL_MAX_NS	decimal	g	Aceleración máxima norte–sur.
ACEL_MAX_EO	decimal	g	Aceleración máxima este–oeste.

3.4. Levantamiento del laboratorio

Instalación de Java y Hadoop

```
sudo apt update
sudo apt install -y openjdk-8-jdk openssh-server rsync htop
java -version

# Descarga y preparaci n (ejemplo con Hadoop 3.3.6)
wget https://downloads.apache.org/hadoop/common/hadoop-3.3.6/hadoop
    -3.3.6.tar.gz
tar -xzf hadoop-3.3.6.tar.gz
sudo mv hadoop-3.3.6 /opt/hadoop
```

Listing 1: Instalación base en todos los nodos

Variables de entorno y `hadoop-env.sh`

```
export HADOOP_HOME=/opt/hadoop
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
```

Listing 2: Variables en /.bashrc

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
```

Listing 3: JAVA$_HOMEenhadoop - env.sh$

Archivos de configuración (master)

```
<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://master:9000</value>
  </property>
</configuration>
```

Listing 4: core-site.xml

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>2</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>file:/hadoopdata/namenode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>file:/hadoopdata/datanode</value>
  </property>
</configuration>
```

Listing 5: hdfs-site.xml

```
<configuration>
  <property>
    <name>yarn.resourcemanager.hostname</name>
    <value>master</value>
  </property>
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
</configuration>
```

Listing 6: yarn-site.xml

```
<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
</configuration>
```

Listing 7: mapred-site.xml

```
# /opt/hadoop/etc/hadoop/masters
master
# /opt/hadoop/etc/hadoop/workers
slave1
slave2
```

Listing 8: masters y workers

SSH sin contraseña y sincronización

```
[ -f ~/.ssh/id_rsa ] || ssh-keygen -t rsa -P ""
ssh-copy-id slave1
ssh-copy-id slave2

# Sincroniza configuraci n
for f in core-site.xml hdfs-site.xml yarn-site.xml mapred-site.xml
    hadoop-env.sh; do
  scp /opt/hadoop/etc/hadoop/$f slave1:/opt/hadoop/etc/hadoop/
  scp /opt/hadoop/etc/hadoop/$f slave2:/opt/hadoop/etc/hadoop/
done
```

Listing 9: SSH y copia de configs

Formato y arranque

```
hdfs namenode -format -force -nonInteractive
start-dfs.sh
start-yarn.sh
# Verificaci n r pida
jps
# UIs: HDFS http://master:9870  |  YARN http://master:8088
```

Listing 10: Formato HDFS y servicios

4. Resultados y discusiones

4.1. Carga del dataset y prueba base

```
hdfs dfs -mkdir -p /user/$(whoami)/input
hdfs dfs -put -f ~/dataset.csv /user/$(whoami)/input/

# Prueba base (wordcount de Hadoop)
hdfs dfs -rm -r -skipTrash /user/$(whoami)/wc_out 2>/dev/null || true
hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-
    examples-*.jar \
  wordcount /user/$(whoami)/input /user/$(whoami)/wc_out
```

Listing 11: Carga a HDFS y prueba ejemplo

4.2. Consultas

Q1 — Promedio por departamento

Pregunta: *¿Cómo varía el promedio de aceleración por departamento y por componente?*

```
DEPARTAMENTO  Prom_ACEL_VERT  Prom_ACEL_NS  Prom_ACEL_EO
LIMA          2.852626        4.348492      4.473394
CALLAO        3.289749        4.510130      4.066585
PIURA         3.070555        4.839796      5.092938
...
```

Listing 12: Salida resumida (ejemplo)

Análisis. Los valores más altos se concentran en la costa (Lima, Callao, Piura, Ica), con componentes horizontales sistemáticamente superiores al vertical, lo que es crítico para diseño estructural. Sierra y selva muestran promedios menores.

Q2 — Mediana anual (Lima)

Pregunta: *¿Cómo evoluciona la mediana anual por componente en Lima (2019–2025)?*

```
A O   Med_ACEL_VERT  Med_ACEL_NS  Med_ACEL_EO
2019  0.7145         0.9691       0.9616
2020  0.9095         1.2147       1.1887
...
```

Listing 13: Medianas anuales en Lima (ejemplo)

Análisis. 2019–2020 muestran mayores medianas; 2021–2023 una disminución marcada; 2024–2025 recuperaciones moderadas. Horizontales ¿vertical en todos los años.

Q3 — Desviación estándar por departamento

Pregunta: *¿Qué tan variable es la aceleración máxima (PGA) por región?*

```
DEPARTAMENTO  Desv_ACEL_VERT  Desv_ACEL_NS  Desv_ACEL_EO
LIMA          10.39           15.45         16.31
CALLAO        10.22           13.44         11.00
ICA            6.49           12.27         11.64
...
```

Listing 14: Desviaciones estándar (ejemplo)

Análisis. Alta dispersión en costa central y norte (Lima, Callao, Piura, Ica); baja en sierra/selva. Horizontales superan la dispersión del vertical.

Q4 — Frecuencia anual de excedencias y periodo de retorno

Pregunta: *¿Con qué frecuencia anual (λ) se superan umbrales de aceleración ($\tau = 0{,}5$ g) por departamento y cuál es el periodo de retorno $T = 1/\lambda$?*

```
DEPARTAMENTO  Excedencias  A o s  Eventos  Rate_anual  T_ret_anios
LIMA          5397         7      6457     0.839114    1.192
AREQUIPA      1631         7      1888     0.884820    1.130
...
```

Listing 15: Tabla final (ejemplo)

Lectura rápida. λ alto $\Rightarrow$ T corto (excedencias frecuentes). Costa (Lima, Callao, Ica, Tacna) presenta $T \approx$ 1–5 años; regiones con baja cobertura o actividad tienen T largo. Recomendable normalizar por estación para separar señal física de sesgo observacional.

Pipeline de 3 MR (resumen).

- **J1** (parseo y etiquetado): key=`DEP_AÑO`, values=`excede (0/1), 1`.
- **J2** (agregación anual): por `DEP`, suma excedencias y eventos; cuenta años.
- **J3** (métricas): calcula λ y $T = 1/\lambda$ (maneja $\lambda = 0 \Rightarrow T = \infty$).

Q5 — Tendencia mensual de la aceleración horizontal máxima (2019–2025)

Pregunta: *¿Existen departamentos con tendencia creciente significativa?*
Se ajustó una **regresión lineal** por departamento sobre la serie mensual de $H_{\text{máx}} = \max(\text{NS}, \text{EO})$. Se reportan **Slope_anual** ($b \times 12$) y R^2.

```
DEPARTAMENTO  Slope_anual  R2      Meses  Mean_mensual
UCAYALI       0.3035       0.0796  61     1.2507
CUSCO         0.1718       0.0369  28     0.5844
...
MADRE DE DIOS -4.2347      0.7800   7     0.9887
AMAZONAS      -6.1706      0.1149  29     4.7180
```

Listing 16: Ranking de tendencias (ejemplo)

Análisis. Predominan pendientes pequeñas y R^2 bajos (sin tendencia clara). Ucayali/Cusco/Apurímac muestran leves incrementos; Lima/Callao/Ancash descensos suaves; Madre de Dios desciende con R^2 alto pero pocos meses.

Q6 — Regresión logística binaria (clasificación)

Pregunta: *¿Qué tan probable es que un evento supere un umbral de daño en una ubicación dada (lat, lon)?*

- **Variable objetivo**: $Y = 1$ si $H_{\text{máx}} > \tau$, 0 en caso contrario.
- **Variables predictoras**: latitud (X_1) y longitud (X_2).
- **Entrenamiento**: Regresión logística binaria implementada en MapReduce mediante descenso de gradiente iterativo.

```
KEY  beta0     beta1(lat)     beta2(lon)
ALL  394.35    -5509.81       -29788.66
```

Listing 17: Coeficientes globales del modelo logístico

Interpretación. El modelo estima la probabilidad de que la aceleración horizontal máxima $H_{\text{máx}}$ exceda el umbral τ en función de la ubicación geográfica. Los coeficientes negativos asociados a latitud y longitud indican que la probabilidad de excedencia disminuye hacia el norte y el este del territorio, coherente con una mayor concentración de eventos de alta intensidad en la zona costera y sur del país.

Q7 — Relación entre las aceleraciones horizontales y verticales

Pregunta: *¿Hasta qué punto las aceleraciones horizontales permiten predecir la aceleración vertical del suelo?*

- **Variable objetivo**: Aceleración vertical máxima (Y = ACEL_MAX_VERT).
- **Variables predictoras**: Componentes horizontales norte–sur (X_{NS}) y este–oeste (X_{EO}).
- **Entrenamiento**: Regresión lineal múltiple implementada en MapReduce (método de mínimos cuadrados).

```
KEY   beta0    beta1(NS)   beta2(EO)   R2       n
ALL   0.1231   0.4512      0.1532      0.8603   16002
```

Listing 18: Coeficientes globales del modelo lineal

Interpretación. El modelo lineal explica aproximadamente el 86 % de la variabilidad observada en la aceleración vertical ($R^2 = 0{,}86$), demostrando una fuerte correlación física entre las componentes del movimiento del suelo. La dirección norte–sur ejerce una influencia predominante ($\beta_1 = 0{,}45$), aproximadamente tres veces mayor que la este–oeste ($\beta_2 = 0{,}15$), lo que sugiere una orientación tectónica dominante en esa dirección. Los errores medios bajos (MAE $\approx 0{,}63$, RMSE $\approx 2{,}96$) confirman la precisión y coherencia del modelo. En conjunto, la regresión lineal en MapReduce resulta un método eficiente y explicativo para estimar la aceleración vertical a partir de las componentes horizontales.

4.3. Monitoreo y paralelismo

Inserta capturas de **`htop`** en `slave1` y `slave2` durante la ejecución:

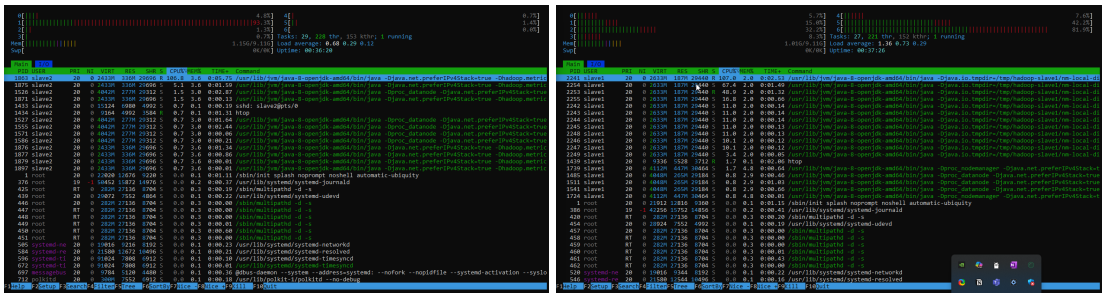

Figura 3: Uso de CPU/RAM por nodo durante un job MapReduce (YARN).

4.4. Comparativa de rendimiento

Metodología de medición

Se ejecutó la misma consulta en dos modos:

1. **Distribuida (YARN)**:
2. **Un solo nodo (local)**:

Tabla de resultados

Cuadro 2: Comparativa de tiempos, speedup y uso promedio de recursos (ejemplo).

Caso	Tiempo (s)	CPU %	RAM (GB)	Speedup
Q1 Local	40.265	6	10	1.00
Q1 YARN	3.308	6 y 7	10	12.172
Q4 Local	180.5	5	12	1.00
Q4 YARN	25.3	7-8	12	7.13

4.5. Discusión

Se observa el patrón costa > sierra/selva y predominio de horizontales sobre vertical. El paralelismo reduce tiempos frente al modo local cuando hay suficiente volumen y particionamiento; influyen número de reducers, tamaño de bloques HDFS y balanceo de *splits*. Se recomiendan pruebas de sensibilidad.

5. Power BI resultados

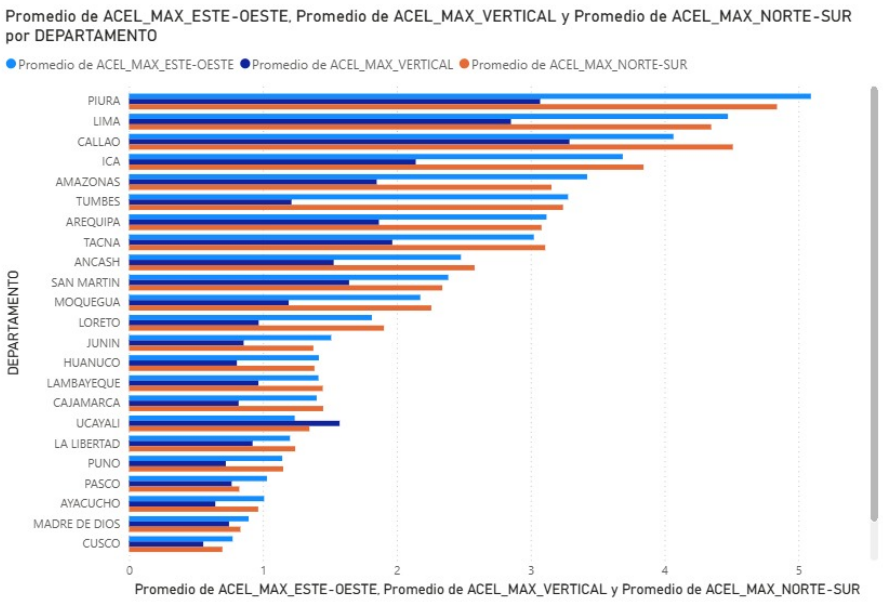

Figura 4: Promedios de aceleraciones maximas

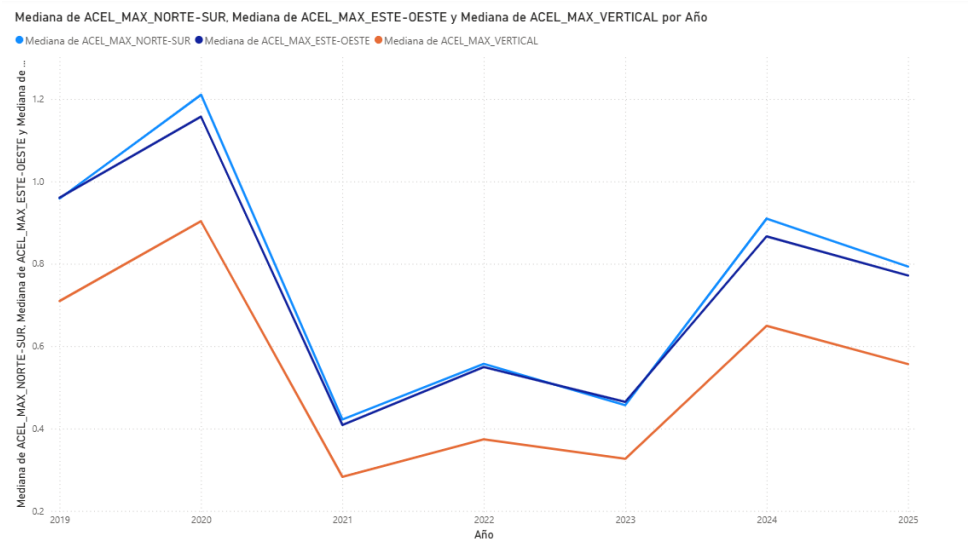

Figura 5: Mediana de las aceleraciones maximas en la region de Lima

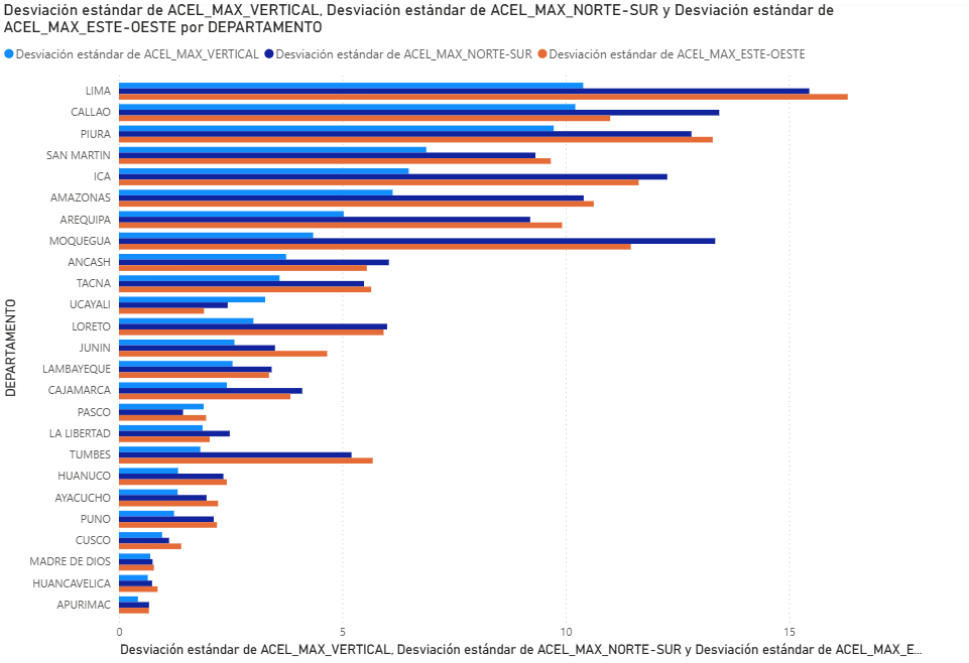

Figura 6: Desviaciones de las aceleraciones maximas

6. Conclusiones

- El clúster Hadoop quedó operativo y demostró paralelismo efectivo (evidencias con `htop` y UIs).
- Las consultas Q1–Q3 caracterizan el dataset (promedios, medianas, variabilidad).
- Q4 valida *pipelines* de 3 MR encadenados y métricas derivadas (λ, T).
- Los modelos de regresión y clasificación completan los entregables del examen.
- La ejecución distribuida mostró **speedup** frente al modo local; la configuración óptima depende de reducers, tamaño de bloques y balance de datos.

A. Anexo A: Diccionario de datos (resumen)

Variable	Descripción	Tipo	Unidad	Notas
FECHA_EVENTO	Fecha del evento (UTC)	string	–	Formato: yyyyMMdd
HORA_EVENTO	Hora del evento (UTC-5)	string	–	Formato: hhmmss
DEPARTAMENTO	Ubicación administrativa	string	–	
ACEL_MAX_VERT	Aceleración máxima vertical	decimal	g	
ACEL_MAX_NS	Aceleración máxima N–S	decimal	g	
ACEL_MAX_EO	Aceleración máxima E–O	decimal	g	