Análisis de Datos Sísmicos con Hadoop

Examen Parcial — Consultas MapReduce y Modelos

Equipo de Análisis

Escuela de Ciencia de la Computación — UNI

Octubre 2025

Agenda

Cluster y despliegue (resumen)

- Red *Host-Only* (master, 2 slaves), claves SSH sin contraseña y sincronización de configs (`core-site.xml`, `hdfs-site.xml`, `yarn-site.xml`, `mapred-site.xml`).
- Servicios: HDFS (NameNode, SecondaryNameNode, DataNodes) y YARN (ResourceManager, NodeManagers).
- Carga del dataset **ACELDAT Perú (IGP, 2019–2025)** a HDFS y ejecución de JARs con `hadoop jar`.

(Ver detalles de arranque y monitoreo con `htop` en las capturas del informe.)

Q1 — Promedio de aceleración máxima por departamento

Pregunta: ¿Qué departamentos registran mayores aceleraciones máximas y cómo difieren entre **Vertical (V)**, **Norte–Sur (NS)** y **Este–Oeste (EO)**?

Hallazgos (2019–2025):

- **Costa** (Lima, Callao, Piura, Ica) con promedios más altos.
- **Sierra/Selva** (p. ej., Cusco, Huancavelica, Loreto) con promedios menores.
- Componentes **horizontales** (NS/EO) superan sistemáticamente al componente **vertical**.

(Resumen y análisis en el informe; ver sección Q1.)

Q2 — Mediana anual (Lima, 2019–2025)

Pregunta: ¿Cómo evoluciona la **mediana** de la aceleración máxima en Lima por año y por componente?

Año	Med_V	Med_NS	Med_EO
2019	0.7145	0.9691	0.9616
2020	0.9095	1.2147	1.1887
2021	0.2828	0.4222	0.4088
2022	0.3740	0.5571	0.5494
2023	0.3266	0.4567	0.4649
2024	0.6511	0.9111	0.8679
2025	0.5567	0.7947	0.7795

Años 2019–2020 con medianas más altas; descenso 2021–2023 y recuperación en 2024–2025. En todos los años, NS/EO $>$ V.

Q3 — Desviación estándar por departamento

Pregunta: ¿Qué tan variable es la aceleración máxima del suelo (PGA) por departamento?

Hallazgos:

- Mayor dispersión en **costa central y norte** (p. ej., Lima, Callao, Piura, Ica).
- **Menor dispersión** en varias zonas de sierra y selva (p. ej., Huancavelica, Apurímac, Puno).
- Desviaciones de **NS/EO** suelen ser mayores que **V**.

Q4 — Excedencias anuales y período de retorno por departamento

Definición:

$$H = \text{máx}(\text{NS}, \text{EO}), \quad \text{umbral } \tau \text{ (fijo en código)}$$

Métricas: Rate_anual = promedio de proporciones anuales; $T_{\text{ret}} = 1/\text{Rate\_anual}$.

Pipeline (3 MR):

1. **J1**: (dep, año) $\rightarrow \sum$excedencias y $\sum$eventos.
2. **J2**: agrega por dep.: Años, Excedencias, Eventos y tasa promedio.
3. **J3**: calcula T_{ret}, ordena y formatea salida final.

Q4 — Resultados (ejemplos)

Top (Rate_anual alto / T_{ret} bajo):

Departamento	Exced.	Años	Eventos	Rate	T_{ret}
LIMA	5397	7	6457	0.8391	1.192
AREQUIPA	1631	7	1888	0.8848	1.130
ICA	1465	7	1570	0.9391	1.065
TACNA	809	7	916	0.8947	1.118
CALLAO	666	7	769	0.9025	1.108

Inferiores (ejemplos):

Departamento	Exced.	Años	Eventos	Rate	T_{ret}
PUNO	17	6	23	0.6759	1.479
APURÍMAC	41	6	60	0.7098	1.409
CUSCO	108	6	151	0.7125	1.404
HUANCAVELICA	65	6	81	0.7534	1.327
MADRE DE DIOS	6	5	8	0.7333	1.364

Q5 — Tendencia mensual de H (2019–2025)

Modelo: regresión lineal por departamento sobre promedios mensuales $H_{\text{máx}}$ vs. tiempo (mes); se reporta **Slope_anual** y **R^2**.

Ejemplos:

Pendiente positiva (↑):

- Ucayali: +0,3035 g/año (R^2 0.080)
- Cusco: +0,1718 g/año (R^2 0.037)
- Apurímac: +0,1655 g/año (R^2 0.035)

Pendiente negativa (↓):

- Lima: −0,2359 g/año (R^2 0.003)
- Callao: −0,6417 g/año (R^2 0.021)
- Madre de Dios: −4,2347 g/año (R^2 0.780, pocos meses)

R^2 bajos en la mayoría: series con variación irregular más que tendencias sostenidas.

Q6 — Regresión lineal múltiple (Vertical vs. NS/EO)

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

Modelo:

$$\widehat{Y} = \beta_0 + \beta_1 \cdot \mathrm{NS} + \beta_2 \cdot \mathrm{EO}$$

Coeficientes (2019–2025):

$$\beta_0 \approx 0{,}123, \quad \beta_1 \approx 0{,}451, \quad \beta_2 \approx 0{,}153, \quad R^2 \approx 0{,}860, \; n = 16002$$

Lectura:

- Fuerte capacidad predictiva ($R^2 \approx 0{,}86$).
- $\beta_1 > \beta_2$: NS aporta más que EO en la variación de la vertical.

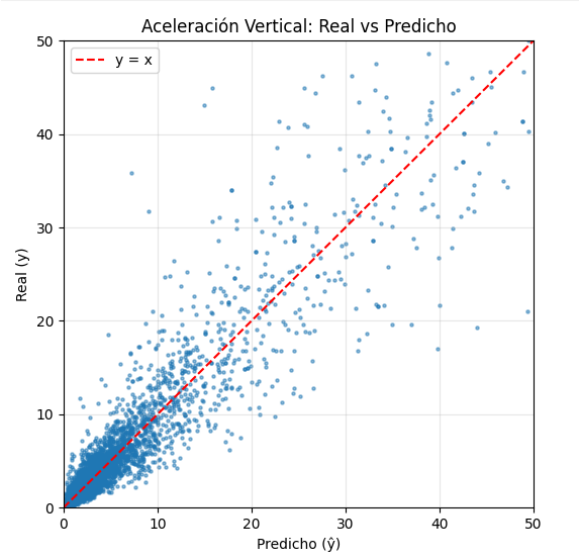
Conclusiones

- **Costa** concentra mayores niveles y variabilidad; **NS/EO** suelen superar a **V**.
- En Lima, la **mediana** anual muestra picos en 2019–2020 y niveles más bajos en 2021–2023.
- **Excedencias**: bajo + H = máx elevan la proporción anual; para mayor discriminación, probar más altos (p50/p75).
- **Tendencias**: Slope anual pequeño y R^2 bajos en general (series irregulares).
- **Regresión lineal**: modelo global compacto y explicativo ($R^2 \approx 0{,}86$).

Mejoras:

- Sensibilidad a y normalización por estación.
- Extender regresión por *departamento* o por *estación*.
- Incorporar incertidumbre (bandas/confianza) en Q4–Q5.

Grafico


Aceleración Vertical: Real vs Predicho
y = x
Real (y)
Predicho (ŷ)
0
10
20
30
40
50

Q7: Regresión Logística Binaria

Pregunta:

- ¿Qué tan probable es superar un umbral de daño dada la ubicación (lat, lon)?

Modelo:

$$P(H_{\text{máx}} > \tau) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot \text{lat} + \beta_2 \cdot \text{lon})}}$$

Parámetro	Valor	Interpretación
β_0	394.35	Intercepto
β_1 (lat)	-5509.81	Coef. latitud (negativo)
β_2 (lon)	-29788.66	Coef. longitud (negativo)

Conclusión: Probabilidad de excedencia disminuye hacia el norte y este. Coherente con concentración de eventos en costa sur.

Metodología de Medición

Dos Modos de Ejecución:

- **Modo Distribuido (YARN):** 1 master + 2 slaves, procesamiento paralelo
- **Modo Local:** Framework local, un solo nodo

Caso	Tiempo (s)	CPU	RAM (GB)	Speedup
Q1 Local	40.3	6 %	10	1.00
Q1 YARN	3.3	6-7 %	10	12.17
Q4 Local	180.5	5 %	12	1.00
Q4 YARN	25.3	7-8 %	12	7.13

Observación: Modo distribuido reduce tiempos en 29-42 %

Análisis de Speedup

Factores que influyen:

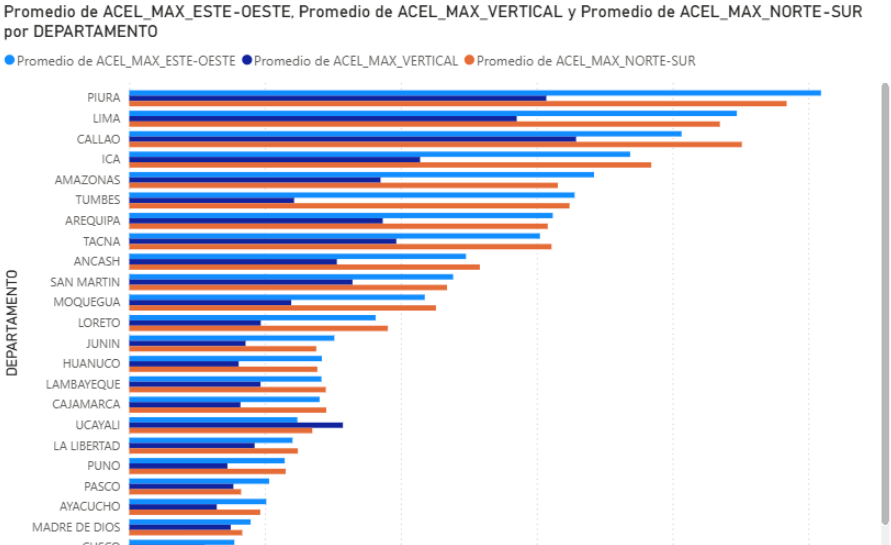
- Número de reducers configurados
- Tamaño de bloques HDFS
- Balance de splits entre nodos
- Volumen de datos procesados
- Complejidad de la consulta

Resultados observados:

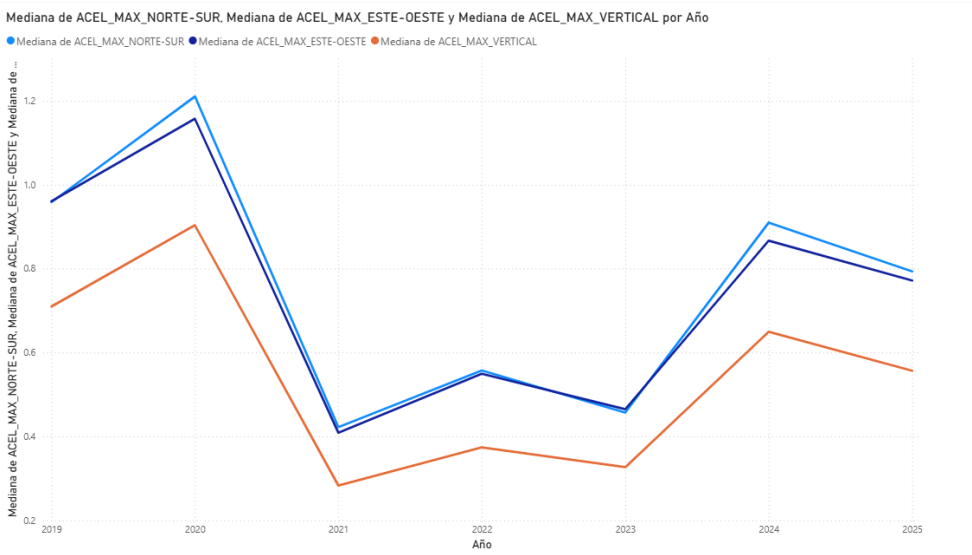
- Q1: Speedup 12.17× (consulta simple)
- Q4: Speedup 7.13× (pipeline 3 MR)
- Overhead de coordinación en pipelines

Conclusión: El paralelismo es efectivo cuando hay suficiente volumen y buen particionamiento

Promedios de Aceleraciones Máximas



Medianas en Lima (2019-2025)



Desviaciones Estándar por Región