Earthquake MST3 – Testing & Evidence Document

Author: Mikołaj Malec (298828)

Functional Tests

Test Case 1: Historical Data Ingestion (USGS & Terraquake)

Objective: Verify that the Apache NiFi ingestion pipelines correctly retrieved the full historical dataset for the year 2025.

Test Steps:

1. Trigger the bulk ingestion process for the start date 2025-01-01 for USGS and Terraquake.
2. Inspect the HDFS Raw directories to confirm file presence for all months/days for both sources and for the clean data.
3. Using Hive, export example data from both sources.

Expected Result: A complete sequence of raw Parquet files covering the entire requested year without gaps and example hive output form the loaded tables.

Actual Result:

HDFS Raw directories output is stored in Test_appendix, they show the dates from 2025-01-01 to 2025-12-22 for both sources and for the clean data.

```
hive> SELECT * FROM usgs_raw LIMIT 1;
OK
tx2025abvp {"mag":1.6,"place":"35 km WSW of Ackerly, Texas","time":1735775714006,"title":"M 1.6 - 35 km WSW of Ackerly, Texas"} {"type":"Point","coord
inates":[-102.059,32.401,8.374]} 2025-01-01
1 row selected (7.373 seconds)
hive> SELECT * FROM terraquake_raw LIMIT 1;
OK
Feature {"eventId":41269422,"originId":133577801,"time":"2025-01-01T22:47:10.750000","author":"SURVEY-INGV","magType":"ML","mag":1.5,"magAuthor":"--",
"type":"earthquake","place":"1 km E Robecco sul Naviglio (MI)","version":100,"geojson_creationTime":"2025-12-29T16:06:47"} {"type":"Point","coordinate
s":[8.9,45.4407,11.6]} 2025-01-01
1 row selected (2.429 seconds)
```

```
vagrant@node1:~$ cqlsh -e "SELECT id, place, region_id, cluster_id FROM earthquakes.events;"

 id           | place                                    | region_id | cluster_id
--------------+------------------------------------------+-----------+------------
 ak2025znsthm |              92 km SE of Kokhanok, Alaska |         1 |          2
     44853052 |                     4 km NE Anghiari (AR) |         9 |         93
   us7000rl6y |       153 km S of Severo-Kuril'sk, Russia |         5 |         50
     44853002 |          2 km NE Cerreto di Spoleto (PG) |         9 |         93
   nc75287501 |       11 km ESE of Hidden Valley Lake, CA |         2 |          6
     44852952 |                     5 km SW Frontone (PU) |         9 |         93
   nc75287496 |                4 km ESE of San Ramon, CA |         2 |        142
   us7000rl6w | 103 km WSW of San Vicente de Cañete, Peru |         3 |         27
     44852872 |                        4 km S Lioni (AV) |         9 |         22
     44852822 |                     4 km NE Anghiari (AR) |         9 |         93
     44852632 |        1 km E Santo Stefano di Magra (SP) |         9 |         93
     44852512 |                        5 km SE Preci (PG) |         9 |         93
     44852152 |                     5 km SW Frontone (PU) |         9 |         93
     44852092 |                     4 km NE Anghiari (AR) |         9 |         93
     44851992 |                     3 km NE Anghiari (AR) |         9 |         93

(15 rows)
vagrant@node1:~$ cqlsh -e "SELECT id, place, region_id, cluster_id FROM earthquakes.events;"

 id           | place                                    | region_id | cluster_id
--------------+------------------------------------------+-----------+------------
     44862232 |                                  Vesuvio |         9 |        100
     44862162 |       1 km NE Tavarnelle Val di Pesa (FI) |         9 |        163
     44862112 |                  6 km W Sansepolcro (AR) |         9 |        163
     44862082 |                  5 km W Sansepolcro (AR) |         9 |        163
     44862032 |                     4 km NE Anghiari (AR) |         9 |        163
     44862002 |                     5 km NE Anghiari (AR) |         9 |        163
     44861902 |                     5 km NE Anghiari (AR) |         9 |        163
     44861872 |                        5 km SW Preci (PG) |         9 |          1
     44861742 |                  5 km W Sansepolcro (AR) |         9 |        163
     44861702 |            8 km E Città di Castello (PG) |         9 |          1
 ak2025znsthm |              92 km SE of Kokhanok, Alaska |         1 |          2
     44853052 |                     4 km NE Anghiari (AR) |         9 |         93
   us7000rl6y |       153 km S of Severo-Kuril'sk, Russia |         5 |         50
     44853002 |          2 km NE Cerreto di Spoleto (PG) |         9 |         93
   nc75287501 |       11 km ESE of Hidden Valley Lake, CA |         2 |          6
     44852952 |                     5 km SW Frontone (PU) |         9 |         93
   nc75287496 |                4 km ESE of San Ramon, CA |         2 |        142
   us7000rl6w | 103 km WSW of San Vicente de Cañete, Peru |         3 |         27
     44852872 |                        4 km S Lioni (AV) |         9 |         22
     44852822 |                     4 km NE Anghiari (AR) |         9 |         93
     44852632 |        1 km E Santo Stefano di Magra (SP) |         9 |         93
     44852512 |                        5 km SE Preci (PG) |         9 |         93
     44852152 |                     5 km SW Frontone (PU) |         9 |         93
     44852092 |                     4 km NE Anghiari (AR) |         9 |         93
     44851992 |                     3 km NE Anghiari (AR) |         9 |         93

(25 rows)
```

Status: Passed

Test Case 2: Speed Layer Stream Enrichment

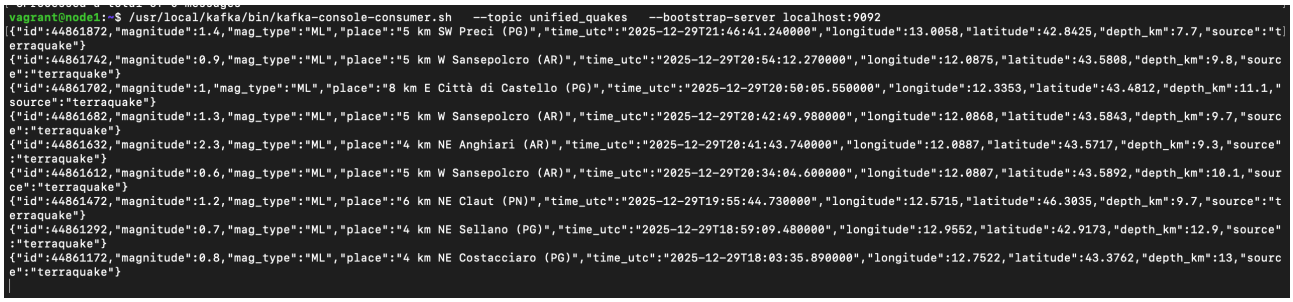
Objective: Confirm that real-time events processed by the Speed Layer are correctly enriched with geospatial metadata (Region ID and Cluster ID) and stored in Apache Cassandra.

Test Steps:

1. Observe the Kafka topic for a new incoming event.
2. Allow the Spark Streaming job to process the event.
3. Query the Apache Cassandra table earthquake_realtime to retrieve the record.

Expected Result: Kafka consumer will intercept the event in the JSON format and the record in Cassandra for the same event must have the region_id and cluster_id columns populated (not null).

Actual Result:



Status: Passed

Test Case 3: Cross-Source Deduplication Validation

Objective: Verify that the Batch Layer deduplication logic successfully merged events from USGS and Terraquake. The Clean Layer should not contain any two events that are within 0.1 degrees of each other and within 10 minutes of each other.

Test Steps:

1. Execute the following QA query in Apache Hive to check for residual duplicates:

```
SELECT
    -- IDs and Time
    a.id AS event_1,
    b.id AS event_2,
    a.time_utc AS time_1,
    b.time_utc AS time_2,

    -- Sources
    a.source_system AS source_1,
    b.source_system AS source_2,

    -- Magnitudes
    a.magnitude AS mag_1,
    b.magnitude AS mag_2,

    -- Location & Diff
    a.place,
    (a.time_utc - b.time_utc) AS time_diff_ms

FROM earthquake_clean a
JOIN earthquake_clean b
ON a.dt = b.dt
WHERE
    -- 1. Prevent self-matches and duplicate pairs
    a.id < b.id

    -- 2. Ensure different sources
    AND a.source_system != b.source_system
```

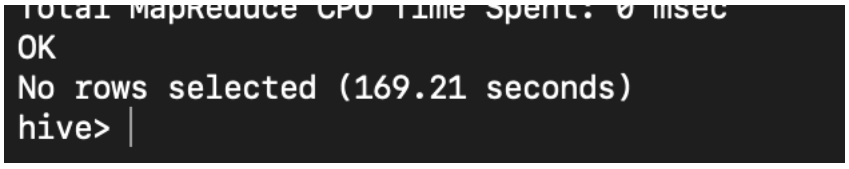
```
    -- 3. Spatial filter (approx 0.1 degrees)
    AND ABS(a.latitude - b.latitude) <= 0.1
    AND ABS(a.longitude - b.longitude) <= 0.1

    -- 4. Time filter: 60,000 ms = 1 Minute
    AND ABS(a.time_utc - b.time_utc) <= 60000

LIMIT 10;
```

Expected Result: The query should return 0 rows, indicating that all overlapping events between the two sources were successfully merged into a single source of truth.

Actual Result:



Status: Passed

Analytical Module & ML Model Evaluation

Test Case 4: K-Means Optimization & Stability Validation

Objective: Verify that the ML pipeline performs robust hyper-parameter tuning by iterating through a cluster range of k = 300 to 600. The test ensures that for each k, 3 random-seed trials are executed to ensure stability, and the final model is automatically selected based on the highest Calinski-Harabasz (CH) Index score.

Test Steps:

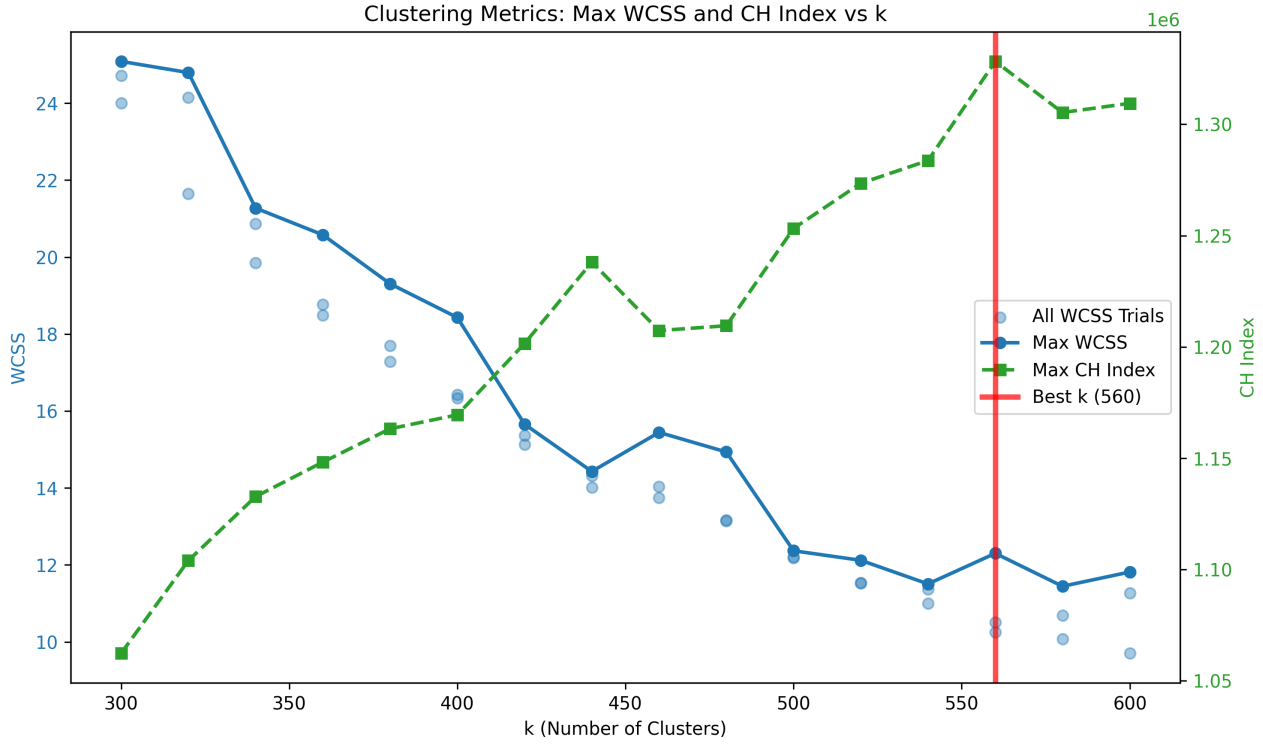
1. Review the PySpark training logs to confirm that multiple trials (iterations with different seeds) were executed for each value of k.

2. Extract the performance metrics table listing k values (50–700) and their associated CH scores.
3. Verify that the system selected the model corresponding to the maximum CH score.
4. Plot the Within-Cluster Sum of Squares (WCSS) against the CH Index and observe the trend.

Expected Result:

- The logs must show multiple distinct WCSS values for every k, confirming that stability trials were run.
- The final selected k must correspond to the peak Calinski-Harabasz score.
- The selected k should geometrically align with the "elbow" (inflection point) of the WCSS curve, indicating where the variance reduction stabilises.

Actual Result:



Status: Passed

Test Case 5: Centroid Back-Projection

Objective: Validate the mathematical accuracy of converting the machine learning model's 3D Cartesian centroids (X, Y, Z) back into Geographic coordinates (Latitude, Longitude) by performing a manual calculation check.

Test Steps:

1. Get a sample centroid vector from the model output.
2. Check if the model's output Lat/Lon matches the manual calculation.

Expected Result: The model output should match the manual calculation.

Actual Result:

```
hive> select * from dim_cluster_metadata limit 1;
OK
0 -0.2660453564903092 -0.6517040619291885 0.7095140480559485 45.22690837241916 -112.20677165688359 767.1438528108063
1 row selected (1.021 seconds)
```

$$r = \sqrt{x^2 + y^2 + z^2}$$

$$x^2 \approx 0.07078$$

$$y^2 \approx 0.42472$$

$$z^2 \approx 0.50341$$

$$r = \sqrt{0.99891} \approx \mathbf{0.99945}$$

$$\phi = \arcsin\left(\frac{z}{r}\right)$$

$$\phi = \arcsin\left(\frac{0.709514}{0.999454}\right) \approx \arcsin(0.70990)$$

$$\phi \approx 0.7893 \text{ radians}$$

Convert to degrees:

$$0.7893 \times \frac{180}{\pi} \approx \mathbf{45.223}^\circ$$

$$\lambda = \operatorname{atan2}(-0.651704, -0.266045) \approx -1.9584 \text{ radians}$$

Convert to degrees:

$$-1.9584 \times \frac{180}{\pi} \approx \mathbf{-112.209}^\circ$$

The results are inside the rounding error +-0.03° which corresponds to ~300 meters
Status: Passed