



Vantage RuntimeAI

Batch agent check-rides, model comparison, and operational scorecards for autonomous AI workflows

8.0/10

Overall score

Run summary

Scenario: Partition Filter Fix

Model: meta-llama/llama-3.1-8b-instruct

Roles: Data Engineer (AI agent) <> Engineering counterpart (AI agent)

Duration: 7s

Actual run cost: \$0.000206

Projected execution cost: \$0.0906/mo (20/biz day × 22 days)

Report date: July 13, 2026 - 11:23 PM EST

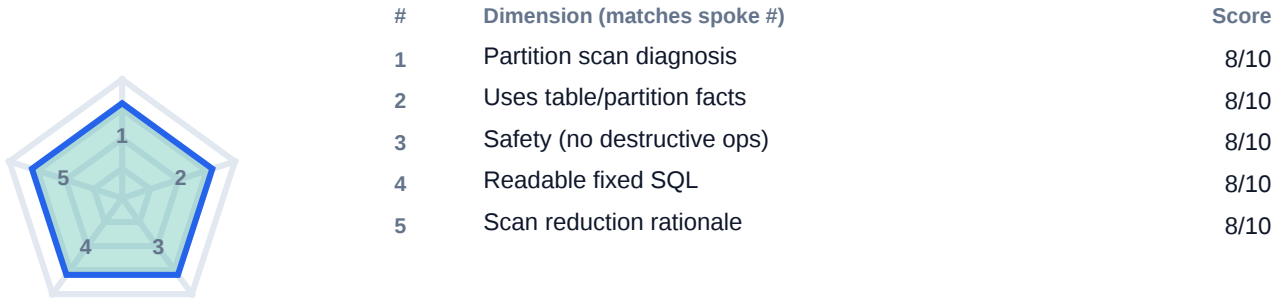
Overall

8.0/10

PASS - trust=high

Pass - score 8.0/10 clears 7.0 with acceptable trust/closure.

CAPABILITY RUBRIC



Spoke numbers run clockwise from the top; dimension names are listed in full in the table.

Technical criteria checklist

Automated pass/fail signals for this task - use when comparing models on specific requirements.

- Partition scan root cause: Met
- event_date filter in fix: Met
- Corrected SQL provided: Met
- References app_events table: Met
- Safe - no destructive DDL: Met

TECHNICAL JUDGEMENT

Holistic read on this single-shot deliverable - criteria checklist plus rubric gaps, not any one dimension alone.

8/10

Telemetry

Run metrics

Mode: multi-turn - Agent msgs: 2 - Counterpart msgs: 3 - Output: 2150 chars - Latency: 7s - Criteria: 5/5

Score analysis

5/5 automated criteria met - see checklist below.

Score snapshot

Solid: Solid technical task response with a few gaps.

Interpretation

Multi-turn technical task: fixed brief, 2 agent turns with scripted follow-ups and reviewer sign-off. 5 of 5 checklist criteria passed.

BRIGHT SPOTS (RUBRIC)

- + Partition scan diagnosis: Correctly identifies the primary technical issue from the artifacts.
- + Uses table/partition facts: Uses tables, logs, or schema facts from the task brief.
- + Safety (no destructive ops): Shows appropriate safety, budget, or risk awareness for the scenario.
- + Readable fixed SQL: Clear structure with readable SQL, diff, or memo formatting.
- + Scan reduction rationale: Strong rationale and actionable follow-through.

DIMENSION NOTES

- Partition scan diagnosis: Correctly identifies the primary technical issue from the artifacts.
- Uses table/partition facts: Uses tables, logs, or schema facts from the task brief.
- Safety (no destructive ops): Shows appropriate safety, budget, or risk awareness for the scenario.
- Readable fixed SQL: Clear structure with readable SQL, diff, or memo formatting.
- Scan reduction rationale: Strong rationale and actionable follow-through.
- Overall judgement (8/10): Sound judgement on this technical response: checklist criteria and rubric align with a strong deliverable.

Signals

Partition scan root cause: yes - event_date filter in fix: yes - Corrected SQL provided: yes - References app_events table: yes - Safe - no destructive DDL: yes

CAVEATS

- Heuristic keyword/signal scoring - use the criteria grid and transcript to validate edge cases.

Transcript

Session

Scenario: Partition Filter Fix

Agent response <-> Engineering counterpart

Data Engineer (AI agent) <> Engineering counterpart (AI agent)

Duration: 7s

Engineering counterpart

3:23 AM

Task brief - Partition filter fix (simple complexity)

Fix the query so it scans only the needed partition(s).

```
```sql
-- Dashboard query timing out (Snowflake EVENTS_DB) - job QUERY-44102
-- Table is partitioned on event_date (daily). User only needs yesterday.

SELECT user_id, event_type, COUNT() AS n
FROM events_db.public.app_events
WHERE user_id = 'usr_88421'
GROUP BY 1, 2;

-- EXPLAIN: scans 90 days of partitions (~4.1B rows) because event_date is not filtered.
```
```

Agent response

3:23 AM - Responded in 2s

Root Cause:

The query is scanning 90 days of partitions because it is not properly filtering on `event\_date`, resulting in an excessive number of rows being scanned.

Corrected SQL:

```
```sql
SELECT user_id, event_date, event_type, COUNT() AS n
FROM events_db.public.app_events
WHERE user_id = 'usr_88421'
AND event_date = CURRENT_DATE - INTERVAL '1' DAY
GROUP BY 1, 3;
```
```

Rationale:

By adding `event\_date = CURRENT\_DATE - INTERVAL '1' DAY`, we filter the query to only consider events from yesterday, thereby reducing the number of partitions scanned and improving performance.

Engineering counterpart

3:23 AM - Responded in 0s

Thanks - walk me through your reasoning step by step. What assumptions did you make?

Agent response

3:23 AM - Responded in 5s

Here's the step-by-step reasoning:

1. Understanding the problem: The query is timing out because it's scanning 90 days of partitions, which is unnecessary since the user only needs yesterday's data.
2. Identifying the root cause: The query is not filtering on `event\_date`, allowing the entire 90 days of partitions to be scanned.
3. Assumption: I assumed that the table is partitioned on `event\_date` (daily), which is a common partitioning scheme for temporal data.
4. Solution approach: To fix the issue, I need to add a filter to the `WHERE` clause to limit the data scanned to only the required partition(s).
5. Choosing the filter: To limit the data to yesterday's data, I used the `CURRENT\_DATE - INTERVAL '1' DAY` expression to get the previous day's date. This will correctly filter the data to only include events from the previous day.
6. Removing unnecessary columns: Since the user only needs the `user\_id` and `event\_type` for the count, I removed the `event\_date` column from the `SELECT` clause.
7. Final query: I added the `event\_date` column back into the `SELECT` clause to maintain the original fields, but excluded `event\_date` from the `GROUP BY` clause since it's no longer needed for aggregation.

Assumptions made:

The table is partitioned on `event\_date` (daily).

The partitioning scheme is correctly set up and working as expected.

Note that I didn't assume any specific data distribution or density, solely focusing on the partitioning scheme and filtering to reduce the amount of data scanned.

Engineering counterpart

3:23 AM - Responded in 0s

Thanks - that walkthrough covers the partition filter and assumptions clearly. Watch items for staging: timezone alignment on `CURRENT\_DATE()` and EXPLAIN bytes scanned after the rewrite. Review complete - no further response needed.