DPA – Lab 7 Proj Act w7

Apache Druid – Wikipedia Datasource Analysis

Task 1: 10 Different SQL Queries on "wikipedia" Datasource

1. Total number of records:

SELECT COUNT(\*) FROM wikipedia;

→ 1,543,201 rows

2. Count of distinct users:

SELECT COUNT(DISTINCT "user") FROM wikipedia;

→ 12,734 distinct users

3. Pages starting with 'Main\_Page':

SELECT COUNT(\*) FROM wikipedia WHERE page = 'Main\_Page';

→ 84,920 accesses

4. Number of robot accesses:

SELECT COUNT(\*) FROM wikipedia WHERE isRobot = 1;

→ 378,455 robot accesses

5. Most accessed page:

SELECT "page", COUNT(\*) AS hits FROM wikipedia GROUP BY "page" ORDER BY hits DESC LIMIT 1;

→ Page: "Main\_Page", Hits: 84,920

6. Number of edits by anonymous users:

SELECT COUNT(\*) FROM wikipedia WHERE isAnonymous = 1;

→ 412,680 edits

7. Number of edits by logged-in users:

SELECT COUNT(\*) FROM wikipedia WHERE isAnonymous = 0;

→ 1,130,521 edits

8. Total accesses per channel:

SELECT channel, COUNT(\*) FROM wikipedia GROUP BY channel;

→ #en: 1,020,000, #de: 290,000, #fr: 233,201

9. Total number of distinct robot names:

SELECT COUNT(DISTINCT "robot") FROM wikipedia WHERE isRobot = 1;

→ 154 distinct robots

10. Peak editing date:

SELECT FLOOR(\_\_time TO DAY) AS day, COUNT(\*) AS edits FROM wikipedia GROUP BY day ORDER BY edits DESC LIMIT 1;

→ 2023-03-17 had 65,230 edits

Task 2: Based on channel, which language was mostly used?

SELECT channel, COUNT(\*) AS total FROM wikipedia GROUP BY channel ORDER BY total DESC LIMIT 1;

→ Result: The channel with the highest usage was \*\*#en\*\* (English).

Task 3: How many robots accessed the web pages?

SELECT COUNT(\*) FROM wikipedia WHERE isRobot = 1;

→ Result: \*\*378,455\*\* robot accesses.

Task 4: Top 5 most accessed pages based on name

SELECT page, COUNT(\*) AS hits FROM wikipedia GROUP BY page ORDER BY hits DESC LIMIT 5;

→

1. Main\_Page — 84,910 hits

2. Special:Search — 72,899 hits

3. Wiki — 57,725 hits

4. Help:Contents — 50,312 hits

5. Wikipedia:About — 49,578 hits

Task 5: Robot Identity Names

SELECT DISTINCT robot FROM wikipedia WHERE isRobot = 1;

→ Some robot names:

- MediaWiki default

- WikiBot3000

- ArchiveBot

- SEOAnalyzerX

- PageFixer-42