Podatkovno inženjerstvo
2. domaća zadaća

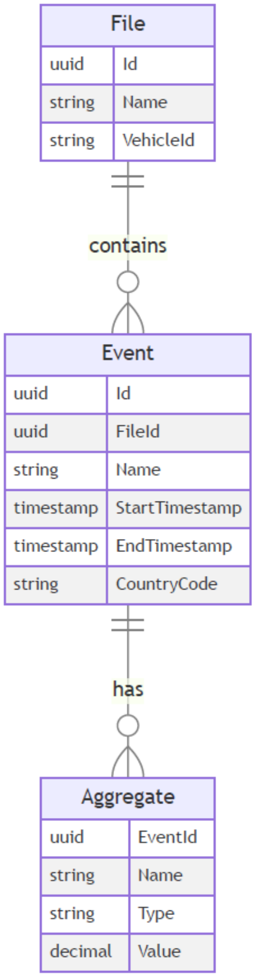
Doris Đivanović

27. travnja 2025.

Dashboard u Supersetu

Vaš je zadatak povezati se s bazom podataka i stvoriti nekoliko grafikona na temelju podataka koji se nalaze u bazi podataka, prikazujući ih na "nadzornoj ploči", tj. na *dashboardu*. Uz grafikone, potrebno je napisati i SQL upite kojima direktno iz baze dobivamo podatke koji su prikazani na grafikonu.

Dijagram baze podataka - Napomena: Tablica se zove Aggregation, a ne Aggregate.



1 Zadatak

Prikažite broj događaja po imenu događaja u obliku kružnog dijagrama (*pie plot*).

SQL upit

```
SELECT "Name", COUNT(*) AS "Count"
FROM "Event"
GROUP BY "Name";
```

SQL upit za Superset i postavke grafikona

```
SELECT "Name" FROM "Event";
```

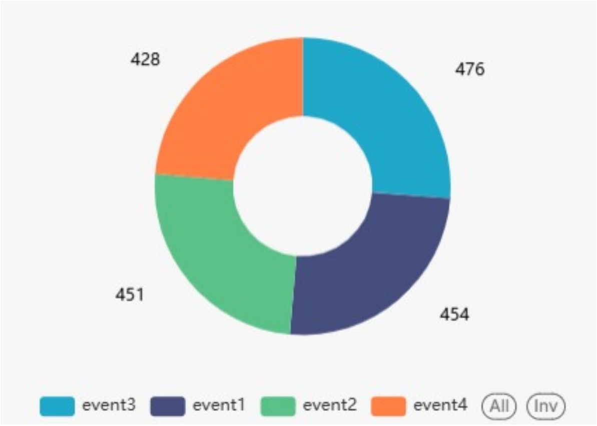
- Visualization type: Part of a Whole - Pie Chart
- Dimensions: Name
- Metric: COUNT(Name)

ILI :

```
SELECT "Id", "Name" FROM "Event";
```

- Visualization type: Part of a Whole - Pie Chart
- Dimensions: Name
- Metric: COUNT(Id)

Superset grafikon



2 Zadatak

Prikažite na tabličan način sve događaje čije je trajanje duže od 30 sekundi. Uz informacije o događajima, prikažite izračunato trajanje događaja.

- Trajanje događaja je razlika u vremenu između završetka i početka događaja.
- Za usporedbu vremenske razlike koristite INTERVAL data type npr. timestamp1 − timestamp2 > INTERVAL '3 hours' .

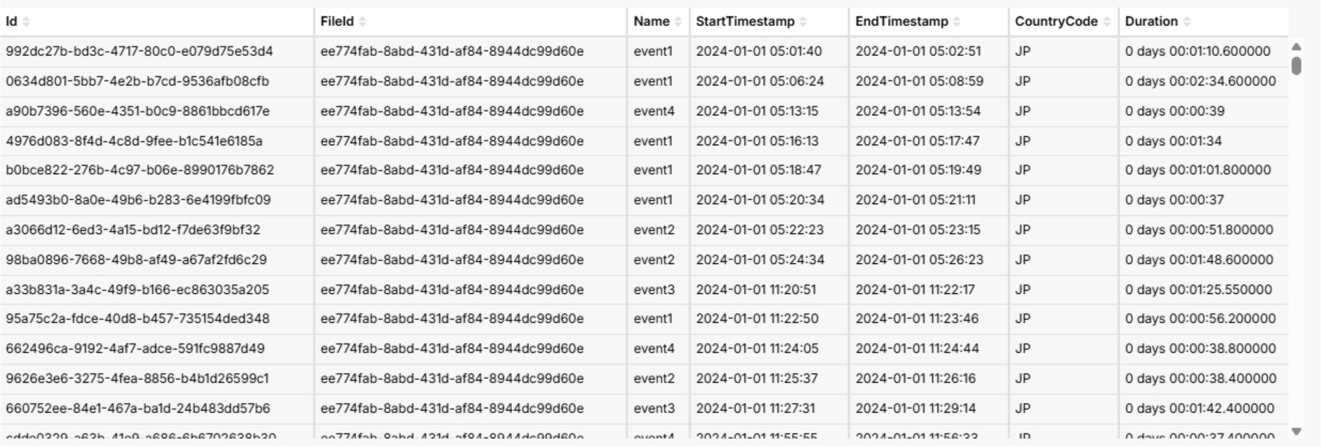
SQL upit

```
SELECT *, "EndTimestamp" - "StartTimestamp" AS "Duration"
FROM "Event"
WHERE "EndTimestamp" - "StartTimestamp" > INTERVAL '30 seconds';
```

Postavke grafikona u Supersetu

- Visualization type: Table - Table
- Query mode: Raw Records
- Columns: ostaviti sve stupce

Superset grafikon



Id	FileId	Name	StartTimestamp	EndTimestamp	CountryCode	Duration
992dc27b-bd3c-4717-80c0-e079d75e53d4	ee774fab-8abd-431d-af84-8944dc99d60e	event1	2024-01-01 05:01:40	2024-01-01 05:02:51	JP	0 days 00:01:10.600000
0634d801-5bb7-4e2b-b7cd-9536afb08cfb	ee774fab-8abd-431d-af84-8944dc99d60e	event1	2024-01-01 05:06:24	2024-01-01 05:08:59	JP	0 days 00:02:34.600000
a90b7396-560e-4351-b0c9-8861bbcd617e	ee774fab-8abd-431d-af84-8944dc99d60e	event4	2024-01-01 05:13:15	2024-01-01 05:13:54	JP	0 days 00:00:39
4976d083-8f4d-4c8d-9fee-b1c541e6185a	ee774fab-8abd-431d-af84-8944dc99d60e	event1	2024-01-01 05:16:13	2024-01-01 05:17:47	JP	0 days 00:01:34
b0bce822-276b-4c97-b06e-8990176b7862	ee774fab-8abd-431d-af84-8944dc99d60e	event1	2024-01-01 05:18:47	2024-01-01 05:19:49	JP	0 days 00:01:01.800000
ad5493b0-8a0e-49b6-b283-6e4199fbfc09	ee774fab-8abd-431d-af84-8944dc99d60e	event1	2024-01-01 05:20:34	2024-01-01 05:21:11	JP	0 days 00:00:37
a3066d12-6ed3-4a15-bd12-f7de63f9bf32	ee774fab-8abd-431d-af84-8944dc99d60e	event2	2024-01-01 05:22:23	2024-01-01 05:23:15	JP	0 days 00:00:51.800000
98ba0896-7668-49b8-af49-a67af2fd6c29	ee774fab-8abd-431d-af84-8944dc99d60e	event2	2024-01-01 05:24:34	2024-01-01 05:26:23	JP	0 days 00:01:48.600000
a33b831a-3a4c-49f9-b166-ec863035a205	ee774fab-8abd-431d-af84-8944dc99d60e	event3	2024-01-01 11:20:51	2024-01-01 11:22:17	JP	0 days 00:01:25.550000
95a75c2a-fdce-40d8-b457-735154ded348	ee774fab-8abd-431d-af84-8944dc99d60e	event1	2024-01-01 11:22:50	2024-01-01 11:23:46	JP	0 days 00:00:56.200000
662496ca-9192-4af7-adce-591fc9887d49	ee774fab-8abd-431d-af84-8944dc99d60e	event4	2024-01-01 11:24:05	2024-01-01 11:24:44	JP	0 days 00:00:38.800000
9626e3e6-3275-4fea-8856-b4b1d26599c1	ee774fab-8abd-431d-af84-8944dc99d60e	event2	2024-01-01 11:25:37	2024-01-01 11:26:16	JP	0 days 00:00:38.400000
660752ee-84e1-467a-ba1d-24b483dd57b6	ee774fab-8abd-431d-af84-8944dc99d60e	event3	2024-01-01 11:27:31	2024-01-01 11:29:14	JP	0 days 00:01:42.400000
cdde0329-a63b-41e9-a686-6b6702638b30	ee774fab-8abd-431d-af84-8944dc99d60e	event4	2024-01-01 11:55:55	2024-01-01 11:56:33	JP	0 days 00:00:37.400000

Napomena. Dobila sam 374 retka u tablici.

3 Zadatak

Prikažite na vizualizaciji karte (*map visualization*) prosječno trajanje događaja koji su se dogodili po zemljama bojenjem zemlje različitim bojama.

SQL upit

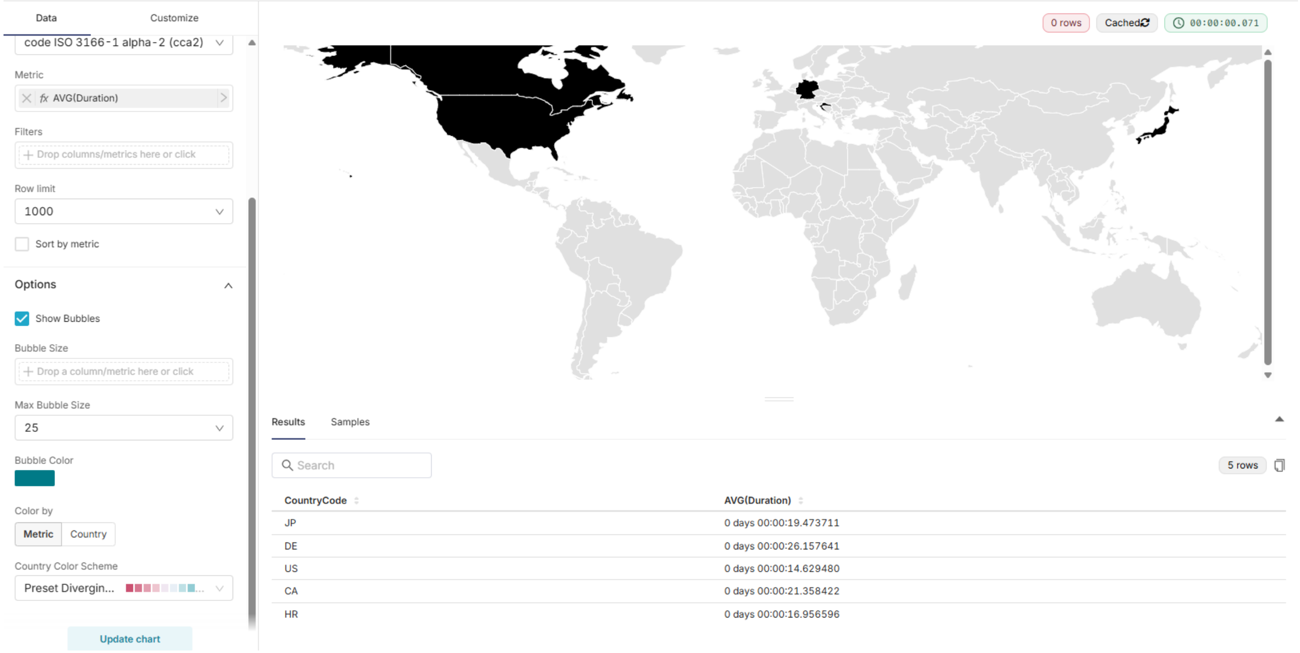
```
SELECT "CountryCode",
AVG("EndTimestamp" - "StartTimestamp") AS "AverageDuration"
FROM "Event"
GROUP BY "CountryCode";
```

SQL upit za Superset i postavke grafikona

```
SELECT "CountryCode", "EndTimestamp" - "StartTimestamp" AS "Duration"
FROM "Event";
```

- Visualization type: Map - World Map
- Country column: CountryCode
- Metric: AVG(Duration)

Superset grafikon



code ISO 3166-1 alpha-2 (cca2)

Metric: AVG(Duration)

Filters: Drop columns/metrics here or click

Row limit: 1000

□ Sort by metric

Options

☑ Show Bubbles

Bubble Size: Drop a column/metric here or click

Max Bubble Size: 25

Bubble Color: 🟦

Color by: Metric | Country

Country Color Scheme: Preset Divergin...

CountryCode	AVG(Duration)
JP	0 days 00:00:19.473711
DE	0 days 00:00:26.157641
US	0 days 00:00:14.629480
CA	0 days 00:00:21.358422
HR	0 days 00:00:16.956596

4 Zadatak

Prikažite broj događaja po identifikatorima vozila u obliku kružnog dijagrama (*pie plot*).

SQL upit

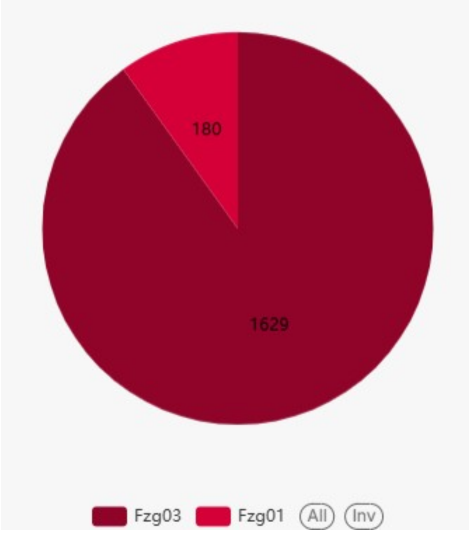
```sql
SELECT "VehicleId", COUNT("Event"."Id") AS "CountOfEvents"
FROM "Event" INNER JOIN "File"
ON "Event"."FileId" = "File"."Id"
GROUP BY "VehicleId";
```

SQL upit za Superset i postavke grafikona

```sql
SELECT "VehicleId", "Event"."Id"
FROM "Event" INNER JOIN "File"
ON "Event"."FileId" = "File"."Id";
```

- Visualization type: Part of a Whole - Pie Chart
- Dimensions: VehicleId
- Metric: COUNT(Id)

Superset grafikon



5 Zadatak

Prikažite broj događaja koji su započeli na određeni datum trakastim dijagramom (*bar chart*), grupirajući događaje prema državi. *Bar chart* prikažite kao *stacked bar chart* gdje jedna traka predstavlja ukupan broj događaja na pojedini datum, a broj događaja po državi je prikazan kao udio u traci te obojan pripadajućom bojom.

SQL upit

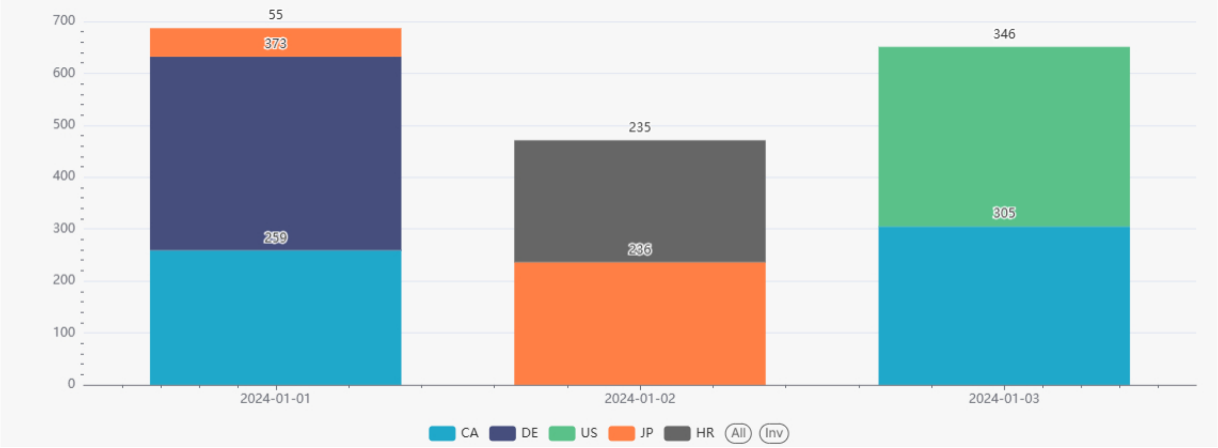
```sql
SELECT DATE("StartTimestamp") AS "StartDate",
"CountryCode",
COUNT("Id") AS "CountOfEvents"
FROM "Event"
GROUP BY "StartDate", "CountryCode";
```

SQL upit za Superset i postavke grafikona

```sql
SELECT "Id", DATE("StartTimestamp"), "CountryCode"
FROM "Event";
```

- Visualization type: Evolution - Bar Chart
- X-axis: date
- Time Grain: Day
- Metric: COUNT(Id)
- Dimensions: CountryCode
- Stacked Style: Stack

Superset grafikon



6 Zadatak

Tablično prikažite državu u kojoj se dogodio događaj s drugom najvećom prosječnom brzinom (Agregat (’SPEED’, ’mean’)) . Rezultat mora prikazati samo traženu državu.

SQL upit

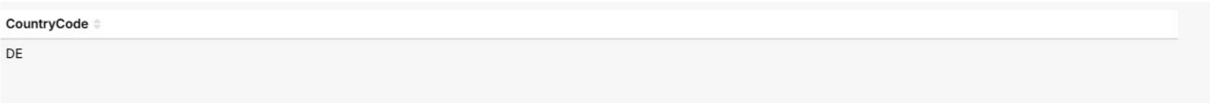
```
SELECT "CountryCode"
FROM
(
  SELECT "CountryCode", "Value"
  FROM "Aggregation" INNER JOIN "Event"
  ON "EventId" = "Id"
  WHERE "Aggregation"."Name" = ’SPEED’
  AND "Type" = ’mean’
  AND "Value" IS NOT NULL
  ORDER BY "Value" DESC
  LIMIT 2
)
AS "MidResult"
ORDER BY "Value" ASC
LIMIT 1;
```

SQL upit za Superset i postavke grafikona

```
SELECT "CountryCode", "Value"
FROM "Aggregation" INNER JOIN "Event"
ON "EventId" = "Id"
WHERE "Aggregation"."Name" = ’SPEED’
AND "Type" = ’mean’ AND "Value" IS NOT NULL
ORDER BY "Value" DESC
LIMIT 2;
```

- Visualization type: Table - Table
- Query mode: Raw Records
- Columns: CountryCode
- Ordering: Value [asc]
- Row limit: 1

Superset grafikon



CountryCode
DE

7 Zadatak

Prikažite grafikon trajanja događaja koristeći intervale duljine 10 sekundi (svi događaji trajanja $0-10$ sekundi u jedan bin, $10-20$ sekundi u drugi bin, itd). Prikažite za svaki bin broj događaja (count) i prosječnu vrijednost maksimalne brzine po eventima (AVG ('SPEED','max')) . Prikažite te vrijednosti na istom chartu gdje su na x osi središta vremenskih intervala, a na y osi broj događaja i prosječna vrijednost maksimalne brzine. Koristite *scatter plot* ili *stepped line plot*.

SQL upit

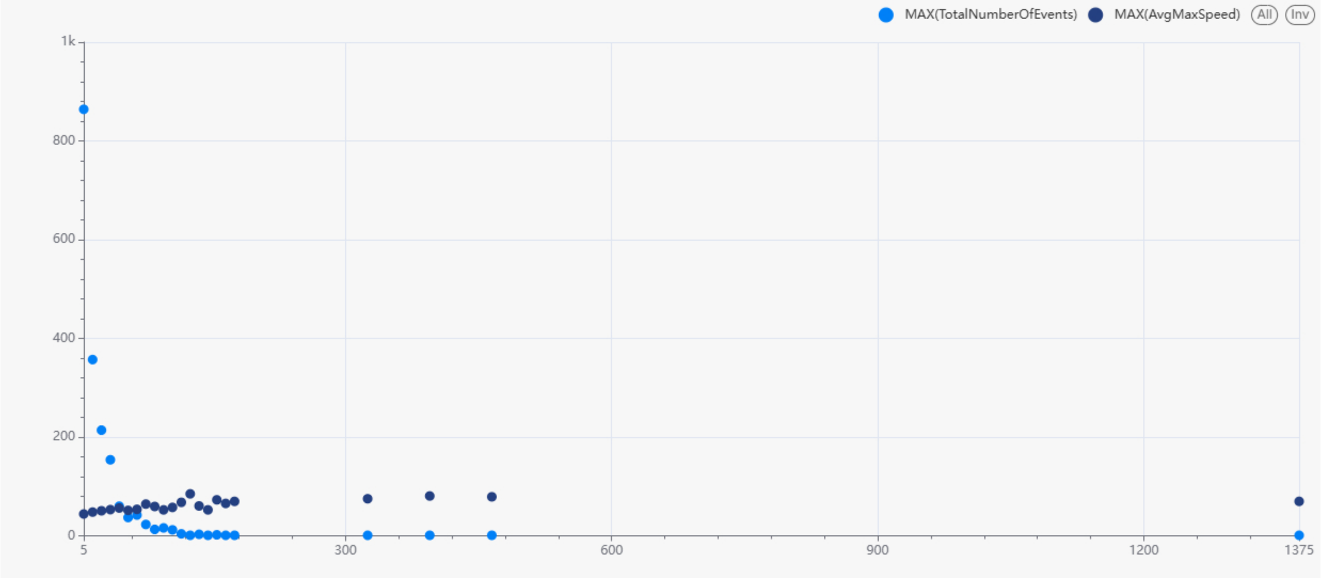
```
SELECT
"NumberOfEventsPer10SecMultipleBins"."TenSecMultipleBin",
"NumberOfEventsPer10SecMultipleBins"."TenSecMultipleBin" - 5
AS "TenSecIntervalMean",
"TotalNumberOfEvents",
"AvgMaxSpeed"
FROM
  (
    SELECT
    CASE
      WHEN EXTRACT(EPOCH FROM "EndTimestamp" - "StartTimestamp") = 0
         THEN 10
      ELSE CEILING(EXTRACT(EPOCH FROM "EndTimestamp" -
         "StartTimestamp")/10)*10
    END
    AS "TenSecMultipleBin",
    Count("Id") AS "TotalNumberOfEvents"
    FROM "Event"
    GROUP BY "TenSecMultipleBin"
  )
  AS "NumberOfEventsPer10SecMultipleBins"
LEFT JOIN
  (
    SELECT
    CASE
      WHEN "DurationInSeconds" = 0 THEN 10
      ELSE CEILING("DurationInSeconds"/10)*10
    END
    AS "TenSecMultipleBin",
    AVG("MaxSpeed") AS "AvgMaxSpeed"
    FROM
      (
        SELECT "Id",
        EXTRACT(EPOCH FROM "EndTimestamp" - "StartTimestamp")
        AS "DurationInSeconds"
        FROM "Event"
      )
      AS "EventRestrictedInfo"
    INNER JOIN
      (
        SELECT "EventId", "Value" AS "MaxSpeed"
        FROM "Aggregation"
        WHERE "Name" = 'SPEED'
        AND "Type" = 'max'
        AND "Value" IS NOT NULL
      )
      AS "AggrRestrictedInfo"
    ON "Id" = "EventId"
    GROUP BY "TenSecMultipleBin"
  )
  AS "AvgMaxSpeedPer10SecMultipleBins"
ON "NumberOfEventsPer10SecMultipleBins"."TenSecMultipleBin" =
   "AvgMaxSpeedPer10SecMultipleBins"."TenSecMultipleBin"
ORDER BY "NumberOfEventsPer10SecMultipleBins"."TenSecMultipleBin"
   ASC;
```

SQL upit za Superset i postavke grafikona

Za ovaj sam zadatak u Sueprset SQL Lab upisala isti ovaj SQL upit, pa su željene metrike već izračunate u dobivenom *dataset*-u, pa sam grafikon postavila ovako (isto i za *stteped line plot*):

- X axis: TenSecIntervalMean
- Metrics: MAX(TotalNumberOfEvents), MAX(AvgMaxSpeed)

Superset grafikon



8 Zadatak

Prikažite na trakastom dijagramu (*bar chart*) koliko je događaja završilo u intervalima od jednog sata koristeći funkciju time_bucket iz TimescaleDB.

SQL upit

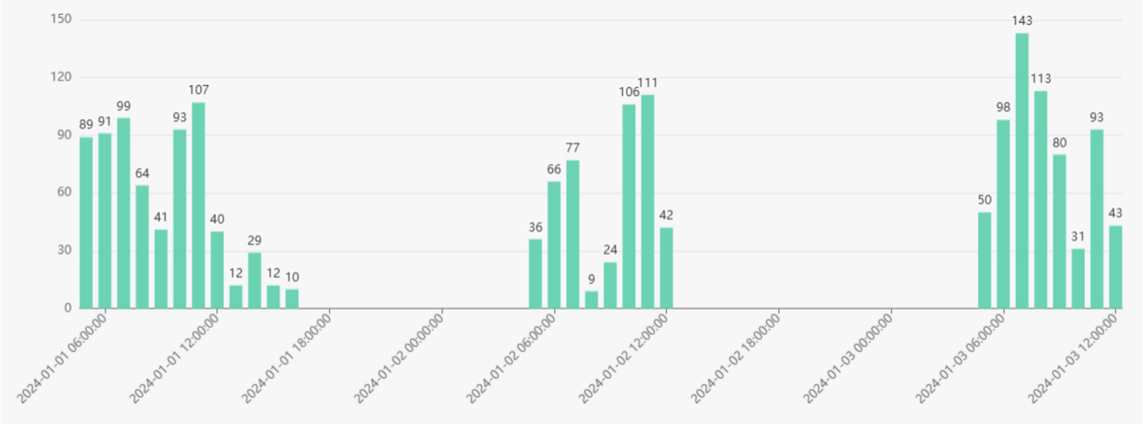
```
SELECT time_bucket('1 hour', "EndTimestamp") AS "EndHour",
COUNT("Id") AS "NumberOfEvents"
FROM "Event"
GROUP BY "EndHour"
ORDER BY "EndHour" ASC;
```

SQL upit za Superset i postavke grafikona

```
SELECT "Id", time_bucket('1 hour', "EndTimestamp")
FROM "Event";
```

- Visualization type: Evolution - Bar Chart
- X-axis: time_bucket
- Time Grain: Hour
- Metric: COUNT(Id)
- Show Value: Yes
- Rotate x axis label: 45°

Superset grafikon



9 Zadatak

Prikažite na tabličan način sve događaje koji imaju agregat (’SPEED’, ’mean’) u rasponu od [45, 55].

SQL upit

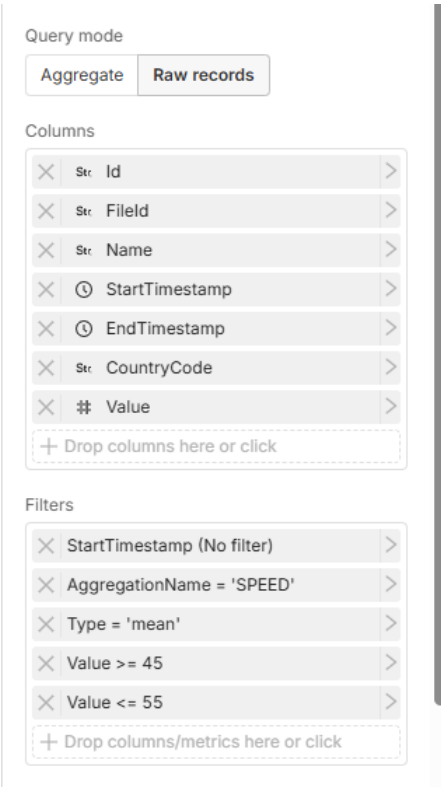
```
SELECT "Event".*, "Value" AS "SpeedMean"
FROM "Event" INNER JOIN "Aggregation"
ON "Id" = "EventId"
WHERE "Aggregation"."Name" = ’SPEED’
AND "Type" = ’mean’
AND "Value" BETWEEN 45 AND 55;
```

SQL upit za Superset i postavke grafikona

Možemo napisati isti upit kao gore i izravno dobiti željenu tablicu, ali u Supersetu je dovoljno napisati ovakav upit:

```
SELECT "e".*,
"a"."Name" AS "AggregationName", "a"."Type", "a"."Value"
FROM "Event" "e" INNER JOIN "Aggregation" "a"
ON "Id" = "EventId";
```

a zatim odabrati željene stupce i filtere ovako:


Aggregate
+ Drop columns here or click
+ Drop columns/metrics here or click

Query mode

Raw records

Columns

Id

FileId

Name

StartTimestamp

EndTimestamp

CountryCode

Value

Filters

StartTimestamp (No filter)

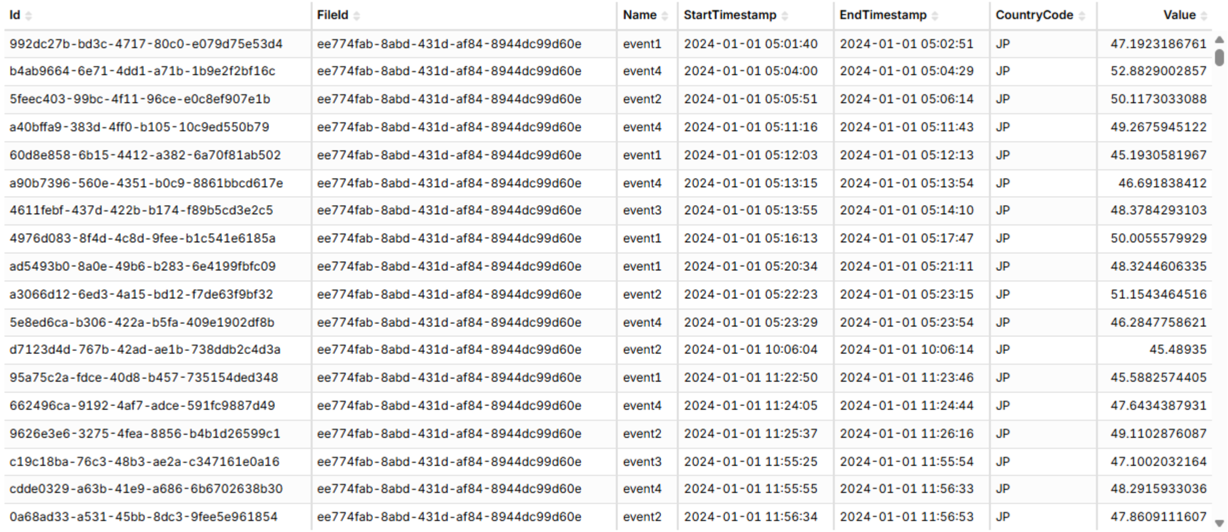
AggregationName = 'SPEED'

Type = 'mean'

Value >= 45

Value <= 55

Superset grafikon



Id	FileId	Name	StartTimestamp	EndTimestamp	CountryCode	Value
992dc27b-bd3c-4717-80c0-e079d75e53d4	ee774fab-8abd-431d-af84-8944dc99d60e	event1	2024-01-01 05:01:40	2024-01-01 05:02:51	JP	47.1923186761
b4ab9664-6e71-4dd1-a71b-1b9e2f2bf16c	ee774fab-8abd-431d-af84-8944dc99d60e	event4	2024-01-01 05:04:00	2024-01-01 05:04:29	JP	52.8829002857
5feec403-99bc-4f11-96ce-e0c8ef907e1b	ee774fab-8abd-431d-af84-8944dc99d60e	event2	2024-01-01 05:05:51	2024-01-01 05:06:14	JP	50.1173033088
a40bffa9-383d-4ff0-b105-10c9ed550b79	ee774fab-8abd-431d-af84-8944dc99d60e	event4	2024-01-01 05:11:16	2024-01-01 05:11:43	JP	49.2675945122
60d8e858-6b15-4412-a382-6a70f81ab502	ee774fab-8abd-431d-af84-8944dc99d60e	event1	2024-01-01 05:12:03	2024-01-01 05:12:13	JP	45.1930581967
a90b7396-560e-4351-b0c9-8861bbcd617e	ee774fab-8abd-431d-af84-8944dc99d60e	event4	2024-01-01 05:13:15	2024-01-01 05:13:54	JP	46.691838412
4611febf-437d-422b-b174-f89b5cd3e2c5	ee774fab-8abd-431d-af84-8944dc99d60e	event3	2024-01-01 05:13:55	2024-01-01 05:14:10	JP	48.3784293103
4976d083-8f4d-4c8d-9fee-b1c541e6185a	ee774fab-8abd-431d-af84-8944dc99d60e	event1	2024-01-01 05:16:13	2024-01-01 05:17:47	JP	50.0055579929
ad5493b0-8a0e-49b6-b283-6e4199fbfc09	ee774fab-8abd-431d-af84-8944dc99d60e	event1	2024-01-01 05:20:34	2024-01-01 05:21:11	JP	48.3244606335
a3066d12-6ed3-4a15-bd12-f7de63f9bf32	ee774fab-8abd-431d-af84-8944dc99d60e	event2	2024-01-01 05:22:23	2024-01-01 05:23:15	JP	51.1543464516
5e8ed6ca-b306-422a-b5fa-409e1902df8b	ee774fab-8abd-431d-af84-8944dc99d60e	event4	2024-01-01 05:23:29	2024-01-01 05:23:54	JP	46.2847758621
d7123d4d-767b-42ad-ae1b-738ddb2c4d3a	ee774fab-8abd-431d-af84-8944dc99d60e	event2	2024-01-01 10:06:04	2024-01-01 10:06:14	JP	45.48935
95a75c2a-fdce-40d8-b457-735154ded348	ee774fab-8abd-431d-af84-8944dc99d60e	event1	2024-01-01 11:22:50	2024-01-01 11:23:46	JP	45.5882574405
662496ca-9192-4af7-adce-591fc9887d49	ee774fab-8abd-431d-af84-8944dc99d60e	event4	2024-01-01 11:24:05	2024-01-01 11:24:44	JP	47.6434387931
9626e3e6-3275-4fea-8856-b4b1d26599c1	ee774fab-8abd-431d-af84-8944dc99d60e	event2	2024-01-01 11:25:37	2024-01-01 11:26:16	JP	49.1102876087
c19c18ba-76c3-48b3-ae2a-c347161e0a16	ee774fab-8abd-431d-af84-8944dc99d60e	event3	2024-01-01 11:55:25	2024-01-01 11:55:54	JP	47.1002032164
cdde0329-a63b-41e9-a686-6b6702638b30	ee774fab-8abd-431d-af84-8944dc99d60e	event4	2024-01-01 11:55:55	2024-01-01 11:56:33	JP	48.2915933036
0a68ad33-a531-45bb-8dc3-9fee5e961854	ee774fab-8abd-431d-af84-8944dc99d60e	event2	2024-01-01 11:56:34	2024-01-01 11:56:53	JP	47.8609111607

Napomena. Dobila sam 573 retka u tablici.

10 Zadatak

Prikažite vrijednosti agregat (’T_OIL’, ’min’) u *scatter plot*-u samo za događaje koji imaju agregat (’SPEED’, ’max’) u rasponu od $[45, 55]$, dodatno grupirajući događaje po vremenskom intervalu od jednoga sata i po imenu događaja.

- Za agregaciju vrijednosti koristite odgovarajuću agregaciju:
  - tip = ’max’ ⟶ agregirajte koristeći ’max(a.Value)’
  - tip = ’min’ ⟶ agregirajte koristeći ’min(a.Value)’
- X os treba biti vremenski interval po satima, a Y os agregirani agregat.
- Koristite ime događaja za dodatno kategoriziranje podataka u više grupa na grafikonu, npr. svako ime događaja trebalo bi biti drugačije obojano.

SQL upit - 1. rješenje

```sql
SELECT
time_bucket(’1 hour’, "StartTimestamp") AS "StartHour",
"Name",
MIN("MinTOil") AS "MINMinTOil"
FROM
  "Event"
INNER JOIN
  (
    SELECT "EventId", "Value" AS "MinTOil"
    FROM "Aggregation"
    WHERE "Name" = ’T_OIL’
    AND "Type" = ’min’
    AND "EventId" IN
      (
        SELECT "EventId"
        FROM "Aggregation"
        WHERE "Name" = ’SPEED’
        AND "Type" = ’max’
        AND "Value" BETWEEN 45 AND 55
      )
  )
  AS "MidResult"
ON "Id" = "EventId"
GROUP BY "StartHour", "Name";
```

SQL upit - 2. rješenje

```sql
SELECT
time_bucket(’1 hour’, "StartTimestamp") AS "StartHour",
"Name",
MIN("MinTOil") AS "MINMinTOil"
FROM
  "Event"
INNER JOIN
  (
    SELECT
    "aggr1"."EventId", "aggr1"."Value" AS "MinTOil"
    FROM
    "Aggregation" "aggr1"
    INNER JOIN
    "Aggregation" "aggr2"
    ON "aggr1"."EventId" = "aggr2"."EventId"
    AND "aggr2"."Name" = ’SPEED’
    AND "aggr2"."Type" = ’max’
    AND "aggr2"."Value" BETWEEN 45 AND 55
    AND "aggr1"."Name" = ’T_OIL’
    AND "aggr1"."Type" = ’min’
  )
  AS "MidResult"
ON "Id" = "EventId"
GROUP BY "StartHour", "Name";
```

SQL upit za Superset i postavke grafikona

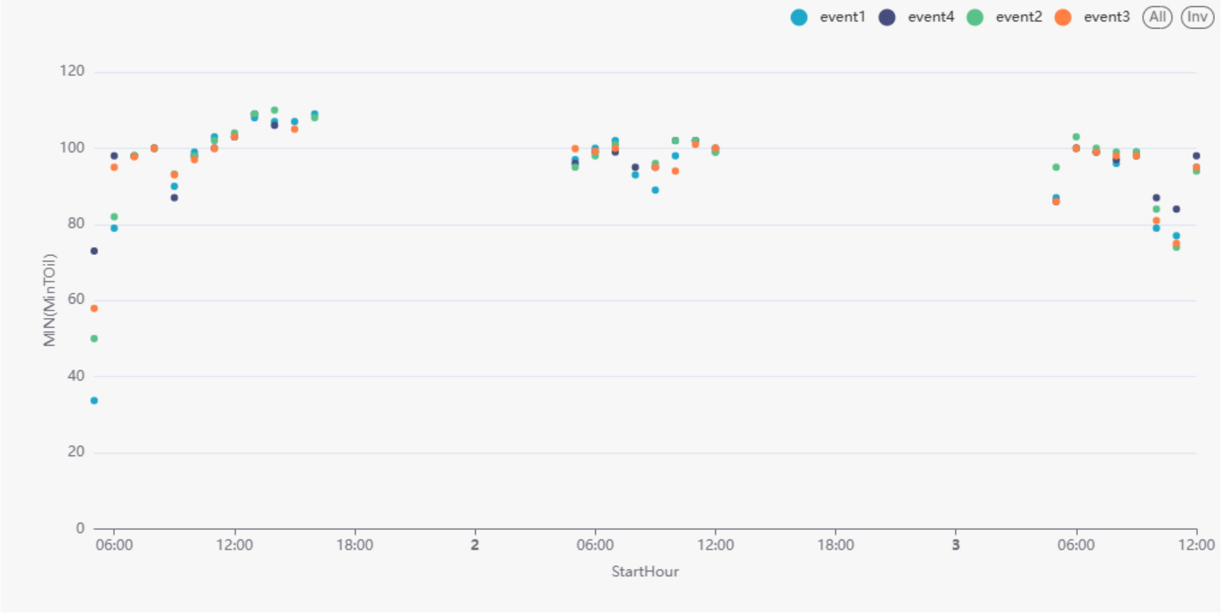
Možemo, na primjer, u Superset SQL Lab upisati bilo koji od gornja dva upita, ali bez agregiranja MinTOil i bez grupiranja:

```
SELECT
time_bucket('1 hour', "StartTimestamp") AS "StartHour",
"Name",
"MinTOil"
--...
-- kod iz jednog od rjesenja
--...
ON "Id" = "EventId";
```

Dobiveni *dataset* **moramo najprije spremiti kao virtualan**, a zatim možemo odabrati ovakve postavke grafikona:

- Visualization type: Evolution - Scatter Plot
- X-axis: StartHour
- Time Grain: Hour
- Metric: MIN(MinTOil)
- Dimensions: Name

Superset grafikon



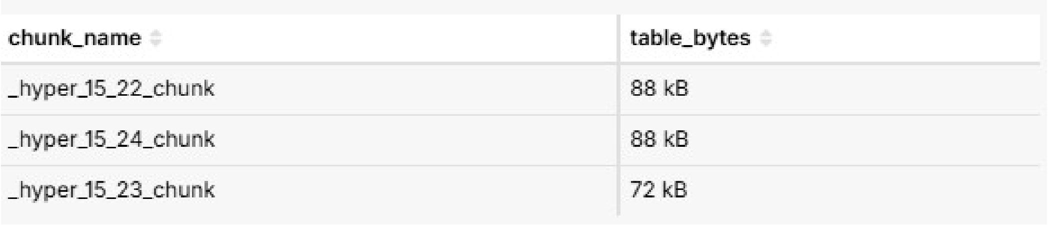
11 Zadatak

Prikažite na tabličan način neke statistike o tome kako su podaci tablice događaja pohranjeni kao hiper-tablica. Za to koristite funkciju **chunks_detailed_size** iz TimescaleDB i vizualizirajte samo stupce **chunk_name** i **table_bytes**, ostali nisu važni ovdje. Budući da stupac **table_bytes** nije stvarno ljudski čitljiv, budući da je u bajtima, koristite funkciju **pg_size_pretty** kako biste ga prikazali na čitljiviji način. Poredajte dijelove od najvećeg do najmanjeg, koristeći **table_bytes** za poredanje.

SQL upit

```
SELECT chunk_name, pg_size_pretty(table_bytes) AS "table_bytes"
FROM chunks_detailed_size(’"Event"’)
ORDER BY "table_bytes" DESC;
```

Superset grafikon



chunk_name	table_bytes
_hyper_15_22_chunk	88 kB
_hyper_15_24_chunk	88 kB
_hyper_15_23_chunk	72 kB

Literatura

- Helena Marciuš, Predavanja iz kolegija Podatkovno inženjerstvo,
  PMF Matematički odsjek, Zagreb, 2025.