

به نام خدا

گزارش دستور کار ۹

آزمایشگاه پایگاه داده - استاد مجتبی بنائی

معین شیردل - ۸۱۰۱۹۷۵۳۵

بهمنماه ۱۴۰۰

## بخش اول: نصب و راه اندازی

ابتدا، دروید را روی پورت ۸۸۸۸ سرور محلی راه اندازی کردم و پس از آن، همین کار را برای کافکا روی پورت ۸۰۰۰ انجام دادم. سپس در کافکا تاپیک events\_topic نیز ایجاد شد. پس از آن، به کمک کد پایتونی موجود در فایلضمیمه، دیتاهای کافکا لود شدند و سپس اقدام به خواندن داده‌های کافکا به کمک دروید کردم.

در تصویر پایین، تsek طراحی شده برای خواندن و وارد کردن داده‌ها از کافکا مشاهده می‌شود.

The screenshot shows the Druid UI interface. In the top navigation bar, there are tabs for druid, Load data, Ingestion, Datasources, Segments, Services, and Query. The main area is titled 'Supervisors'. It has a table with columns: Datasource, Type, Topic/Stream, Status, and Actions. One row is visible: 'events' (kafka) -> 'events\_topic' (Status: RUNNING). Below this is a 'Tasks' section with a table showing an indexing task: 'Task ID: index\_kafka\_events', 'Group ID: "index\_kafka\_events"', 'Type: index\_kafka', 'Datasource: events', 'Location: localhost:8100', 'Created time: 2022-01-28T21:16:55.618Z', and 'Status: RUNNING'.

در این تصویر نیز، دیتا سورس حاصل از این عملیات که حاوی فایل‌های واردہ از کافکاست را می‌بینیم.

The screenshot shows the Druid UI interface. In the top navigation bar, there are tabs for druid, Load data, Ingestion, Datasources, Segments, Services, and Query. The main area is titled 'Datasources'. It has a table with columns: Datasource name, Availability, Availability detail, Total data size, Segment size (rows) minimum / average / maximum, Segment granularity, Total rows, Avg. row size (bytes), Replicated size, Compaction, and % Compressed bytes / segment. One row is visible: 'events' (Availability: Fully available (3 segments), Total data size: 0.00 B, Total rows: 7,285, Avg. row size: 0.00 B, Compaction: Not enabled).

همچنین تصویر تاپیک ایجاد شده در کافکا نیز مشخص است.

The screenshot shows the Kafka UI interface. In the top navigation bar, there are tabs for Topics, Consumers, and Logs. The main area is titled 'Topics'. It has a table with columns: Topics, Partitions, Replications, and Consumer Groups. One topic is listed: 'events\_topic' (Count: 7666, Size: 1.751 MB, Last Record: 5 seconds ago, Total Partitions: 2, Factor: 1, In Sync: 1, Consumer Groups: None). At the bottom right, there is a 'Create a topic' button.

## بخش دوم: Rollup ها

اولین Rollup یک روزانه است که به ازای هر روز و هر event\_id (به علت عدم امکان حذف ستون event\_id) کمترین میزان event\_value در ستون min\_event\_value ذخیره می شود. حاصل این Rollup در تصویر زیر قابل مشاهده است و تعداد داده های تجمعی شده در هر گروه نیز مشخص است.

_time long (time column)	event_type string	count count	min_event_value doubleSum
2021-11-22T00:00:00.000Z	idle_5	1	2841.703
2021-10-13T00:00:00.000Z	hover	1	9033.929
2021-11-13T00:00:00.000Z	hover	1	2889.34
2021-10-25T00:00:00.000Z	hover	2	1531.259
2021-10-09T00:00:00.000Z	click	1	6022.874
2021-12-09T00:00:00.000Z	click	1	1637.433
2021-11-24T00:00:00.000Z	click	2	9316.655999999999
2021-10-19T00:00:00.000Z	click	1	3350.811
2021-11-06T00:00:00.000Z	click	1	4607.362
2021-11-23T00:00:00.000Z	idle_5	1	7533.202
2021-10-22T00:00:00.000Z	click	1	818.724
2021-10-21T00:00:00.000Z	hover	1	8912.735
2021-11-15T00:00:00.000Z	click	1	2255.037
2021-11-11T00:00:00.000Z	hover	1	6070.766

نتیجه ی این Rollup یک دیتا سورس جدید خواهد بود که تسك محاسبه و به دست آوردن و ایجاد این دیتاسورس جدید در تصویر زیر مشخص است:

Supervisors					
Datasource	Type	Topic/Stream	Status	Actions	
events	kafka	events_topic	RUNNING		
events_rollup_day	kafka	events_topic	RUNNING		

Tasks								
Task ID	Group ID	Type	Datasource	Location	Created time	Status		
index_kafka_events_rollup_day_3c69bcd2b42df2_pgbppmjo	index_kafka_events_rollup_day	index_kafka	events_rollup...	localhost:8101	2022-01-28T21:49:17.847Z	RUNNING		
index_kafka_events_6fe026814ccc35d_njhepkmg	index_kafka_events	index_kafka	events	localhost:8100	2022-01-28T21:16:55.618Z	RUNNING		

همچنین اضافه شدن این دیتاسورس به لیست دیتاسورس های کلستر دروید ما نیز در تصویر زیر مشخص است:

Datasources										
DataSource name	Availability	Availability detail	Total data size	Segment size (rows) minimum / average / maximum	Segment granularity	Total rows	Avg. row size (bytes)	Replicated size	Compaction	% Compacted bytes / segme
events_one_day	Fully available (150 segments)	No segments...	1.09 MB	5 - 47 - 107	Day	7,136	153	1.09 MB	Not enabled	-
events_rollup_day	Fully available (75 segments)	No segments...	0.00 B	0 - 0 - 0	Day	300	0	0.00 B	Not enabled	-

دومین Rollup یک ماهانه است که به ازای هر روز و هر event\_id (به علت عدم امکان حذف ستون event\_id) کمترین میزان event\_value در ستون min\_event\_value ذخیره می شود. باقی موارد مشابه قسمت قبل است و تصاویر این دیتابورس جدید در ادامه قابل مشاهده است:

_time long (time column)	event_type string	count count	sum_event_value doubleSum
2021-11-01T00:00:00.000Z	idle_5	4	24299.59100000004
2021-10-01T00:00:00.000Z	hover	4	33259.263
2021-11-01T00:00:00.000Z	hover	2	8960.106
2021-10-01T00:00:00.000Z	click	4	17306.123
2021-12-01T00:00:00.000Z	click	1	1637.433
2021-11-01T00:00:00.000Z	click	5	20756.34900000002

Supervisors	Type	Topic/Stream	Status	Actions
events	kafka	events_topic	RUNNING	
events_rollup_day	kafka	events_topic	RUNNING	
events_rollup_month	kafka	events_topic	RUNNING	

Tasks	Group by	None	Group ID	Type	DataSource	Status	Created time	Status
index_kafka_events_rollup_month_9308f4af4aff50f_igoiaomm			index_kafka_events_rollup_month	index_kafka	events_rollu...	localhost:8100	2022-01-28T21:57:59.754Z	RUNNING
index_kafka_events_rollup_day_3c69bcd2b42df2_pgbbpmjo			index_kafka_events_rollup_day	index_kafka	events_rollu...	localhost:8101	2022-01-28T21:49:17.847Z	RUNNING

Datasource name	Availability	Availability detail	Total data size	Segment size (rows) minimum / average / maximum	Segment granularity	Total rows	Avg. row size (bytes)	Replicated size	Compaction	% Compacted bytes / segme
events_one_day	Fully available (150 segments)	No segment...	1.09 MB	5 47 107	Day	7,136	153	1.09 MB	Not enabled	-
events_rollup_day	Fully available (75 segments)	No segment...	0.00 B	0 0 0	Day	300	0	0.00 B	Not enabled	-
events_rollup_month	Fully available (3 segments)	No segment...	0.00 B	0 0 0	Month	12	0	0.00 B	Not enabled	-

## بخش سوم: کوئری ها

- ۱- در این قسمت، تعداد event های کاربرانی که نام آن ها با b شروع می شود را می بینیم. به کمک فیلتر Like کاربران با حرف b در ابتدای نام کاربریشان را انتخاب کرده ایم و سپس با گروه بندی روی نام کاربری، تعداد event های هر کاربر را شمرده ایم:

The screenshot shows the Druid interface with a query editor and a results table. The query is:

```
SELECT
    user_name,
    COUNT(*) AS "Count"
FROM events_one_day
WHERE user_name LIKE 'b%'
GROUP BY 1
ORDER BY 2 DESC
```

The results table shows the count of events for users whose names start with 'b':

user_name	Count
baldwinwillie	12
bailykimberly	10
brandonturner	9
barbaraatkinson	8
bquinn	8
barbaralane	7
bauerkenneth	7
bfranco	7
bridgetparks	7

- ۲- در این قسمت، تعداد event های مشاهده شده در بازه زمانی دو ماه گذشته، به تفکیک صفحه ای که آن در آن رخ داده است محاسبه شده است. (به کمک ساختن یک جدول کمکی متشکل از تمام event های دو ماه اخیر)

The screenshot shows the Druid interface with a query editor and a results table. The query is:

```
SELECT
    page,
    COUNT(*) AS "Count"
FROM (SELECT * FROM events_one_day WHERE __time >= CURRENT_TIMESTAMP - INTERVAL '2' MONTH )
GROUP BY 1
ORDER BY 2 DESC
```

The results table shows the count of events for different pages over the last two months:

page	Count
tag	49
blog	45
tags	44
list	43
main	42
search	42
categories	40
app	39
explore	36

۳- در این قسمت، تعداد event ها را به تفکیک نوع آن event شمرده ایم و به ترتیب تعداد مرتب کرده ایم:

The screenshot shows the Druid interface with a sidebar containing a tree view of data sources and their metrics. The main area displays a query result table.

Query:

```
1 SELECT
2   event_type,
3   COUNT(*) AS "Count"
4   FROM events_one_day
5   GROUP BY 1
6   ORDER BY 2 DESC
```

Table:

event_type	Count
hover	1842
buy	1817
idle_5	1742
click	1735

۴- در این قسمت، تعداد event ها را به تفکیک نوع محصول شمرده ایم. در داده ممکن است product\_id null داشته باشد که با Filter Not Null این داده ها را حذف می کنیم و سایر product\_id ها را می شماریم و در نهایت حاصل را بر اساس تعداد event مرتب می کنیم.

The screenshot shows the Druid interface with a sidebar containing a tree view of data sources and their metrics. The main area displays a query result table.

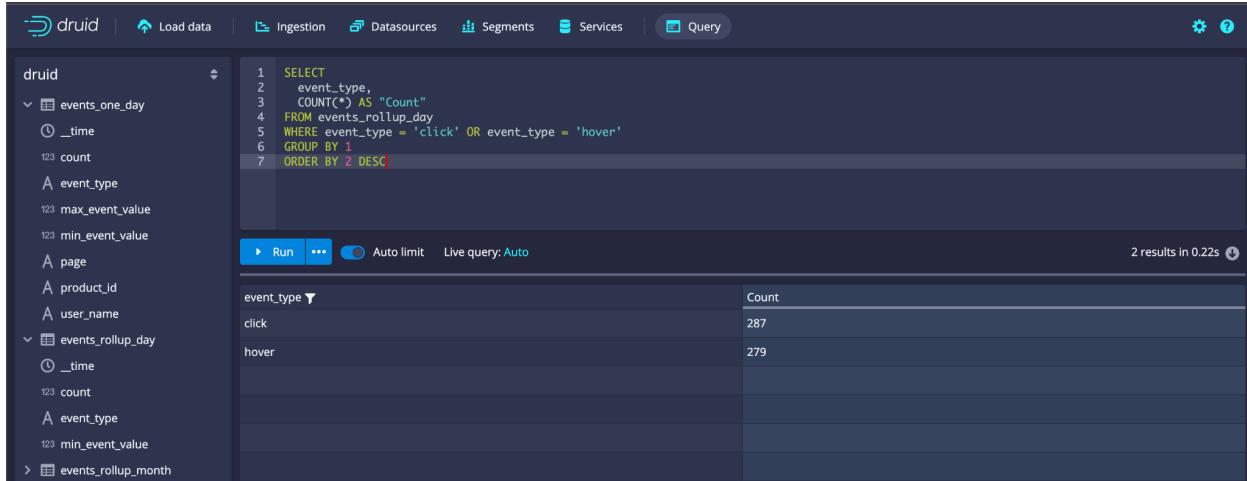
Query:

```
1 SELECT
2   product_id,
3   COUNT(*) AS "Count"
4   FROM events_one_day
5   WHERE product_id IS NOT NULL
6   GROUP BY 1
7   ORDER BY 2 DESC
```

Table:

product_id	Count
product05	639
product02	634
product10	630
product04	623
product07	620
product01	611
product03	610
product09	605
product06	599

۵- در این قسمت، برخلاف قسمت قبل که روی داده‌ی تمام event‌ها کار کردیم، روی دیتا سورس حاصل از rollup یک روزه کار کردیم. به این صورت که تعداد تمامی event‌های از نوع click و hover (که event‌های مربوط به mouse هستند) را شمرده ایم و مقایسه کرده ایم و حاصل به شکل زیر در آمده است.



The screenshot shows the Druid UI interface. On the left, there's a sidebar with a tree view of data sources and their metrics. The main area is a query editor with a SQL-like query and its execution results.

```

SELECT
  event_type,
  COUNT(*) AS "Count"
FROM events_rollup.day
WHERE event_type = 'click' OR event_type = 'hover'
GROUP BY 1
ORDER BY 2 DESC
  
```

Below the query, there are buttons for 'Run', '...', 'Auto limit', and 'Live query: Auto'. To the right, it says '2 results in 0.22s'.

event_type	Count
click	287
hover	279