metatron-doc-user Documentation

Release 0.4.3

metatron team

May 21, 2019

METATRON DISCOVERY

CHAPTER

ONE

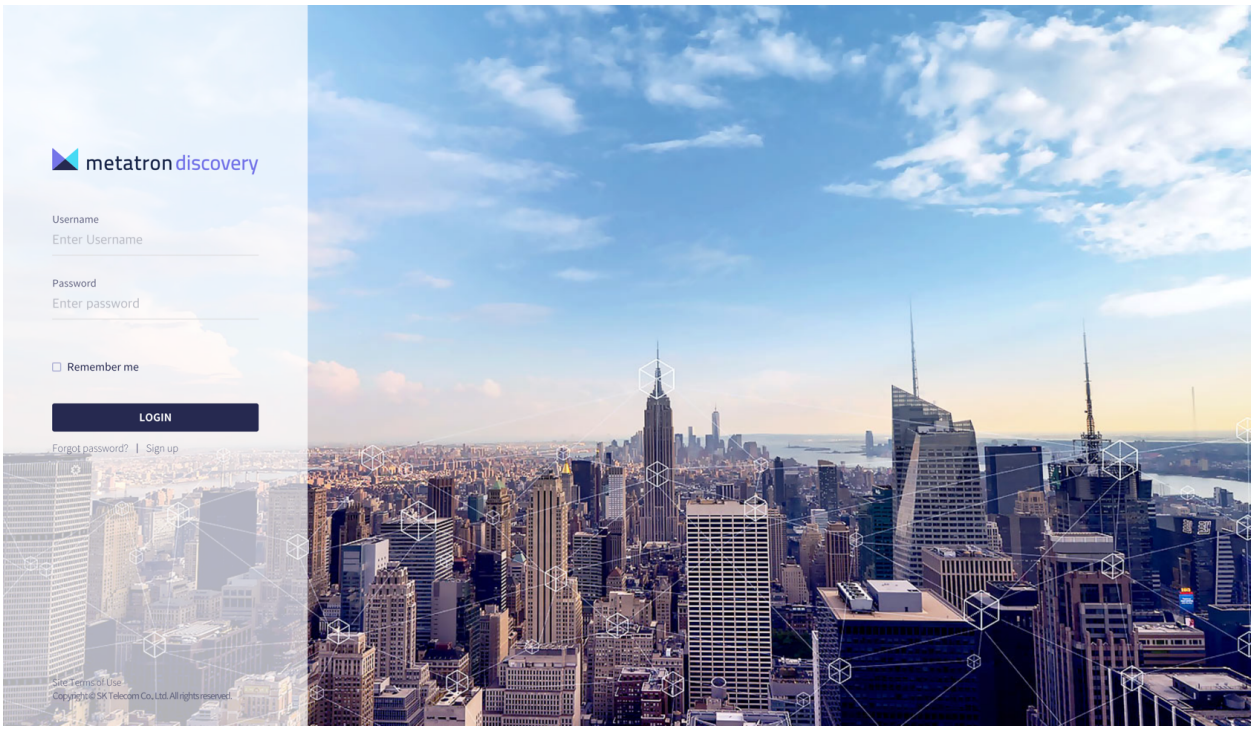
DISCOVERY QUICK GUIDE

Metatron Discovery is an all-in-one solution that enables rapid loading, pre-processing, and analysis of large amounts all together. With Metatron Discovery, business users without technical knowledge can directly work with data and gain insights from rapid visualization.

You can perform data analysis with Metatron Discovery using the two methods below:

- **Method 1:** Run Metatron Discovery demo site. Enter "metatron" as your ID and password.
- **Method 2:** Download the single-mode Metatron Discovery to your local PC. Download is provided in three ways.
  - Custom install: Download the source code from the Github repository, or directly run the build file.
  - Virtual machine: Run the virtual machine image. This is also available in the Windows OS.
  - Docker: Run the Docker image for a quick installation.

Do you see the screen below? Congratulations! You are now ready for quick and easy data analysis with Metatron Discovery.



For a quick start, follow the three-step tutorial below:

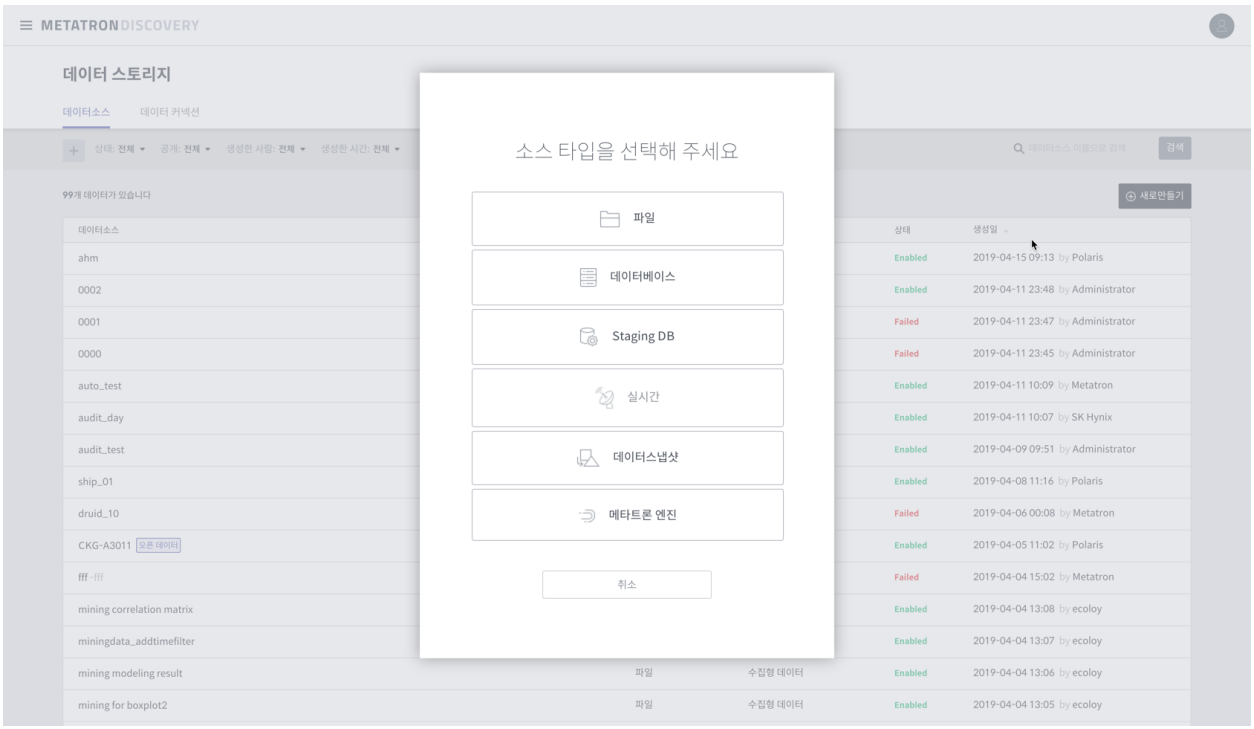
1.1 Step 1. Create a data source

The first step in data analysis is ingesting your data into the system. Metatron Discovery allows you to easily ingest various data sources.

The example in this tutorial shows you how to ingest data from your local directory. First, prepare data. An Excel file (.xls, .xlsx) or .csv file will suffice. This tutorial uses sales data. Download it from the link below:

```
sample data (.csv)
```

Data sources can be viewed and ingested from **Management > Data Storage > Data Source**. To create a new data source, click the **New** button on the upper right of the data source list.

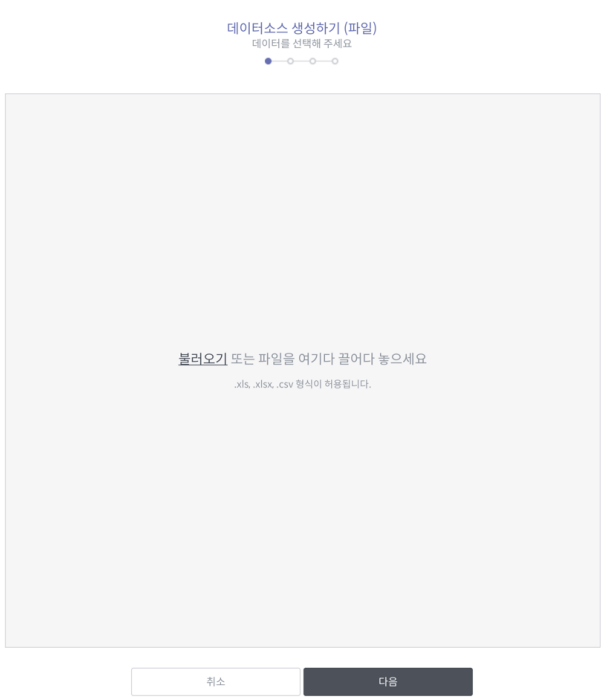


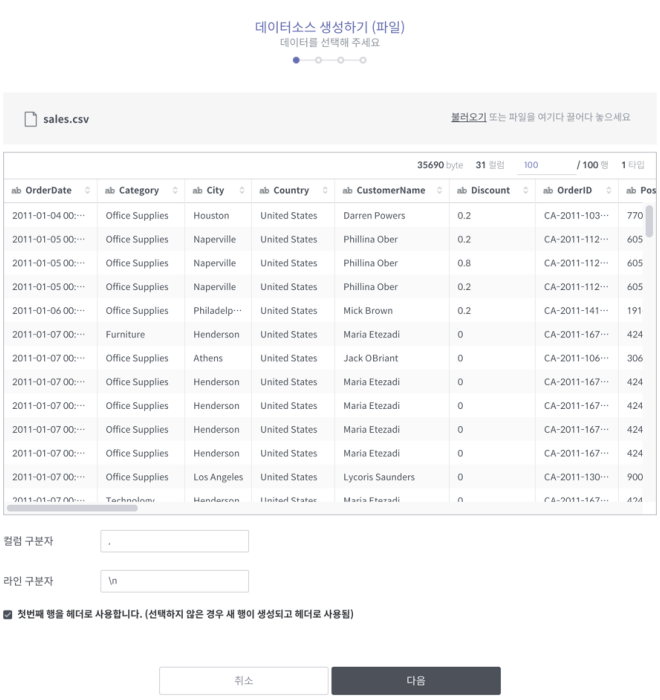
In this tutorial, click **File** to retrieve the data from your local directory. See *Create a data source* for details on creating a data source from other sources.

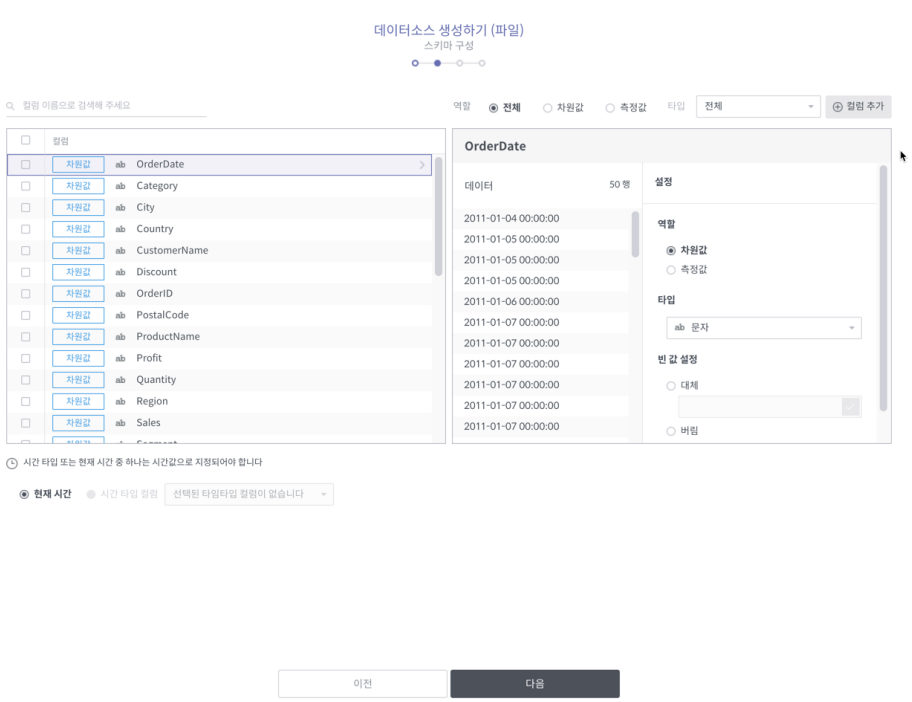
Drag and drop the data you wish to analyze, or retrieve it from the directory.

Drag your cursor over the sales data to view up to 100 rows of data with detection of the column delimiter and line separator. This data is properly displayed using the default delimiter and separator. Click **Next**.

While viewing the data, adjust the column types properly. This task is called **data schema configuration**.


데이터소스 생성하기 (파일)
데이터를 선택해 주세요
불러오기 또는 파일을 여기다 끌어다 놓으세요
.xls, .xlsx, .csv 형식이 허용됩니다.
취소
다음


데이터소스 생성하기 (파일)
데이터를 선택해 주세요
sales.csv
불러오기 또는 파일을 여기다 끌어다 놓으세요
35690 byte 31 컬럼 100 / 100 행 1 타입
OrderDate Category City Country CustomerName Discount OrderID
2011-01-04 00:··· Office Supplies Houston United States Darren Powers 0.2 CA-2011-103··· 770
2011-01-05 00:··· Office Supplies Naperville United States Phillina Ober 0.2 CA-2011-112··· 605
2011-01-05 00:··· Office Supplies Naperville United States Phillina Ober 0.8 CA-2011-112··· 605
2011-01-05 00:··· Office Supplies Naperville United States Phillina Ober 0.2 CA-2011-112··· 605
2011-01-06 00:··· Office Supplies Philadelp··· United States Mick Brown 0.2 CA-2011-141··· 191
2011-01-07 00:··· Furniture Henderson United States Maria Etezadi 0 CA-2011-167··· 424
2011-01-07 00:··· Office Supplies Athens United States Jack OBriant 0 CA-2011-106··· 306
2011-01-07 00:··· Office Supplies Henderson United States Maria Etezadi 0 CA-2011-167··· 424
2011-01-07 00:··· Office Supplies Henderson United States Maria Etezadi 0 CA-2011-167··· 424
2011-01-07 00:··· Office Supplies Henderson United States Maria Etezadi 0 CA-2011-167··· 424
2011-01-07 00:··· Office Supplies Henderson United States Maria Etezadi 0 CA-2011-167··· 424
2011-01-07 00:··· Office Supplies Los Angeles United States Lycoris Saunders 0 CA-2011-130··· 900
컬럼 구분자 ,
라인 구분자 \n
첫번째 행을 헤더로 사용합니다. (선택하지 않은 경우 새 행이 생성되고 헤더로 사용됨)
취소
다음



Each column functions as a "dimension" or "measure." See *"Dimensions" and "Measures"* for further details. In this data, the `Discount`, `Profit`, `Quantity`, `Sales`, `DaystoShipActual`, `SalesForecast`, `DaystoShipScheduled`, `SalesperCustomer`, and `ProfitRatio` columns must be converted into measures.

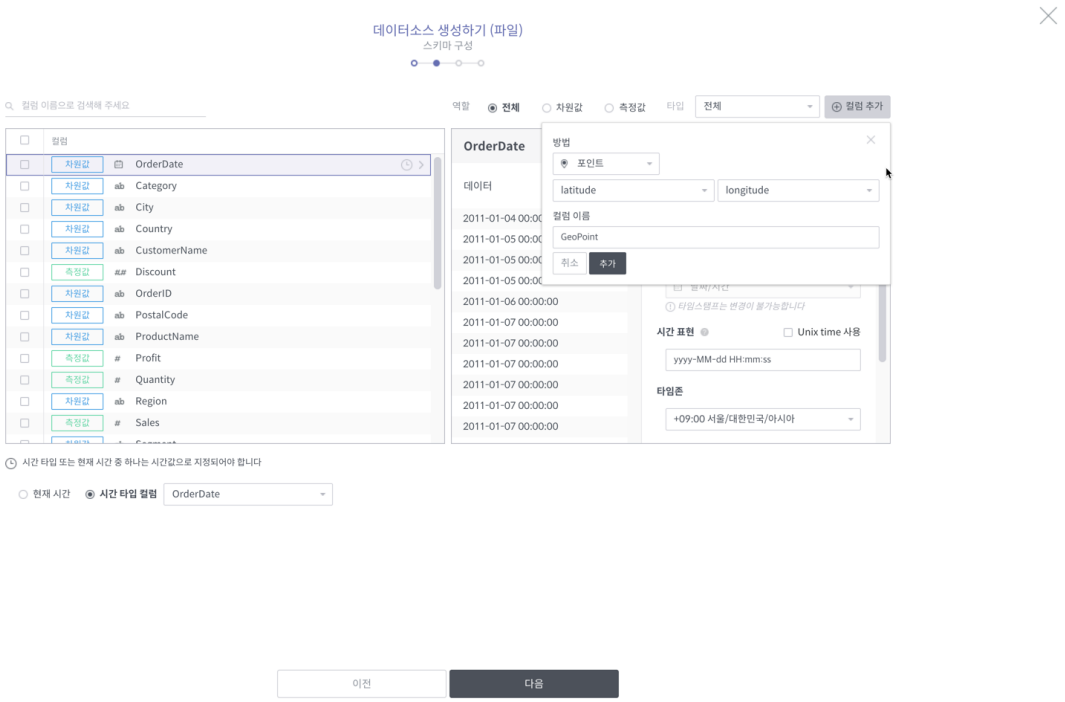
Next, the data types of columns must be adjusted properly. The string type is the default setting for dimensions, and the integer type for measures. While viewing the sample, change the data type settings properly. Below is a list of items to be modified in this data.

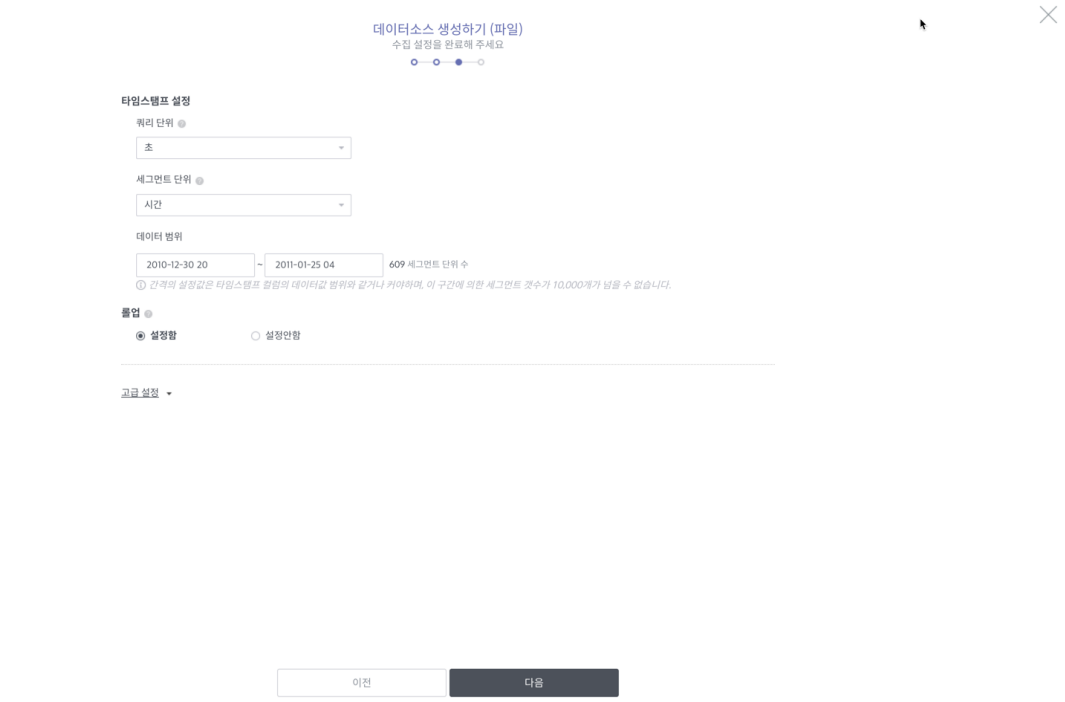
- `Orderdate` : Date/Time
- `Discount` : Decimal
- `ShipDate` : Date/Time (Change the time format to yyyy. MM. dd. and click the checkbox to validate)
- `SalesperCustomer` : Decimal
- `ProfitRatio` : Decimal
- `latitude` : Latitude
- `longitude` : Longitude

Lastly, you should create a new column. Since we already have columns for latitude and longitude, we can create a point type column. Click the **Add column** button on the upper right. Select the `latitude` column for the **Latitude** column, and the `longitude` column for the **Longitude** column. Name the columns appropriately, and click **Add**. A new point type column is created!

Once you are done with schema configuration, click **Next**. If necessary, you can change the settings for ingestion into Druid. The default settings are sufficient for now.

Lastly, enter the **Name** and **Description** for the data source. Click **Done** to proceed to the data source details page.


데이터소스 생성하기 (파일)
스키마 구성
컬럼 이름으로 검색해 주세요
역할
전체
차원값
측정값
타입
전체
컬럼 추가
컬럼
OrderDate
Category
City
Country
CustomerName
Discount
OrderID
PostalCode
ProductName
Profit
Quantity
Region
Sales
방법
포인트
latitude
longitude
컬럼 이름
GeoPoint
취소
추가
데이터
시간 표현
Unix time 사용
yyyy-MM-dd HH:mm:ss
타임존
+09:00 서울/대한민국/아시아
시간 타입 또는 현재 시간 중 하나는 시간값으로 지정되어야 합니다
현재 시간
시간 타입 컬럼
OrderDate
이전
다음


데이터소스 생성하기 (파일)
수집 설정을 완료해 주세요
타임스탬프 설정
쿼리 단위
초
세그먼트 단위
시간
데이터 범위
2010-12-30 20
2011-01-25 04
609 세그먼트 단위 수
롤업
설정함
설정안함
고급 설정
이전
다음

In the data source details page, you can view the ingestion status in real time. The screen below appears after a few minutes, indicating success. A histogram is displayed. If you encounter an error while ingesting another data source, click **Details** to view the Druid ingestion log. Ingestion may be unsuccessful due to a duplicate column name or mismatch between column types and their data. Try ingestion again after addressing the issue.

To make the data source available to other users, check the checkbox next to **Allow all workspaces to use this datasource** under **Publish**. To make the data source available only to specific users, click **Edit** and select individual users' or teams' workspaces as desired.

In this example, we will choose **Open Data** to make it available to all users.

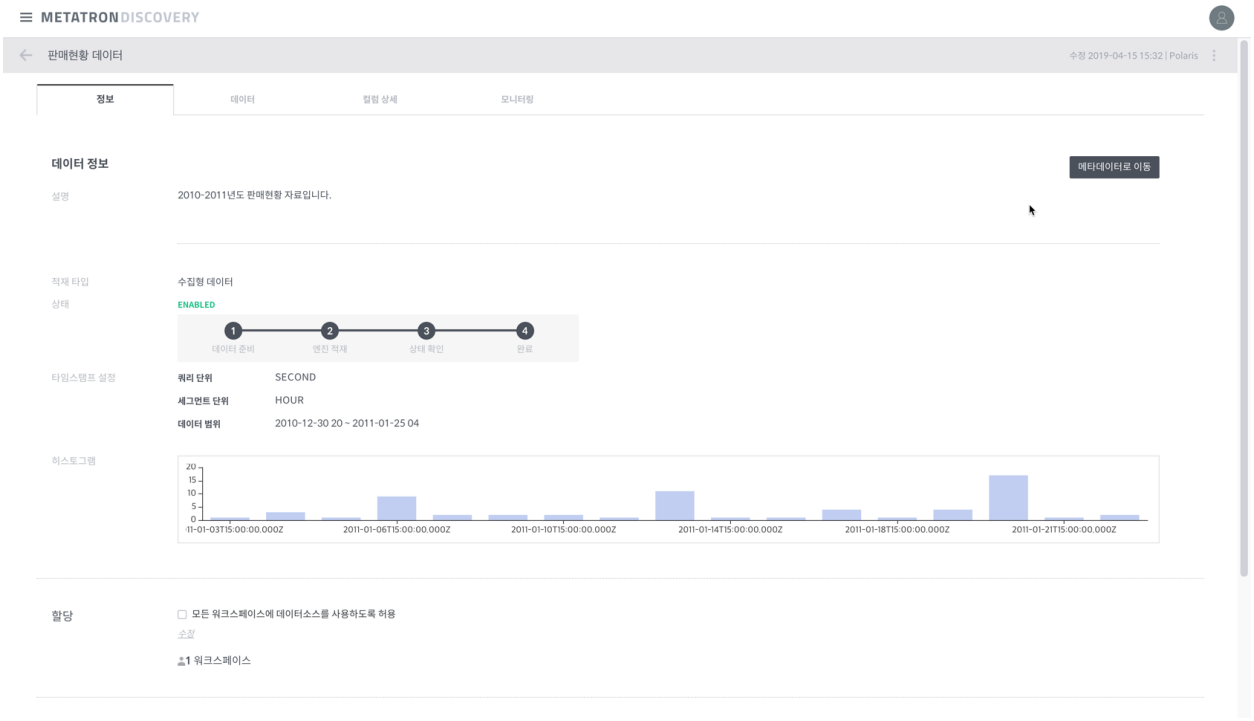
The ingested data can be viewed under the **Data** tab.

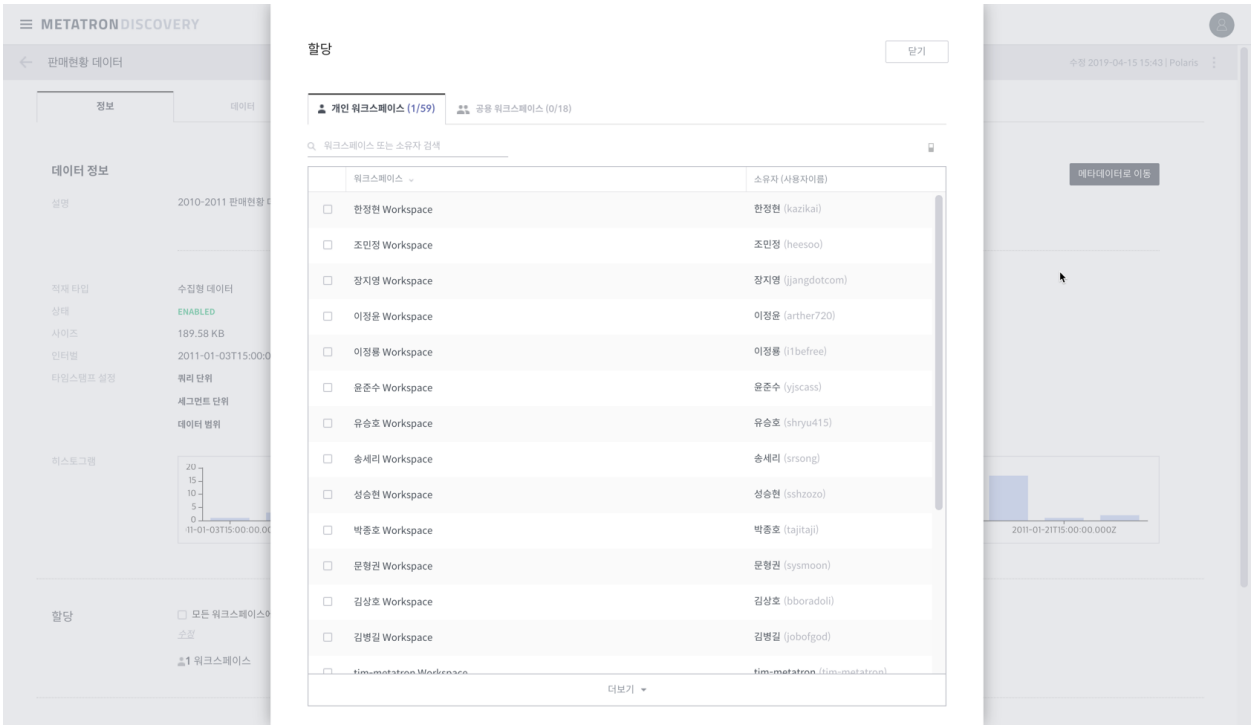
Congratulations! Now, it's time to use the data source. Let's proceed to the next step.

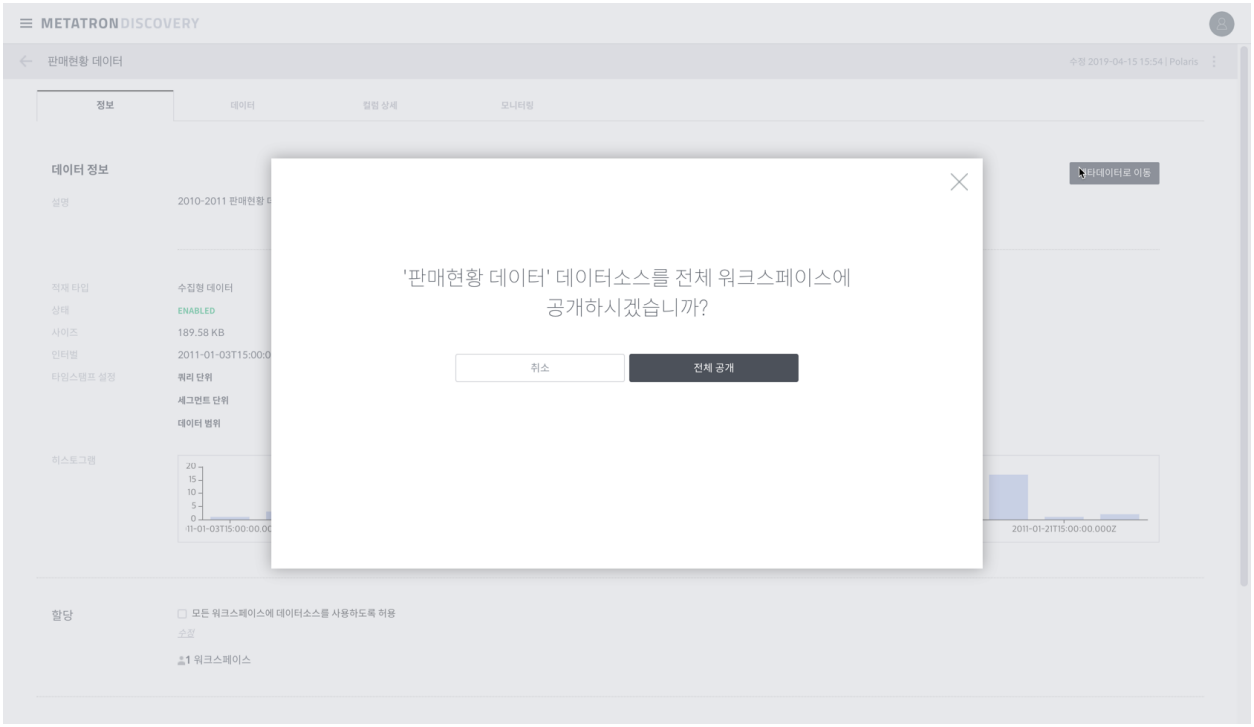
- *Step 2. Create a workbook*

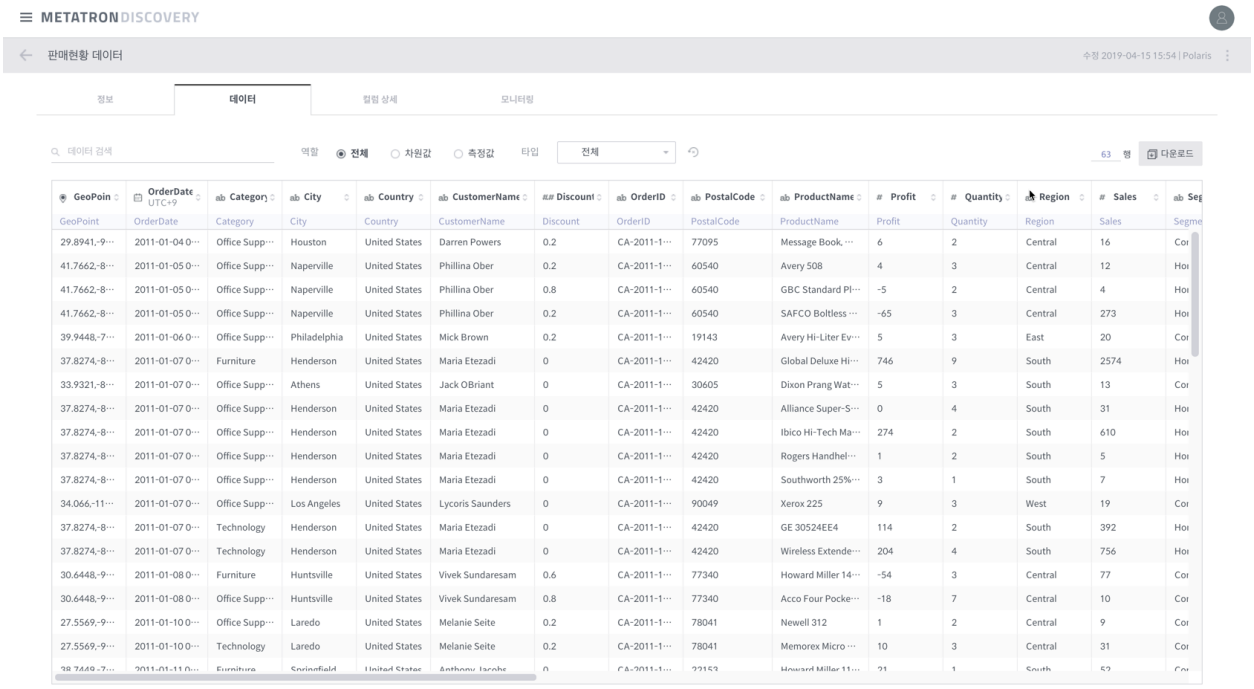
1.2 Step 2. Create a workbook

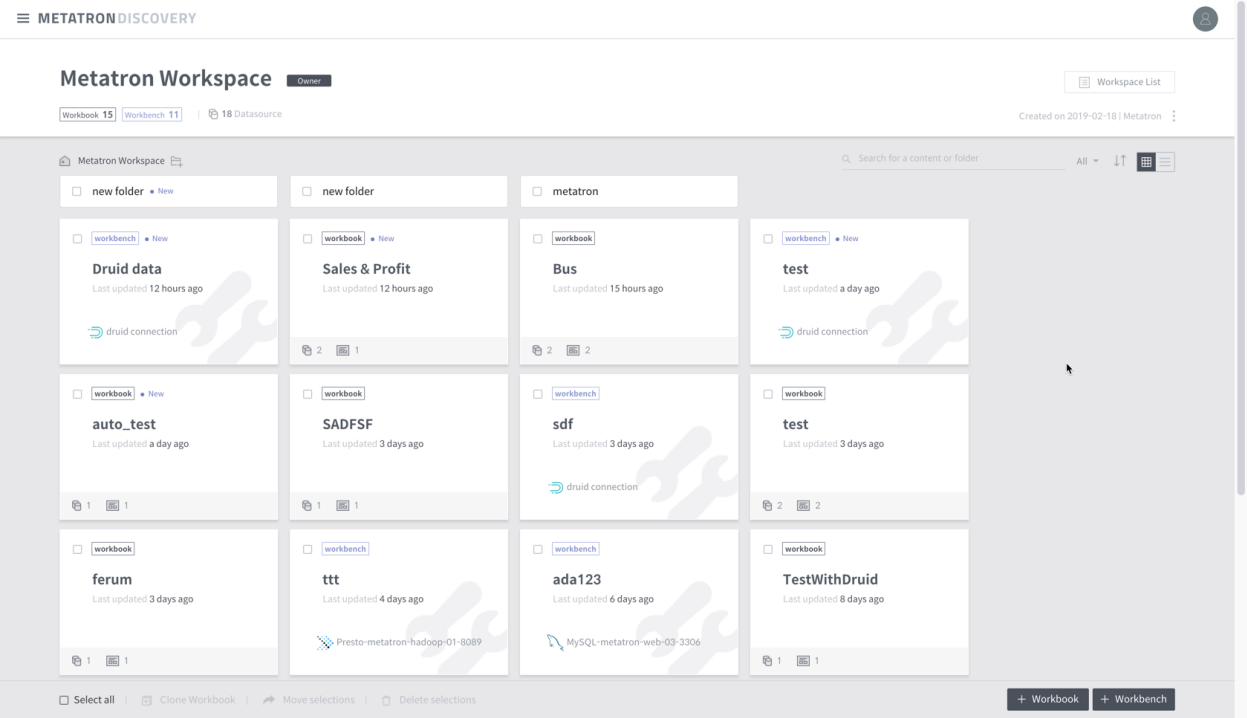
Do you have the data ready for analysis? Now, it's time to create a workbook. The Workbook module supports the visualization of data. Click the Metatron Discovery logo on the upper left to enter your personal workspace.


METATRONDISCOVERY
판매현황 데이터
수정 2019-04-15 15:32 | Polaris
정보
데이터
컬럼 상세
모니터링
데이터 정보
메타데이터로 이동
설명
2010-2011년도 판매현황 자료입니다.
적재 타입
수집형 데이터
상태
ENABLED
1
2
3
4
데이터 준비
엔진 적재
상태 확인
완료
타임스탬프 설정
쿼리 단위
SECOND
세그먼트 단위
HOUR
데이터 범위
2010-12-30 20 ~ 2011-01-25 04
히스토그램
20
15
10
5
0
2011-01-03T15:00:00.000Z
2011-01-06T15:00:00.000Z
2011-01-10T15:00:00.000Z
2011-01-14T15:00:00.000Z
2011-01-18T15:00:00.000Z
2011-01-21T15:00:00.000Z
할당
모든 워크스페이스에 데이터소스를 사용하도록 허용
수정
1 워크스페이스


METATRONDISCOVERY
판매현황 데이터
정보
데이터
데이터 정보
설명
2010-2011 판매현황
적재 타입
수집형 데이터
상태
ENABLED
사이즈
189.58 KB
인터벌
2011-01-03T15:00:0
타임스탬프 설정
쿼리 단위
세그먼트 단위
데이터 범위
히스토그램
20
15
10
5
0
2011-01-03T15:00:00.00
할당
모든 워크스페이스에
수정
1 워크스페이스
할당
닫기
개인 워크스페이스 (1/59)
공용 워크스페이스 (0/18)
워크스페이스 또는 소유자 검색
워크스페이스
소유자 (사용자이름)
한정현 Workspace
한정현 (kazikai)
조민정 Workspace
조민정 (heesoo)
장지영 Workspace
장지영 (jjangdotcom)
이정윤 Workspace
이정윤 (arther720)
이정룡 Workspace
이정룡 (i1befree)
윤준수 Workspace
윤준수 (yjscass)
유승호 Workspace
유승호 (shryu415)
송세리 Workspace
송세리 (srsong)
성승현 Workspace
성승현 (sshzozo)
박종호 Workspace
박종호 (tajitaji)
문형권 Workspace
문형권 (sysmoon)
김상호 Workspace
김상호 (bboradoli)
김병길 Workspace
김병길 (jobofgod)
더보기
수정 2019-04-15 15:43 | Polaris
메타데이터로 이동
2011-01-21T15:00:00.000Z


METATRONDISCOVERY
판매현황 데이터
수정 2019-04-15 15:54 | Polaris
정보
데이터
컬럼 상세
모니터링
데이터 정보
메타데이터로 이동
설명
2010-2011 판매현황
적재 타입
수집형 데이터
상태
ENABLED
사이즈
189.58 KB
인터벌
2011-01-03T15:00:0
타임스탬프 설정
쿼리 단위
세그먼트 단위
데이터 범위
히스토그램
2011-01-21T15:00:00.000Z
'판매현황 데이터' 데이터소스를 전체 워크스페이스에
공개하시겠습니까?
취소
전체 공개
할당
모든 워크스페이스에 데이터소스를 사용하도록 허용
수정
1 워크스페이스


METATRONDISCOVERY
판매현황 데이터
수정 2019-04-15 15:54 | Polaris
정보
데이터
컬럼 상세
모니터링
데이터 검색
역할
전체
차원값
측정값
타입
전체
63 행
다운로드
GeoPoin
OrderDate UTC+9
Category
City
Country
CustomerName
Discount
OrderID
PostalCode
ProductName
Profit
Quantity
Region
Sales
GeoPoint
OrderDate
Category
City
Country
CustomerName
Discount
OrderID
PostalCode
ProductName
Profit
Quantity
Region
Sales
29.8941,-9··· 2011-01-04 0··· Office Supp··· Houston United States Darren Powers 0.2 CA-2011-1··· 77095 Message Book, ··· 6 2 Central 16
41.7662,-8··· 2011-01-05 0··· Office Supp··· Naperville United States Phillina Ober 0.2 CA-2011-1··· 60540 Avery 508 4 3 Central 12
41.7662,-8··· 2011-01-05 0··· Office Supp··· Naperville United States Phillina Ober 0.8 CA-2011-1··· 60540 GBC Standard Pl··· -5 2 Central 4
41.7662,-8··· 2011-01-05 0··· Office Supp··· Naperville United States Phillina Ober 0.2 CA-2011-1··· 60540 SAFCO Boltless ··· -65 3 Central 273
39.9448,-7··· 2011-01-06 0··· Office Supp··· Philadelphia United States Mick Brown 0.2 CA-2011-1··· 19143 Avery Hi-Liter Ev··· 5 3 East 20
37.8274,-8··· 2011-01-07 0··· Furniture Henderson United States Maria Etezadi 0 CA-2011-1··· 42420 Global Deluxe Hi··· 746 9 South 2574
33.9321,-8··· 2011-01-07 0··· Office Supp··· Athens United States Jack OBriant 0 CA-2011-1··· 30605 Dixon Prang Wat··· 5 3 South 13
37.8274,-8··· 2011-01-07 0··· Office Supp··· Henderson United States Maria Etezadi 0 CA-2011-1··· 42420 Alliance Super-S··· 0 4 South 31
37.8274,-8··· 2011-01-07 0··· Office Supp··· Henderson United States Maria Etezadi 0 CA-2011-1··· 42420 Ibico Hi-Tech Ma··· 274 2 South 610
37.8274,-8··· 2011-01-07 0··· Office Supp··· Henderson United States Maria Etezadi 0 CA-2011-1··· 42420 Rogers Handhel··· 1 2 South 5
37.8274,-8··· 2011-01-07 0··· Office Supp··· Henderson United States Maria Etezadi 0 CA-2011-1··· 42420 Southworth 25%··· 3 1 South 7
34.066,-11··· 2011-01-07 0··· Office Supp··· Los Angeles United States Lycoris Saunders 0 CA-2011-1··· 90049 Xerox 225 9 3 West 19
37.8274,-8··· 2011-01-07 0··· Technology Henderson United States Maria Etezadi 0 CA-2011-1··· 42420 GE 30524EE4 114 2 South 392
37.8274,-8··· 2011-01-07 0··· Technology Henderson United States Maria Etezadi 0 CA-2011-1··· 42420 Wireless Extende··· 204 4 South 756
30.6448,-9··· 2011-01-08 0··· Furniture Huntsville United States Vivek Sundaresam 0.6 CA-2011-1··· 77340 Howard Miller 14··· -54 3 Central 77
30.6448,-9··· 2011-01-08 0··· Office Supp··· Huntsville United States Vivek Sundaresam 0.8 CA-2011-1··· 77340 Acco Four Pocke··· -18 7 Central 10
27.5569,-9··· 2011-01-10 0··· Office Supp··· Laredo United States Melanie Seite 0.2 CA-2011-1··· 78041 Newell 312 1 2 Central 9
27.5569,-9··· 2011-01-10 0··· Technology Laredo United States Melanie Seite 0.2 CA-2011-1··· 78041 Memorex Micro ··· 10 3 Central 31



Let's begin by clicking the **+ Workbook** button on the bottom right. Enter the name and description for the workbook. The checkbox is marked by default for you to create a dashboard once a workbook is created. A single workbook contains multiple dashboards, and each single dashboard contains multiple charts.

Proceed with creating a dashboard. A dashboard requires a data source for visualization. This data source can be either a single source, or joined data sources. See *Create a dashboard* for further details. This tutorial uses `Sales Report`, ingested previously in Step 1.

Click the **+ Add data source** button for the data source selection popup. Search `Sales Report`, or select the **Show open data only** checkbox and choose from the results.

Finally, enter the **Name** and **Description** for the dashboard.

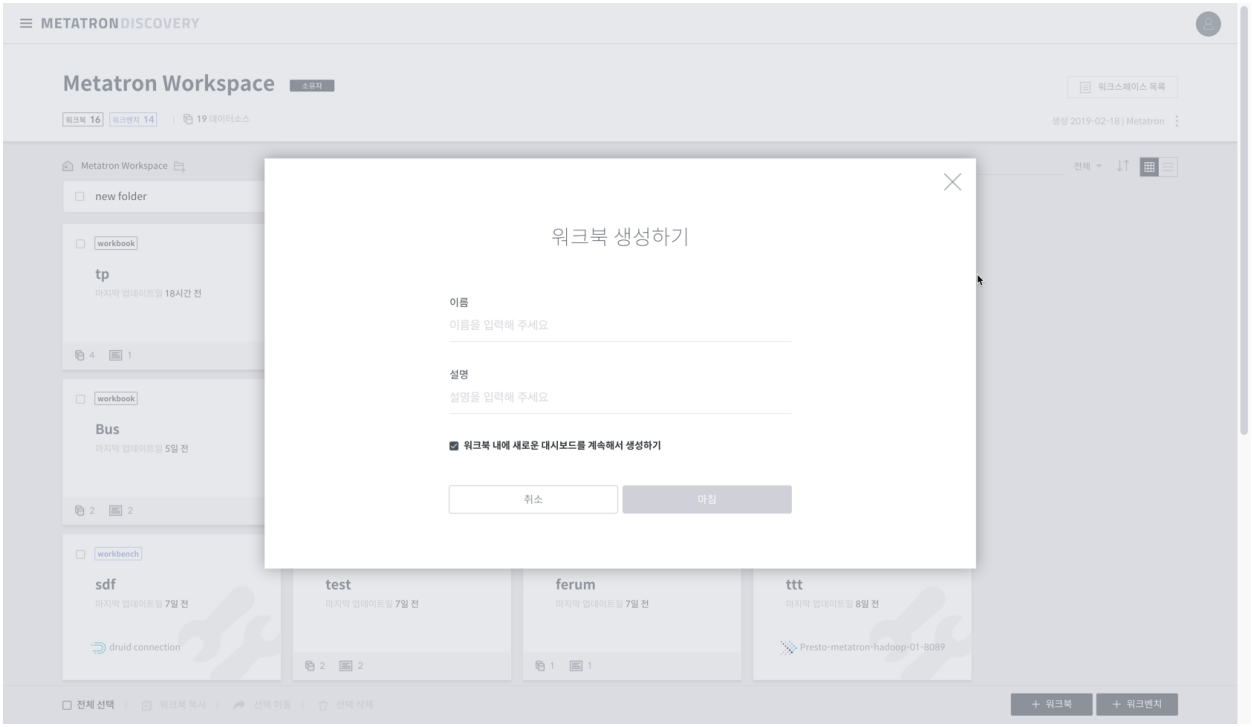
You have created a dashboard in the workbook. Now, you can add widgets to the dashboard.

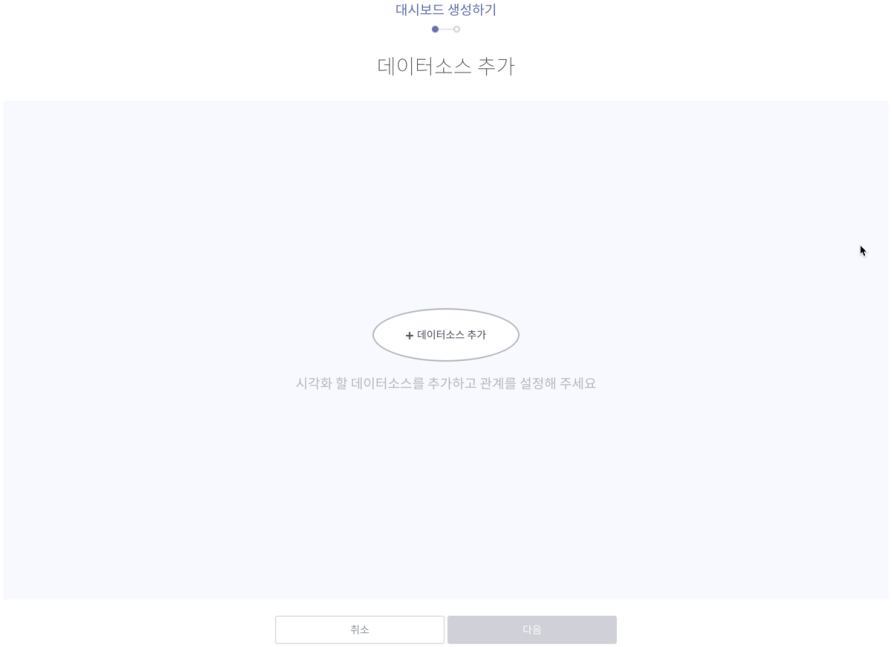
Let's proceed to the next step.

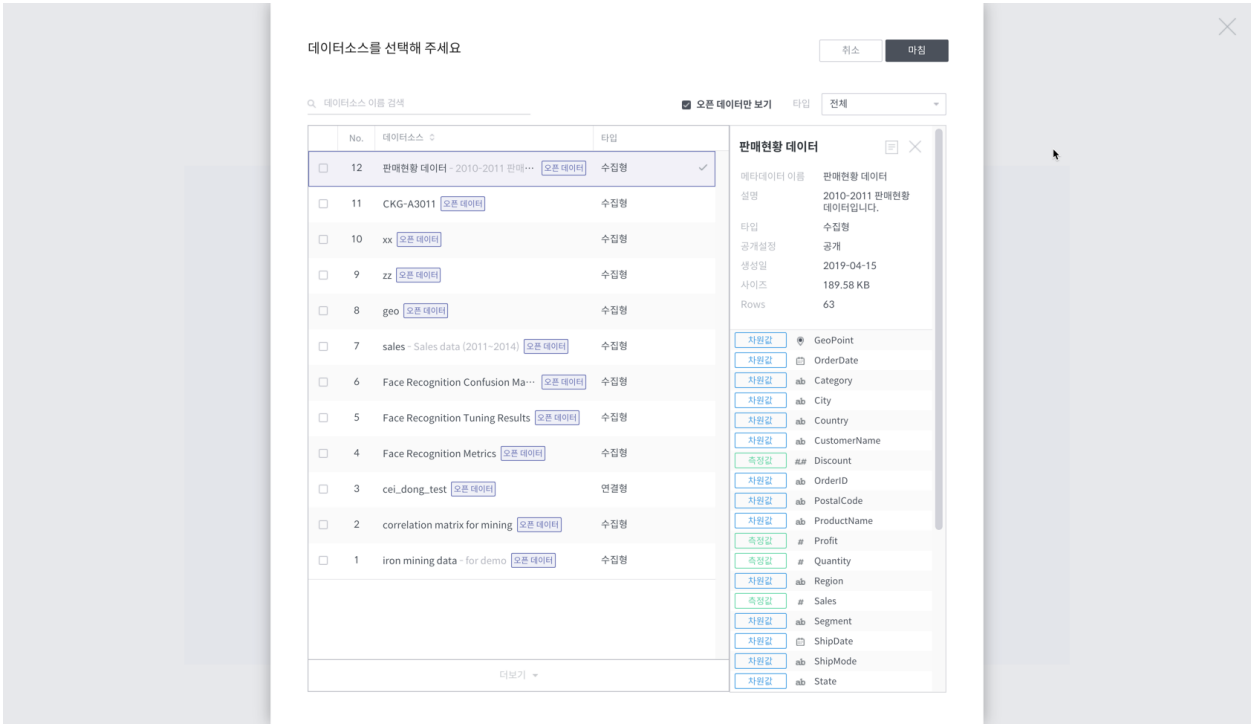
- *Step 3. Organize a dashboard*

1.3 Step 3. Organize a dashboard

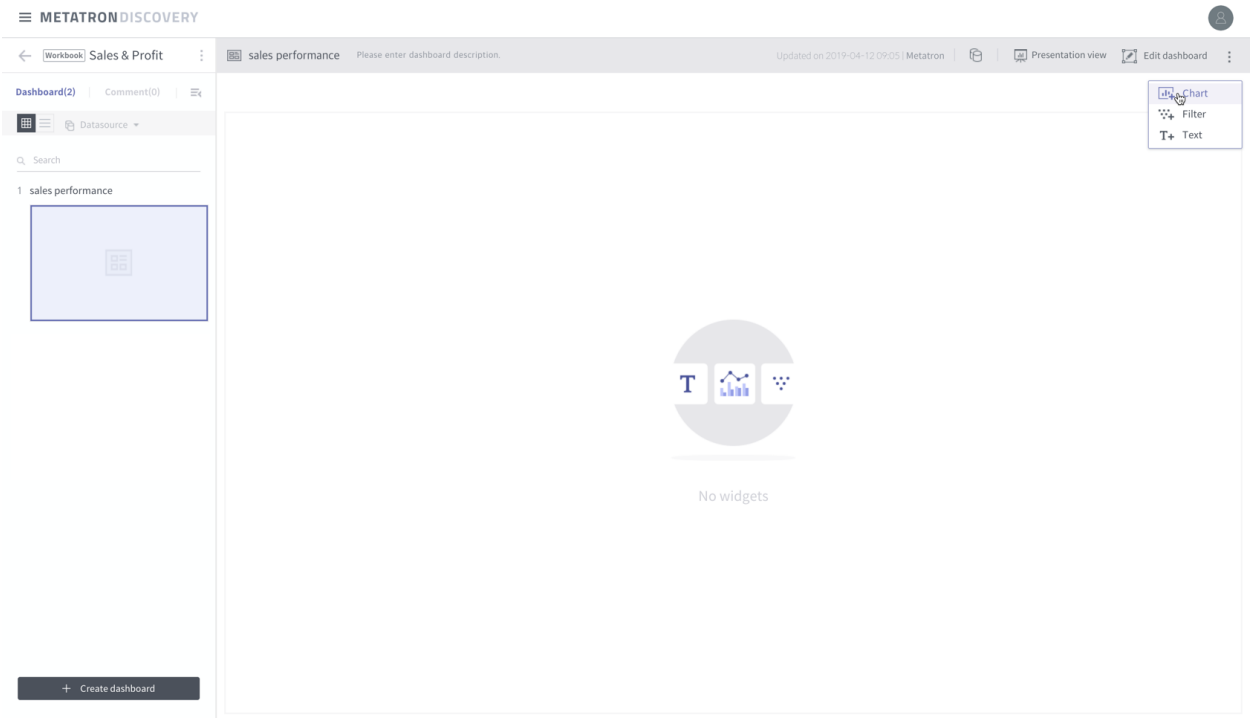
The final step is to create chart widgets, text widgets, and filter widgets to fill the empty dashboard. The dashboard can be edited in the following order:

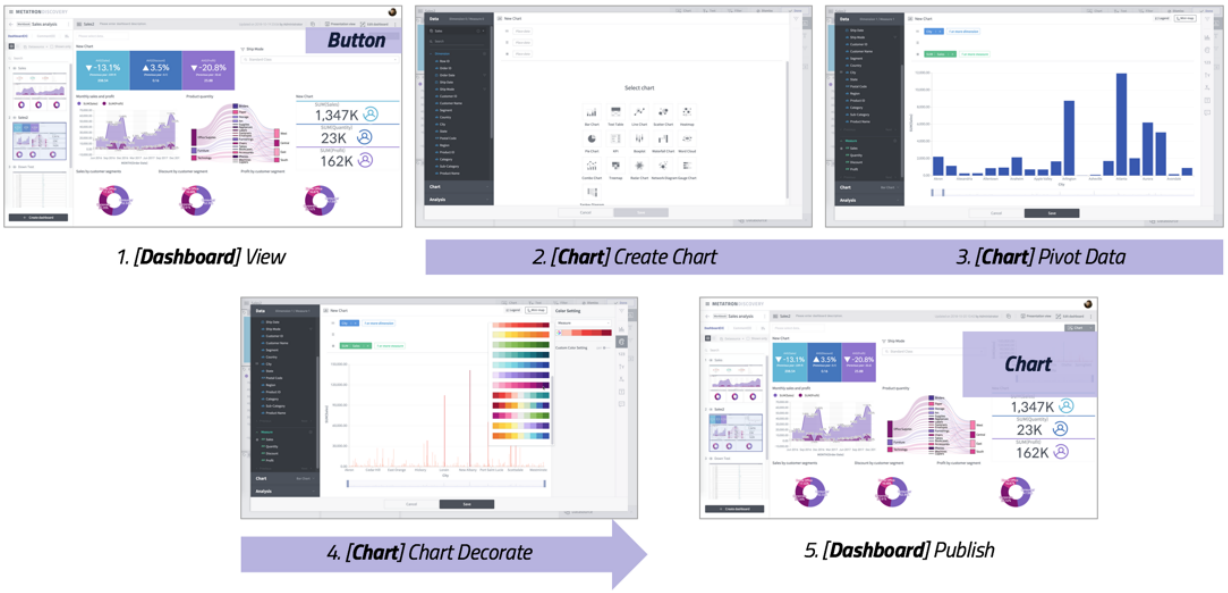

METATRONDISCOVERY
Metatron Workspace
워크스페이스 목록
Metatron Workspace
new folder
workbook
tp
마지막 업데이트일 18시간 전
workbook
Bus
마지막 업데이트일 5일 전
workbench
sdf
마지막 업데이트일 7일 전
druid connection
test
마지막 업데이트일 7일 전
ferum
마지막 업데이트일 7일 전
ttt
마지막 업데이트일 8일 전
Presto-metatron-hadoop-01-8089
워크북 생성하기
이름
이름을 입력해 주세요
설명
설명을 입력해 주세요
워크북 내에 새로운 대시보드를 계속해서 생성하기
취소
마침
전체 선택
+ 워크북
+ 워크벤치


대시보드 생성하기
데이터소스 추가
+ 데이터소스 추가
시각화 할 데이터소스를 추가하고 관계를 설정해 주세요
취소
다음

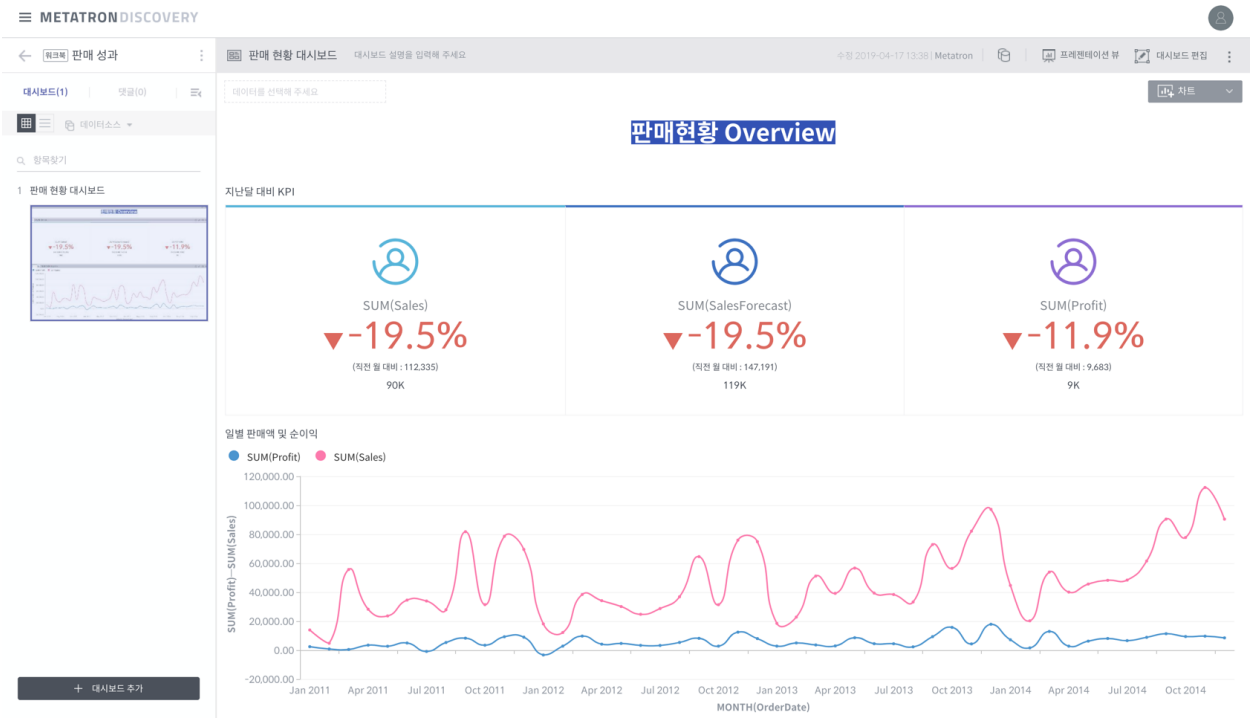

데이터소스를 선택해 주세요
취소
마침
데이터소스 이름 검색
오픈 데이터만 보기
타입
전체
No.
데이터소스
타입
12 판매현황 데이터 - 2010-2011 판매··· 오픈 데이터 수집형
11 CKG-A3011 오픈 데이터 수집형
10 xx 오픈 데이터 수집형
9 zz 오픈 데이터 수집형
8 geo 오픈 데이터 수집형
7 sales - Sales data (2011-2014) 오픈 데이터 수집형
6 Face Recognition Confusion Ma··· 오픈 데이터 수집형
5 Face Recognition Tuning Results 오픈 데이터 수집형
4 Face Recognition Metrics 오픈 데이터 수집형
3 cei_dong_test 오픈 데이터 연결형
2 correlation matrix for mining 오픈 데이터 수집형
1 iron mining data - for demo 오픈 데이터 수집형
더보기
판매현황 데이터
메타데이터 이름 판매현황 데이터
설명 2010-2011 판매현황 데이터입니다.
타입 수집형
공개설정 공개
생성일 2019-04-15
사이즈 189.58 KB
Rows 63
차원값 GeoPoint
차원값 OrderDate
차원값 Category
차원값 City
차원값 Country
차원값 CustomerName
측정값 Discount
차원값 OrderID
차원값 PostalCode
차원값 ProductName
측정값 Profit
측정값 Quantity
차원값 Region
측정값 Sales
차원값 Segment
차원값 ShipDate
차원값 ShipMode
차원값 State


대시보드 생성하기
대시보드 생성을 완성해 주세요
워크북 Sales & Profit
데이터소스 판매현황 데이터
이름
이름을 입력해 주세요
설명
설명을 입력해 주세요
이전
마침


METATRONDISCOVERY
Workbook Sales & Profit
sales performance
Please enter dashboard description.
Presentation view
Edit dashboard
Dashboard(2)
Comment(0)
Datasource
Search
1 sales performance
Create dashboard
Chart
Filter
Text
No widgets


Button
1,347K
23K
162K
1. [Dashboard] View
2. [Chart] Create Chart
3. [Chart] Pivot Data
Chart
1,347K
23K
162K
4. [Chart] Chart Decorate
5. [Dashboard] Publish

Using the `Sales Report` created earlier, let's add a key performance indicator chart and a line chart to the dashboard.



In the empty dashboard, click the **Chart** button to create a chart.

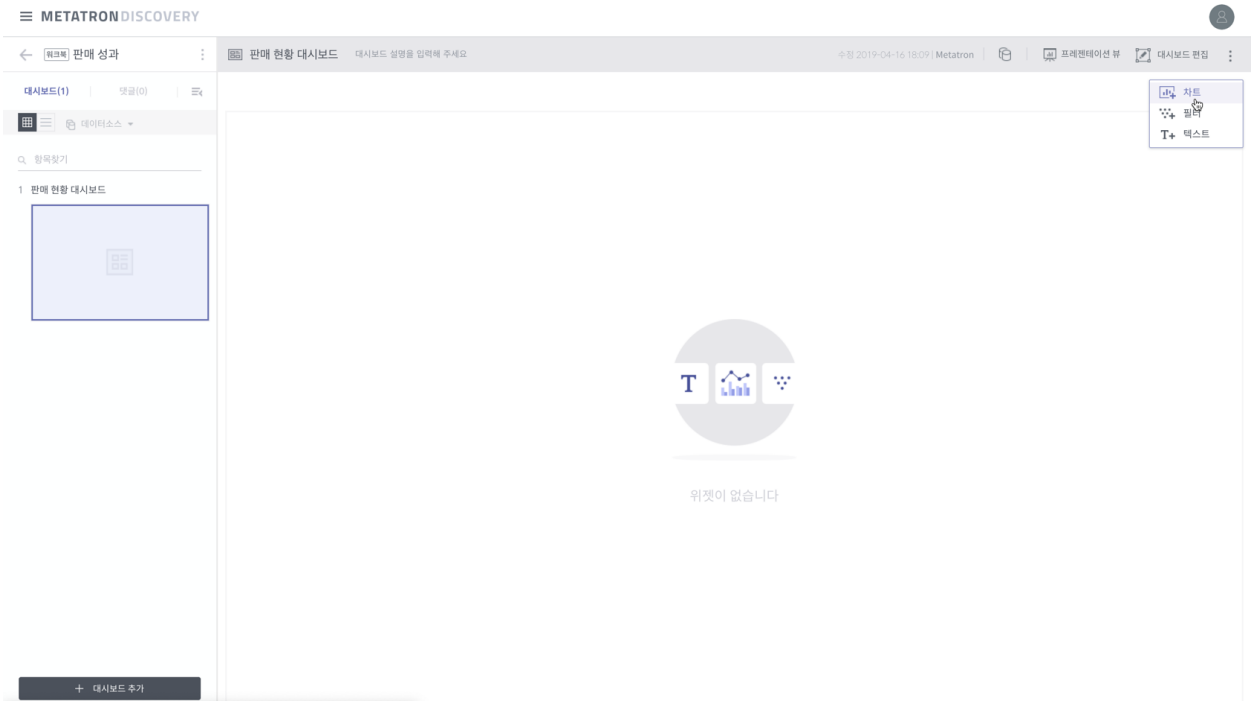
1.3.1 Creating a key performance indicator chart

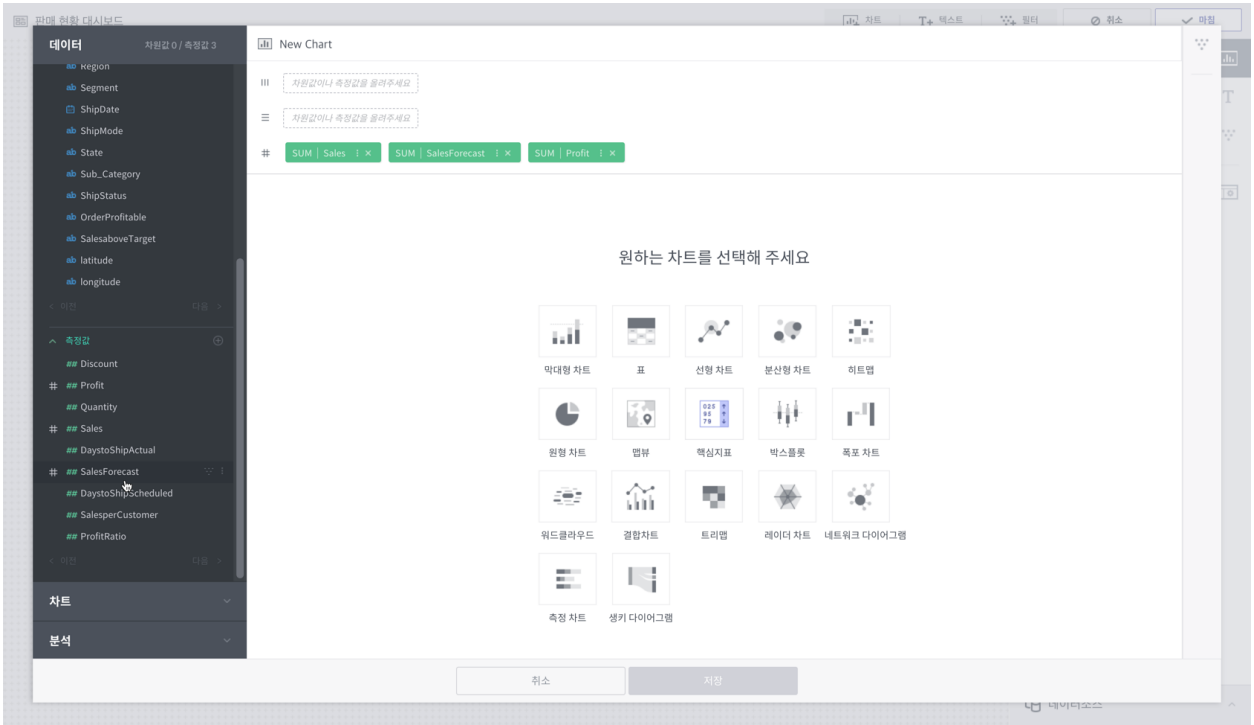
The first chart you will be creating is a key performance indicator (KPI) chart. The KPI chart is a simple yet powerful chart that displays the goals of an organization in an intuitive manner. The goal of our dashboard is to clearly present sales data. As such, the KPI chart should include total sales, sales forecast, and profit. What should we do? Simply click the three measurement columns named "Sales," "SalesForecast," and "Profit" under the Data menu. This task is called pivoting. The pivoted columns are automatically aggregated and placed on shelves. Once columns are on shelves, suitable charts are recommended. How about clicking the recommended **KPI** chart?

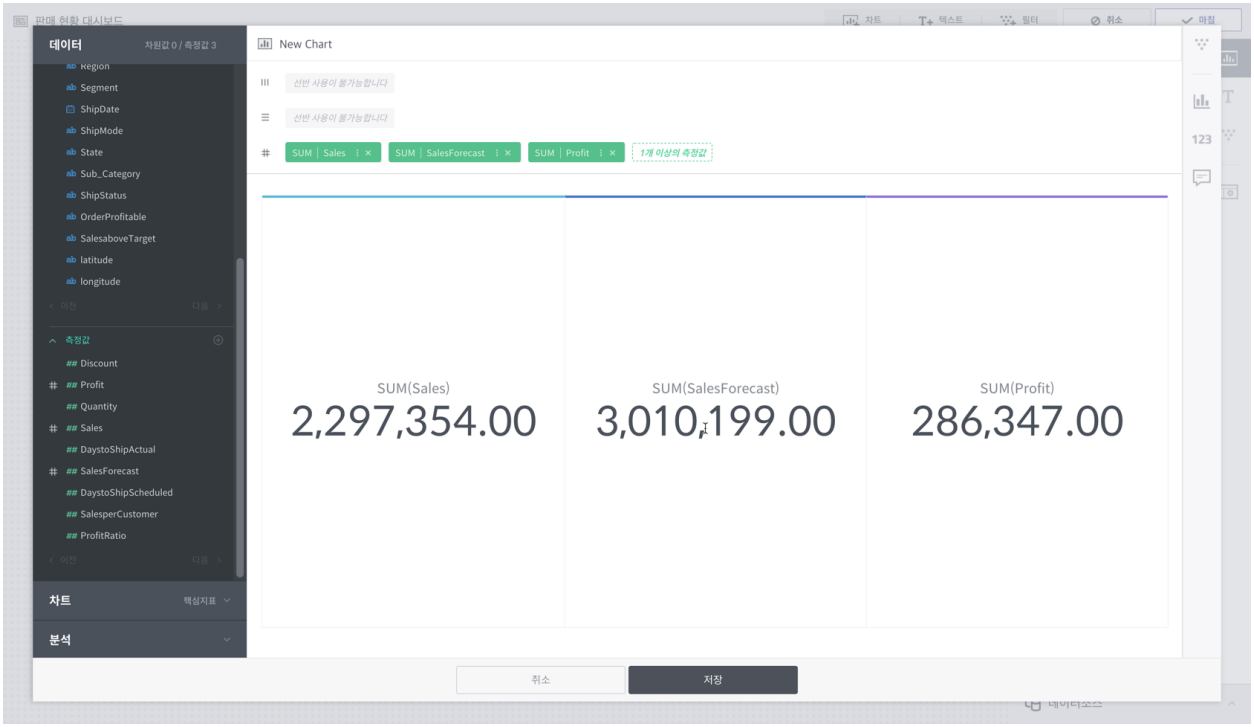
The KPI chart is created as follows: To make it more presentable, let's enter the chart properties menu on the right.

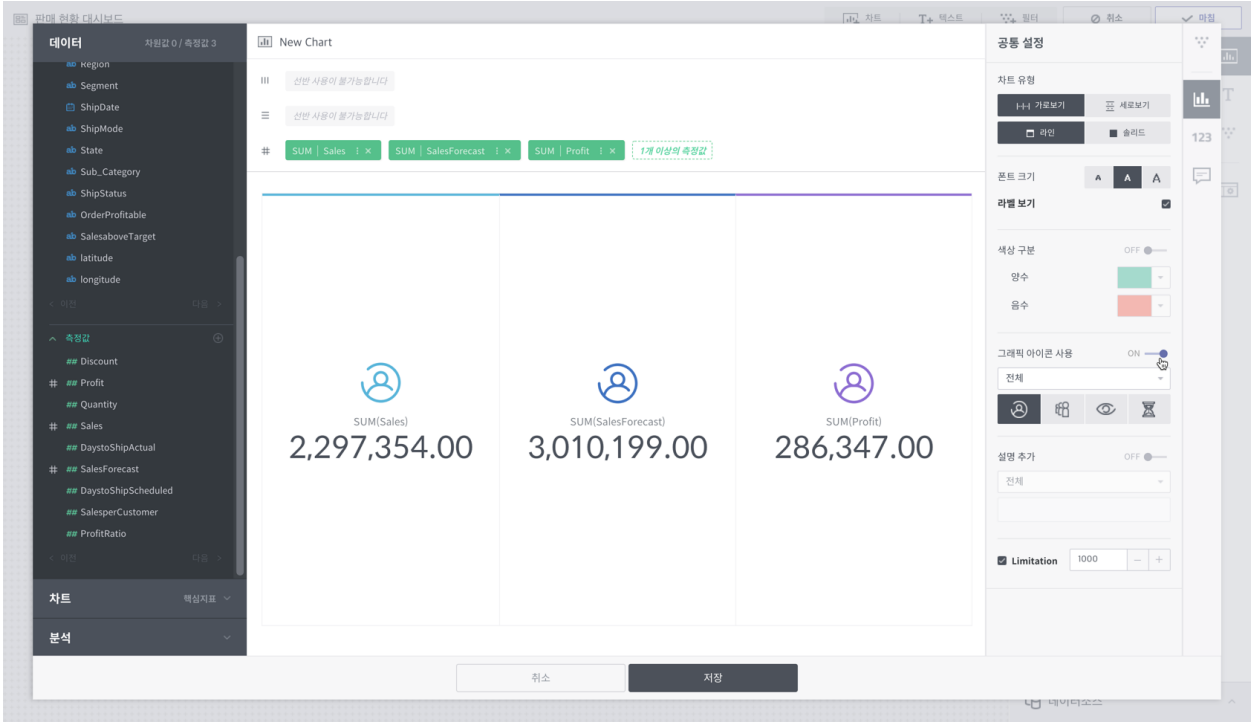
Click  to enter the **Common Setting** panel and add an icon to each measure column.

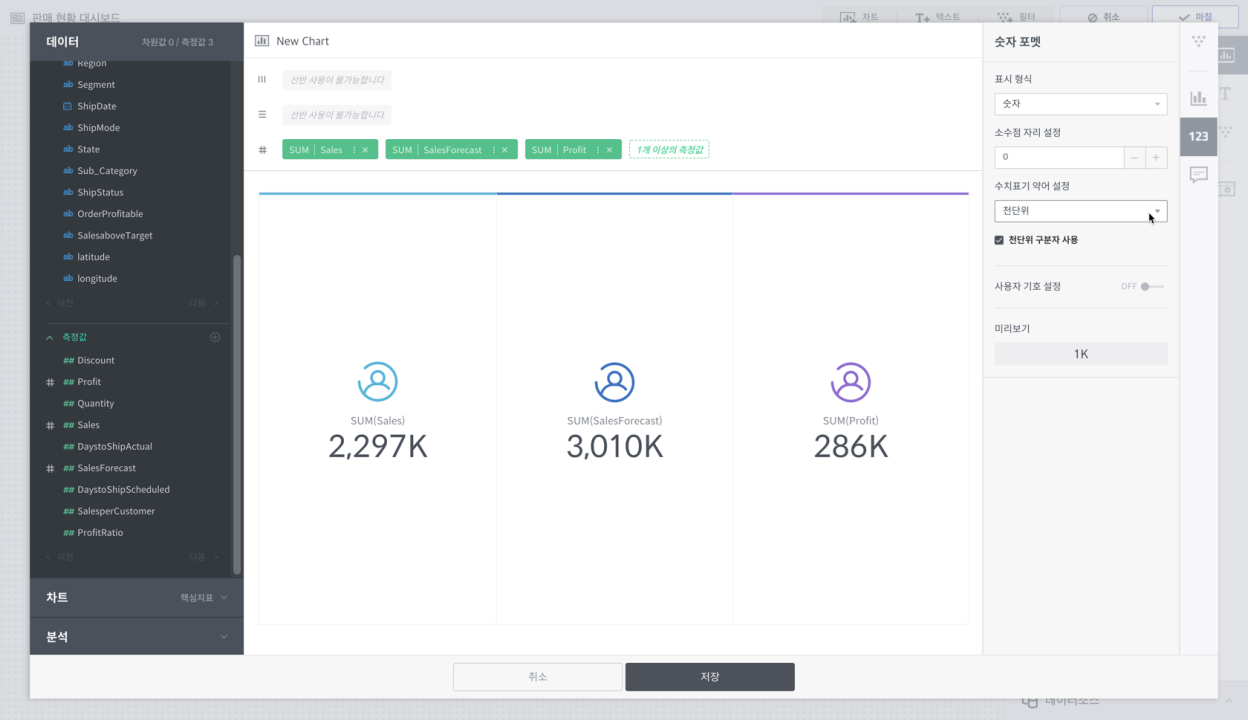
Click  to enter the **Number Format** panel and change the decimal place and abbreviation display.


METATRONDISCOVERY
판매 성과
판매 현황 대시보드
대시보드(1)
댓글(0)
데이터소스
항목찾기
1 판매 현황 대시보드
대시보드 추가
차트
필터
텍스트
위젯이 없습니다


데이터
New Chart
Region
Segment
ShipDate
ShipMode
State
Sub_Category
ShipStatus
OrderProfitable
SalesaboveTarget
latitude
longitude
측정값
Discount
Profit
Quantity
Sales
DaystoShipActual
SalesForecast
DaystoShipScheduled
SalesperCustomer
ProfitRatio
SUM | Sales
SUM | SalesForecast
SUM | Profit
원하는 차트를 선택해 주세요
막대형 차트
표
선형 차트
분산형 차트
히트맵
원형 차트
맵뷰
핵심지표
박스플롯
폭포 차트
워드클라우드
결합차트
트리맵
레이더 차트
네트워크 다이어그램
측정 차트
생키 다이어그램
차트
분석
취소
저장


데이터
New Chart
SUM | Sales
SUM | SalesForecast
SUM | Profit
SUM(Sales)
2,297,354.00
SUM(SalesForecast)
3,010,199.00
SUM(Profit)
286,347.00
차트
분석
취소
저장


데이터
New Chart
공통 설정
SUM | Sales
SUM | SalesForecast
SUM | Profit
SUM(Sales)
2,297,354.00
SUM(SalesForecast)
3,010,199.00
SUM(Profit)
286,347.00
차트
분석
취소
저장



The most important feature of the KPI chart is comparing present achievements with past performance. Click  to enter the **Set up secondary indicators** panel. Set a secondary indicator, and check the % improvement in performance compared to the previous month. If you wish, you can emphasize the secondary indicator instead of the original indicator.

Click **Save** to display the chart in the dashboard.

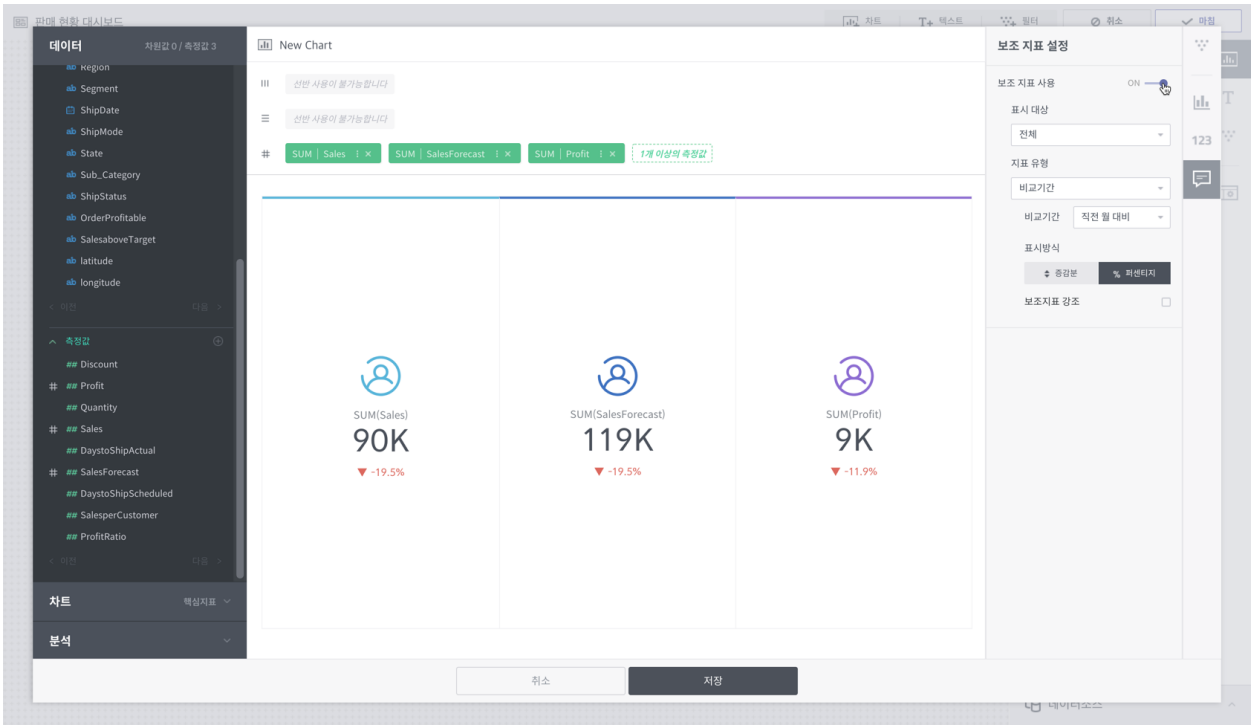
1.3.2 Creating a line chart

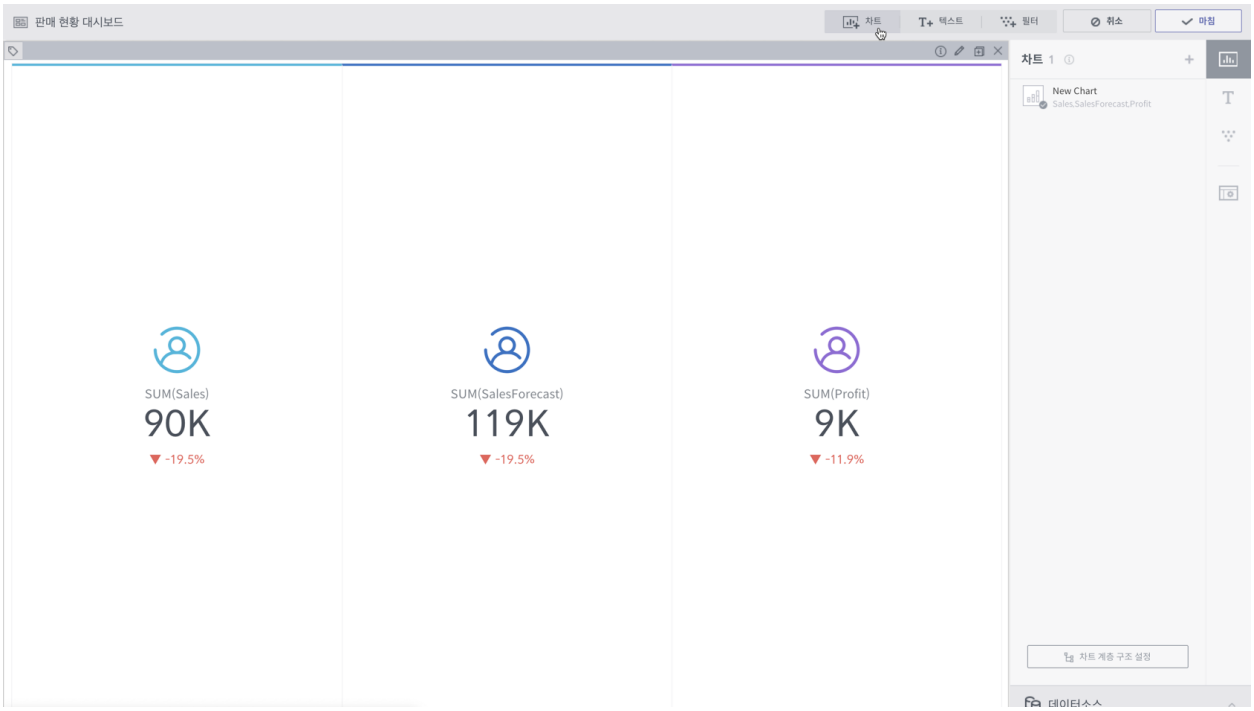
Next, let's create a line chart, the most basic type of chart. Shall we take a look at how sales and profit change over time? Again, click the **Chart** button to begin drawing a new chart. Click the `OrderDate`, `Profit`, and `Sales` columns to see how the values change over time. Click the recommended **Line Chart**.

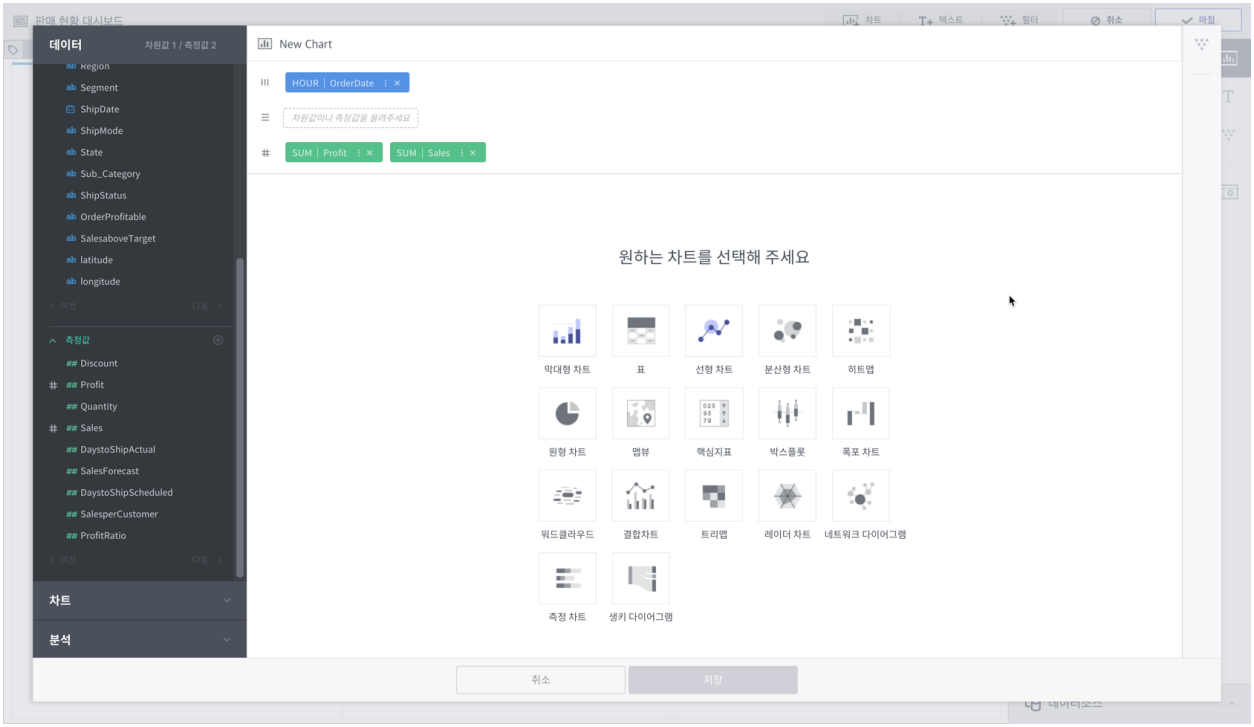
A line chart is drawn. Open the chart properties panel, and change the line shape to "round."

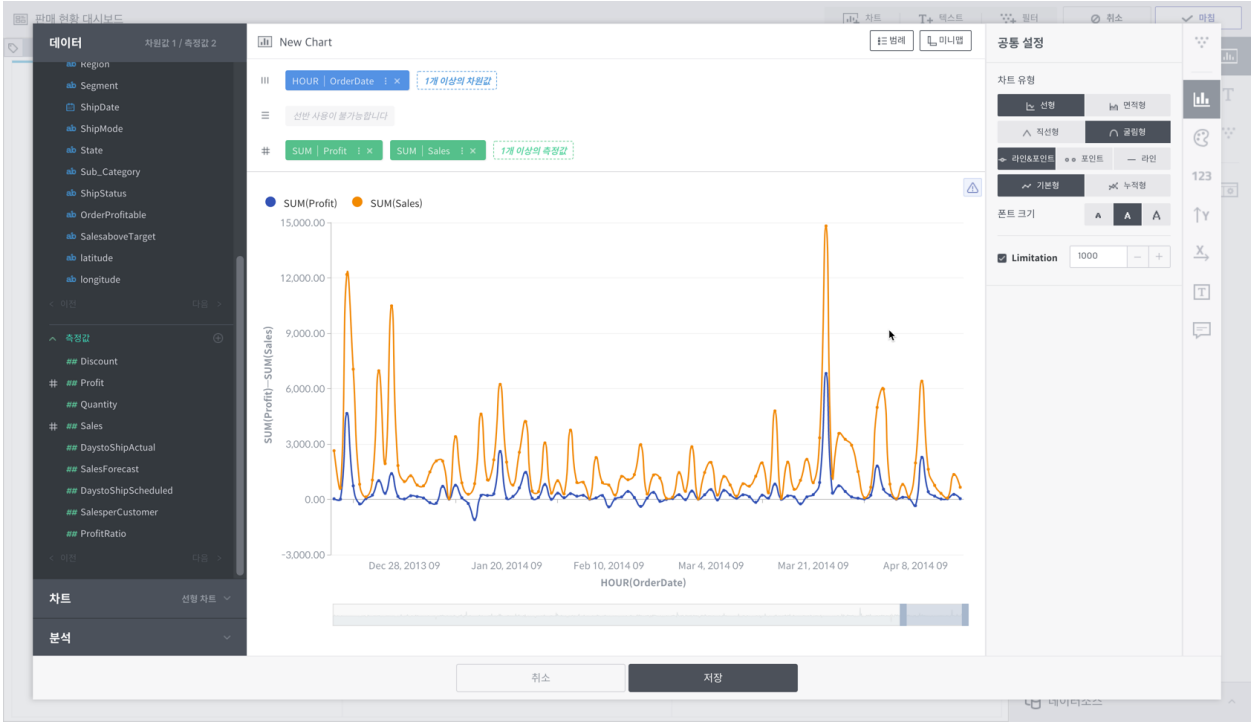
There is too much data as `OrderDate` is aggregated on an hourly basis. To view by month, go to the menu of the `OrderDate` column, and set **Granularity** as **Month**. The entire data is displayed now! Click **Mini Map** on the upper right to remove the mini map from the chart.

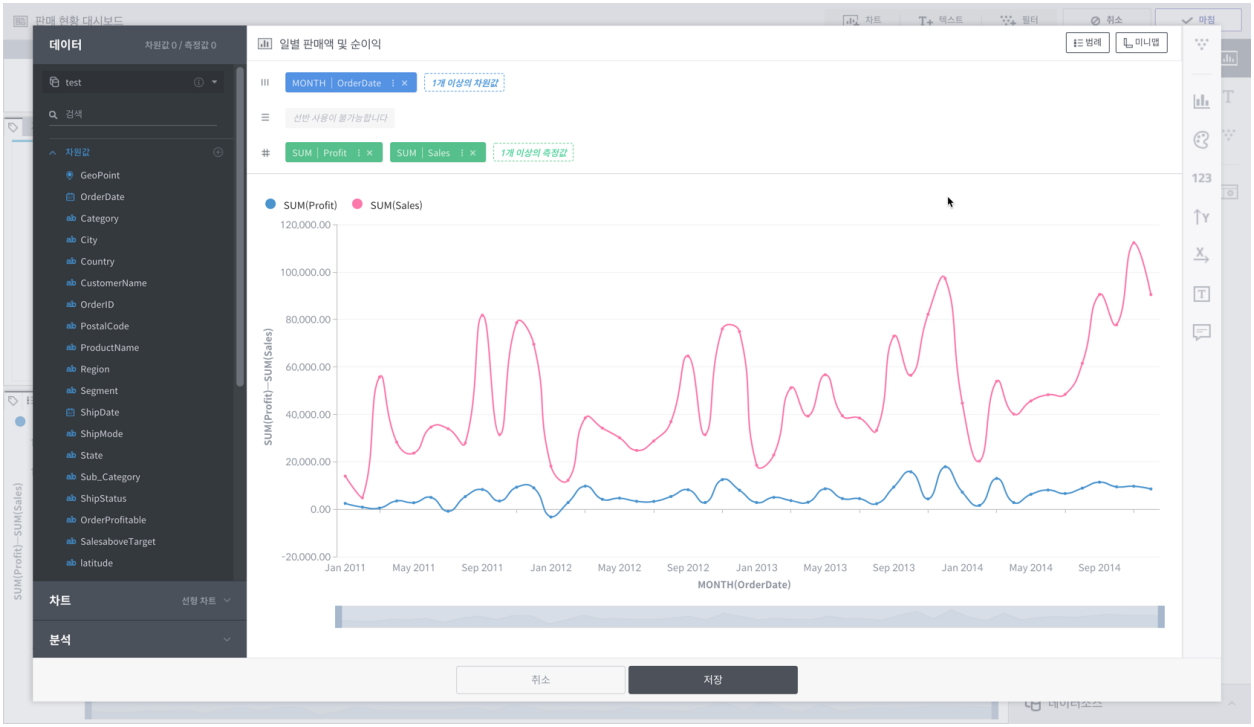
Click  on the right menu, and change colors using the **Color Setting** panel.

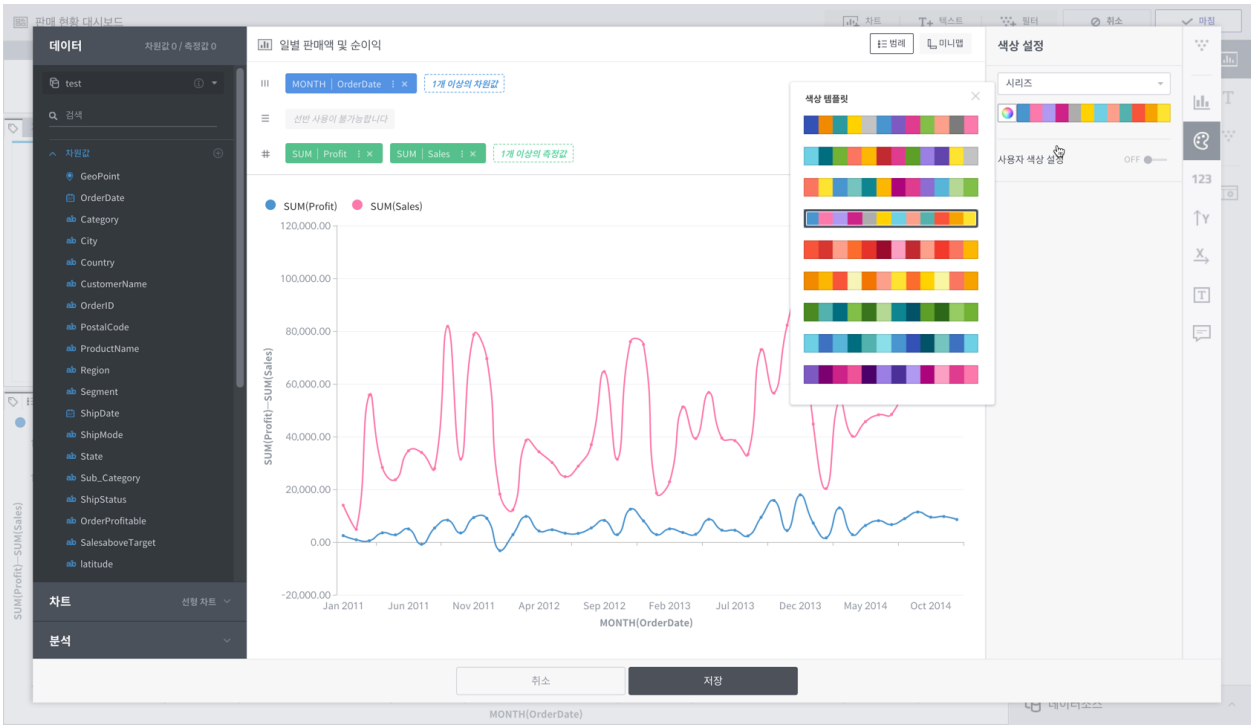

판매 현황 대시보드
데이터
New Chart
보조 지표 설정
SUM(Sales)
90K
▼ -19.5%
SUM(SalesForecast)
119K
▼ -19.5%
SUM(Profit)
9K
▼ -11.9%
취소
저장


판매 현황 대시보드
차트
텍스트
필터
취소
마침
차트 1
New Chart
SUM(Sales)
90K
▼ -19.5%
SUM(SalesForecast)
119K
▼ -19.5%
SUM(Profit)
9K
▼ -11.9%

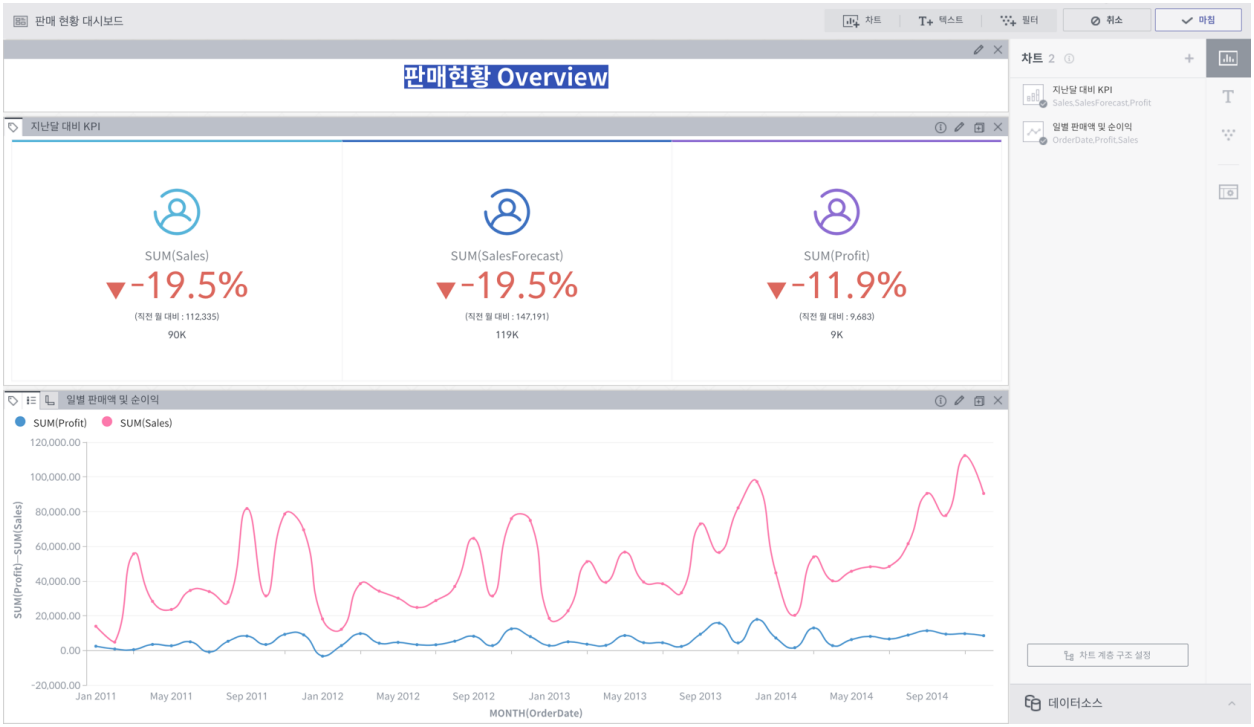

데이터
New Chart
HOUR | OrderDate
SUM | Profit
SUM | Sales
원하는 차트를 선택해 주세요
막대형 차트
표
선형 차트
분산형 차트
히트맵
원형 차트
맵뷰
핵심지표
박스플롯
폭포 차트
워드클라우드
결합차트
트리맵
레이더 차트
네트워크 다이어그램
측정 차트
생키 다이어그램
차트
분석
취소
저장


데이터
New Chart
공통 설정
HOUR | OrderDate
SUM | Profit
SUM | Sales
SUM(Profit)
SUM(Sales)
15,000.00
12,000.00
9,000.00
6,000.00
3,000.00
0.00
-3,000.00
Dec 28, 2013 09
Jan 20, 2014 09
Feb 10, 2014 09
Mar 4, 2014 09
Mar 21, 2014 09
Apr 8, 2014 09
HOUR(OrderDate)
차트 유형
Limitation
1000
차트
분석
취소
저장


데이터
일별 판매액 및 순이익
MONTH | OrderDate
SUM | Profit
SUM | Sales
SUM(Profit)
SUM(Sales)
MONTH(OrderDate)
차트
분석
취소
저장


데이터
일별 판매액 및 순이익
색상 설정
색상 템플릿
MONTH | OrderDate
SUM | Profit
SUM | Sales
SUM(Profit)
SUM(Sales)
MONTH(OrderDate)
차트
분석
취소
저장

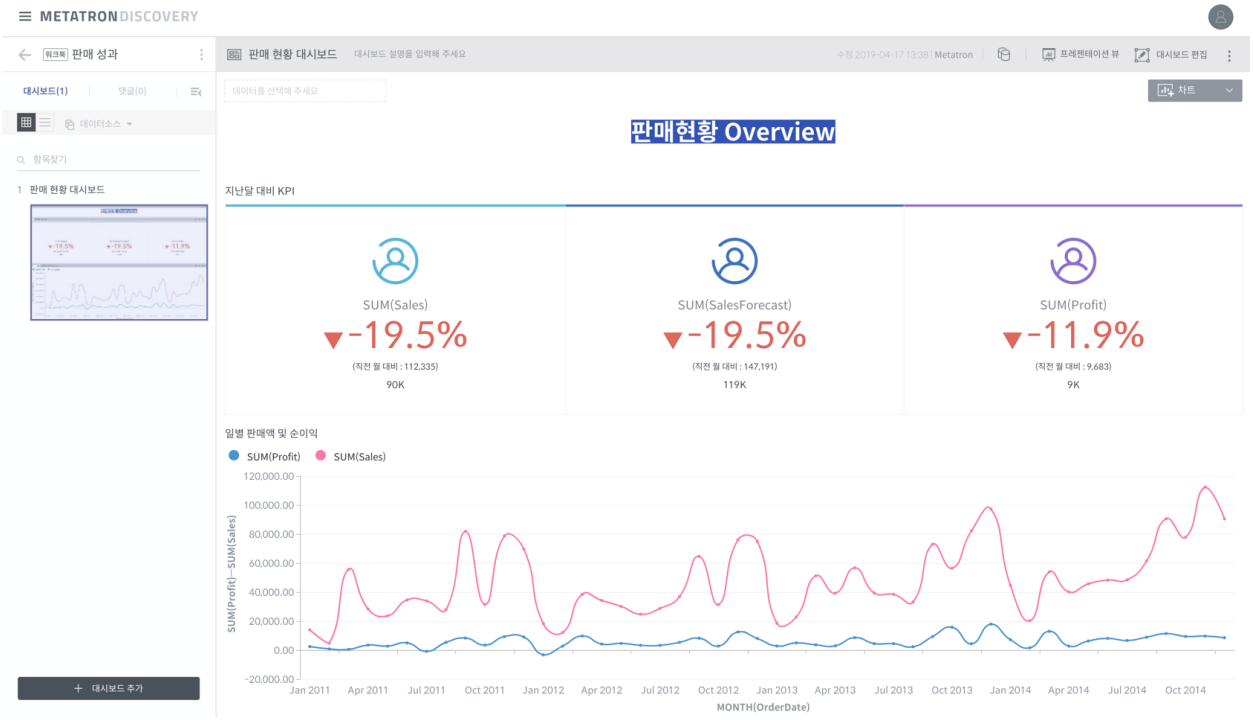
Click Save, and drag and drop the chart to the desired position. Add information to the dashboard by adding a **text widget**. Click **Done** to finish dashboard editing.



In this tutorial, you learned how to draw two chart types. Using the interactive dashboard, you can select a chart or add filters to present data as desired. You can also modify, add, or delete charts if required.

Are you ready to learn more about Metatron Discovery?

- *Overview of Metatron Discovery*
- *Components of Metatron Discovery*
- *Metatron engine: Druid*


METATRONDISCOVERY
판매 성과
판매 현황 대시보드
대시보드 설명을 입력해 주세요
프레젠테이션 뷰
대시보드 편집
대시보드(1)
댓글(0)
데이터소스
항목찾기
1 판매 현황 대시보드
대시보드 추가
데이터를 선택해 주세요
차트
판매현황 Overview
지난달 대비 KPI
SUM(Sales)
▼-19.5%
(직전 월 대비 : 112,335)
90K
SUM(SalesForecast)
▼-19.5%
(직전 월 대비 : 147,191)
119K
SUM(Profit)
▼-11.9%
(직전 월 대비 : 9,683)
9K
일별 판매액 및 순이익
SUM(Profit)
SUM(Sales)
120,000.00
100,000.00
80,000.00
60,000.00
40,000.00
20,000.00
0.00
-20,000.00
SUM(Profit), SUM(Sales)
Jan 2011
Apr 2011
Jul 2011
Oct 2011
Jan 2012
Apr 2012
Jul 2012
Oct 2012
Jan 2013
Apr 2013
Jul 2013
Oct 2013
Jan 2014
Apr 2014
Jul 2014
Oct 2014
MONTH(OrderDate)

CHAPTER

TWO

INTRODUCTION OF METATRON DISCOVERY

Metatron Discovery is a solution that analyzes data ingested into the Metatron server cluster in a simple, sophisticated manner, and visualizes the results in the user PC in the form of charts and reports. A web-based application, it is highly accessible such that it can be remotely accessed by from any PC.

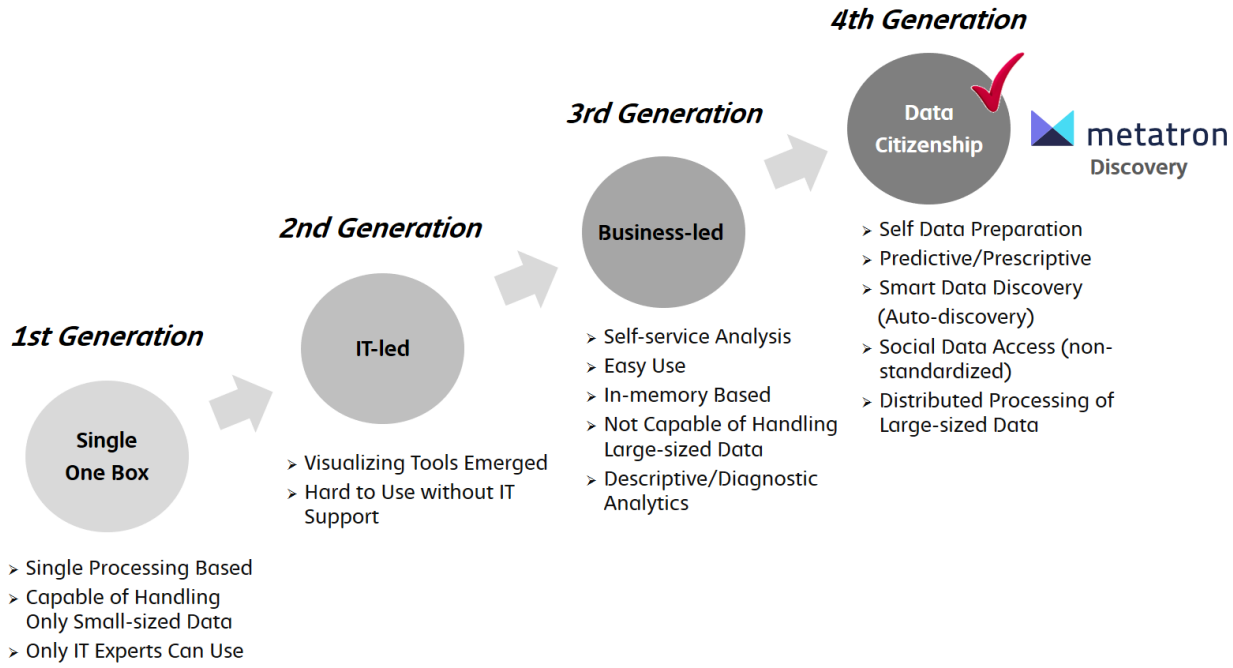
This section introduces the technical background and structure of Metatron Discovery, and the Druid engine powering Metatron.

2.1 Overview of Metatron Discovery

Metatron Discovery is a 4th-generation OLAP-based business intelligence (BI) solution that combines OLAP, visualization, and machine learning technologies for even non-experts to quickly and easily derive higher-level value from data.

2.1.1 4th-generation BI solution

The figure below shows BI trends from the 1st to 4th generation.



The mainstream products in the current BI market belong to the 2nd and 3rd generations, and 4th generation products are beginning to come under the spotlight. As a 4th generation BI solution, Metatron Discovery supports self & ad-hoc data discovery and guarantees rapid response to big data.

2.1.2 Built on Big OLAP

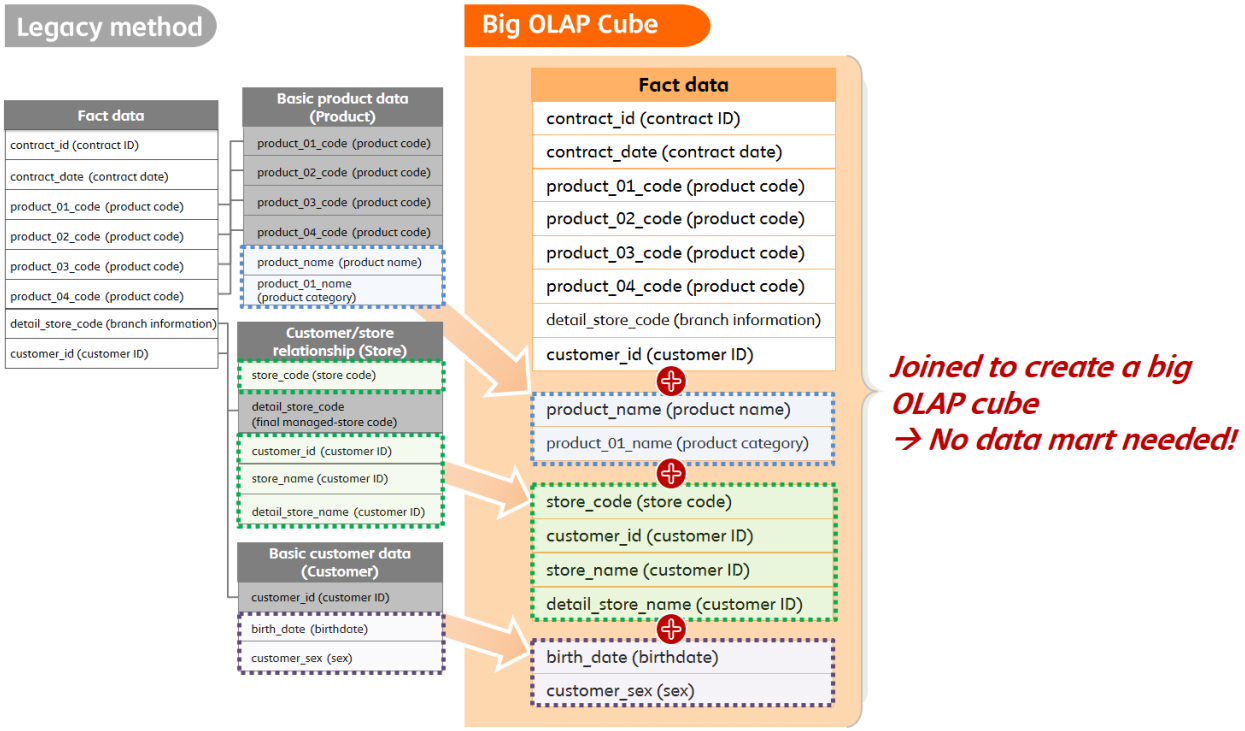
Metatron Discovery combines data of various dimensions for large-sized fact data to produce a single big OLAP cube (data mart).

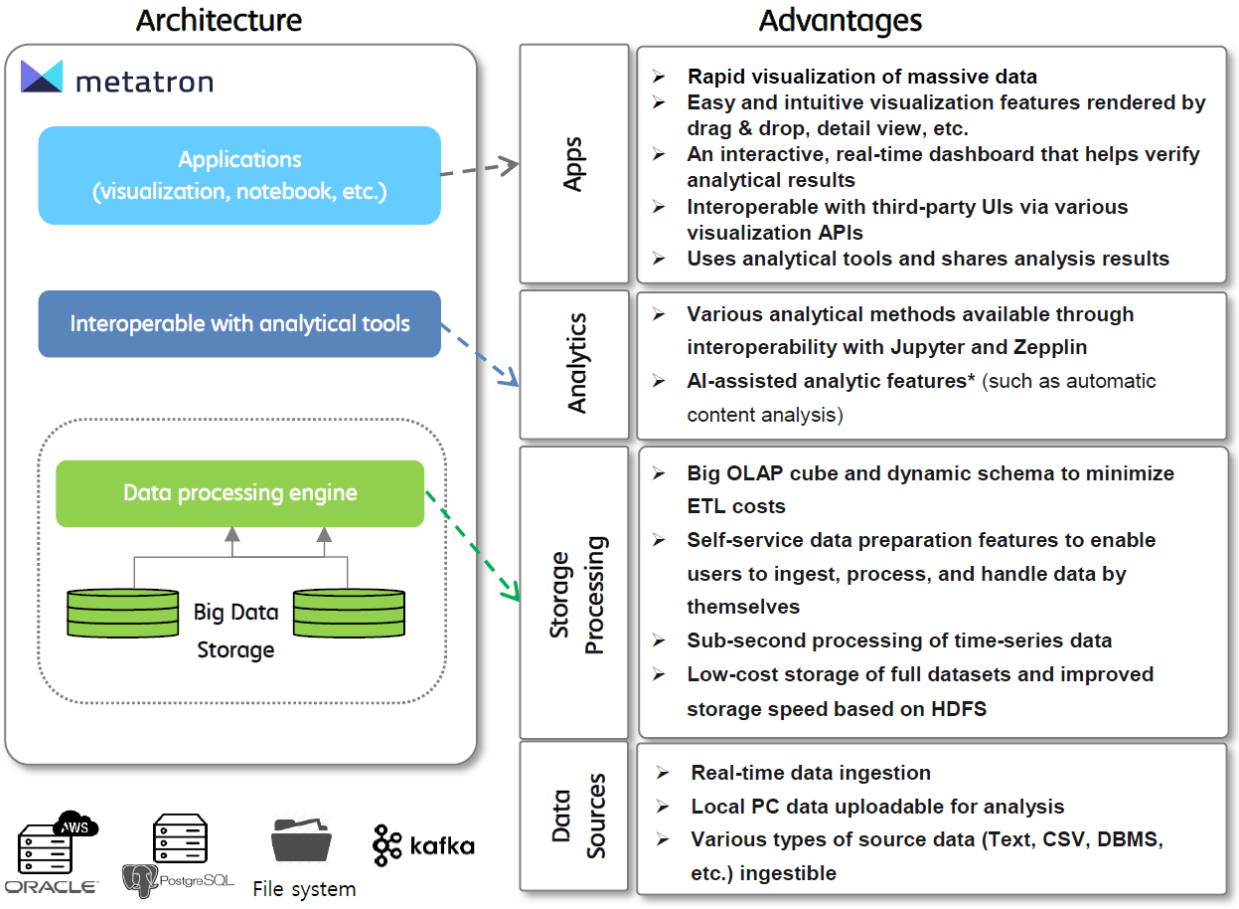
The use of a big OLAP cube offers the following advantages:

- Minimizes the number of data marts.
  - Lower ETL cost for data mart production.
  - Influence of structural change can be minimized.
  - Satisfies diverse demands by saving all fact data.
- Distributed architecture allows storing of large-scale data and ensures fast data processing.
- With a dynamic schema approach, schema changes do not require schema redefining.
- Data can be processed at the record level in real time as tables are saved with no data loss.

2.1.3 Architecture of Metatron Discovery

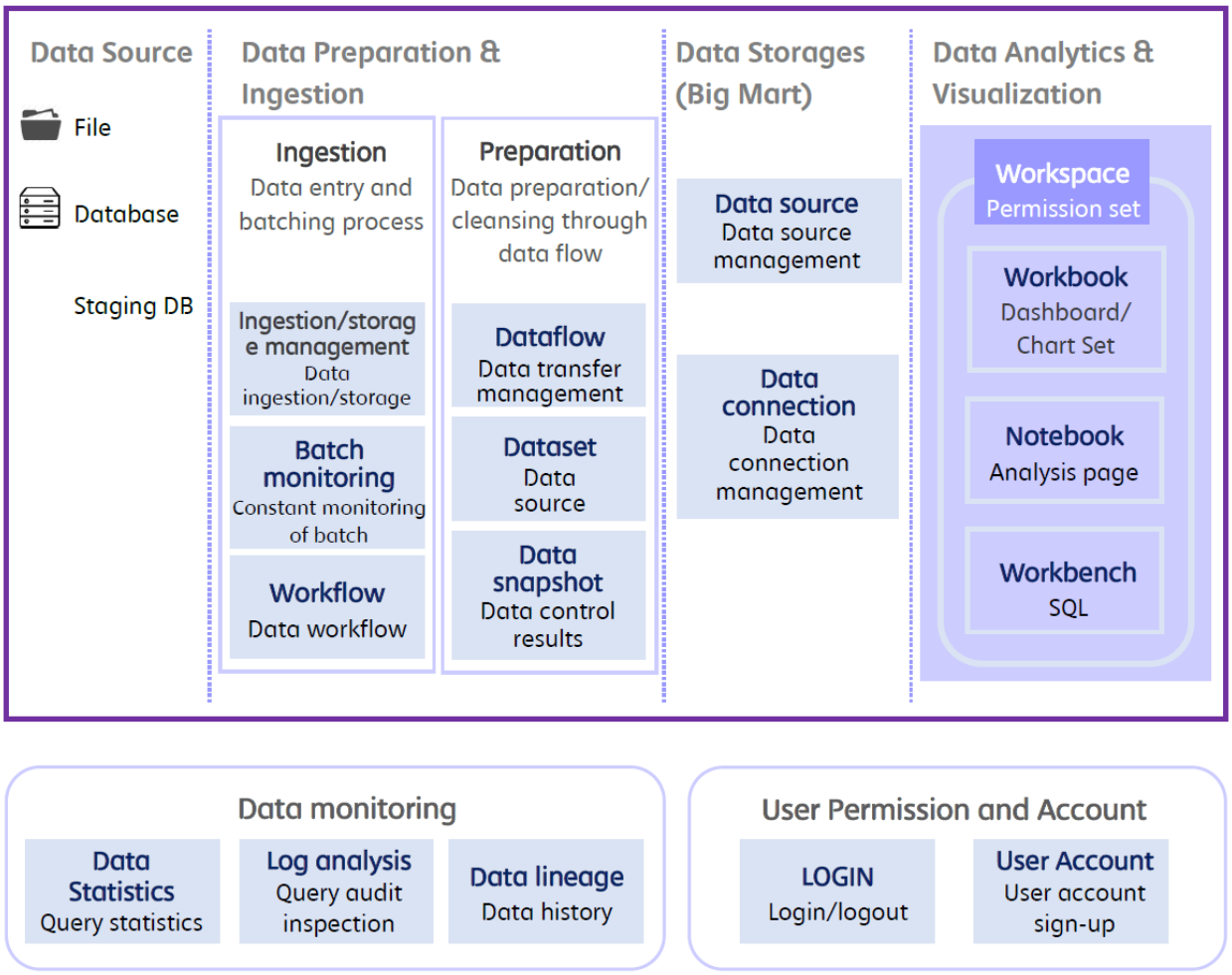
Metatron Discovery is an end-to-end solution that supports the entire process of data discovery, from preparation of large-scale data to data visualization and exploration and to advanced analytics. The figure below is a summary of Metatron's architecture and key features.


Legacy method
Big OLAP Cube
Fact data
contract_id (contract ID)
contract_date (contract date)
product_01_code (product code)
product_02_code (product code)
product_03_code (product code)
product_04_code (product code)
detail_store_code (branch information)
customer_id (customer ID)
Basic product data (Product)
product_01_code (product code)
product_02_code (product code)
product_03_code (product code)
product_04_code (product code)
product_name (product name)
product_01_name (product category)
Customer/store relationship (Store)
store_code (store code)
detail_store_code (final managed-store code)
customer_id (customer ID)
store_name (customer ID)
detail_store_name (customer ID)
Basic customer data (Customer)
customer_id (customer ID)
birth_date (birthdate)
customer_sex (sex)
Joined to create a big OLAP cube
→ No data mart needed!


Architecture
Advantages
metatron
Applications (visualization, notebook, etc.)
Interoperable with analytical tools
Data processing engine
Big Data Storage
ORACLE
PostgreSQL
File system
kafka
Apps
Rapid visualization of massive data
Easy and intuitive visualization features rendered by drag & drop, detail view, etc.
An interactive, real-time dashboard that helps verify analytical results
Interoperable with third-party UIs via various visualization APIs
Uses analytical tools and shares analysis results
Analytics
Various analytical methods available through interoperability with Jupyter and Zepplin
AI-assisted analytic features* (such as automatic content analysis)
Storage Processing
Big OLAP cube and dynamic schema to minimize ETL costs
Self-service data preparation features to enable users to ingest, process, and handle data by themselves
Sub-second processing of time-series data
Low-cost storage of full datasets and improved storage speed based on HDFS
Data Sources
Real-time data ingestion
Local PC data uploadable for analysis
Various types of source data (Text, CSV, DBMS, etc.) ingestible

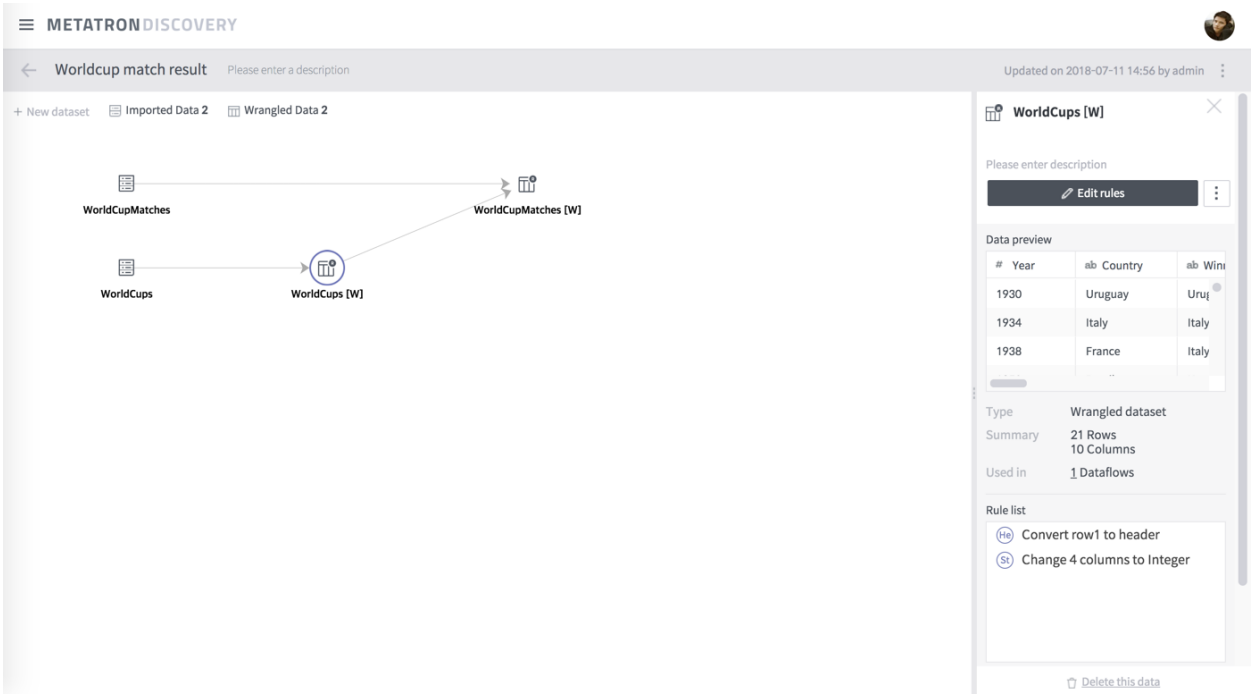
2.2 Components of Metatron Discovery

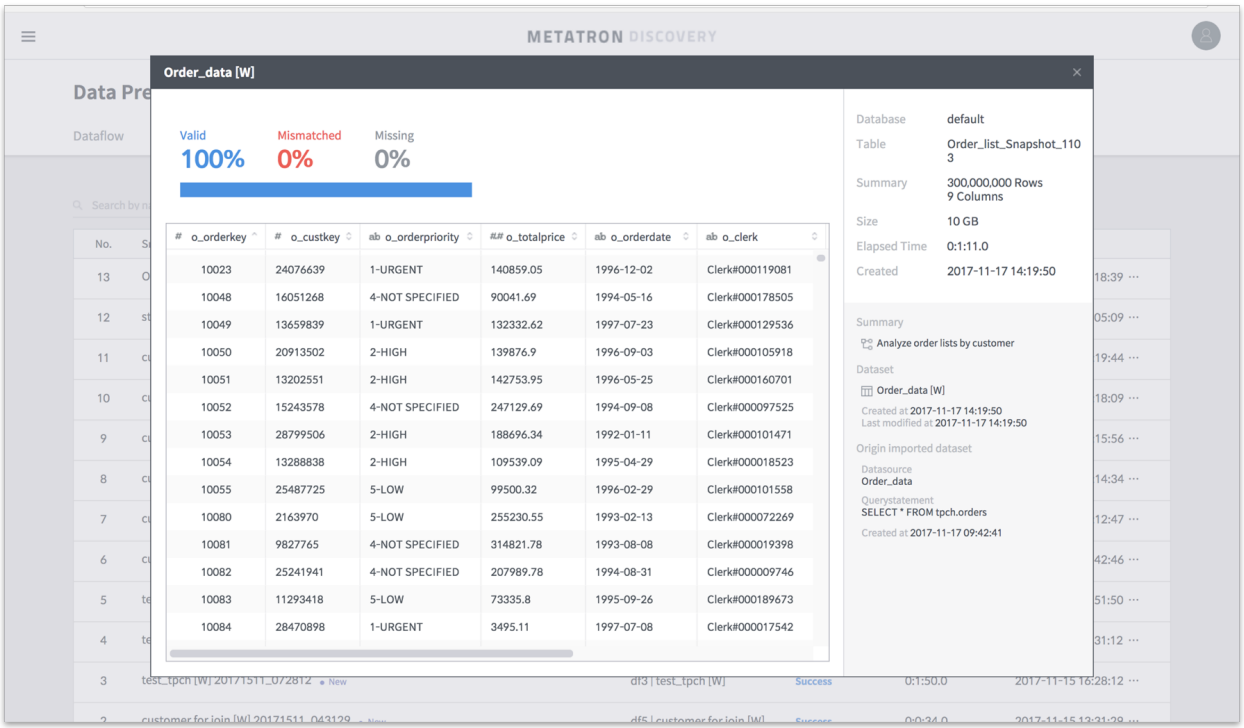
Metatron Discovery performs analytics on its ingested data sources or other external data sources using various analytical tools and outputs analytical results in charts and reports. To utilize this system, you must understand its overall structure shown below:



2.2.1 Data Preparation

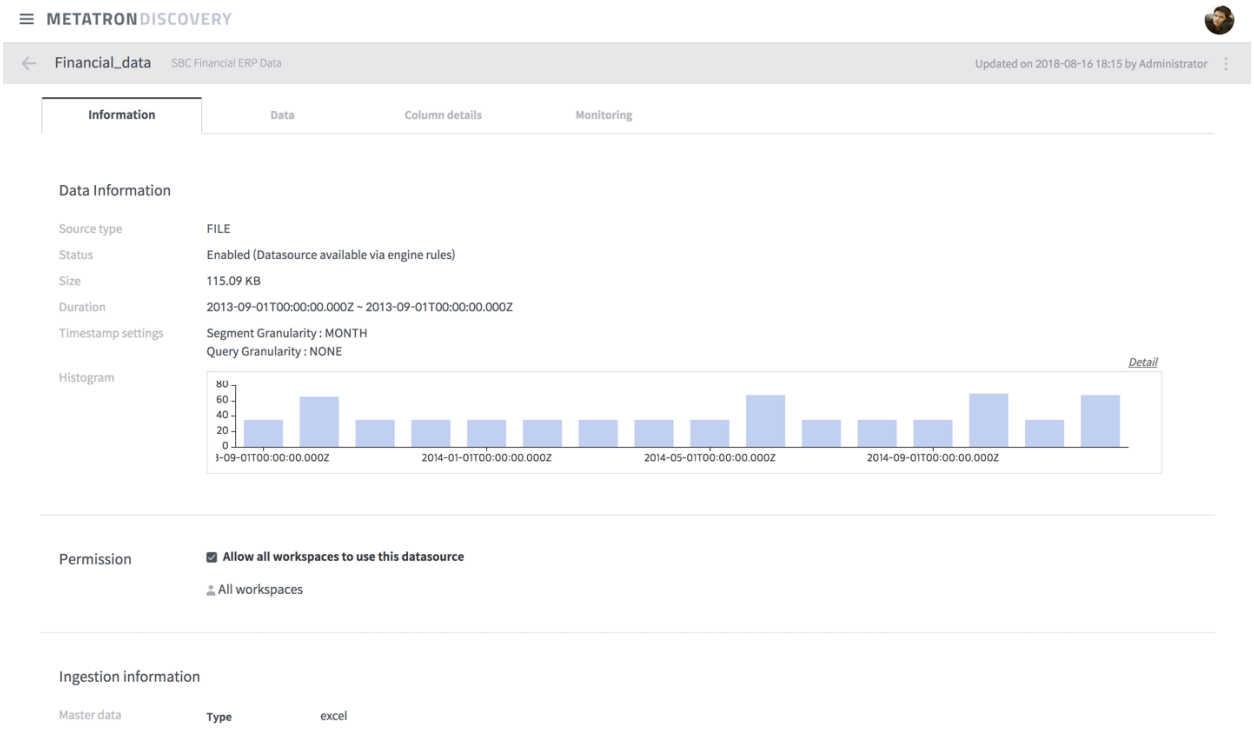
Data Preparation refines data from source data to be ingested into Metatron. See *Data Preparation* for details on data preparation.


METATRON DISCOVERY
Worldcup match result
Please enter a description
Updated on 2018-07-11 14:56 by admin
+ New dataset
Imported Data 2
Wrangled Data 2
WorldCupMatches
WorldCupMatches [W]
WorldCups
WorldCups [W]
WorldCups [W]
Please enter description
Edit rules
Data preview
Year
Country
1930
Uruguay
1934
Italy
1938
France
Type
Wrangled dataset
Summary
21 Rows
10 Columns
Used in
1 Dataflows
Rule list
Convert row1 to header
Change 4 columns to Integer
Delete this data


METATRON DISCOVERY
Order_data [W]
Valid
100%
Mismatched
0%
Missing
0%
o_orderkey
o_custkey
o_orderpriority
o_totalprice
o_orderdate
o_clerk
10023 24076639 1-URGENT 140859.05 1996-12-02 Clerk#000119081
10048 16051268 4-NOT SPECIFIED 90041.69 1994-05-16 Clerk#000178505
10049 13659839 1-URGENT 132332.62 1997-07-23 Clerk#000129536
10050 20913502 2-HIGH 139876.9 1996-09-03 Clerk#000105918
10051 13202551 2-HIGH 142753.95 1996-05-25 Clerk#000160701
10052 15243578 4-NOT SPECIFIED 247129.69 1994-09-08 Clerk#000097525
10053 28799506 2-HIGH 188696.34 1992-01-11 Clerk#000101471
10054 13288838 2-HIGH 109539.09 1995-04-29 Clerk#000018523
10055 25487725 5-LOW 99500.32 1996-02-29 Clerk#000101558
10080 2163970 5-LOW 255230.55 1993-02-13 Clerk#000072269
10081 9827765 4-NOT SPECIFIED 314821.78 1993-08-08 Clerk#000019398
10082 25241941 4-NOT SPECIFIED 207989.78 1994-08-31 Clerk#000009746
10083 11293418 5-LOW 73335.8 1995-09-26 Clerk#000189673
10084 28470898 1-URGENT 3495.11 1997-07-08 Clerk#000017542
Database default
Table Order_list_Snapshot_1103
Summary 300,000,000 Rows 9 Columns
Size 10 GB
Elapsed Time 0:1:11.0
Created 2017-11-17 14:19:50
Summary
Analyze order lists by customer
Dataset
Order_data [W]
Created at 2017-11-17 14:19:50
Last modified at 2017-11-17 14:19:50
Origin imported dataset
Datasource
Order_data
Querystatement
SELECT * FROM tpch.orders
Created at 2017-11-17 09:42:41

2.2.2 Data Storage

Data Storage manages data ingested into the Metatron engine for analysis and visualization. See *Data Management* for details on data management.



2.2.3 Data analysis and visualization

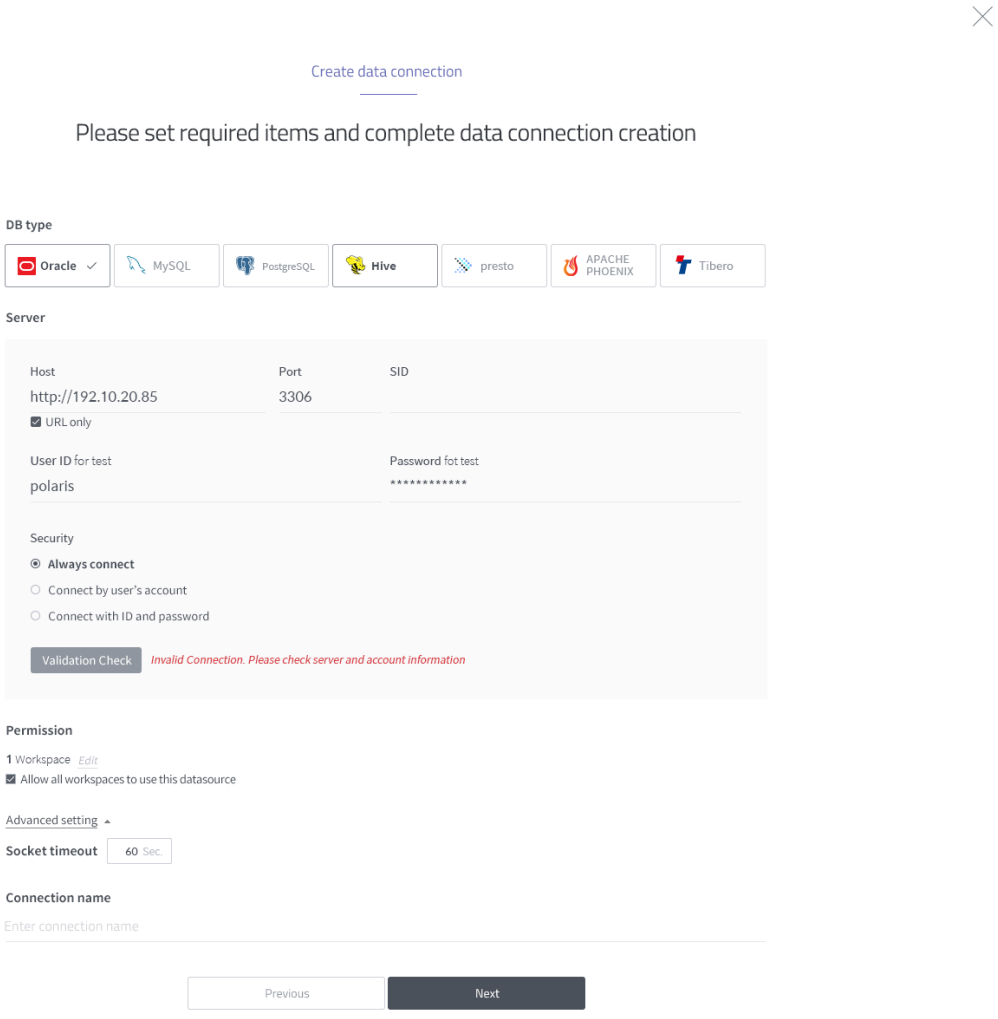
Each module below allows users to perform visualization-based exploration and analysis of stored data.

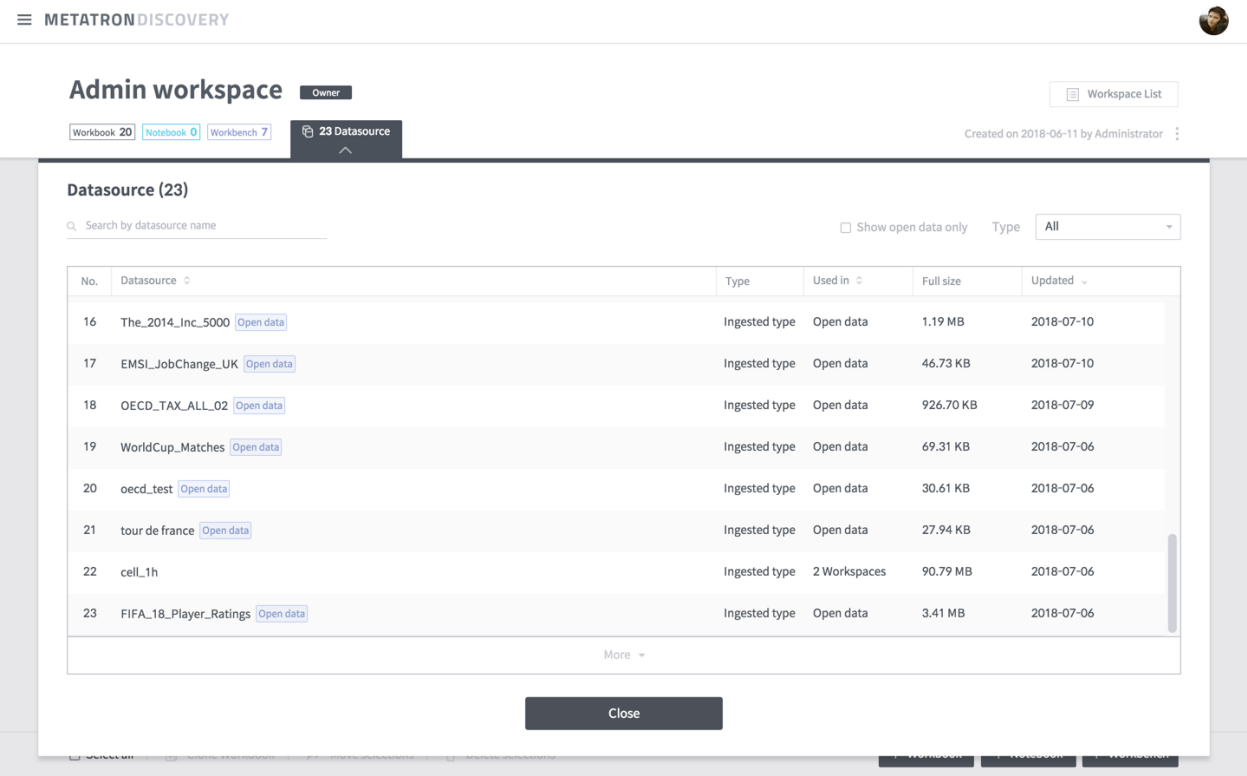
Workspace

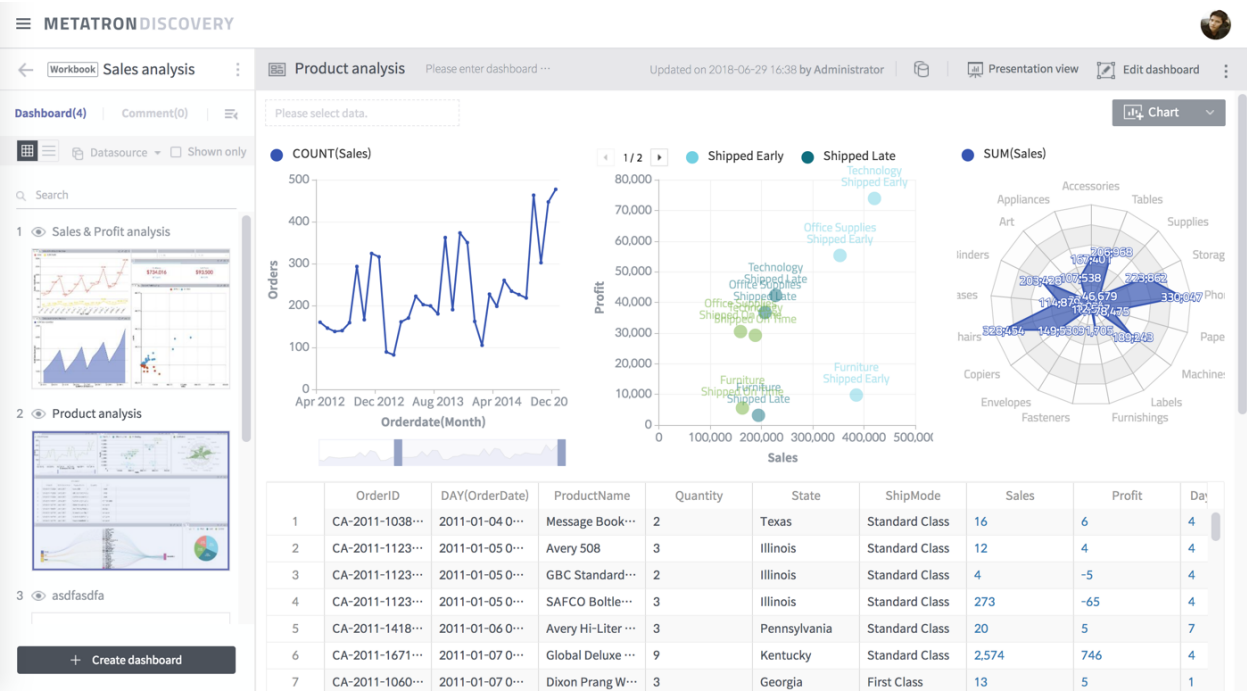
Workspace provides an interface to manage its workbooks, workbenches, and notebooks used in an organization according to user access. See *Workspace* for details on the use of the workspace.

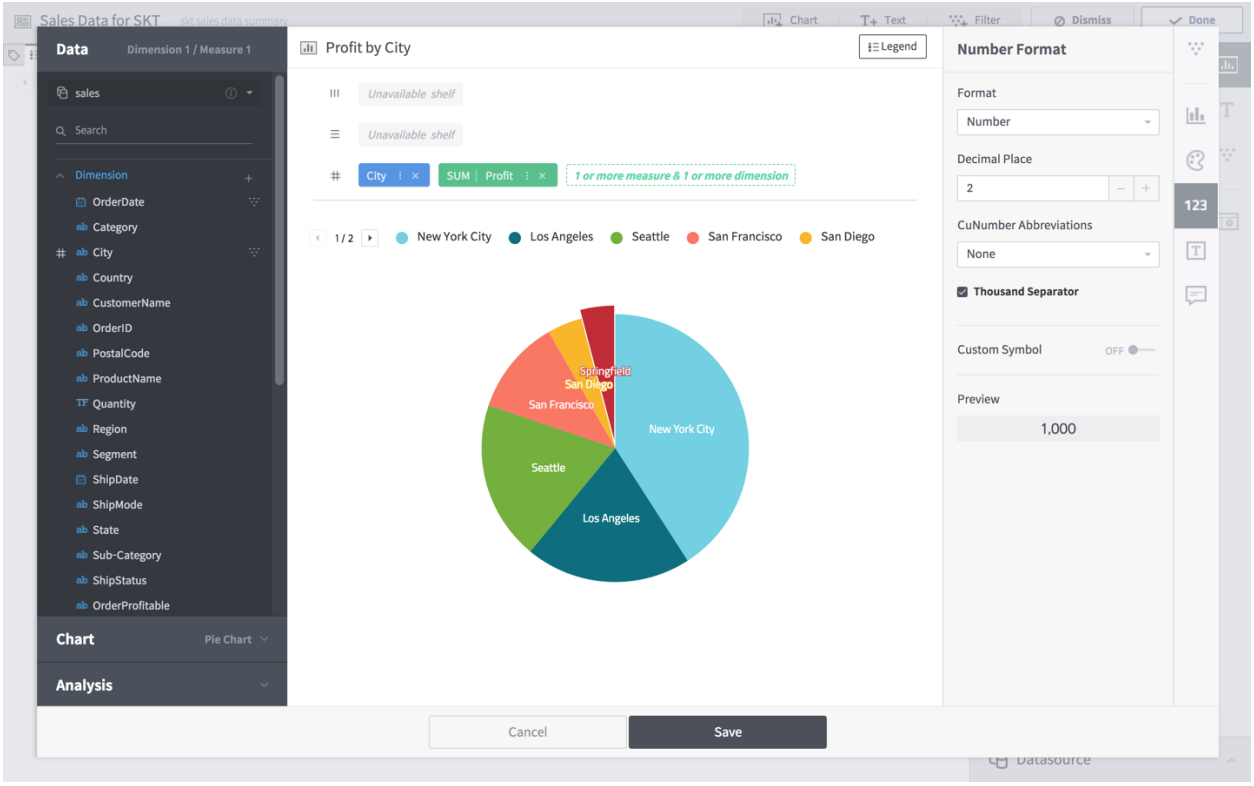
Workbook, dashboard, chart

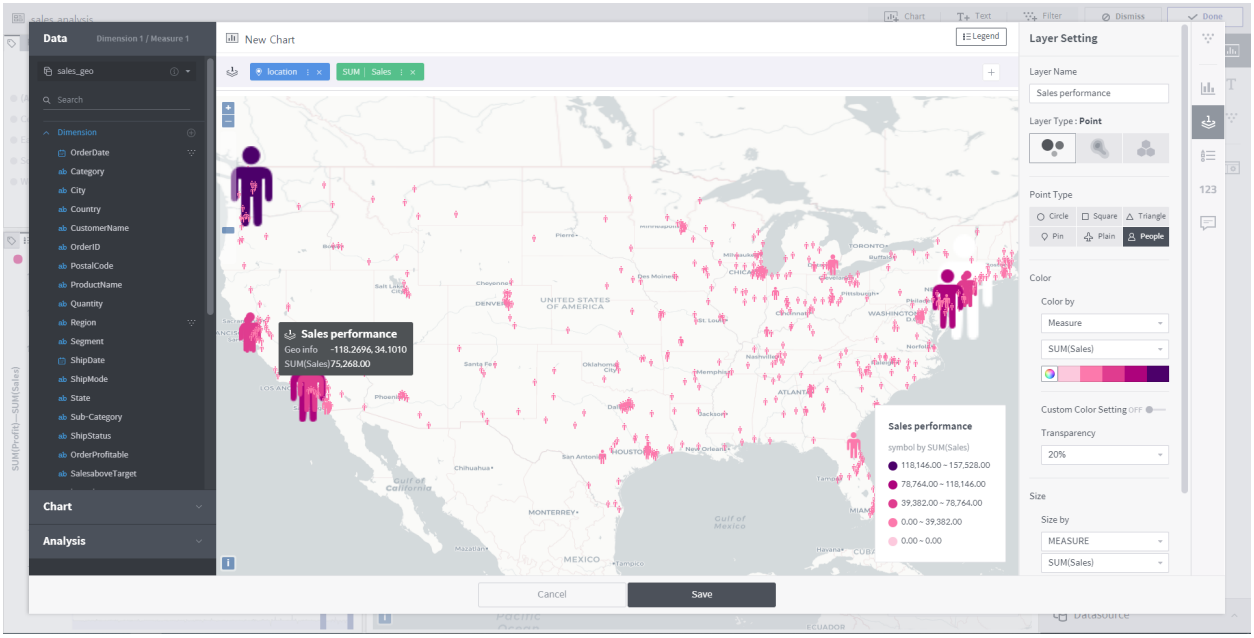
Workbook supports working on, sharing, and making a presentation with dashboards and charts using a PowerPoint-like interface. See *Workbook* for details on the workbook module.


Create data connection
Please set required items and complete data connection creation
DB type
Oracle
MySQL
PostgreSQL
Hive
presto
APACHE PHOENIX
Tibero
Server
Host
http://192.10.20.85
Port
3306
SID
URL only
User ID for test
polaris
Password for test
Security
Always connect
Connect by user's account
Connect with ID and password
Validation Check
Invalid Connection. Please check server and account information
Permission
1 Workspace Edit
Allow all workspaces to use this datasource
Advanced setting
Socket timeout
60 Sec.
Connection name
Enter connection name
Previous
Next


METATRONDISCOVERY
Admin workspace
Owner
Workbook 20
Notebook 0
Workbench 7
23 Datasource
Workspace List
Created on 2018-06-11 by Administrator
Datasource (23)
Search by datasource name
Show open data only
Type
All
No.
Datasource
Type
Used in
Full size
Updated
16 The_2014_Inc_5000 Open data Ingested type Open data 1.19 MB 2018-07-10
17 EMSI_JobChange_UK Open data Ingested type Open data 46.73 KB 2018-07-10
18 OECD_TAX_ALL_02 Open data Ingested type Open data 926.70 KB 2018-07-09
19 WorldCup_Matches Open data Ingested type Open data 69.31 KB 2018-07-06
20 oecd_test Open data Ingested type Open data 30.61 KB 2018-07-06
21 tour de france Open data Ingested type Open data 27.94 KB 2018-07-06
22 cell_1h Ingested type 2 Workspaces 90.79 MB 2018-07-06
23 FIFA_18_Player_Ratings Open data Ingested type Open data 3.41 MB 2018-07-06
More
Close

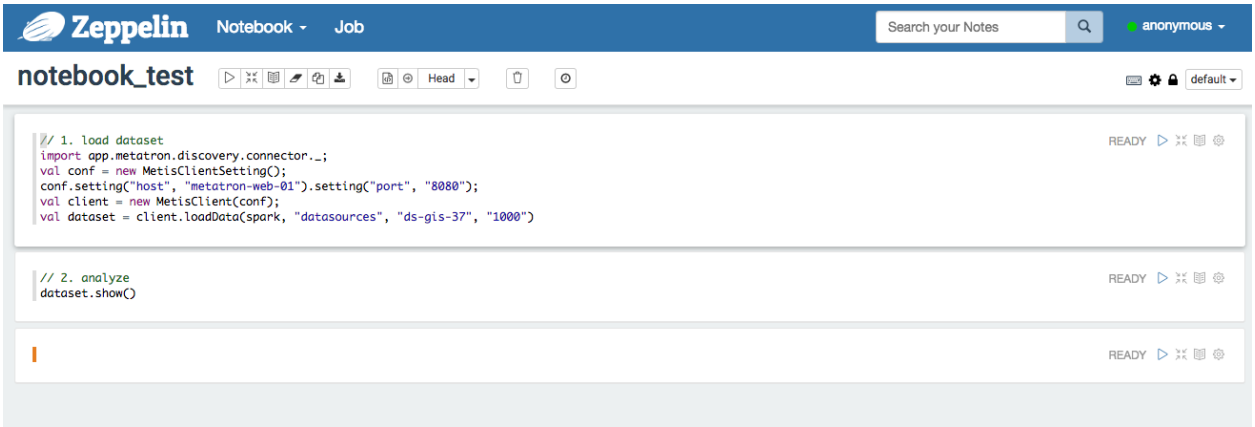

METATRONDISCOVERY
Workbook
Sales analysis
Product analysis
Please enter dashboard
Updated on 2018-06-29 16:38 by Administrator
Presentation view
Edit dashboard
Dashboard(4)
Comment(0)
Please select data.
Chart
Datasource
Shown only
Search
1 Sales & Profit analysis
2 Product analysis
3 asdfasdfa
Create dashboard
COUNT(Sales)
Orders
Orderdate(Month)
Shipped Early
Shipped Late
Profit
Sales
SUM(Sales)
OrderID
DAY(OrderDate)
ProductName
Quantity
State
ShipMode
Sales
Profit
1 CA-2011-1038··· 2011-01-04 0··· Message Book··· 2 Texas Standard Class 16 6 4
2 CA-2011-1123··· 2011-01-05 0··· Avery 508 3 Illinois Standard Class 12 4 4
3 CA-2011-1123··· 2011-01-05 0··· GBC Standard··· 2 Illinois Standard Class 4 -5 4
4 CA-2011-1123··· 2011-01-05 0··· SAFCO Boltle··· 3 Illinois Standard Class 273 -65 4
5 CA-2011-1418··· 2011-01-06 0··· Avery Hi-Liter ··· 3 Pennsylvania Standard Class 20 5 7
6 CA-2011-1671··· 2011-01-07 0··· Global Deluxe ··· 9 Kentucky Standard Class 2,574 746 4
7 CA-2011-1060··· 2011-01-07 0··· Dixon Prang W··· 3 Georgia First Class 13 5 1


Sales Data for SKT
Chart
Text
Filter
Dismiss
Done
Data
Dimension 1 / Measure 1
sales
Search
Dimension
OrderDate
Category
City
Country
CustomerName
OrderID
PostalCode
ProductName
Quantity
Region
Segment
ShipDate
ShipMode
State
Sub-Category
ShipStatus
OrderProfitable
Chart
Pie Chart
Analysis
Profit by City
Legend
Unavailable shelf
Unavailable shelf
City
SUM | Profit
1 or more measure & 1 or more dimension
1 / 2
New York City
Los Angeles
Seattle
San Francisco
San Diego
Springfield
San Diego
San Francisco
New York City
Seattle
Los Angeles
Number Format
Format
Number
Decimal Place
2
123
CuNumber Abbreviations
None
Thousand Separator
Custom Symbol
OFF
Preview
1,000
Cancel
Save
Datasource


Data
Dimension 1 / Measure 1
sales_geo
Search
Dimension
OrderDate
Category
City
Country
CustomerName
OrderID
PostalCode
ProductName
Quantity
Region
Segment
ShipDate
ShipMode
State
Sub-Category
ShipStatus
OrderProfitable
SalesaboveTarget
Chart
Analysis
New Chart
Legend
location
SUM | Sales
Sales performance
Geo info -118.2696, 34.1010
SUM(Sales)75,268.00
Sales performance
symbol by SUM(Sales)
118,146.00 ~ 157,528.00
78,764.00 ~ 118,146.00
39,382.00 ~ 78,764.00
0.00 ~ 39,382.00
0.00 ~ 0.00
Layer Setting
Layer Name
Sales performance
Layer Type : Point
Point Type
Circle
Square
Triangle
Pin
Plain
People
Color
Color by
Measure
SUM(Sales)
Custom Color Setting
Transparency
20%
Size
Size by
MEASURE
SUM(Sales)
Cancel
Save

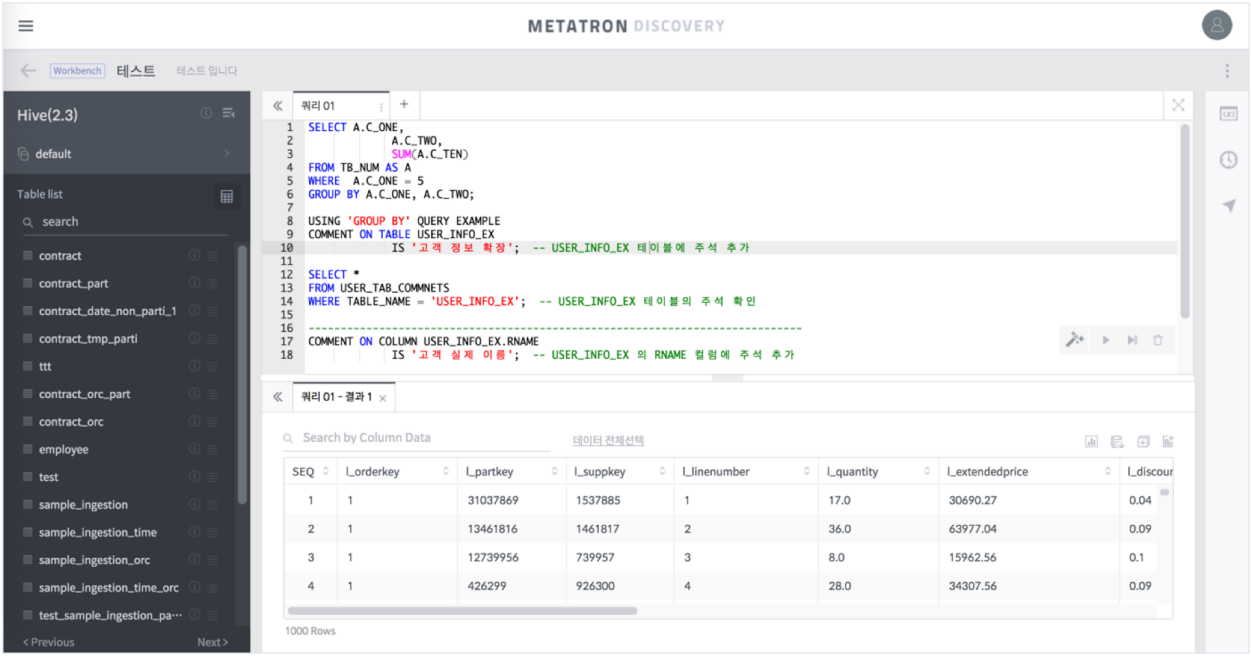
Notebook

Notebook enables advanced analytics based on machine learning. See *Notebook* for details on the notebook module.



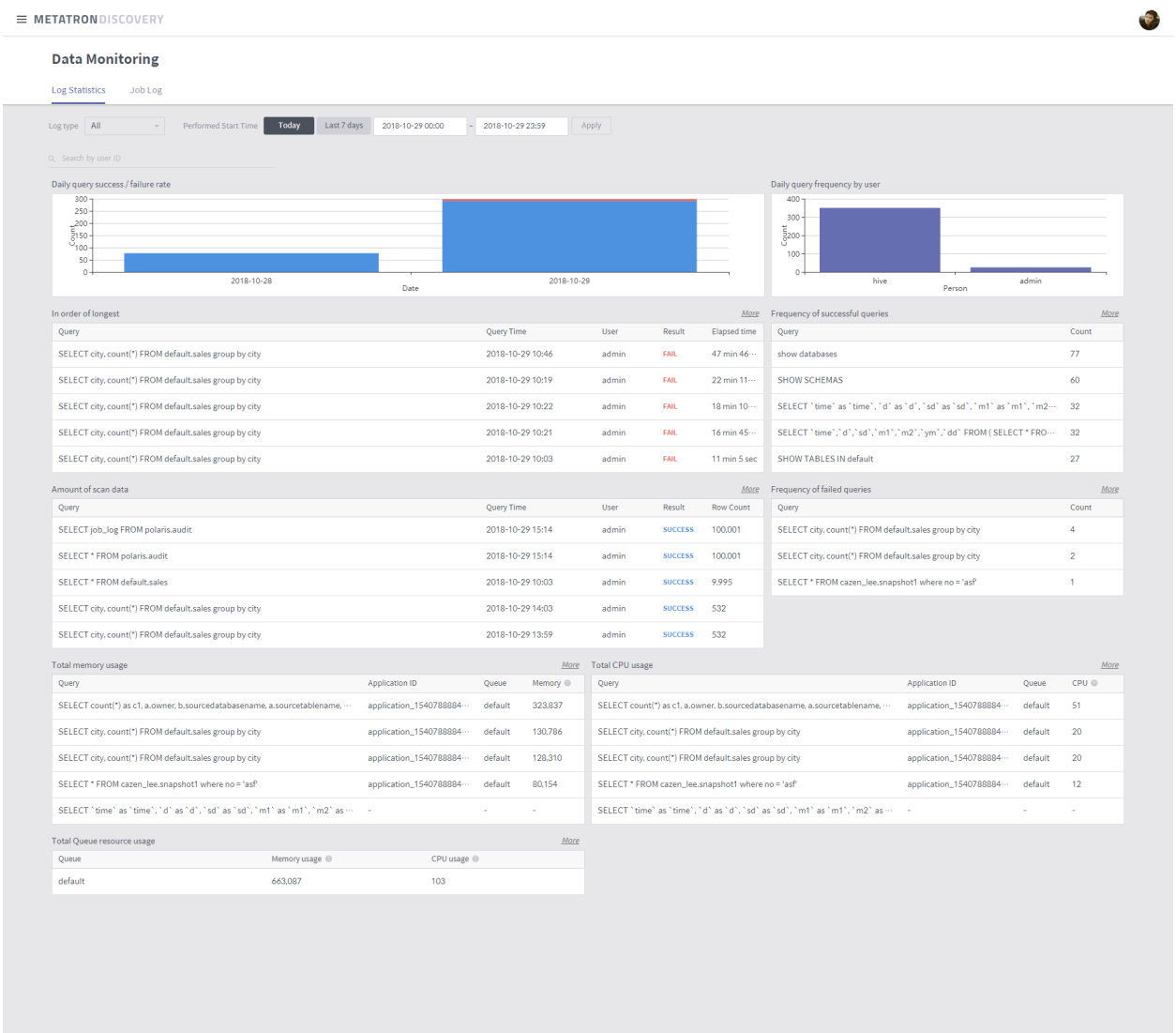
Workbench

Workbench enables SQL data analytics. See *Workbench* for details on the workbench module.



2.2.4 Data Monitoring

This function monitors data use based on data query statistics and query logs. See *Data Monitoring* for details on the data monitoring functionality.


METATRONDISCOVERY
Data Monitoring
Log Statistics
Job Log
Log type
All
Performed Start Time
Today
Last 7 days
2018-10-29 00:00
2018-10-29 23:59
Apply
Search by user ID
Daily query success / failure rate
Count
Date
2018-10-28
2018-10-29
Daily query frequency by user
Count
Person
hive
admin
In order of longest
More
Query
Query Time
User
Result
Elapsed time
SELECT city, count(*) FROM default.sales group by city
2018-10-29 10:46
admin
FAIL
47 min 46···
SELECT city, count(*) FROM default.sales group by city
2018-10-29 10:19
admin
FAIL
22 min 11···
SELECT city, count(*) FROM default.sales group by city
2018-10-29 10:22
admin
FAIL
18 min 10···
SELECT city, count(*) FROM default.sales group by city
2018-10-29 10:21
admin
FAIL
16 min 45···
SELECT city, count(*) FROM default.sales group by city
2018-10-29 10:03
admin
FAIL
11 min 5 sec
Frequency of successful queries
More
Query
Count
show databases
77
SHOW SCHEMAS
60
SELECT `time` as `time`, `d` as `d`, `sd` as `sd`, `m1` as `m1`, `m2`···
32
SELECT `time`, `d`, `sd`, `m1`, `m2`, `ym`, `dd` FROM (SELECT * FRO···
32
SHOW TABLES IN default
27
Amount of scan data
More
Query
Query Time
User
Result
Row Count
SELECT job_log FROM polaris.audit
2018-10-29 15:14
admin
SUCCESS
100,001
SELECT * FROM polaris.audit
2018-10-29 15:14
admin
SUCCESS
100,001
SELECT * FROM default.sales
2018-10-29 10:03
admin
SUCCESS
9,995
SELECT city, count(*) FROM default.sales group by city
2018-10-29 14:03
admin
SUCCESS
532
SELECT city, count(*) FROM default.sales group by city
2018-10-29 13:59
admin
SUCCESS
532
Frequency of failed queries
More
Query
Count
SELECT city, count(*) FROM default.sales group by city
4
SELECT city, count(*) FROM default.sales group by city
2
SELECT * FROM cazen_lee.snapshot1 where no = 'asf'
1
Total memory usage
More
Query
Application ID
Queue
Memory
SELECT count(*) as c1, a.owner, b.sourcedatabasename, a.sourcetablename, ···
application_1540788884···
default
323,837
SELECT city, count(*) FROM default.sales group by city
application_1540788884···
default
130,786
SELECT city, count(*) FROM default.sales group by city
application_1540788884···
default
128,310
SELECT * FROM cazen_lee.snapshot1 where no = 'asf'
application_1540788884···
default
80,154
SELECT `time` as `time`, `d` as `d`, `sd` as `sd`, `m1` as `m1`, `m2` as ···
-
-
-
Total CPU usage
More
Query
Application ID
Queue
CPU
SELECT count(*) as c1, a.owner, b.sourcedatabasename, a.sourcetablename, ···
application_1540788884···
default
51
SELECT city, count(*) FROM default.sales group by city
application_1540788884···
default
20
SELECT city, count(*) FROM default.sales group by city
application_1540788884···
default
20
SELECT * FROM cazen_lee.snapshot1 where no = 'asf'
application_1540788884···
default
12
SELECT `time` as `time`, `d` as `d`, `sd` as `sd`, `m1` as `m1`, `m2` as ···
-
-
-
Total Queue resource usage
More
Queue
Memory usage
CPU usage
default
663,087
103

2.2.5 User permission and account management

You can add/delete users or manage user permission.

2.3 Metatron engine: Druid

The development of information and communications technology has been accompanied by a rapid increase in the amount of data generated, highlighting the importance of efficient data collection, management, and utilization. However, RDBMS-based legacy tools are unable to process mass amounts of multidimensional data. This has led to the emergence of new methodologies and solutions aimed at satisfying the demand for big data.

Metamarkets, a technology startup based in Silicon Valley, launched a column-oriented distributed data store known as Druid in 2011, and open sourced it in October 2012. Many companies have turned to Druid for their backend technology because it offers various advantages, including fast and efficient data processing.

As a B2C telecommunications service provider, SK Telecom recognized the need to effectively manage and analyze the vast amounts of network data generated by its users every minute. Metatron, an end-to-end business intelligence solution with Druid as the underlying engine, was thus developed and launched in 2016.

The following sections discuss the features of Druid that make it suitable for time-series data processing, and introduce how they were adapted and improved by SK Telecom for Metatron.

2.3.1 Background of Druid development

Druid was originally designed to satisfy the following needs around ingesting and exploring large quantities of transactional events (log data):

- The developers wanted to be able to rapidly and arbitrarily slice and dice data and drill into that data effectively without any restrictions, along with sub-second queries over any arbitrary combination of dimensions. These capabilities were needed to allow users of their data dashboard to arbitrarily and interactively explore and visualize event streams.
- The developers wanted to be able to ingest events and make them exportable almost immediately after their occurrence. This was crucial to enable users to collect and analyze data in real time for timely situational assessments, predictions and business decisions. Popular open source data warehousing systems such as Hadoop were unable to provide the sub-second data ingestion latencies as required.
- The developers wanted to ensure multitenancy and high availability for their solution services. Their systems needed to be constantly up and be able to withstand all sorts of potential failures without going down or taking any downtime. Downtime is costly and many businesses cannot afford to wait if a system is unavailable in the face of software upgrades or network failure.

2.3.2 Druid features

Data table components

Data tables in Druid (called data sources) are collections of timestamped events designed for OLAP queries. A data source is composed of three distinct types of columns (here we use an example dataset from online advertising).

Timestamp column	Dimension columns				Metric columns	
timestamp	publisher	advertiser	gender	country	click	price
2011-01-01T01:01:35Z	bieberfever.com	google.com	Male	USA	0	0.65
2011-01-01T01:03:63Z	bieberfever.com	google.com	Male	USA	0	0.62
2011-01-01T01:04:51Z	bieberfever.com	google.com	Male	USA	1	0.45
2011-01-01T01:00:00Z	ultratrimfast.com	google.com	Female	UK	0	0.87
2011-01-01T02:00:00Z	ultratrimfast.com	google.com	Female	UK	0	0.99
2011-01-01T02:00:00Z	ultratrimfast.com	google.com	Female	UK	1	1.53

Fig. 1: Source: http://druid.io

- **Timestamp column:** Druid treats timestamp separately in a data source because all its queries center around the time axis (If non-time series data is ingested in batch, all records are timestamped with the current time for use in Druid).
- **Dimension columns:** Dimensions are string attributes of an event, and the columns most commonly used in filtering the data. Four dimensions are involved in the example dataset: publisher, advertiser, gender, and country. They each represent an axis of the data chosen to slice across.
- **Metric columns:** Metrics are columns used in aggregations and computations. In the example, the metrics are clicks and price. Metrics are usually numeric values, and computations include operations such as count, sum, and mean (Metatron has extended supported Druid data types).

Data ingestion

Druid supports real-time and batch ingestion.

One major characteristic of Druid is real-time ingestion, which is enabled by real-time nodes (For details, see *Real-time nodes*). Events ingested in real-time from a data stream get indexed in seconds to become queryable in the Druid cluster.

Data roll-up

The individual events in our example dataset are not very interesting because there may be trillions of such events. However, summarizations of this type of data by time interval can yield many useful insights. Druid summarizes this raw data when ingesting it using an optional process called "roll-up." Below is an example of roll-up.

timestamp	domain	gender	clicked
2011-01-01T00:01:35Z	bieber.com	Female	1
2011-01-01T00:03:03Z	bieber.com	Female	0
2011-01-01T00:04:51Z	ultra.com	Male	1
2011-01-01T00:05:33Z	ultra.com	Male	1
2011-01-01T00:05:53Z	ultra.com	Female	0
2011-01-01T00:06:17Z	ultra.com	Female	1
2011-01-01T00:23:15Z	bieber.com	Female	0
2011-01-01T00:38:51Z	ultra.com	Male	1
2011-01-01T00:49:33Z	bieber.com	Female	1
2011-01-01T00:49:53Z	ultra.com	Female	0

➔

timestamp	domain	gender	clicked
2011-01-01T00:00:00Z	bieber.com	Female	1
2011-01-01T00:00:00Z	ultra.com	Female	2
2011-01-01T00:00:00Z	ultra.com	Male	3

Fig. 2: Source: Interactive Exploratory Analytics with Druid | DataEngConf SF '17

The table on the left lists the domain click events that occurred from 00:00:00 to 01:00:00 on January 1, 2011. Since individual events recorded in seconds do not have much significance from the analyst's perspective, the data was compiled at a granularity of one hour. This results in the more meaningful table on the right, which shows the number of clicks by gender for the same time period.

In practice, rolling up data can dramatically reduce the size of data that needs to be stored (up to a factor of 100), thereby saving on storage resources and enabling faster queries.

But, as data is rolled up, individual events can no longer be queried; the rollup granularity is the minimum granularity you will be able to explore data at and events are floored to this granularity. The unit of granularity can be set as desired by users. If necessary, the roll-up process may be disabled to ingest every individual event.

Data sharding

A data source is a collection of timestamped events and partitioned into a set of shards. A shard is called a segment in Druid and each segment is typically 5?10 million rows. Druid partitions its data sources into well-defined time intervals, typically an hour or a day, and may further partition on values from other columns to achieve the desired segment size.

The example below shows a data table segmented by hour:

Segment `sampleData_2011-01-01T01:00:00Z_2011-01-01T02:00:00Z_v1_0`:

```
2011-01-01T01:00:00Z   ultratrimfast.com   google.com   Male   USA   1800   25   15.70
2011-01-01T01:00:00Z   bieberfever.com     google.com   Male   USA   2912   42   29.18
```

Segment `sampleData_2011-01-01T02:00:00Z_2011-01-01T03:00:00Z_v1_0`:

```
2011-01-01T02:00:00Z   ultratrimfast.com   google.com   Male   UK   1953   17   17.31
2011-01-01T02:00:00Z   bieberfever.com     google.com   Male   UK   3194   170  34.01
```

This segmentation by time can be achieved because every single event in a data source is timestamped.

Segments represent the fundamental storage unit in Druid and replication and distribution are done at a segment level. They are designed to be immutable, which means that once a segment is created, it cannot be edited. This ensures no contention between reads and writes. Druid segments are just designed to be read very fast.

In addition, this data segmentation is key to parallel processing in Druid's distributed environment: As one CPU can scan one segment at a time, data partitioned into multiple segments can be scanned by multiple CPUs simultaneously in parallel, thereby ensuring fast query returns and stable load balancing.

Data storage format and indexing

The way Druid stores data contributes to its data structures highly optimized for analytic queries. This section uses the Druid table below as an example:

Timestamp	Page	Username	Gender	City	Characters Added	Characters Removed
2011-01-01T01:00:00Z	Justin Bieber	Boxer	Male	San Francisco	1800	25
2011-01-01T01:00:00Z	Justin Bieber	Reach	Male	Waterloo	2912	42
2011-01-01T02:00:00Z	Ke$ha	Helz	Male	Calgary	1953	17
2011-01-01T02:00:00Z	Ke$ha	Xeno	Male	Taiyuan	3194	170

Fig. 3: Source: Druid: A Real-time Analytical Data Store

Columnar storage and indexing

Druid is a column store, which means each individual column is stored separately. Given that Druid is best used for aggregating event streams, column storage allows for more efficient CPU usage as only the columns pertaining to a query are actually loaded and scanned in that query. In a row oriented data store, all columns associated with a row must be scanned as part of an aggregation. The additional scan time can introduce significant performance degradations. In the example above, the page, user, gender, and city columns only contain strings. Storing strings directly is unnecessarily costly; instead, they can be mapped into unique integer identifiers. For example,

```
Justin Bieber -> 0
Ke$ha -> 1
```

This mapping allows the page column to be represented as an integer array where the array indices correspond to the rows of the original dataset. For the page column, we can represent the unique pages as follows:

```
[0, 0, 1, 1]
```

Thus, strings are replaced by fixed-length integers in storage, which are much easier to compress. Druid indexes data on a per-shard (segment) level.

Indices for filtering data

Druid creates additional lookup indices that facilitate filtering on string columns. Let us consider the above example table again. A query might be: "How many Wikipedia edits were done by users in San Francisco who are also male?" This example query involves two dimensions: City (San Francisco) and Gender (Male). For each dimension, a binary

array is created where the array indices represent whether or not their corresponding rows match the query filter, as shown below:

```
San Francisco (City) -> rows [1] -> [1][0][0][0]
Male (Gender) -> rows [1, 2, 3, 4] -> [1][1][1][1]
```

And the query filter performs the AND operation between the two arrays:

```
[1][0][0][0] AND [1][1][1][1] = [1][0][0][0]
```

As a result, only row 1 is subject to scanning, which retrieves only the filtered rows and eliminates unnecessary workload. And these binary arrays are very easy to compress as well.

This lookup can be used for the OR operation too. If a query filters on San Francisco or Calgary, array indices will be for each dimension value:

```
San Francisco (City) -> rows [1] -> [1][0][0][0]
Calgary (City) -> rows [3] -> [0][0][1][0]
```

And then the OR operation is performed on the two arrays:

```
[1][0][0][0] OR [0][0][1][0] = [1][0][1][0]
```

Thus the query scans rows 1 and 3 only.

This approach of performing Boolean operations on large bitmap sets is commonly used in search engines.

Query languages

Druid's native query language is JSON over HTTP. Druid queries include:

- Group By
- Time-series roll-ups
- Arbitrary Boolean filters
- Sum, Min, Max, Avg and other aggregation functions
- Dimensional Search

In addition to these, query libraries in numerous languages, including SQL, are developed and shared.

2.3.3 Druid cluster architecture

A Druid cluster consists of different types of nodes and each node type is designed to perform a specific set of things:

Real-time nodes

Real-time nodes function to ingest and query event streams. The nodes are only concerned with events for some small time range and periodically hand them off to the deep storage in the following steps:

1. Incoming events are indexed in memory and immediately become available for querying.
2. The in-memory data is regularly persisted to disk and converted into an immutable, columnar storage format.

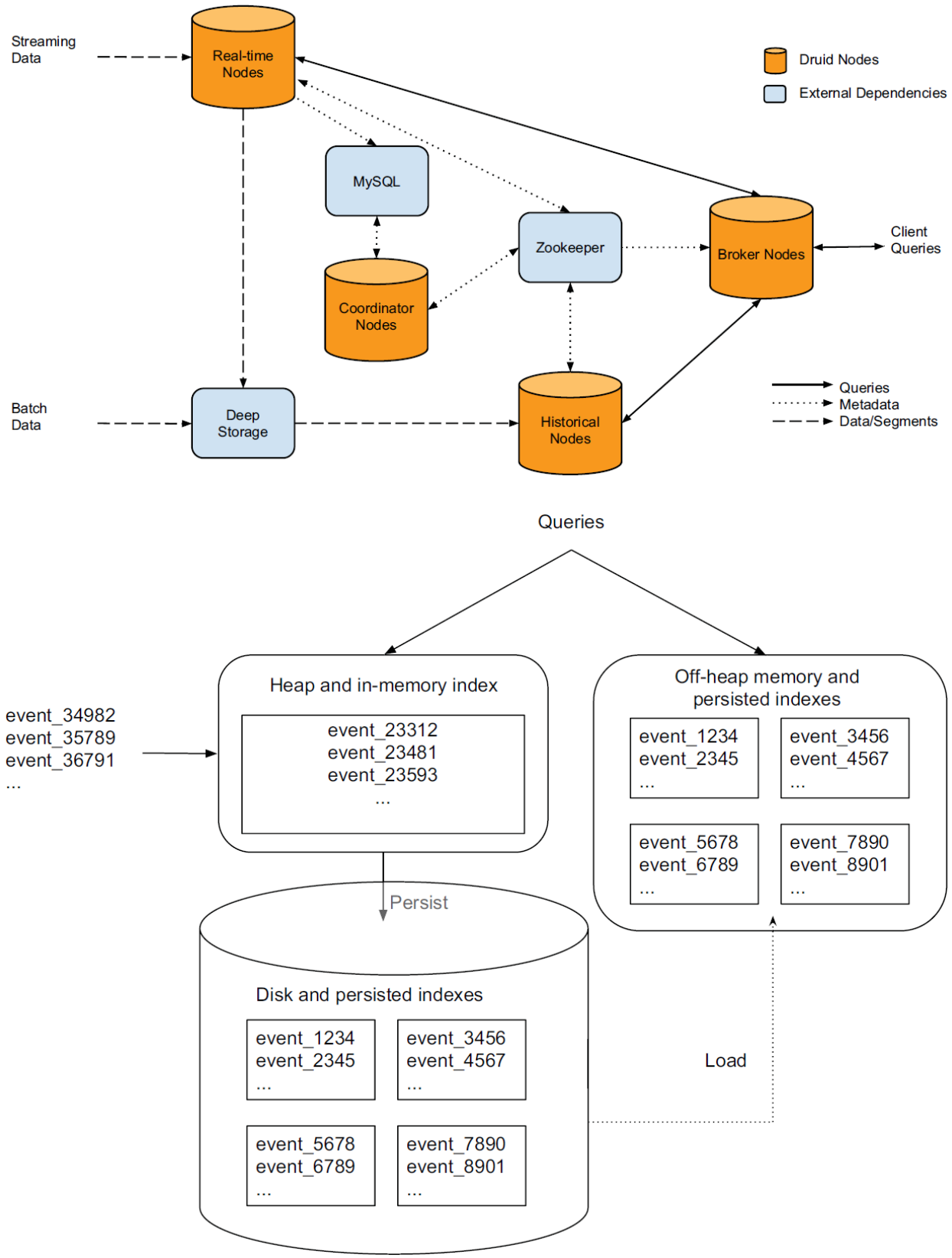


Fig. 4: Source: Druid: A Real-time Analytical Data Store

3. The persisted data is loaded into off-heap memory to be still queryable.
4. On a periodic basis, the persisted indexes are merged together to form a "segment" of data and then get handed off to deep storage.

In this way, all events ingested into real-time nodes, regardless before or after persisted, are present in memory (either on- or off-heap) and thus can be queried (queries hit both the in-memory and persisted indexes). This functionality of real-time nodes enables Druid to conduct real-time data ingestion meaning that events can be queried almost as soon as they occur. In addition, there is no data loss during these steps. In addition, there is no data loss during these steps.

Real-time nodes announce their online state and the data they serve in Zookeeper (see *External dependencies*) for the purpose of coordination with the rest of the Druid cluster.

Historical nodes

Historical nodes function to load and serve the immutable blocks of data (segments) created by real-time nodes. These nodes download immutable segments locally from the deep storage and serve queries over those segments (e.g., data aggregation/filtering). The nodes are operationally simple based on a shared-nothing architecture; they have no single point of contention and simply load, drop, and serve segments as instructed by Zookeeper.

A historical node's process of serving a query is as follows:

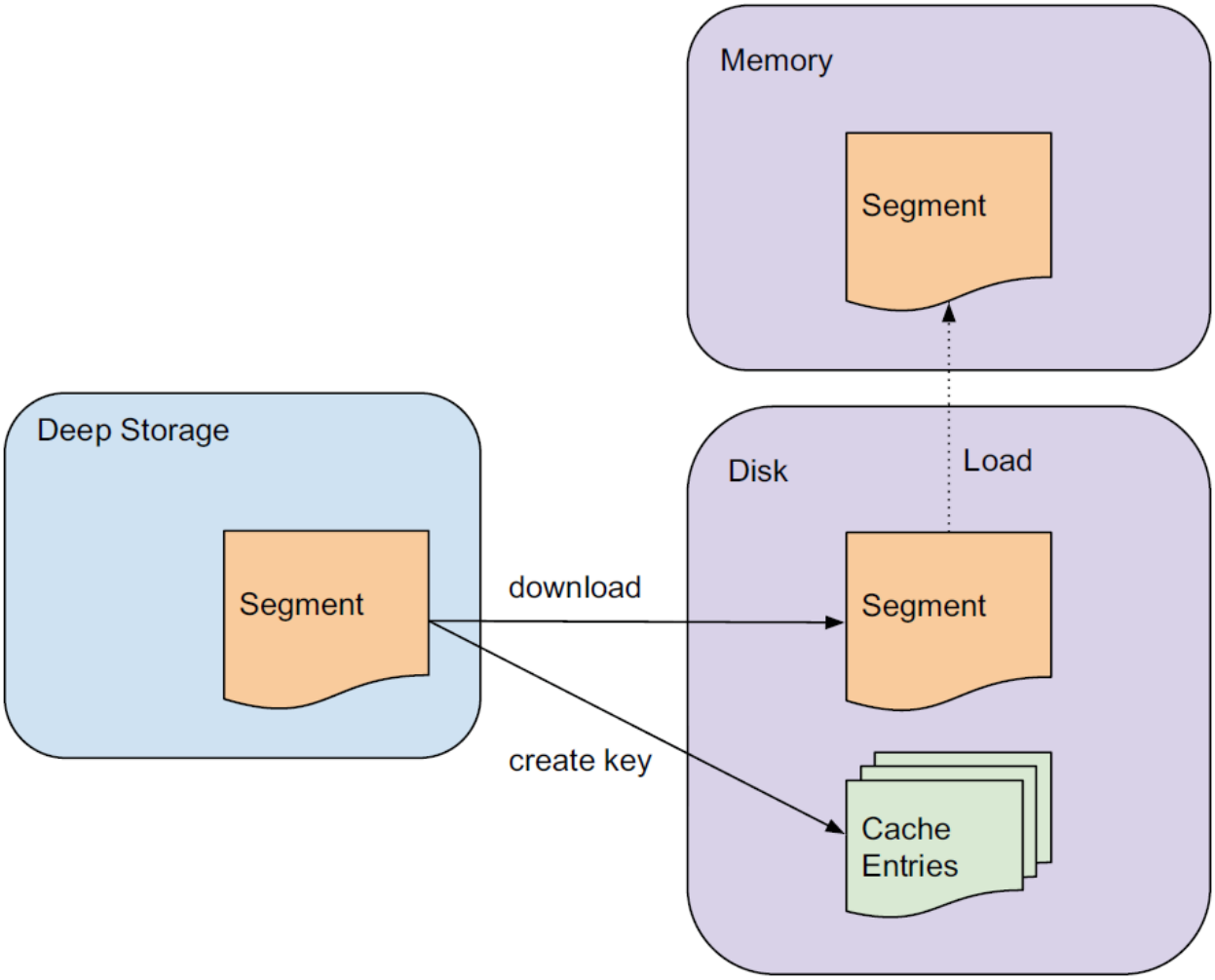


Fig. 5: Source: Druid: A Real-time Analytical Data Store

Once a query is received, the historical node first checks a local cache that maintains information about what segments already exist on the node. If information about a segment in question is not present in the cache, the node will proceed to download the segment from deep storage. On the completion of the processing, the segment is announced in Zookeeper to become queryable and the node performs the requested query on the segment.

Historical nodes can support read consistency because they only deal with immutable data. Immutable data blocks also enable a simple parallelization model: historical nodes can concurrently scan and aggregate immutable blocks without blocking.

Similar to real-time nodes, historical nodes announce their online state and the data they are serving in Zookeeper.

Broker nodes

Broker nodes understand the metadata published in Zookeeper about what segments are queryable and where those segments are located. Broker nodes route incoming queries such that the queries hit the right historical or real-time nodes. Broker nodes also merge partial results from historical and real-time nodes before returning a final consolidated result to the caller.

Broker nodes use a cache for resource efficiency as follows:

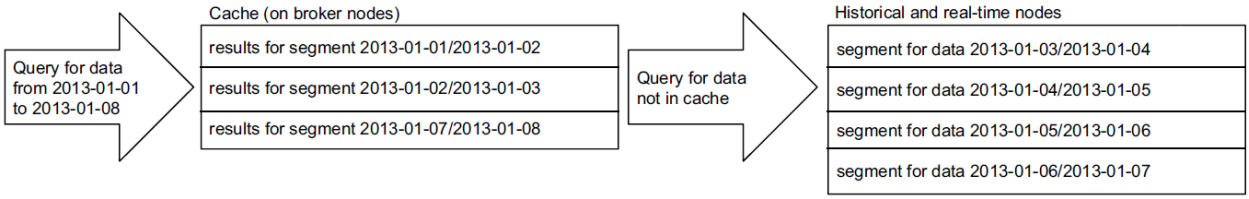


Fig. 6: Source: Druid: A Real-time Analytical Data Store

Once a broker node receives a query involving a number of segments, it checks for segments already existing in the cache. For any segments absent in the cache, the broker node will forward the query to the correct historical and real-time nodes. Once historical nodes return their results, the broker will cache these results on a per-segment basis for future use. Real-time data is never cached and hence requests for real-time data will always be forwarded to real-time nodes. Since real-time data is perpetually changing, caching the results is unreliable.

Coordinator nodes

Coordinator nodes are primarily in charge of data management and distribution on historical nodes. The coordinator nodes determine which historical nodes perform queries on which segments and tell them to load new data, drop outdated data, replicate data, and move data to load balance. This enables fast, efficient, and stable data processing in a distributed group of historical nodes.

As with all Druid nodes, coordinator nodes maintain a Zookeeper connection for current cluster information. Coordinator nodes also maintain a connection to a MySQL database that contains additional operational parameters and configurations, including a rule table that governs how segments are created, destroyed, and replicated in the cluster.

Coordinator nodes undergo a leader-election process that determines a single node that runs the coordinator functionality. The remaining coordinator nodes act as redundant backups.

External dependencies

Druid has a couple of external dependencies for cluster operations.

- **Zookeeper:** Druid relies on Zookeeper for intra-cluster communication.
- **Metadata storage:** Druid relies on a metadata storage to store metadata about segments and configuration. MySQL and PostgreSQL are popular metadata stores for production.
- **Deep storage:** Deep storage acts as a permanent backup of segments. Services that create segments upload segments to deep storage and historical nodes download segments from deep storage. S3 and HDFS are popular deep storages.

High availability characteristics

Druid is designed to have no single point of failure. The different node types operate fairly independent of each other and there is minimal interaction among them. Hence, intra-cluster communication failures have minimal impact on data availability. To run a highly available Druid cluster, you should have at least two nodes of every node type running.

Architecture extensibility

Druid features a modular, extensible platform that allows various external modules to be added to its basic architecture. An example of how Druid's architecture can be extended with modules is shown below:

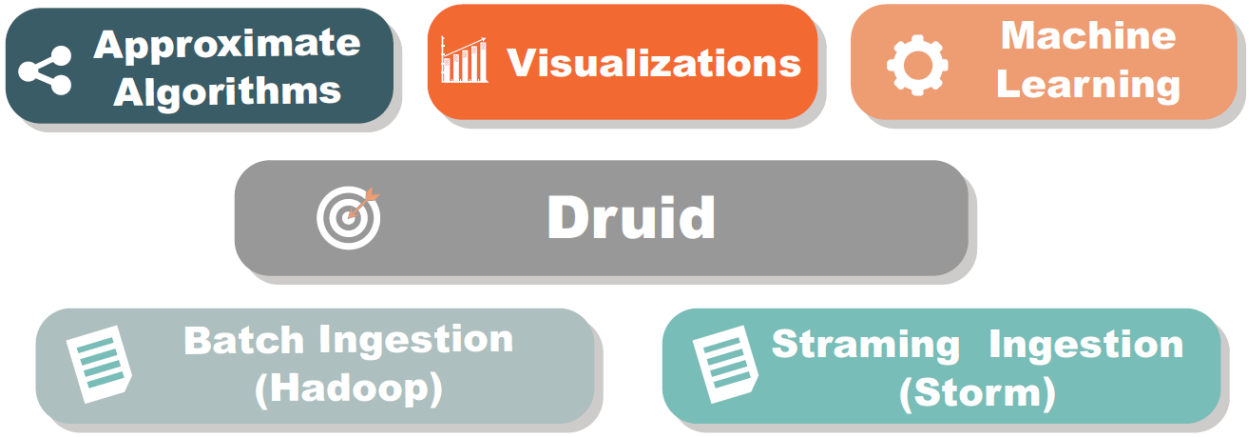


Fig. 7: Source: MetaMarkets - Introduction to Druid by Fangjin Yang

Metatron, an end-to-end business intelligence solution to be introduced in this paper, was also built by adding various modules to the Druid engine.

2.3.4 Druid performance assessments

With Druid being a data store that supports real-time data exploration, its quantitative assessments are focused on two key aspects:

- Query latency
- Ingestion latency

This is because the key to achieving "real-time" performance is to minimize the time spent on query processing and ingestion. A number of organizations and individuals, including the developers of Druid, have established benchmarks for Druid performance assessment based on the two key aspects, and shared how Druid compares to other database management systems.

Self-assessment by Druid developers

Druid: A Real-time Analytical Data Store[1] was published by the developers in 2014. Chapter 6. Performance contains details of Druid assessment, with a particular focus on query and ingestion latencies. The benchmarks of Druid performance are briefly introduced in the following sections.

Query latency

Regarding Druid's query latency, the paper discusses two performance assessments?one was conducted on eight data sources that had been most queried at Metamarkets and the other was on TPC-H datasets. In this section, we review the latter assessment. The latencies from querying on TPC-H datasets were measured by comparing with MySQL, and the cluster environment was as follows:

- **Druid historical nodes:** Amazon EC2 m3.2xlarge instance types (Intel® Xeon® E5-2680 v2 @ 2.80GHz)
- **Druid broker nodes:** c3.2xlarge instances (Intel® Xeon® E5-2670 v2 @ 2.50GHz)
- **Pledged mountain draw converting** (subtract soft a3.2analysed repurchase pairs)

The figure below shows the query latencies resulting from Druid and MySQL when tested on the 1GB and 100GB TPC-H datasets:

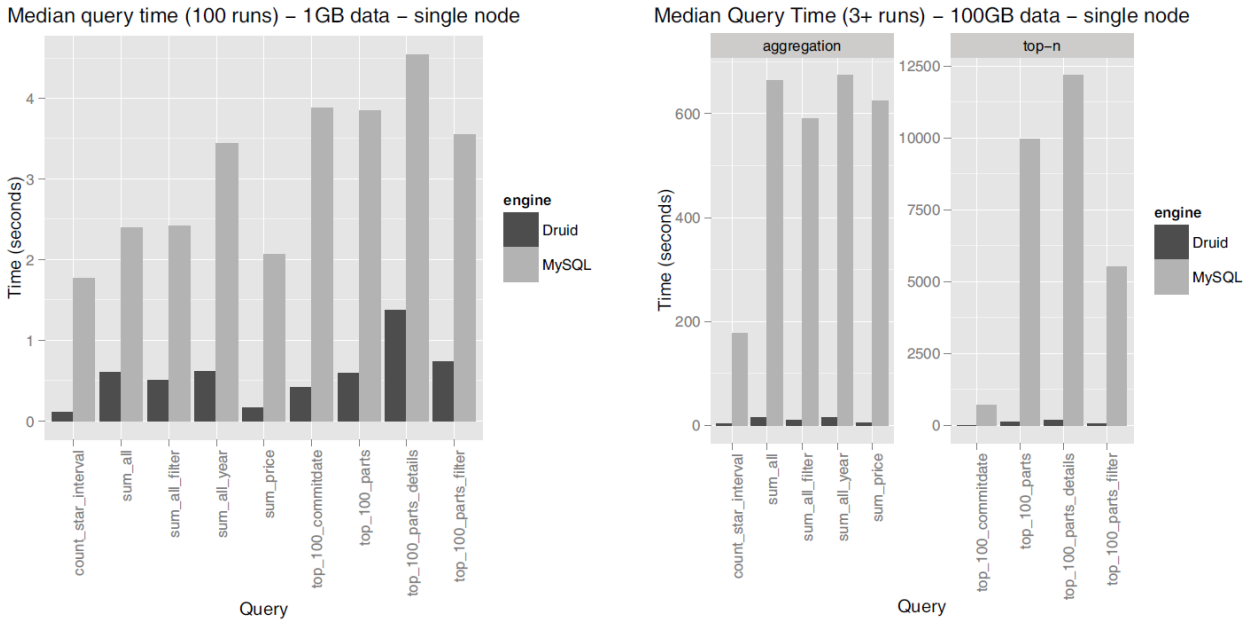


Fig. 8: Source: Druid: A Real-time Analytical Data Store

[1] F. Yang, E. Tschetter, X. Léauté, N. Ray, G. Merlino, and D. Ganguli. (2014). *Druid: a real-time analytical data store*. Retrieved from http://druid.io/docs/0.12.1/design/index.html.

By showcasing these results, the paper suggests that Druid is capable of extremely faster query returns compared to legacy relational database systems.

The Druid paper also presents how faster query returns are achieved when multiple nodes are joined together in a cluster. When tested on the TPC-H 100 GB dataset, the performance difference between a single node (8 cores) and six-node cluster (48 cores) was as follows:

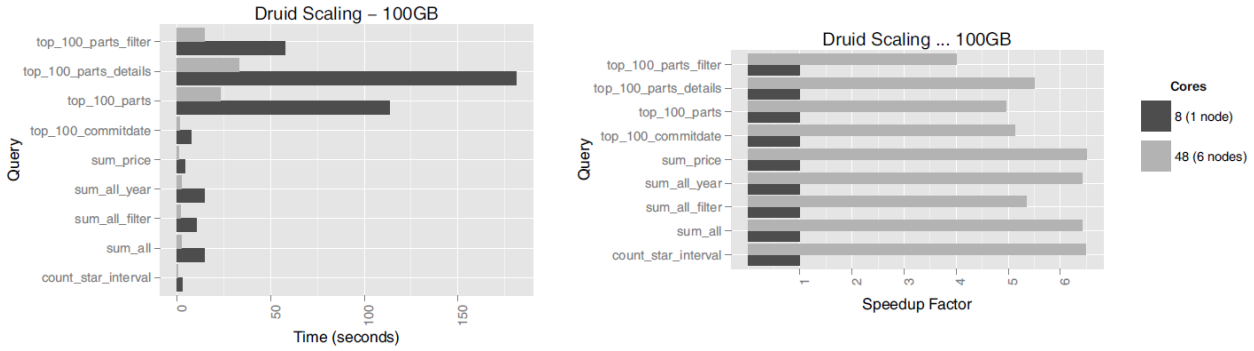


Fig. 9: Source: Druid: A Real-time Analytical Data Store

It was observed that not all types of queries achieve linear scaling, but the simpler aggregation queries do, ensuring a speed increment almost proportional to the number of the cores (SK Telecom's Metatron has made improvements to achieve much more obvious linear scalability).

Ingestion latency

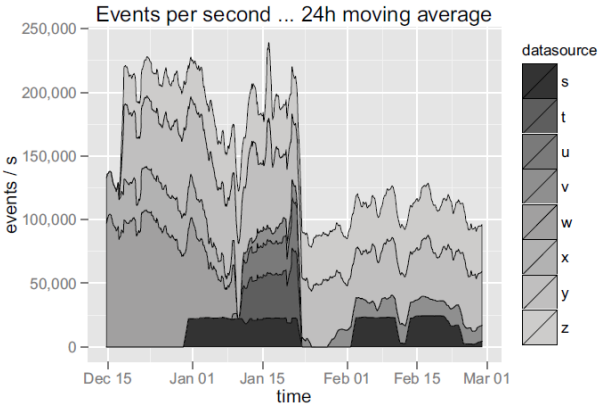
The paper also assessed Druid's data ingestion latency on a production ingestion setup consisting of:

- 6 nodes, totalling 360GB of RAM and 96 cores (12 x Intel®Xeon®E5-2670).

A total of eight production data sources were selected for this assessment. The characteristics of each data source and their ingestion results are shown below. Note that in this setup, several other data sources were being ingested and many other Druid related ingestion tasks were running concurrently on the machines.

Data Source	Dimensions	Metrics	Peak events/s
s	7	2	28334.60
t	10	7	68808.70
u	5	1	49933.93
v	30	10	22240.45
w	35	14	135763.17
x	28	6	46525.85
y	33	24	162462.41
z	33	24	95747.74

Ingestion characteristics of various data sources



Combined cluster ingestion rates

Fig. 10: Source: Druid: A Real-time Analytical Data Store

Druid's data ingestion latency is heavily dependent on the complexity of the dataset being ingested, but the latency measurements present here are sufficient to demonstrate that Druid well addresses the stated problems of interactivity.

Druid performance assessment by SK Telecom

SK Telecom also measured the query and ingestion latencies of Druid as detailed below:

Query latency test

The conditions of query latency measurement were as follows:

- Data: TPC-H 100G dataset (900 million rows)
- Pre-aggregation granularity: day
- Servers: r3.4xlarge nodes, (2.5GHz * 16, 122G, 320G SSD) * 6
- No. of historical nodes: 6
- No. of broker nodes: 1

The query times for five queries of the TPC-H 100G dataset were as follows (the query times in Hive were also measured as a reference):

Fig. 11: Source: SK Telecom T-DE WIKI Metatron Project

Note: The reasons why the Hive benchmark performed poorly include that some processes were performed through Thrift and the dataset wasn't partitioned.

Ingestion latency test

The conditions of ingestion latency measurement were as follows:

- Ingestion data size: 30 million rows/day, 10 columns
- Memory: 512 GB
- CPU: Intel (R) Xeon (R) Gold 5120 CPU @ 2.20 GHz (56 cores)
- No. of historical nodes: 100
- No. of broker nodes: 2
- Jobs performed by three out of ten middle-manager nodes
- Ingestion tool: Apache Kafka

Data ingestion was performed 100 times under the conditions specified above, and the average ingestion latency was 1.623439 seconds. As illustrated below, ingestion latency was computed as the sum of Kafka ingestion latency, Druid ingestion latency, and Druid query latency.

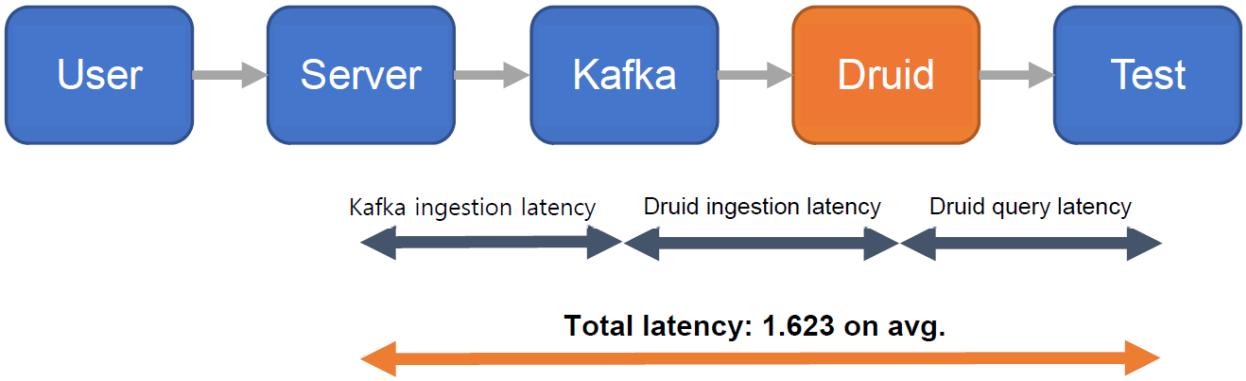


Fig. 12: Source: SK Telecom T-DE WIKI Metatron Project

Druid assessments by third parties

Druid assessment by Outlyer

In the Outlyer blog, twenty open source time-series database systems were assessed in a post[2] titled Top 10 Time Series Databases and published on August 26, 2016. The author Steven Acreman ranked Druid in the 8th place, and his set of criteria was as follows:

[2] Steven Acreman. (2016, Aug 26). *Top 10 Time Series Databases*. Retrieved from https://blog.outlyer.com/top10-open-source-time-series-databases.

Table 1: A summary of Druid assessment by Outlyer

Items	Druid performance
Write performance - single node	25k metrics/sec Source: https://groups.google.com/forum/#!searchin/druid-user/benchmark%7C
Write performance - 5-node cluster	100k metrics / sec (calculated)
Query performance	Moderate
Maturity	Stable
Pro's	Good data model and cool set of analytics features. Mostly designed for fast queries over large
Con's	Painful to operate, not very fast write throughput. Real time ingestion is tricky to setup.

Druid assessment by DB-Engines

DB-Engines[3], an online website, publishes a list of database management systems ranked by their current popularity every months. To measure the popularity of a system, it uses the following parameters:

- Number of mentions of the system on websites: It is measured as the number of results in queries of the search engines Google, Bing and Yandex.
- General interest in the system: For this measurement, the frequency of searches in Google Trends is used.
- Frequency of technical discussions about the system: The ranking list uses the number of related questions and the number of interested users on the well-known IT-related Q&A sites Stack Overflow and DBA Stack Exchange.
- Number of job offers, in which the system is mentioned: The ranking list uses the number of offers on the leading job search engines Indeed and Simply Hired.
- Number of profiles in professional networks, in which the system is mentioned: The ranking list uses the internationally most popular professional networks LinkedIn and Upwork.
- Relevance in social networks. The ranking list counts the number of Twitter tweets, in which the system is mentioned.

As of July 2018, Druid ranked 118th out of a total of 343 systems, and 7th out of 25 time-series database systems.

Comparison with Apache Spark

Comparing Druid with Apache Spark is meaningful because both technologies are emerging as next-generation solutions for large-scale analytics and their different advantages make them very complementary when combined together. Metatron also makes use of this combination: Druid as the data storage/processing engine and Spark as an advanced analytics module.

This section briefly introduces a report comparing the performance of Druid and Spark[4][5] published by Harish Butani, the founder of Sparkline Data Inc. Prior to the performance comparison, the report states that the two solutions are in complementary relations, rather than competitors.

[3] DB-Engines website. https://db-engines.com, July 2018.

[4] Harish Butani. (2018, Sep 18). Combining Druid and Spark: Interactive and Flexible Analytics at Scale. Retrieved from https://www.linkedin.com/pulse/combining-druid-spark-interactiveflexible-analytics-scale-butani.

[5] Harish Butani. (2015, Aug 28). TPCH Benchmark. Retrieved from https://github.com/SparklineData/spark-druid-olap/blob/master/docs/benchmark/BenchMarkDetails.pdf.

Apache Spark characteristics

Apache Spark is an open-source cluster computing framework providing rich APIs in Java, Scala, Python, and R. Spark's programming model is used to build analytical solutions that combine SQL, machine learning, and graph processing. Spark supports powerful functions to process large-scale and/or complex data manipulation workflows, but it isn't necessarily optimized for interactive queries.

Dataset, queries, performance results

For the benchmark, the 10G TPC-H dataset was used. The 10G star schema was converted into a flattened (denormalized) transaction dataset and reorganized to be queryable in Druid and Spark. The sizes of the resulting datasets were:

- TPCH Flat TSV: 46.80GB
- Druid Index in HDFS: 17.04GB
- TPCH Flat Parquet: 11.38GB
- TPCH Flat Parquet Partition by Month: 11.56GB

And then, a number of queries were chosen to test the performance differences in various aspects as shown below:

Table 2: Queries used for query latency comparison between Druid Apache Spark

Query	Interval	Filters	Group By	Ag
Basic Aggregation.	None	None	ReturnFlag LineStatus	Co
Ship Date Range	1995-12/1997-09	None	ReturnFlag LineStatus	Co
SubQry Nation, pType ShpDt Range	1995-12/1997-09	P_Type S_Nation + C_Nation	S_Nation	Co
TPCH Q1	None	None	ReturnFlag LineStatus	Co
TPCH Q3	1995-03-15-	O_Date MktSegment	Okey Odate ShipPri	Su
TPCH Q5	None	O_Date Region	S_Nation	Su
TPCH Q7	None	S_Nation + C_Nation	S_Nation C_Nation ShipDate.Year	Su
TPCH Q8	None	Region Type O_Date	ODate.Year	Su

The test results are as follows:

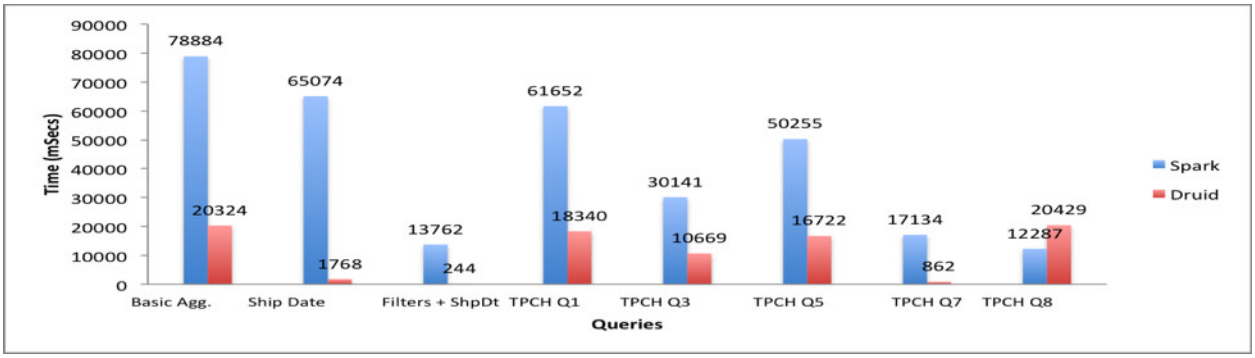


Fig. 13: Source: Combining Druid and Spark: Interactive and Flexible Analytics at Scale

- The Filters + Ship Date query provides the greatest performance gain (over 50 times over Spark) when Druid is used. This is not surprising as this query is a typical slice-and-dice query tailor-made

for Druid. Along the same lines, TPCH Q7 shows a significant performance boost when running on Druid: milliseconds on Druid vs. 10s of seconds on Spark.

- For TPCH Q3, Q5, and Q8 there is an improvement, but not to the same level as Q7. This is because the OrderDate predicate is translated to a JavaScript filter in Druid, which is significantly slower than a native Java filter.
- The Basic Aggregation and TPCH Q1 queries definitely show improvement. The Count-Distinct operation is translated to a cardinality aggregator in Druid, which is an approximate count. This is definitely an advantage for Druid, especially for large cardinality dimensions.

These results can vary with testing conditions, but one thing is clear: Queries that have time partitioning or dimensional predicates (like those commonly found in OLAP workflows) are significantly faster in Druid.

Implications

The testing results showcase that combining the analytic capabilities with Spark and the OLAP and low latency capabilities of Druid can create great synergy. Druid ingests, explores, filters, and aggregates data efficiently and interactively, while the rich programming APIs of Spark enable in-depth analytics. By leveraging these different capabilities, we can build a more powerful, flexible, and extremely low latency analytics solution.

References

2.3.5 Metatron powered by Druid

As explained previously, Metatron employs Druid as its underlying engine and has made developments and improvements of Druid for its own uses. This section introduces the background, progress, and results of the adoption of Druid to Metatron.

Metatron development background and Druid integration

Metatron as a big data analytics solution

As a telecommunications service provider with the most number of subscribers in South Korea, SK Telecom has exerted significant efforts to establish a stable network environment through by using the mass amounts of network data logs generated by its users.

Due to the limitations of existing IT infrastructure in mass data processing, SK Telecom needed a big-data warehousing system (Apache Hadoop) and a big-data analytics solution compatible with the system. The company built its own Hadoop infrastructure to store mass amounts of data at low cost, but faced the following limitations:

- Network data generated by the countless users could not be analyzed in real time. Although it was possible to store and process big data, visualizations could be implemented only with a sampled subset of data in the same way as on legacy systems.
- Having different solutions and different managers support each stage of data analytics, such as ETL, DW, and BI, not only involved significant time and costs, but also resulted in poor data accessibility. An end-to-end solution was needed to analyze all stages at once in a simple and quick manner.

Why the Druid engine

Druid was the optimal engine for the Metatron solution because it fulfilled the aforementioned needs with the features below:

- Druid collects mass amounts of data in real time and indexes them into a queryable format, ensuring very fast data aggregations (a few seconds at the slowest) based on distributed processing.
- Druid's OLAP time-series data format enables analysts to perform data exploration, filtering, and visualization as desired. Such free and flexible data exploration is essential for users to intuitively select the required data and determine correlations between different dimensions on it.
- Druid's extensible architecture allows modules to be easily added.

Built on this architecture, Metatron is an end-to-end solution that embraces all layers of data collection, storage, processing, analysis, and visualization.

Druid engine integration

The Druid engine was integrated in Metatron as follows:

- With Druid as the basic engine for processing/analytics, the GUI was designed to support users in different professional domains and big-data analysts in data-related tasks such as data preparation, analytics, and visualization, as well as the sharing of results.
- IT administrators can manage/monitor data sources in Druid, and they can establish data preparation rules if data sources of higher quality are required.

Druid functions reinforced in Metatron

The open-source Druid, despite its strengths in data collection and processing, had to be improved for Metatron to properly function as an end-to-end solution. This section examines the limitations of the open-source Druid and the functions reinforced in Metatron.

Limitations of the open-source Druid

The open-source Druid has the following limitations:

- Since Druid does not yet have full support for joins, Metatron uses another SQL engine for data preparation.
- Druid supports only a subset of SQL queries.
- For a data lake, a traditional SQL engine is more appropriate.
- Druid cannot append to or update already indexed segments, except for in some unusual cases.
- Nulls are not allowed.
- Filtering is not supported for metric columns.
- Linear scalability is not ensured. Increasing the number of servers doesn't improve the performance as much.
- Only a few data types are supported and it is difficult to add a new one.
- The management and monitoring tools are not powerful enough.

Druid functions reinforced in Metatron

The following functions of Druid were strengthened in Metatron:

Query functionality improvements

- Improved the functionality of the GroupBy query type.

- Slightly improved the functionality of other types of queries.

Features added

- Virtual columns (map, expression. etc.)
- New metric types (double, string, array, etc.)
- New expression functions
- Druid query results can be stored on the HDFS or exported into a file.
- Queries for meta information and statistics
- New aggregate functions (variance, correlation, etc.)
- (Limited) Window functions (lead, lar, running aggregations, etc.)
- (Limited) Joins
- (Limited) Sub-queries
- Temporary data sources
- Complex queries (data source summarization, correlation between data sources, k-means, etc.)
- Custom columns grouping
- Geographic information system (GIS) supported
- Columnar histograms
- Bit-slice indexing

Index structure improvements

- Histograms for filtering on metrics
- Lucene format supported for text filtering

Connectability with other systems

- Hive storage handler
- Ingestion into Hive tables (based on connection with the Hive metastore)
- Ingestion into the ORC format
- RDBMS data ingestion via based on JDBC
- (Limited) SQL support backported

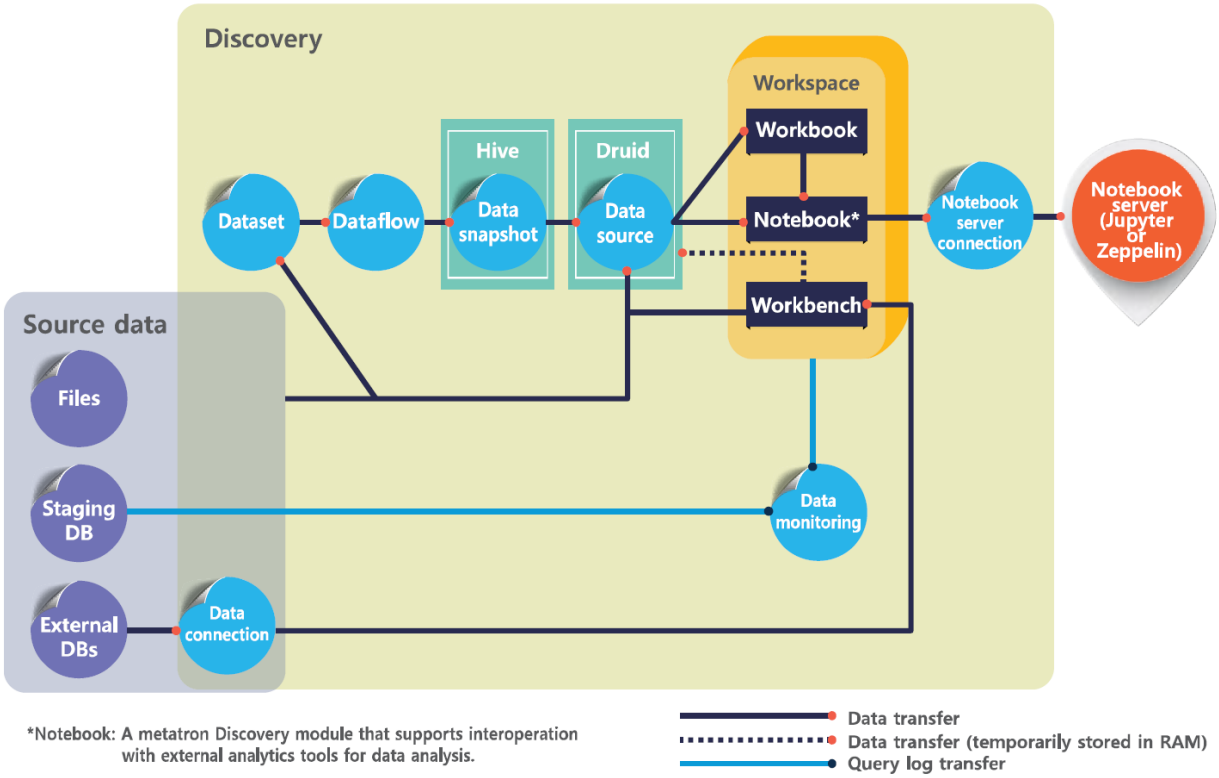
Miscellaneous improvements

- Bug fixes (+50) and minor improvements

CHAPTER

THREE

DATA MANAGEMENT



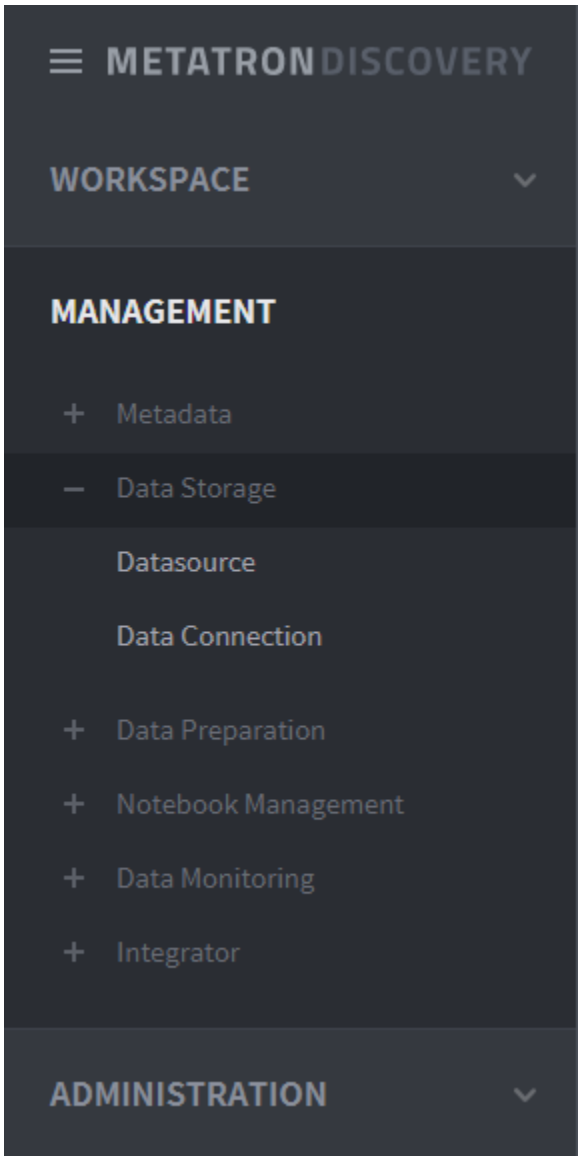
As shown above, data used by the three Discovery modules (workbook, notebook and workbench) is prepared from various types of source data, engines, and storages. For these operations, data flows need to be standardized and managed, and different types of source data need to be linked.

Source data required for analysis and visualization is either ingested into the Metatron engine as a **data source**, or linked directly from an external database with a **data connection**. Data usage can be monitored and tracked using **data monitoring**.

3.1 Data Source

In Metatron Discovery, a "data source" refers to a Druid database table into which data is ingested. Based on these data sources, workbooks and notebooks perform data analytics and visualization.

The Data Source menu can be accessed under **MANAGEMENT > Data Storage > Data Source** on the left-hand panel of the main screen.

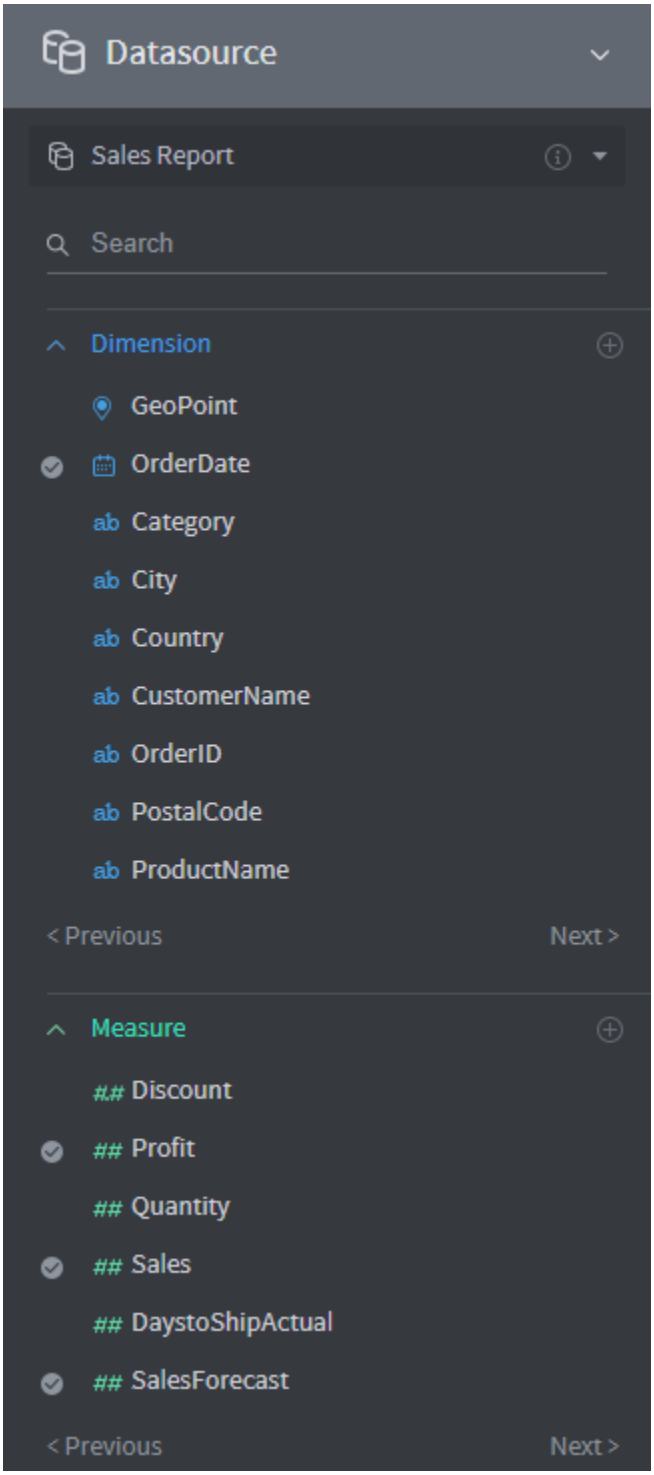


3.1.1 "Dimensions" and "Measures"

The columns of a data source linked to the dashboard are categorized into **dimension** and **measure** columns as explained below. To make full use of Discovery's data analysis and visualization features, you must understand the concepts of dimensions and measures clearly.

Dimension column

A column containing categorical data with the following characteristics:


Datasource
Sales Report
Search
Dimension
GeoPoint
OrderDate
Category
City
Country
CustomerName
OrderID
PostalCode
ProductName
< Previous
Next >
Measure
Discount
Profit
Quantity
Sales
DaystoShipActual
SalesForecast
< Previous
Next >

- The values in this type of column are not for aggregation but to be categorized (e.g.: Category, Region, Organization)
- By each of these categories, measure values are aggregated.

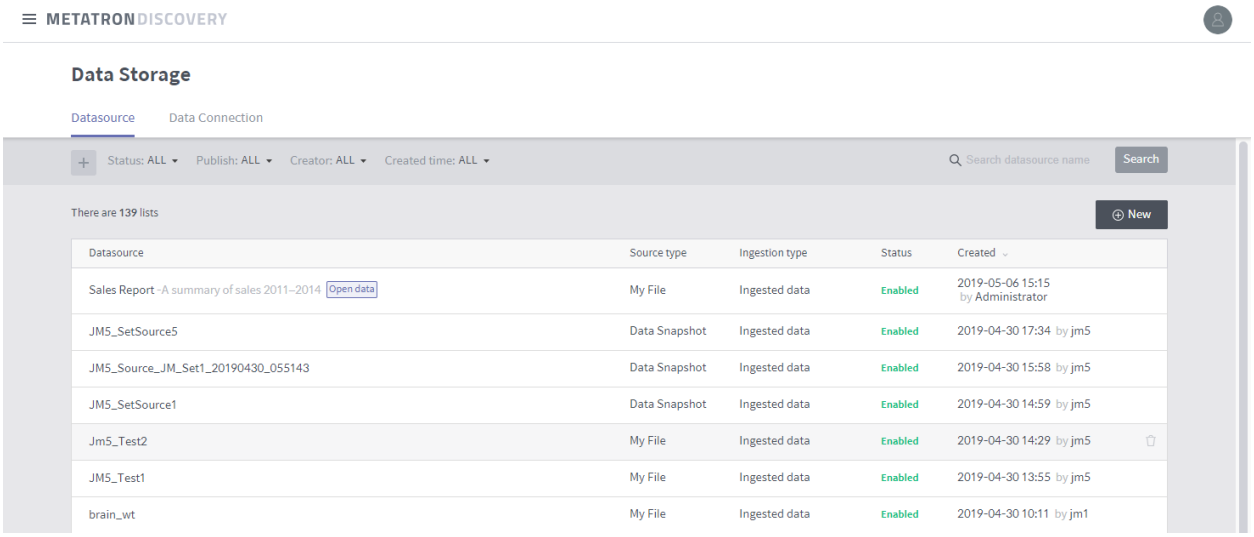
Measure columns

A column containing quantitative data with the following characteristics:

- The values in this type of column are subject to aggregation or contain quantitative information (e.g.: Sales)
- These values are aggregated based on dimensions.

3.1.2 Data source management home

On this home page, you can create, edit and view data sources.

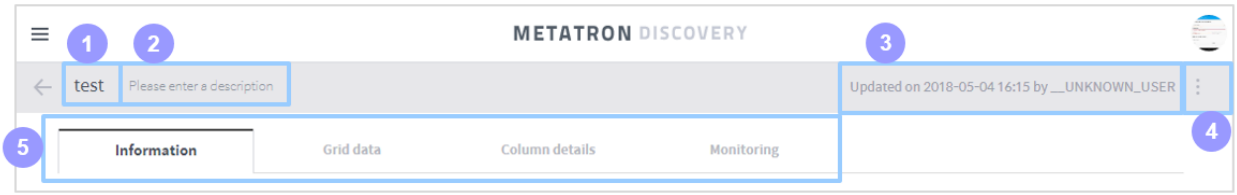


1. **Status:** Filters the data source list by the availability of data sources stored in the data storage.
   - **Enable:** Displays data sources that have been ingested and are available in workbooks or workbenches.
   - **Preparing:** Displays new data sources whose ingestion is in progress.
   - **Failed:** Displays data sources that have not been created properly.
   - **Disabled:** Displays data sources that have been ingested but are not available because of an error in a certain Druid process.
2. **Publish:** Filter the data source list by public workspace.
   - **Open Data:** Displays only data sources publicly available in all workspaces.
   - **Admin Workspace:** Displays only data sources available in the administrator workspace.
   - **Shared workspaces:** Displays only data sources available in the selected shared workspaces.
3. **Creator:** Filters the data source list by user or group that created the source data.
4. **Created time:** Determines whether the data source list is filtered by created or updated time. You can choose from among All, Today, and Last 7 days or specify a time range to display only those entries that were created/updated within the range.

5. **Search by name of data source:** Searches the data source list for the name you type in.
6. **Data source list:** Lists data sources filtered by specified criteria. Click an entry in the list to view its details. (Refer to *Data source details*)
7. **Delete:** Hover the mouse over a data source to display a trash icon. Click the icon to delete the data source.

3.1.3 Data source details

Click a data source listed in the data source management home to view various attributes of that data source. The following subsections describe each area of the data source details. Note that a data source represents a Druid database table stored in Metatron and necessarily includes a timestamp column as a time-series table.



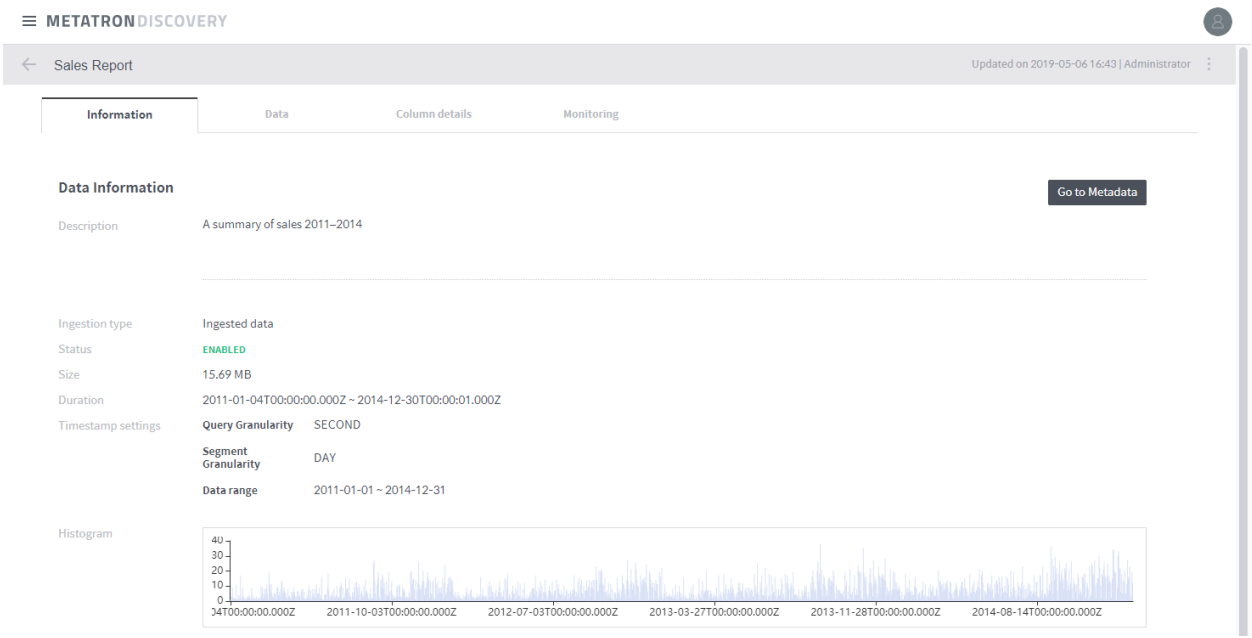
Common top area

1. **Name:** Name of the data source. Click on it if you want modify it.
2. **Description:** Description of the data source. Click on it if you want modify it.
3. **Last update:** Shows who and when last updated the data source.
4. **Delete:** Click this icon to display a menu that allows you to delete the data source.
5. **Tab selection:** Each tab displays a specific set of attributes of the data source. Depending upon the type of data source, not all of the three tabs may be displayed. For details on each tab, refer to the relevant subsection below.

Data information area

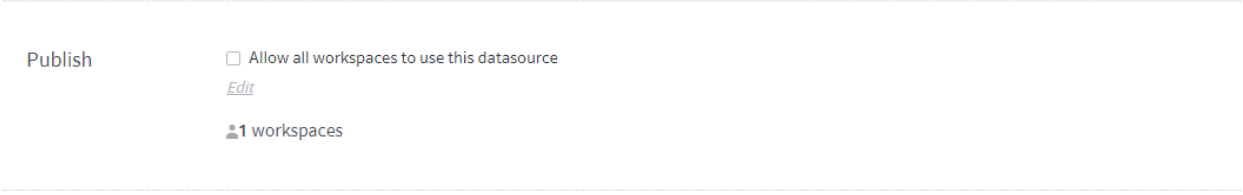
This area displays basic information of the data source.

1. **Data type:** Type of the imported source data from which the data source has been created.
2. **Status:** Displays the availability of the data source.
3. **Size:** Displays the size of the data source.
4. **Duration:** Displays the time range of the timestamps included in the data source.
5. **Timestamp setting:** Displays the granularities defined when the data source was created.
   - **Query Granularity:** Defines the minimum time period by which data is queried. This ensures faster returns by aggregating data per granularity interval.
   - **Segment Granularity:** In Druid, a data source is stored into multiple segments to be processed over multiple nodes in the distributed cluster environment. This granularity setting defines the time intervals into which the data source is partitioned.
   - **Histogram:** A graph displaying the size of the data stored within each time interval in Kbytes. This histogram is can be rendered because the Druid engine timestamps every table record.



Publish area

In this area, you can check and set which workspaces have access to the data source.



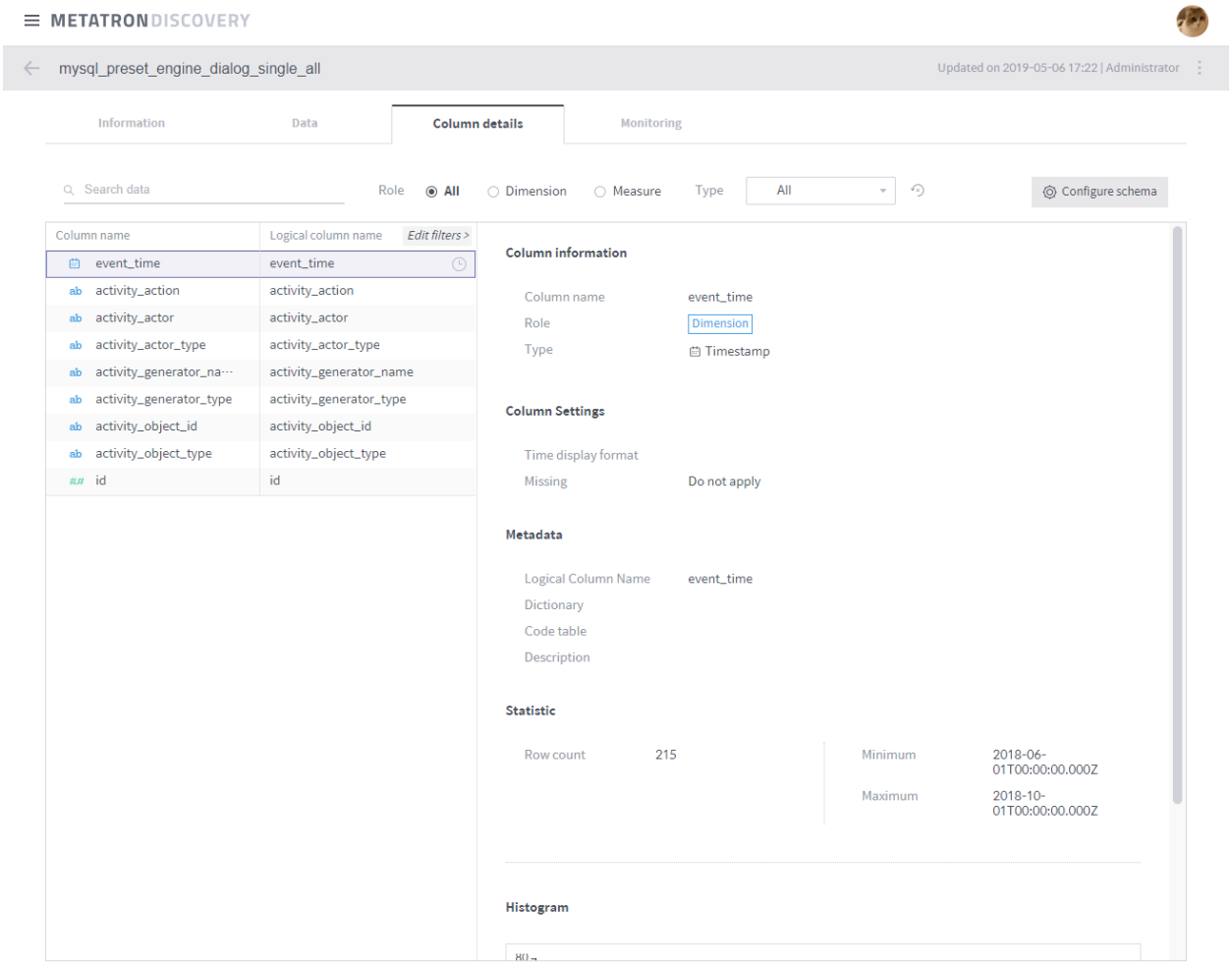
1. **Allow all workspaces to use this data source:** Select this check box to make the data source available in all workspaces.
2. **Edit:** Used to allow specific workspaces to access the data source. This button will disappear if the data source is set as open data.
3. **Number of shared workspaces:** Displays how many workspaces have access to the data source.

Change data schema

The top section of the column details tab provides a user interface to filter columns by the criteria you define. Columns that meet the criteria are displayed on the left. You can also edit column settings.

Column view/settings

1. **Search data:** Searches for columns by the column name you type in.
2. **Role:** Displays all, dimension, or measure columns.
3. **Type:** Displays the columns whose data type is selected.
4. **View all:** Clears all filter settings in the Search data, Role, and Type options and returns to view all columns.
5. **Configure schema:** Click this button to prompt a window to edit the current column settings.


METATRONDISCOVERY
mysql_preset_engine_dialog_single_all
Updated on 2019-05-06 17:22 | Administrator
Information
Data
Column details
Monitoring
Search data
Role
All
Dimension
Measure
Type
All
Configure schema
Column name
Logical column name
Edit filters >
event_time
activity_action
activity_actor
activity_actor_type
activity_generator_na···
activity_generator_type
activity_object_id
activity_object_type
id
Column information
Column name
event_time
Role
Dimension
Type
Timestamp
Column Settings
Time display format
Missing
Do not apply
Metadata
Logical Column Name
event_time
Dictionary
Code table
Description
Statistic
Row count
215
Minimum
2018-06-01T00:00:00.000Z
Maximum
2018-10-01T00:00:00.000Z
Histogram

6. **Column list:** Lists table columns.
7. **Column information:** Displays attributes of the selected column.
8. **Column settings:** Displays the metadata of the selected column.
9. **Statistics:** Displays the row count and other statistical values of the selected column.

Configure the schema

Provides a user interface for editing the name and type of columns.

1. **Role:** Displays whether the column is a dimension or measure.
2. **Name:** Displays the actual name of the column.
3. **Logical name:** Allows you to edit the logical name of the column displayed in the system.
4. **Type:** Allows you to edit the logical type (character/integer/date, etc.) of the column.
5. **Format:** Allows you to edit the display format of the column in the case of the column being a timestamp type.
6. **Description:** Allows you to add a detailed description of the column.

Analyze data statistics

The Monitoring tab reports the usage of the data source.

Change of transaction

Displays the trend of data source transactions over time.

Changes of data size

Displays the trend of the data source size over time.

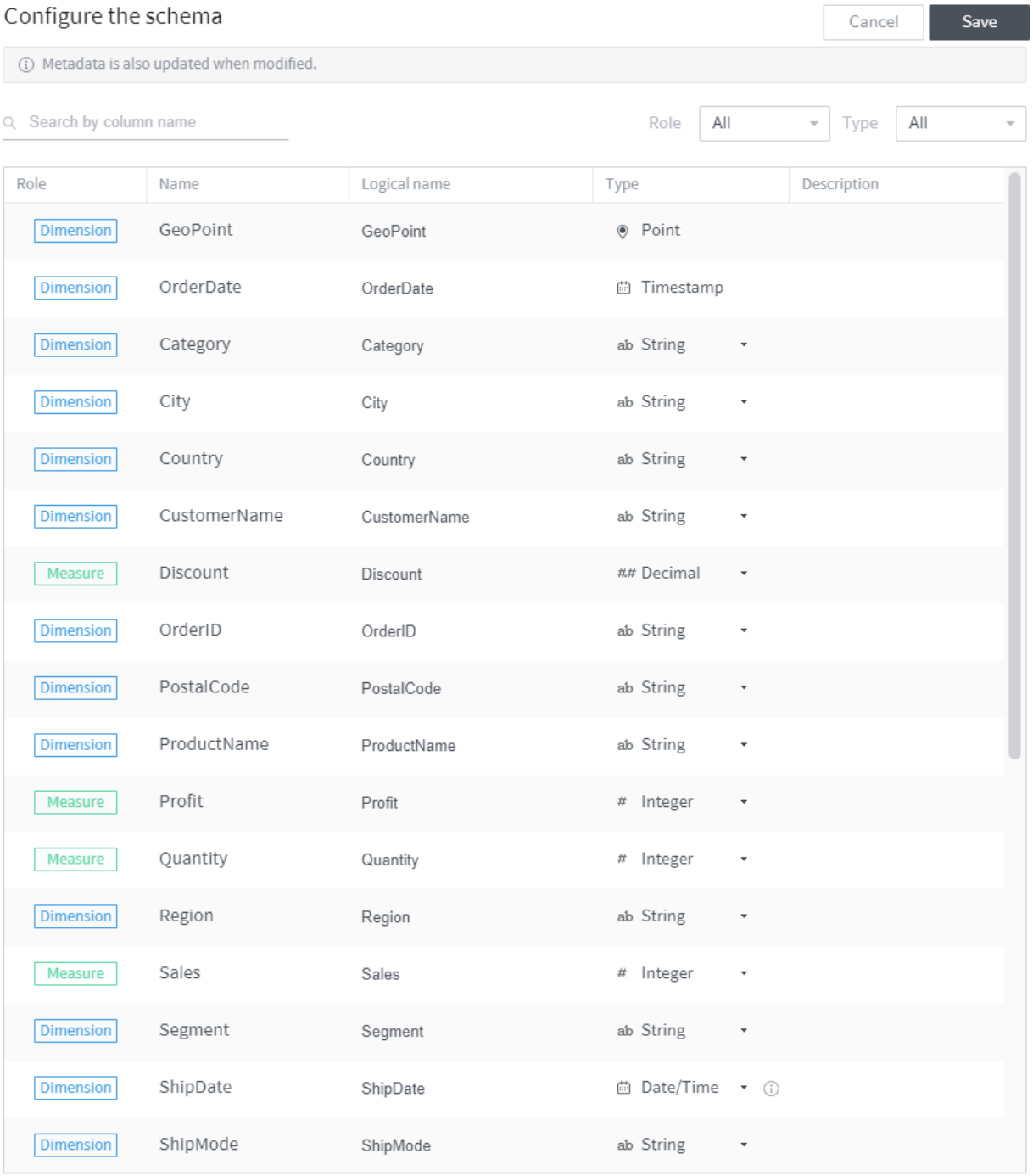
Query distribution (during last one week)

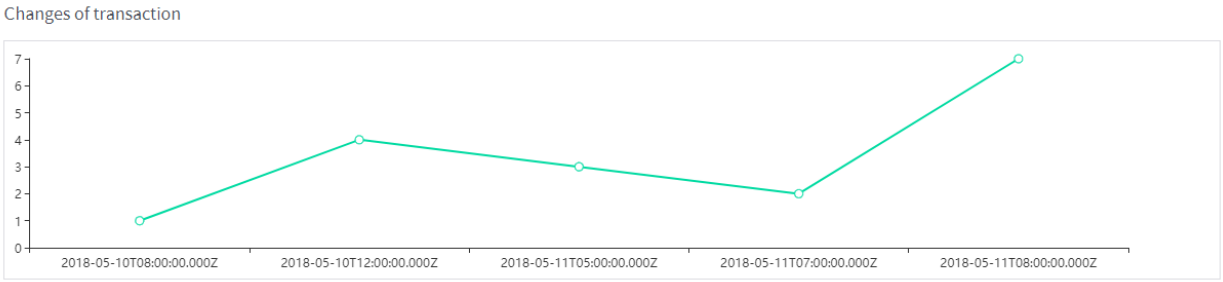
- **Query distribution by user (during last one week):** Displays a pie chart of query percentages by user for the past week.
- **Query distribution by elapsed time (during last one week):** Displays a pie chart of query percentages by execution time for the past week.

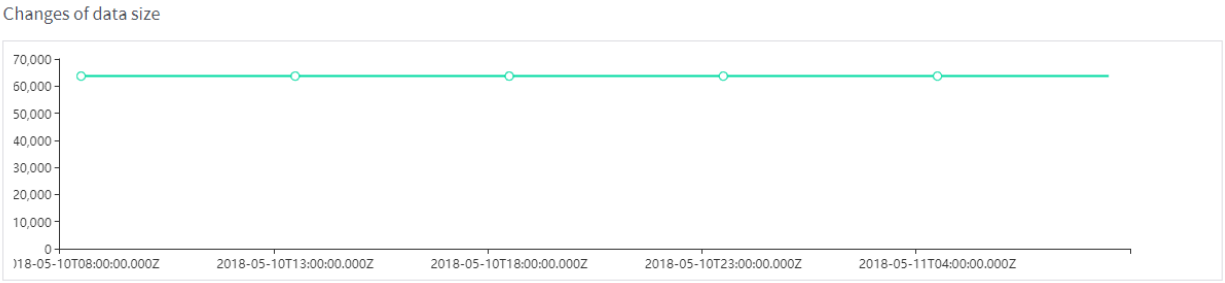
Query log

Used to view a detailed history of each performed query.

1. **Date:** Set a time range to display only those queries that were last executed within this time range.
2. **Query type:** Filters the performed queries by type.
3. **Status:** Displays all, succeeded, or failed queries.
4. **Query list:** Lists queries filtered by specified criteria.
5. **Detail:** Click on it to view the query statement.


Configure the schema
Cancel
Save
Metadata is also updated when modified.
Search by column name
Role
All
Type
All
Role
Name
Logical name
Type
Description
Dimension GeoPoint GeoPoint Point
Dimension OrderDate OrderDate Timestamp
Dimension Category Category ab String
Dimension City City ab String
Dimension Country Country ab String
Dimension CustomerName CustomerName ab String
Measure Discount Discount #.# Decimal
Dimension OrderID OrderID ab String
Dimension PostalCode PostalCode ab String
Dimension ProductName ProductName ab String
Measure Profit Profit # Integer
Measure Quantity Quantity # Integer
Dimension Region Region ab String
Measure Sales Sales # Integer
Dimension Segment Segment ab String
Dimension ShipDate ShipDate Date/Time
Dimension ShipMode ShipMode ab String





3.1.4 Create a data source

This section explains the process of ingesting various types of source data into the Metatron engine and converting them into data sources.

To create a data source, click the **+ New** button at the top right of the **Data Source** home screen.

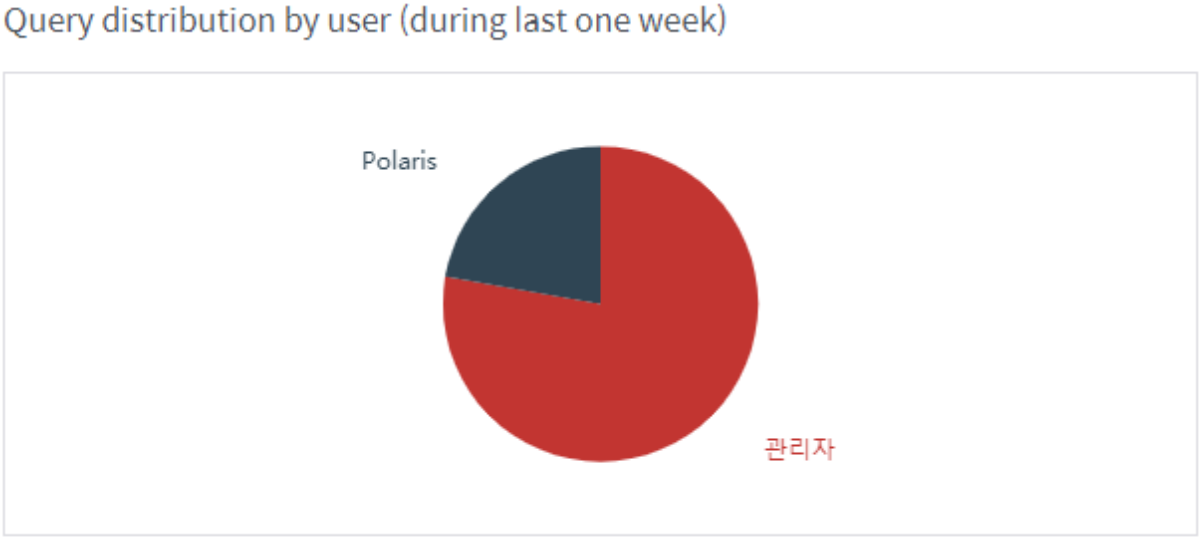
Then, select the type of source data.

- **File:** Creates a data source from a file stored on your local PC (for details, refer to *Create a data source from a file*).
- **Database:** Creates a data source from an external database (for details, refer to *Create a data source from a database*).
- **Staging DB:** Creates a data source from Metatron's internal Hive database (for details, refer to *Create a data source from a staging database*).
- **Stream:** This function is not currently supported.
- **Data Snapshot:** This function is not currently supported.
- **Metatron Engine:** Migrates a data source stored in a previous Metatron version (for details, refer to *Add a data source with the Metatron engine*).

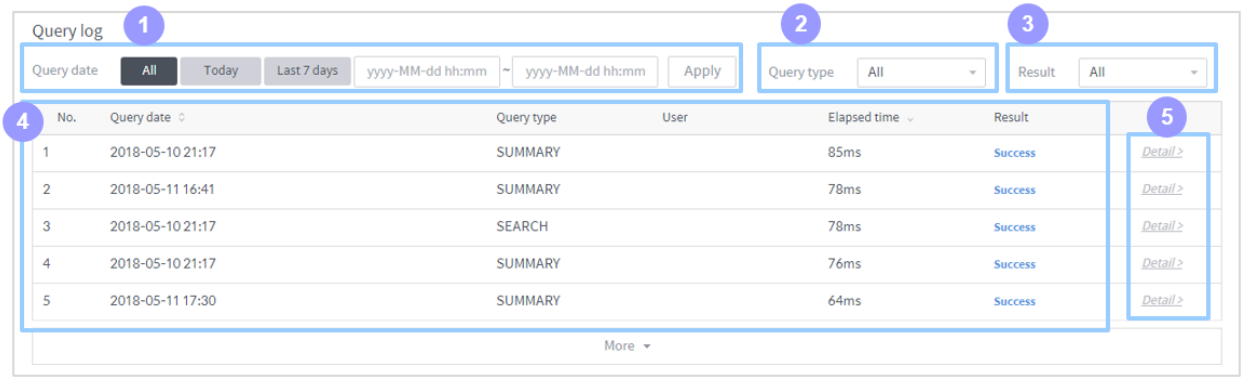
Create a data source from a file

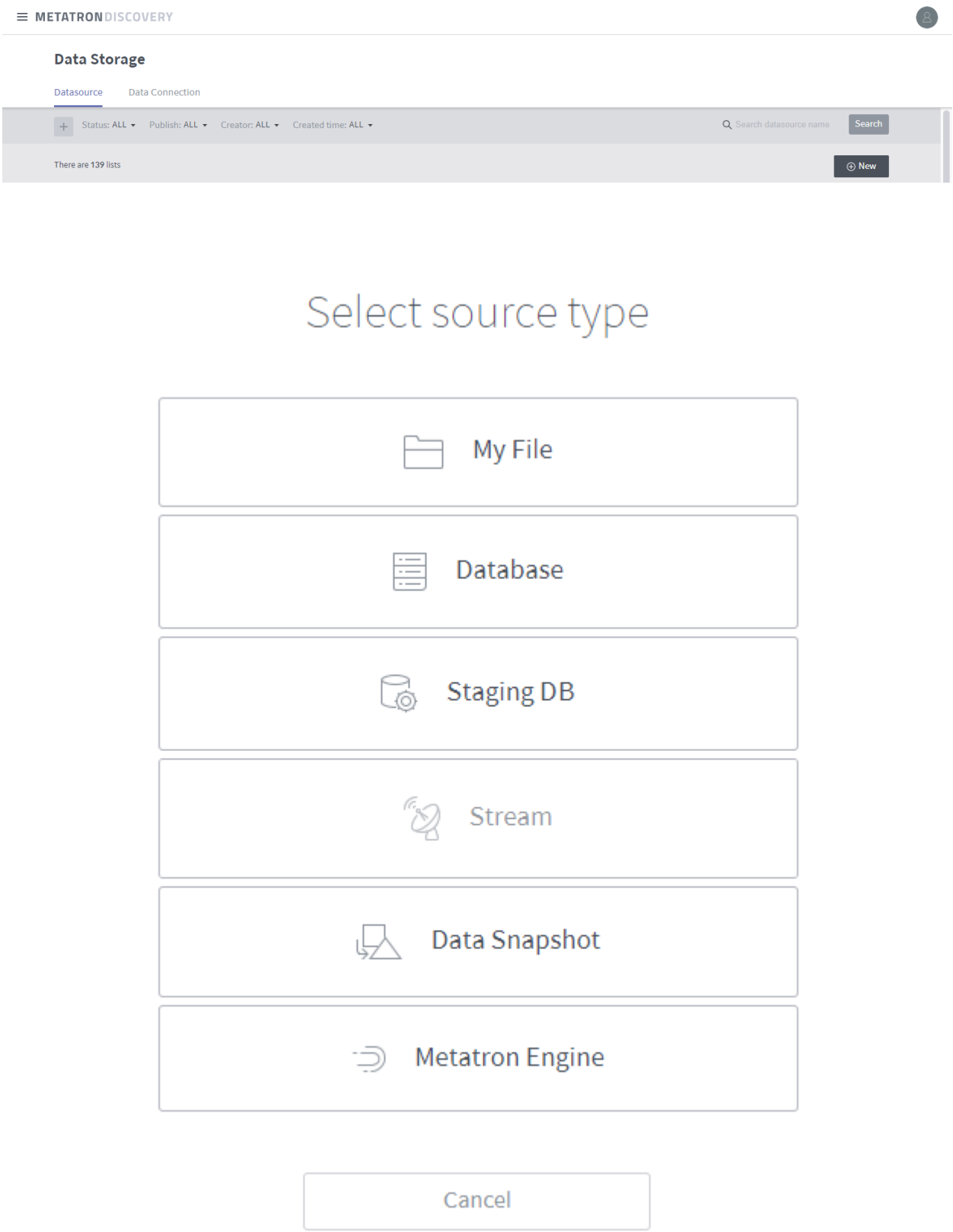
Creates a data source from a file stored on your local PC.

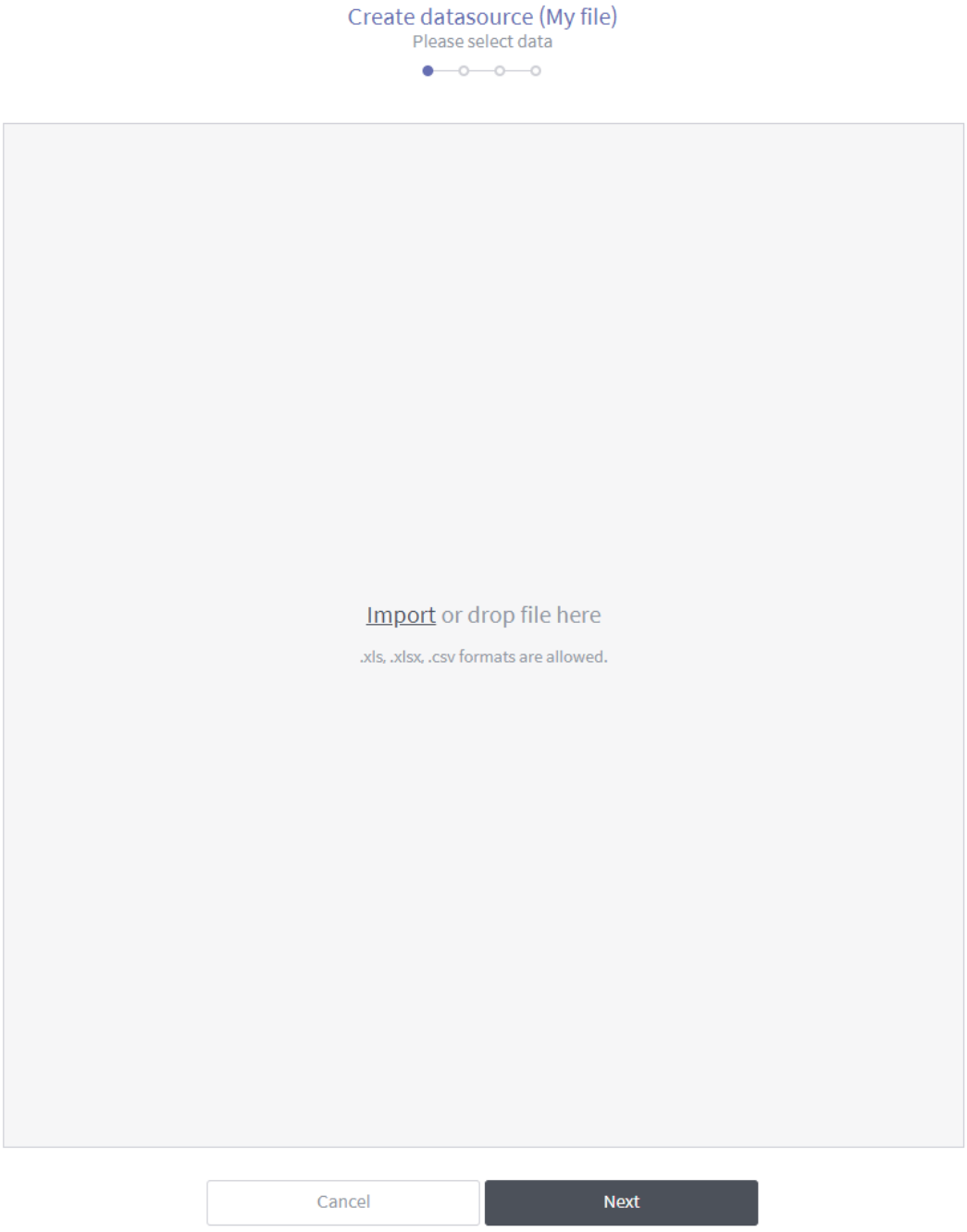
1. On the source data type selection page, select **File**.
2. Select a file to be used as a data source from your local PC. You can either click the **Import** button and select the file, or drag and drop a file to the box. Once a file is selected, click Next.
3. From the file, select the sheet to be included in the data source.


Query distribution by user (during last one week)
Polaris
관리자


Query distribution by elapsed time (during last one week)
Within 3 Seconds


Query log
1
Query date
All
Today
Last 7 days
yyyy-MM-dd hh:mm
~
yyyy-MM-dd hh:mm
Apply
2
Query type
All
3
Result
All
4
No.
Query date
Query type
User
Elapsed time
Result
5
1 2018-05-10 21:17 SUMMARY 85ms Success Detail >
2 2018-05-11 16:41 SUMMARY 78ms Success Detail >
3 2018-05-10 21:17 SEARCH 78ms Success Detail >
4 2018-05-10 21:17 SUMMARY 76ms Success Detail >
5 2018-05-11 17:30 SUMMARY 64ms Success Detail >
More


METATRONDISCOVERY
Data Storage
Datasource
Data Connection
Status: ALL
Publish: ALL
Creator: ALL
Created time: ALL
Search datasource name
Search
There are 139 lists
New
Select source type
My File
Database
Staging DB
Stream
Data Snapshot
Metatron Engine
Cancel


Create datasource (My file)
Please select data
Import or drop file here
.xls, .xlsx, .csv formats are allowed.
Cancel
Next

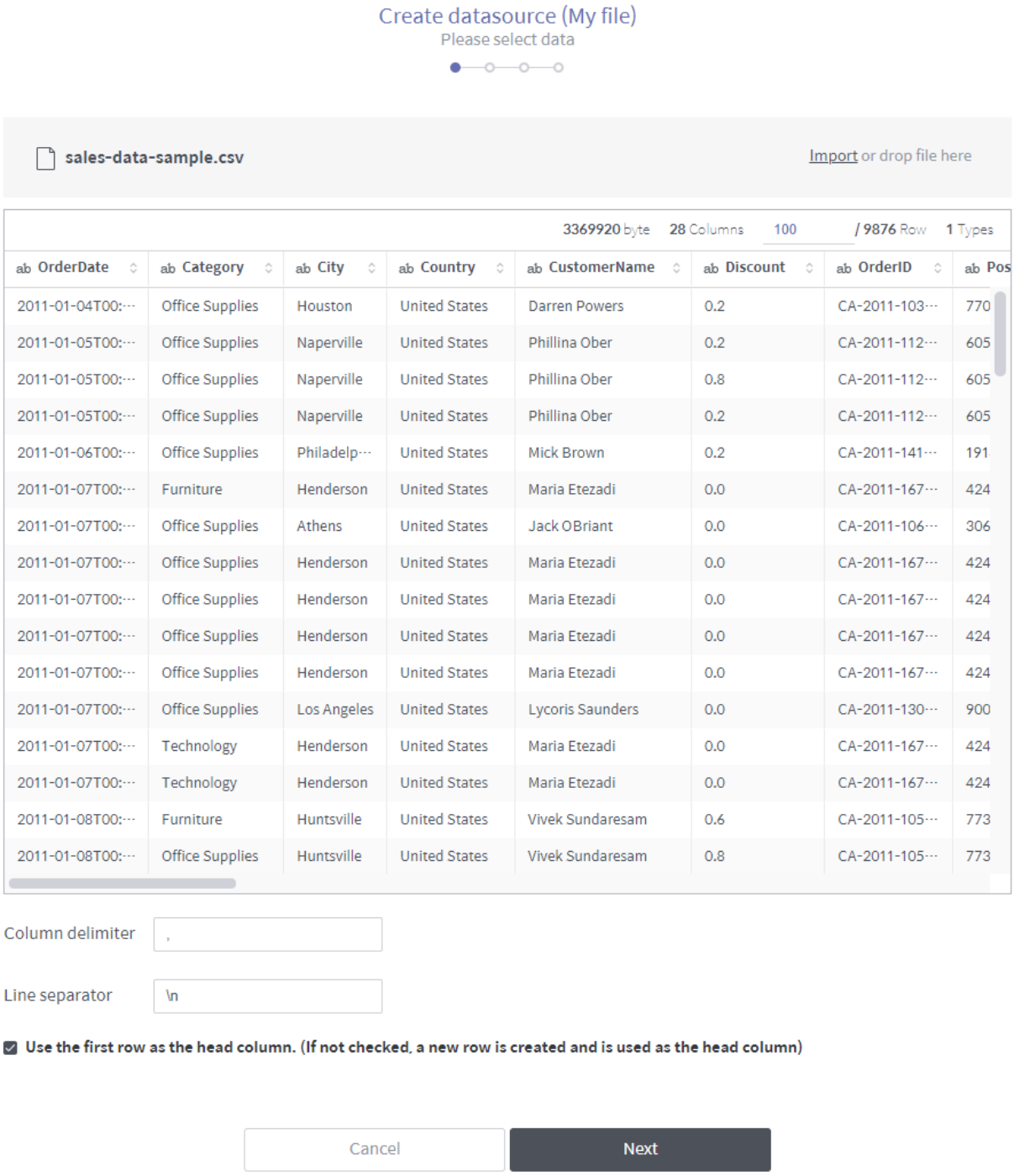
Note: If the "No preview data" message is shown in spite of there being data, check whether the **Column delimiter** and **Line Separator** have been configured correctly. In this example, the **Line Separator** must be set to "r"? the carriage return for MS Windows.

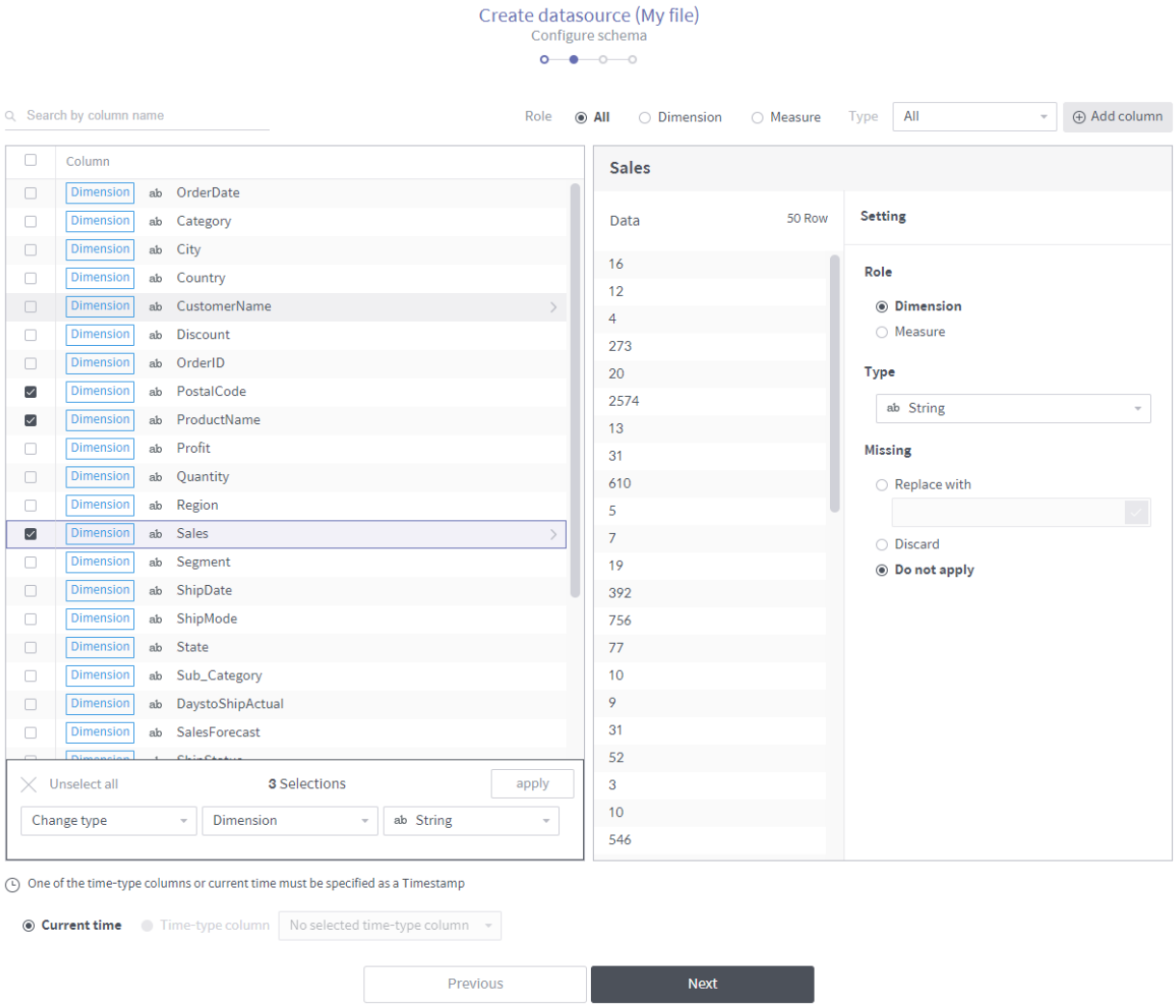
- **File name:** Name of the imported file. You can replace it with another file.
- **File sheet list:** Displays the sheets included in the imported file. Select the sheet from which you want to create a data source.
- **File sheet name:** Name of the currently selected sheet.
- **Size:** Size of the imported file.
- **Column:** Number of columns in the imported file.
- **Row:** Displayed number of rows and total number of rows in the imported file. Enter the number of rows to be displayed on the page.
- **Type:** Displays how many data types are recognized from the columns. The data type of each column can be modified later.
- **Use the first row as the head column:** Select the check box to use the first row of the file as column headers. If you don't select it, a new row is inserted as a column header row.

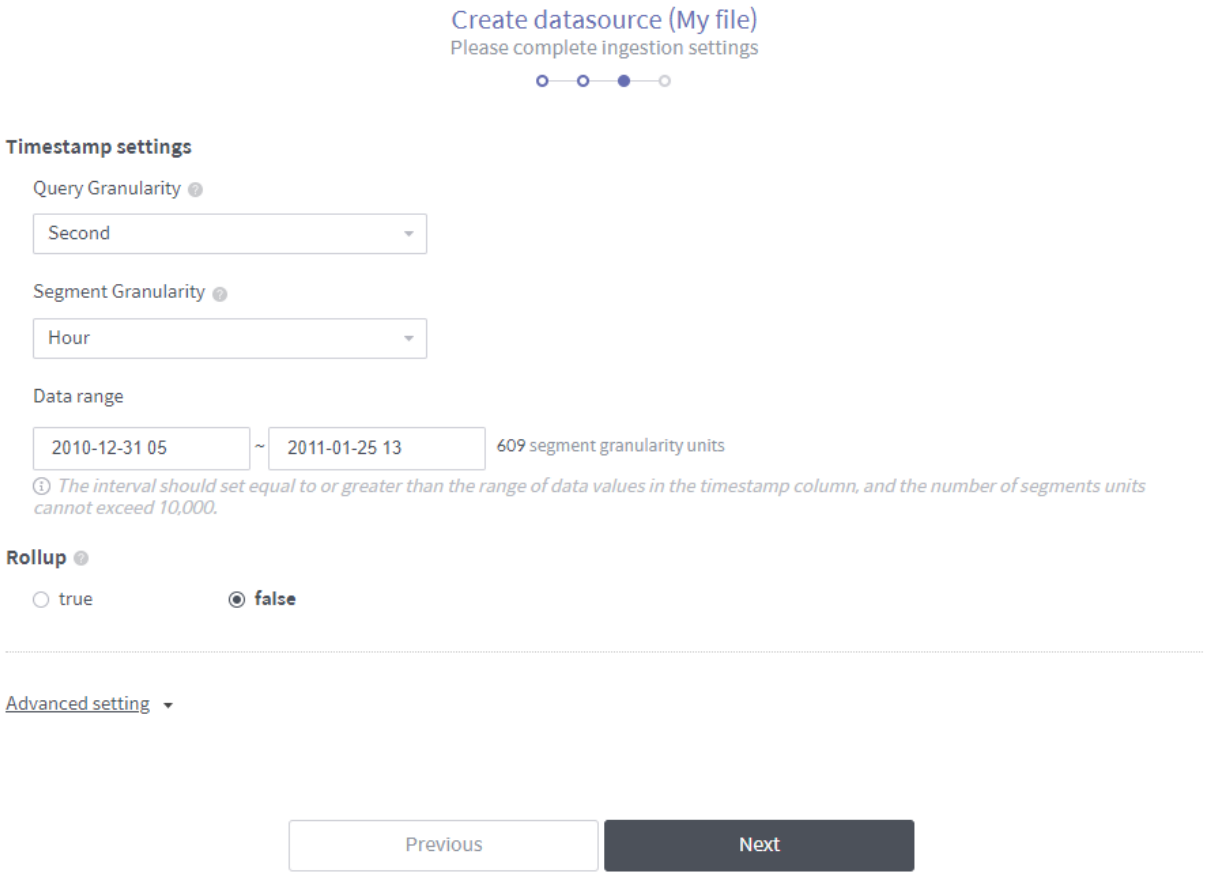
4. Configure the schema of the data source.
    - **Search by column header:** Searches the imported file for columns by name.
    - **Role:** Displays all, dimension, or measure columns from the imported file.
    - **Recommended filters:** Displays columns to which a top-priority filter is applied.
    - **Type:** Filters the columns in the imported file by field type.
    - **Column list section:** Lists columns filtered by specified criteria. Once you have selected columns, a panel appears at the bottom of the screen. After selecting your desired batch action in the panel, click **Apply** to perform the batch action on the selected columns.
    - **Individual column settings section:** This area is used to set the attributes of a column selected from the column list. **Missing** is used to set nulls in the column.
        - **Replace with:** Replaces the nulls with the value typed in.
        - **Discard:** Discards the nulls.
        - **Do not set:** Leaves the nulls as nulls. However, the nulls in the timestamp column are mandatorily discarded.
    - **Timestamp setting:** Determines how to timestamp each row. You can either designate an existing time-type column as a timestamp column, or create a new time-type column whose values are all timestamped with the current time.

      Note: Metatron Druid is a time-series engine that requires a timestamp for each row when a data source is created.

    - **:** , `Point` .
5. Configure data source ingestion and click Next.


Create datasource (My file)
Please select data
sales-data-sample.csv
Import or drop file here
3369920 byte 28 Columns 100 / 9876 Row 1 Types
ab OrderDate ab Category ab City ab Country ab CustomerName ab Discount ab OrderID ab Pos
2011-01-04T00:··· Office Supplies Houston United States Darren Powers 0.2 CA-2011-103··· 770
2011-01-05T00:··· Office Supplies Naperville United States Phillina Ober 0.2 CA-2011-112··· 605
2011-01-05T00:··· Office Supplies Naperville United States Phillina Ober 0.8 CA-2011-112··· 605
2011-01-05T00:··· Office Supplies Naperville United States Phillina Ober 0.2 CA-2011-112··· 605
2011-01-06T00:··· Office Supplies Philadelp··· United States Mick Brown 0.2 CA-2011-141··· 191
2011-01-07T00:··· Furniture Henderson United States Maria Etezadi 0.0 CA-2011-167··· 424
2011-01-07T00:··· Office Supplies Athens United States Jack OBriant 0.0 CA-2011-106··· 306
2011-01-07T00:··· Office Supplies Henderson United States Maria Etezadi 0.0 CA-2011-167··· 424
2011-01-07T00:··· Office Supplies Henderson United States Maria Etezadi 0.0 CA-2011-167··· 424
2011-01-07T00:··· Office Supplies Henderson United States Maria Etezadi 0.0 CA-2011-167··· 424
2011-01-07T00:··· Office Supplies Henderson United States Maria Etezadi 0.0 CA-2011-167··· 424
2011-01-07T00:··· Office Supplies Los Angeles United States Lycoris Saunders 0.0 CA-2011-130··· 900
2011-01-07T00:··· Technology Henderson United States Maria Etezadi 0.0 CA-2011-167··· 424
2011-01-07T00:··· Technology Henderson United States Maria Etezadi 0.0 CA-2011-167··· 424
2011-01-08T00:··· Furniture Huntsville United States Vivek Sundaresam 0.6 CA-2011-105··· 773
2011-01-08T00:··· Office Supplies Huntsville United States Vivek Sundaresam 0.8 CA-2011-105··· 773
Column delimiter ,
Line separator \n
Use the first row as the head column. (If not checked, a new row is created and is used as the head column)
Cancel
Next

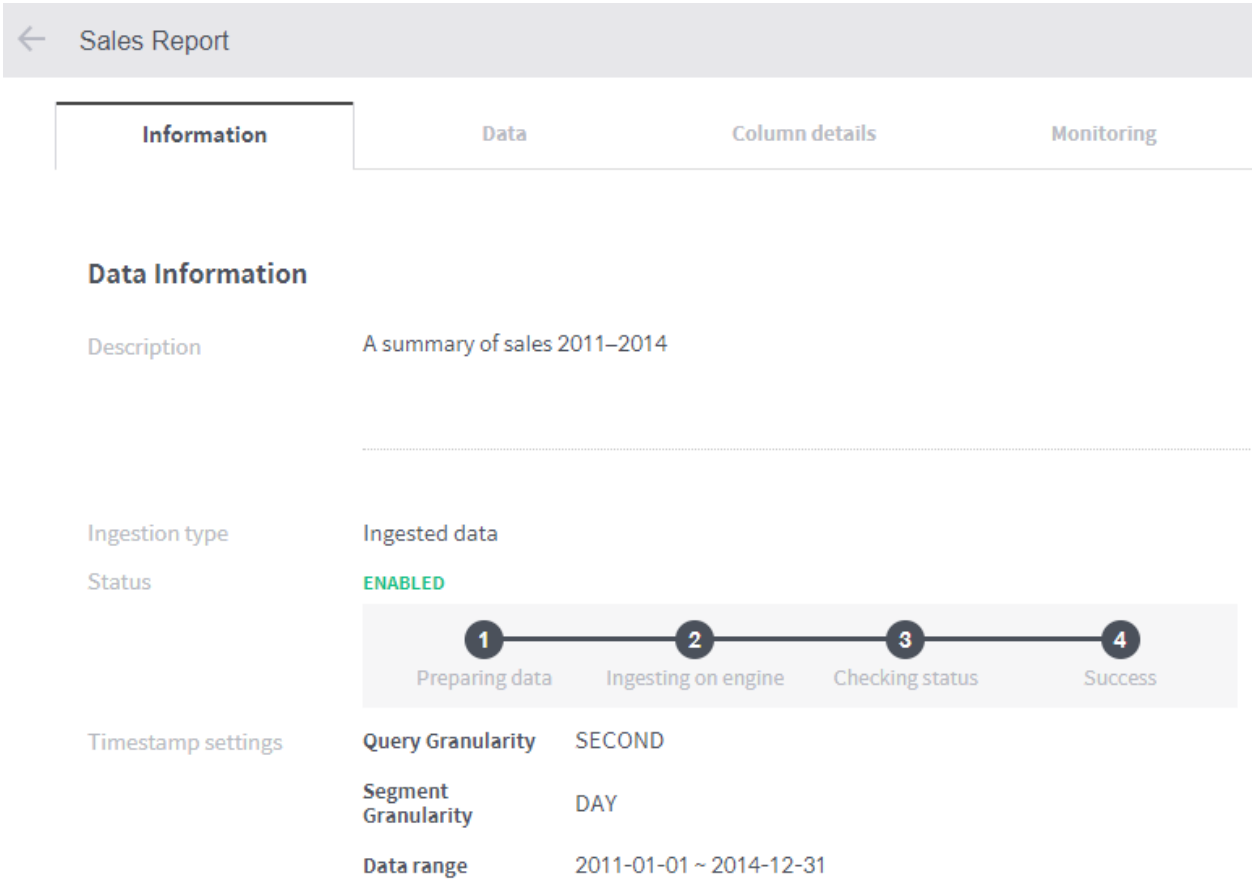

Create datasource (My file)
Configure schema
Search by column name
Role
All
Dimension
Measure
Type
All
Add column
Column
Dimension ab OrderDate
Dimension ab Category
Dimension ab City
Dimension ab Country
Dimension ab CustomerName
Dimension ab Discount
Dimension ab OrderID
Dimension ab PostalCode
Dimension ab ProductName
Dimension ab Profit
Dimension ab Quantity
Dimension ab Region
Dimension ab Sales
Dimension ab Segment
Dimension ab ShipDate
Dimension ab ShipMode
Dimension ab State
Dimension ab Sub_Category
Dimension ab DaystoShipActual
Dimension ab SalesForecast
Unselect all
3 Selections
apply
Change type
Dimension
ab String
Sales
Data
50 Row
16
12
4
273
20
2574
13
31
610
5
7
19
392
756
77
10
9
31
52
3
10
546
Setting
Role
Dimension
Measure
Type
ab String
Missing
Replace with
Discard
Do not apply
One of the time-type columns or current time must be specified as a Timestamp
Current time
Time-type column
No selected time-type column
Previous
Next


Create datasource (My file)
Please complete ingestion settings
Timestamp settings
Query Granularity
Second
Segment Granularity
Hour
Data range
2010-12-31 05
~
2011-01-25 13
609 segment granularity units
The interval should set equal to or greater than the range of data values in the timestamp column, and the number of segments units cannot exceed 10,000.
Rollup
true
false
Advanced setting
Previous
Next

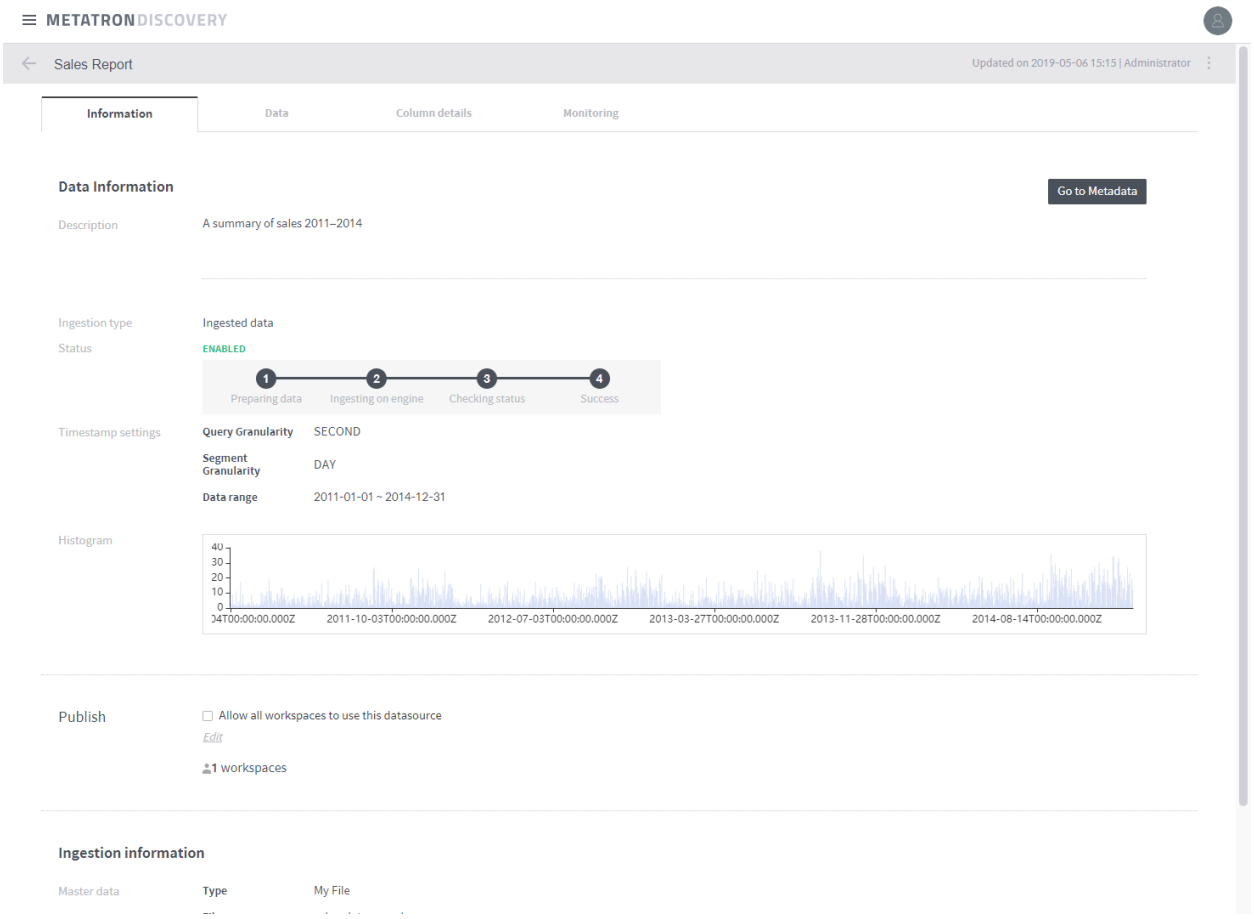
- **Segment Granularity:** In Druid, a data source is stored into multiple segments to be processed over multiple nodes in the distributed cluster environment. This granularity setting defines the time intervals into which the data source is partitioned.
- **Query Granularity:** Defines the minimum time period by which data is queried. This ensures faster returns by aggregating data per granularity interval.
- **Rollup:** "Data rollup" summarizes data based on its dimension (for details on the concept of data rollup, refer to *Data roll-up*). A summarization rule might be summing up all values in each column or applying a set of expressions such as `profit=sales=expenses`.
- **Advanced settings:** Configures how to ingest data. Type in the text box in the JSON format. For example,

```
{maxRowsInMemory : 75000,
maxOccupationInMemory : -1,
maxShardLength : -2147483648,
leaveIntermediate : false,
cleanupOnFailure : true,
overwriteFiles : false,
ignoreInvalidRows : false,
assumeTimeSorted : false}
```

6. Confirm the information about the data set from the imported file, enter the **Name** and **Description**, and click **Done** to create a data source. It may take a few seconds or minutes depending on the amount of data as the source data is ingested into the internal Metatron engine (Druid).



7. After data ingestion is complete, you can check the status. In the example below, the status is set to **ENABLED** and a histogram is displayed.



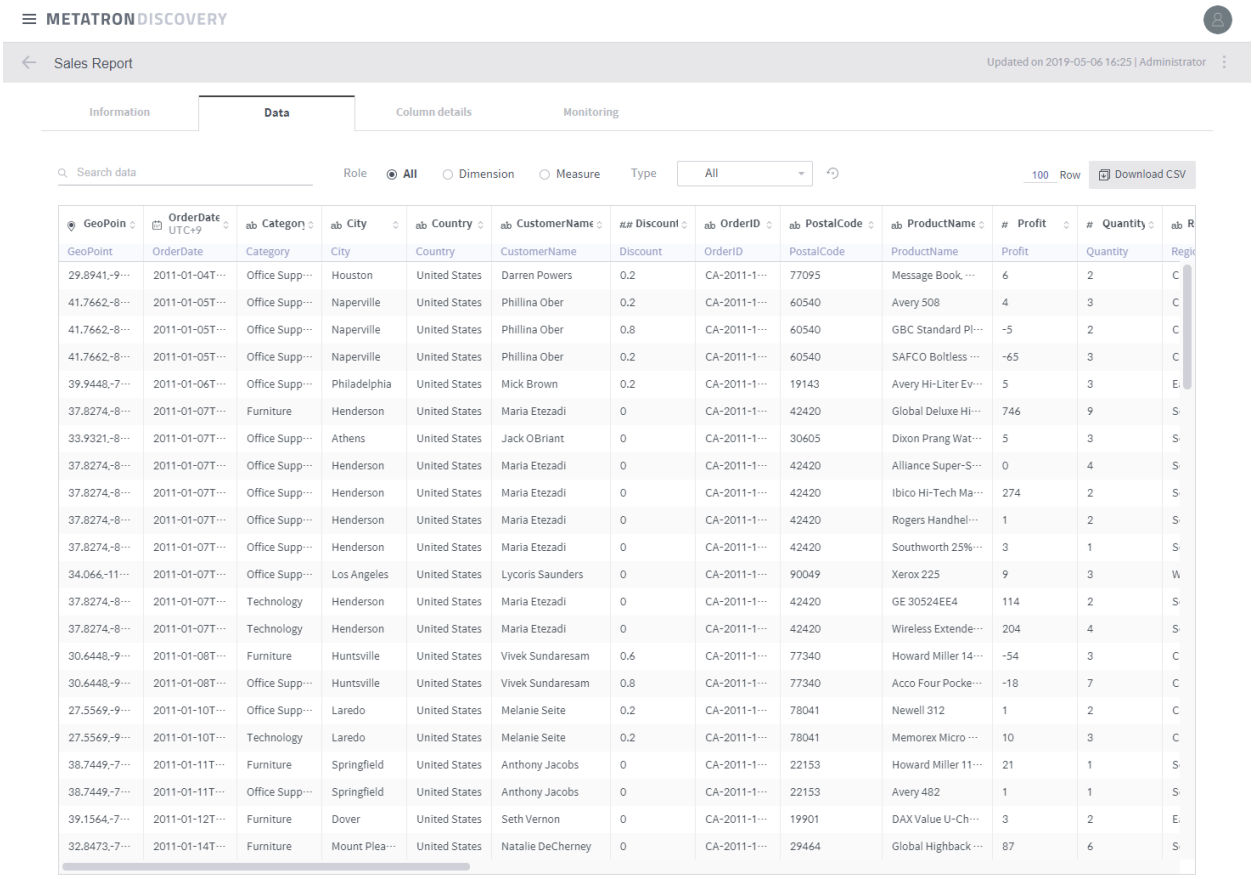
8. In the **Data** tab, you can check the ingested data in the form of a table.

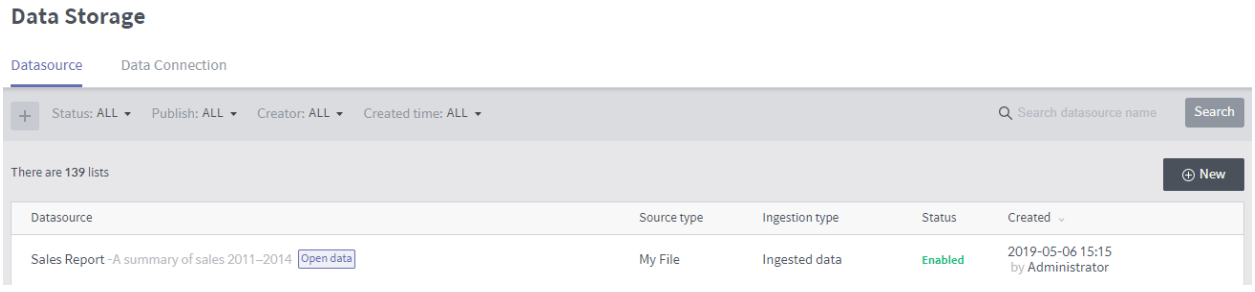
9. On the **Data Source** management home screen, you will find a newly-created data source. While data is being ingested, the status is displayed as **Disabled** as shown below; the status changes to **Enabled** once ingestion is complete. After that, you can use the data source.

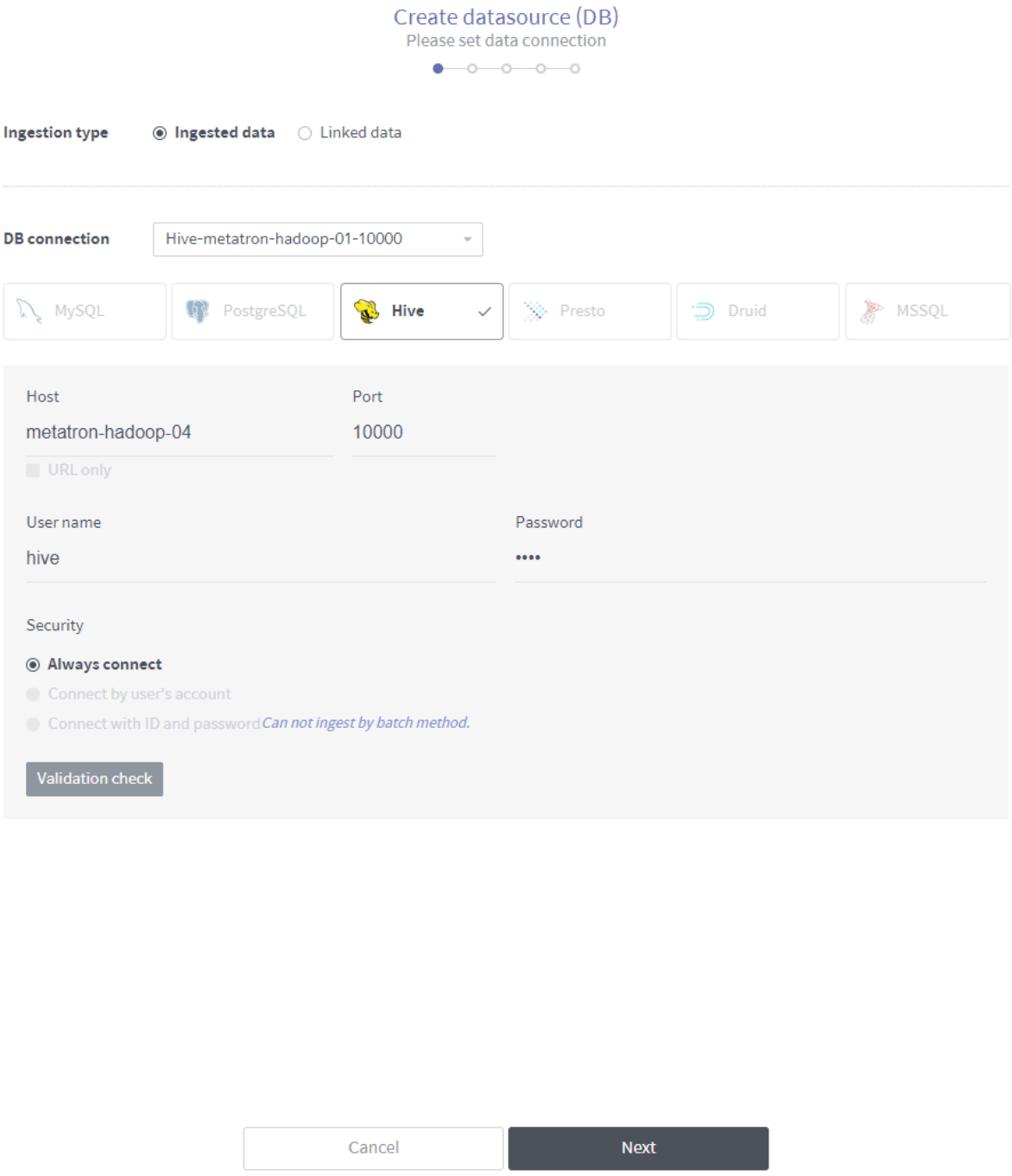
Create a data source from a database

Creates a data source from an external database.

1. On the source data type selection page, select **Database**.
2. Enter the information to connect the database.
    - **Ingestion type:** Select how to ingest data into the data source.
        - **Ingested data:** Displays data sources that contain data ingested into the Metatron storage.


METATRONDISCOVERY
Sales Report
Updated on 2019-05-06 16:25 | Administrator
Information
Data
Column details
Monitoring
Search data
Role
All
Dimension
Measure
Type
All
100 Row
Download CSV
GeoPoin
OrderDate UTC+9
Categor
City
Country
CustomerName
Discount
OrderID
PostalCode
ProductName
Profit
Quantity
GeoPoint
OrderDate
Category
City
Country
CustomerName
Discount
OrderID
PostalCode
ProductName
Profit
Quantity
29.8941,-9··· 2011-01-04T··· Office Supp··· Houston United States Darren Powers 0.2 CA-2011-1··· 77095 Message Book ··· 6 2
41.7662,-8··· 2011-01-05T··· Office Supp··· Naperville United States Phillina Ober 0.2 CA-2011-1··· 60540 Avery 508 4 3
41.7662,-8··· 2011-01-05T··· Office Supp··· Naperville United States Phillina Ober 0.8 CA-2011-1··· 60540 GBC Standard Pl··· -5 2
41.7662,-8··· 2011-01-05T··· Office Supp··· Naperville United States Phillina Ober 0.2 CA-2011-1··· 60540 SAFCO Boltless ··· -65 3
39.9448,-7··· 2011-01-06T··· Office Supp··· Philadelphia United States Mick Brown 0.2 CA-2011-1··· 19143 Avery Hi-Liter Ev··· 5 3
37.8274,-8··· 2011-01-07T··· Furniture Henderson United States Maria Etezadi 0 CA-2011-1··· 42420 Global Deluxe Hi··· 746 9
33.9321,-8··· 2011-01-07T··· Office Supp··· Athens United States Jack OBriant 0 CA-2011-1··· 30605 Dixon Prang Wat··· 5 3
37.8274,-8··· 2011-01-07T··· Office Supp··· Henderson United States Maria Etezadi 0 CA-2011-1··· 42420 Alliance Super-S··· 0 4
37.8274,-8··· 2011-01-07T··· Office Supp··· Henderson United States Maria Etezadi 0 CA-2011-1··· 42420 Ibico Hi-Tech Ma··· 274 2
37.8274,-8··· 2011-01-07T··· Office Supp··· Henderson United States Maria Etezadi 0 CA-2011-1··· 42420 Rogers Handhel··· 1 2
37.8274,-8··· 2011-01-07T··· Office Supp··· Henderson United States Maria Etezadi 0 CA-2011-1··· 42420 Southworth 25%··· 3 1
34.066,-11··· 2011-01-07T··· Office Supp··· Los Angeles United States Lycoris Saunders 0 CA-2011-1··· 90049 Xerox 225 9 3
37.8274,-8··· 2011-01-07T··· Technology Henderson United States Maria Etezadi 0 CA-2011-1··· 42420 GE 30524EE4 114 2
37.8274,-8··· 2011-01-07T··· Technology Henderson United States Maria Etezadi 0 CA-2011-1··· 42420 Wireless Extende··· 204 4
30.6448,-9··· 2011-01-08T··· Furniture Huntsville United States Vivek Sundaresam 0.6 CA-2011-1··· 77340 Howard Miller 14··· -54 3
30.6448,-9··· 2011-01-08T··· Office Supp··· Huntsville United States Vivek Sundaresam 0.8 CA-2011-1··· 77340 Acco Four Pocke··· -18 7
27.5569,-9··· 2011-01-10T··· Office Supp··· Laredo United States Melanie Seite 0.2 CA-2011-1··· 78041 Newell 312 1 2
27.5569,-9··· 2011-01-10T··· Technology Laredo United States Melanie Seite 0.2 CA-2011-1··· 78041 Memorex Micro ··· 10 3
38.7449,-7··· 2011-01-11T··· Furniture Springfield United States Anthony Jacobs 0 CA-2011-1··· 22153 Howard Miller 11··· 21 1
38.7449,-7··· 2011-01-11T··· Office Supp··· Springfield United States Anthony Jacobs 0 CA-2011-1··· 22153 Avery 482 1 1
39.1564,-7··· 2011-01-12T··· Furniture Dover United States Seth Vernon 0 CA-2011-1··· 19901 DAX Value U-Ch··· 3 2
32.8473,-7··· 2011-01-14T··· Furniture Mount Plea··· United States Natalie DeCherney 0 CA-2011-1··· 29464 Global Highback··· 87 6


Data Storage
Datasource
Data Connection
Status: ALL
Publish: ALL
Creator: ALL
Created time: ALL
Search datasource name
Search
There are 139 lists
New
Datasource
Source type
Ingestion type
Status
Created
Sales Report -A summary of sales 2011–2014 Open data
My File
Ingested data
Enabled
2019-05-06 15:15
by Administrator


Create datasource (DB)
Please set data connection
Ingestion type
Ingested data
Linked data
DB connection
Hive-metatron-hadoop-01-10000
MySQL
PostgreSQL
Hive
Presto
Druid
MSSQL
Host
metatron-hadoop-04
Port
10000
URL only
User name
hive
Password
••••
Security
Always connect
Connect by user's account
Connect with ID and password
Can not ingest by batch method.
Validation check
Cancel
Next

  - **Linked data:** Displays data sources that load data from linked databases whenever necessary.

- **Load a data connection:** Automatically loads access information for a database that is already registered as a data connection. However, you must verify the connection by clicking the **Validation check** button.
- **DB type:** Select the type of the database to be connected.
- **Host:** Enter the hostname to connect to the database.
- **Port:** Enter the port to connect to the database.
- **User name:** Enter the username of the database.
- **Password:** Enter the password of the database.
- **Validation check:** Once you fill out all fields, the Test button becomes active. Click on it to verify if the connection is valid: The validity of the connection appears below the button.

3. Select data. You can either select a table from the connected database, or write a query yourself.
   - **Table:** Select a database and a table to display the table's data. Once the data being ingested has been displayed, confirm the data and click **Next**.
   - **Query:** Write a query to import the data you want, and click **Run** to display the data in the lower section. Confirm the data and click **Next**.
4. The rest of the process is identical to *Create a data source from a file*. However, when creating a data source from a database, you must configure additional **ingestion settings** as follows.
   - **Ingest once:** Ingest the data currently stored in the database only this once. When selecting the **Limited record count**, you can specify how many rows are to be ingested from the first row.
   - **Ingest periodically:** Saves data on a regular basis.

Create a data source from a staging database

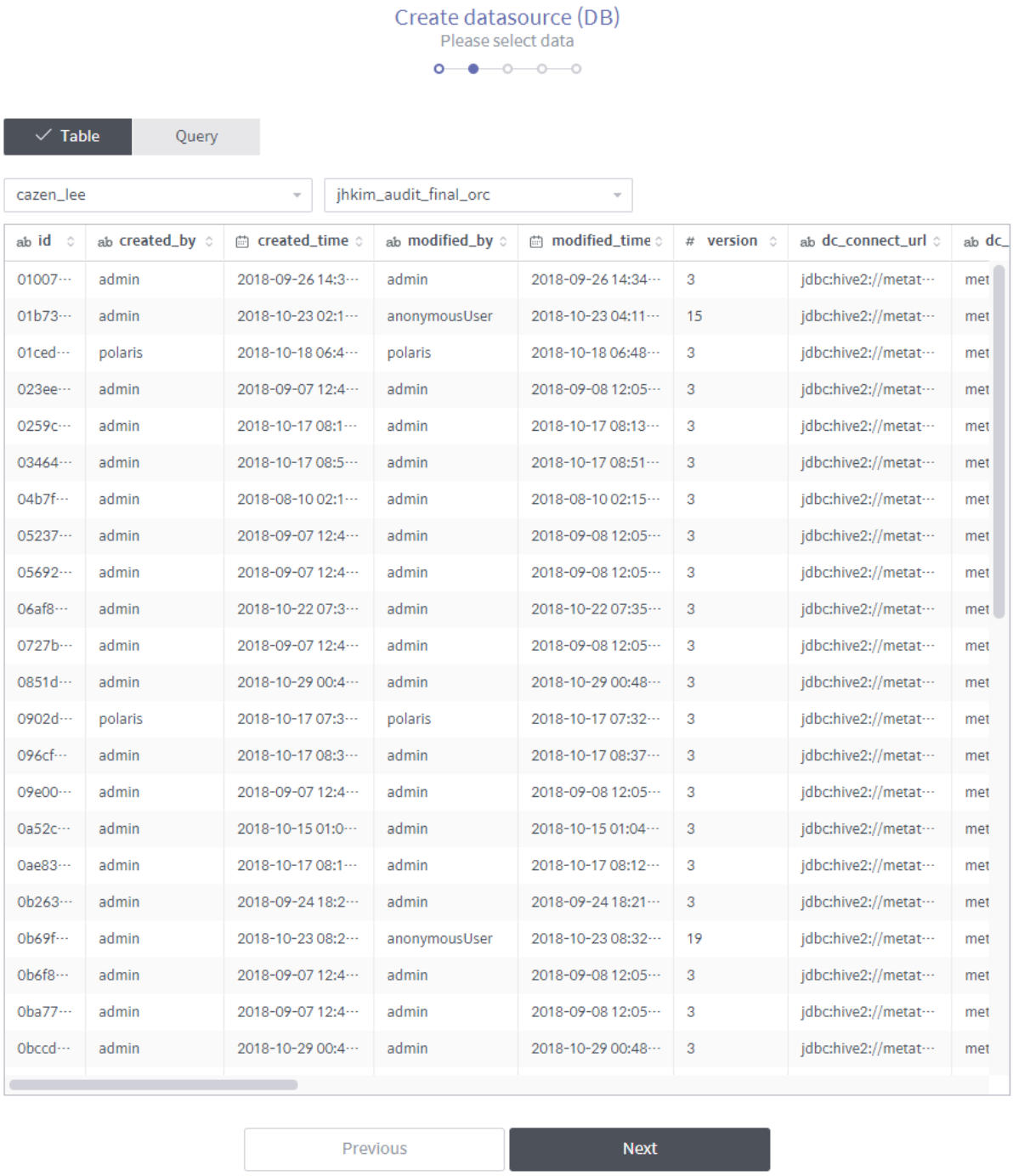
Creates a data source from Metatron's internal Hive database.

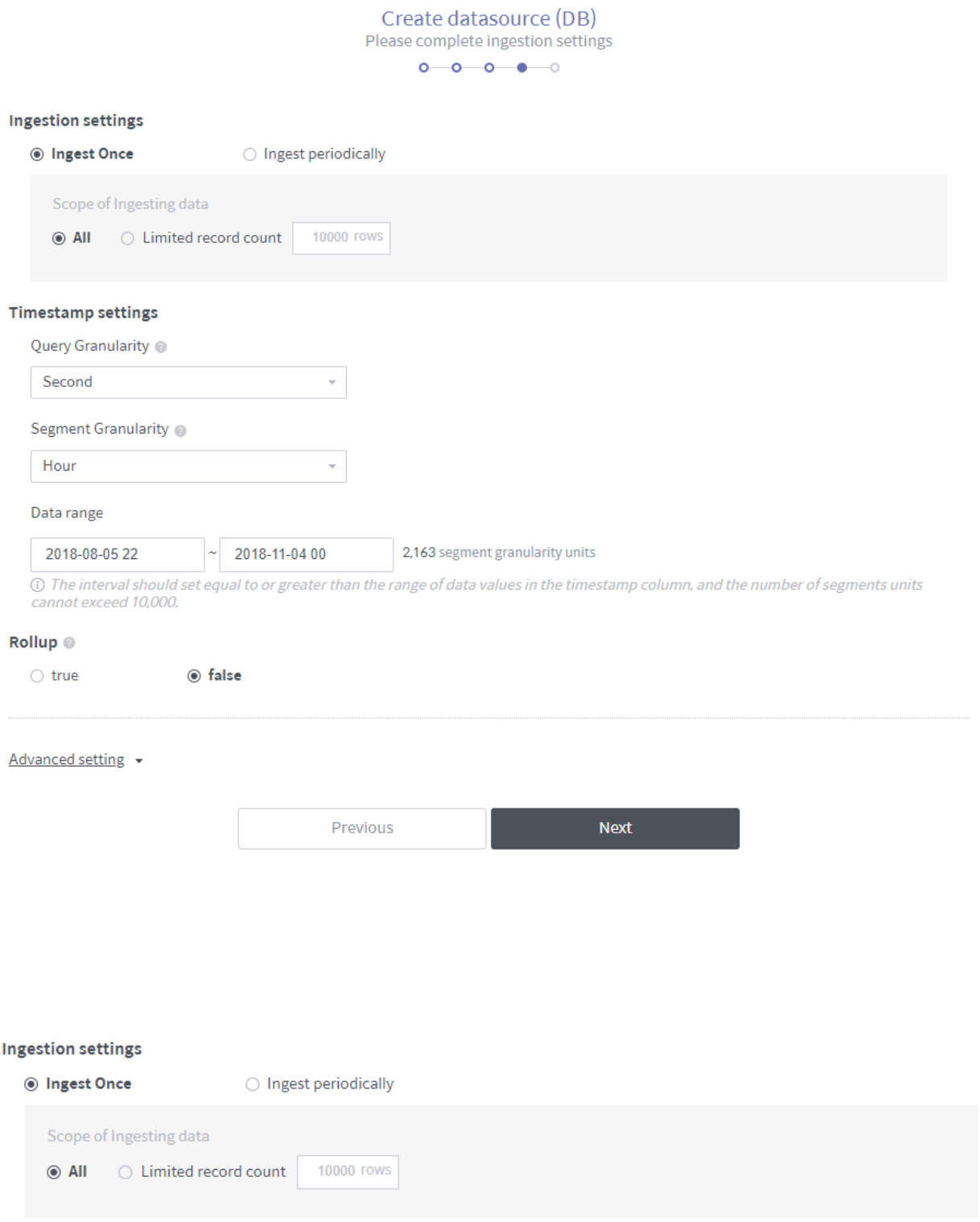
1. On the source data type selection page, select **Staging DB**.
2. Once you select the database and its table to connect, the data is displayed.
3. The rest of the process is identical to *Create a data source from a database*.

Add a data source with the Metatron engine

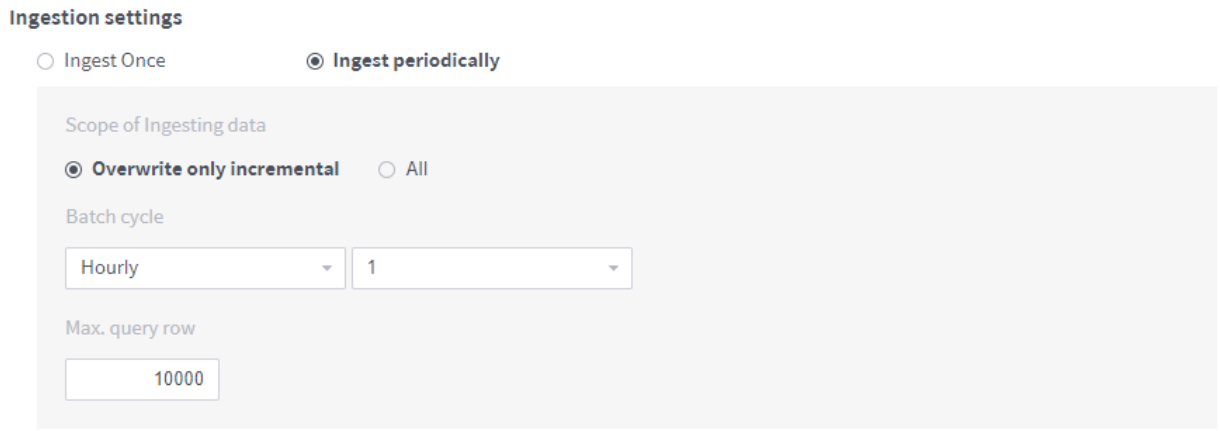
Migrates a data source stored in a previous Metatron version.

1. On the source data type selection page, select **Metatron Engine**.
2. When data sources created in a previous version of Metatron are listed on the left as shown below, select the check boxes of the data sources you want to migrate to the current version.


Create datasource (DB)
Please select data
Table
Query
cazen_lee
jhkim_audit_final_orc
id
created_by
created_time
modified_by
modified_time
version
dc_connect_url
dc_
01007··· admin 2018-09-26 14:3··· admin 2018-09-26 14:34··· 3 jdbc:hive2://metat··· met
01b73··· admin 2018-10-23 02:1··· anonymousUser 2018-10-23 04:11··· 15 jdbc:hive2://metat··· met
01ced··· polaris 2018-10-18 06:4··· polaris 2018-10-18 06:48··· 3 jdbc:hive2://metat··· met
023ee··· admin 2018-09-07 12:4··· admin 2018-09-08 12:05··· 3 jdbc:hive2://metat··· met
0259c··· admin 2018-10-17 08:1··· admin 2018-10-17 08:13··· 3 jdbc:hive2://metat··· met
03464··· admin 2018-10-17 08:5··· admin 2018-10-17 08:51··· 3 jdbc:hive2://metat··· met
04b7f··· admin 2018-08-10 02:1··· admin 2018-08-10 02:15··· 3 jdbc:hive2://metat··· met
05237··· admin 2018-09-07 12:4··· admin 2018-09-08 12:05··· 3 jdbc:hive2://metat··· met
05692··· admin 2018-09-07 12:4··· admin 2018-09-08 12:05··· 3 jdbc:hive2://metat··· met
06af8··· admin 2018-10-22 07:3··· admin 2018-10-22 07:35··· 3 jdbc:hive2://metat··· met
0727b··· admin 2018-09-07 12:4··· admin 2018-09-08 12:05··· 3 jdbc:hive2://metat··· met
0851d··· admin 2018-10-29 00:4··· admin 2018-10-29 00:48··· 3 jdbc:hive2://metat··· met
0902d··· polaris 2018-10-17 07:3··· polaris 2018-10-17 07:32··· 3 jdbc:hive2://metat··· met
096cf··· admin 2018-10-17 08:3··· admin 2018-10-17 08:37··· 3 jdbc:hive2://metat··· met
09e00··· admin 2018-09-07 12:4··· admin 2018-09-08 12:05··· 3 jdbc:hive2://metat··· met
0a52c··· admin 2018-10-15 01:0··· admin 2018-10-15 01:04··· 3 jdbc:hive2://metat··· met
0ae83··· admin 2018-10-17 08:1··· admin 2018-10-17 08:12··· 3 jdbc:hive2://metat··· met
0b263··· admin 2018-09-24 18:2··· admin 2018-09-24 18:21··· 3 jdbc:hive2://metat··· met
0b69f··· admin 2018-10-23 08:2··· anonymousUser 2018-10-23 08:32··· 19 jdbc:hive2://metat··· met
0b6f8··· admin 2018-09-07 12:4··· admin 2018-09-08 12:05··· 3 jdbc:hive2://metat··· met
0ba77··· admin 2018-09-07 12:4··· admin 2018-09-08 12:05··· 3 jdbc:hive2://metat··· met
0bccd··· admin 2018-10-29 00:4··· admin 2018-10-29 00:48··· 3 jdbc:hive2://metat··· met
Previous
Next


Create datasource (DB)
Please complete ingestion settings
Ingestion settings
Ingest Once
Ingest periodically
Scope of Ingesting data
All
Limited record count
10000 rows
Timestamp settings
Query Granularity
Second
Segment Granularity
Hour
Data range
2018-08-05 22
~
2018-11-04 00
2,163 segment granularity units
The interval should set equal to or greater than the range of data values in the timestamp column, and the number of segments units cannot exceed 10,000.
Rollup
true
false
Advanced setting
Previous
Next

Ingestion settings
Ingest Once
Ingest periodically
Scope of Ingesting data
All
Limited record count
10000 rows



3. Click **Done** to migrate the selected data sources.

3.2 Data Connection

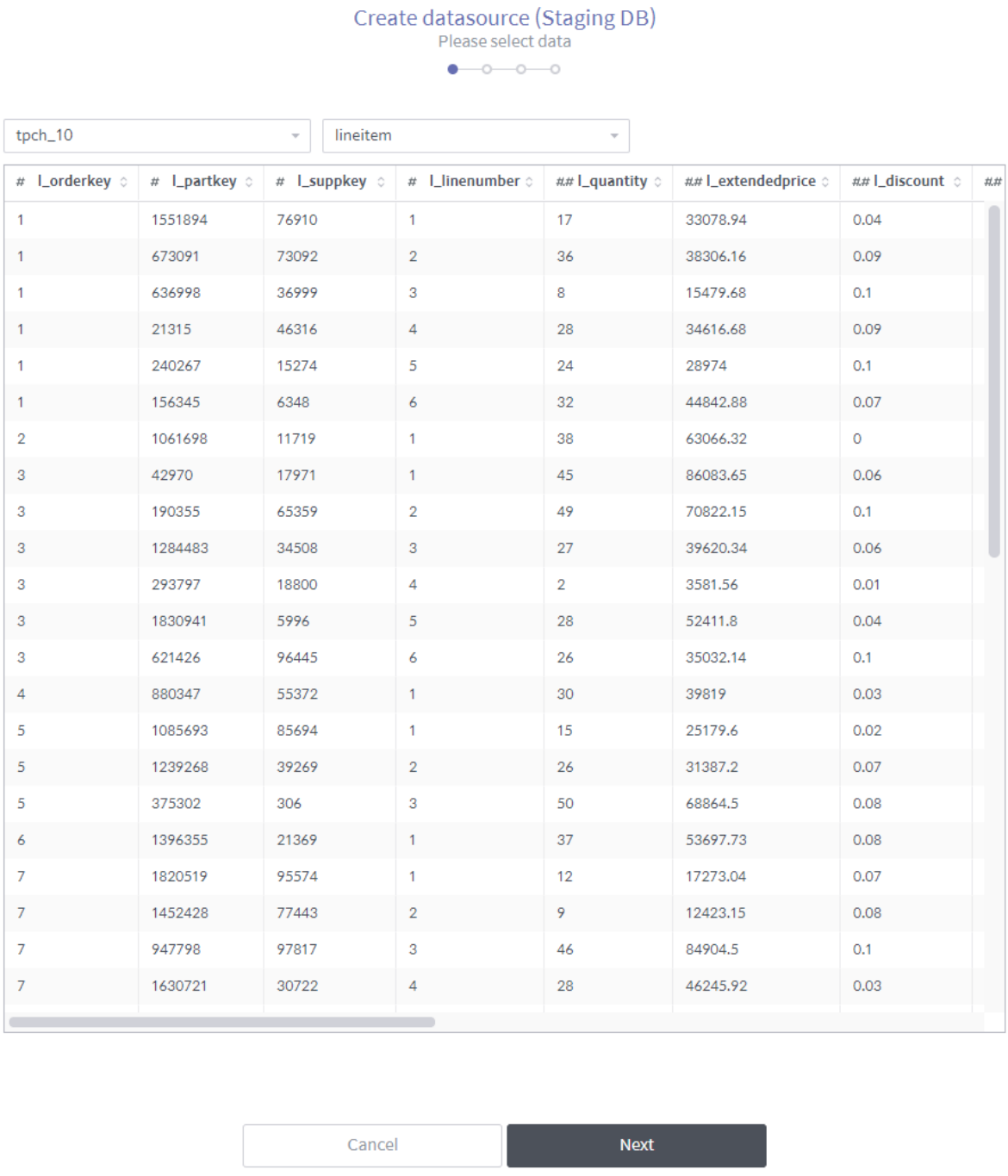
Metatron Discovery can connect to an external database directly. To connect to an external database, you must create and manage a data connection containing the access information to that database. By registering such a data connection, you don't need to enter the access information each time you connect to the same database.

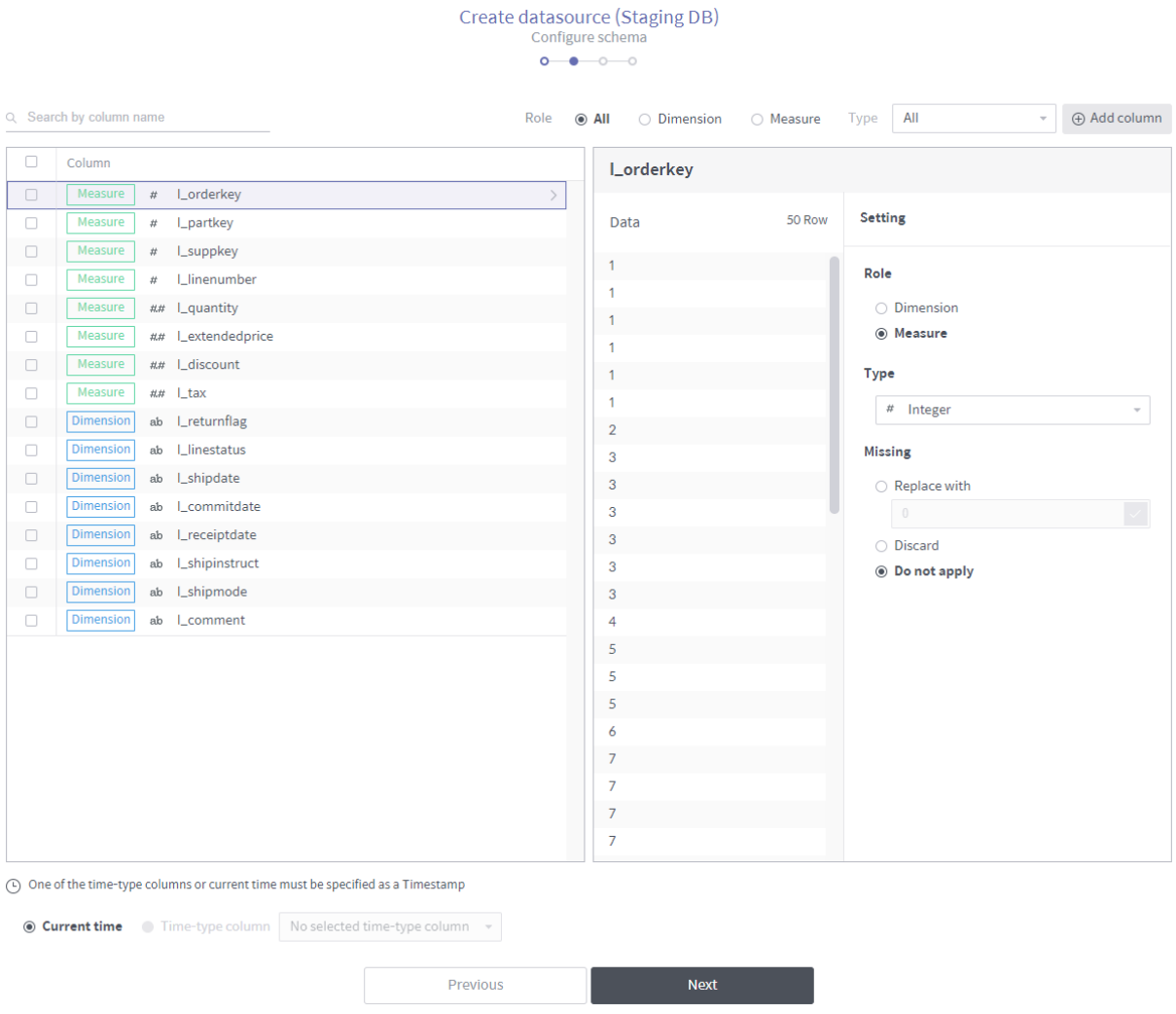
The Data Connection menu can be accessed under **MANAGEMENT > Data Storage > Data Connection** on the left-hand panel of the main screen.

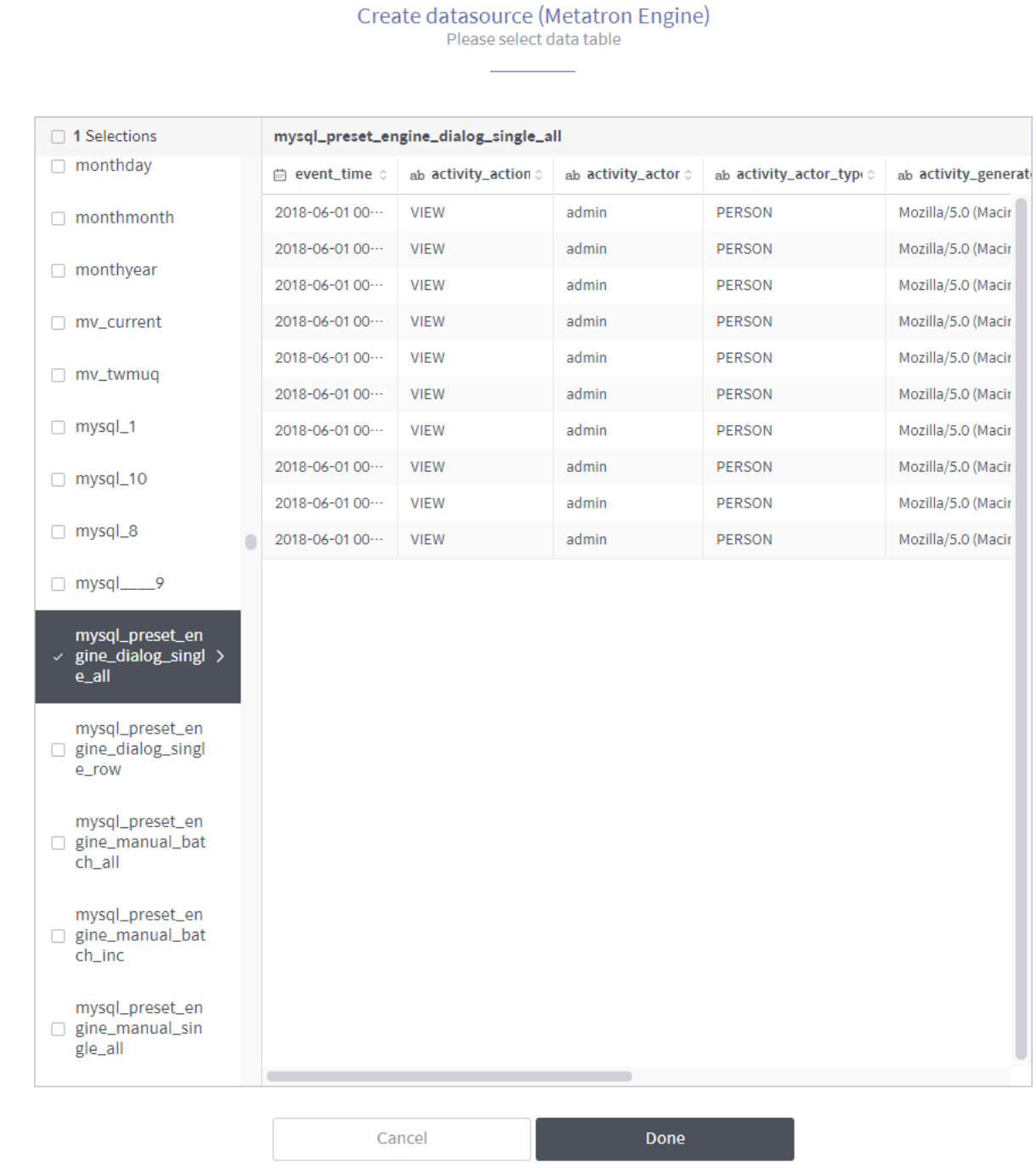
3.2.1 Data connection management home

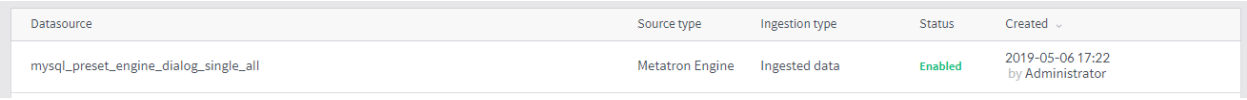
On the **Data Connection** home page, you can create, edit and view database connections.

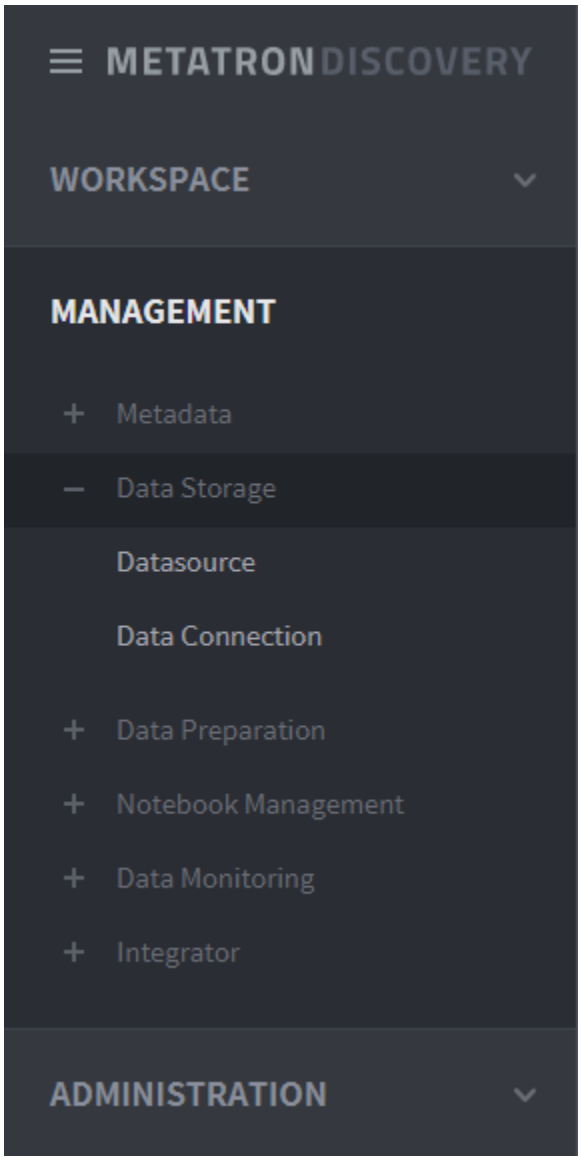
- **Publish:** Filter the data connection list by public workspace.
- **Creator:** Filter the data connection list by creator.
- **DB type:** Filter the data connection list by database type (MySQL, PostgreSQL, Hive, or Presto).
- **Security:** Filter the data connection list by security type (Always connect, connect by user's account, or connect with ID and password).
- **Created time:** Filter the data connection list by time of creation (Today, Last 7 days, or Between).
- **Search:** Search the data connection list by data connection name.
- **Number of data connections:** Displays how many data connections are returned in the list.
- **New:** Click on it to create a new data connection.
- **Delete:** Hover the mouse over a data connection to display a recycle bin icon. Click the icon to delete the data connection.

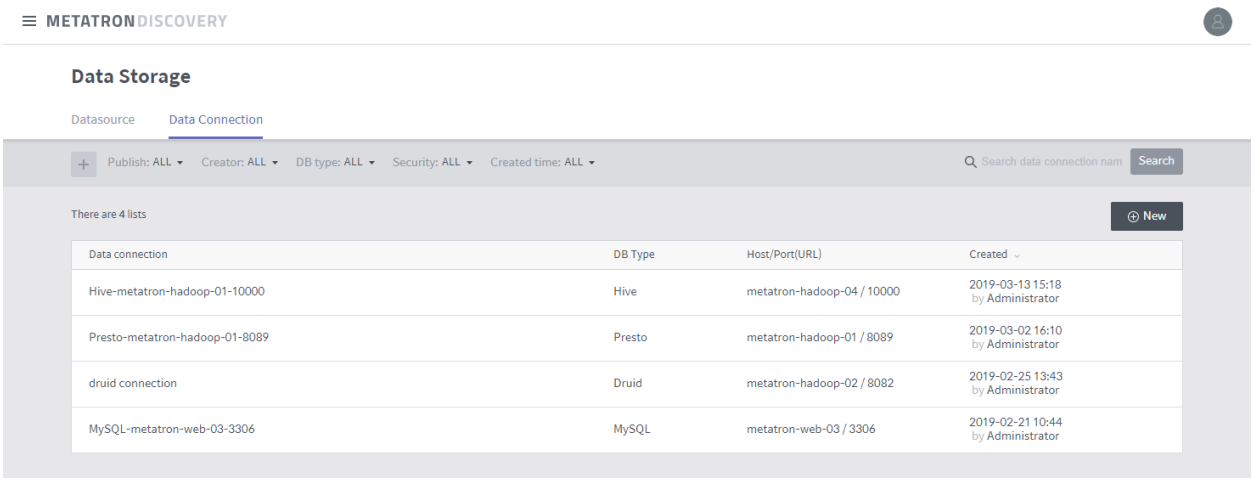

Create datasource (Staging DB)
Please select data
tpch_10
lineitem
l_orderkey
l_partkey
l_suppkey
l_linenumber
l_quantity
l_extendedprice
l_discount
1 1551894 76910 1 17 33078.94 0.04
1 673091 73092 2 36 38306.16 0.09
1 636998 36999 3 8 15479.68 0.1
1 21315 46316 4 28 34616.68 0.09
1 240267 15274 5 24 28974 0.1
1 156345 6348 6 32 44842.88 0.07
2 1061698 11719 1 38 63066.32 0
3 42970 17971 1 45 86083.65 0.06
3 190355 65359 2 49 70822.15 0.1
3 1284483 34508 3 27 39620.34 0.06
3 293797 18800 4 2 3581.56 0.01
3 1830941 5996 5 28 52411.8 0.04
3 621426 96445 6 26 35032.14 0.1
4 880347 55372 1 30 39819 0.03
5 1085693 85694 1 15 25179.6 0.02
5 1239268 39269 2 26 31387.2 0.07
5 375302 306 3 50 68864.5 0.08
6 1396355 21369 1 37 53697.73 0.08
7 1820519 95574 1 12 17273.04 0.07
7 1452428 77443 2 9 12423.15 0.08
7 947798 97817 3 46 84904.5 0.1
7 1630721 30722 4 28 46245.92 0.03
Cancel
Next


Create datasource (Staging DB)
Configure schema
Search by column name
Role
All
Dimension
Measure
Type
All
Add column
Column
Measure # l_orderkey
Measure # l_partkey
Measure # l_suppkey
Measure # l_linenumber
Measure ## l_quantity
Measure ## l_extendedprice
Measure ## l_discount
Measure ## l_tax
Dimension ab l_returnflag
Dimension ab l_linestatus
Dimension ab l_shipdate
Dimension ab l_commitdate
Dimension ab l_receiptdate
Dimension ab l_shipinstruct
Dimension ab l_shipmode
Dimension ab l_comment
l_orderkey
Data
50 Row
1
1
1
1
1
1
2
3
3
3
3
3
3
3
4
5
5
5
6
7
7
7
7
Setting
Role
Dimension
Measure
Type
Integer
Missing
Replace with
0
Discard
Do not apply
One of the time-type columns or current time must be specified as a Timestamp
Current time
Time-type column
No selected time-type column
Previous
Next


Create datasource (Metatron Engine)
Please select data table
1 Selections
monthday
monthmonth
monthyear
mv_current
mv_twmuq
mysql_1
mysql_10
mysql_8
mysql____9
mysql_preset_engine_dialog_single_all
mysql_preset_engine_dialog_single_row
mysql_preset_engine_manual_batch_all
mysql_preset_engine_manual_batch_inc
mysql_preset_engine_manual_single_all
mysql_preset_engine_dialog_single_all
event_time
activity_action
activity_actor
activity_actor_type
activity_generat
2018-06-01 00···
VIEW
admin
PERSON
Mozilla/5.0 (Macir
Cancel
Done


Datasource
Source type
Ingestion type
Status
Created
mysql_preset_engine_dialog_single_all
Metatron Engine
Ingested data
Enabled
2019-05-06 17:22
by Administrator


METATRONDISCOVERY
WORKSPACE
MANAGEMENT
Metadata
Data Storage
Datasource
Data Connection
Data Preparation
Notebook Management
Data Monitoring
Integrator
ADMINISTRATION


METATRONDISCOVERY
Data Storage
Datasource
Data Connection
Publish: ALL
Creator: ALL
DB type: ALL
Security: ALL
Created time: ALL
Search data connection nam
Search
There are 4 lists
New
Data connection
DB Type
Host/Port(URL)
Created
Hive-metatron-hadoop-01-10000
Hive
metatron-hadoop-04 / 10000
2019-03-13 15:18 by Administrator
Presto-metatron-hadoop-01-8089
Presto
metatron-hadoop-01 / 8089
2019-03-02 16:10 by Administrator
druid connection
Druid
metatron-hadoop-02 / 8082
2019-02-25 13:43 by Administrator
MySQL-metatron-web-03-3306
MySQL
metatron-web-03 / 3306
2019-02-21 10:44 by Administrator

3.2.2 Create a data connection

On the **Create data connection** screen, enter the required information to create a connection.

- **DB type:** Four database types are currently supported. (MySQL, PostgreSQL Hive, Presto)
- **Host:** Enter the hostname to connect to the database.
- **Port:** Enter the port to connect to the database.
- **URL only:** Enter a database URL instead of a host and port.
- **User name:** Enter the username of the database.
- **Password:** Enter the password of the database.
- **Security:** Set the type of security to be applied while using the data connection.
  - **Always connect:** Logs in using the account information the user has entered to create a new data connection.
  - **Connect by user's account:** Logs in using the account information registered in Metatron Discovery.
  - **Connect with ID and password:** Requires to enter the account information every time the data connection is used.
- **Validation check:** Checks whether the connection information entered is valid; the result is shown next to the button. The validity of the connection appears below the button.
- **Advanced settings:** You can add a custom property key and value as options.
- **Publish:** Set which workspaces have access to the data connection.
  - **Allow all workspaces to use this data connection:** Select this check box to make the data connection available in all workspaces.
  - **Edit:** Used to allow specific workspaces to access the data connection. This button will disappear if the data connection is set as open data.
  - **Number of shared workspaces:** Displays how many workspaces have access to the data connection.

3.3 Data Monitoring

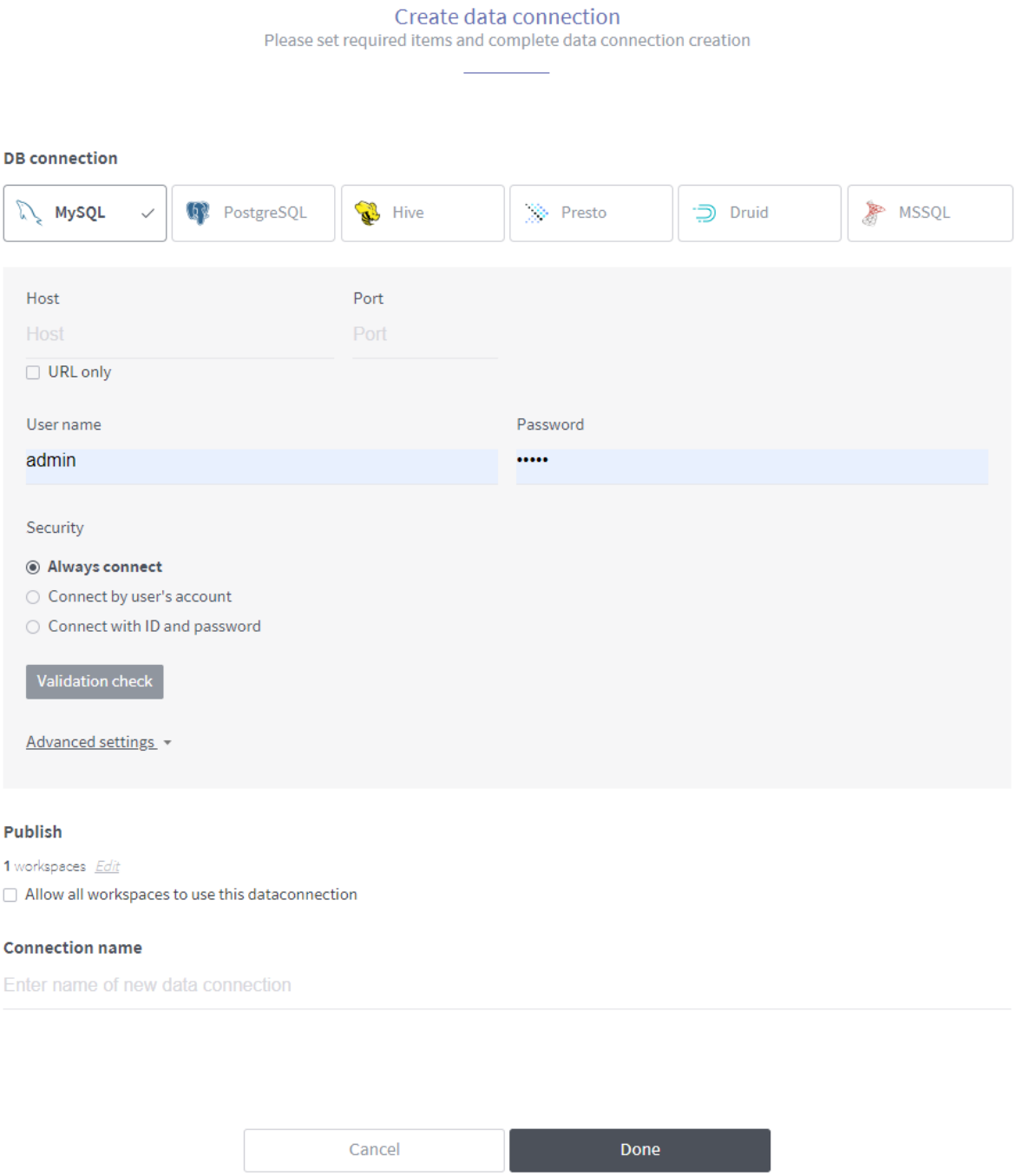
Data monitoring supports monitoring the logs of all queries submitted by users in Metatron Workbench to the staging database (internal Hive database) and external databases connected to Metatron.

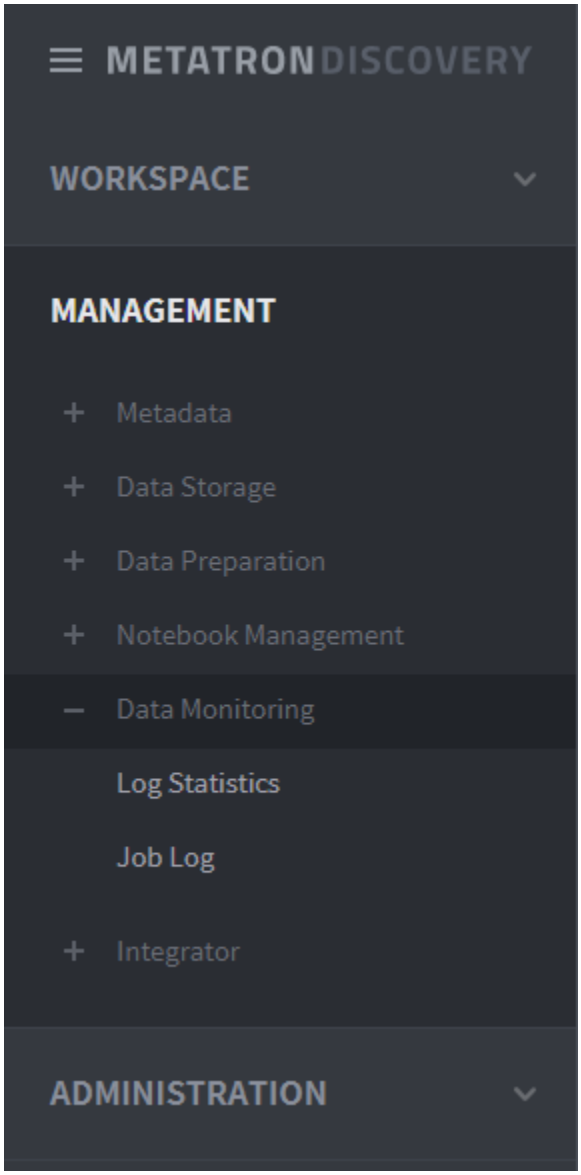
The Data Monitoring menu can be accessed under **MANAGEMENT > Data Storage > Data Monitoring** on the left-hand panel of the main screen.

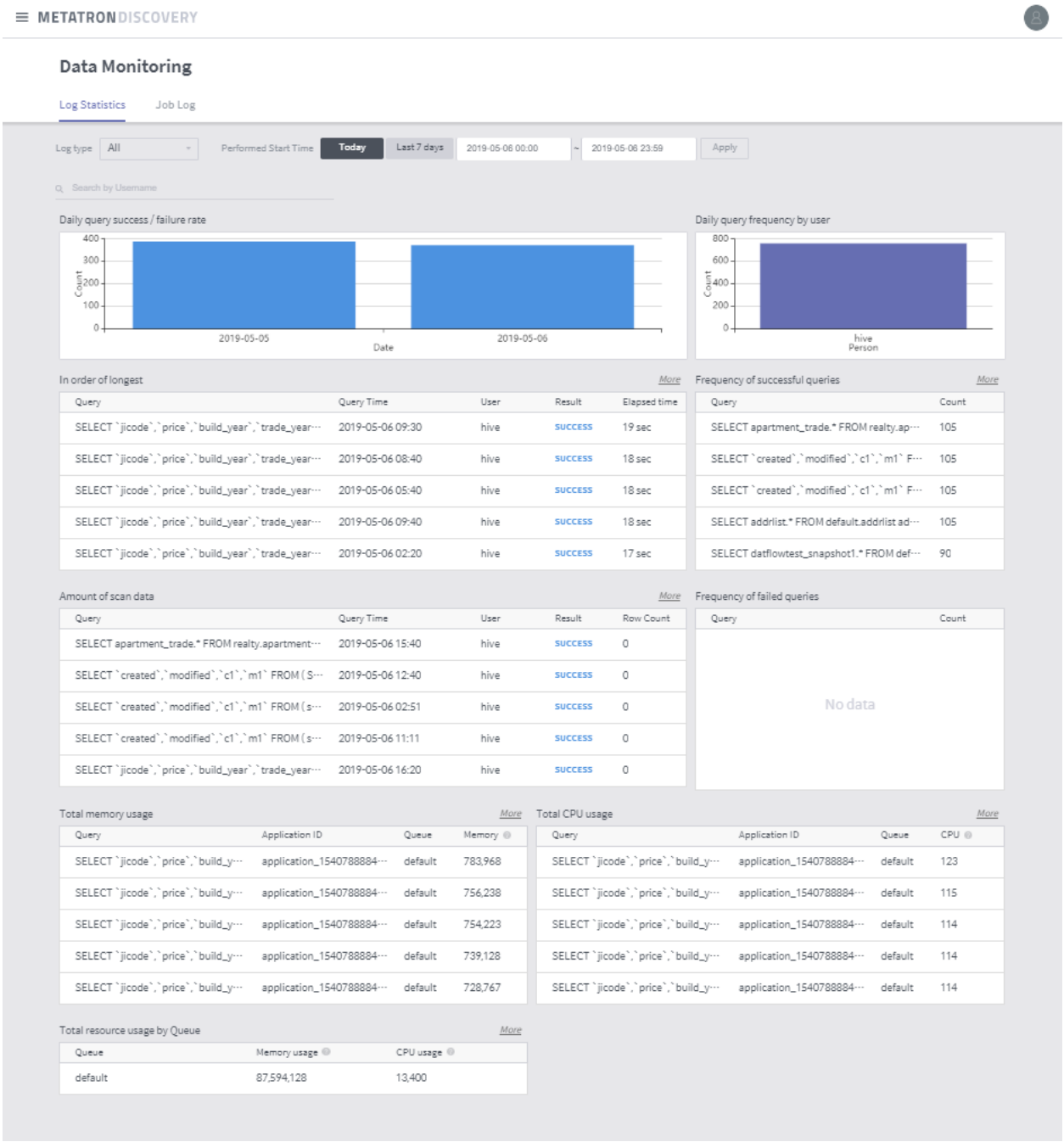
3.3.1 Log Statistics

This page collects and reports various statistics related to the performance of queries in Metatron Discovery. You can view the following nine types of basic statistics.

1. **Query success/failure rate:** Displays the daily success/failure rates of queries performed in Metatron.


Create data connection
Please set required items and complete data connection creation
DB connection
MySQL
PostgreSQL
Hive
Presto
Druid
MSSQL
Host
Port
Host
Port
URL only
User name
Password
admin
Security
Always connect
Connect by user's account
Connect with ID and password
Validation check
Advanced settings
Publish
1 workspaces Edit
Allow all workspaces to use this dataconnection
Connection name
Enter name of new data connection
Cancel
Done

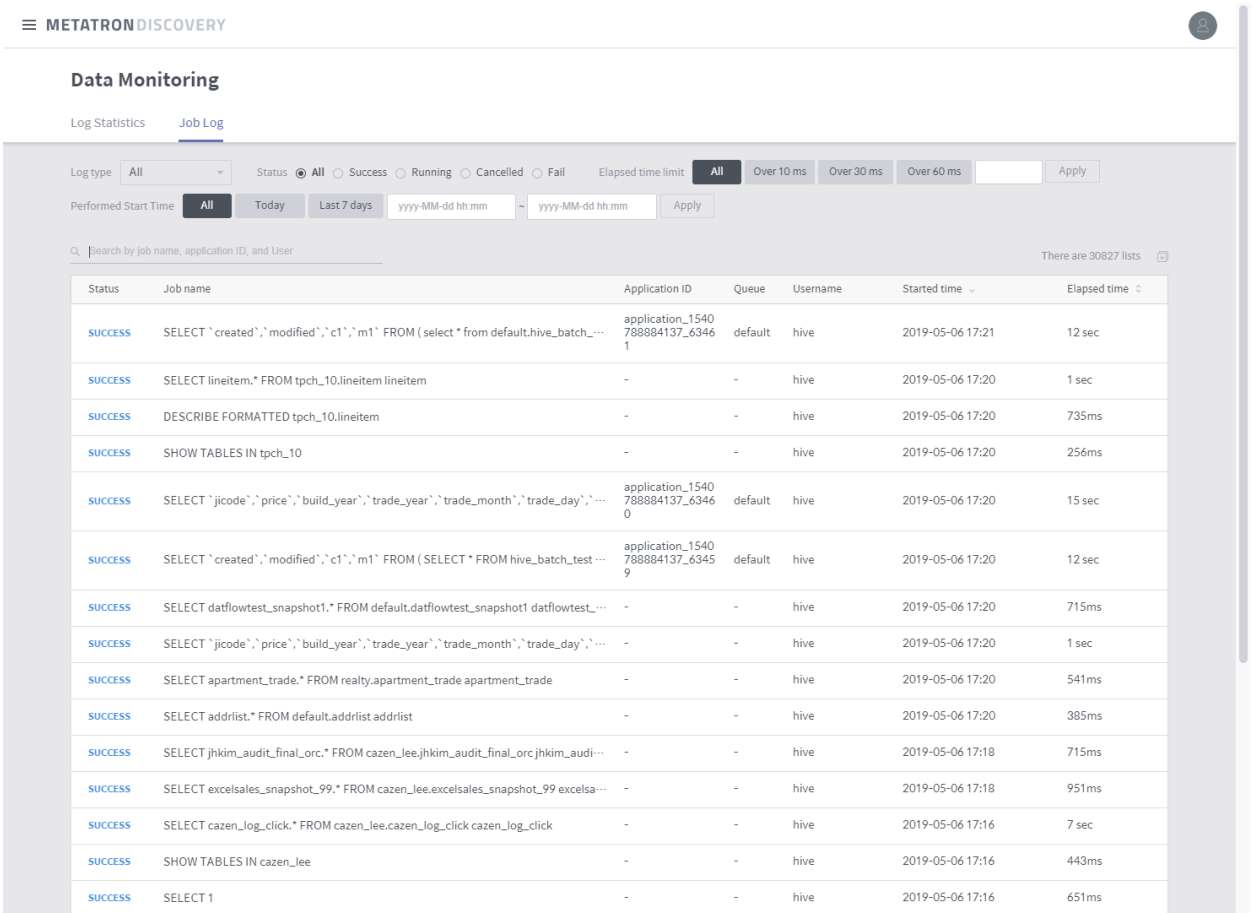

METATRONDISCOVERY
WORKSPACE
MANAGEMENT
Metadata
Data Storage
Data Preparation
Notebook Management
Data Monitoring
Log Statistics
Job Log
Integrator
ADMINISTRATION


METATRONDISCOVERY
Data Monitoring
Log Statistics
Job Log
Log type
All
Performed Start Time
Today
Last 7 days
2019-05-06 00:00
~
2019-05-06 23:59
Apply
Search by Username
Daily query success / failure rate
Count
Date
2019-05-05
2019-05-06
Daily query frequency by user
Count
hive
Person
In order of longest
More
Query
Query Time
User
Result
Elapsed time
SELECT `jicode`,`price`,`build_year`,`trade_year··· 2019-05-06 09:30 hive SUCCESS 19 sec
SELECT `jicode`,`price`,`build_year`,`trade_year··· 2019-05-06 08:40 hive SUCCESS 18 sec
SELECT `jicode`,`price`,`build_year`,`trade_year··· 2019-05-06 05:40 hive SUCCESS 18 sec
SELECT `jicode`,`price`,`build_year`,`trade_year··· 2019-05-06 09:40 hive SUCCESS 18 sec
SELECT `jicode`,`price`,`build_year`,`trade_year··· 2019-05-06 02:20 hive SUCCESS 17 sec
Frequency of successful queries
More
Query
Count
SELECT apartment_trade.* FROM realty.ap··· 105
SELECT `created`,`modified`,`c1`,`m1` F··· 105
SELECT `created`,`modified`,`c1`,`m1` F··· 105
SELECT addrlist.* FROM default.addrlist ad··· 105
SELECT datflowtest_snapshot1.* FROM def··· 90
Amount of scan data
More
Query
Query Time
User
Result
Row Count
SELECT apartment_trade.* FROM realty.apartment··· 2019-05-06 15:40 hive SUCCESS 0
SELECT `created`,`modified`,`c1`,`m1` FROM (S··· 2019-05-06 12:40 hive SUCCESS 0
SELECT `created`,`modified`,`c1`,`m1` FROM (s··· 2019-05-06 02:51 hive SUCCESS 0
SELECT `created`,`modified`,`c1`,`m1` FROM (s··· 2019-05-06 11:11 hive SUCCESS 0
SELECT `jicode`,`price`,`build_year`,`trade_year··· 2019-05-06 16:20 hive SUCCESS 0
Frequency of failed queries
Query
Count
No data
Total memory usage
More
Query
Application ID
Queue
Memory
SELECT `jicode`,`price`,`build_y··· application_1540788884··· default 783,968
SELECT `jicode`,`price`,`build_y··· application_1540788884··· default 756,238
SELECT `jicode`,`price`,`build_y··· application_1540788884··· default 754,223
SELECT `jicode`,`price`,`build_y··· application_1540788884··· default 739,128
SELECT `jicode`,`price`,`build_y··· application_1540788884··· default 728,767
Total CPU usage
More
Query
Application ID
Queue
CPU
SELECT `jicode`,`price`,`build_y··· application_1540788884··· default 123
SELECT `jicode`,`price`,`build_y··· application_1540788884··· default 115
SELECT `jicode`,`price`,`build_y··· application_1540788884··· default 114
SELECT `jicode`,`price`,`build_y··· application_1540788884··· default 114
SELECT `jicode`,`price`,`build_y··· application_1540788884··· default 114
Total resource usage by Queue
More
Queue
Memory usage
CPU usage
default 87,594,128 13,400

2. **Query frequency by user:** Graph indicating how many queries were performed by each user. Click a bar to view the job log for the user.
3. **In order of longest:** Displays the performed queries in the order of the longest running time.
4. **Amount of scan data:** Displays the performed queries in the order of the highest amount of scanned data.
5. **Frequency of successful queries:** Displays the performed queries in the order of the highest frequency of success.
6. **Frequency of failed queries:** Displays the performed queries in the order of the highest frequency of failure.
7. **Total memory usage:** Displays the performed queries in the order of the largest memory usage in total.
8. **Total CPU usage:** Displays the performed queries in the order of the largest CPU usage in total.
9. **Resource usage by queue:** Displays the resource usage in each YARN queue in the Hadoop environment.

3.3.2 Job Log

This page reports the history of all queries performed in Metatron. You can easily view previous jobs by searching the history of queries with your customized filters. The following are the filters applicable to job searching.

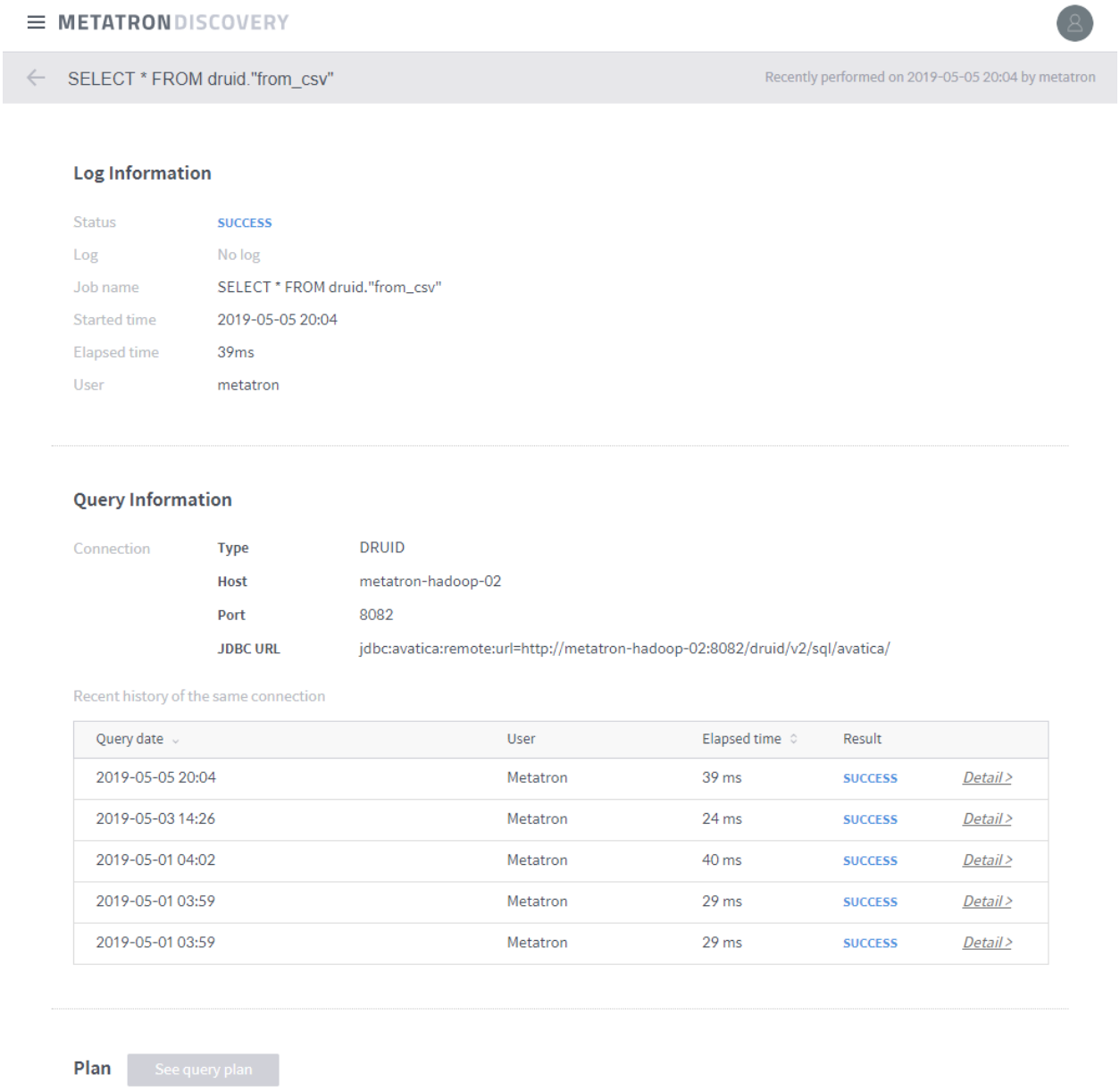


1. **Status:** Filters queries by whether they were successful or failed.
2. **Limited elapsed time:** Filters queries by long running time. You can set a reference time for this filtering.
3. **Performed start time:** Determines a time range by which to filter queries. This time range is based on when each query started running.
4. **Search by job or application:** Searches the query history by query statement or application ID.
5. **Number of entries:** Displays how many queries are returned in the list.
6. **Job list:** Lists queries filtered by specified criteria. Click an entry in the list to view its details.

Query details

Click a query listed in the job log home to view details on that query. The following information can be viewed in the details page.

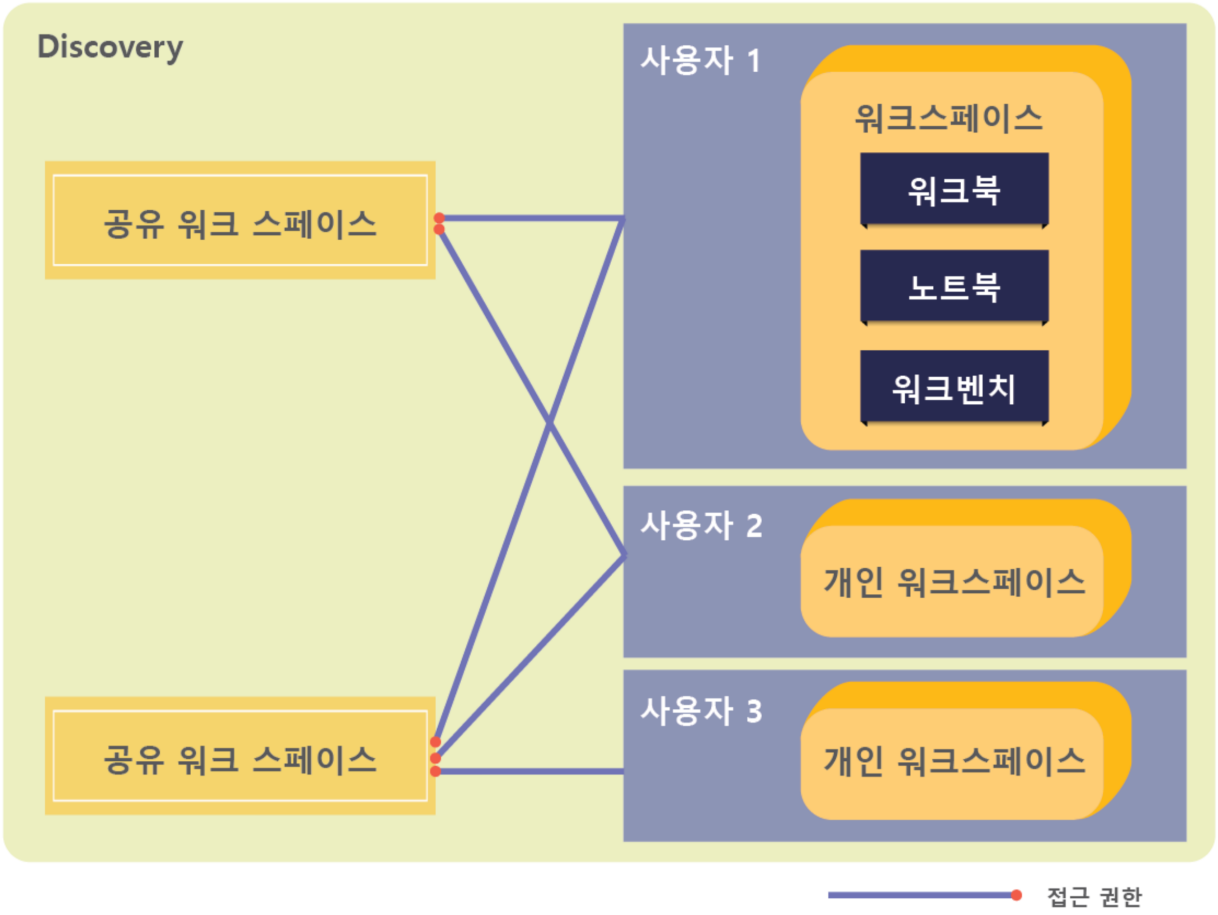
1. **Status:** Displays whether the query was successful or failed.
2. **Job name:** Statement used to perform the query.
3. **Start time:** Time when the query started running.
4. **Elapsed time:** Time taken to perform the query.
5. **User:** User ID who performed the query.
6. **Connection:** For a query performed in a workbench, the connection information of the database is displayed.
7. **Recent history of the same connection:** For a query performed in a workbench, the latest five queries performed in the database and their results are displayed. Click Detail to pop up a window showing the query statement.
8. **Plan:** Implements the query plan.


METATRONDISCOVERY
SELECT * FROM druid."from_csv"
Recently performed on 2019-05-05 20:04 by metatron
Log Information
Status SUCCESS
Log No log
Job name SELECT * FROM druid."from_csv"
Started time 2019-05-05 20:04
Elapsed time 39ms
User metatron
Query Information
Connection
Type DRUID
Host metatron-hadoop-02
Port 8082
JDBC URL jdbc:avatica:remote:url=http://metatron-hadoop-02:8082/druid/v2/sql/avatica/
Recent history of the same connection
Query date
User
Elapsed time
Result
2019-05-05 20:04 Metatron 39 ms SUCCESS Detail >
2019-05-03 14:26 Metatron 24 ms SUCCESS Detail >
2019-05-01 04:02 Metatron 40 ms SUCCESS Detail >
2019-05-01 03:59 Metatron 29 ms SUCCESS Detail >
2019-05-01 03:59 Metatron 29 ms SUCCESS Detail >
Plan
See query plan

CHAPTER

FOUR

WORKSPACE



A workspace stores Metatron Discovery's analytics entities such as workbooks, notebooks, and workbenches. There are two types of workspaces: personal and shared workspaces.

- **Personal workspace:** A private workspace assigned to each Discovery member. It is accessible only to the owner.
- **Shared workspace:** A public workspace shared by multiple users. It is used for users to share analytics processes and results with each other. The owner or administrator of a shared workspace can grant various levels of access to Discovery members.

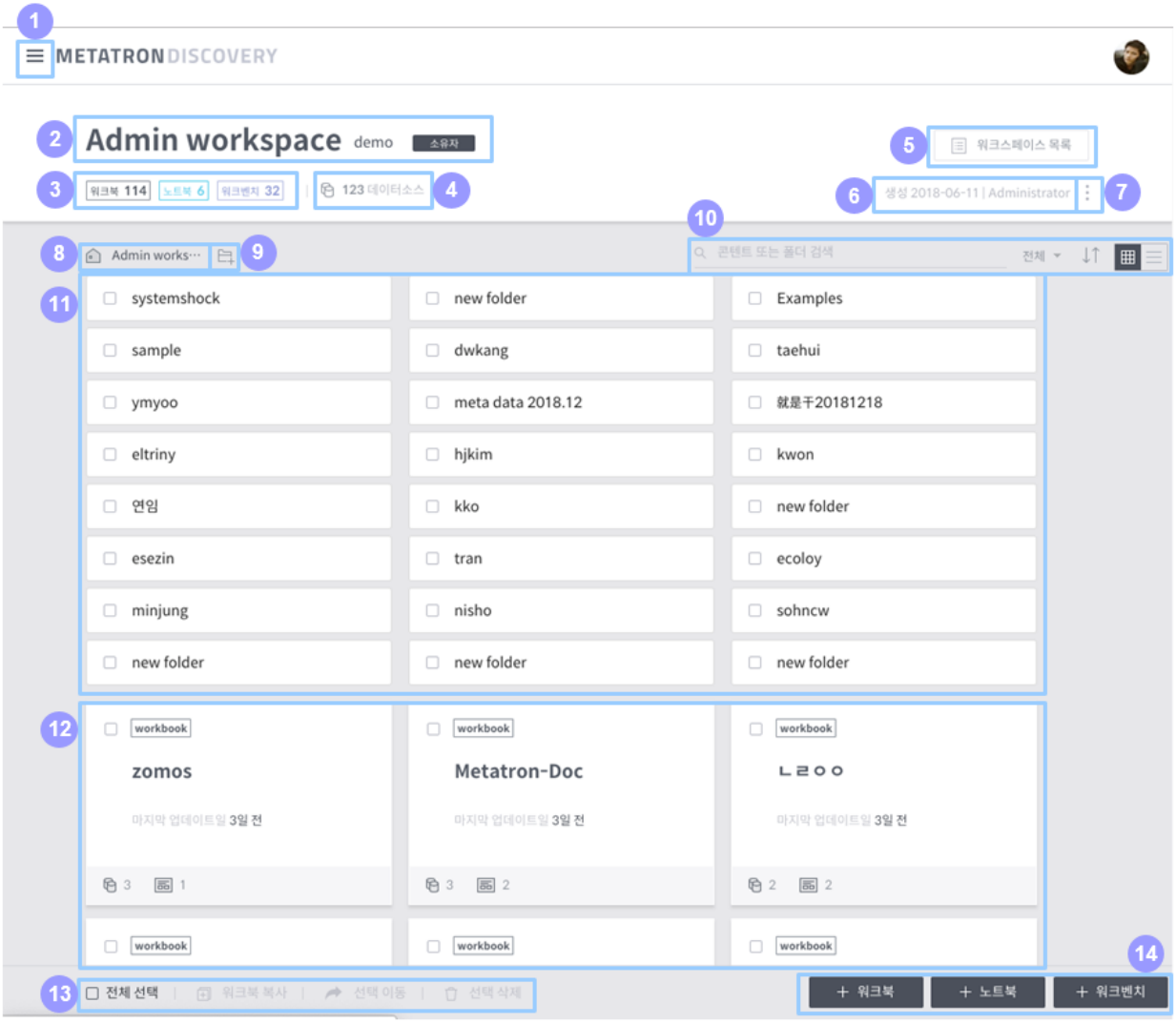
This chapter introduces **workspace home page and UI**, and then how to use **shared workspaces**.

4.1 Workspace home

On the workspace home page, you can perform manage the Metatron Discovery entities (workbooks, notebooks and workbenches) contained in the workspace.

4.1.1 Composition of the workspace home

The overall composition of the workspace home is as follows:



1. **Main menu button:** Click this button to open a panel to access another workspace.
2. **Workspace information:** Displays the name and description of the workspace. If the logged-in user owns the workspace, an Owner icon will be displayed next to the name of the workspace.
3. **Registered entities:** Displays the number of entities registered in the workspace by entity type.
4. **Data source:** Displays the number of data sources used in the workspace. Click this area to show a list of these data sources.

5. **Workspace list:** Click this button to show a list of shared workspaces. (See *Shared workspace list* for how to handle it.)
6. **Creation information:** Displays who and when created the workspace.
7. **More:** Edit the settings of the workspace.
   - **Edit the name and description:** Edits the name and description of the workspace.
   - **Set shared member & group:** Sets the users and groups who can access the workspace. (See *Set access permissions for a shared workspace* for details.)
   - **Set notebook server:** Sets access information for external analytics tool servers used by the Notebook module.
   - **Set permission schema:** Sets the access permission of each user role for the workspace. (See *Set access permissions for a shared workspace* for details.)
   - **Change owner:** Changes the owner of the workspace.
   - **Delete workspace:** Deletes the workspace.
8. **Path in the workspace:** Displays the current location in the workspace. Click on a parent folder listed in the path to move to that folder.
9. **Create a folder:** Click on it to create a new folder in the current location.
10. **Filter/sort the entity list:**
    - **Search:** Searches for an entity or folder in the workspace by name.
    - **Entity type:** Displays only your selected type of entities among workbooks, notebooks, and workbenches.
    - **Sort:** Sorts folders and entities by their name or when they were last updated.
    - **View type:** Select either the grid view or list view as the format of how the entities are listed in the workspace.
11. **Folder list:** Displays folders that meet search criteria in the current location. Click one to enter that folder. (For details on individual folders, see *Folder items*)
12. **Entity list:** Displays entities that meet search or sorting criteria in the current location. Click an entity to enter its home. (For details on individual entities, see *Entity items*)
13. **Select/clone/move/delete entity:** Select all entities, or clone, move or delete an entity. (See *Select/clone/move/delete folder and entity* for details.)
14. **Create an entity:** Buttons used to create a specific type of entity in the workspace. (For details, see *Create a workbook*, *Create a notebook*, and *Create a workbench*, respectively.)

4.1.2 Folder items

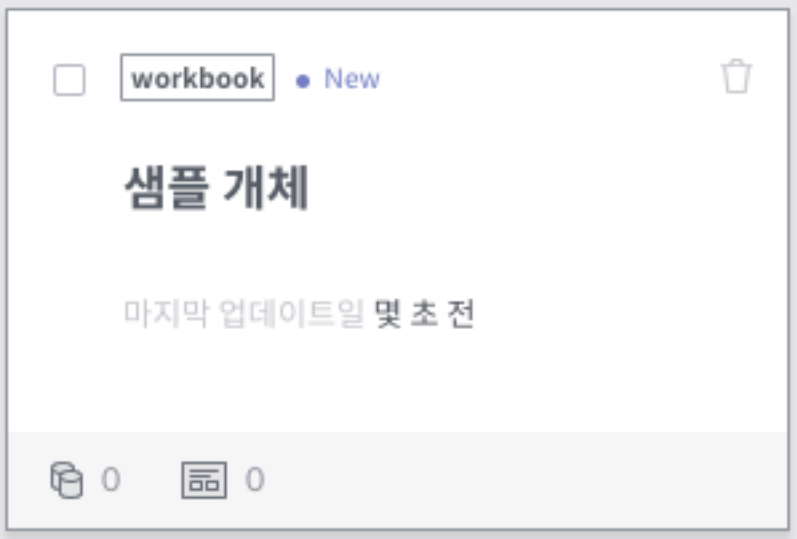
When the mouse cursor is over a folder, it is shown as follows:

- **Check box:** Used to select the folder. You can clone, move or delete the selected folder.
- **Name:** Name of the folder.

- **Edit:** Click on it to modify the name of the folder. This button is displayed only when you hover the mouse over the folder item.
- **Delete:** Click on it to delete the folder. This button is displayed only when you hover the mouse over the folder item.

4.1.3 Entity items

When the mouse cursor is over an entity, it is shown as follows:

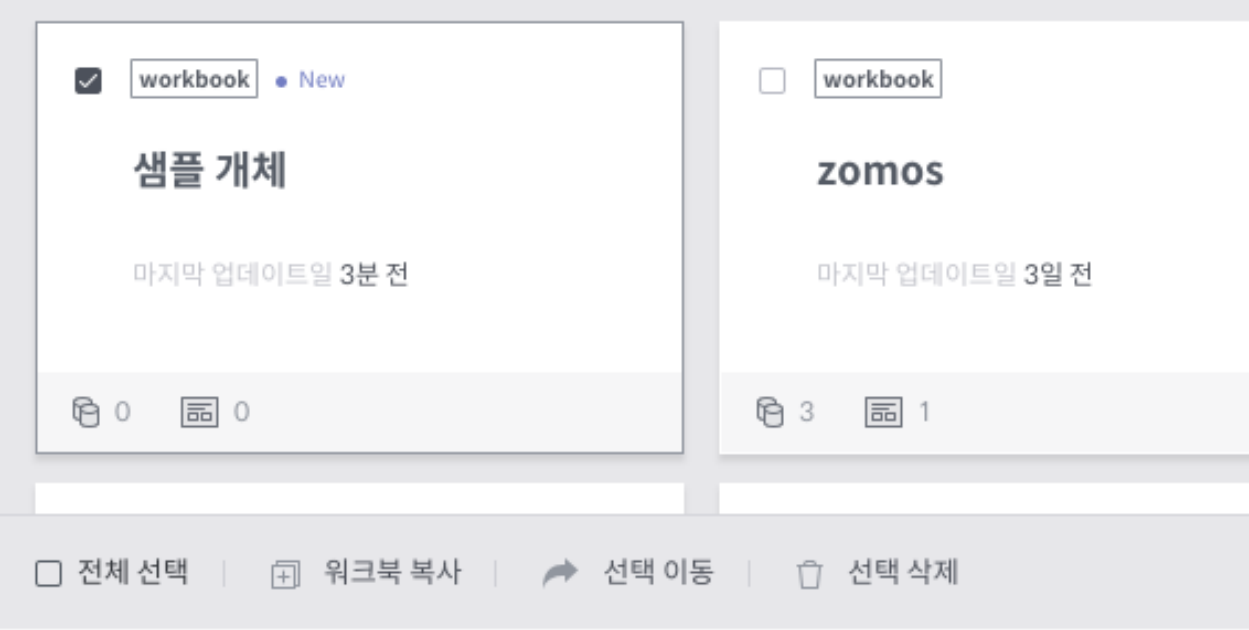


- **Check box:** Used to select the entity. You can clone, move or delete the selected entity.
- **Entity type:** Displays the type of the entity (workbook/notebook/workbench).
- **Delete:** Click on it to delete the entity. This button is displayed only when you hover the mouse over the entity item.
- **Name:** Name of the entity.
- **Last updated:** Displays when the entity was last updated.
- **Number of data sources/dashboards:** This is an exclusive area for the workbook type.
  - The number next to the icon refers to how many data sources are connected to the workbook.
  - The number next to the icon refers to how many dashboards are registered in the workbook.

4.1.4 Select/clone/move/delete folder and entity

You can clone, move or delete folders and entities in the workspace. Once you select a folder or entity, the clone, move, and delete buttons in the lower-left corner of the workspace home become active.

- **Select all:** Selects all items in the current folder and entity list.
- **Clone workbook:** This is exclusive for the workbook type. Click this button to clone the selected workbooks.



- **Move selections:** Moves the selected folders and entities. Workbooks can be moved to another workspace, and other types of items can be moved to another folder in the same workspace. However, it is impossible to move selections when workbooks and other types of entities are selected together.
- **Delete:** Deletes the selected folders and entities.

4.2 Shared workspace

A shared workspace is designed for access and use by multiple users. The following subsections describe how to view and create shared workspaces, and explain "permission schema," which sets which users or groups are allowed to access shared workspaces.

4.2.1 Shared workspace list

The shared workspace list page is used to view a list of all shared workspaces accessible to the logged-in user and to move to a specific workspace. This page can be accessed via two methods:

- Click the ⋮ button at the top-left of the Discovery screen to open the main panel, and click **Workspace list >>**.
- Click **Workspace list** at the top-right of the workspace home.

The shared workspace list page is composed as follows:

1. **Number of shared workspaces:** Displays how many shared workspaces are listed.
2. **Add a shared workspace:** Click this button to move to the page to add a shared workspace. (See *Create a shared workspace* for a deailed procedure)
3. **Personal workspace:** Click this button to move to the personal workspace owned by the logged-in user.
4. **Search:** Searches the shared workspace list by the name you typed in.

5. **Favorites:** Displays only those workspaces designated as favorites.
6. **Public only:** Displays only those workspaces set as public.
7. **I'm the owner:** Displays only those workspaces for which the logged-in user is the administrator.
8. **Name ascending/descending:** Sorts the shared workspace list by name ascending/descending.
9. **Workspace list:** Lists workspaces filtered by specified criteria. Click one to move to enter that workspace.

4.2.2 Create a shared workspace

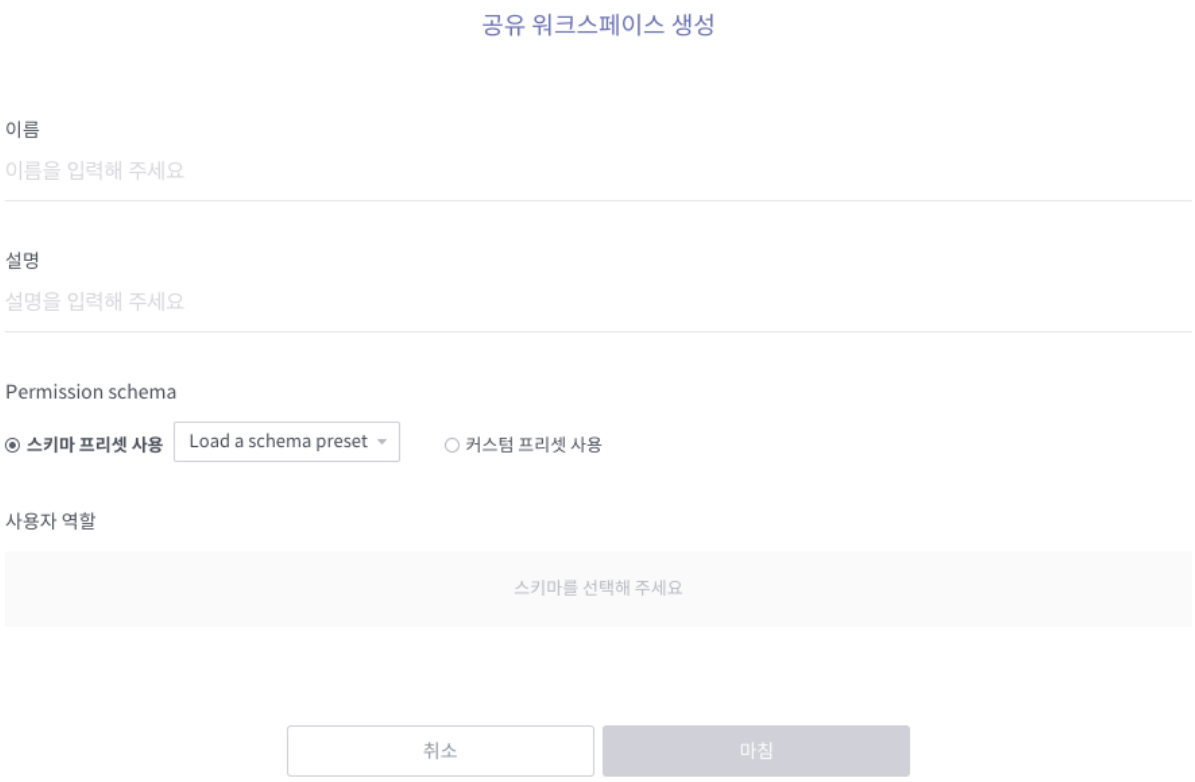
A new shared workspace is created as follows:

1. Click the ⊕ button on the shared workspace list page to move the page to create a new shared workspace.
2. Enter a **Name** and **Description**, and then set up the **Permission schema** by referring to the descriptions below:
   - **Use a preset schema:** Load the permission schema defined by the administrator.
   - **Use a custom schema:** Define a new permission schema. (See *Set access permissions for a shared workspace* for how to define a new permission schema.)
3. Click **Done** to finish creating a workspace.

4.2.3 Set access permissions for a shared workspace

Setting the access permission for a shared workspace is conducted in the following two steps:

- Set an access permission for each user role (See *Set permission schema*)



- Grant a role to each user or user group (See *Set shared members & groups*)

Set permission schema

View permission schema

Click the ⋮ icon at the top-right of the shared workspace home and click **Set permission schema** to view the defined permission schema as follows:

In the above example, Manager, Editor, Watcher, and Guest are defined as user roles. As shown in this example, a permission schema is a set of user roles defining different access permissions.

What each column determines is as follows:

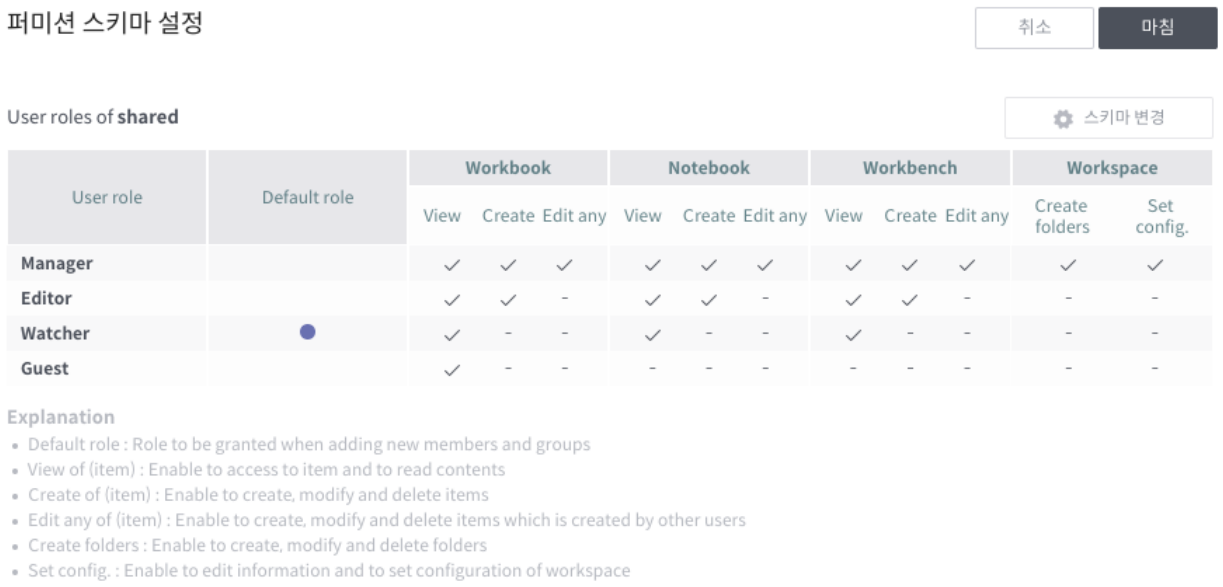
Default role

When a new user or user group is added, it is assigned the default role.

Permission for each entity type (workbook/notebook/workbench)

- **View:** Allows to access and view data in entities of the type.
- **Create:** Allows to create, edit, and delete entities of the type.
- **Edit any:** Allows to edit or delete entities of the type created by another user.

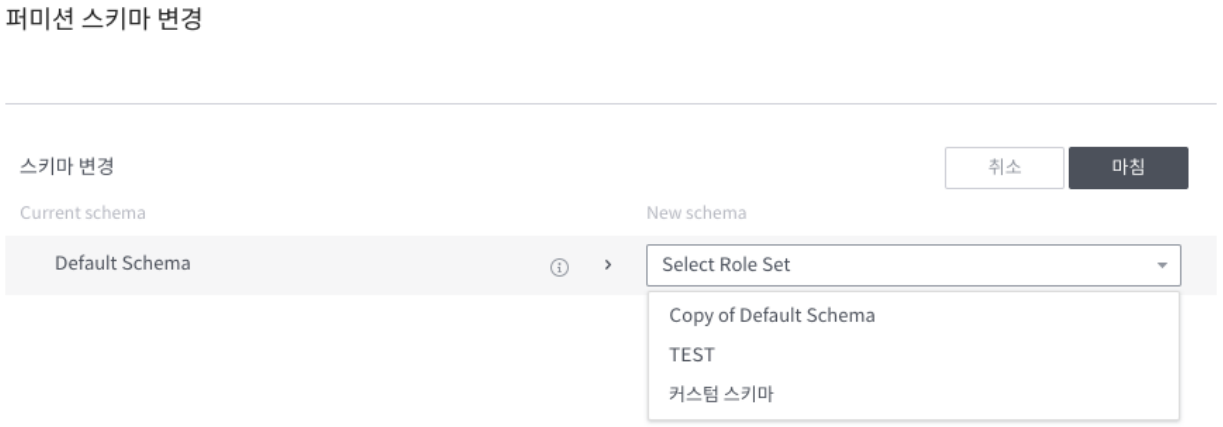
Workspace permission



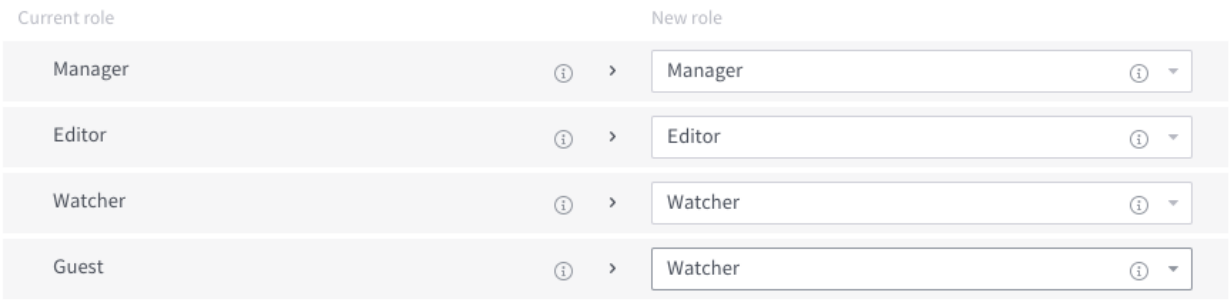
- Create folders: Allows to create, edit, and delete folders in the workspace.
- Set config.: Allows to modify the name and description of the workspace and to change the workspace permission schema.

Change permission schema

Click the **Change schema** button on the permission schema view page to move to a page to change the defined permission schema as follows:



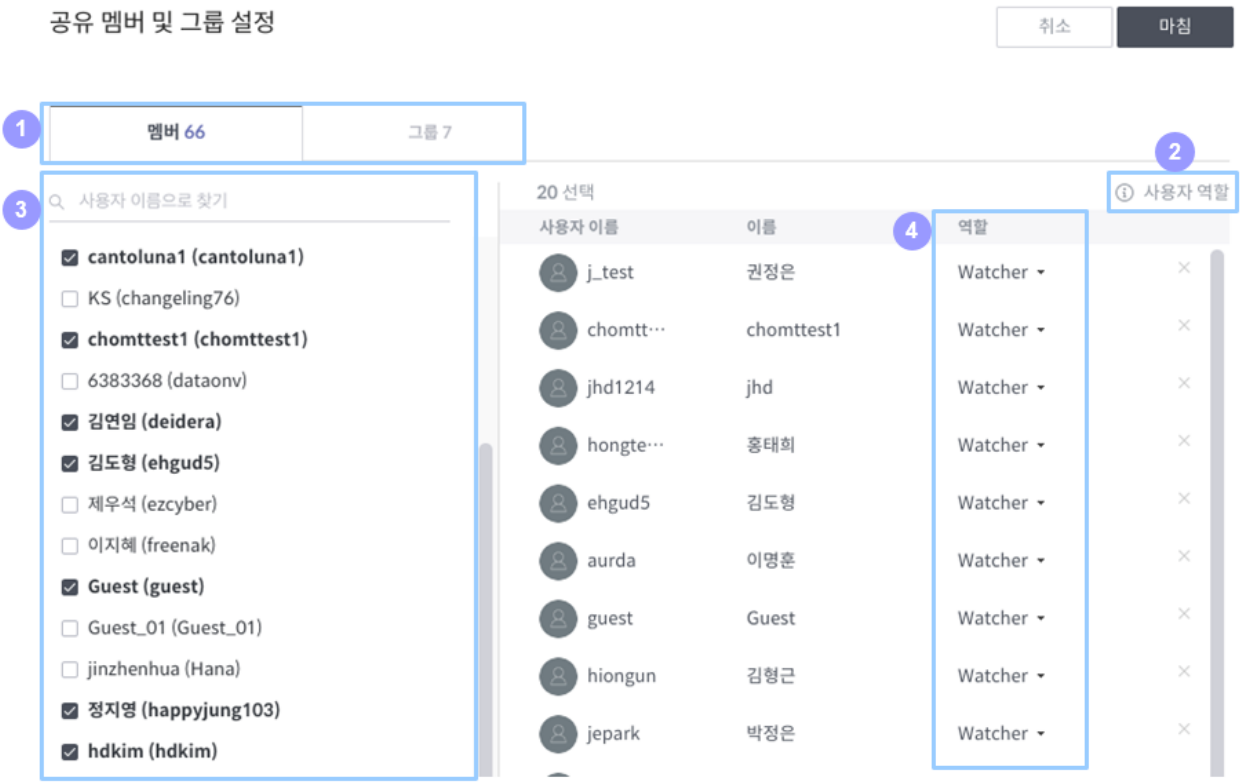
Click **Select Role Set** combo box on the right to display the permission schema defined by the administrator. **Custom schema** at the bottom of the list allows you to set new user roles. Select one to display the following section. (If you select **Custom schema**, you must first define a permission for each user role. Click the button at the right of New schema to move to the permission setting page, and set a permission for each user role by referring to *View permission schema*)



Here, each user role of the current permission schema is substituted with the user role defined in the new permission schema. Hover the mouse over the ⓘ icon next to the name of a user role to display the permission assigned to the user role. Click **Done** to finish setting the permission schema.

Set shared members & groups

Click the icon at the top-right of the shared workspace home, and click **Set shared member & group** to move to a page to set members and groups for the shared workspace as follows: On this page, each user or user group is assigned a user role defined in the permission schema. Assign user roles by referring to the following explanation, and click **Done** to finish setting workspace access permissions.



1. **Select whether to assign user roles individually or in groups**
   - **Member tab:** Assign user roles to individual users.

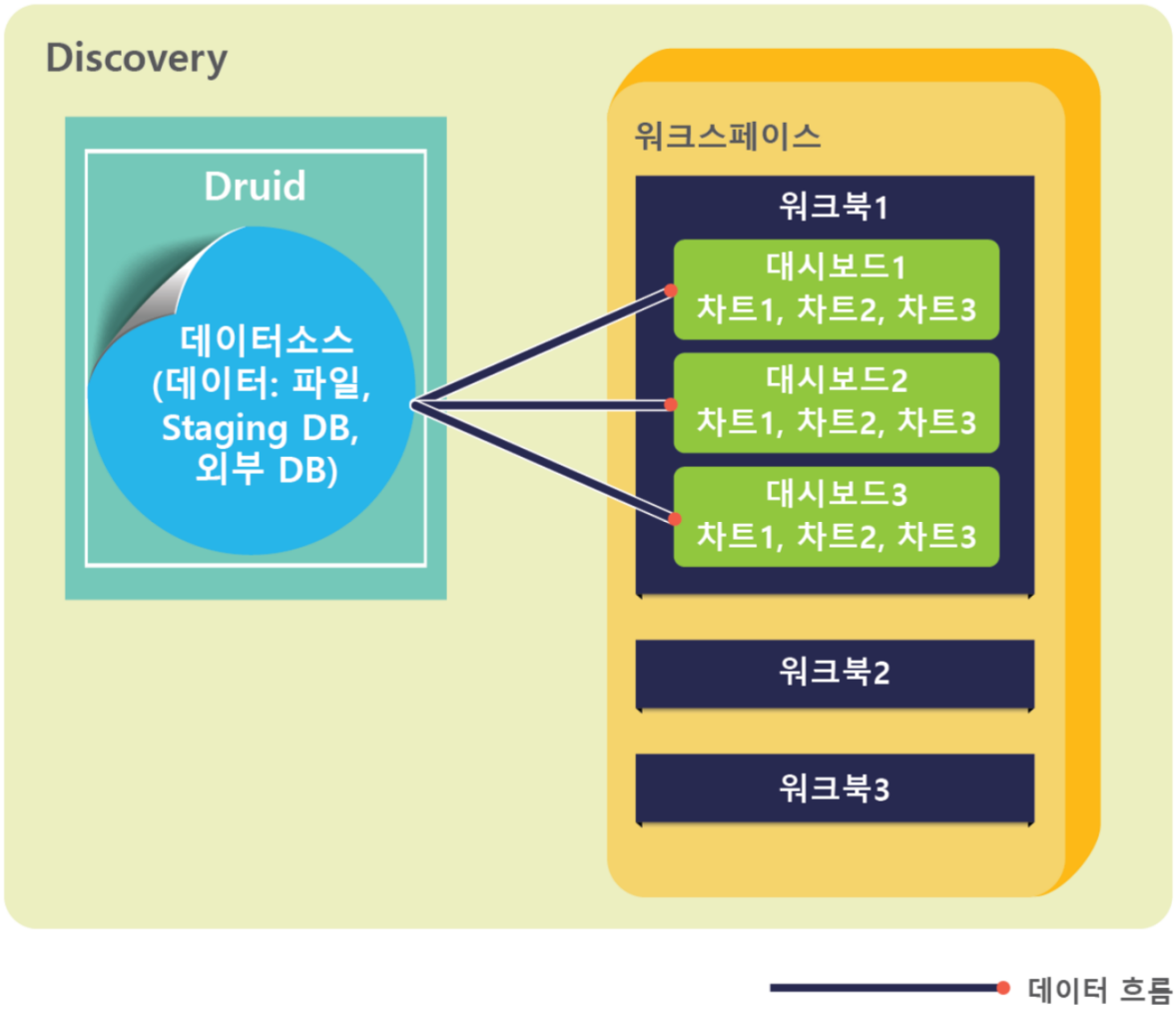
- **Group tab:** Assign user roles in groups. (A user group can be established by administrator permission.)

2. **User roles:** Click on it to pop up a dialog box showing the permission schema, which defines a permission for each user role.
3. **Member/group list:** Lists the users (groups in the case of the group tab) registered in Discovery. Click a user (group) in the list to add it to the role assignment section on the right. Click an added user (group) to remove it from the section on the right.
4. **Assign a user role:** Click this combo box to display user roles defined in the active permission schema. Select the role you want to assign to the user (group).

CHAPTER

FIVE

WORKBOOK



Workbook is a data visualization module powered by the Metatron Druid engine. As shown in the diagram above, each **workbook**?a standalone report?consists of multiple **dashboards**, while each dashboard consists of various **charts** showing a visualization of source data analysis.

The main features of Workbook are as follows:

- Fast and flexible data analytics over time-series multidimensional data sources.

- Dashboards contain a variety of visualized charts and texts to be compiled into a report for presentations.
- Frequently used algorithms such as clustering, prediction lines, and trend lines can be implemented through a GUI (graphical user interface).

This chapter consists of:

5.1 Create a workbook

In Metatron Discovery, a **workbook** functions as a standalone data analytics report. Once a workbook is created, you can store a number of **dashboard** slides in the workbook and present them in the proper order.

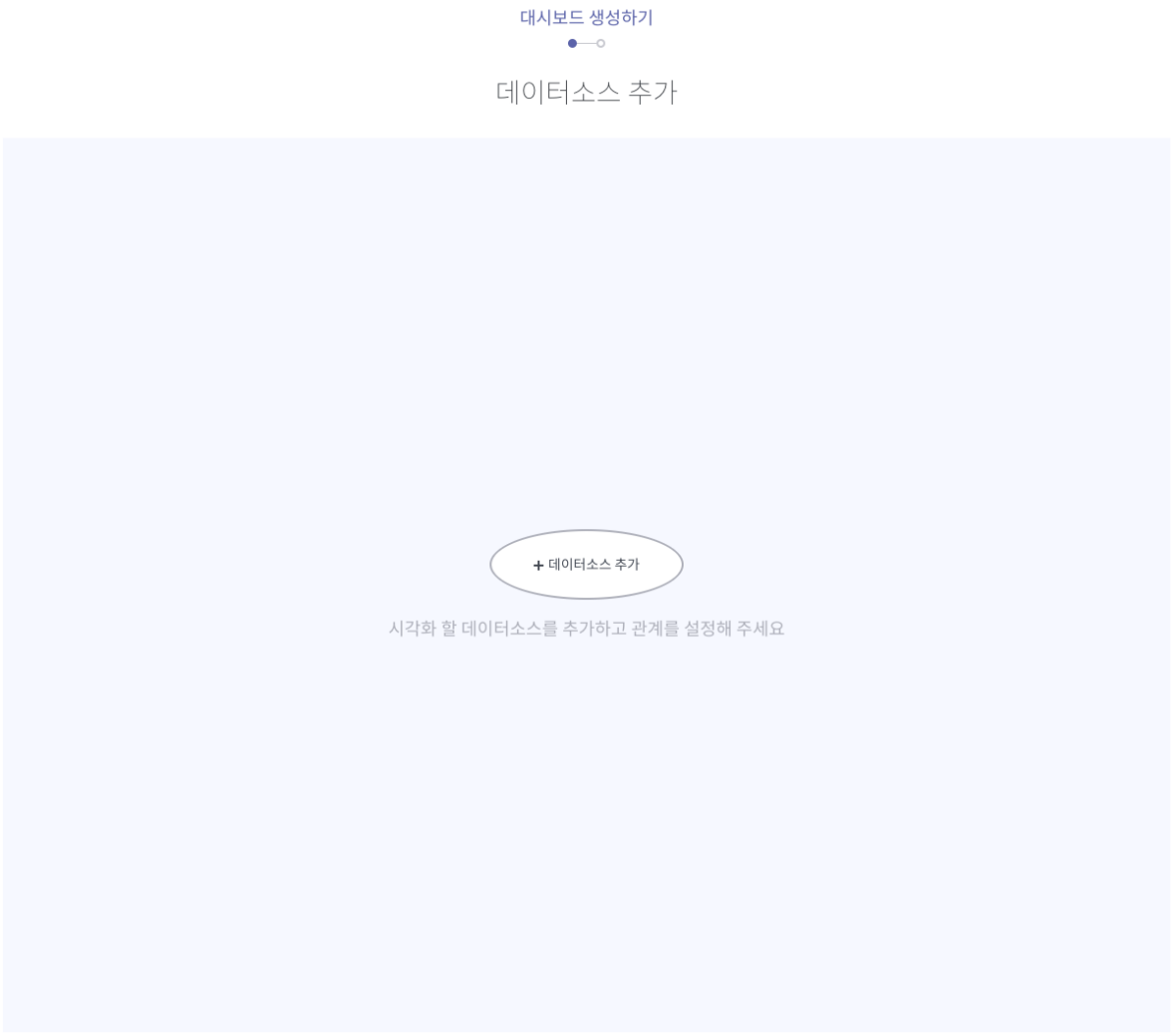
A workbook is created as follows:

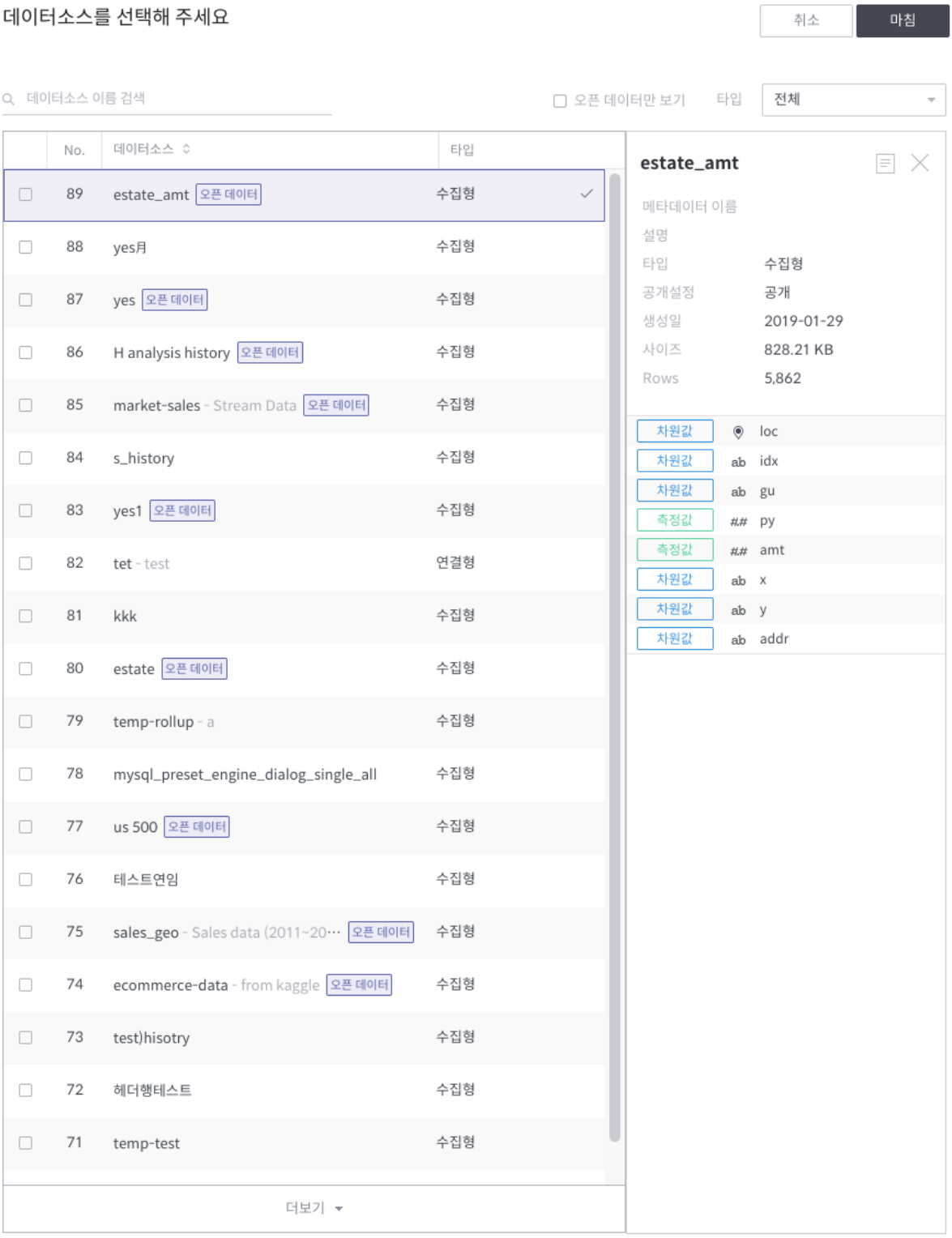
1. Click the **+ Workbook** button at the bottom of the workspace to move to the workbook creation page.

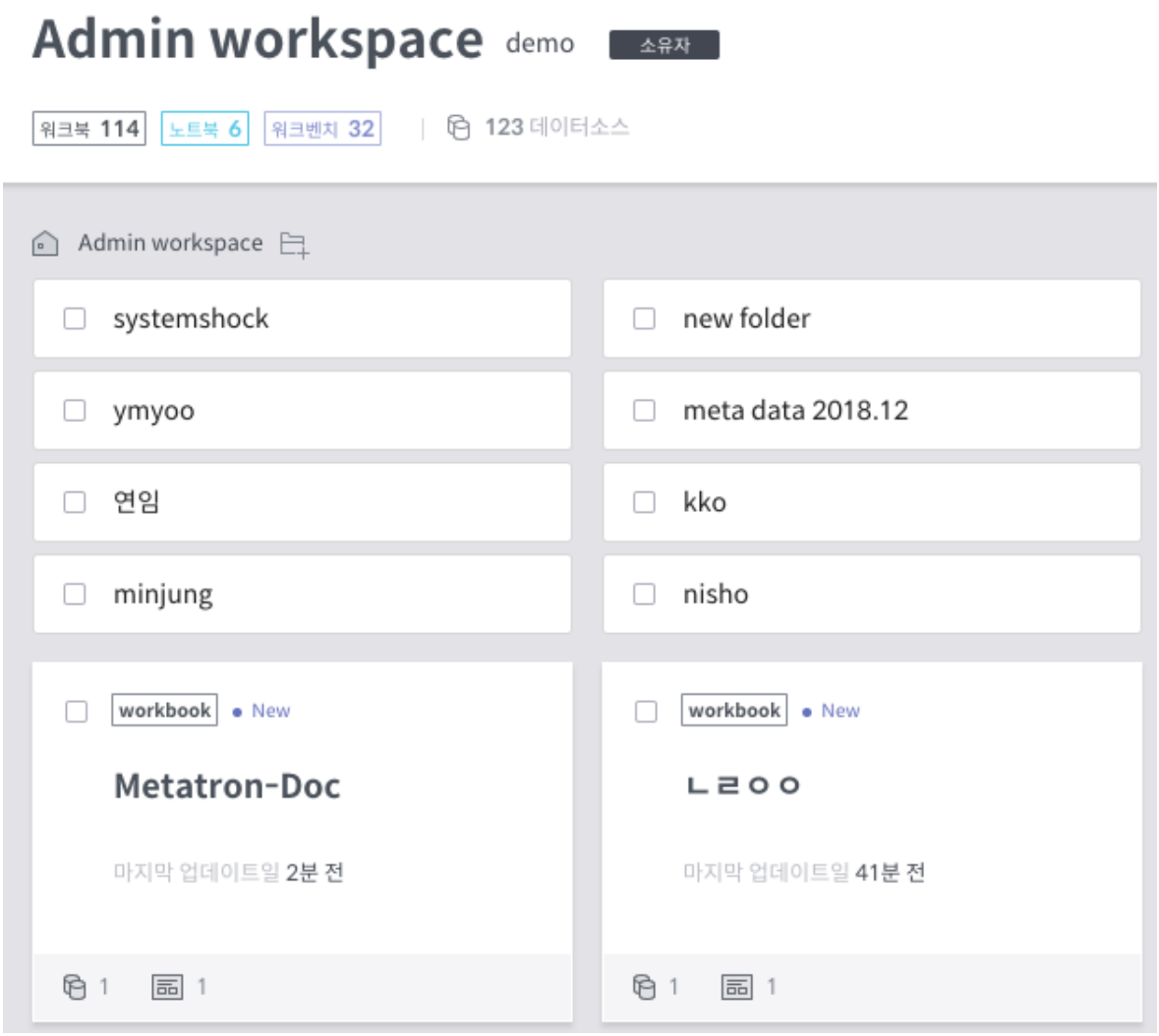
2. Enter a name (required) and description for the workbook to be created and click **Done**. If you select **Continue to create a dashboard of a new workbook**, you'll proceed directly to the **Create Dashboard** page. This option is provided because a workbook cannot work without dashboards in it.

3. After clicking the "+ Add Data Source" button in the middle of the screen, select a data source to create a dashboard. For details on how to create a dashboard, refer to *Create a dashboard*.

4. You can check the new workbook in the workspace home as shown below. Click the workbook to enter it.

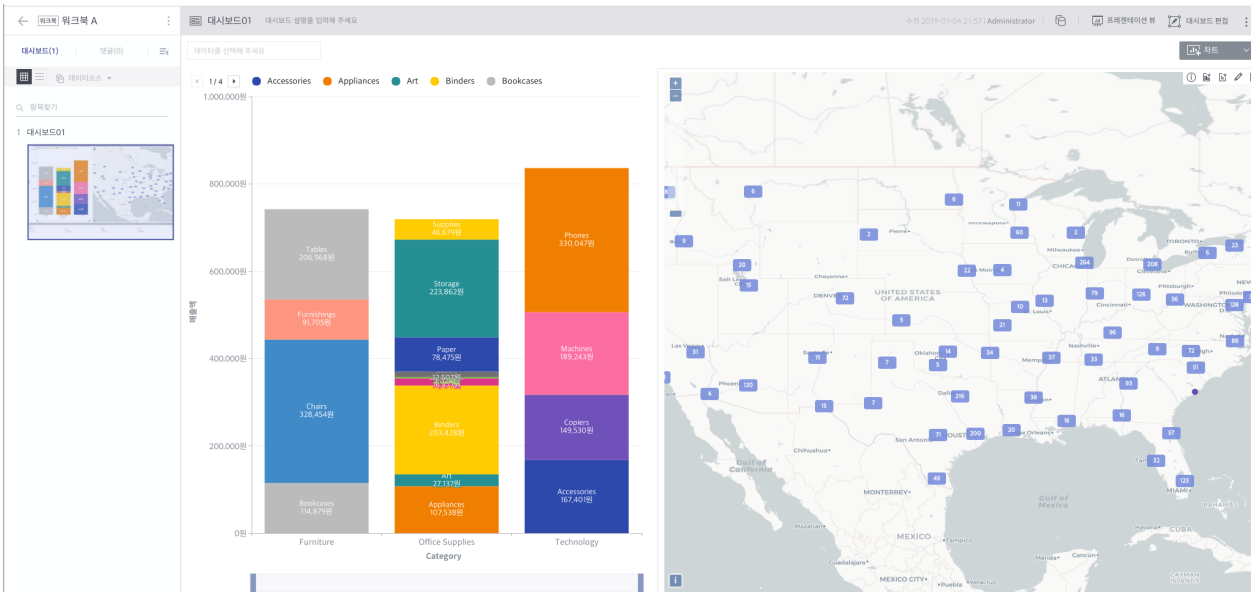

대시보드 생성하기
데이터소스 추가
+ 데이터소스 추가
시각화 할 데이터소스를 추가하고 관계를 설정해 주세요


데이터소스를 선택해 주세요
취소
마침
데이터소스 이름 검색
오픈 데이터만 보기
타입
전체
No.
데이터소스
타입
89 estate_amt 오픈 데이터 수집형
88 yes月 수집형
87 yes 오픈 데이터 수집형
86 H analysis history 오픈 데이터 수집형
85 market-sales - Stream Data 오픈 데이터 수집형
84 s_history 수집형
83 yes1 오픈 데이터 수집형
82 tet - test 연결형
81 kkk 수집형
80 estate 오픈 데이터 수집형
79 temp-rollup - a 수집형
78 mysql_preset_engine_dialog_single_all 수집형
77 us 500 오픈 데이터 수집형
76 테스트연임 수집형
75 sales_geo - Sales data (2011~20··· 오픈 데이터 수집형
74 ecommerce-data - from kaggle 오픈 데이터 수집형
73 test)hisotry 수집형
72 헤더행테스트 수집형
71 temp-test 수집형
더보기
estate_amt
메타데이터 이름
설명
타입 수집형
공개설정 공개
생성일 2019-01-29
사이즈 828.21 KB
Rows 5,862
차원값 loc
차원값 ab idx
차원값 ab gu
측정값 #.# py
측정값 #.# amt
차원값 ab x
차원값 ab y
차원값 ab addr


Admin workspace
demo
소유자
워크북 114
노트북 6
워크벤치 32
123 데이터소스
Admin workspace
systemshock
new folder
ymyoo
meta data 2018.12
연임
kko
minjung
nisho
workbook
New
Metatron-Doc
마지막 업데이트일 2분 전
1
1
workbook
New
ㄴㄹㅇㅇ
마지막 업데이트일 41분 전
1
1

5.2 Dashboard

Stored in a workbook, a **dashboard** provides functions to analyze and visualize its connected data source as needed. Therefore, an important step to create a dashboard is connecting to a data source.



You can visualize analyses of various data sources into charts and texts; those visualizations are customizable using pivoting, chart mapping, and filtering.

5.2.1 Create a dashboard

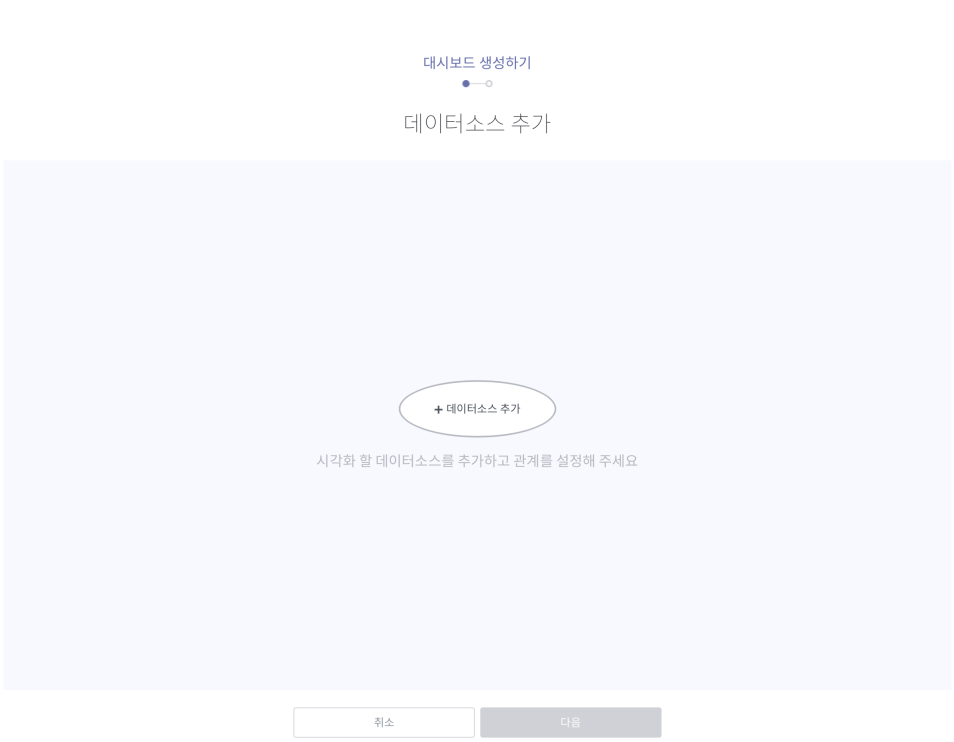
A dashboard is created as follows:

1. Click **+ Add data source** on the workbook screen.

2. From the list of data sources accessible to the workspace, select the master data sources to which you want to connect the dashboard. In a subsequent step, you can select additional data sources to be joined to these master data sources selected here.

    - **Search by data source name:** Search for a data source accessible to the workspace by name.
    - **Show open data only:** Displays only those designated as "open data sources."
    - **Type:** Displays only those data sources that are the connection or collection type.
    - **Data source list:** Lists data sources filtered by specified criteria.
    - **Data source information:** Displays brief information of the data source selected in the list.

3. If you have selected more than one data source, you can associate them by dragging one data source to another. Associated data sources can be filtered by each other. If you do not want data source association, simply click **Done**.



4. Once you drag a data source to another one, a new window will pop up to prompt you to configure the data source association. Select a column on each table as an association key by which to filter the other data source. And click **Done**.

5. Once you have finished setting up associations between the master data sources, click **Done**.

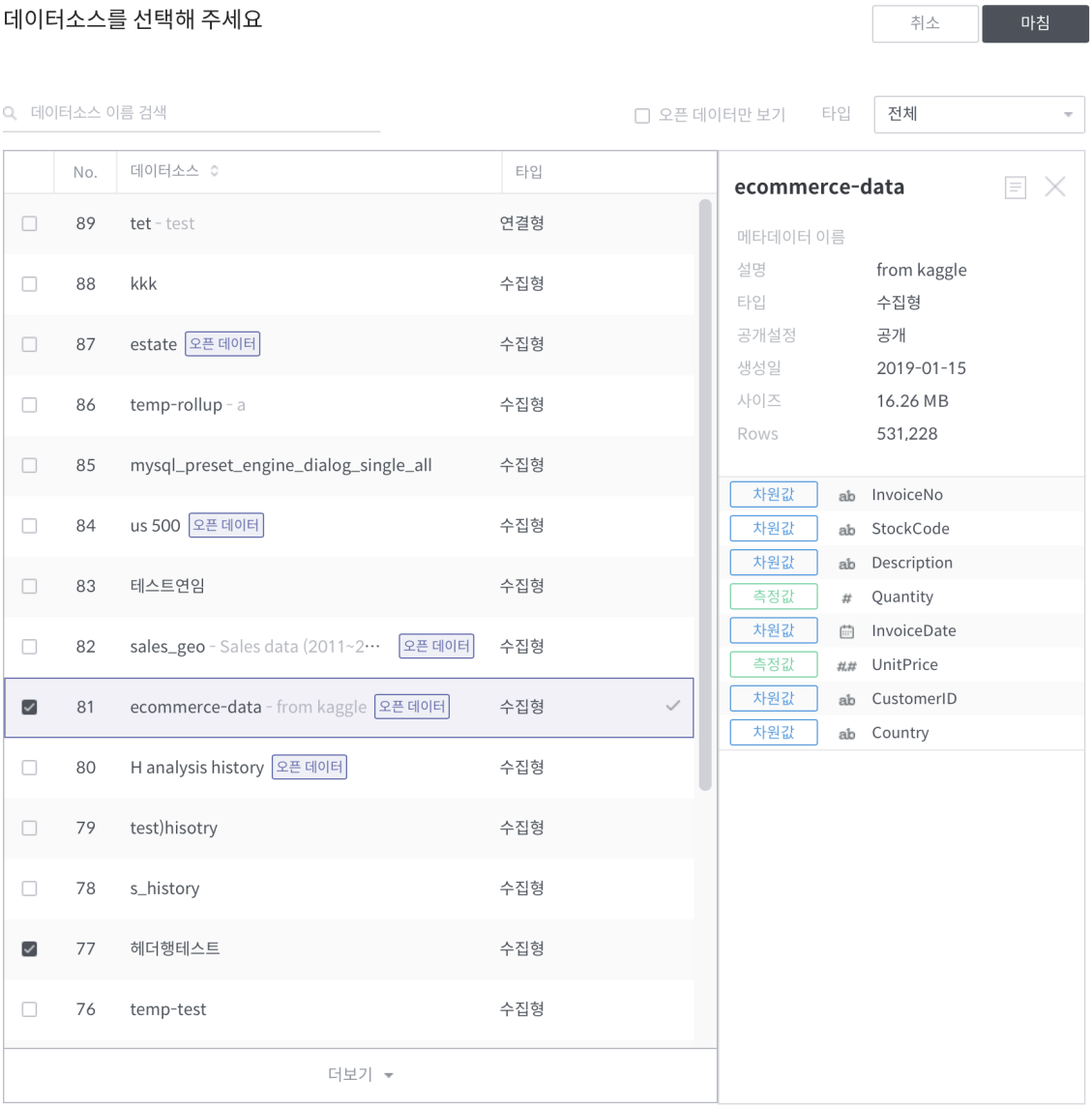
6. Re-configure master data source associations or add other data sources to be joined to the top data source selected above as described below:

   Master data source association view

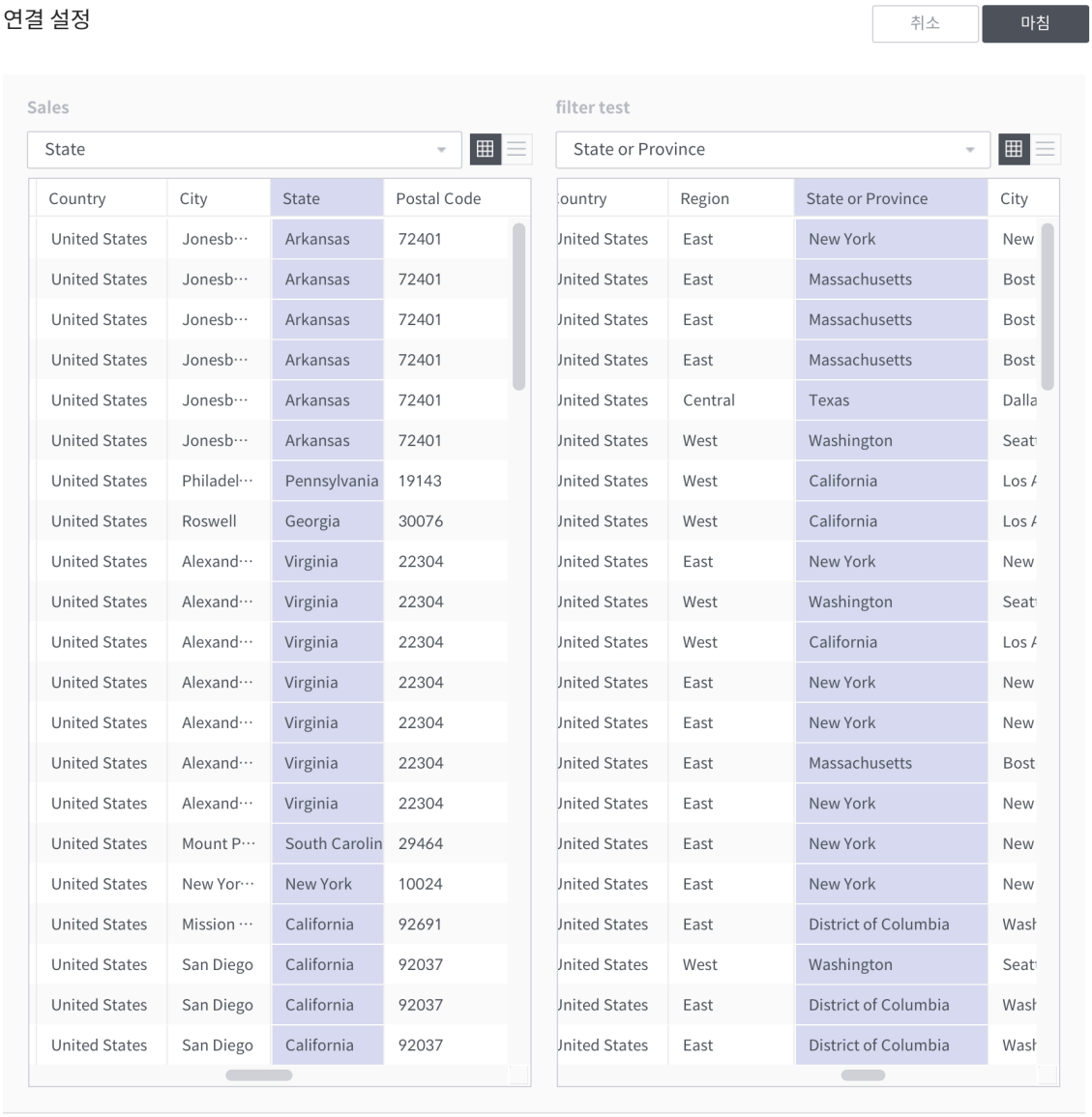
   -  : Click on it to add a new master data source.
   - **Edit association:** Click on it to edit an established data source association.

   Settings panel for individual master data sources (click one of the ovals corresponding to a master data source on the diagram to open it)

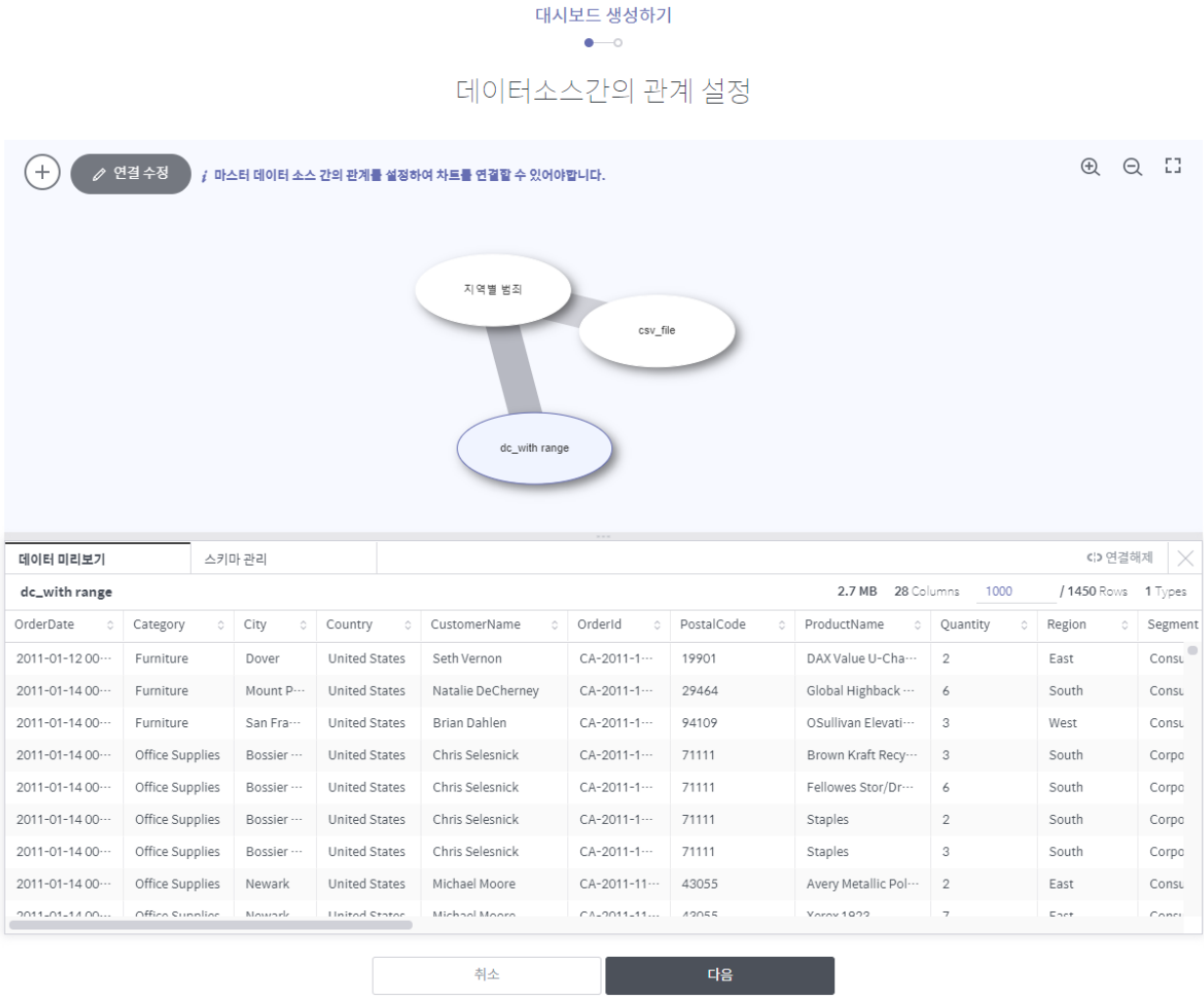
   - **Data preview:** Displays the data table resulting from data source joins.
   - **Manage schema:** Allows you to manage joins to the selected data source (for a detailed procedure, refer to the next step).
   - **Unlink:** Click on it to remove the selected data source.
   -  : Click on it to close the panel.


데이터소스를 선택해 주세요
취소
마침
데이터소스 이름 검색
오픈 데이터만 보기
타입
전체
No.
데이터소스
타입
89 tet - test 연결형
88 kkk 수집형
87 estate 오픈 데이터 수집형
86 temp-rollup - a 수집형
85 mysql_preset_engine_dialog_single_all 수집형
84 us 500 오픈 데이터 수집형
83 테스트연임 수집형
82 sales_geo - Sales data (2011~2… 오픈 데이터 수집형
81 ecommerce-data - from kaggle 오픈 데이터 수집형
80 H analysis history 오픈 데이터 수집형
79 test)hisotry 수집형
78 s_history 수집형
77 헤더행테스트 수집형
76 temp-test 수집형
더보기
ecommerce-data
메타데이터 이름
설명 from kaggle
타입 수집형
공개설정 공개
생성일 2019-01-15
사이즈 16.26 MB
Rows 531,228
차원값 InvoiceNo
차원값 StockCode
차원값 Description
측정값 Quantity
차원값 InvoiceDate
측정값 UnitPrice
차원값 CustomerID
차원값 Country


대시보드 생성하기
데이터소스간의 관계 설정
Guide
연결 수정
마침
데이터 소스를 클릭하고 가장자리를 다른 데이터 소스로 드래그하여 연결합니다.
지역별 범죄
csv_file
dc_with range


연결 설정
취소
마침
Sales
State
Country City State Postal Code
United States Jonesb··· Arkansas 72401
United States Jonesb··· Arkansas 72401
United States Jonesb··· Arkansas 72401
United States Jonesb··· Arkansas 72401
United States Jonesb··· Arkansas 72401
United States Jonesb··· Arkansas 72401
United States Philadel··· Pennsylvania 19143
United States Roswell Georgia 30076
United States Alexand··· Virginia 22304
United States Alexand··· Virginia 22304
United States Alexand··· Virginia 22304
United States Alexand··· Virginia 22304
United States Alexand··· Virginia 22304
United States Alexand··· Virginia 22304
United States Alexand··· Virginia 22304
United States Alexand··· Virginia 22304
United States Mount P··· South Carolin 29464
United States New Yor··· New York 10024
United States Mission ··· California 92691
United States San Diego California 92037
United States San Diego California 92037
United States San Diego California 92037
filter test
State or Province
ountry Region State or Province City
Jnited States East New York New
Jnited States East Massachusetts Bost
Jnited States East Massachusetts Bost
Jnited States East Massachusetts Bost
Jnited States Central Texas Dalla
Jnited States West Washington Seat
Jnited States West California Los A
Jnited States West California Los A
Jnited States East New York New
Jnited States West Washington Seat
Jnited States West California Los A
Jnited States East New York New
Jnited States East New York New
Jnited States East Massachusetts Bost
Jnited States East New York New
Jnited States East New York New
Jnited States East New York New
Jnited States East District of Columbia Wash
Jnited States West Washington Seat
Jnited States East District of Columbia Wash
Jnited States East District of Columbia Wash


대시보드 생성하기
데이터소스간의 관계 설정
Guide
연결 수정
마침
데이터 소스를 클릭하고 가장자리를 다른 데이터 소스로 드래그하여 연결합니다.
지역별 범죄
csv_file
dc_with range


대시보드 생성하기
데이터소스간의 관계 설정
연결 수정
마스터 데이터 소스 간의 관계를 설정하여 차트를 연결할 수 있어야합니다.
지역별 범죄
csv_file
dc_with range
데이터 미리보기
스키마 관리
연결해제
dc_with range
2.7 MB 28 Columns 1000 / 1450 Rows 1 Types
OrderDate Category City Country CustomerName OrderId PostalCode ProductName Quantity Region Segment
2011-01-12 00··· Furniture Dover United States Seth Vernon CA-2011-1··· 19901 DAX Value U-Cha··· 2 East Consu
2011-01-14 00··· Furniture Mount P··· United States Natalie DeCherney CA-2011-1··· 29464 Global Highback ··· 6 South Consu
2011-01-14 00··· Furniture San Fra··· United States Brian Dahlen CA-2011-1··· 94109 OSullivan Elevati··· 3 West Consu
2011-01-14 00··· Office Supplies Bossier ··· United States Chris Selesnick CA-2011-1··· 71111 Brown Kraft Recy··· 3 South Corpo
2011-01-14 00··· Office Supplies Bossier ··· United States Chris Selesnick CA-2011-1··· 71111 Fellowes Stor/Dr··· 6 South Corpo
2011-01-14 00··· Office Supplies Bossier ··· United States Chris Selesnick CA-2011-1··· 71111 Staples 2 South Corpo
2011-01-14 00··· Office Supplies Bossier ··· United States Chris Selesnick CA-2011-1··· 71111 Staples 3 South Corpo
2011-01-14 00··· Office Supplies Newark United States Michael Moore CA-2011-11··· 43055 Avery Metallic Pol··· 2 East Consu
취소
다음

7. To join one of the master data sources to other data sources, click the corresponding oval on the diagram → click the **Manage Schema** tab on the panel at the bottom → click **+ Add a data source to join**.

8. Refer to the description below to set up data joins.

   - **Master data source:** Displays information on the master data source to which you want to join another data source.
   - **Datasource to join:** Select a data source to be joined to the master data source.
   - **Add to join keys**: A join key defines the join relationship between the master and slave data sources in each column. Select a column to be joined from each data source, and click this button to add a new join key. For this, the two columns must be of the same data type.
   - **Join type:** Select how to join and transform a data source. To help you understand, each join type is explained below using the following tables as an example.

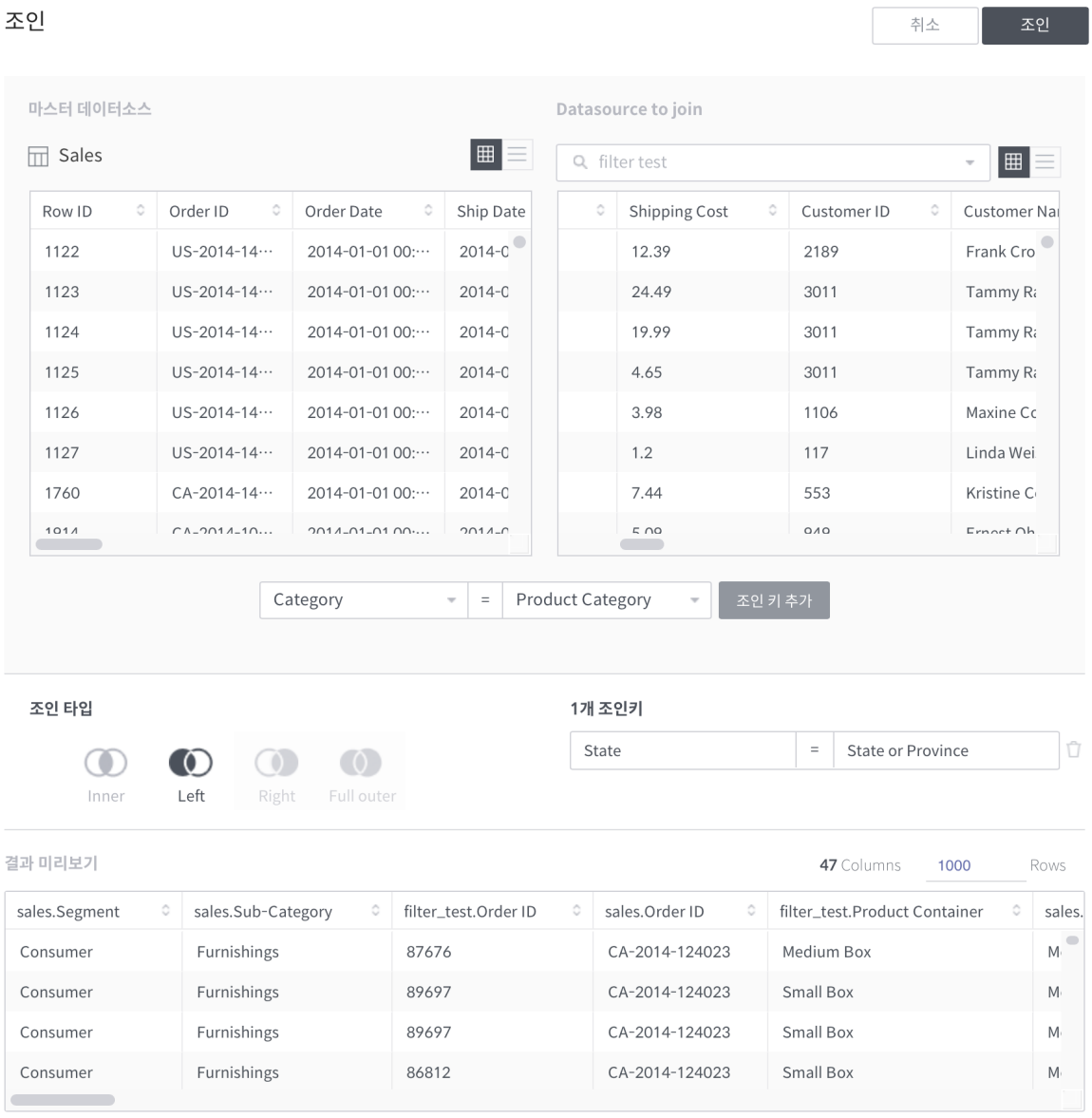

조인
취소
조인
마스터 데이터소스
Sales
Row ID
Order ID
Order Date
Ship Date
1122
US-2014-14···
2014-01-01 00:···
1123
1124
1125
1126
1127
1760
CA-2014-14···
Datasource to join
filter test
Shipping Cost
Customer ID
Customer Na
12.39
2189
Frank Cro
24.49
3011
Tammy Ra
19.99
4.65
3.98
1106
Maxine Co
1.2
117
Linda Wei
7.44
553
Kristine C
Category
=
Product Category
조인 키 추가
조인 타입
Inner
Left
Right
Full outer
1개 조인키
State
=
State or Province
결과 미리보기
47 Columns
1000
Rows
sales.Segment
sales.Sub-Category
filter_test.Order ID
sales.Order ID
filter_test.Product Container
Consumer
Furnishings
87676
CA-2014-124023
Medium Box
89697
Small Box
89697
Small Box
86812
Small Box

Table 1: Master data source

Product name (join key)	Price
A	$22.11
B	$9.23
C	$8.99
D	$10.10

Table 2: Data source to be joined

Product name (join key)	Sales
B	100
D	200
E	50

– **Inner:** Imports those records of each data source whose join key column values are present also in the other data source's join key column, joins them, and stores the joined records in the resulting table. (Intersection between two data sources)

Product name (join key)	Price	Sales
B	$9.23	100
D	$10.10	200

– **Left:** Imports those records of the right data source (data source to be joined) whose join key column values are present also in the join key column of the left data source (master data source to join), joins them to the left data source records, and stores the joined records in the resulting table. Those records from the right data source whose join key column values are not present in the left data source are discarded.

Product name (join key)	Price	Sales
A	$22.11	null
B	$9.23	100
C	$8.99	null
D	$10.10	200

– **Right:** Imports those records of the left data source (master data source to join) whose join key column values are present also in the join key column of the right data source (data source to be joined), joins them to the right data source records, and stores the joined records in the resulting table. Those records from the left data source whose join key column values are not present in the right data source are discarded.

Product name (join key)	Price	Sales
B	$9.23	100
D	$10.10	200
E	$null	50

– **Full Outer:** Imports all records from both data sources, join them, and stores the joined records in the resulting table. (Union between two data sources)

Product name (join key)	Price	Sales
A	$22.11	null
B	$9.23	100
C	$8.99	null
D	$10.10	200
E	null	50

- **Preview results:** Displays the data table resulting from data source joins.

9. Confirm the information on the imported data source, enter the **Name** and **Description**, and click **Done** to create a new dashboard.

10. The new dashboard will be added to the workbook home. Click the dashboard to display its contents.

5.2.2 Change dashboard size and layout

Click **Edit Dashboard** on the basic dashboard page to go to a page for editing the configuration of the dashboard. In this page, you can add a widget, edit the dashboard, set the hierarchy and change the layout.

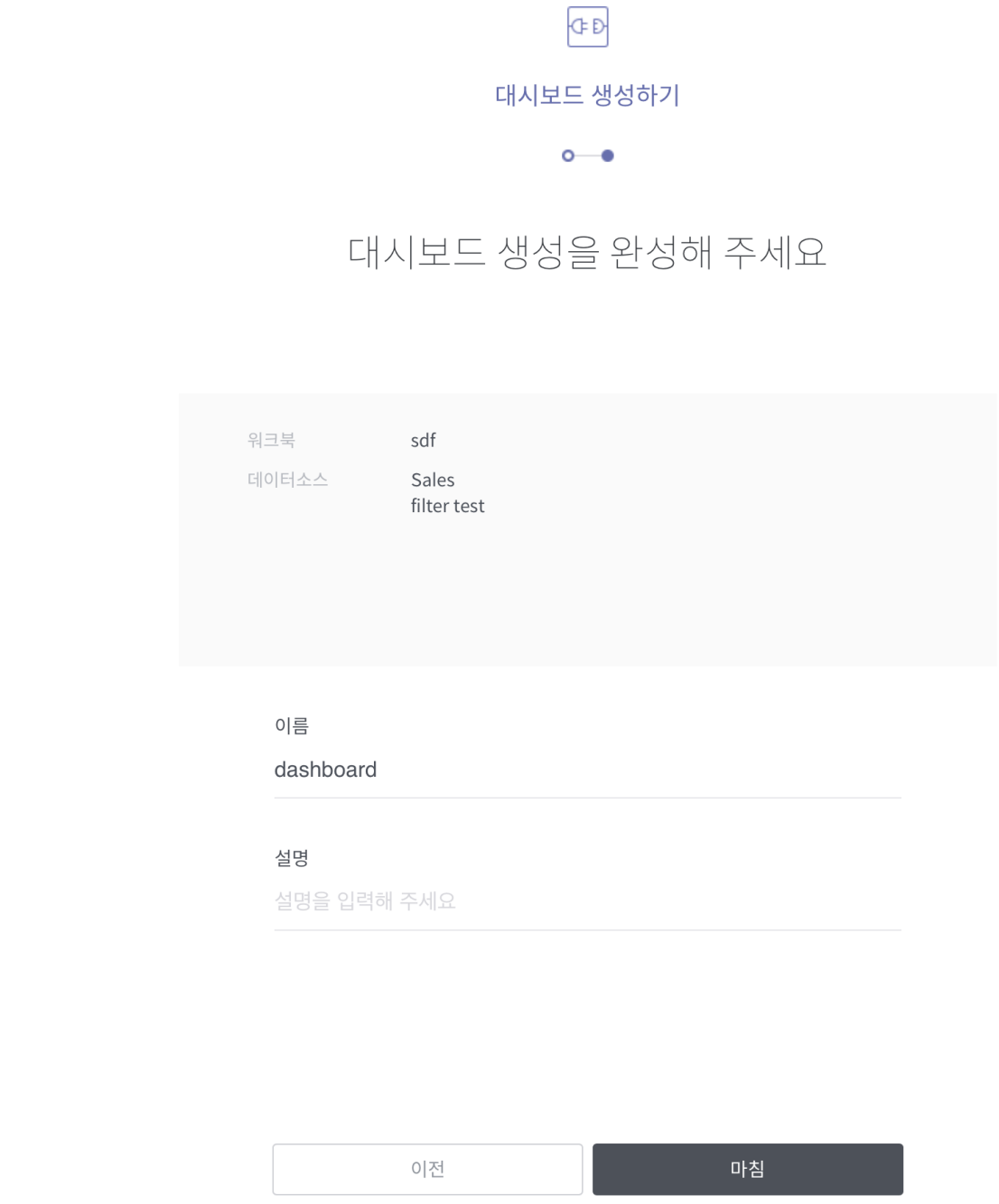
Dashboard widget arrangement settings

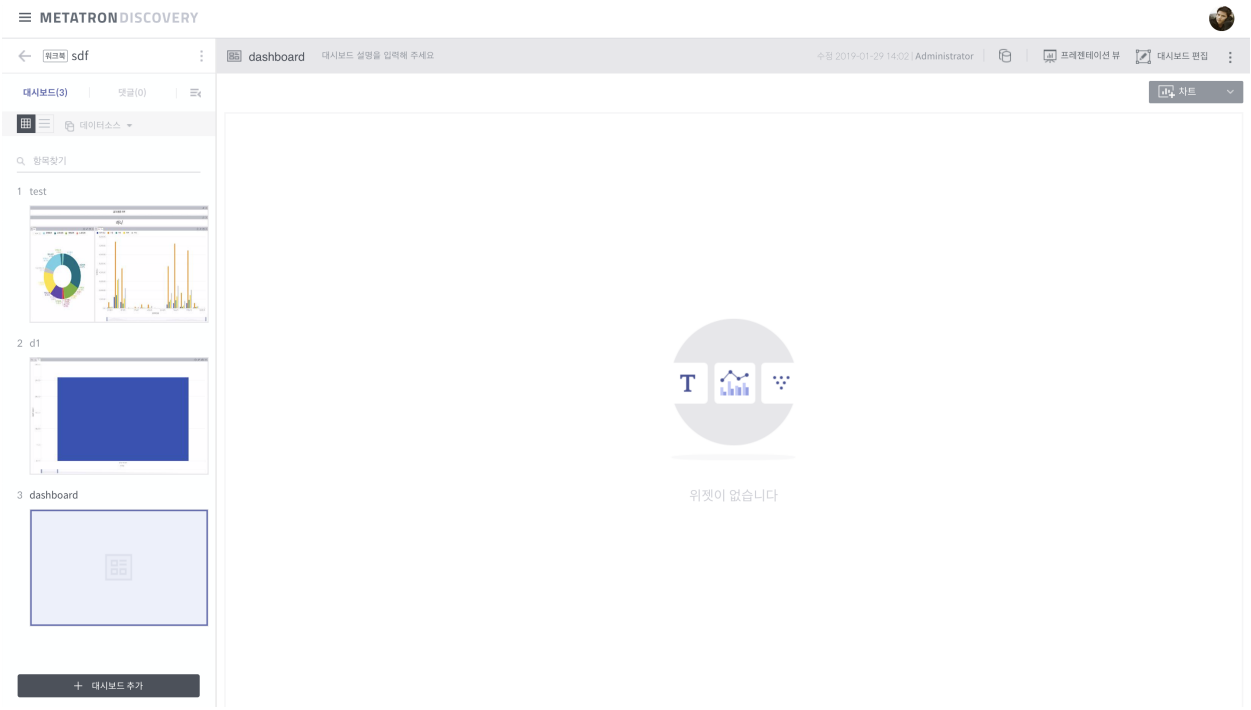
1. **Change widget location:** Drag the title of a widget to move the widget.
2. **Adjust widget width:** Move the distance between widgets to adjust their widths.
3. **Add a widget to the display area:** Drag a widget from the widget list on the right panel to the left widget display area to add the widget to the display area.
4. **Delete a widget from the display area:** Click the X button on a widget shown in the widget display area to delete the widget from the display area.

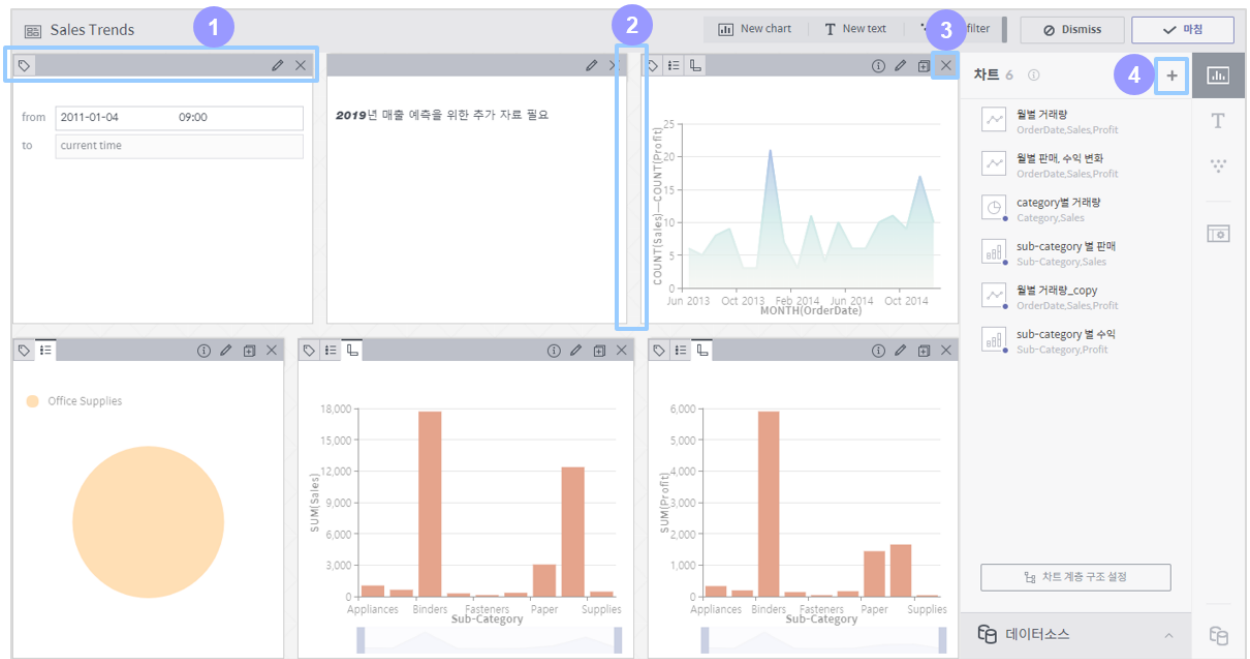
Chart widget panel

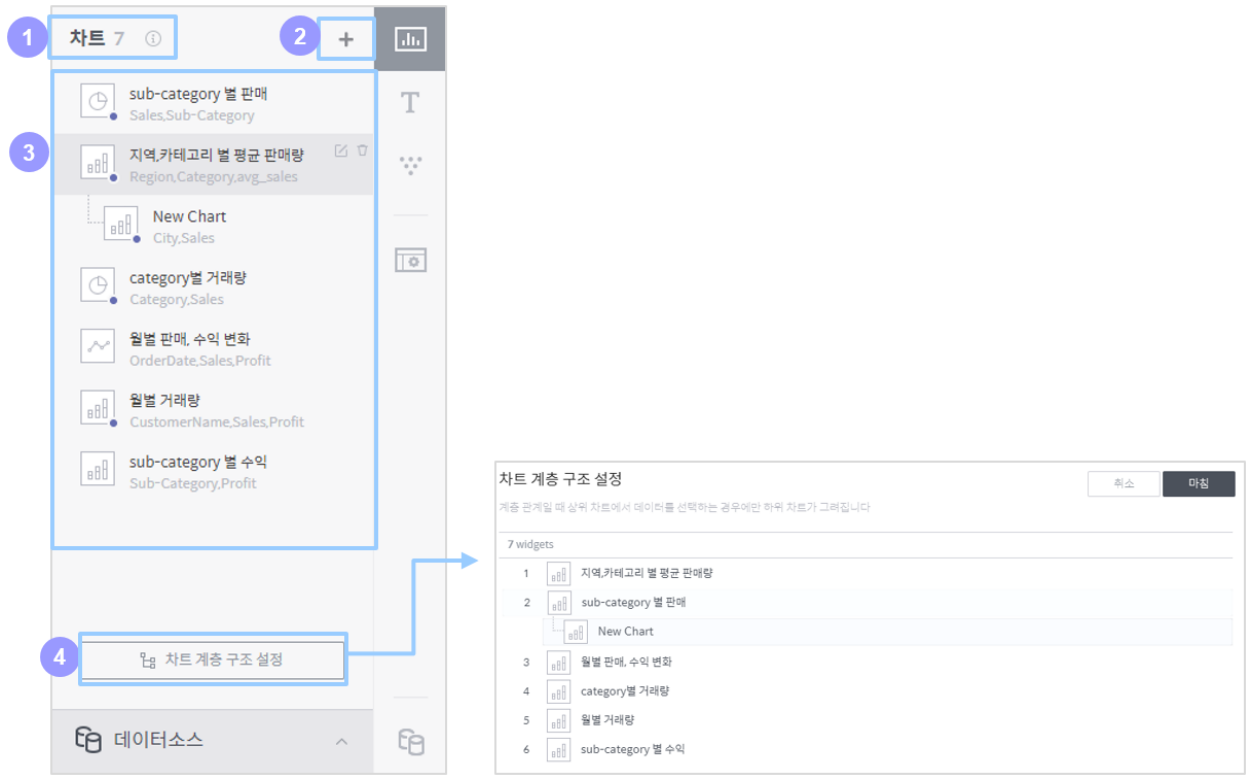
On the chart widget panel, you can add/edit/delete a chart in the dashboard.

1. **Number of chart widgets:** Displays how many chart widgets are registered in the dashboard.
2. **Add a chart widget:** Click on it to create a new chart widget in the dashboard.
3. **Chart widget list:** Lists chart widgets registered in the dashboard. Hover the mouse over a widget to display the edit and delete icons. Drag a widget to the widget display area to display the widget in the display area.
4. **Set chart hierarchy:** Click on it to set parent/child relationships between charts in the dashboard. Selecting a data item from the parent chart filters the child chart by the selection. To set a hierarchy, drag the chart to be set as a child under the chart to be set as a parent. Once you finish setting the chart hierarchy, the chart menu is restructured accordingly.


대시보드 생성하기
대시보드 생성을 완성해 주세요
워크북
sdf
데이터소스
Sales
filter test
이름
dashboard
설명
설명을 입력해 주세요
이전
마침


METATRONDISCOVERY
sdf
dashboard
대시보드(3)
댓글(0)
데이터소스
항목찾기
1 test
2 d1
3 dashboard
대시보드 추가
차트
위젯이 없습니다


Sales Trends
1
2
3
4
New chart
New text
filter
Dismiss
마침
from 2011-01-04 09:00
to current time
2019년 매출 예측을 위한 추가 자료 필요
COUNT(Sales),COUNT(Profit)
MONTH(OrderDate)
Office Supplies
SUM(Sales)
SUM(Profit)
Sub-Category
Appliances
Binders
Fasteners
Paper
Supplies
차트 6
월별 거래량
월별 판매, 수익 변화
category별 거래량
sub-category 별 판매
월별 거래량_copy
sub-category 별 수익
차트 계층 구조 설정
데이터소스



Text widget panel

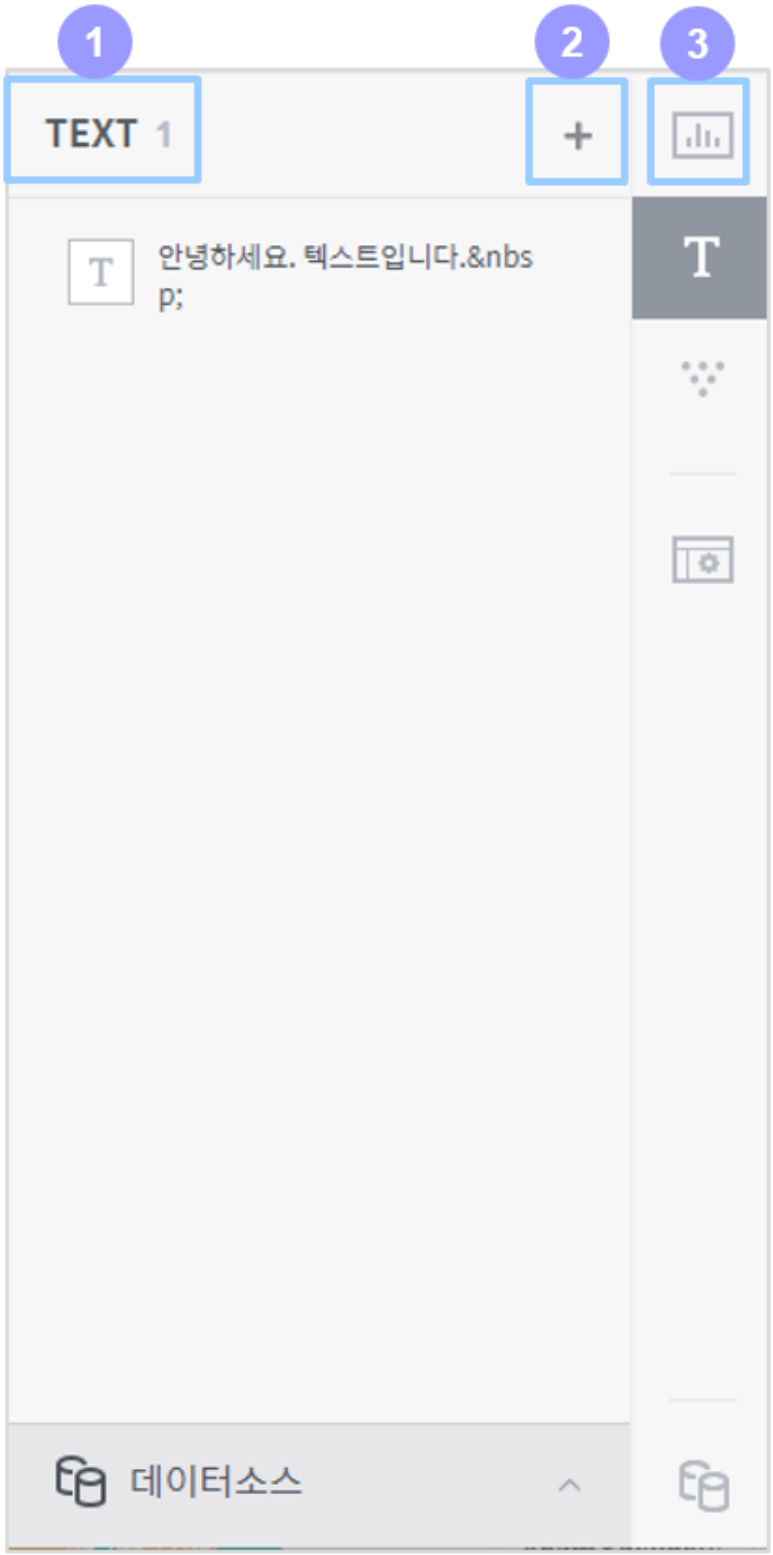
On the text widget panel, you can add/edit/delete a text widget in the dashboard.

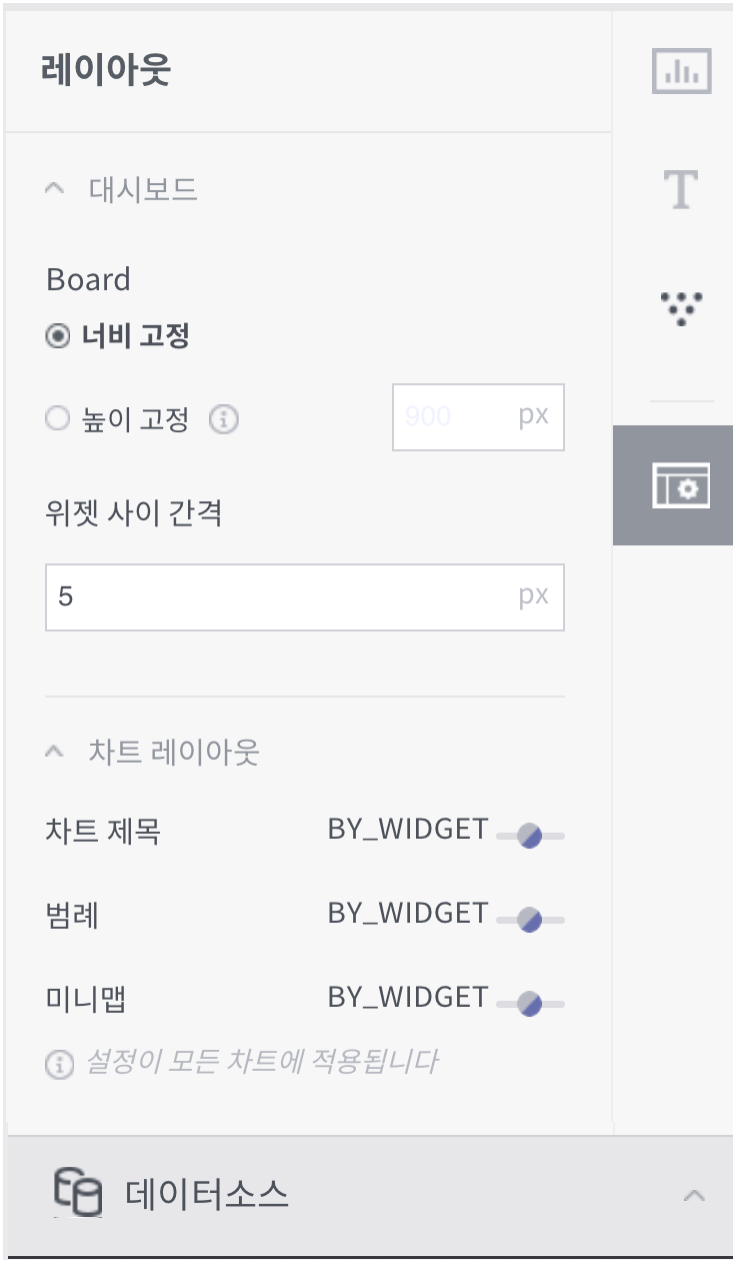
1. **Number of text widgets:** Displays how many text widgets are registered in the dashboard.
2. **Add a text widget:** Click on it to create a new text widget in the dashboard.
3. **Text widget list:** Lists text widgets registered in the dashboard. Hover the mouse over a widget to display the edit and delete icons. Drag a widget to the widget display area to display the widget in the display area.

Layout panel

On the layout panel, you can adjust some settings on how to arrange widgets and display each widget in the widget display area.

- **Set board height**
  - **Fix to screen:** Maximizes the height of the dashboard to fill the screen.
  - **Fix to height:** Set the height of the dashboard to a specific pixel value.
  - **Margin between widgets:** Sets the margin between widgets in the widget display area.
- **Chart title:** Sets whether to display the title of each chart and filter widget in the widget display area.
- **Legend:** Sets whether to display a legend for each chart widget in the widget display area.
- **Mini-map:** Sets whether to display a mini-map for each chart widget in the widget display area.


1
2
3
TEXT 1
+
T
안녕하세요. 텍스트입니다.&nbs
p;
T
데이터소스


레이아웃
대시보드
Board
너비 고정
높이 고정
900
px
위젯 사이 간격
5
px
차트 레이아웃
차트 제목
BY_WIDGET
범례
BY_WIDGET
미니맵
BY_WIDGET
설정이 모든 차트에 적용됩니다
데이터소스

Data source panel

In the data source panel, you can view and edit information on connected data sources, as well as add column filters easily. Click on a filter icon on a dimension or measure on the right-hand side to add a filter.

Please note that the filters you can apply or clear here are global filters applied to the entire dashboard, and those applied or cleared in the chart editor are all chart filters.

5.2.3 Check data sources in a dashboard

Click the  button on the basic dashboard page to display a dialog box displaying information about the data source used in the dashboard. At the top-left corner, you can choose the data source that you want to view. This dialog box consists of three tabs (Data grid, Column detail, Dashboard data information).

Data grid tab

Displays all values in the data source.

Column details tab

Displays details about each column of the data source.

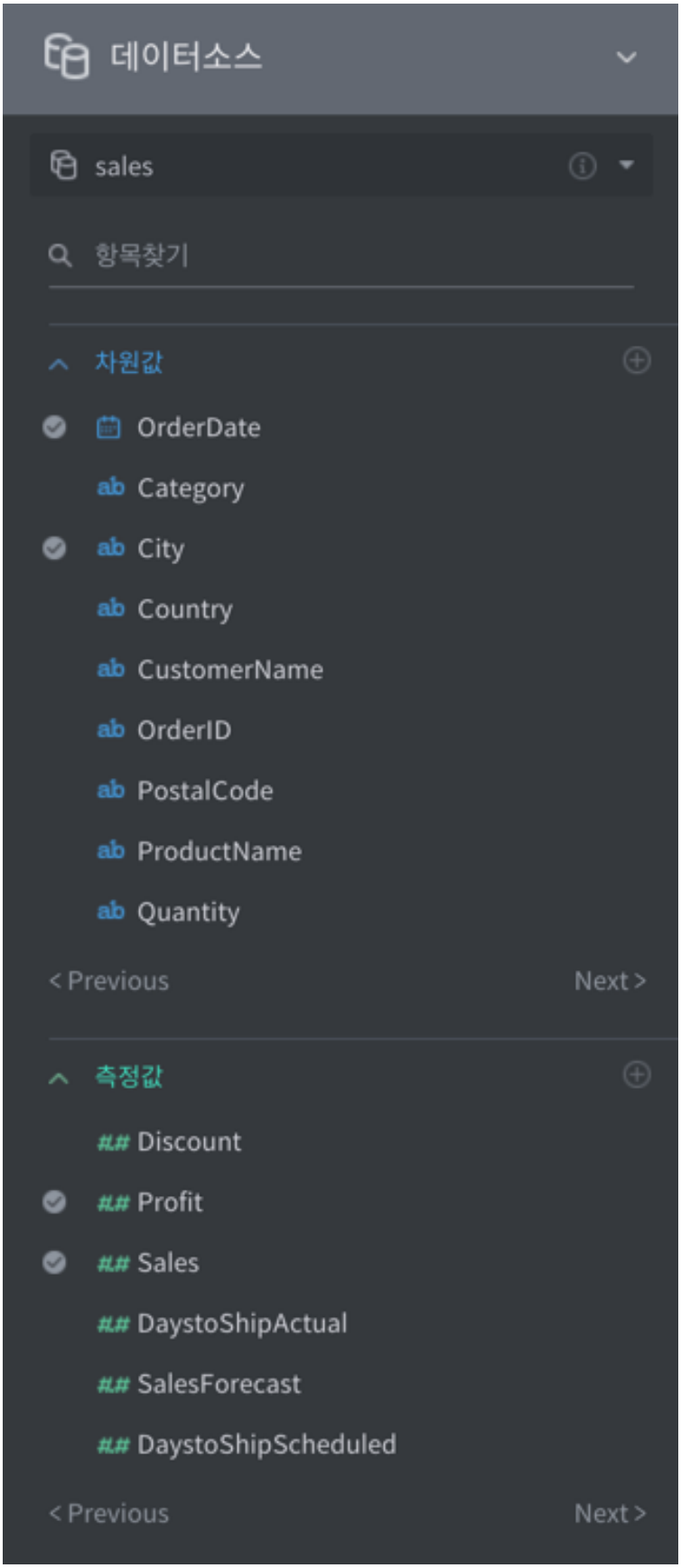
Dashboard data information tab

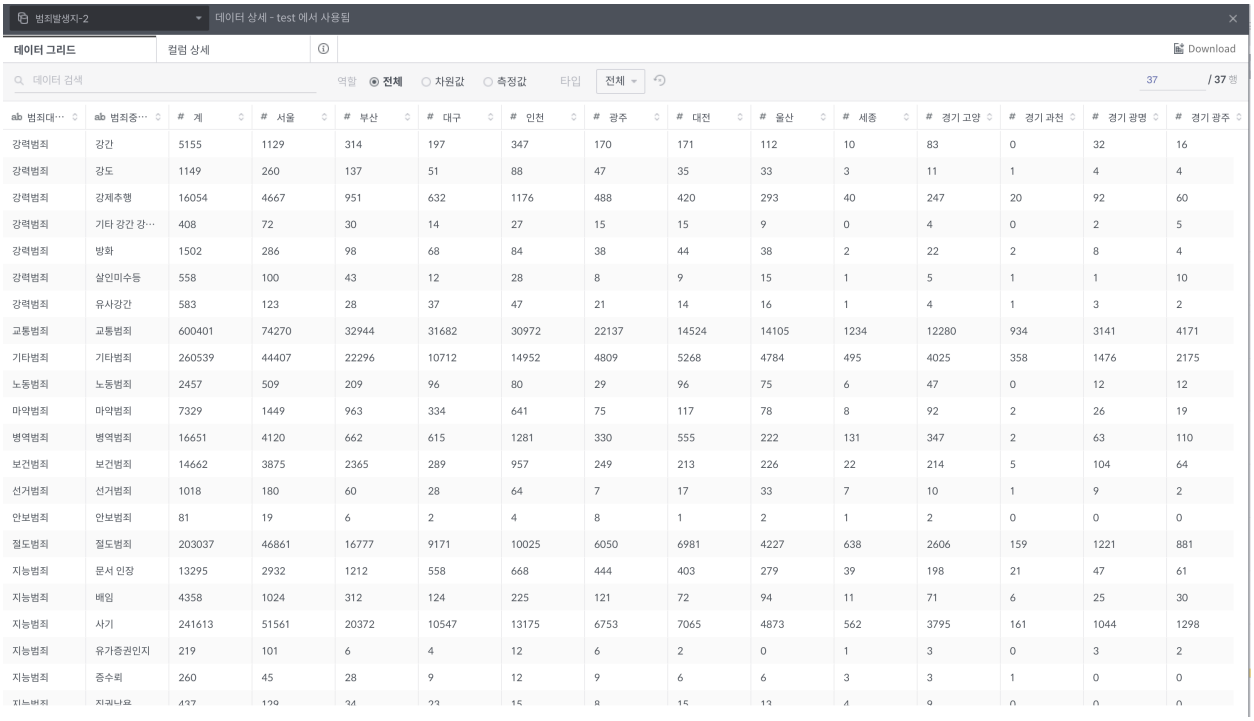
Displays an overview of the data source.

5.2.4 Presentation with a dashboard

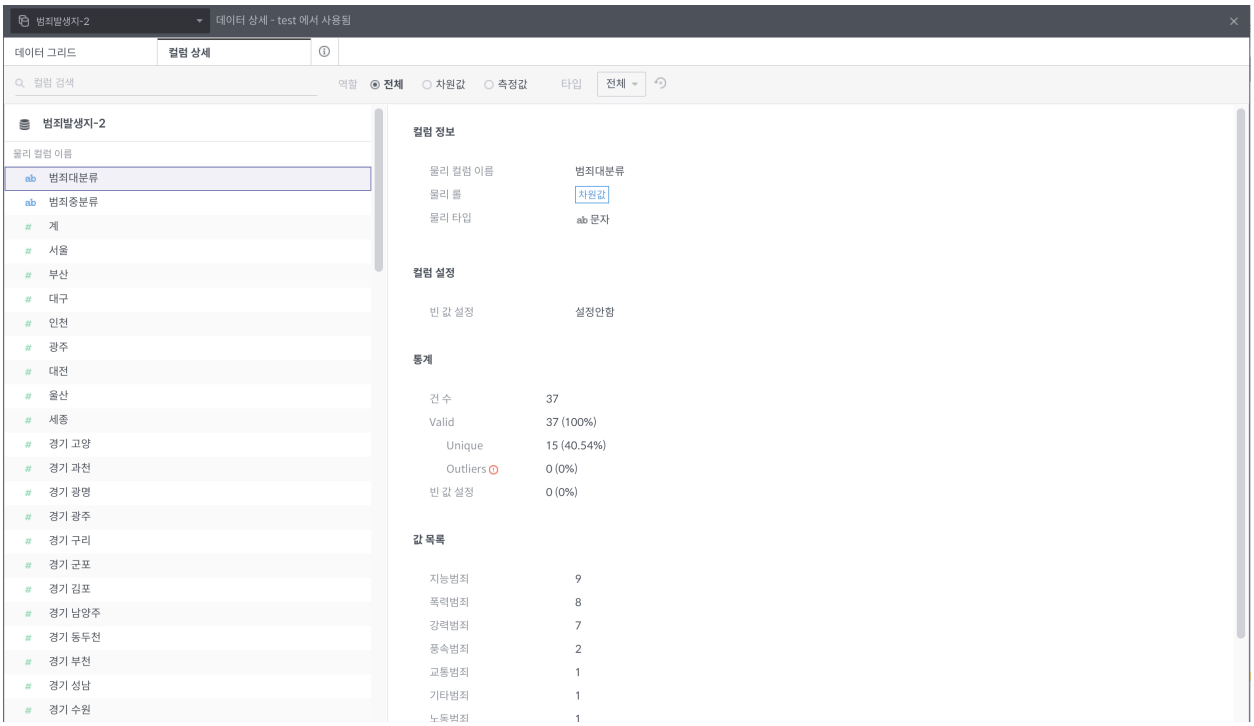
Click **Presentation view** on the basic dashboard page to view workbook dashboards with a presentation UI. In this mode, you can easily report and share data analytics results.

1. **Name:** Name of the current dashboard.
2. **Slide navigation:** Each circle represents a different dashboard in the workbook. For example, if you click the 4th circle, the 4th dashboard slide will be displayed with that circle highlighted.
3. **Auto slide show settings:** Select a duration for each slide and click PLAY to start an auto slide show.
4. **Exit:** Closes the presentation view and returns to the workbook/dashboard basic page.


데이터소스
sales
항목찾기
차원값
OrderDate
Category
City
Country
CustomerName
OrderID
PostalCode
ProductName
Quantity
< Previous
Next >
측정값
Discount
Profit
Sales
DaystoShipActual
SalesForecast
DaystoShipScheduled
< Previous
Next >

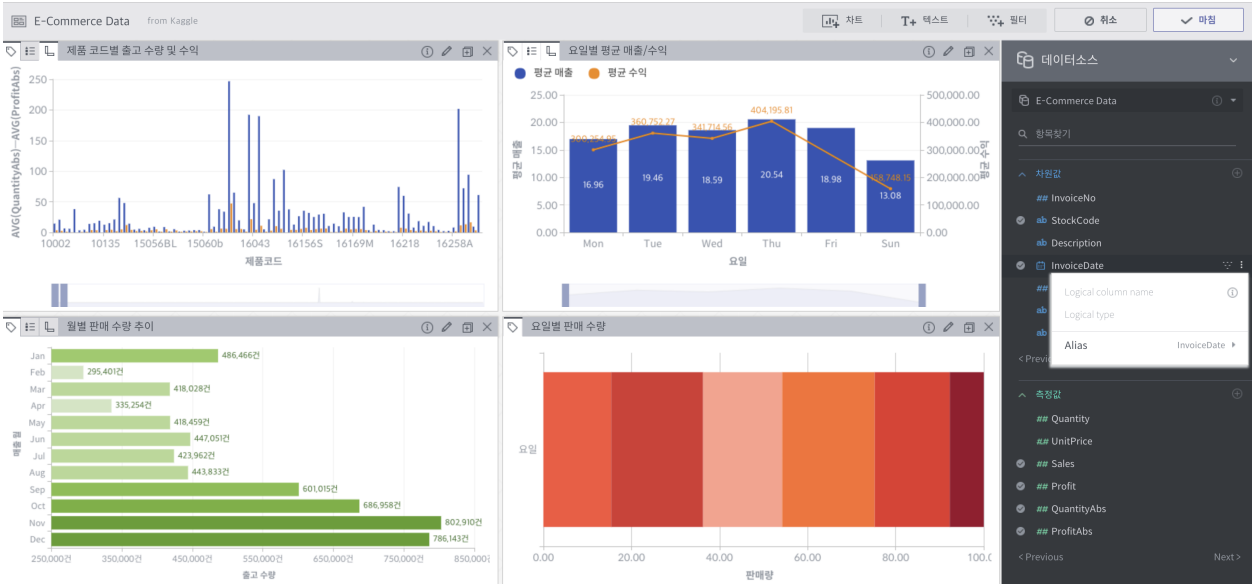


범죄대…	범죄중…	계	서울	부산	대구	인천	광주	대전	울산	세종	경기 고양	경기 과천	경기 광명	경기 광주
강력범죄	강간	5155	1129	314	197	347	170	171	112	10	83	0	32	16
강력범죄	강도	1149	260	137	51	88	47	35	33	3	11	1	4	4
강력범죄	강제추행	16054	4667	951	632	1176	488	420	293	40	247	20	92	60
강력범죄	기타 강간 강…	408	72	30	14	27	15	15	9	0	4	0	2	5
강력범죄	방화	1502	286	98	68	84	38	44	38	2	22	2	8	4
강력범죄	살인미수등	558	100	43	12	28	8	9	15	1	5	1	1	10
강력범죄	유사강간	583	123	28	37	47	21	14	16	1	4	1	3	2
교통범죄	교통범죄	600401	74270	32944	31682	30972	22137	14524	14105	1234	12280	934	3141	4171
기타범죄	기타범죄	260539	44407	22296	10712	14952	4809	5268	4784	495	4025	358	1476	2175
노동범죄	노동범죄	2457	509	209	96	80	29	96	75	6	47	0	12	12
마약범죄	마약범죄	7329	1449	963	334	641	75	117	78	8	92	2	26	19
병역범죄	병역범죄	16651	4120	662	615	1281	330	555	222	131	347	2	63	110
보건범죄	보건범죄	14662	3875	2365	289	957	249	213	226	22	214	5	104	64
선거범죄	선거범죄	1018	180	60	28	64	7	17	33	7	10	1	9	2
안보범죄	안보범죄	81	19	6	2	4	8	1	2	1	2	0	0	0
절도범죄	절도범죄	203037	46861	16777	9171	10025	6050	6981	4227	638	2606	159	1221	881
지능범죄	문서 인장	13295	2932	1212	558	668	444	403	279	39	198	21	47	61
지능범죄	배임	4358	1024	312	124	225	121	72	94	11	71	6	25	30
지능범죄	사기	241613	51561	20372	10547	13175	6753	7065	4873	562	3795	161	1044	1298
지능범죄	유가증권인지	219	101	6	4	12	6	2	0	1	3	0	3	2
지능범죄	증수뢰	260	45	28	9	12	9	6	6	3	3	1	0	0




대시보드 데이터 정보
사이즈 40.6 KB
요약 90 컬럼 / 37 Rows
역할 2 Dimensions / 88 Measures
컬럼 타입 ab 2 # 88
데이터소스 범죄발생지-2

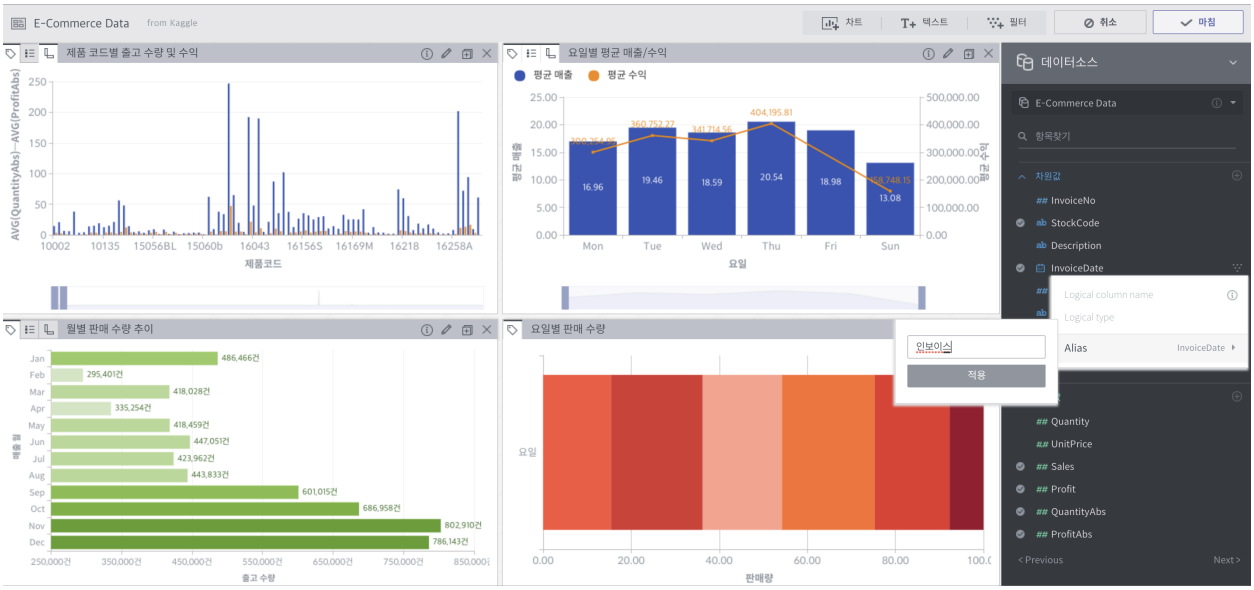

1
E-Commerce Data
2
3
3 sec.
PLAY
4
제품 코드별 출고 수량 및 수익
요일별 평균 매출/수익
평균 매출
평균 수익
월별 판매 수량 추이
요일별 판매 수량


E-Commerce Data
from Kaggle
차트
텍스트
필터
취소
마침
제품 코드별 출고 수량 및 수익
요일별 평균 매출/수익
월별 판매 수량 추이
요일별 판매 수량
데이터소스
E-Commerce Data
항목찾기
차원값
InvoiceNo
StockCode
Description
InvoiceDate
Logical column name
Logical type
Alias
InvoiceDate
측정값
Quantity
UnitPrice
Sales
Profit
QuantityAbs
ProfitAbs
< Previous
Next >

5.2.5 Renaming columns

Hover the mouse over a column name on the data source panel in dashboard editing mode, and click the icon on the right to check the alias of the column.

Hover the mouse over the alias to open a window where you can enter a new column name. After entering the name, click **Apply** to see the change applied.



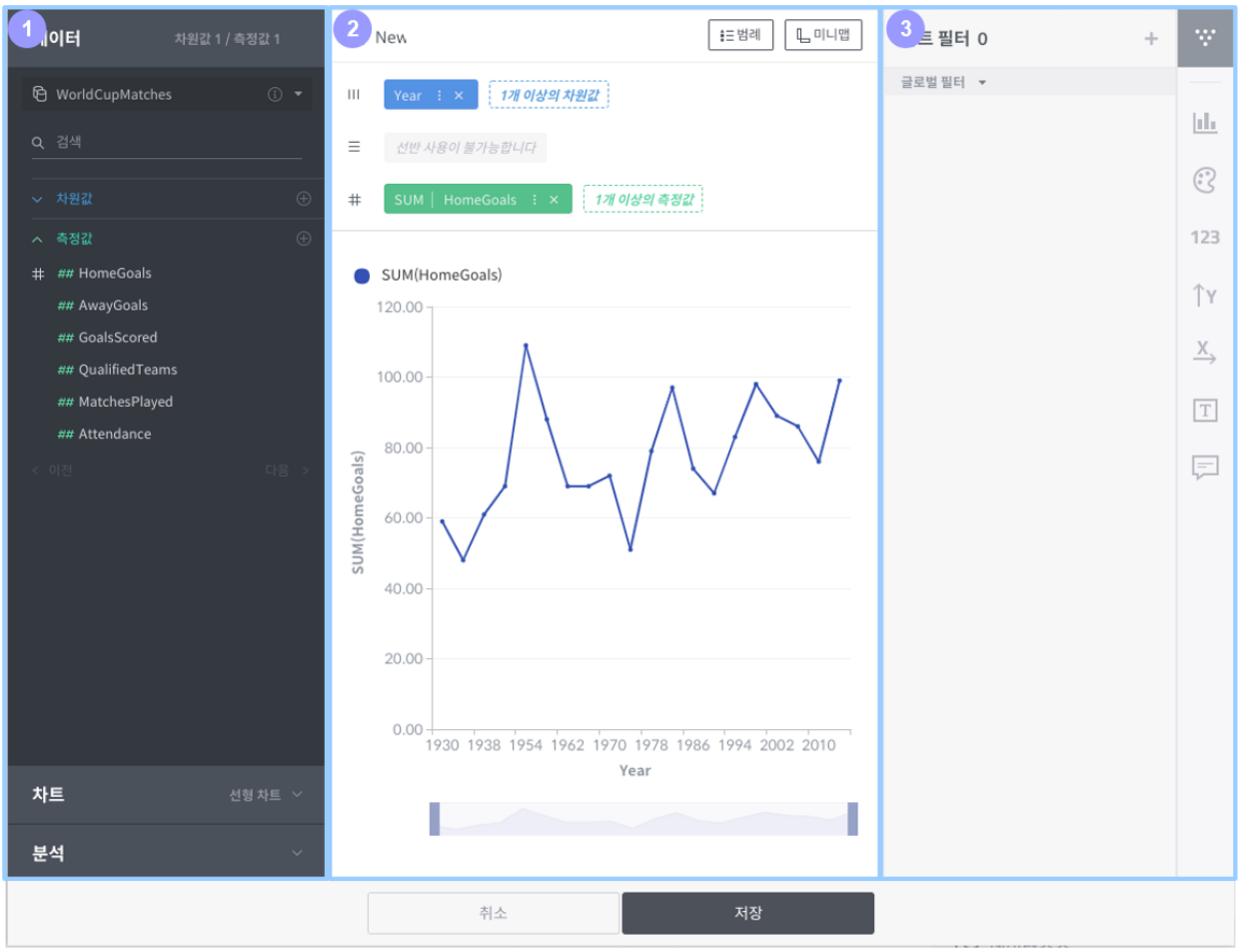
5.3 Chart

Charts that analyze and visualize data are the main components of a dashboard. This section describes some concepts that you need to understand to create a chart for data analytics, as well as the elements that make up the chart configuration UI.

The chart home is divided into the following three sections:

1. **Column/chart selection section:** This section is so organized that you can create a chart step by step. You can either choose columns under the Data menu to have appropriate chart types suggested, or select a chart type under the Chart menu before choosing data columns. In addition, you can configure some analytics settings under the Analytics menu.
2. **Visualization section:** This section is composed of the shelves onto which columns are put and the visualization area where the chart is displayed. Once data and a chart type are selected in the column/chart selection section, the chart is drawn in this area.
3. **Option section:** Used to customize the appearance and display of the chart. Depending on the chart type, the option section may include the filter, palette, axis, numeric format, and chart format areas.

In the subsequent subsections, we will explain how to use this user interface to create and manage various types of charts.

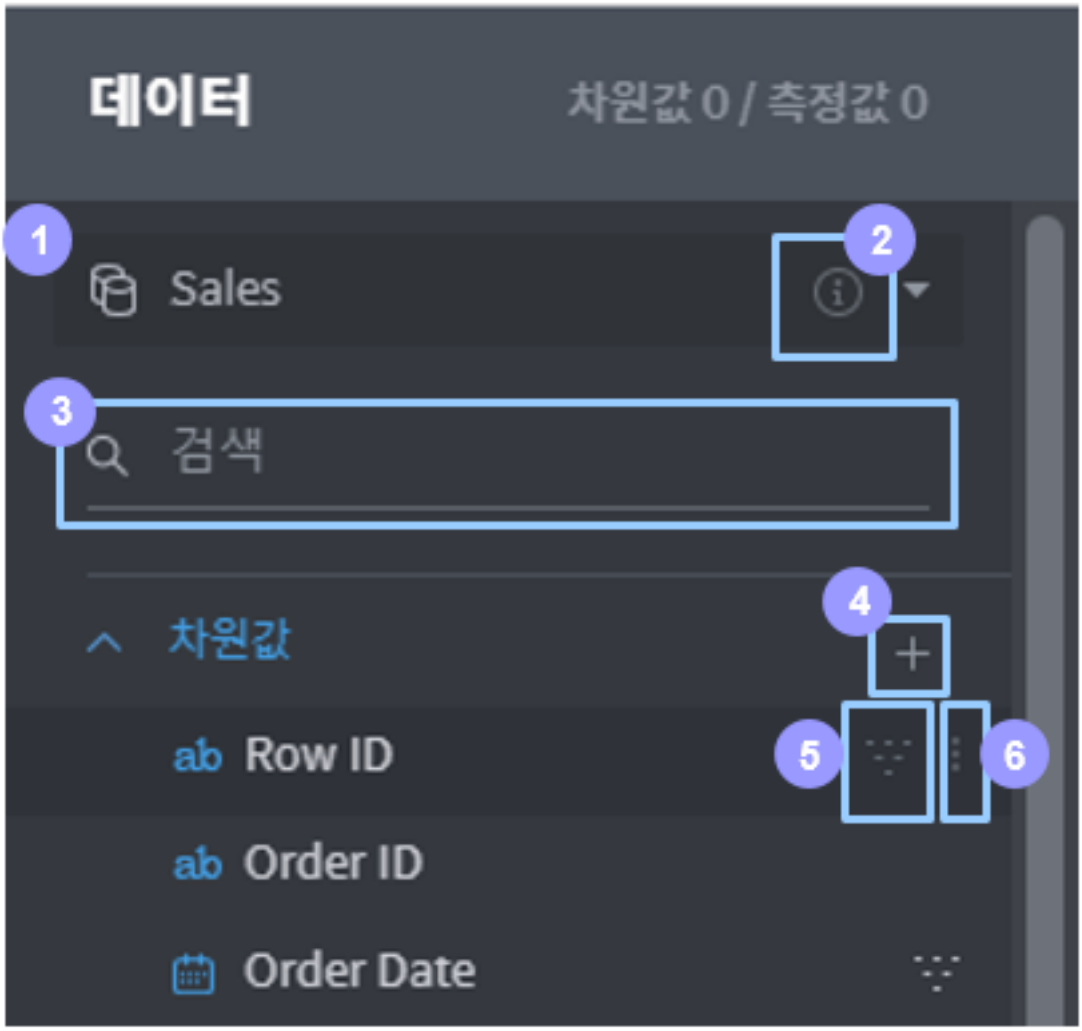

1
차원값 1 / 측정값 1
WorldCupMatches
검색
차원값
측정값
HomeGoals
AwayGoals
GoalsScored
QualifiedTeams
MatchesPlayed
Attendance
이전
다음
차트
선형 차트
분석
2
New
범례
미니맵
Year
1개 이상의 차원값
선반 사용이 불가능합니다
SUM | HomeGoals
1개 이상의 측정값
SUM(HomeGoals)
120.00
100.00
80.00
60.00
40.00
20.00
0.00
SUM(HomeGoals)
1930 1938 1954 1962 1970 1978 1986 1994 2002 2010
Year
3
필터 0
글로벌 필터
123
취소
저장

5.3.1 Data column list

The columns listed in the data column list are categorized into "dimensions" and "measures." For the concept of dimensions and measures, refer to *"Dimensions" and "Measures"*.

Structure of the data column list

In the data column list, you can view and edit information on connected data sources, as well as add or remove column filters easily.



1. **Select/set data source:** Allows you to select a data source or configure its associations and joins.
2. **Data details:** Click on it to pop up a dialog box displaying information about the selected data source.
3. **Search by column name:** Searches the column list by name.
4. **Add custom column:** Click on it to open the dialog box to create a new column by combining/processing data source columns. Custom columns are commonly used throughout the dashboard.

5. **Apply/clear filter:** Hover the mouse over a column to display this button. Click on it to apply a chart filter to the column, and click again to clear the chart filter. For columns to which a filter is applied, the  icon is displayed regardless of the mouse position.
6. **More:** Hover the mouse over a column to display this button. It is used to check additional information on the column and set an alias.
   - : Click on it to pop up a dialog box displaying a summary of the column and its data values.
   - **Logic column name:** Shows the logical name of the column.
   - **Type:** Shows the logical type of the column.
   - **Alias:** Sets a column alias. A regular column name can contain only alphanumeric characters and a limited number of special characters with no spaces allowed. Therefore, setting an alias may help to identify the column for convenient analytics work. Aliases are commonly used throughout the dashboard.
   - **Value alias:** You can also set an alias for each data value in the column. Aliases are commonly used throughout the dashboard.

Add a custom column

Click the + button on the data source column list to open a dialog box for adding a custom column. By applying various formulas to existing columns of the data source, you can create a new column that helps create your desired chart.

1. **Column name:** Fill in a name for the custom column.
2. **Coding box:** Write a code for the custom column. Click a list from the column or formula list below to type your selection in this box automatically.
3. **Add column:** Lists the columns of the data source. Click a column in the list to automatically type your selection in the coding box.
4. **Add formula:** Lists the formulas supported by Metatron. Click a formula in the list to type your selection in the coding box automatically, with the text cursor relocated to where a parameter needs to be inserted. For details on each formula's purpose, use, and examples, see the help box on the right.

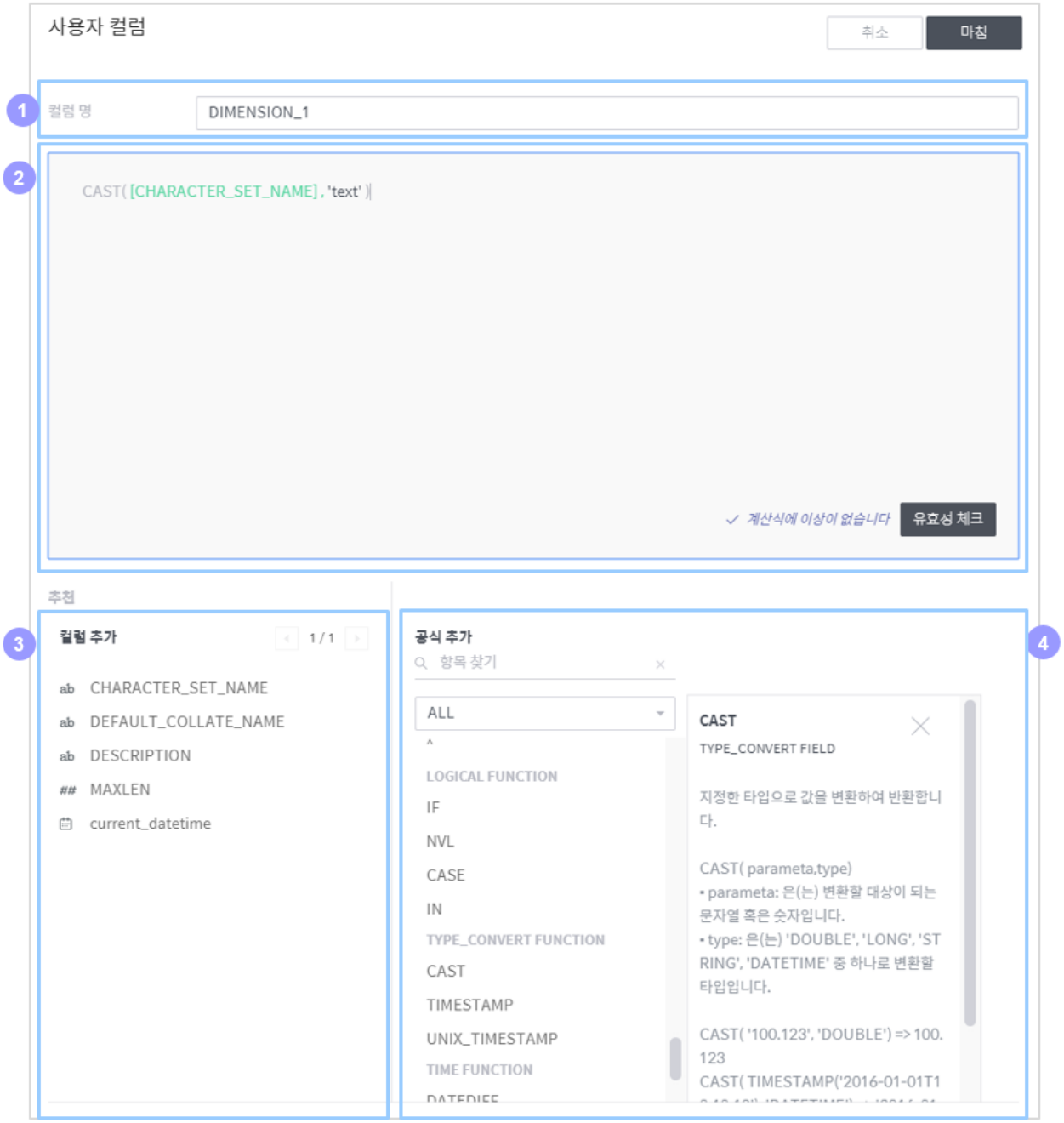
5.3.2 Draw a chart (pivoting)

What is pivoting

Pivoting is a process of grouping the given table by specific columns, thereby helping the analyst view particular aspects of the source data in a graphic or tabular chart. This process includes selecting columns that contain meaningful data and placing them on the column/row/cross shelves.

In the example shown above, two dimension columns are placed on the column shelf and one measure column is placed on the cross shelf. The chart displays data resulting from the columns placed on the shelves in this way.

Mandatory/recommended column types for each shelf vary depending on the chart type. Selecting a chart type before placing columns on a shelf shows the necessary column types for each shelf.


사용자 컬럼
취소
마침
1
컬럼 명
DIMENSION_1
2
CAST([CHARACTER_SET_NAME] , 'text')
계산식에 이상이 없습니다
유효성 체크
추천
3
컬럼 추가
1 / 1
ab CHARACTER_SET_NAME
ab DEFAULT_COLLATE_NAME
ab DESCRIPTION
MAXLEN
current_datetime
4
공식 추가
항목 찾기
ALL
LOGICAL FUNCTION
IF
NVL
CASE
IN
TYPE_CONVERT FUNCTION
CAST
TIMESTAMP
UNIX_TIMESTAMP
TIME FUNCTION
CAST
TYPE_CONVERT FIELD
지정한 타입으로 값을 변환하여 반환합니다.
CAST(parameta,type)
• parameta: 은(는) 변환할 대상이 되는 문자열 혹은 숫자입니다.
• type: 은(는) 'DOUBLE', 'LONG', 'STRING', 'DATETIME' 중 하나로 변환할 타입입니다.
CAST('100.123', 'DOUBLE') => 100.123
CAST(TIMESTAMP('2016-01-01T1

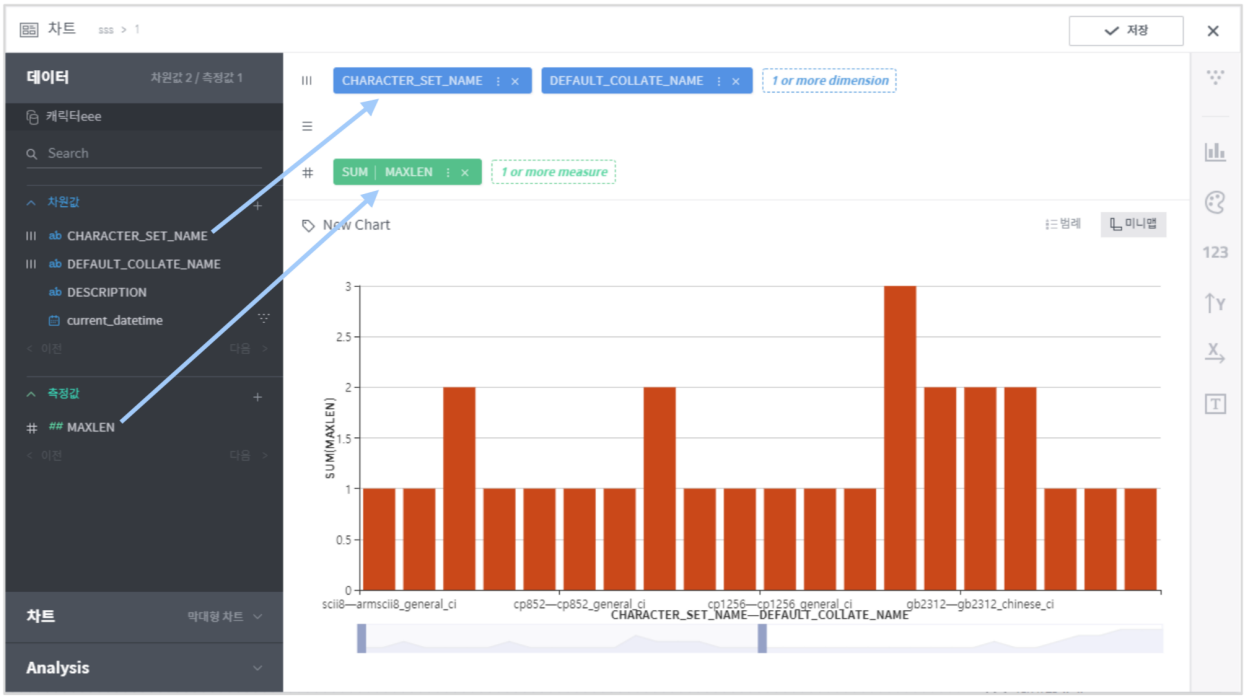
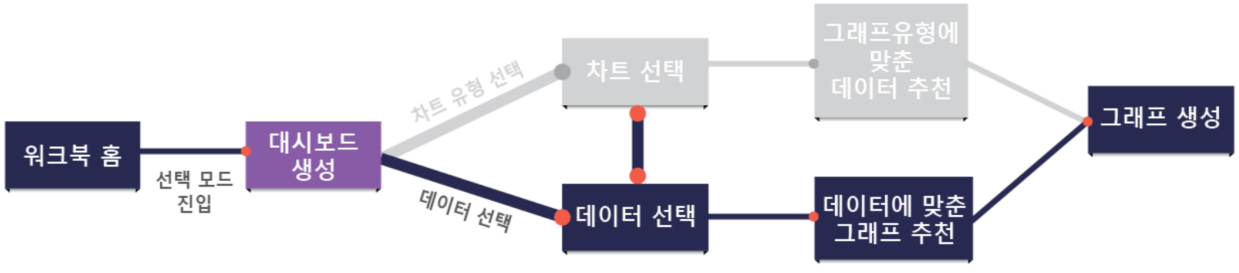
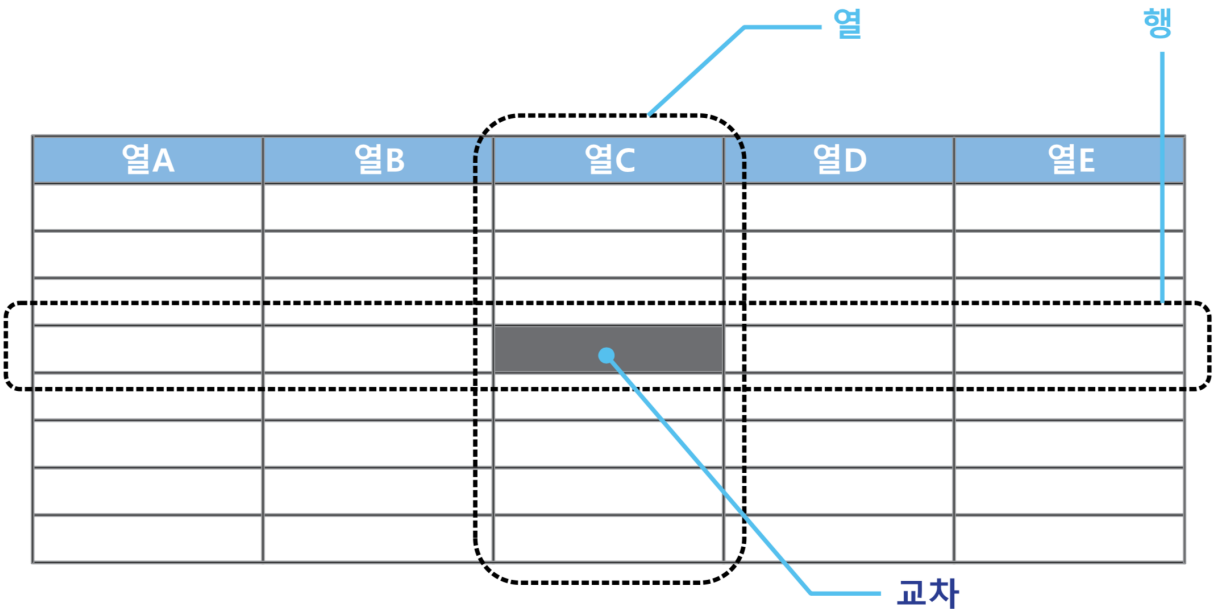


차트
저장
데이터
차원값 2 / 측정값 1
캐릭터eee
Search
차원값
CHARACTER_SET_NAME
DEFAULT_COLLATE_NAME
DESCRIPTION
current_datetime
측정값
MAXLEN
CHARACTER_SET_NAME
DEFAULT_COLLATE_NAME
1 or more dimension
SUM | MAXLEN
1 or more measure
New Chart
범례
미니맵
SUM(MAXLEN)
3
2.5
2
1.5
1
0.5
0
sci8—armsci8_general_ci
cp852—cp852_general_ci
cp1256—cp1256_general_ci
gb2312—gb2312_chinese_ci
CHARACTER_SET_NAME—DEFAULT_COLLATE_NAME
차트
막대형 차트
Analysis

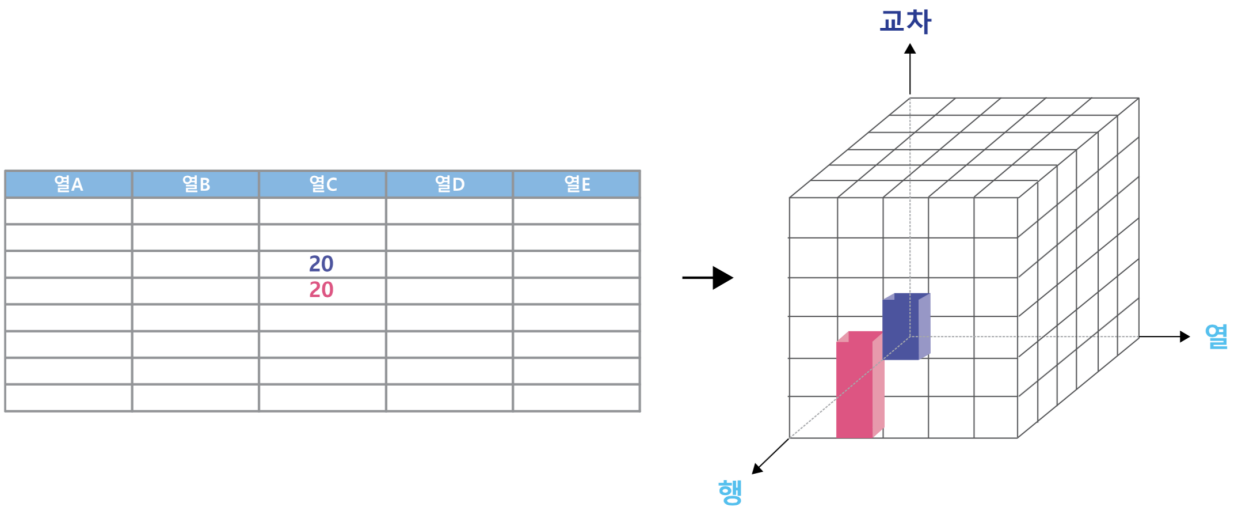

워크북 홈
선택 모드 진입
대시보드 생성
차트 유형 선택
데이터 선택
차트 선택
데이터 선택
그래프유형에 맞춘 데이터 추천
데이터에 맞춘 그래프 추천
그래프 생성

Column/row/cross shelves

Think of the structure of Excel to understand what column/row/cross shelves work for. As shown below, the crossing of each column and row cross contains a value.

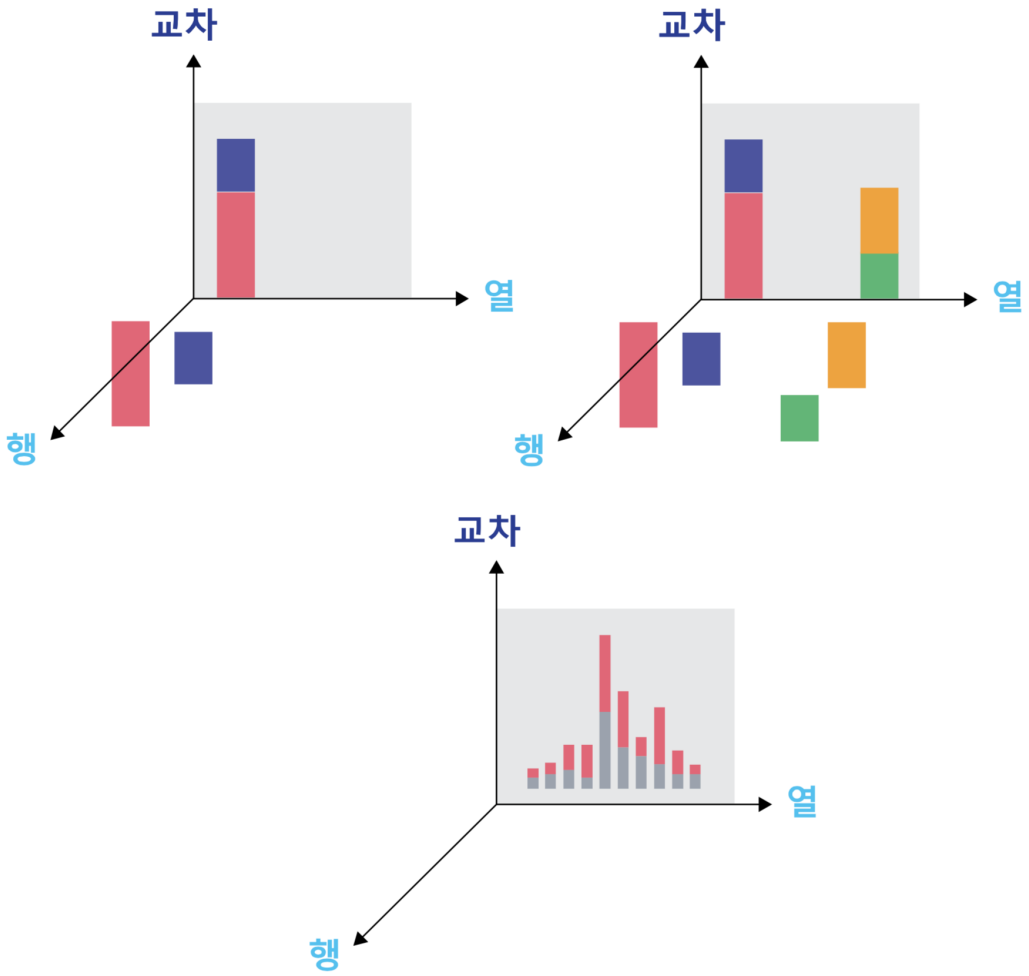


Whereas Excel shows data in a two-dimensional grid composed of columns, rows and crosses, Metatron is an OLAP data discovery tool capable of multidimensional data representation. In the following Metatron chart, the column, row, and crossing axes form a three-dimensional cube.



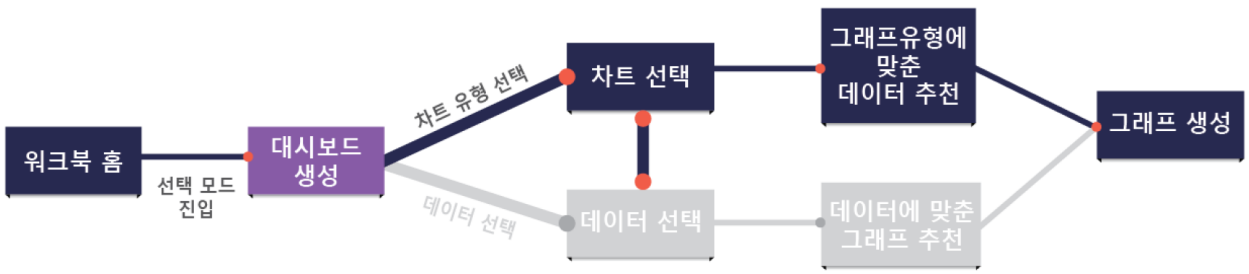
If the values of an Excel grid are displayed in a three-dimensional chart, each crossing value will be represented by a bar. However, Metatron needs to display such a chart two-dimensionally; for this, bars either in the same column or

in the same row get stacked at one point while remaining distinctive from one another. The resulting two-dimensional chart is shown in the gray area of the chart below.

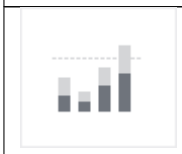
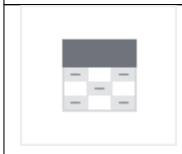
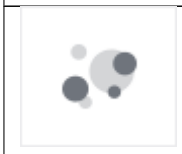


5.3.3 Select a chart type

Metatron Discovery provides about 20 types of charts. If you place columns on shelves before selecting a chart, suitable charts are highlighted in purple.

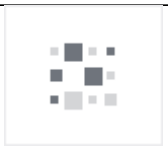
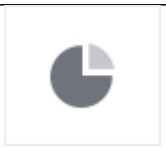
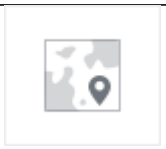
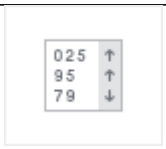


The table below summarizes conditions to create, uses, and examples for each chart.

Chart name/icon	Conditions to create	Characteristics	Uses	Examples
 Bar chart	Column: 1 or more dimensions / Cross: 1 or more measures	Compares the value of each item.	Used to compare groups or view trends over time. Very effective when the trend is significantly fluctuating.	Comparison between products regarding their sales and profits
 Table	Column or row: 1 or more dimensions / Cross: 1 or more measures	Displays the values of crossings between two dimensions as text.	Used to view measure values aggregated by certain criteria. Useful to check exact values rather than a visualization of them.	Sales details by year
 Line chart	Column: 1 or more dimensions / Cross: 1 or more measures	Displays data changes over time.	Used to view trends over time. If changes are moderate, a line chart is more effective than a bar chart.	Monthly sales trend
 Scatter chart	Column: 1 measure / Row: 1 measure / Cross: 1 or more dimensions	Displays relations between items.	Used to define relations between two parameters.	Relations between product sales and profits

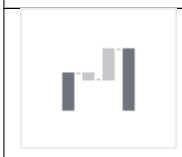
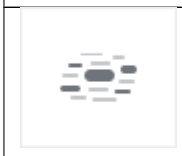
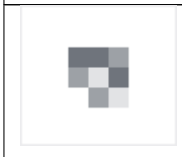
Continued on next page

Table 3 – continued from previous page

Chart name/icon	Conditions to create	Characteristics	Uses	Examples
 Heatmap	Column or row: 1 or more dimensions / Cross: 1 or more measures	Displays the values of crossings between two dimensions in colors and sizes at different points.	Used to provide an intuitive view of relations between two dimensions represented by colors and sizes. Similar to a table chart, but more of a visual type.	Sales of each product by region
 Pie chart	Cross: 1 or more dimensions, 1 or more measures	Shows how much each item accounts for.	Used to compare the compositions of something.	Comparison between web browsers regarding their market share
 Map view	Layer shelf: dimension (location attribute), 1 or more dimensions, 1 or more measures	Displays the data for each location on the map.	Used for intuitive comparisons of variables by using colors for each region. Used to emphasize visual elements.	Comparison of sales of each product by region
 KPI	Cross: 1 or more measures	Displays main indicators along with their trends.	Used to quickly convey information on an organization's current achievement.	An organization's performance index, such as how many customers have been brought in this year

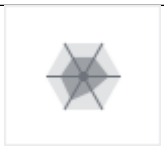
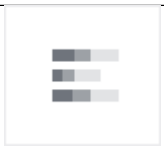
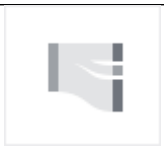
Continued on next page

Table 3 – continued from previous page

Chart name/icon	Conditions to create	Characteristics	Uses	Examples
 Box plot	Column: 1 or more dimensions / Row: 1 dimension / Cross: 1 measure	Indicates increase and decrease in value.	Used to compare groups regarding their share.	Proportion of flight delay accounted for by each airplane model
 Waterfall chart	Column: 1 time-dimension / Cross: 1 measure	Displays cumulative changes resulting from the increase or decrease in value for each time interval.	Used to emphasize increase and decrease in value over time.	Changes in the number of team members for a certain period; stock price trends
 Word cloud	Cross: 1 or more dimensions, 1 measure	Displays words sized in proportion to the number of mentions.	Used to summarize and emphasize important words.	Summary of the voices of customers
 Combo chart	Column: 1 or more dimensions / Cross: 2?4 measures	Compares data by combining bar and line charts.	Used to visualize different types of data simultaneously.	Simultaneous monitoring of product price and sales
 Treemap	Column: 1 dimension / Row: 1 or more dimensions / Cross: 1 measure	Displays hierarchical data using nested rectangles.	Used to visualize hierarchical data.	Monitoring of sales of products classified into major, medium, and minor categories.

Continued on next page

Table 3 – continued from previous page

Chart name/icon	Conditions to create	Characteristics	Uses	Examples
 Radar chart	Cross: 1 dimension, 1 or more measures	Displays different quantitative variables on axes starting from the same point.	Used for a visual comparison among different quantitative variables.	Product quality evaluation in five aspects.
 Network diagram	Subject shelf: 1 dimension / Target shelf: 1 dimension / Connecting shelf: 1 measure	Diagram connecting elements in dependence relations	Used to view data flows regarding where data elements are generated.	Monitoring the task flows of a project
 Gauge chart	Column: Row: 1 or more dimensions / Cross: 1 measure	Visualizes performance for the specified target.	Used to view the proportions of data elements.	Monitoring of profits by region
 Sankey diagram	Column: 3 or more dimensions / Cross: 1 measure	Displays the proportion of each data flow by the width of the connection line.	Used to monitor data flows and their respective sizes.	Monitoring energy flows in a factory

5.3.4 Chart style attributes

Once data is pivoted, an options menu is shown on the right of the screen to allow you to set the chart style. The composition of the menu varies with chart type. This section describes the settings used universally by all chart types and the "Common Setting" items for each chart type.

Chart style settings menu

This section describes how to configure the settings of the chart style settings menu. Note that not all the settings are shown for every chart type.

Color setting

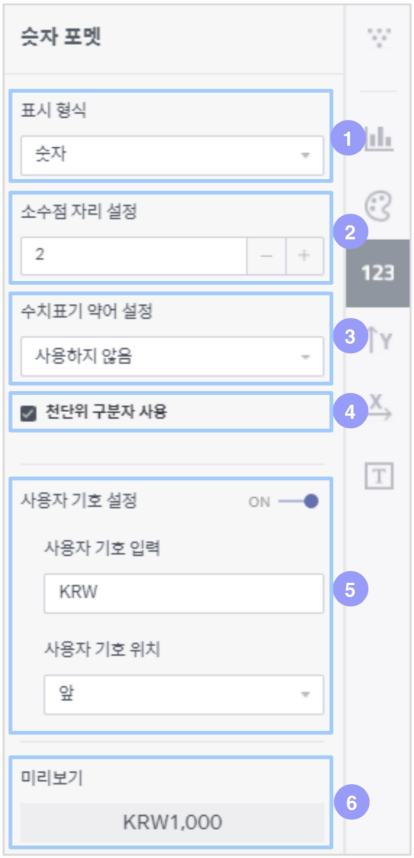
Defines various colors used in the chart.

1. **Graph color setting:** Set criteria to classify data on the chart by color, and select a coloring theme.
   - **Series:** Colors data elements differently with measures.
   - **Dimension:** Colors data elements differently with dimensions.
   - **Measure:** Colors elements differently with the size of each aggregate of measure values.
2. **Setting color range:** This setting is displayed when **Measure** is selected as the criterion to classify data by color. Set "ON" to set colors differently with each range of measure values. The measure data to be colored can be subdivided into as many ranges as you want, starting with the lowest one. To add a new range, adjust the upper limit of the highest range and click **Add new range**.

Number format

Defines how to display numerical text data on the chart graph. To use this function, turn on Show Axis Label in the Data Label Settings Menu.

1. **Format:** Select a display format for numeric values from among number, currency, percent, and exponent.
2. **Decimal place:** Set how many digits to display after the decimal point.
3. **Number abbreviations:** You can use K (thousands), M (millions), or B (billions) as an abbreviation for a large numeric value. Select **Automation** to automatically set the most proper symbol in accordance with the number of digits.



4. **Thousands separator:** Select whether to add thousands separators when displaying numeric data values.
5. **Customer symbol:** Insert a custom text before/after numeric data values.
6. **Preview:** Displays the result of the defined number format.

Y-axis setting (when chart type is vertical)

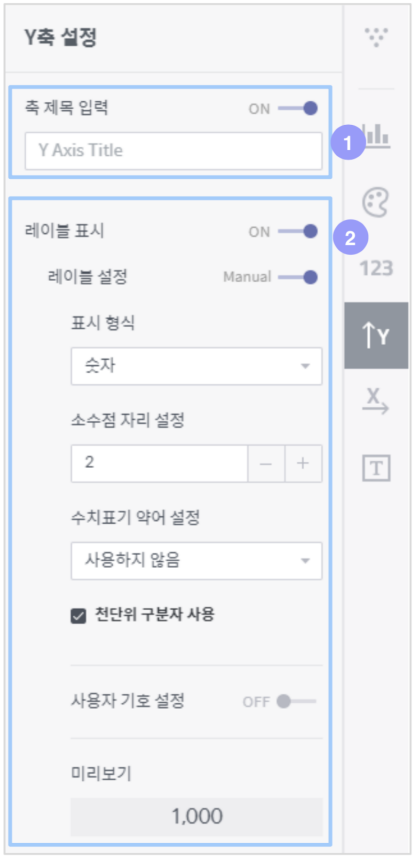
If you set the chart direction **Horizontal** in the Common Setting area, the settings are exchanged between X-axis and Y-axis.

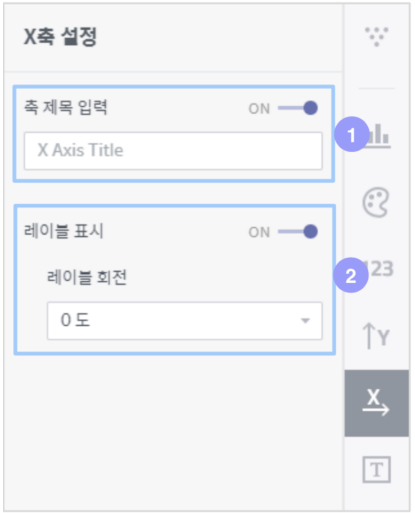
1. **Show axis title:** Used to set a title for the Y-axis of the chart. Disabling this function hides the title of the Y-axis.
2. **Show axis label:** Select whether or not to show the data labels on the Y-axis of the chart. Disabling this function hides the data labels on the Y-axis.
   - **Label setting:** Set the numeric format of the data labels on the Y-axis. Set automatic to import the settings of **Format** or manual to set specific format for the data labels on the Y-axis.

X-axis setting (when chart type is vertical)

Defines how to display the X-axis of the chart. If you set the chart direction **Horizontal** in the Common Setting area, the settings are exchanged between X-axis and Y-axis.

1. **Show axis title:** Used to set a title for the X-axis of the chart. Disabling this function hides the title of the X-axis.
2. **Show axis label:** Select whether or not to show the data labels on the X-axis of the chart. Disabling this function hides the data labels on the X-axis.

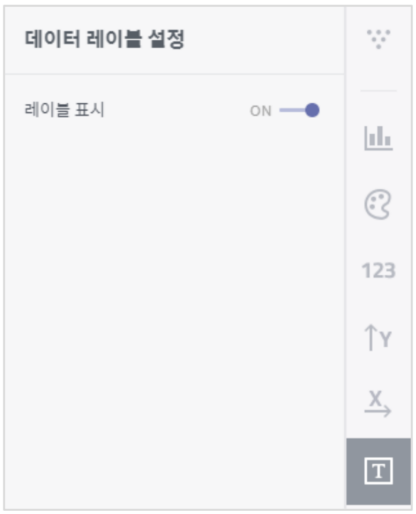

Y축 설정
축 제목 입력
ON
Y Axis Title
1
레이블 표시
ON
2
레이블 설정
Manual
표시 형식
숫자
소수점 자리 설정
2
수치표기 약어 설정
사용하지 않음
천단위 구분자 사용
사용자 기호 설정
OFF
미리보기
1,000
123
Y
X
T


X축 설정
축 제목 입력
ON
X Axis Title
1
레이블 표시
ON
2
레이블 회전
0 도
123
Y
X
T

- **Rotation:** Select an angle for the data labels on the X-axis from among 0, 45, and 90 degrees.

Data label setting

Selects whether to display the data values on the chart graph.



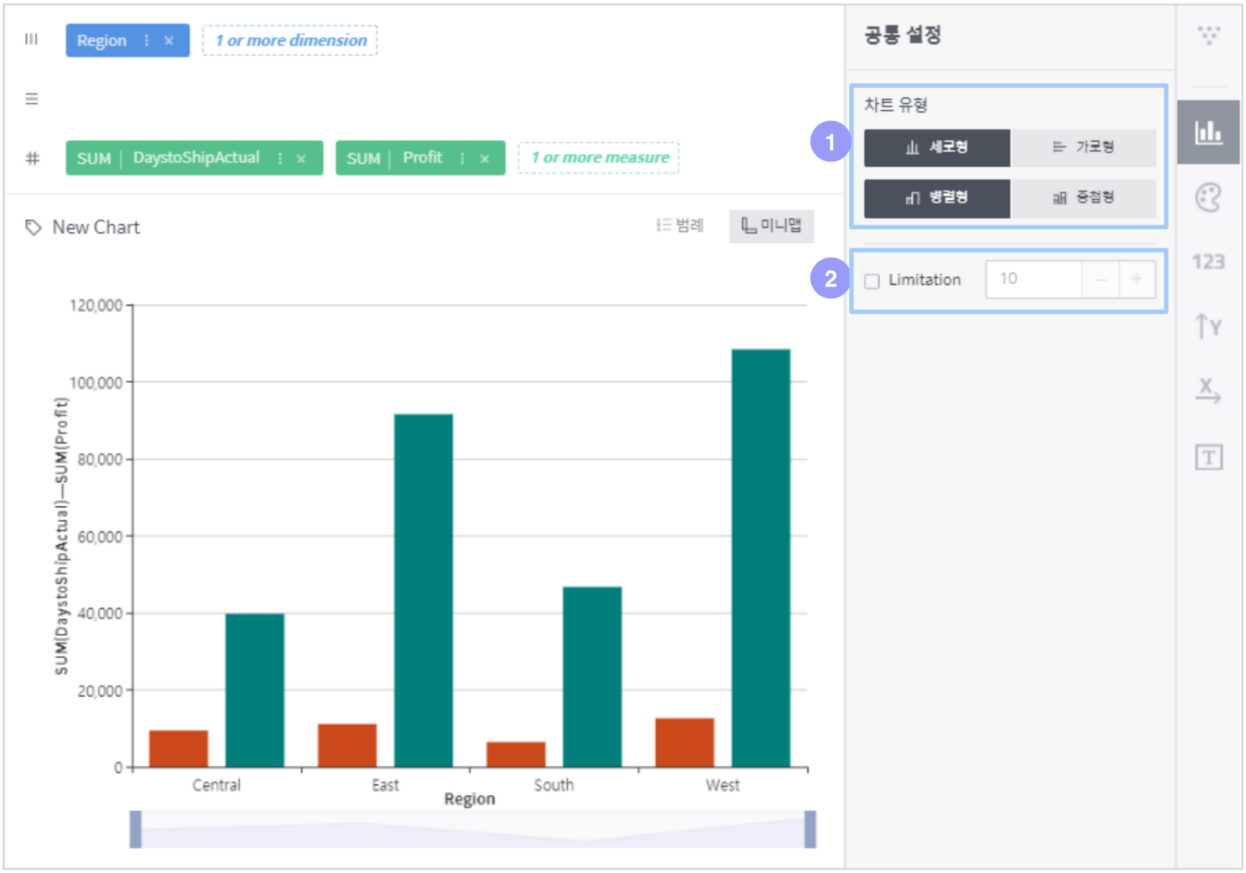
Common settings for each chart type

This section describes how to style the six most popular chart types (bar chart, table, line chart, scatter chart, heatmap, and pie chart).

Bar chart

This type of chart presents data values in each category of a dimension column with rectangular bars.

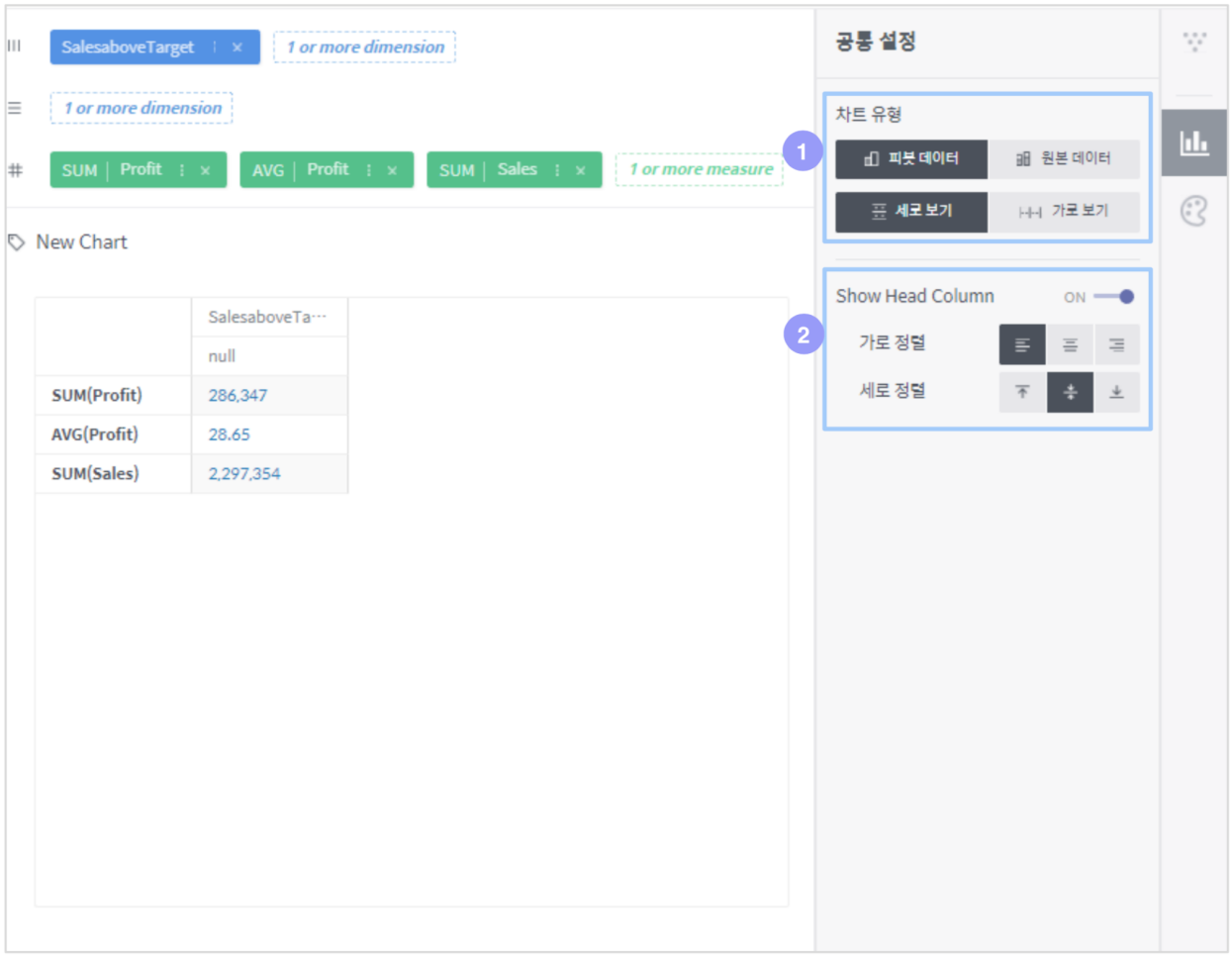
1. **Chart type**
   - **Vertical:** Displays data values as vertical bars with the dimension axis set vertical.
   - **Horizontal:** Displays data values as horizontal bars with the dimension axis set horizontal.
   - **Parallel:** If more than one measure are selected, different bars representing those measures are displayed in parallel.
   - **Stacked:** If more than one measure are selected, different bars representing those measures are stacked at one position.


Region
1 or more dimension
SUM | DaystoShipActual
SUM | Profit
1 or more measure
New Chart
범례
미니맵
공통 설정
차트 유형
세로형
가로형
병렬형
중첩형
1
2
Limitation
10
123
120,000
100,000
80,000
60,000
40,000
20,000
0
SUM(DaystoShipActual)—SUM(Profit)
Central
East
South
West
Region

2. **Limitation:** Set how many columns to display on the chart.

Table

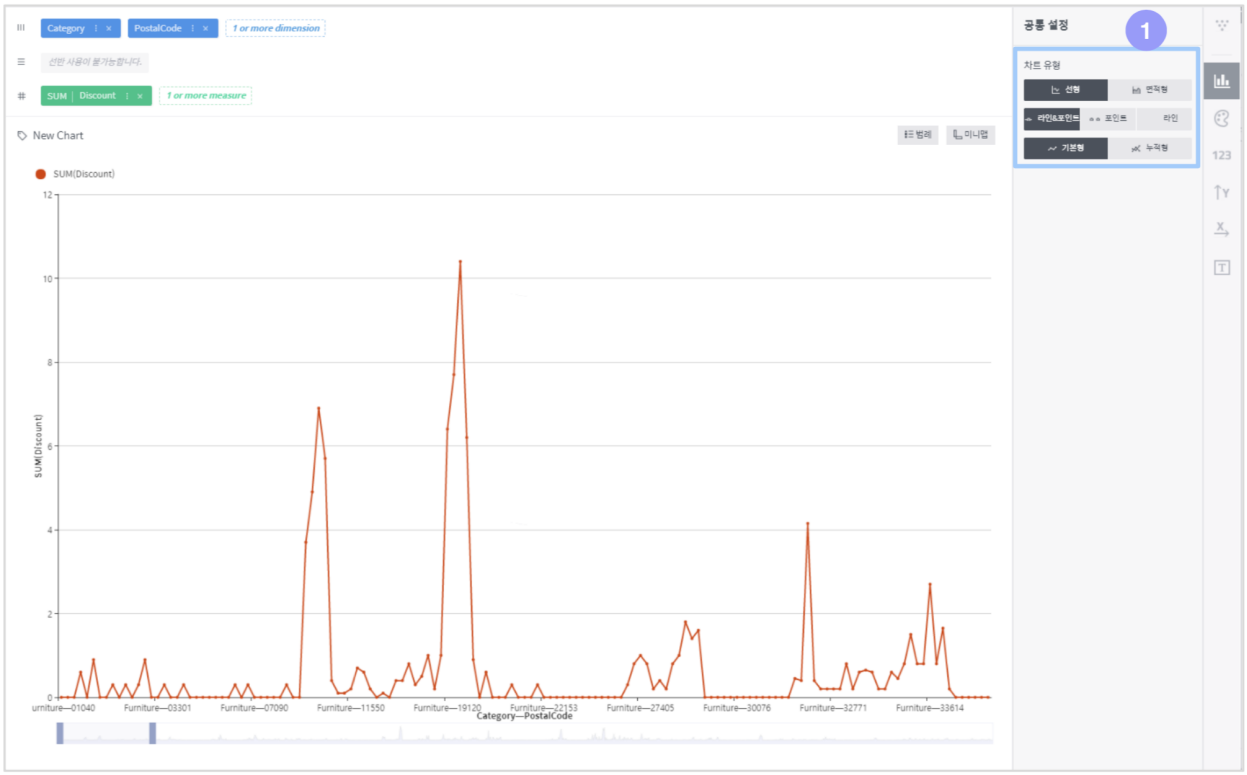
A table block is formed based on the categories into which the dimension columns on the column/row shelves are grouped; accordingly, the values of the measure columns on the cross shelf are displayed as text in the crossings.



1. **Chart type**
   - **Pivot:** Aggregates (`SUM`, `MIN`, `MAX`, etc) measure values for each pair of column and row dimensions into a different cell.
   - **Original:** Displays all original measure values as unaggregated together with the selected dimensions.
   - **Vertical:** Displays measure values vertically in the table. This cannot be used when "Original" is selected for displaying the table.
   - **Horizontal:** Displays the table horizontally when "Pivot" is selected for displaying the table. Displays measure values horizontally in the table.
2. **Show head column:** Set horizontal and vertical text alignment in the column headers. When "Original" is selected, the column headers are necessarily shown. When "Pivot" is selected, you may optionally hide the column headers.

Line chart

This type of chart presents data values in each category of a dimension column with points. Adjacent data points are connected with each other. This type of chart is used to view trends.



1. **Chart type**
    - **Line type:** Displays the chart graph by drawing lines between points that represent measure value aggregates.
    - **Area type:** Colors the area formed by the connecting lines.
    - **Line & point:** Shows both the data points and connecting lines.
    - **Point:** Shows the data points only.
    - **Line:** Shows the connecting lines only.
    - **Basic:** Displays each aggregate as it is on the chart.
    - **Cumulative:** Displays cumulative aggregates on the chart.

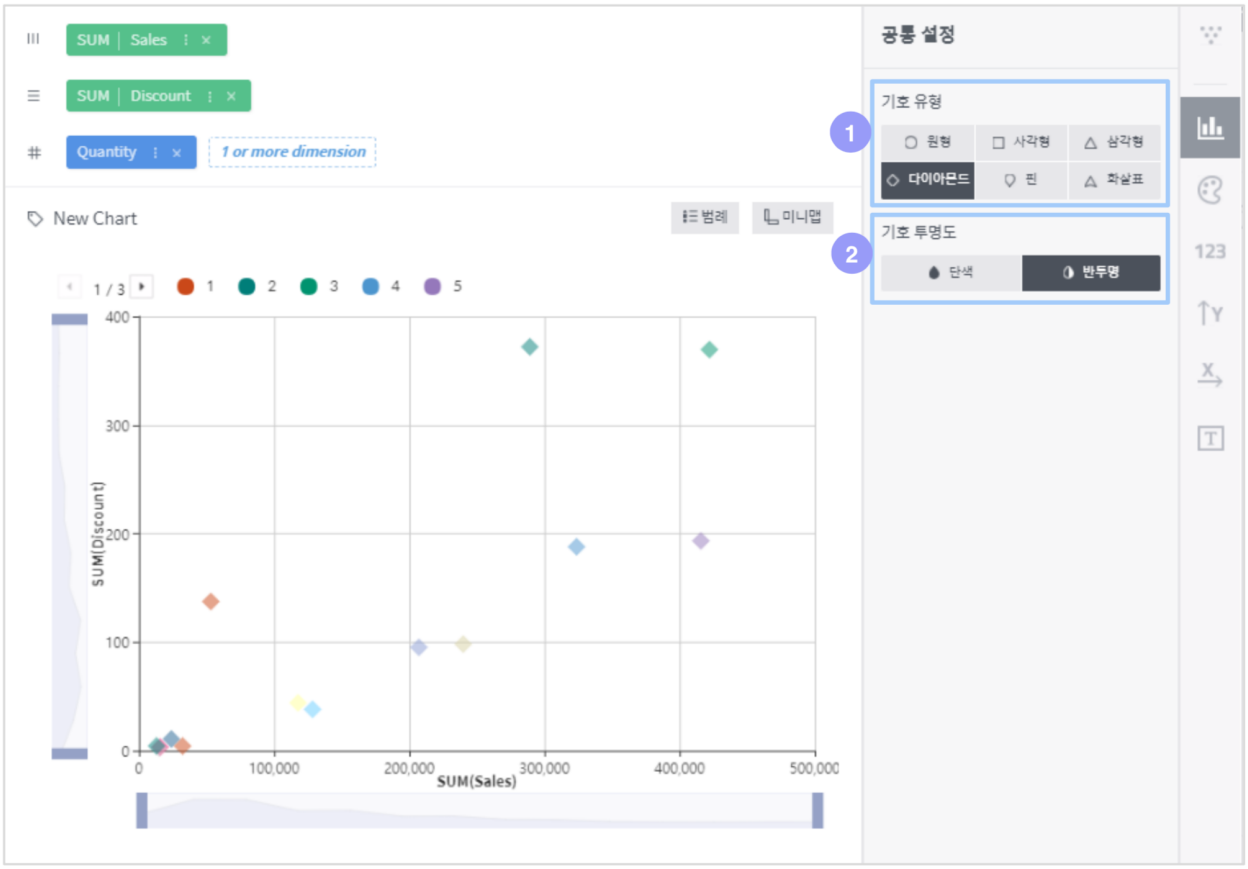
Scatter chart

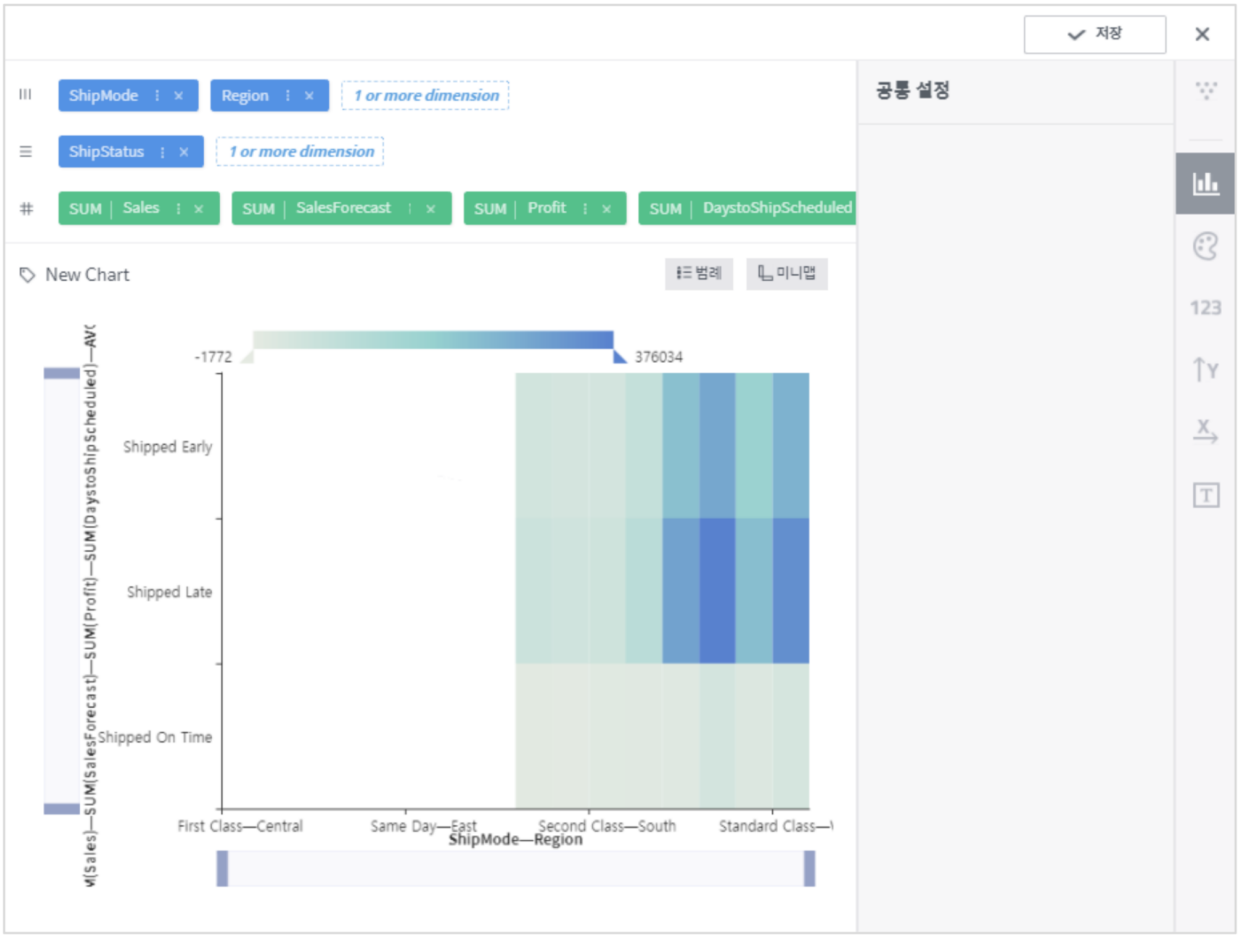
This type of chart presents data values in each category of a dimension column with defined symbols.

1. **Symbol type:** Set the shape of the symbol to be shown on the chart.
2. **Symbol transparency:** Set the transparency of the symbol to be shown on the chart. You can set colors either solid or transparent.

Heatmap

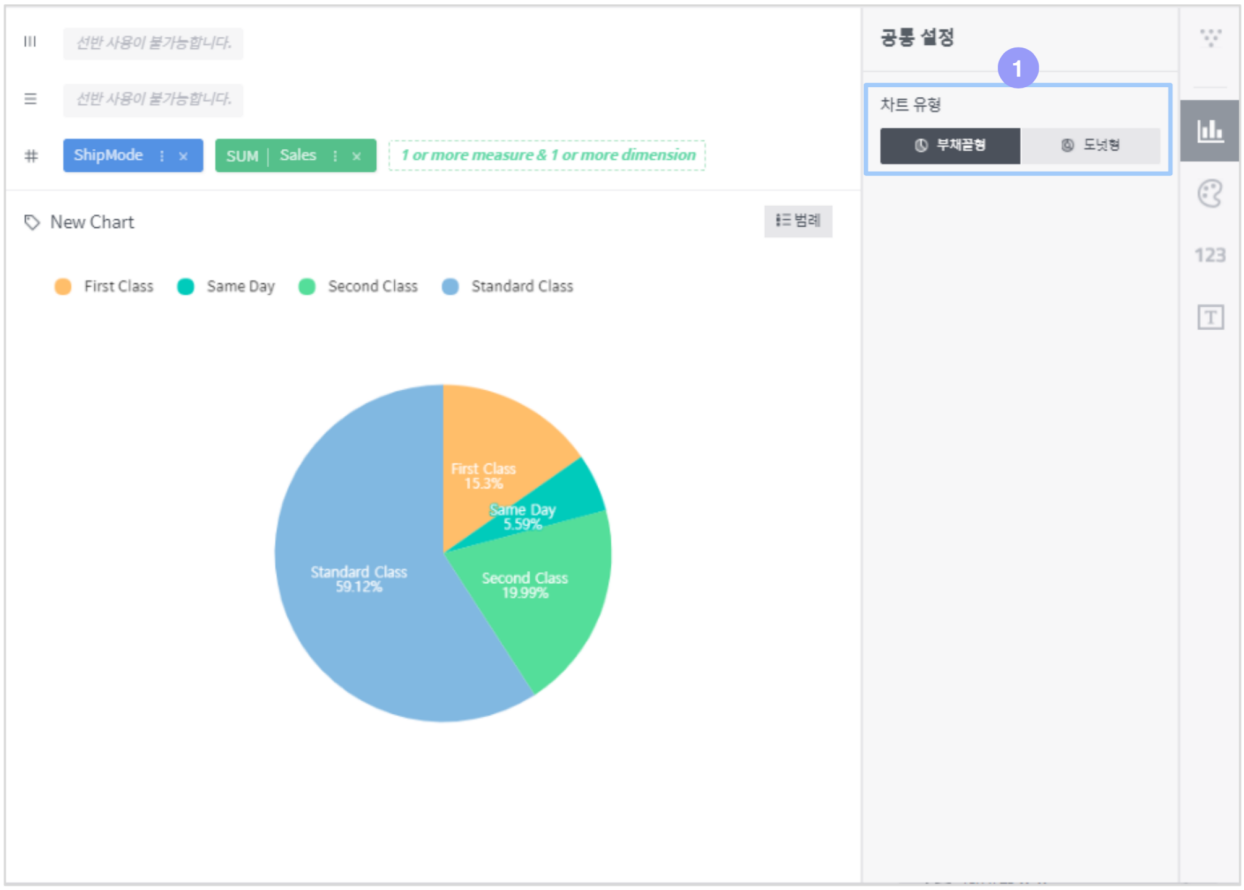
This type of chart displays values aggregated from the measure column placed on the cross shelf by using colors. For a larger aggregated value, a darker color is applied. The heatmap type does not provide any common settings.


SUM | Sales
SUM | Discount
Quantity
1 or more dimension
New Chart
범례
미니맵
1 / 3
1
2
3
4
5
400
300
200
100
0
SUM(Discount)
0
100,000
200,000
300,000
400,000
500,000
SUM(Sales)
공통 설정
기호 유형
원형
사각형
삼각형
다이아몬드
핀
화살표
1
기호 투명도
단색
반투명
2
123
Y
X
T


저장
ShipMode
Region
1 or more dimension
ShipStatus
1 or more dimension
SUM | Sales
SUM | SalesForecast
SUM | Profit
SUM | DaystoShipScheduled
공통 설정
New Chart
범례
미니맵
-1772
376034
Shipped Early
Shipped Late
Shipped On Time
First Class—Central
Same Day—East
Second Class—South
ShipMode—Region
123

Pie chart

This type of chart visualizes the proportion of each category of the dimension column.



1. **Chart type**
   - **Sector:** Displays a pie-shaped chart.
   - **Donut:** Displays a donut-shaped chart.

5.4 Filter

Filters are to display only data matching their preset conditions when forming dashboards and charts. Charts use two types of filters: chart filters and global filters. Chart filters are applied to individual charts, whereas global filters are applied to an entire dashboard.

5.4.1 Chart filters

A chart filter defines what range of data is to be shown on the chart. This chapter describes how to set up and make use of chart filters.

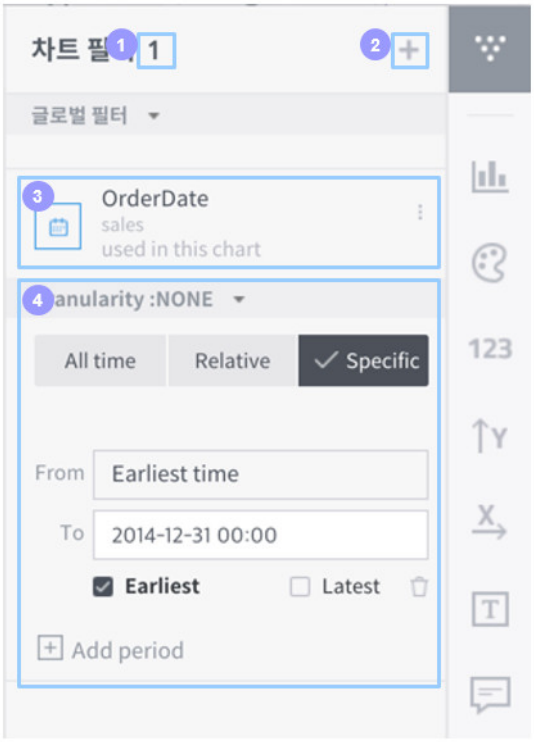
Automatically included filters

The following column filters are included automatically when a chart is created:

- **Timestamp column filter:** As a time-series data store, the Metatron engine necessarily uses a time filter.
- **Recommended filters:** Column filters designated as "recommended filters" during the registration of the data source.
- **Dashboard filters set global:** Filters applied to all charts registered in the dashboard.

Chart filter panel

The chart filter panel is located on the right-hand side of the chart home screen. On this panel, you can easily view and configure registered filters.

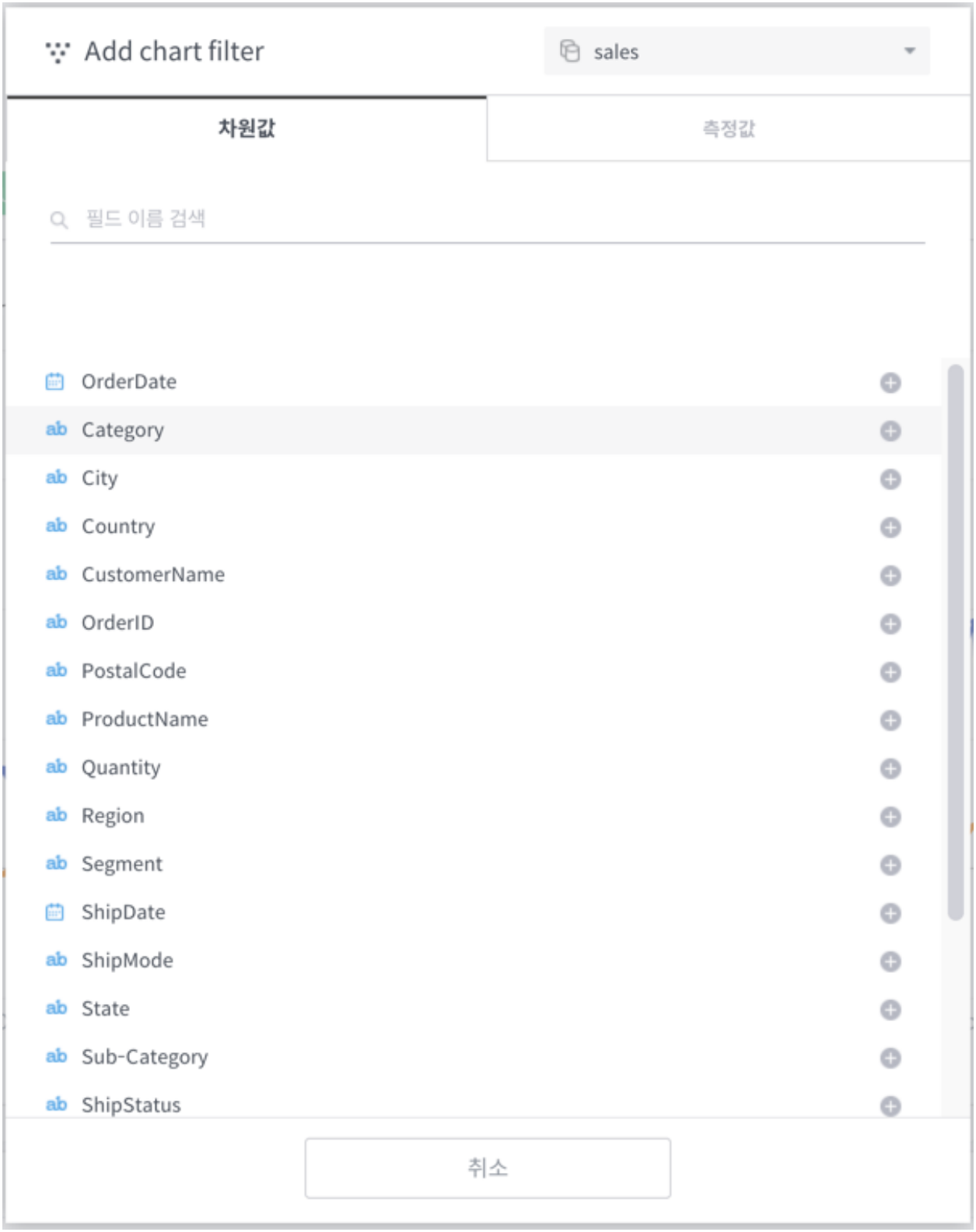


1. **Filter number:** Displays how many filters are registered for the chart.
2. **Add/edit filter:** Click on "+" at the top right to either add a new filter or open a popup for configuring an existing filter.
3. **Columns applied with the filter:** The top part of each individual filter displays which columns are applied with the filter.
4. **Filter settings:** Click the hamburger menu at the top right of an individual filter either to reset the filter or configure the details of the filter.

Chart filter dialog box

Click the button at the top of the chart filter panel or click the button in each filter area to open the chart filter dialog box. With this dialog box, you can add a new filter or configure an existing filter.

The chart filter dialog box is divided into the Dimension and Measure tabs as shown below:


Add chart filter
sales
차원값
측정값
필드 이름 검색
OrderDate
Category
City
Country
CustomerName
OrderID
PostalCode
ProductName
Quantity
Region
Segment
ShipDate
ShipMode
State
Sub-Category
ShipStatus
취소

Dimension filtering

From the connected data source, select a dimension on which to create a filter.

- **Value range:** Select whether to filter the chart by a single or multiple data categories.
  - **Single:** Select one data category by which to filter the chart.
  - **Multiple:** Select multiple data categories by which to filter the chart.
- **Search:** If there are too many elements in the column, this function allows you to limit the results only to those you wish to see.
  - **Search by name:** Search the column element list by name.
  - **Element filtering:** Filters elements either by matching element names with regular expressions or wildcards, or by applying a range condition to a measure.
- **Defined value:** Used to add?as a filter criterion?a data element that is not contained in the column. This allows you to create a filter in advance for a data element that may be added later.

Timestamp column filter settings

Dimensions with a time icon displayed are of a timestamp type for which a timestamp filter can be configured. Although they are set to "All time" by default, you can select Relative or Specific if you wish to display only data from a certain period in the chart.

"Relative" sets a period of time relative to the present and displays only data from the applicable period of time in the chart.

"Specific" directly sets a certain period of time of data and displays only data from the applicable period of time in the chart.

Measure filtering

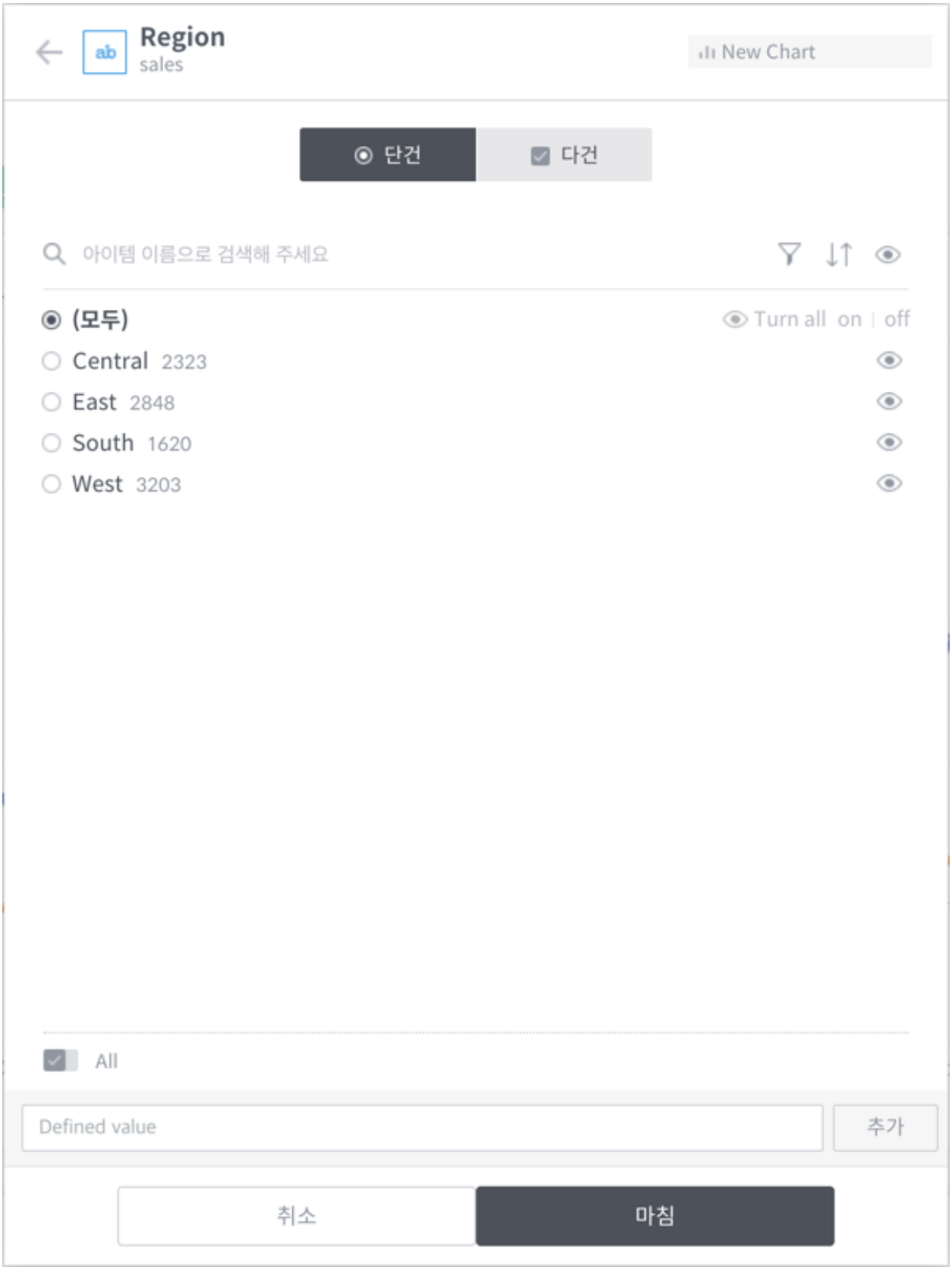
From the connected data source, select a measure on which to create a filter.

Once you have selected a measure, designate the range of values to filter.

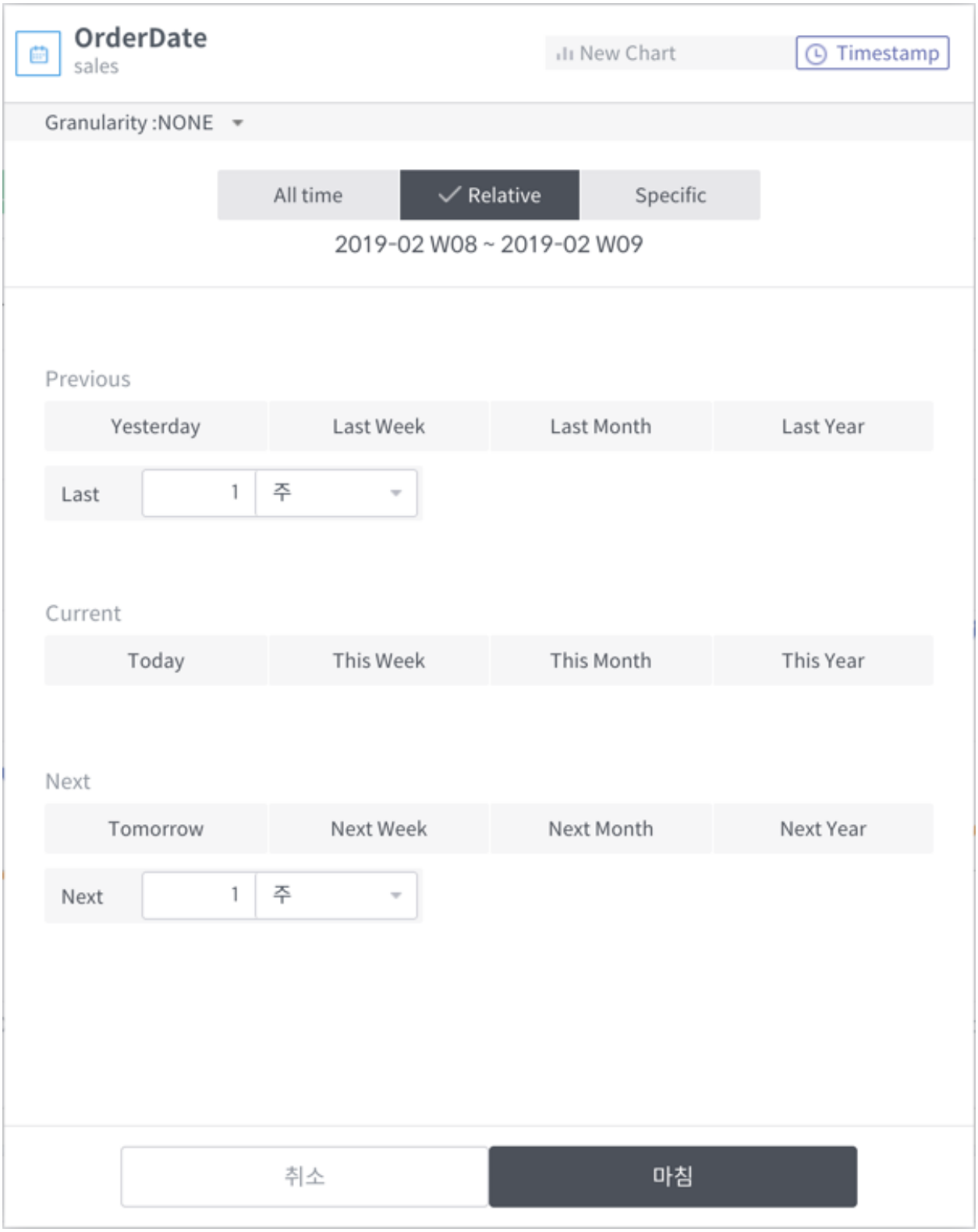
5.4.2 Global filters

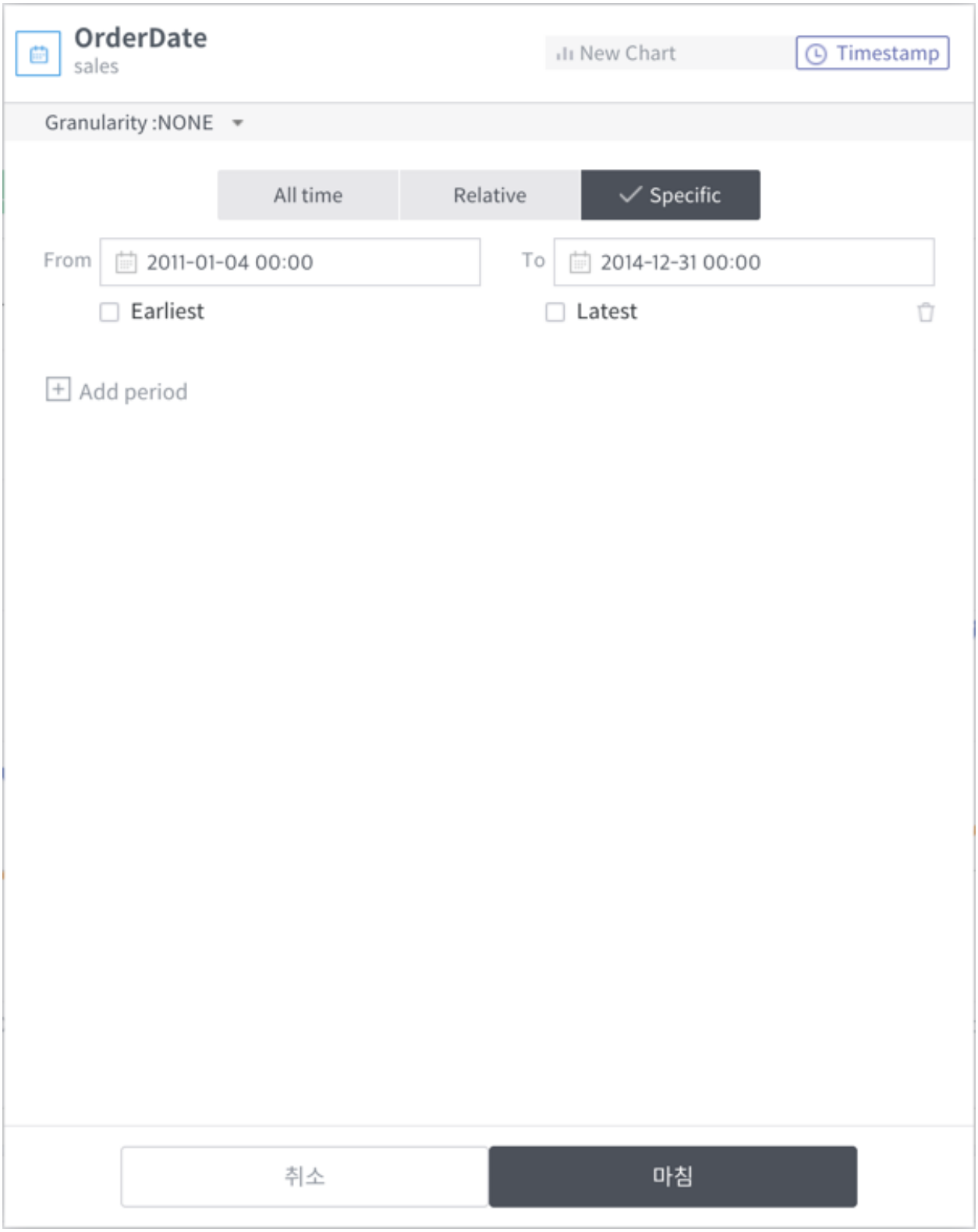
Global filters specify which data is to be displayed in all charts of a dashboard. They can be added, edited, or deleted in the filter panel in the dashboard editing window.

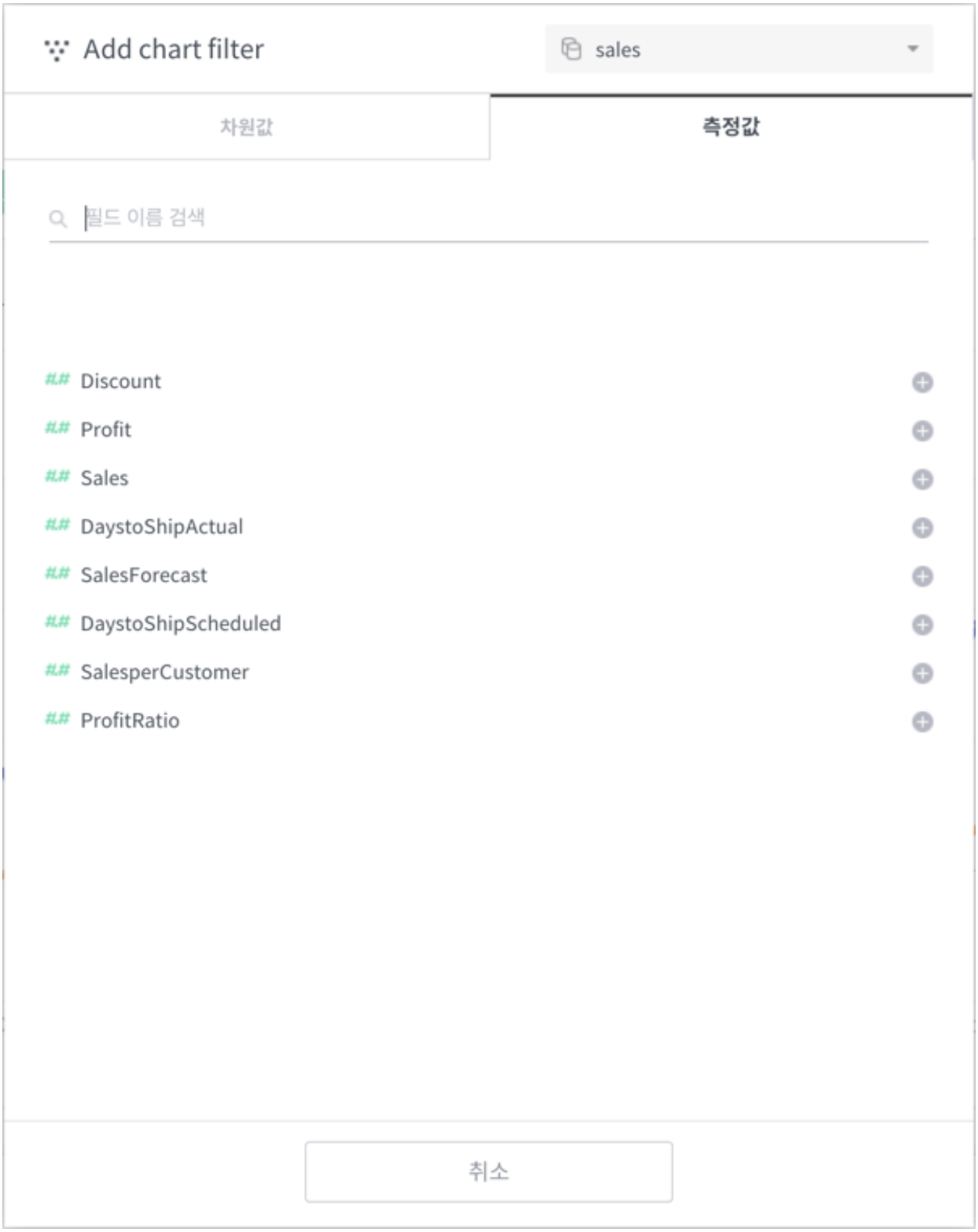
1. **Number of filter widgets:** Displays how many filter widgets are currently registered in the dashboard next to the global filter heading.
2. **Add a filter widget:** Click the "+" icon at the top right to create a new filter widget in the dashboard. The filter creation popup interface and process for creating filters are the same as the process for creating chart filters described in the previous section.


Region
sales
New Chart
단건
다건
아이템 이름으로 검색해 주세요
(모두)
Turn all on | off
Central 2323
East 2848
South 1620
West 3203
All
Defined value
추가
취소
마침

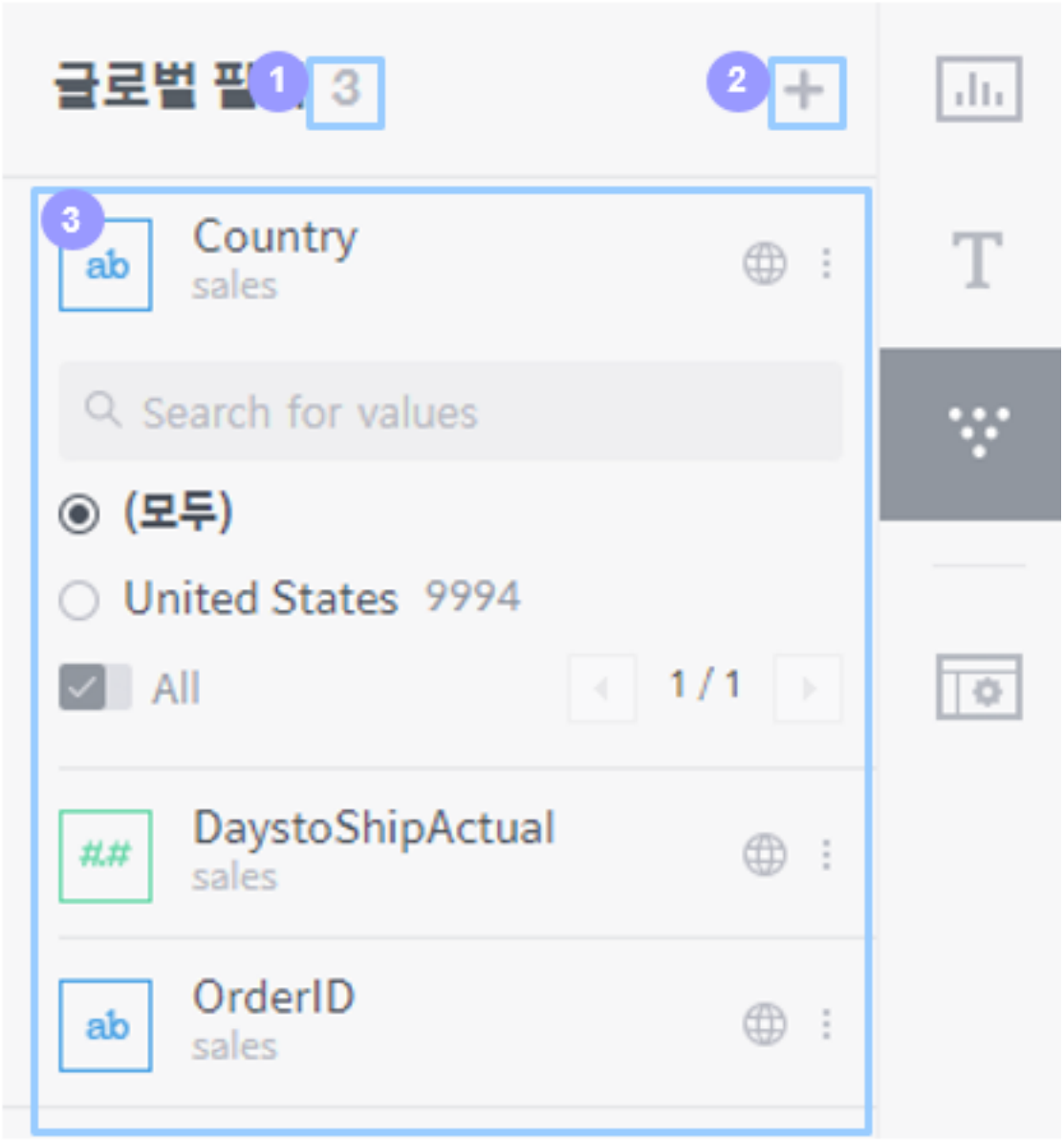

Matcher
와일드카드
정규식
시작하는 값 'c'
c
시작 단어
Condition
Profit 합계 of values is above or equal to 10
Profit
합계
>=
10
Limitation
상위
10
측정값 선택
합계
초기화
적용


OrderDate
sales
New Chart
Timestamp
Granularity :NONE
All time
Relative
Specific
2019-02 W08 ~ 2019-02 W09
Previous
Yesterday
Last Week
Last Month
Last Year
Last
1
주
Current
Today
This Week
This Month
This Year
Next
Tomorrow
Next Week
Next Month
Next Year
Next
1
주
취소
마침


OrderDate
sales
New Chart
Timestamp
Granularity :NONE
All time
Relative
Specific
From
2011-01-04 00:00
To
2014-12-31 00:00
Earliest
Latest
Add period
취소
마침


Add chart filter
sales
차원값
측정값
필드 이름 검색
Discount
Profit
Sales
DaystoShipActual
SalesForecast
DaystoShipScheduled
SalesperCustomer
ProfitRatio
취소


Profit
sales
New Chart
Value range
Minimum
-6600
Maximum
2984
Available input range -6600 ~8400
취소
마침


글로벌 필
1
3
2
3
ab
Country
sales
Search for values
(모두)
United States 9994
All
1 / 1
#.#
DaystoShipActual
sales
ab
OrderID
sales

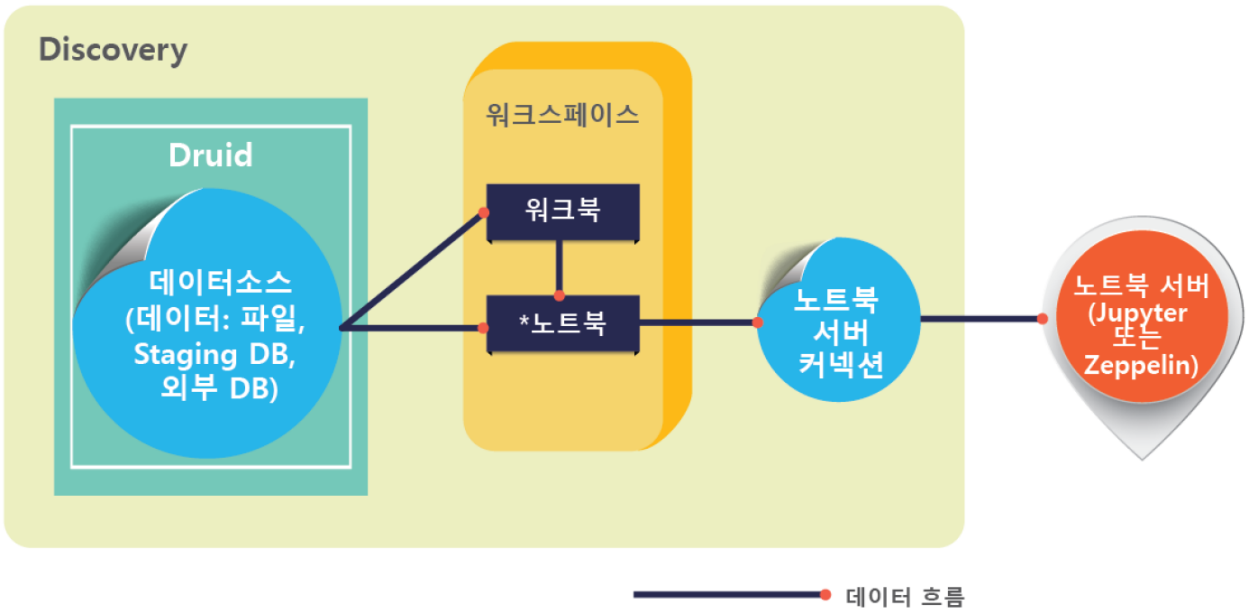
3. **Filter widget list:** Lists filter widgets registered in the dashboard. Hover the mouse over a widget to display the edit and delete icons. Drag a widget to the widget display area to display the widget in the display area.

Global filters applied to the entire dashboard are also listed when creating an individual filter for a new chart. When creating a global filter, if there are any individual chart filters, it intuitively notifies you of which column the filter was created from.

CHAPTER

SIX

NOTEBOOK



Metatron Discovery supports a notebook function. Notebook is a tool for creating and sharing documents that include live codes, equations, visualizations, and descriptive texts. It is mostly used for data cleaning and manipulation, numerical simulations, statistical modeling, and machine learning.

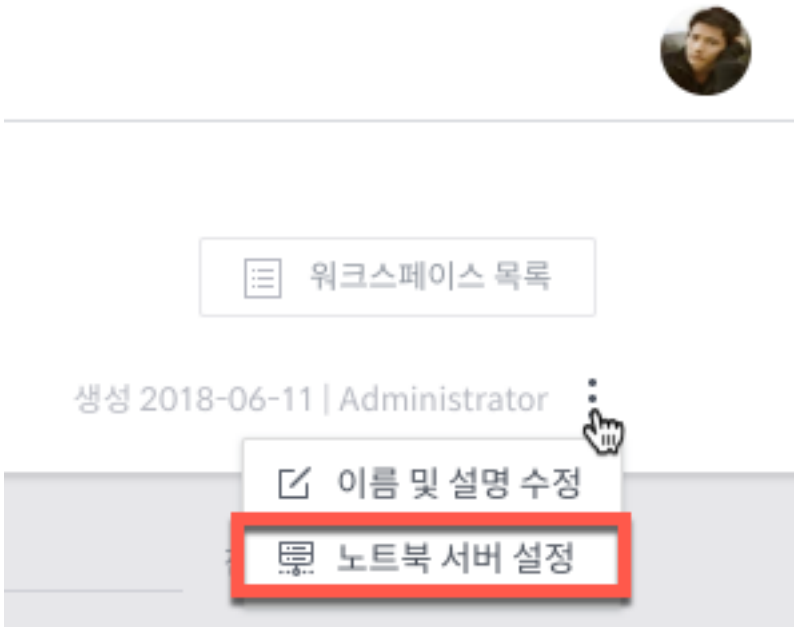
Metatron Discovery allows users to register and use external Jupyter and Zeppelin servers. Jupyter uses Python and R?programming languages commonly used in data science?while Zeppelin uses Spark (Scala) to help with real-time and interactive analysis and visualization of data. Before running the notebook, its server must be set up.

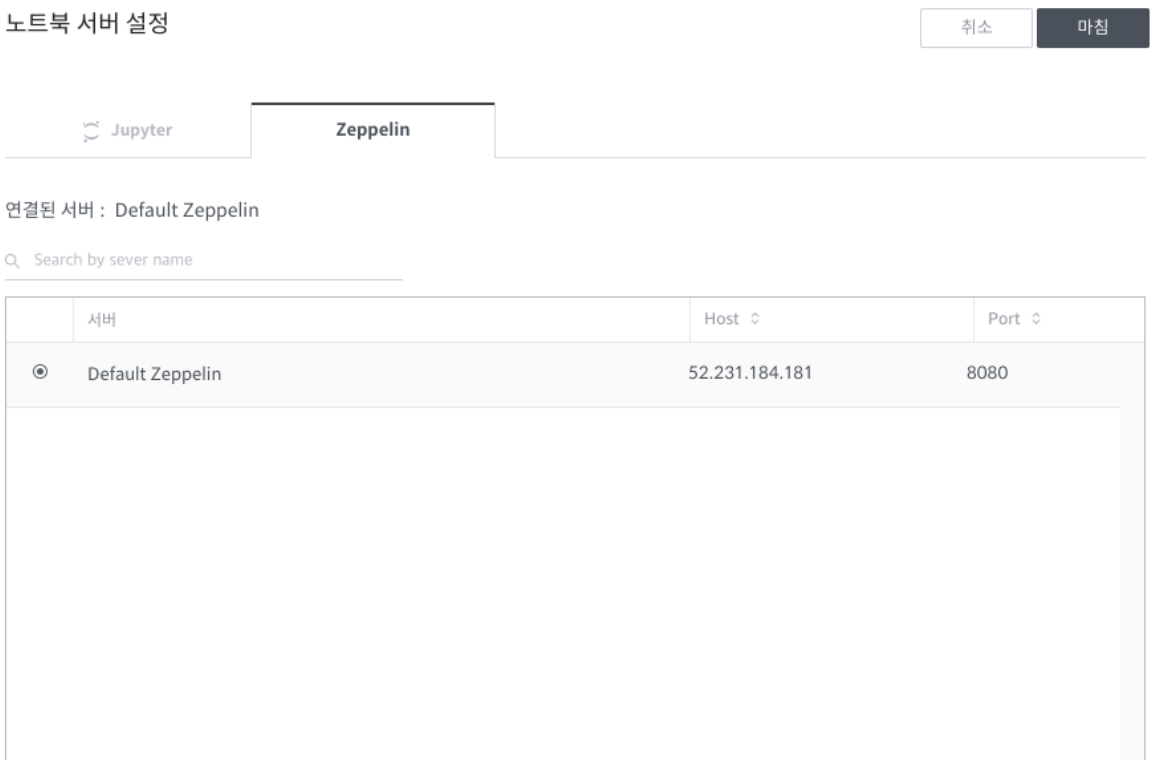
6.1 Register a notebook server

To analyze data in a workspace using a notebook, initial settings are required for the notebook server. The procedure for initial settings for a notebook server is as follows:

1. Click the ⋮ button in the top-right corner of the workspace and select **Set notebook server**.

2. From the list of Jupyter and Zeppelin servers preregistered by the administrator, select the notebook server that you wish to connect to and use in your workspace and click **Done**.


워크스페이스 목록
생성 2018-06-11 | Administrator
이름 및 설명 수정
노트북 서버 설정

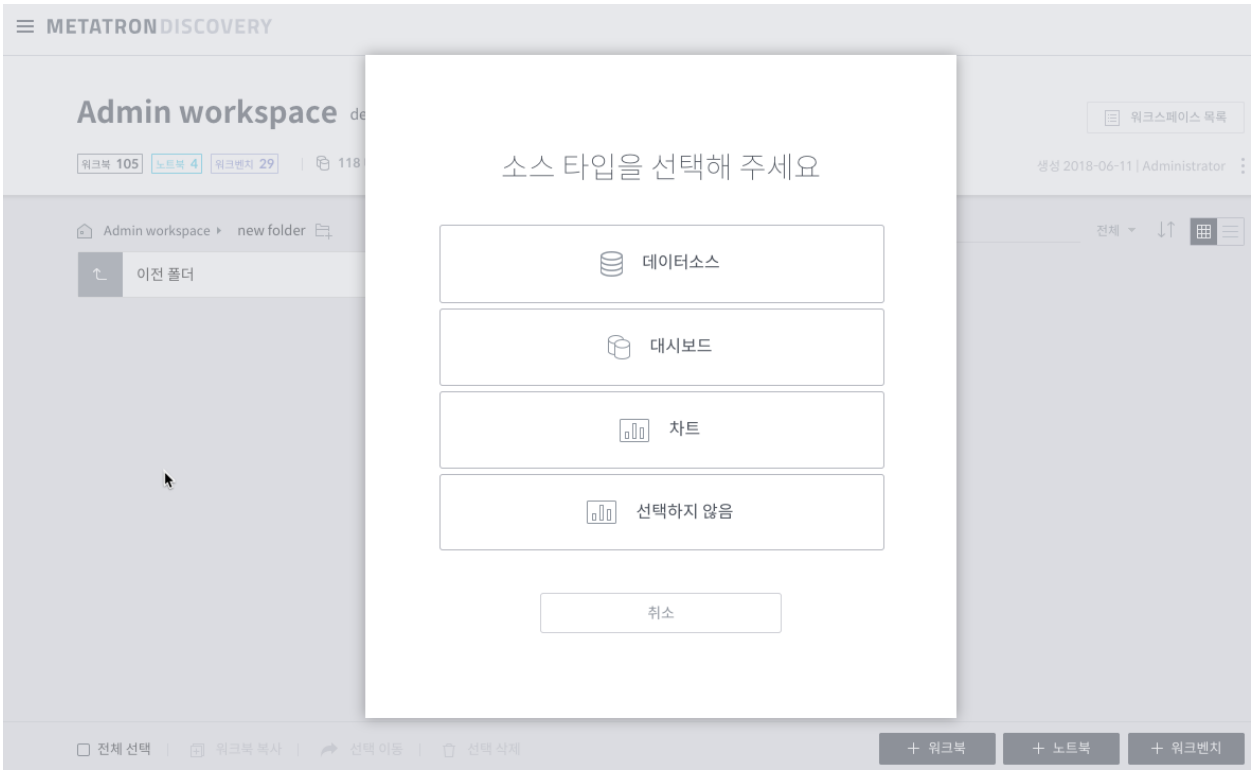

노트북 서버 설정
취소
마침
Jupyter
Zeppelin
연결된 서버 : Default Zeppelin
Search by sever name
서버
Host
Port
Default Zeppelin
52.231.184.181
8080

6.2 Create a notebook

Once the notebook server has been set up, you can create a notebook. A notebook is created as follows:

1. Click the **+ Notebook** button at the bottom of the workspace. You'll be prompted to create a notebook.

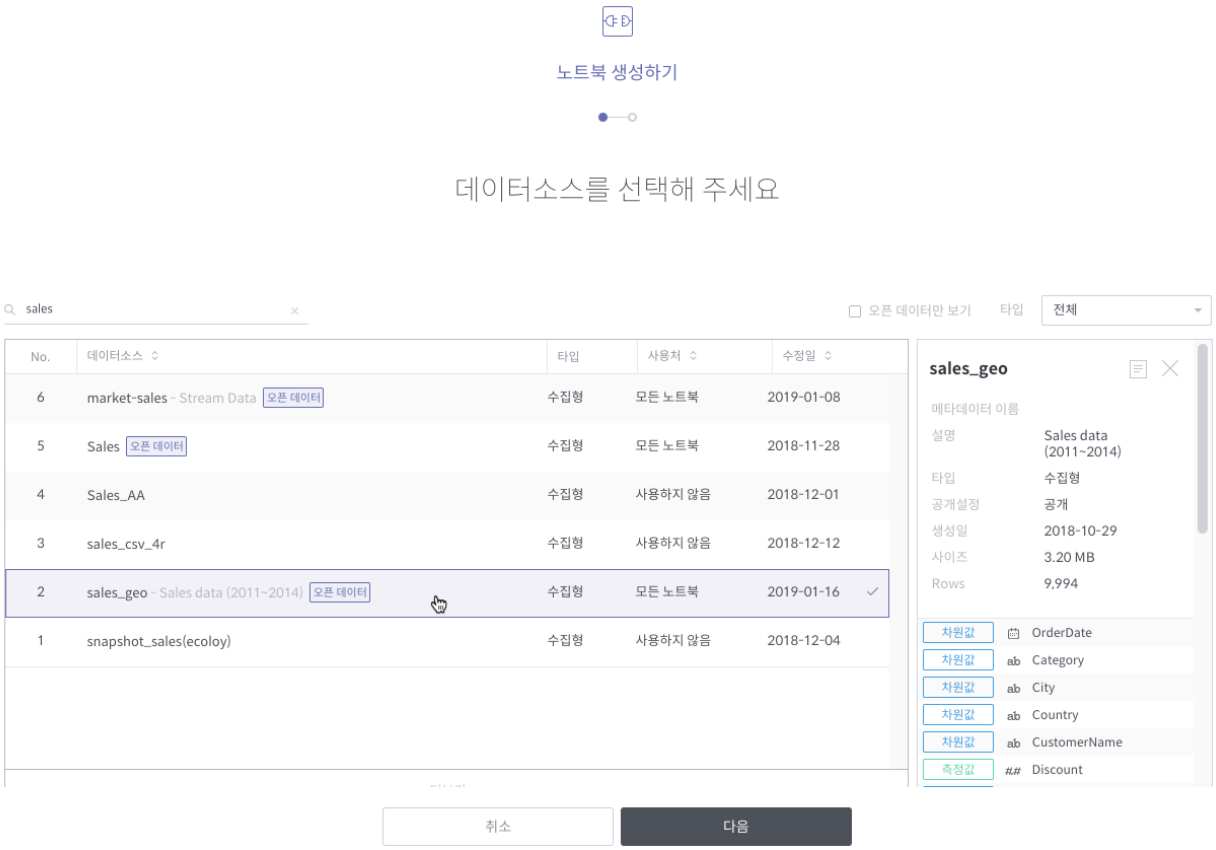
2. Select the type of data set that you wish to analyze in the notebook. You can choose between **Data source**, the unit of data used in Metatron Discovery, **Dashboard**, **Chart**, and **Not selected**. If you want to use Zeppelin, select **Not selected**.



3. After selecting either **Data Source**, **Dashboard**, or **Chart**, you can see a list of data currently registered in Metatron Discovery. Select the data to analyze and click Next.

4. Enter the information about the notebook that you want to use as an analytics tool for data. The **server type** can only be selected for a notebook server connected at the initial notebook server setup. If **Jupyter** is selected, "R" or "Python" can be selected for analysis, whereas "Spark" (Scala) is used when **Zeppelin** is selected.

5. Once a notebook has been created, you can find it in the workspace.



6.3 Use a notebook

In a newly created notebook, you can write a script and serve it through a REST API. A notebook can be used as follows:

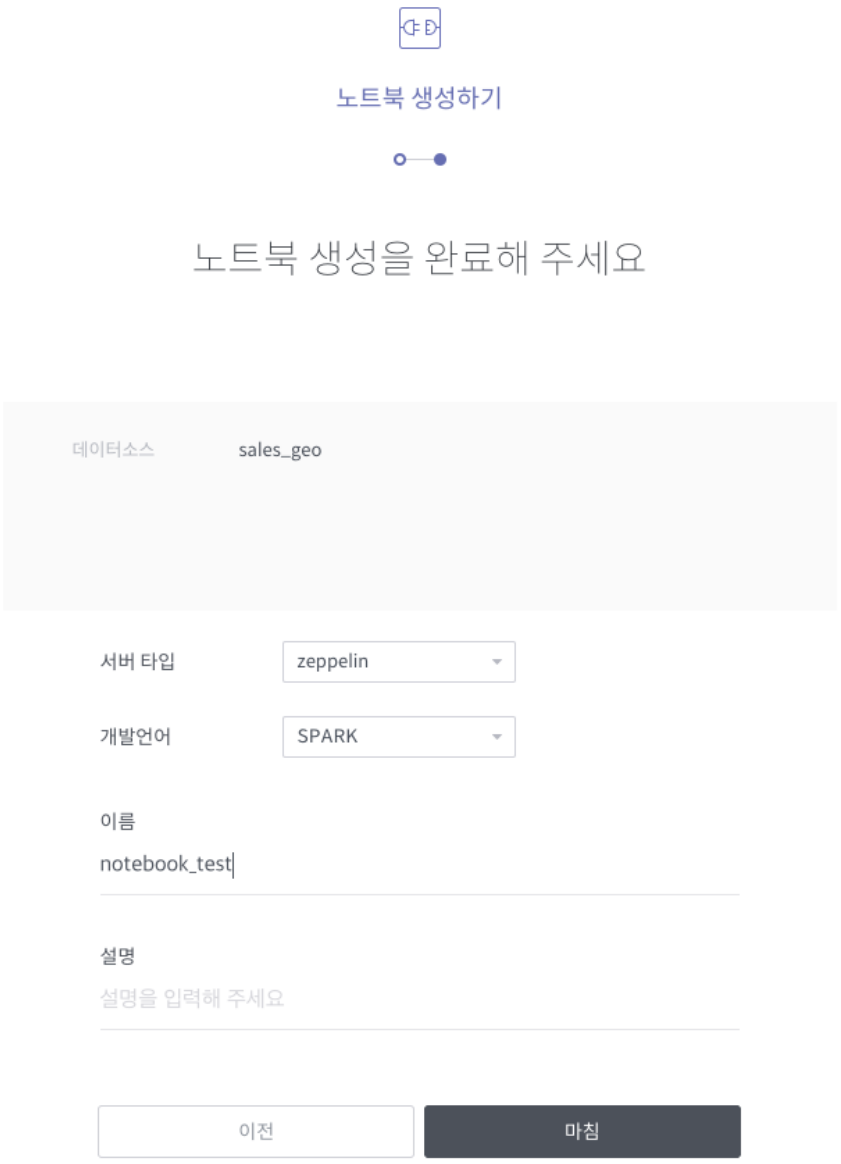
6.3.1 Detailed notebook page

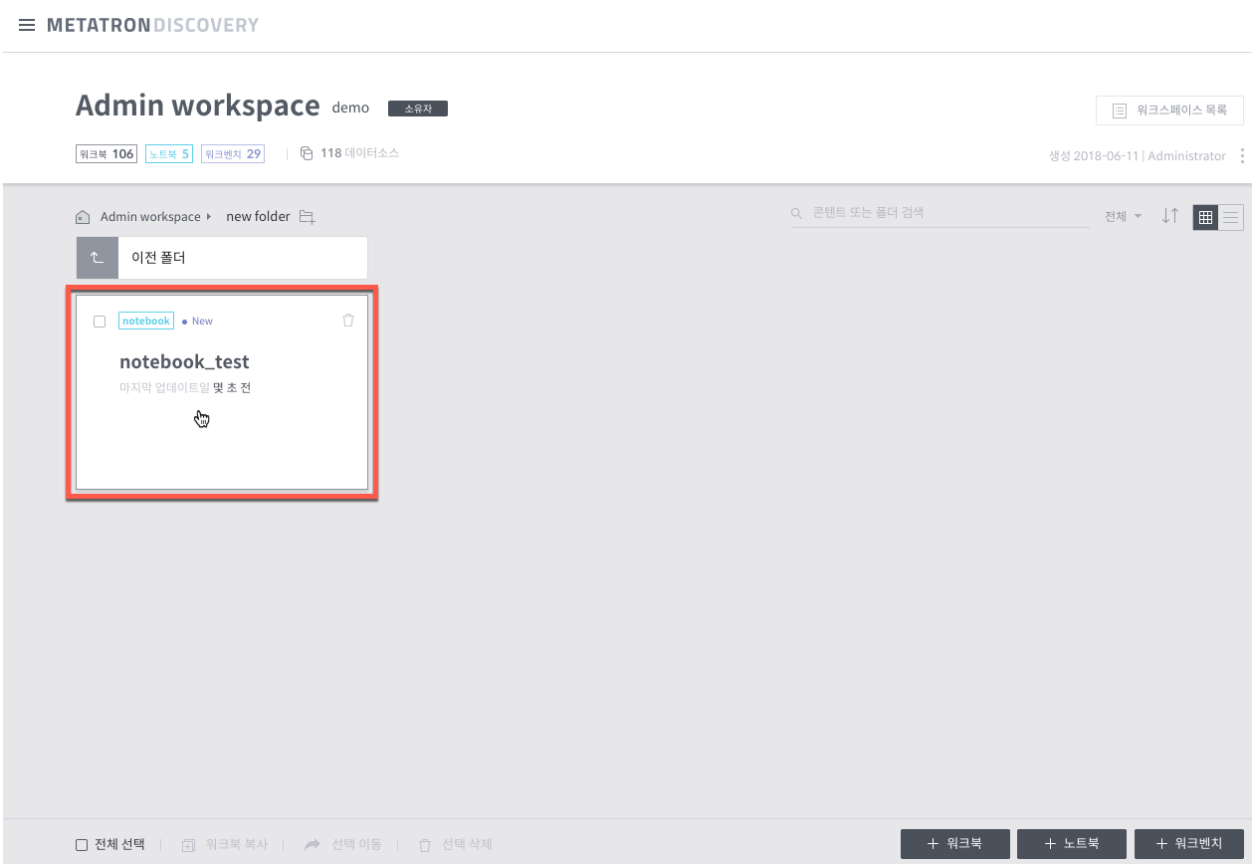
On the workspace screen, select the notebook you want to use as an analytics tool. Then, the following screen with detailed information appears. You'll see basic information on the notebook: data type, data source name, development language, and analytic code, etc.

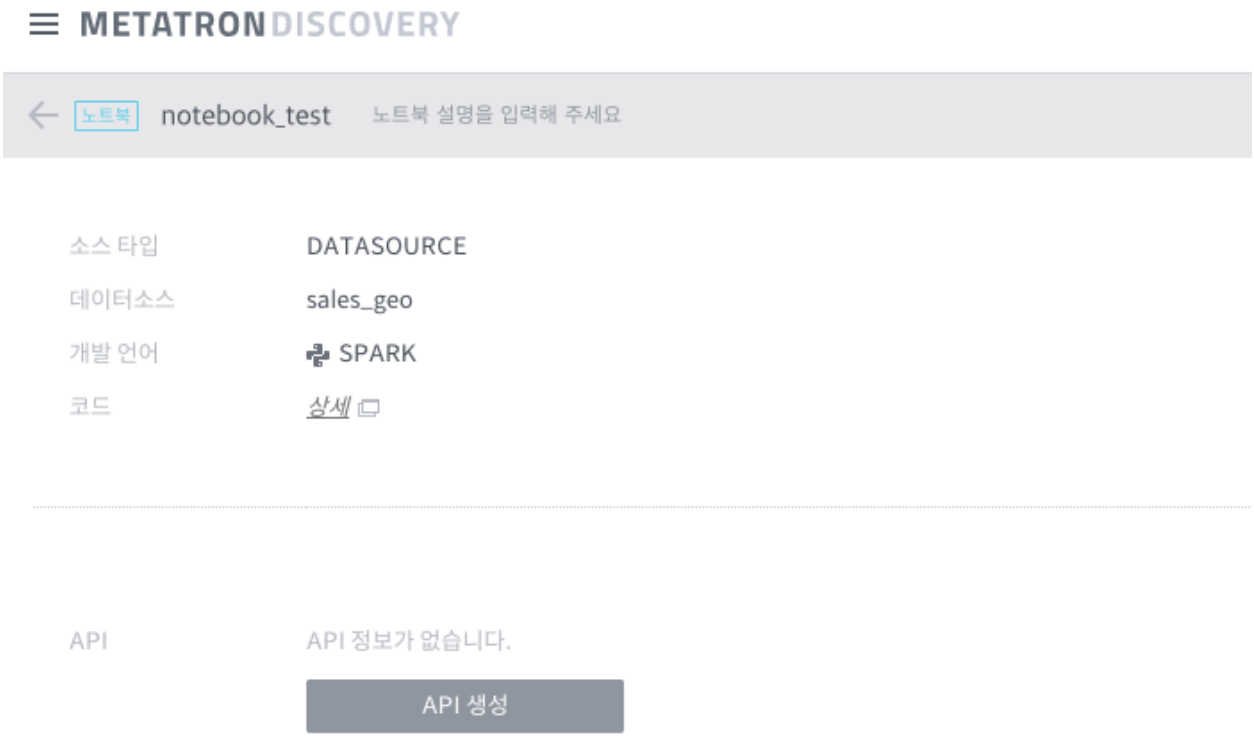
6.3.2 Notebook coding

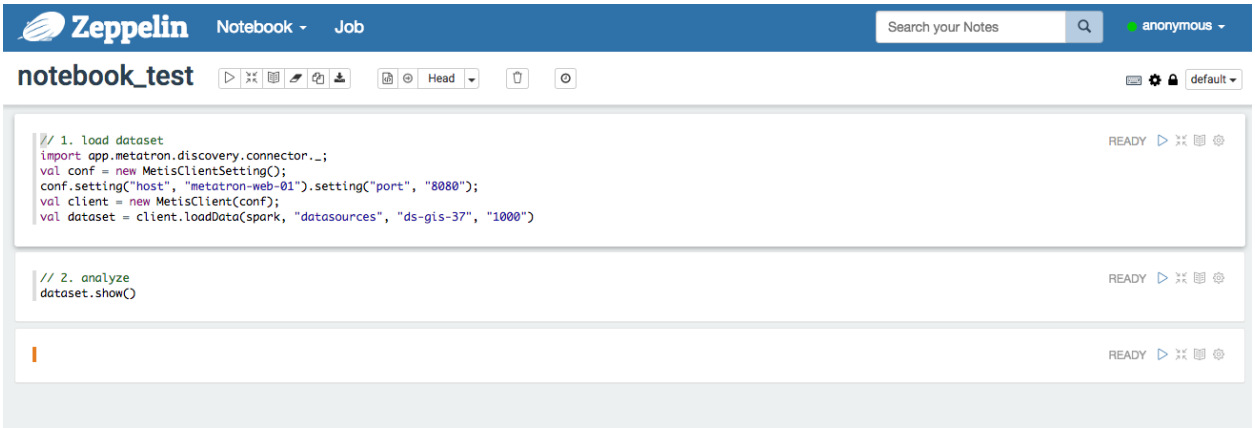
Click **Detail** on the notebook page to pop up a new window for coding in the notebook. At the top of this window, a code to load a dataset is inserted; executing this cell loads a JSON dataset as the dataset object.

The screen above appears when Zeppelin is selected and includes a cell for loading the data selected when the notebook was created. After coding the program starting from the third cell, click **Save** when you are finished.


노트북 생성하기
노트북 생성을 완료해 주세요
데이터소스
sales_geo
서버 타입
zeppelin
개발언어
SPARK
이름
notebook_test
설명
설명을 입력해 주세요
이전
마침

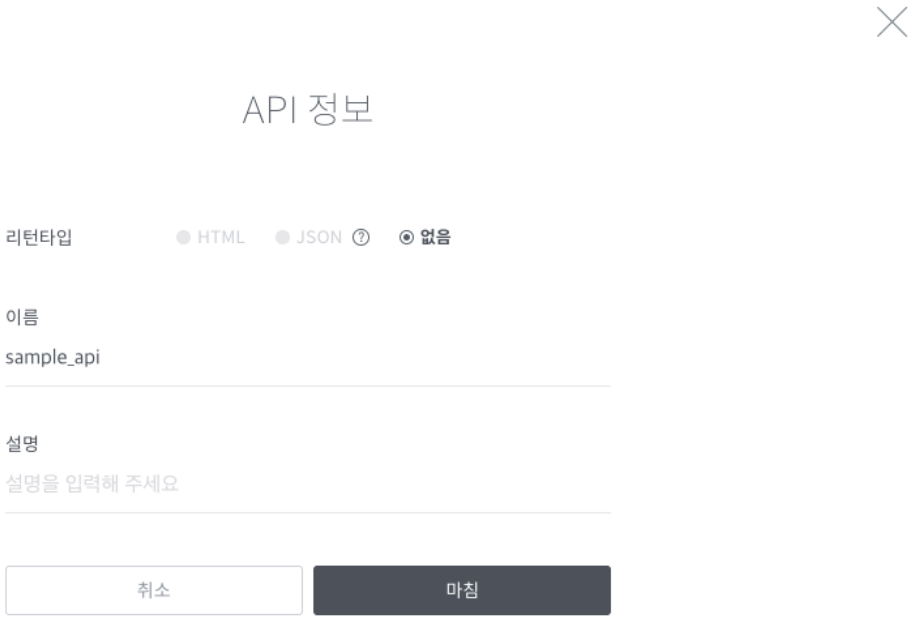

METATRON DISCOVERY
Admin workspace
demo
소유자
워크스페이스 목록
워크북 106
노트북 5
워크벤치 29
118 데이터소스
생성 2018-06-11 | Administrator
Admin workspace
new folder
콘텐트 또는 폴더 검색
전체
이전 폴더
notebook
New
notebook_test
마지막 업데이트일 몇 초 전
전체 선택
워크북 복사
선택 이동
선택 삭제
+ 워크북
+ 노트북
+ 워크벤치


METATRON DISCOVERY
노트북
notebook_test
노트북 설명을 입력해 주세요
소스 타입
DATASOURCE
데이터소스
sales_geo
개발 언어
SPARK
코드
상세
API
API 정보가 없습니다.
API 생성



6.3.3 Register a notebook API

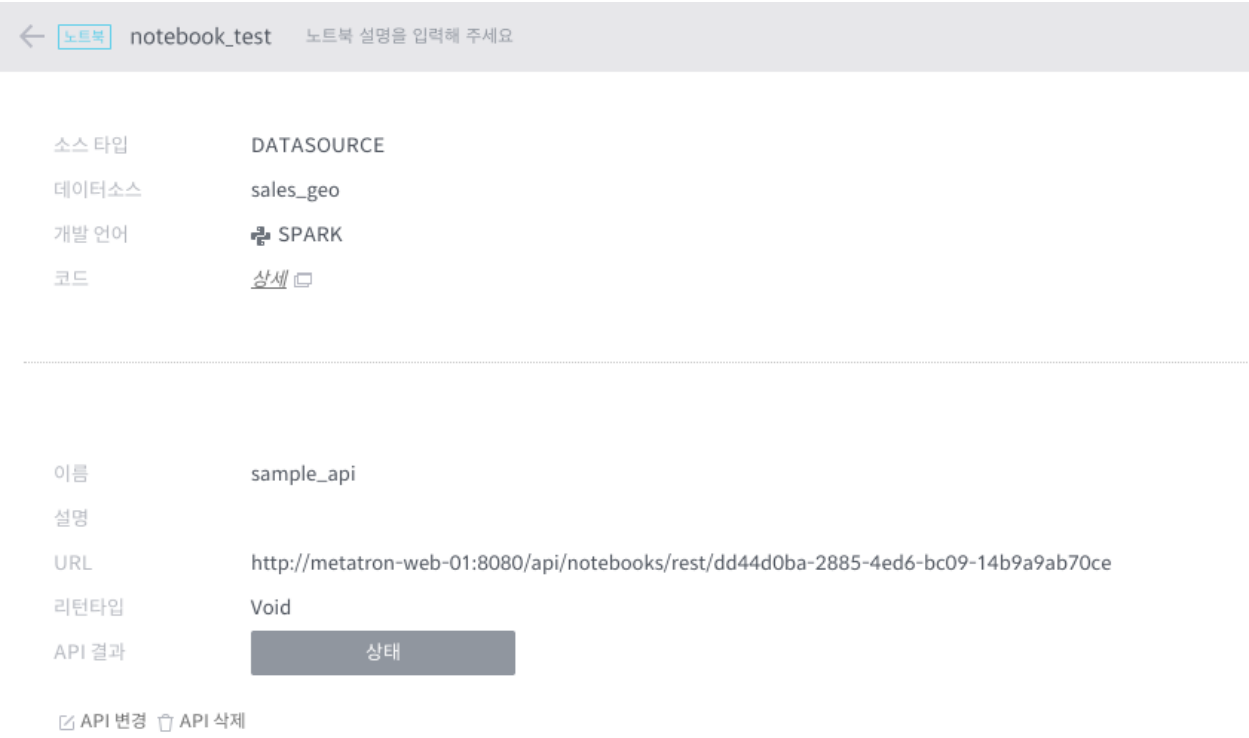
Once you write a notebook code, you can return the results by calling a REST API. Select a **Return type** by referring to the descriptions below, and enter a **Name** and **Description**.



- **HTML**: The results of running the notebook script are returned in HTML.
- **JSON**: The results of running the notebook script are returned in a custom JSON format. In this case, the `response.write(...)` function provided by Metatron Discovery will be used. The following is an example code for using the response.write function:
  - R-based notebook: `response.write(list(coefficient = 2, intercept = 0))`
  - Python-based notebook: `response.write({'coefficient' : 2.5, 'intercept' : 0})`

- **None**: Runs the notebook script but does not provide returns.

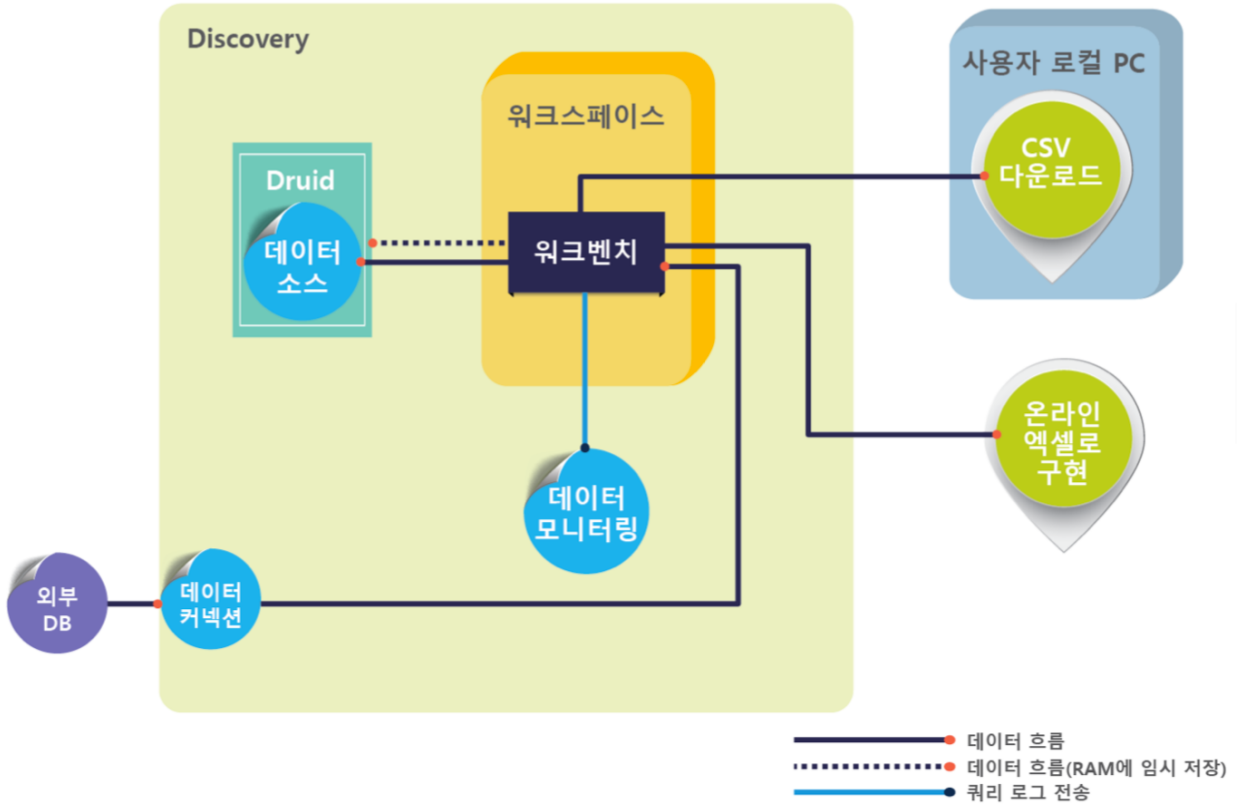
Once you enter API information and click **Done**, the API is created to provide a REST API URL as shown below. Click **Result** to view the URL execution results in a popup window.



CHAPTER

SEVEN

WORKBENCH



Metatron Workbench provides an environment for data preparation and analytics based on SQL. Its main functions are as follows:

- Various external databases can be loaded in one space.
- The user can conveniently navigate/select linked tables and columns and view their details.
- Query edit tools are embedded and query results can be viewed interactively and available for various uses:
  - Query results can be downloaded into a local file or exported to an online Excel.
  - Query results can be interactively visualized to help the analyst see an outline of the resulting data table.
  - Query results can be stored as a data source available for analytics in a workbook or notebook.

Each document that stores SQL-based analytic queries is called a "workbench." This chapter introduce how to **create** and **use** workbenches.

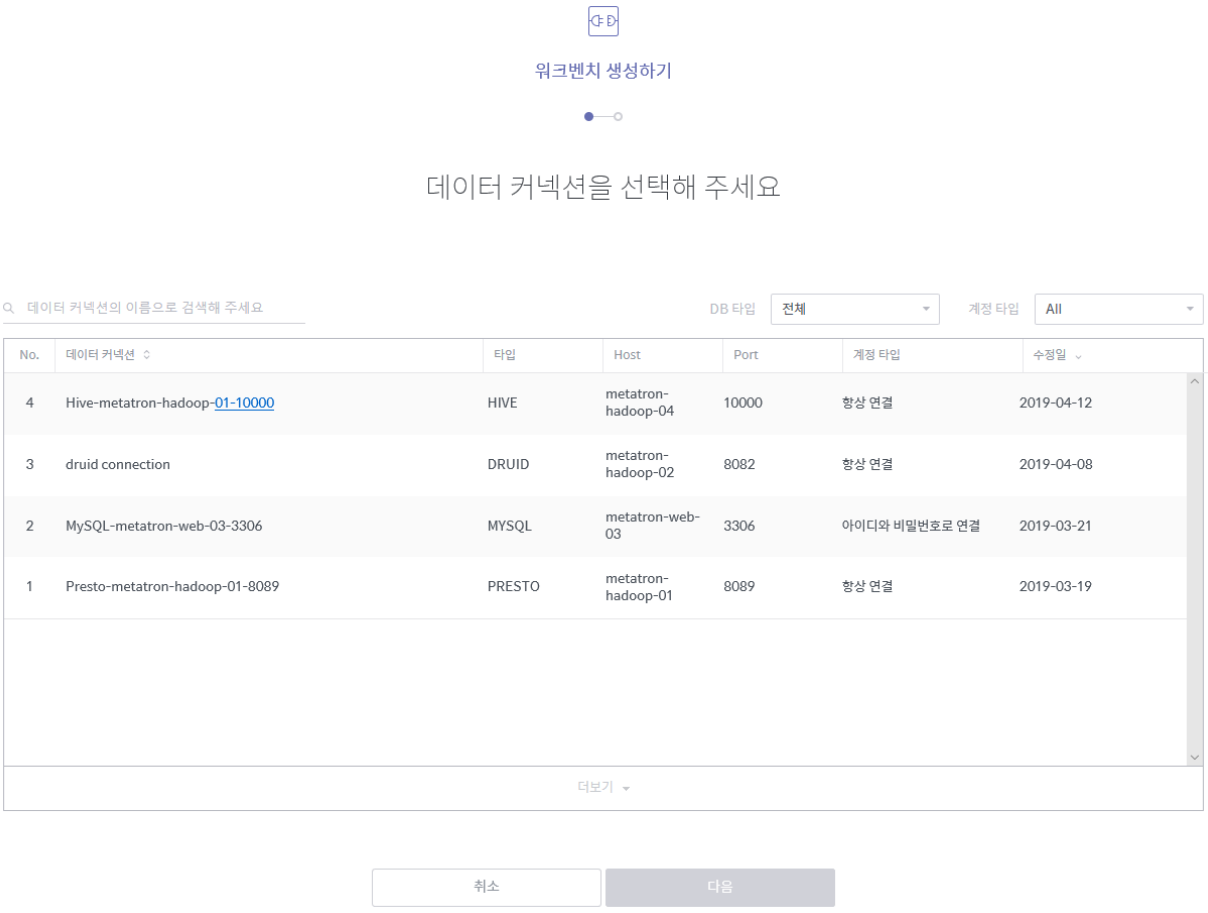
7.1 Create a workbench

To use a workbench in the workspace, a workbench-type data connection must be established. See *Data Connection* for how to handle it.

To create a workbench:

1. Click the **+ Workbench** button at the bottom of the workspace. You'll be prompted to select a data connection for data analytics.

2. Select the workbench-type data connection that connects to the data table you want, and click **Next**.



- **Search by name of data connection:** Searches the list of data connections available to the workspace by the name you type in.
- **DB type:** Filters data connections by database type (Oracle/MySQL/Hive/Presto/Tibero). Select **All** to display data connections regardless of database type.
- **Account type:** Filters data connections by account type (All/Always connect/Connect by user's account/Connect with ID and password). Select **All** to display data connections regardless of account type.

- **Data connection:** Lists data connections filtered by specified criteria.

3. Confirm the information of the selected data connection and enter a name and a description to create a workbench.

4. The created workbench is immediately available.

7.2 Use a workbench

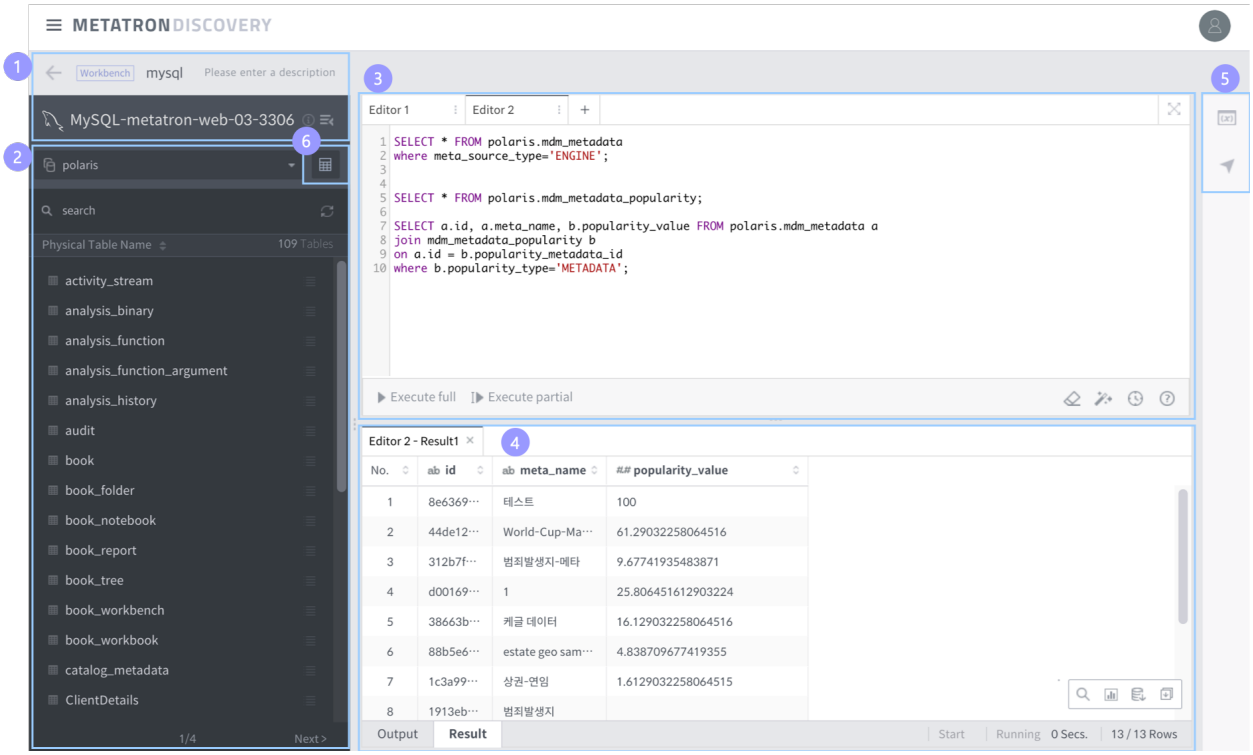
In the workbench, you can edit and manage an SQL database easily, as well as visualize and store the results of a query on it in various forms. The workbench page consists of five sections shown below, and an additional schema browser is provided.

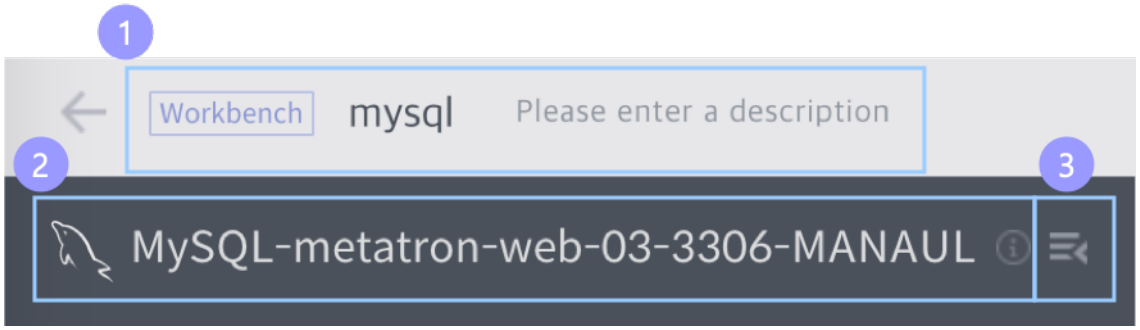
1. Basic information section (See *Basic information section*)
2. Schema and table section (See *Schema and table section*)
3. Query editor section (See *Query editor section*)
4. Query results section (See *Query results section*)
5. Extra tools section (*Extra tools section*)
6. Schema browser (*Schema browser*)

7.2.1 Basic information section

This section displays basic information on the active workbench.

1. **Name:** Name of the workbench. Click on it to change the workbench's name.
2. **Data connection:** Name of the data connection used by the workbench. Click the icon to view its details.

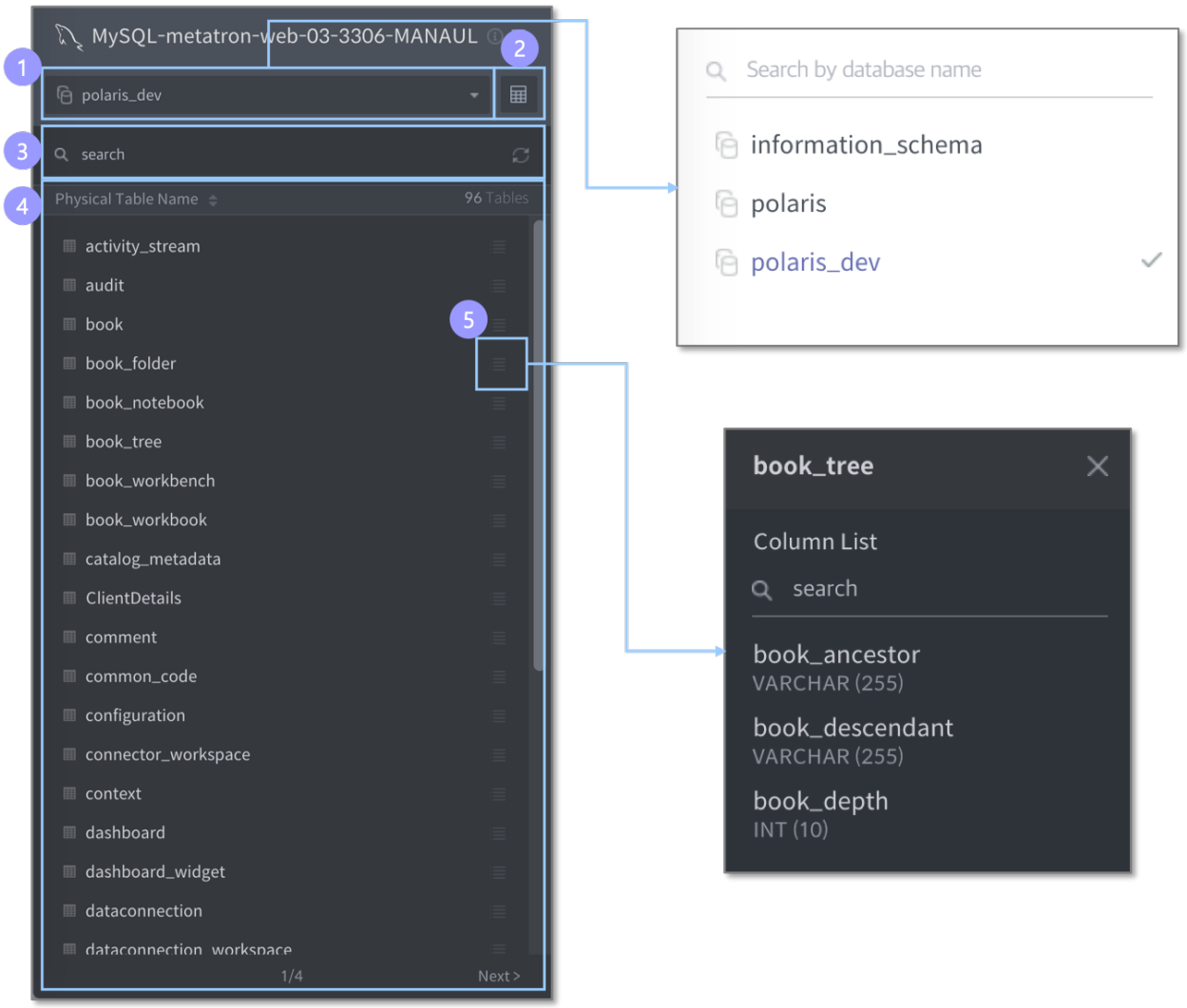

METATRON DISCOVERY
Workbench
mysql
Please enter a description
MySQL-metatron-web-03-3306
polaris
search
Physical Table Name
109 Tables
activity_stream
analysis_binary
analysis_function
analysis_function_argument
analysis_history
audit
book
book_folder
book_notebook
book_report
book_tree
book_workbench
book_workbook
catalog_metadata
ClientDetails
1/4
Next >
Editor 1
Editor 2
SELECT * FROM polaris.mdm_metadata
where meta_source_type='ENGINE';
SELECT * FROM polaris.mdm_metadata_popularity;
SELECT a.id, a.meta_name, b.popularity_value FROM polaris.mdm_metadata a
join mdm_metadata_popularity b
on a.id = b.popularity_metadata_id
where b.popularity_type='METADATA';
Execute full
Execute partial
Editor 2 - Result1
No.
id
meta_name
popularity_value
1 8e6369··· 테스트 100
2 44de12··· World-Cup-Ma··· 61.29032258064516
3 312b7f··· 범죄발생지-메타 9.67741935483871
4 d00169··· 1 25.806451612903224
5 38663b··· 캐글 데이터 16.129032258064516
6 88b5e6··· estate geo sam··· 4.838709677419355
7 1c3a99··· 상권-연임 1.6129032258064515
8 1913eb··· 범죄발생지
Output
Result
Start
Running 0 Secs.
13 / 13 Rows


1
Workbench
mysql
Please enter a description
2
3
MySQL-metatron-web-03-3306-MANAUL

3. : UI button to collapse or expand the panel.

7.2.2 Schema and table section

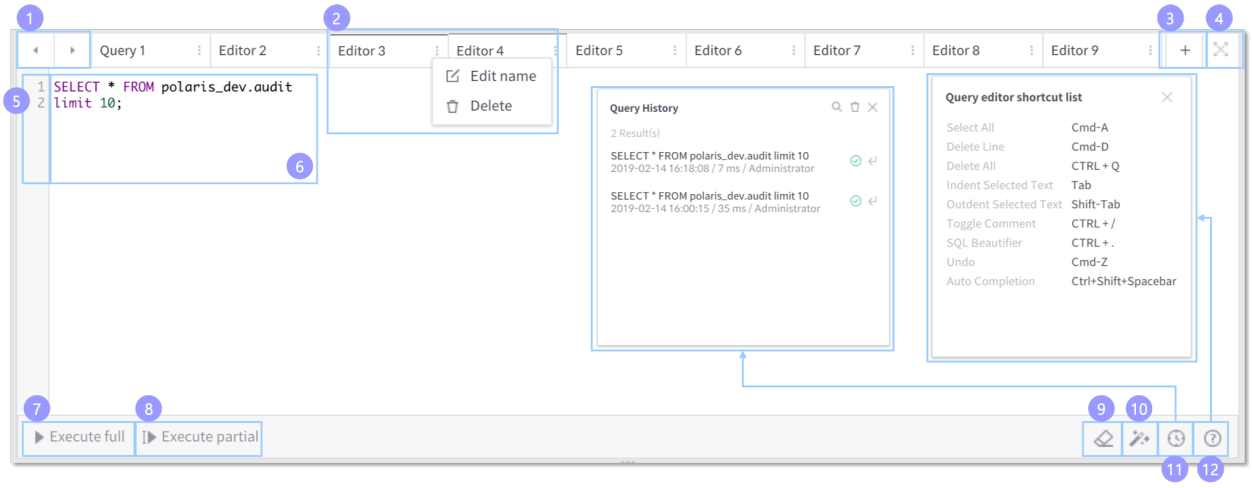
This section provides a UI to conveniently insert the name of a database, table, or column in the query editor.



1. **Database name:** Displays the name of the selected database. By default, the first database of the data connection used by the workbench is selected. Click on it to list all databases included in the data connection. Select a database in the list to replace the currently selected database.
2. **Schema browser:** A popup browser displaying the table list of the selected database, and information of all the columns and records in each table.
3. **Search table:** Searches the list of the tables registered in the selected database by the name you type in.
4. **Table name:** Select a table to automatically insert it in the query editor along with a `SELECT \* FROM {table name}` query.
5. **Column list:** Displays all columns belonging to the table and their respective data types. Click a column name to automatically insert it in the query editor.

7.2.3 Query editor section

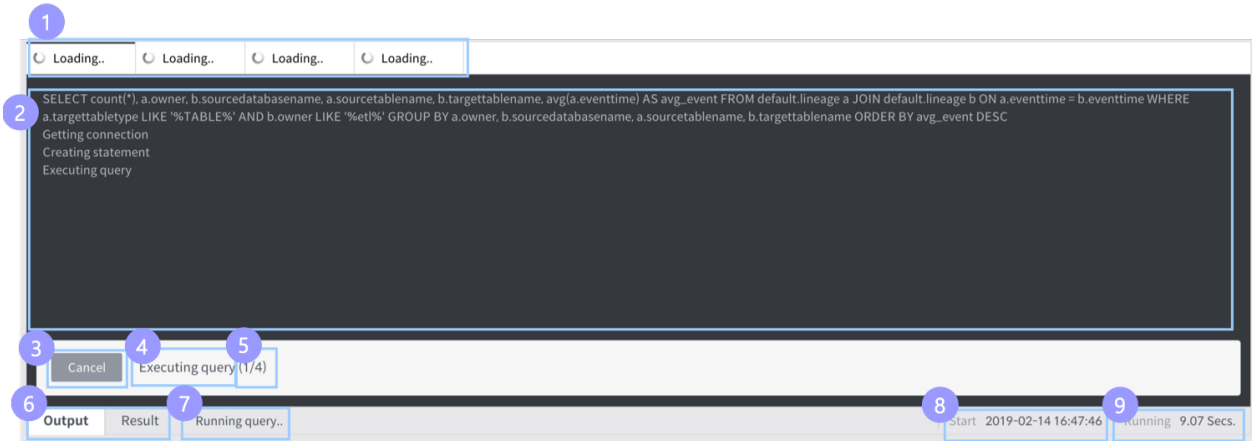
This section allows you to edit and run queries.



1. : Navigates to tabs of previous or subsequent queries when there are too many tabs. If tabs are not many, this button will not appear.
2. **Tab:** You can run or store queries in separate tabs for more efficient management of them. Click the button to edit the tab title or delete the tab.
3. : Click this button to add a new tab.
4. : Click this button to minimize the query editor or maximize it to full screen.
5. **Query lines:** Displays the numbering of the query code lines.
6. **Editor area:** Write query statements in this area. You can run either single or multiple queries. Insert `;` at the end of each query statement to run them separately. Autocomplete is supported.
7. **Execute full:** Execute all queries in the editor. (Shortcut: Ctrl + Enter)
8. **Execute partial:** Executes only the query statement where the cursor is located, or execute queries selected by dragging the mouse. (Shortcut: Command + Enter)
9. **CLEAR SQL:** Clears all query statements.
10. **SQL BEAUTIFIER:** Re-words query statements using standard query syntax.
11. **Query History:** Lists past queries executed in the query editor. If you select a query in the list, it will inserted in the query editor.
12. **Query Editor Shortcuts:** Shows a list of shortcuts available in the query editor.

7.2.4 Query results section

Once a query is executed, its results are displayed in a query results tab. Query results tabs are cumulatively added, and you can selectively delete specific results tabs. Query results are displayed in a text grid, and they can be previewed in charts, stored into data sources, and exported into CSV files.

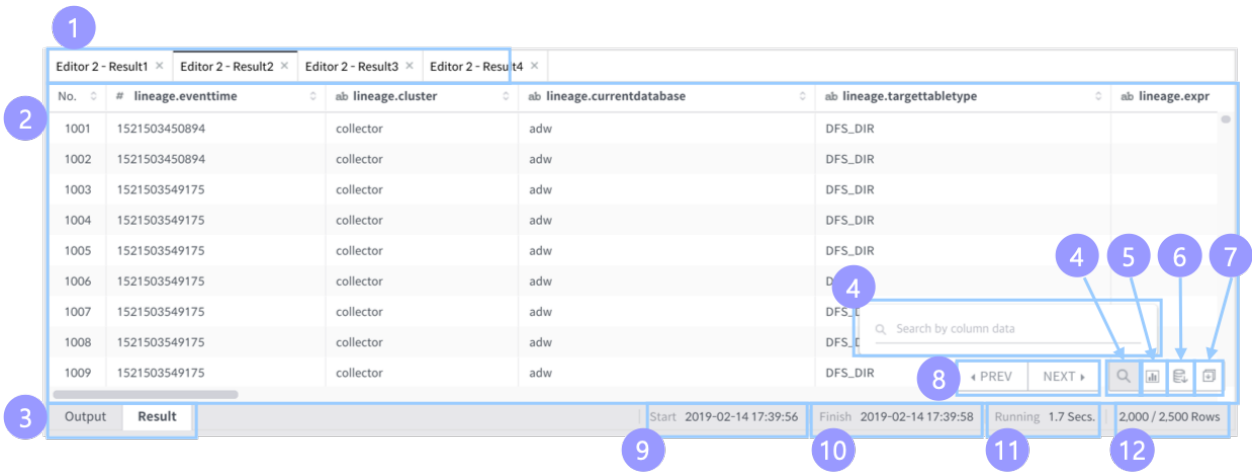


During query execution

1. **Query result tabs:** When multiple queries are executed, a different tab is created for each query to show its result. While a query's execution is in progress, "Loading" is displayed in its tab title.
2. **Query log:** Shows an execution log for the query. In the case of a Hive connection, a Hive job log is additionally displayed.
3. **Cancel:** Cancels the execution of the query. The time taken for cancelation may vary with the DB type.
4. **Query execution phase:** Shows the current phase of query execution. There are a total of five query execution phases.
   - Getting connection
   - Creating statement
   - Executing query
   - Getting result set
   - Done!
5. **No. of the current query:** Shows the number of the currently executed query when multiple queries are executed.
6. **Output/Result tabs:** By clicking either tab, you can switch to the query log/result view.
7. **Query status:** Shows the query's status from among:
   - Running query
   - Query execution failed..
   - Query execution canceled..
8. **Query start time:** Displays when the query execution started.
9. **Query running time:** Displays how long it took to execute the query.

After query execution

1. **Query result tabs:** When multiple queries are executed, a different tab is created for each query to show its result. While a query's execution is in progress, "Loading" is displayed in its tab title.



2. **Data details:** Shows a data table resulting from executing the query. You can copy this data output to the clipboard.
3. **Output/Result tabs:** By clicking either tab, you can switch to the query log/result view.
4. **Search for column data:** Searches for a column or value in the resulting table.
5. **Chart preview:** Draws a virtual chart of the query results. This chart is only for visualization; it is not stored in the workspace. (See *Chart* for how to handle it)
6. **Save as Data source:** Stores the query results into a data source in the workspace. A dialog box will pop up to create a data source, and the resulting table is used instead of selecting a data connection and a table. Therefore, you will be immediately prompted to set the schema definition and ingestion cycle. (See *Create a data source* for how to handle it)
7. **Export CSV file:** Downloads the resulting table into a local file (CSV).
8. **Data page navigation:** If the resulting data includes more than 1,000 rows, you can navigate the data pages using the Prev and Next buttons.
9. **Query start time:** Displays when the query execution started.
10. **Query finish time:** Displays when the query execution finished.
11. **Query running time:** Displays how long it took to execute the query.
12. **Query data rows:** Shows the number of rows of the resulting data and the current page number.

7.2.5 Extra tools section

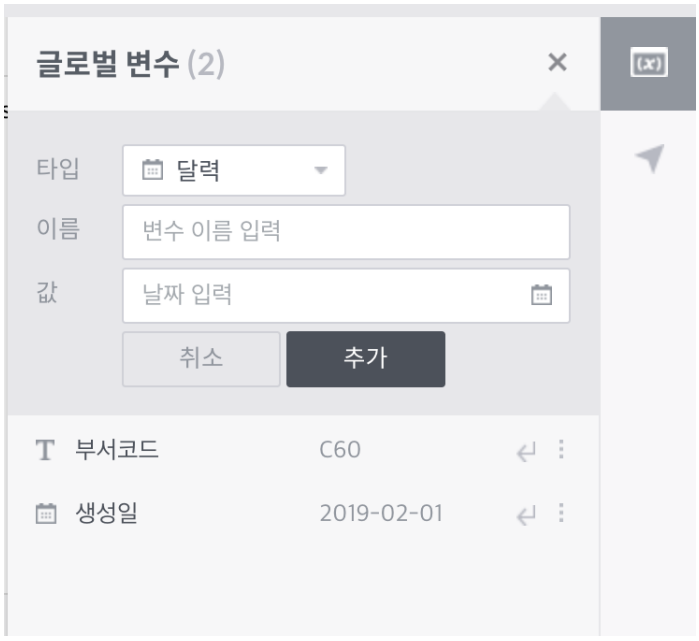
The extra tools section provides useful tools for the workbench.

- Setting up global variables for repeatedly used statements (See *Setting up global variables*)
- Navigation to move to another workbench (See *Workbench Navigation*)

Setting up global variables

If a certain type of statement is repeatedly used with a different value for each query run, set the variable element as a "global variable" for convenient use.

- **Variable type:** You can select either a calendar or text type.



- **Add new variable:** Select the variable type you want and click "Add new variable." A new global variable will be added in the query editor.
- **Name:** Enter a name for the variable.
- **Variable value:** For a calendar variable, select a date; for a text variable, select a text value.

Workbench Navigation

Used to move to another workbench. Click the target workbench to move to.

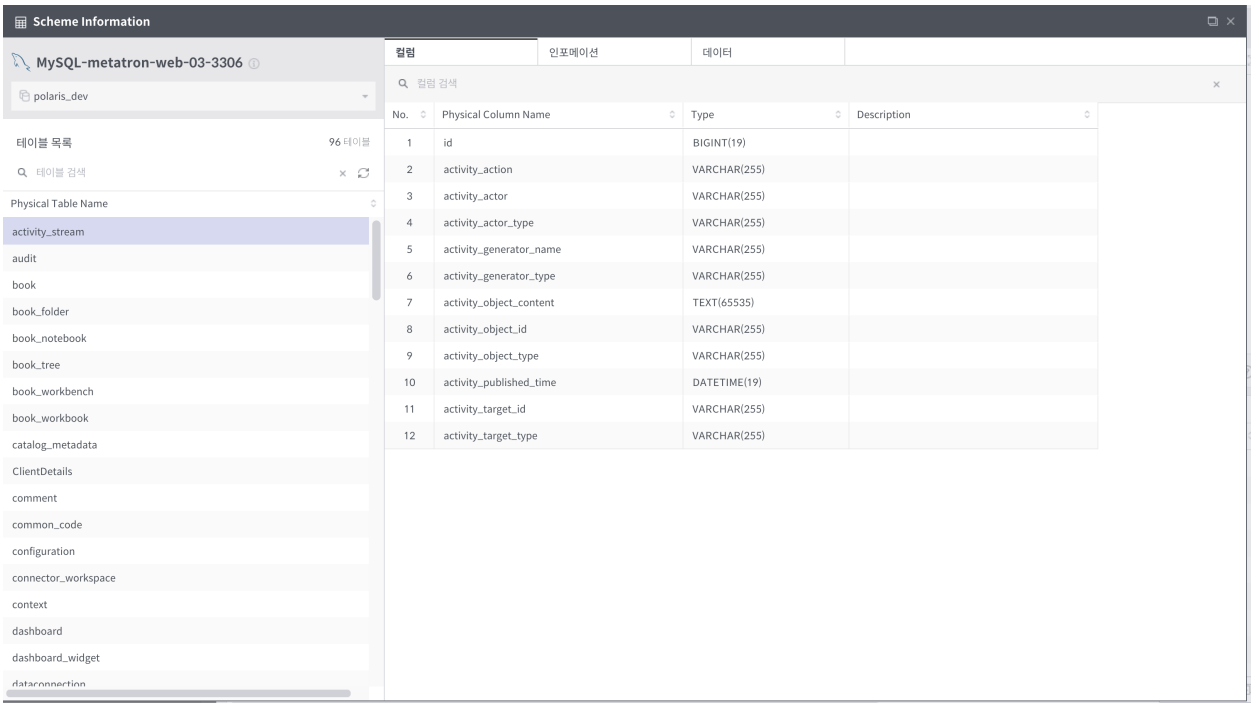
- **Search for workbench:** Search for a workbench stored in the workspace.
- **Workbench list:** Displays all workbenches stored in the workspace. Click a workbench in the list to move to that workbench.

7.2.6 Schema browser

Displays the table list of the selected database, and information of the columns and records in each table.

- **Column:** Shows the names and data types of all columns of the selected table.
- **Information:** Displays attributes of the selected table.
- **Data:** Displays data of the selected table. A maximum of 50 rows can be viewed.


워크벤치 네비게이션 (20)
검색
No.
워크벤치 이름
수정일
32 test sales 2019-02-07
31 asas 2019-01-30
30 ㅁㄴㅇㅁㄴㅇ 2019-01-30
29 aaaaaa 2019-01-29
28 VV 2019-01-24
27 Test 2019-01-22
26 s 2019-01-17
25 Test 2019-01-17
24 test 2019-01-16
23 test 2019-01-15
22 test_abc 2019-01-15
21 Test-Workbench-omelas 2019-01-14
20 Test2-mysql 2019-01-11
19 ㅎㅎ 2019-01-07
18 1234 2019-01-03
17 워크벤치 2018-12-26
16 sample workbench 2018-12-25
15 hive-trip-data 2018-12-21
14 man 2018-12-20
13 Test 2018-12-19
+ 더보기


Scheme Information
MySQL-metatron-web-03-3306
polaris_dev
테이블 목록
96 테이블
테이블 검색
Physical Table Name
activity_stream
audit
book
book_folder
book_notebook
book_tree
book_workbench
book_workbook
catalog_metadata
ClientDetails
comment
common_code
configuration
connector_workspace
context
dashboard
dashboard_widget
datasconnection
컬럼
인포메이션
데이터
컬럼 검색
No.
Physical Column Name
Type
Description
1 id BIGINT(19)
2 activity_action VARCHAR(255)
3 activity_actor VARCHAR(255)
4 activity_actor_type VARCHAR(255)
5 activity_generator_name VARCHAR(255)
6 activity_generator_type VARCHAR(255)
7 activity_object_content TEXT(65535)
8 activity_object_id VARCHAR(255)
9 activity_object_type VARCHAR(255)
10 activity_published_time DATETIME(19)
11 activity_target_id VARCHAR(255)
12 activity_target_type VARCHAR(255)

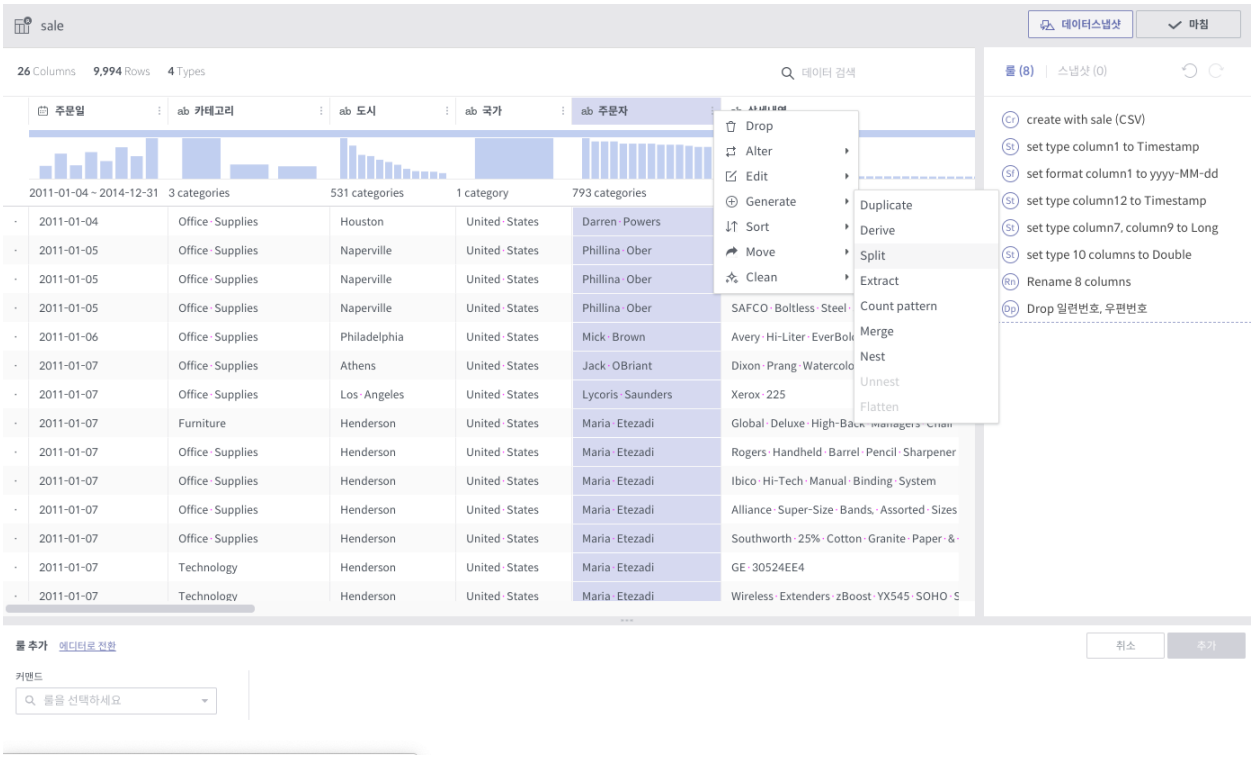
CHAPTER

EIGHT

DATA PREPARATION

Data Preparation is a tool that creates transformation rules to transform files and tables for more convenient analysis of datasets, and saves the results into HDFS or Hive.

Advantages of data preparation in Metatron Discovery



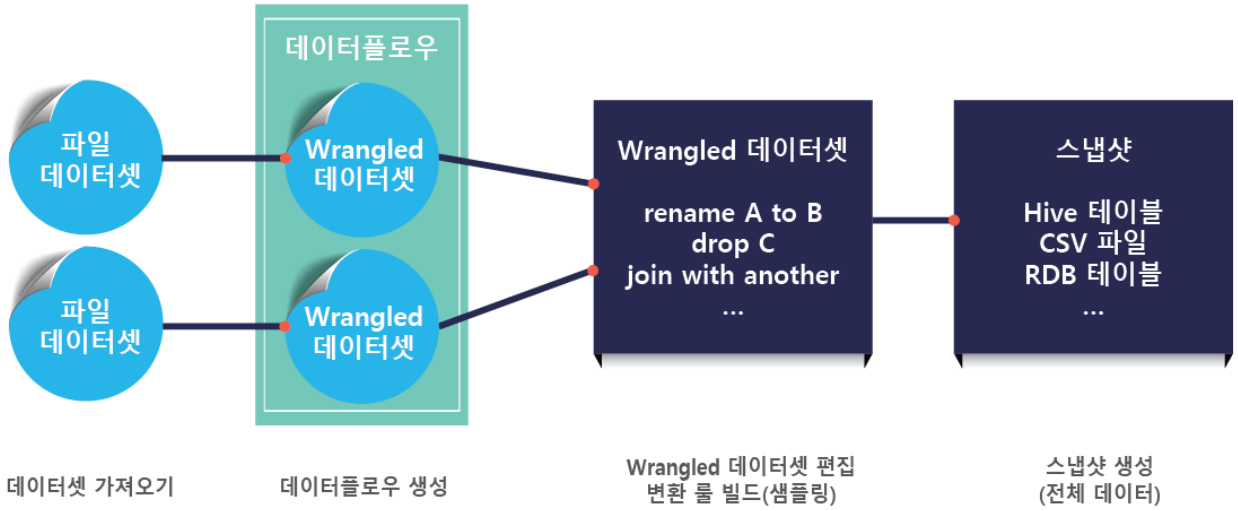
Users can create transformation rules by following the step-by-step process as shown in the above GUI. Since the transformation results from each step are stored in memory together with the data distribution, users can easily check the results through the simple click of a button and perform **undo** and **redo** just like using a text editor.

Based on these characteristics, the data preparation tool offers the following advantages:

- Users unfamiliar with programming or data processing can obtain the desired results.
- Adding a transformation rule usually involves programming or writing an SQL query. However, Metatron Discovery's Data Preparation provides a GUI for **exploratory transformation** that enables the creation of transformation rules simply by clicking a button or typing.

- Basic data transformation is conducted automatically. For instance, a type cast is automatically applied to columns comprised of numerals. This is made possible by the **undo** and rule deletion functions.
- Data of different forms can be combined as desired (e.g. reference file + fact table).
- The results of data refinement can be shared with others, thus reducing the burden of exchanging physical data.
- Storage space is saved and **information life cycle (ILM)** shortened by deleting the actual data and retaining only the transformation rules involved. The actual data can be easily created whenever needed.

Structure of data preparation in Metatron Discovery



As shown in the above figure, data preparation is comprised of a **dataset** built from the target data, a **dataflow** that defines transformation rules for the designated dataset, and a **data snapshot** that shows the transformation results.

8.1 Create a dataset

A **dataset**, which is the basic unit of data preparation, refers to an entity subject to data operations. Datasets are either **imported datasets** and **wrangled datasets**.

- **Imported Dataset:** A source data entity before the implementation of transformation rules
- **Wrangled dataset:** A data entity subject to analysis following the implementation of transformation rules

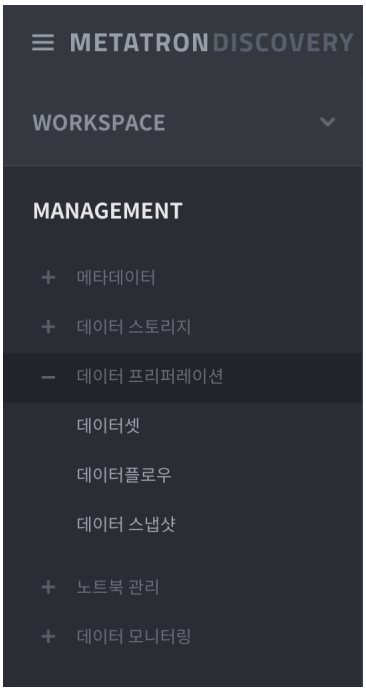
A wrangled dataset is created during the **dataflow** setting process, which defines transformation rules, while an imported dataset is created during this dataset creation procedure.

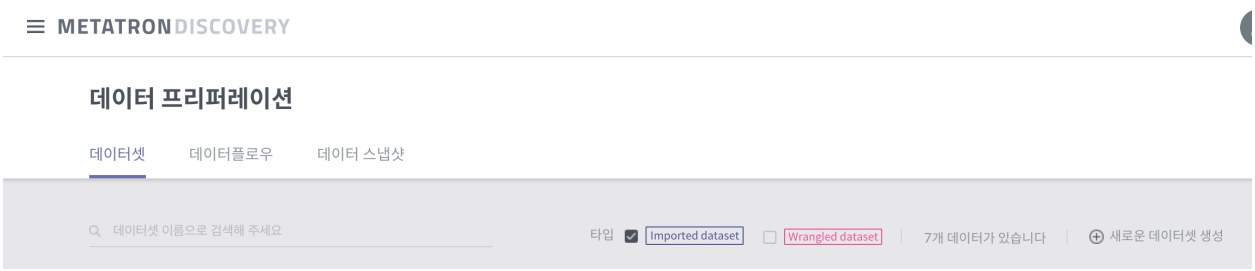
The Dataflow menu can be accessed under **MANAGEMENT > Data Preparation > Dataset** on the left-hand panel of the main screen.

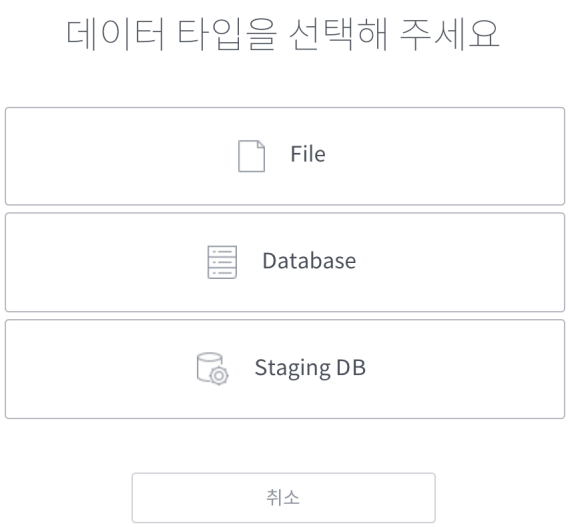
Next, on the upper right of the **dataset** page, click the **+ Generate new dataset** button to create a new dataset.

In the dataset creation page, select the dataset type.

- **My file:** Create a dataset by opening the user's local file or via a URI (upcoming feature) (See *Create a dataset from a file* for a detailed procedure).


METATRONDISCOVERY
WORKSPACE
MANAGEMENT
메타데이터
데이터 스토리지
데이터 프리퍼레이션
데이터셋
데이터플로우
데이터 스냅샷
노트북 관리
데이터 모니터링


METATRONDISCOVERY
데이터 프리퍼레이션
데이터셋
데이터플로우
데이터 스냅샷
데이터셋 이름으로 검색해 주세요
타입
Imported dataset
Wrangled dataset
7개 데이터가 있습니다
새로운 데이터셋 생성


데이터 타입을 선택해 주세요
File
Database
Staging DB
취소

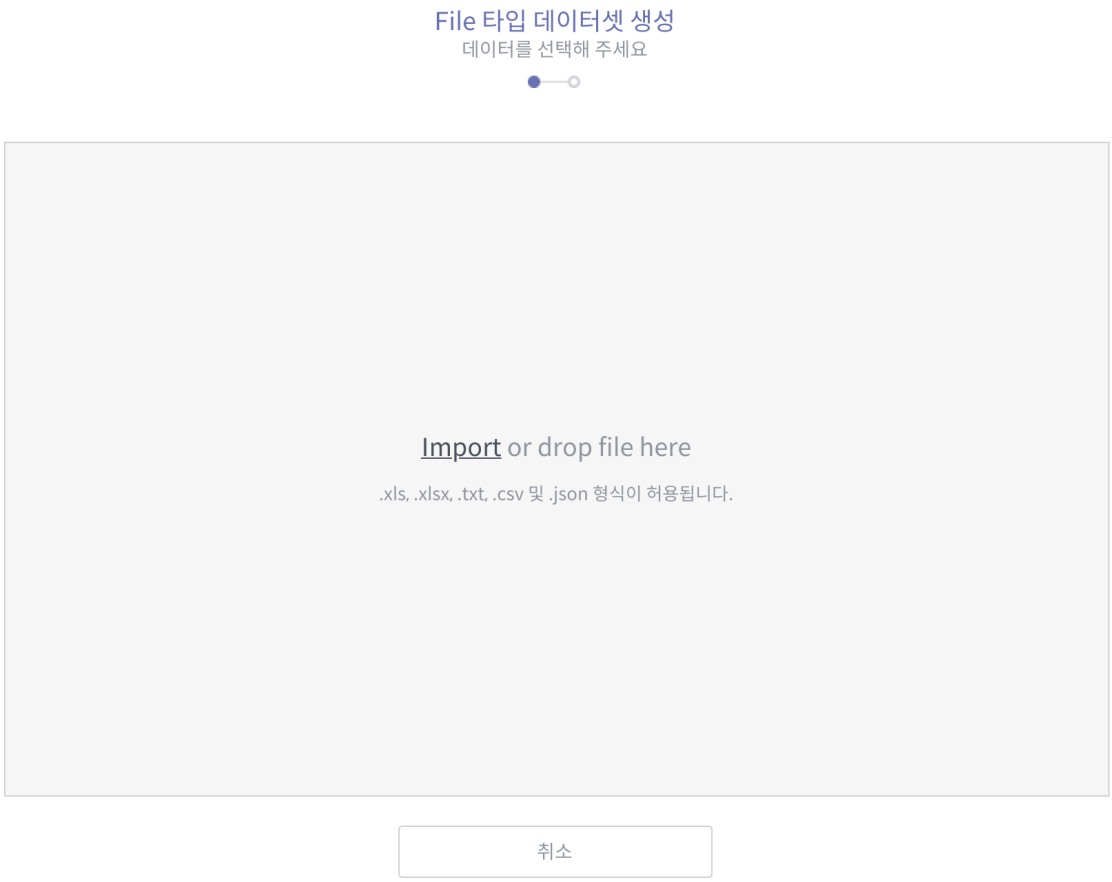
- **Database:** Create a dataset using external database access information and queries (See *Create a dataset from a database* for a detailed procedure).
- **Staging DB:** Create a dataset from the staging DB built in Metatron (See *Create a dataset from staging DB* for a detailed procedure).

Note: The Staging DB is an in-cluster database that stores data temporarily in order to facilitate data loading. Hive is generally used for it.

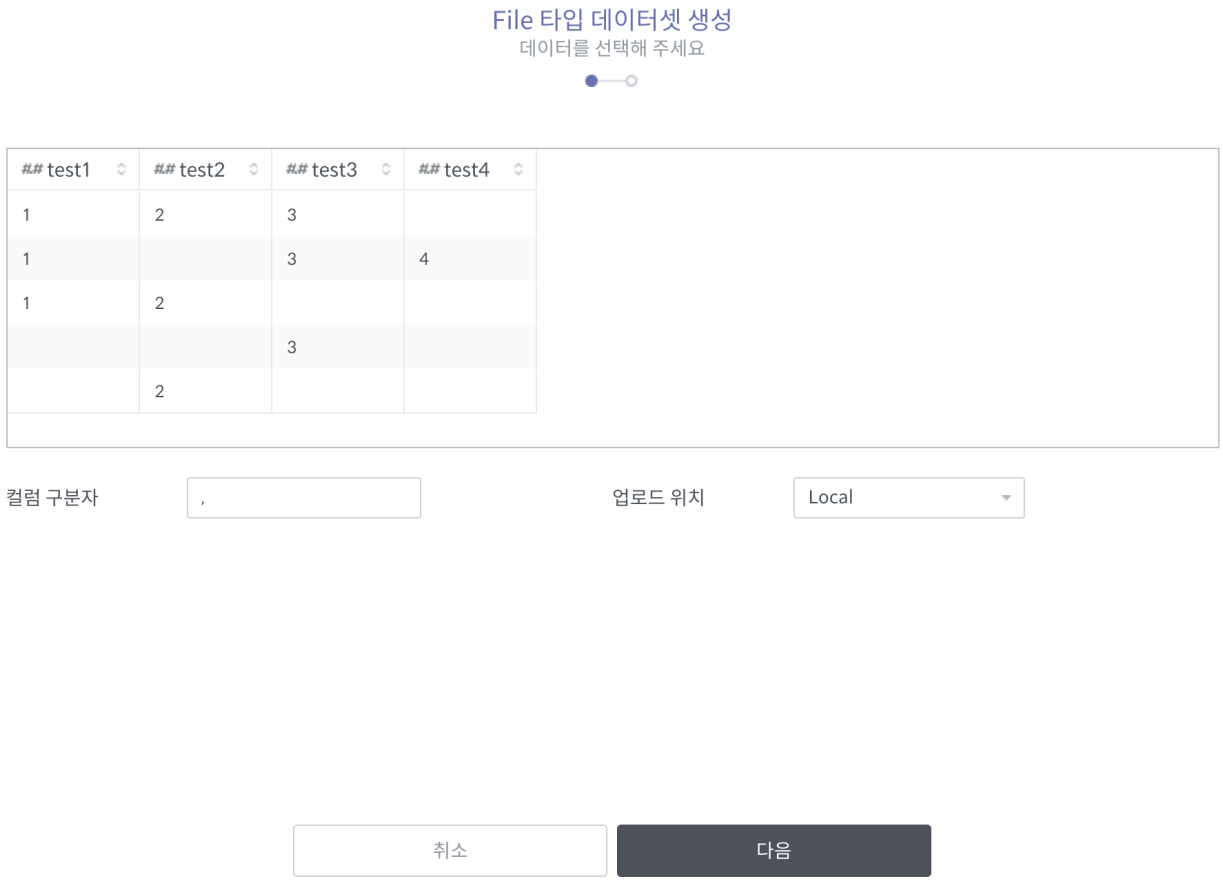
8.1.1 Create a dataset from a file

Create a dataset by opening the user's local file or via a URI (upcoming feature).

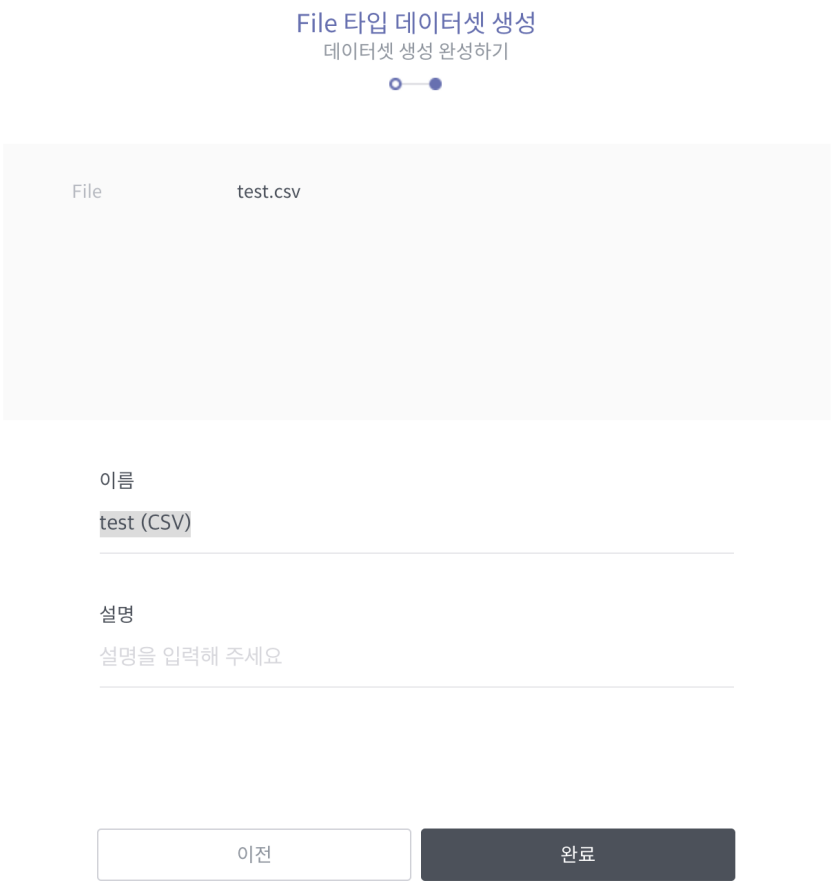
1. On the data type selection page, select **My File**.
2. Select a file to be used as a data source from your local PC. You can click the **Import** button to select a file, or drag and drop the file into the box. Once a file is selected, click Next.



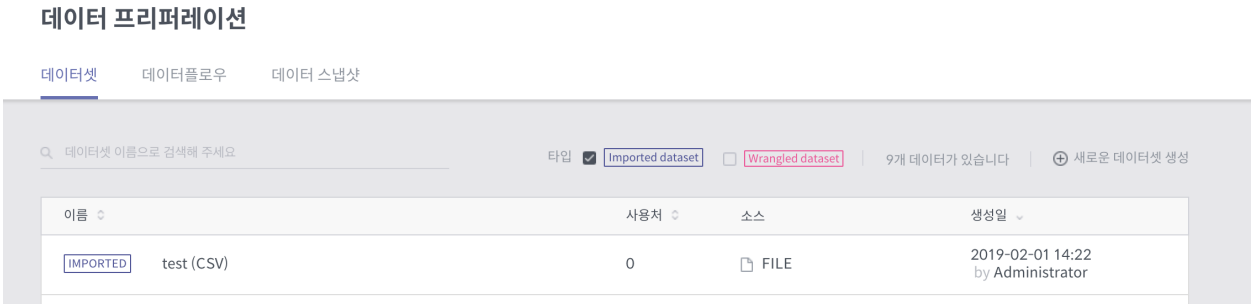
3. Check the grid of the uploaded file, and designate a column delimiter. Proceed if the data is successfully displayed.


File 타입 데이터셋 생성
데이터를 선택해 주세요
test1
test2
test3
test4
1 2 3
1 3 4
1 2
3
2
컬럼 구분자
,
업로드 위치
Local
취소
다음

4. Enter the **Name** and **Description** of the dataset, and click the **Done** button.



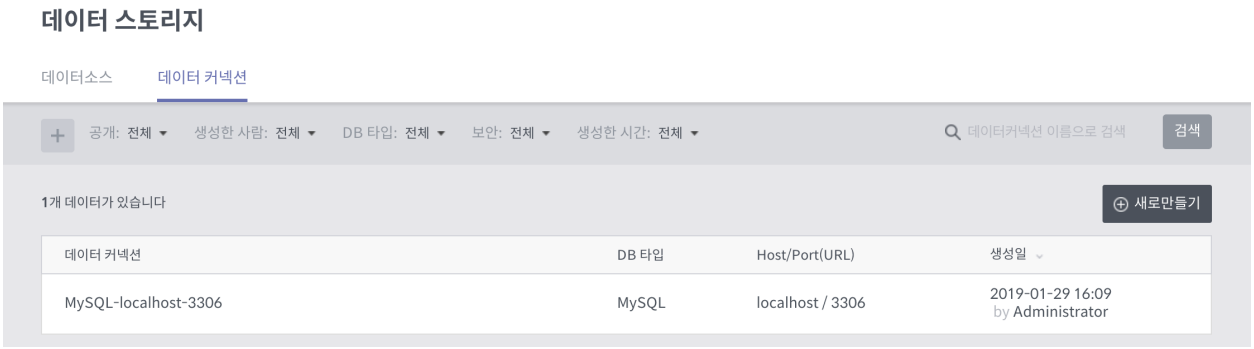
5. Once the dataset is created, the dataset list is displayed. You can check that the list contains the newly created dataset.



8.1.2 Create a dataset from a database

Create a dataset using external database access information and queries.

To create a dataset from a database, you should first create a data connection. See *Create a data connection* for a detailed procedure.



After establishing the data connection, go to **MANAGEMENT > Data Preparation > Dataset > + Generate new dataset**.

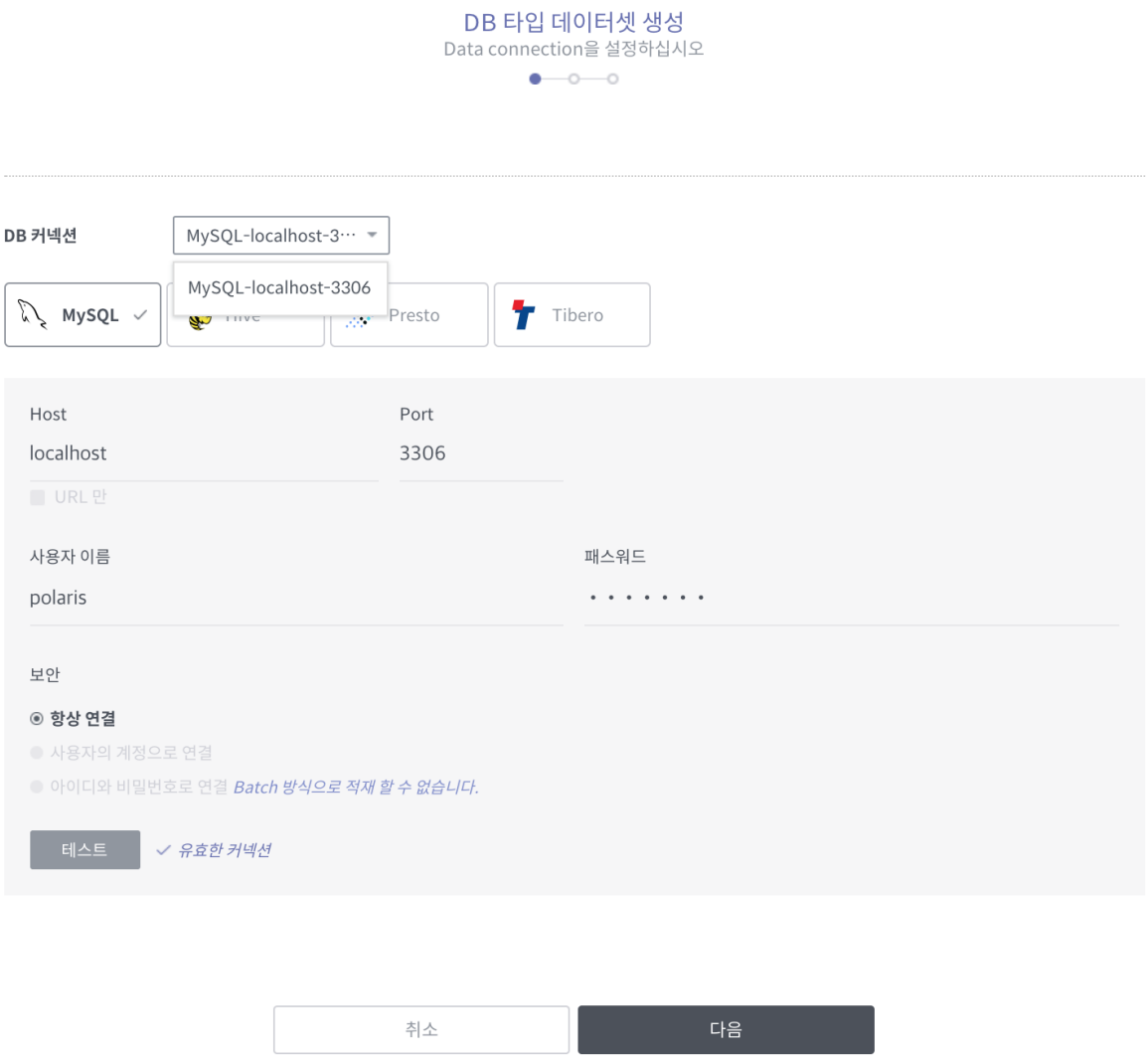
1. On the data type selection page, select **Database**.
2. Select the data connection, and press the **Test** button to check that the connection is valid.
3. Select the data. You can either select a table from the connected database, or write a query yourself.
    - **Table:** Select a database and a table to display the table's data. Once the data being ingested has been displayed, confirm the data and click **Next**.
    - **Query:** Write a query to import the data you want, and click **Run** to display the data in the lower section. Confirm the data and click **Next**.
4. Enter the **Name** and **Description** of the dataset, and click the **Done** button.
5. Once the dataset is created, the dataset list is displayed. You can check that the list contains the newly created dataset.

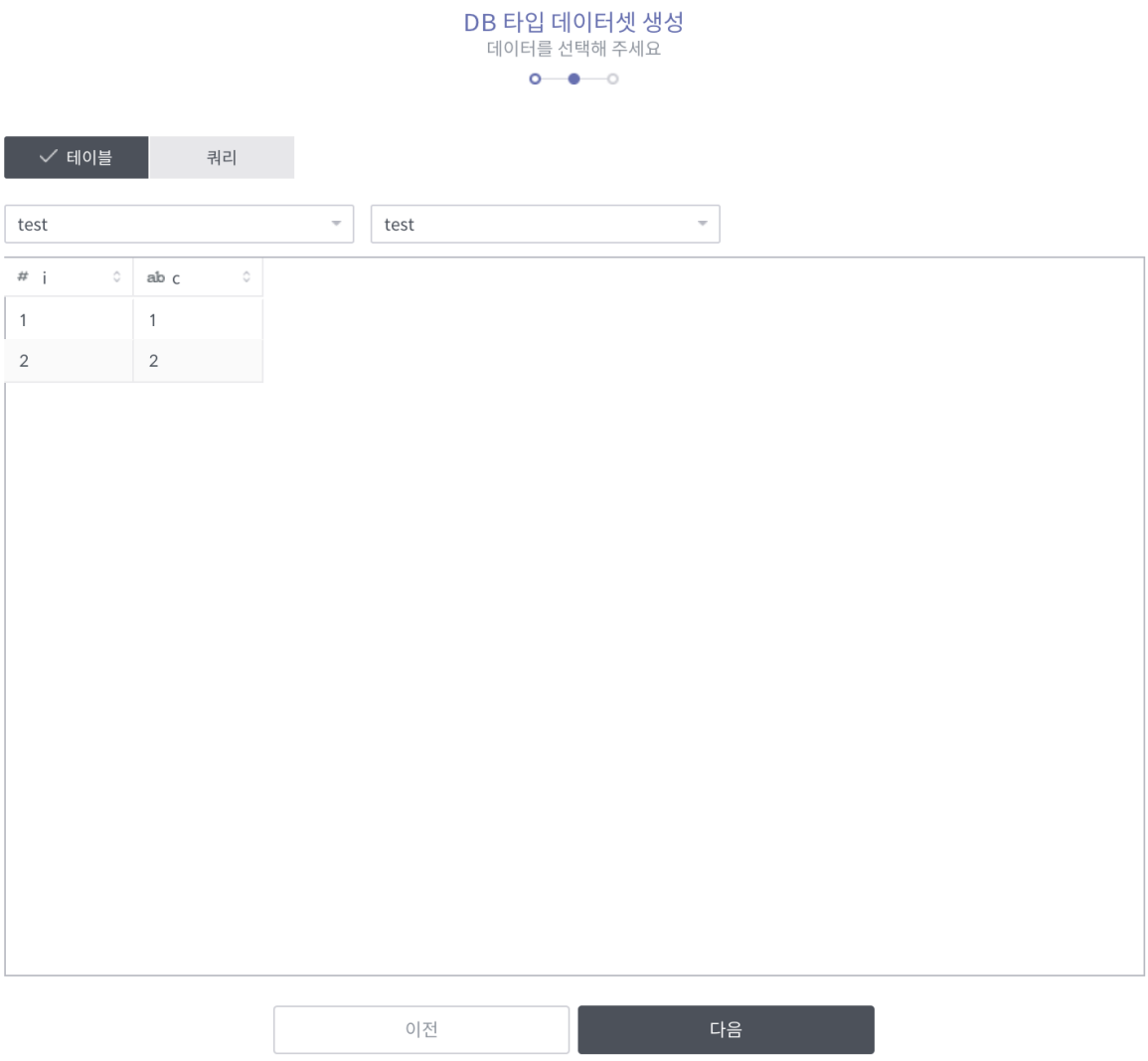
8.1.3 Create a dataset from staging DB

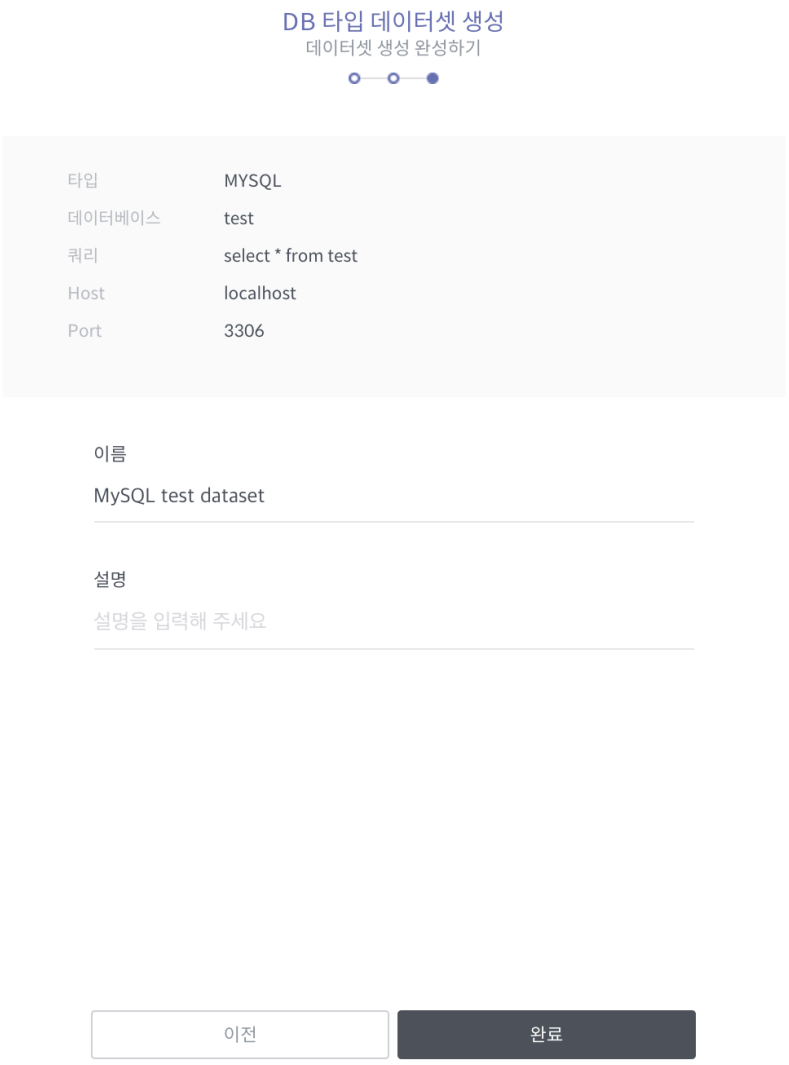
Create a dataset from the staging DB built in Metatron.

The creation of a staging DB dataset is the same as dataset creation from a database, but does not involve the selection of a data connection.

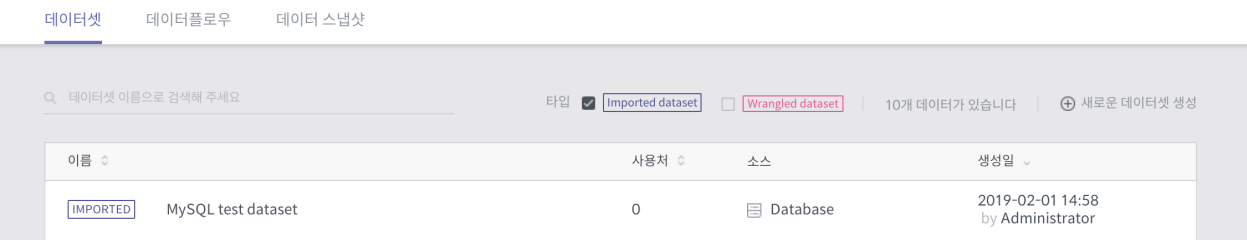
1. On the data type selection page, select **Staging DB**.
2. Select the data. You can either select a table from the connected database, or write a query yourself.
    - **Table:** Select a database and a table to display the table's data. Once the data being ingested has been displayed, confirm the data and click **Next**.
    - **Query:** Write a query to import the data you want, and click **Run** to display the data in the lower section. Confirm the data and click **Next**.
3. Enter the **Name** and **Description** of the dataset, and click the **Done** button.

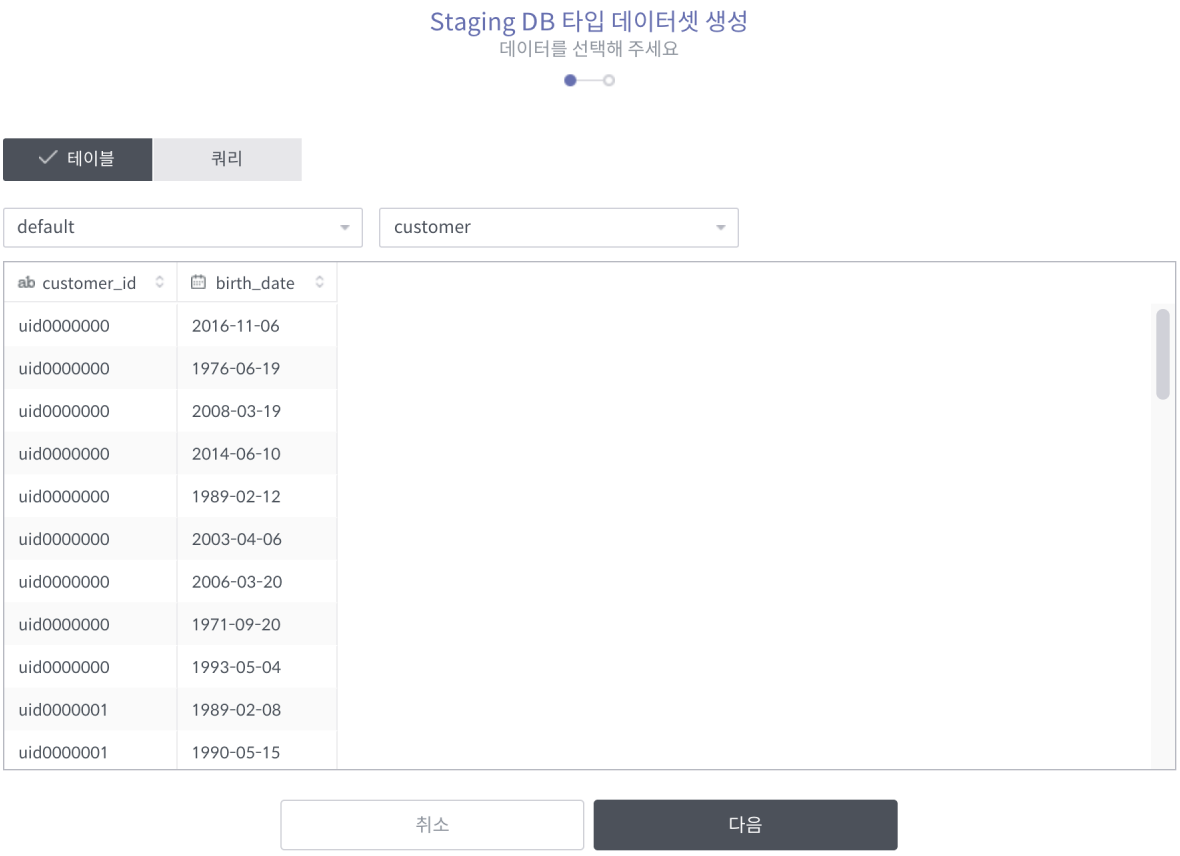

DB 타입 데이터셋 생성
Data connection을 설정하십시오
DB 커넥션
MySQL-localhost-3···
MySQL-localhost-3306
MySQL
Presto
Tibero
Host
localhost
Port
3306
URL 만
사용자 이름
polaris
패스워드
보안
항상 연결
사용자의 계정으로 연결
아이디와 비밀번호로 연결 Batch 방식으로 적재 할 수 없습니다.
테스트
유효한 커넥션
취소
다음

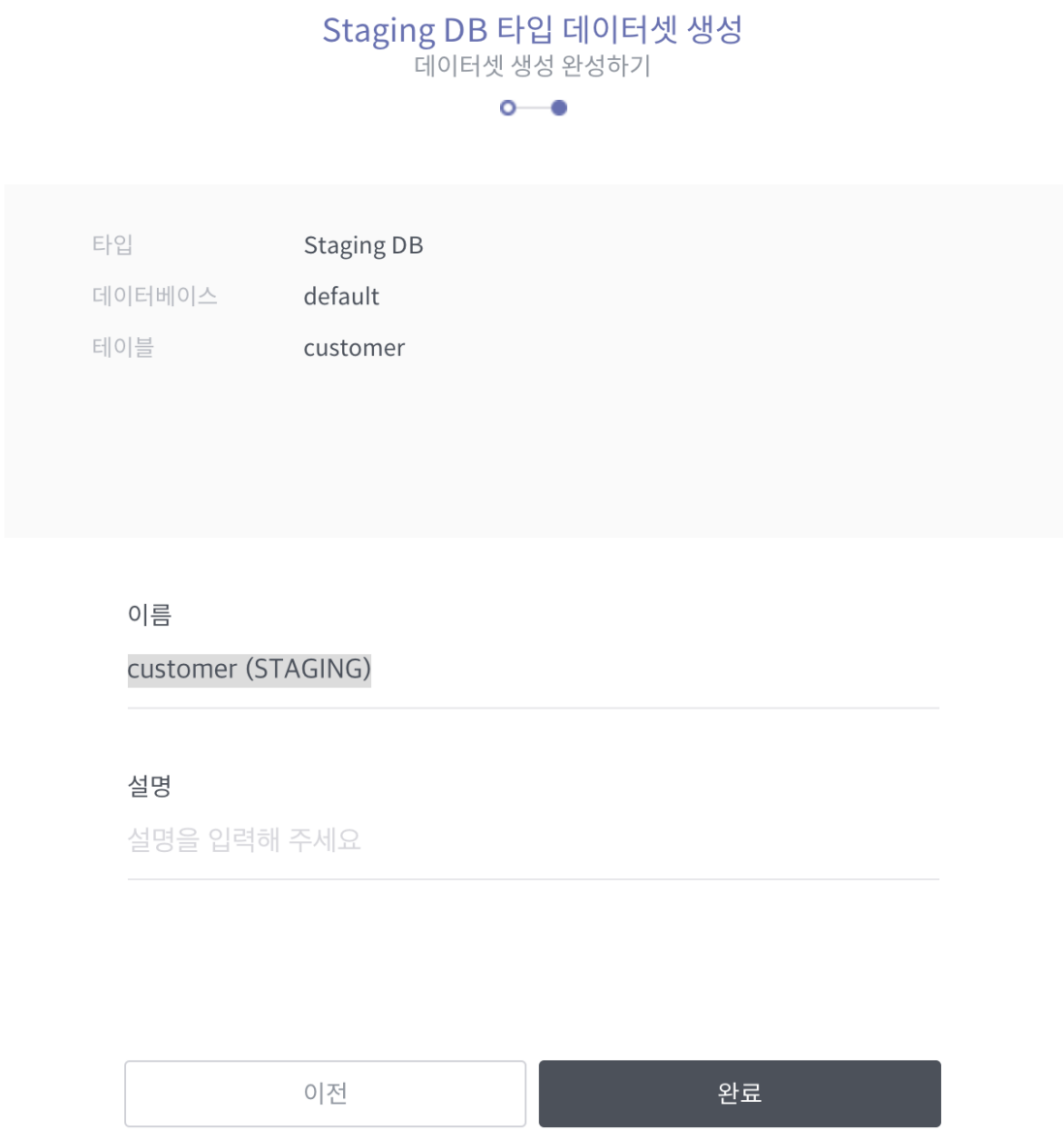

DB 타입 데이터셋 생성
데이터를 선택해 주세요
테이블
쿼리
test
test
i
ab c
1
1
2
2
이전
다음


DB 타입 데이터셋 생성
데이터셋 생성 완성하기
타입 MYSQL
데이터베이스 test
쿼리 select * from test
Host localhost
Port 3306
이름
MySQL test dataset
설명
설명을 입력해 주세요
이전
완료

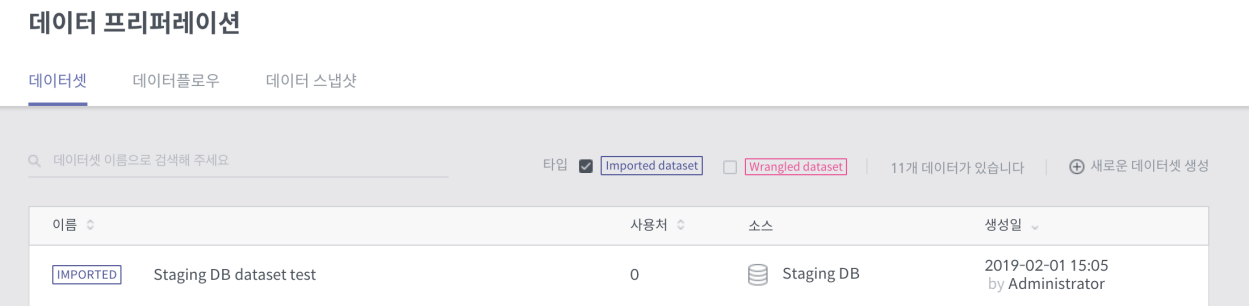
데이터 프리퍼레이션


데이터셋
데이터플로우
데이터 스냅샷
데이터셋 이름으로 검색해 주세요
타입
Imported dataset
Wrangled dataset
10개 데이터가 있습니다
새로운 데이터셋 생성
이름
사용처
소스
생성일
IMPORTED
MySQL test dataset
0
Database
2019-02-01 14:58
by Administrator


Staging DB 타입 데이터셋 생성
데이터를 선택해 주세요
테이블
쿼리
default
customer
customer_id
birth_date
uid0000000 2016-11-06
uid0000000 1976-06-19
uid0000000 2008-03-19
uid0000000 2014-06-10
uid0000000 1989-02-12
uid0000000 2003-04-06
uid0000000 2006-03-20
uid0000000 1971-09-20
uid0000000 1993-05-04
uid0000001 1989-02-08
uid0000001 1990-05-15
취소
다음


Staging DB 타입 데이터셋 생성
데이터셋 생성 완성하기
타입
Staging DB
데이터베이스
default
테이블
customer
이름
customer (STAGING)
설명
설명을 입력해 주세요
이전
완료

4. Once the dataset is created, the dataset list is displayed. You can check that the list contains the newly created dataset.



8.2 Manage a dataflow

A **dataflow** is the unit of processing a **dataset**. A single dataflow can be associated with multiple datasets to perform transformations. That is, a dataset must belong to a dataflow for transformation rules to be applied. It forms a relationship such as a "join" or "union" with other datasets.

As shown below, the dataflow details page shows the dependency among all datasets in a dataflow, and the transformation rules applied to each dataset.

The following subsections cover the processes involved in defining a dataflow, such as **adding a dataset**, **editing transformation rules**, and **creating a data snapshot with transformation results**.

The Dataflow menu can be accessed under **MANAGEMENT > Data Preparation > Dataflow** on the left-hand panel of the main screen.

8.2.1 Add a dataset

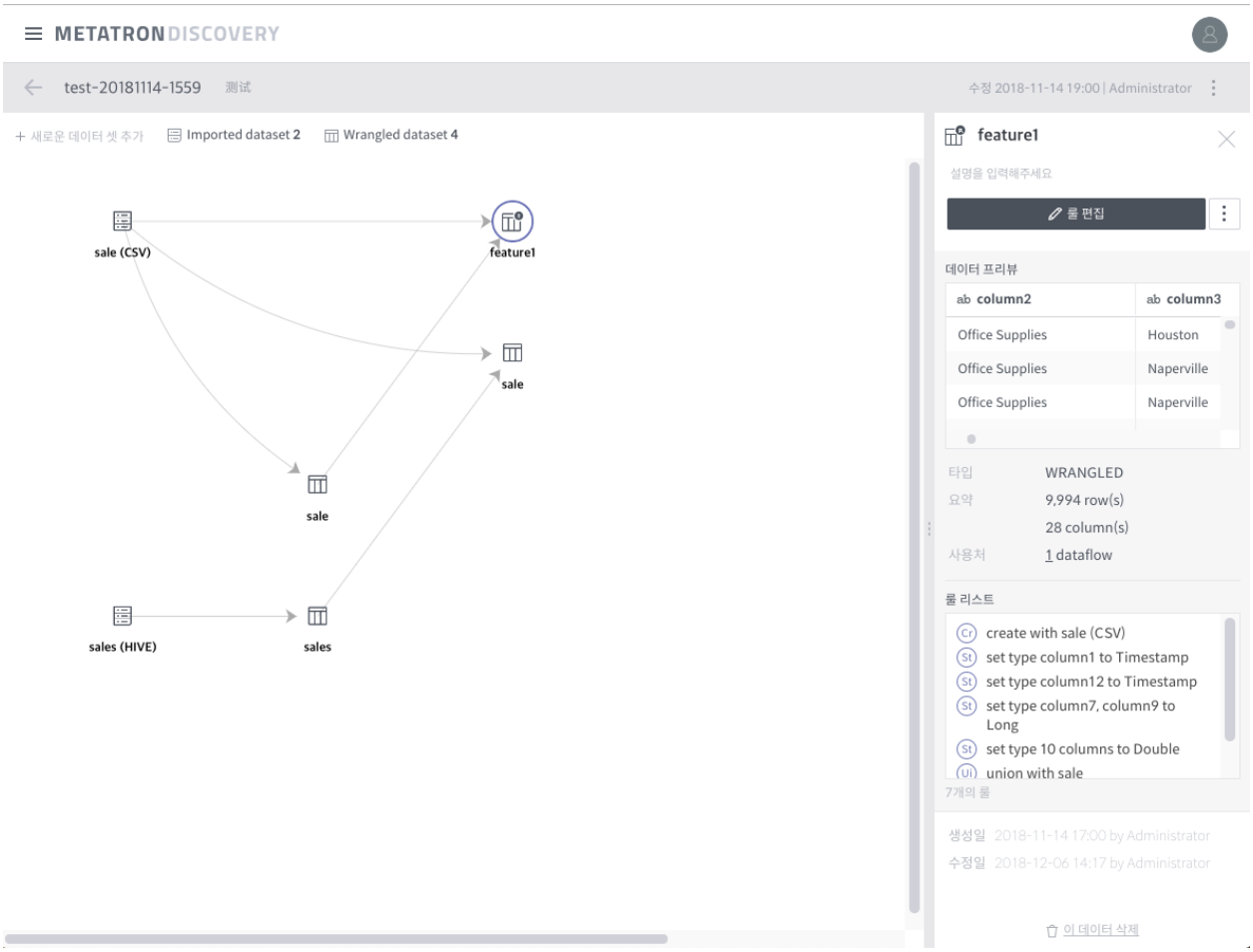
The first step in defining a dataflow is to add a dataset. This can be conducted using the two methods described below:

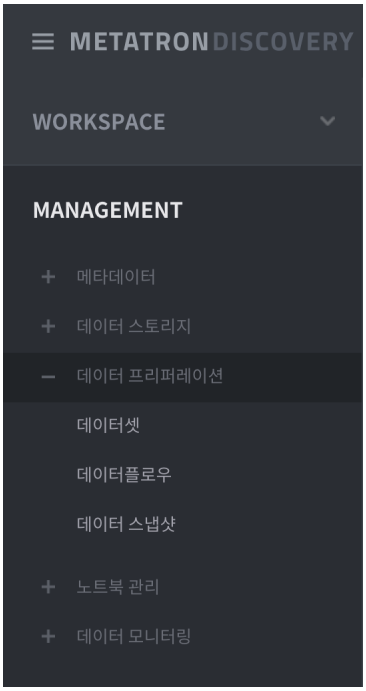
- *Adding a dataset after creating an empty dataflow*
- *Creating a dataflow in the dataset details page*

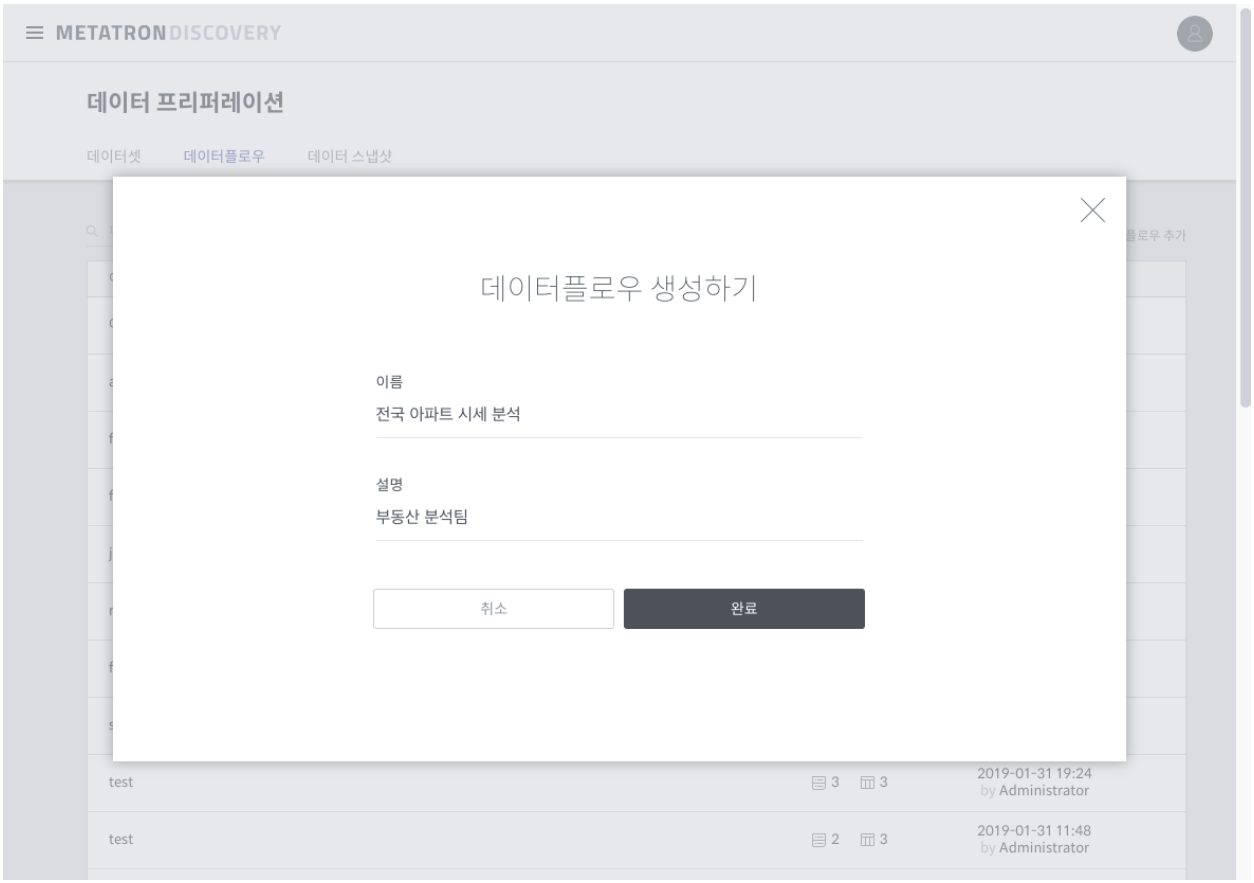
Adding a dataset after creating an empty dataflow

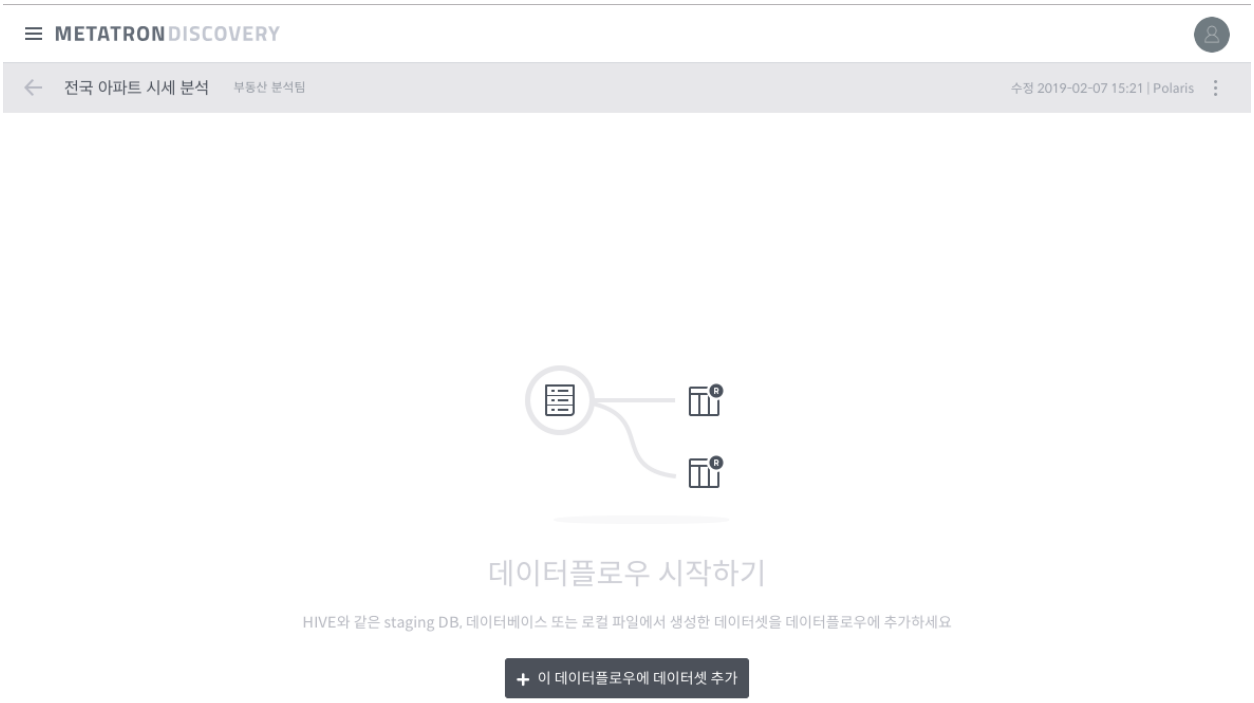
1. Click **Add a dataflow** on the upper right of the **Dataflow** page.
2. Enter the **Name** and **Description** for the dataflow, and click **Done** to create an empty dataflow.

3. Click the **Add dataset to this dataflow** button on the center of the page.


METATRONDISCOVERY
test-20181114-1559
测试
수정 2018-11-14 19:00 | Administrator
+ 새로운 데이터 셋 추가
Imported dataset 2
Wrangled dataset 4
sale (CSV)
feature1
sale
sale
sales (HIVE)
sales
feature1
설명을 입력해주세요
룰 편집
데이터 프리뷰
ab column2
ab column3
Office Supplies
Houston
Office Supplies
Naperville
Office Supplies
Naperville
타입
WRANGLED
요약
9,994 row(s)
28 column(s)
사용처
1 dataflow
룰 리스트
create with sale (CSV)
set type column1 to Timestamp
set type column12 to Timestamp
set type column7, column9 to Long
set type 10 columns to Double
union with sale
7개의 룰
생성일 2018-11-14 17:00 by Administrator
수정일 2018-12-06 14:17 by Administrator
이 데이터 삭제


METATRONDISCOVERY
WORKSPACE
MANAGEMENT
메타데이터
데이터 스토리지
데이터 프리퍼레이션
데이터셋
데이터플로우
데이터 스냅샷
노트북 관리
데이터 모니터링


METATRONDISCOVERY
데이터 프리퍼레이션
데이터셋
데이터플로우
데이터 스냅샷
데이터플로우 생성하기
이름
전국 아파트 시세 분석
설명
부동산 분석팀
취소
완료
test
2019-01-31 19:24
by Administrator
test
2019-01-31 11:48
by Administrator



4. Select the datasets to be added.

5. When an imported dataset and its corresponding wrangled dataset are created, click the **Edit rules** button to edit rules (see *Edit rules* for a detailed procedure).

Creating a dataflow in the dataset details page

In the dataset details page, click the **Create dataflow with this dataset** button to create a dataflow, and proceed until the step before **Edit rules**.

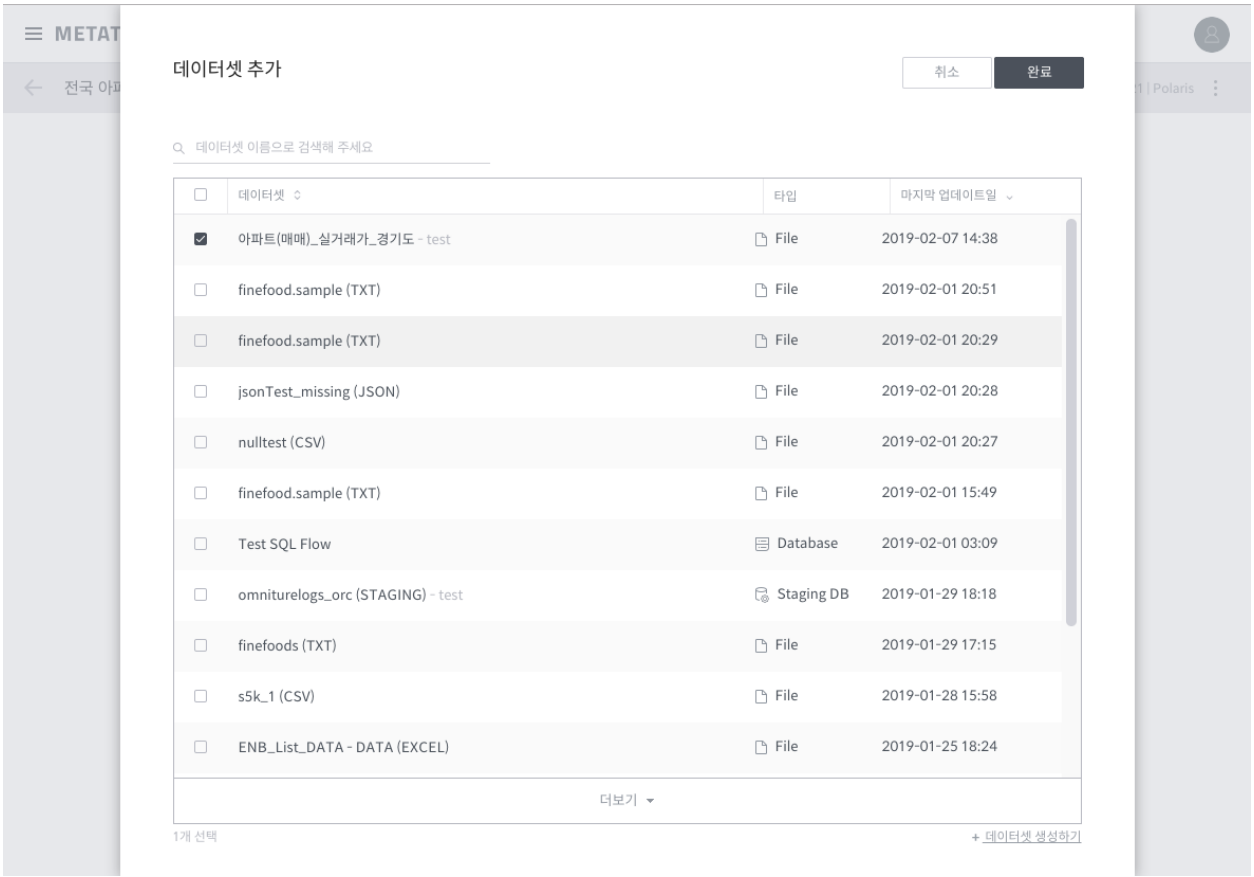
> **Note:** The dataflow is named based on the name of the dataset.

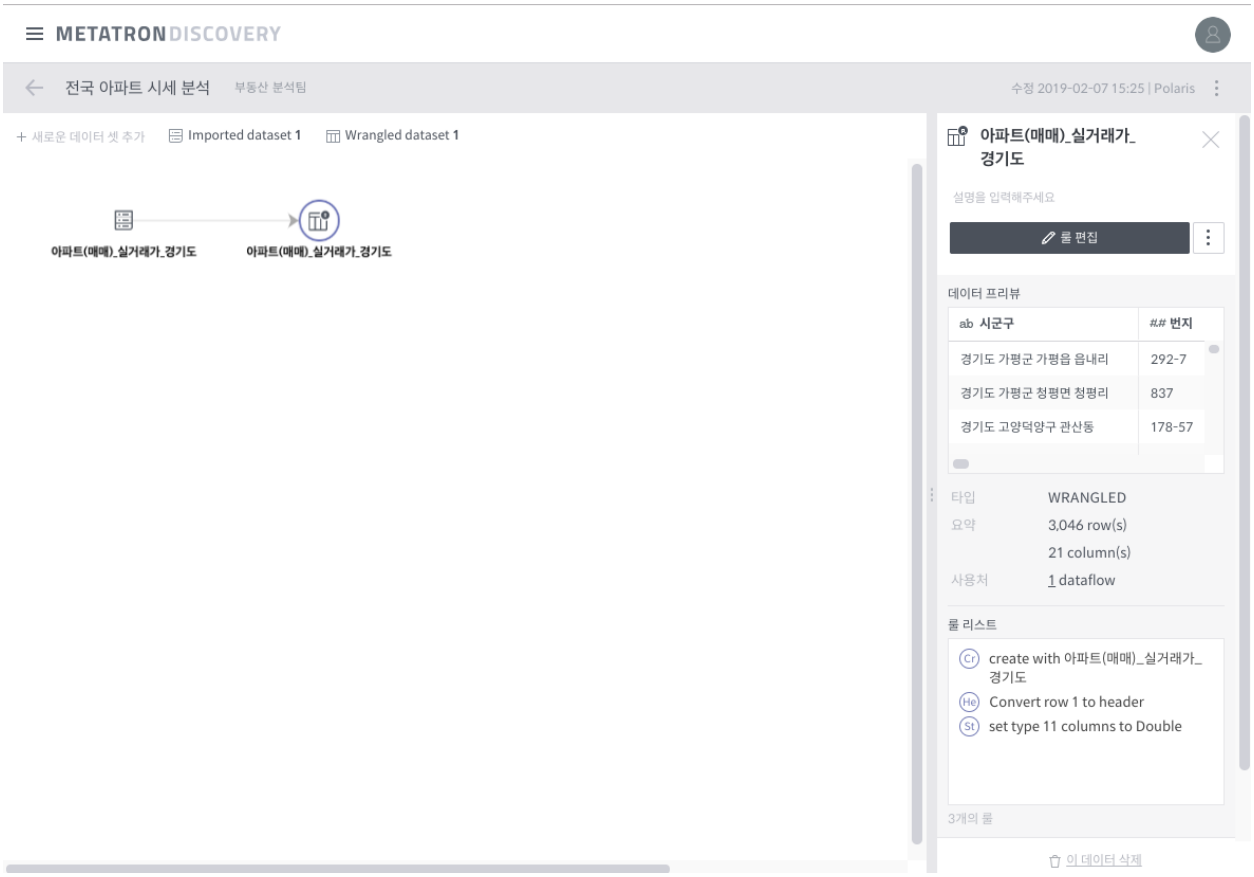
8.2.2 Edit rules

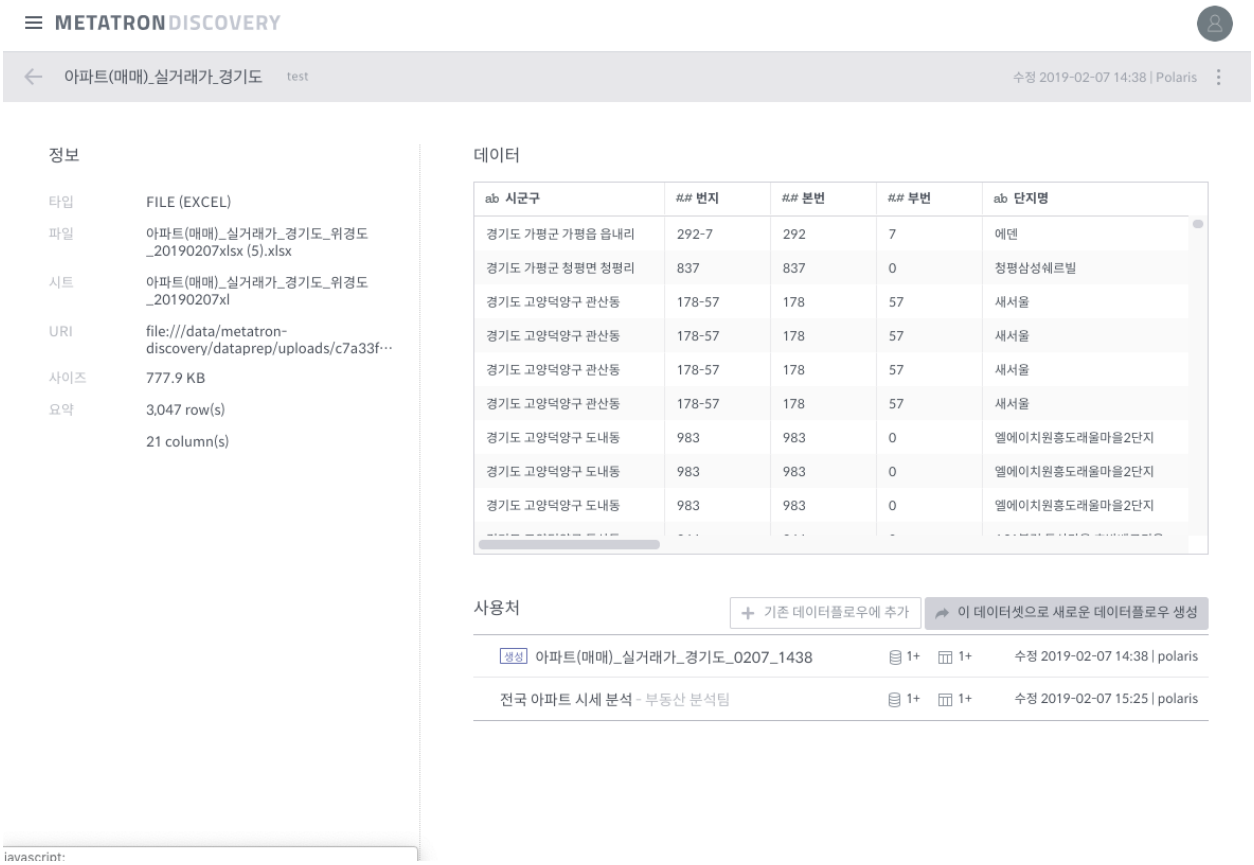
The key task in data preparation is to create rules for data transformation (usually refinement). The transformation rules and input/output specifications are combined to be applied to actual data or other similar data, or scheduling is performed for such tasks.

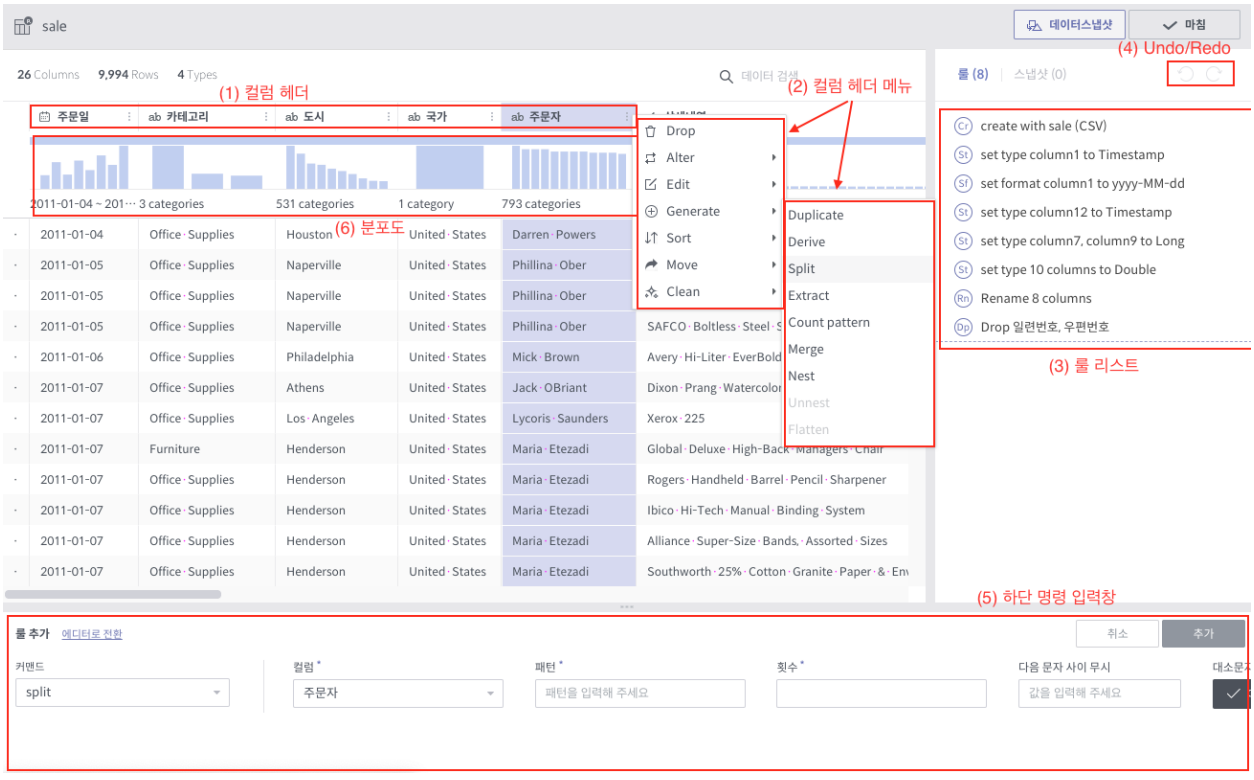
Below are instructions on creating rules, checking the results, and modifying or deleting rules.

The Edit Rules page consists of the following:


데이터셋 추가
취소
완료
데이터셋 이름으로 검색해 주세요
데이터셋
타입
마지막 업데이트일
아파트(매매)_실거래가_경기도 - test
File
2019-02-07 14:38
finefood.sample (TXT)
File
2019-02-01 20:51
finefood.sample (TXT)
File
2019-02-01 20:29
jsonTest_missing (JSON)
File
2019-02-01 20:28
nulltest (CSV)
File
2019-02-01 20:27
finefood.sample (TXT)
File
2019-02-01 15:49
Test SQL Flow
Database
2019-02-01 03:09
omniturelogs_orc (STAGING) - test
Staging DB
2019-01-29 18:18
finefoods (TXT)
File
2019-01-29 17:15
s5k_1 (CSV)
File
2019-01-28 15:58
ENB_List_DATA - DATA (EXCEL)
File
2019-01-25 18:24
더보기
1개 선택
+ 데이터셋 생성하기


METATRONDISCOVERY
전국 아파트 시세 분석
부동산 분석팀
수정 2019-02-07 15:25 | Polaris
+ 새로운 데이터 셋 추가
Imported dataset 1
Wrangled dataset 1
아파트(매매)_실거래가_경기도
아파트(매매)_실거래가_경기도
아파트(매매)_실거래가_경기도
설명을 입력해주세요
룰 편집
데이터 프리뷰
ab 시군구
번지
경기도 가평군 가평읍 읍내리
292-7
경기도 가평군 청평면 청평리
837
경기도 고양덕양구 관산동
178-57
타입
WRANGLED
요약
3,046 row(s)
21 column(s)
사용처
1 dataflow
룰 리스트
create with 아파트(매매)_실거래가_경기도
Convert row 1 to header
set type 11 columns to Double
3개의 룰
이 데이터 삭제


METATRON DISCOVERY
아파트(매매)_실거래가_경기도
test
수정 2019-02-07 14:38 | Polaris
정보
타입 FILE (EXCEL)
파일 아파트(매매)_실거래가_경기도_위경도_20190207xlsx (5).xlsx
시트 아파트(매매)_실거래가_경기도_위경도_20190207xl
URI file:///data/metatron-discovery/dataprep/uploads/c7a33f···
사이즈 777.9 KB
요약 3,047 row(s)
21 column(s)
데이터
ab 시군구 | ## 번지 | ## 본번 | ## 부번 | ab 단지명
경기도 가평군 가평읍 읍내리 292-7 292 7 에덴
경기도 가평군 청평면 청평리 837 837 0 청평삼성쉐르빌
경기도 고양덕양구 관산동 178-57 178 57 새서울
경기도 고양덕양구 관산동 178-57 178 57 새서울
경기도 고양덕양구 관산동 178-57 178 57 새서울
경기도 고양덕양구 관산동 178-57 178 57 새서울
경기도 고양덕양구 도내동 983 983 0 엘에이치원흥도래울마을2단지
경기도 고양덕양구 도내동 983 983 0 엘에이치원흥도래울마을2단지
경기도 고양덕양구 도내동 983 983 0 엘에이치원흥도래울마을2단지
사용처
기존 데이터플로우에 추가
이 데이터셋으로 새로운 데이터플로우 생성
생성 아파트(매매)_실거래가_경기도_0207_1438 수정 2019-02-07 14:38 | polaris
전국 아파트 시세 분석 - 부동산 분석팀 수정 2019-02-07 15:25 | polaris

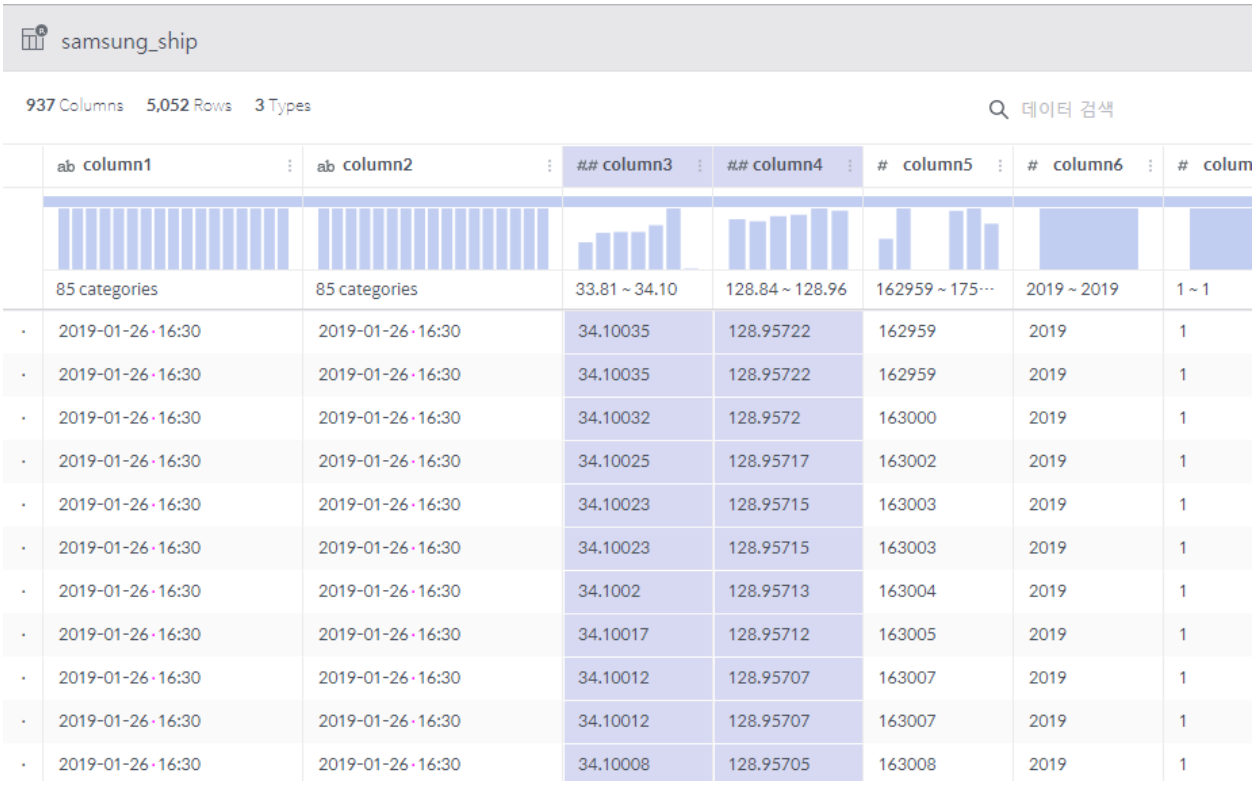

sale
데이터스냅샷
마침
(4) Undo/Redo
26 Columns 9,994 Rows 4 Types
데이터 검색
(1) 컬럼 헤더
(2) 컬럼 헤더 메뉴
주문일 | 카테고리 | 도시 | 국가 | 주문자
2011-01-04 ~ 201··· 3 categories
531 categories
1 category
793 categories
(6) 분포도
Drop
Alter
Edit
Generate
Sort
Move
Clean
Duplicate
Derive
Split
Extract
Count pattern
Merge
Nest
Unnest
Flatten
룰 (8) | 스냅샷 (0)
create with sale (CSV)
set type column1 to Timestamp
set format column1 to yyyy-MM-dd
set type column12 to Timestamp
set type column7, column9 to Long
set type 10 columns to Double
Rename 8 columns
Drop 일련번호, 우편번호
(3) 룰 리스트
(5) 하단 명령 입력창
룰 추가
에디터로 전환
취소
추가
커맨드
split
컬럼 *
주문자
패턴 *
패턴을 입력해 주세요
횟수 *
다음 문자 사이 무시
값을 입력해 주세요
대소문자

1. Column type, name, and menu button
2. Menu for simple rule creation
3. Rule list and insert button (appears when cursor is placed in between rules)
4. Enabled when undo and redo are available
5. Panel to enter rule details
6. Column value distribution, distinct count, type mismatch, null value, etc.

Create a rule

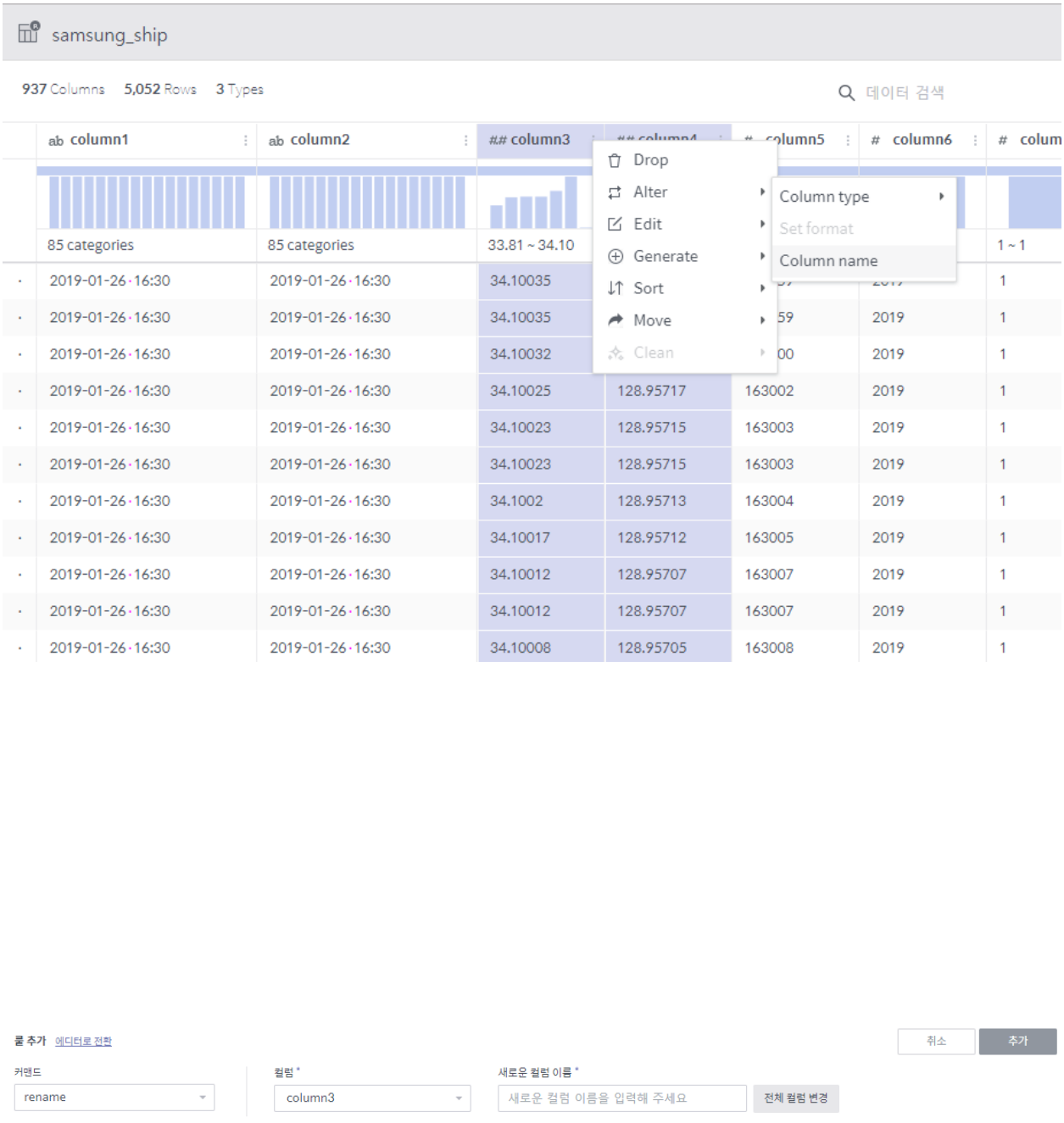
Using the column header menu

1. Select a target column by clicking the column header.
    - Press the function key to select multiple columns.
    - Depending on your OS, click while holding the ^ or key to select/deselect a column (toggle).
    - Click while holding the Shift key to select a range.

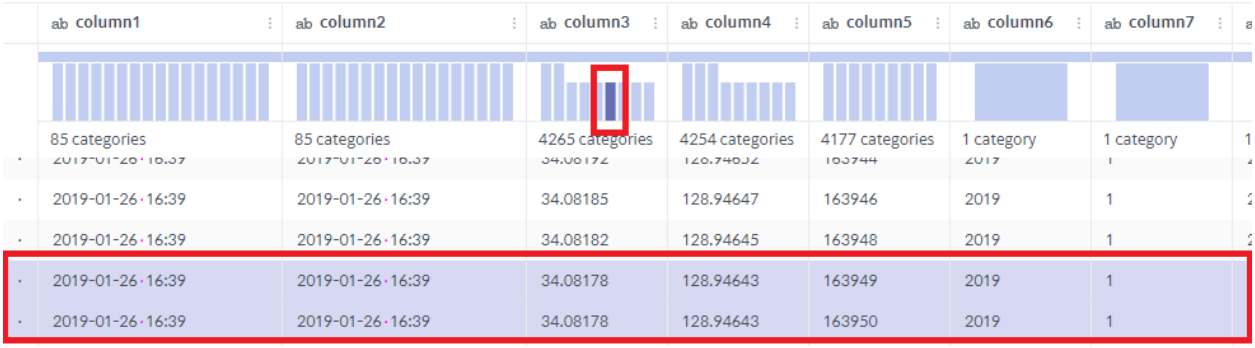


2. Click the ⋮ icon in the header of a selected column to open the header menu, and select a transformation command.
    - Among the commands, **drop** and **settype** are performed upon clicking.
3. To add details, fill out the command input panel below, and click the **Add** button.

4. Some commands can be performed by selecting a distribution bar.

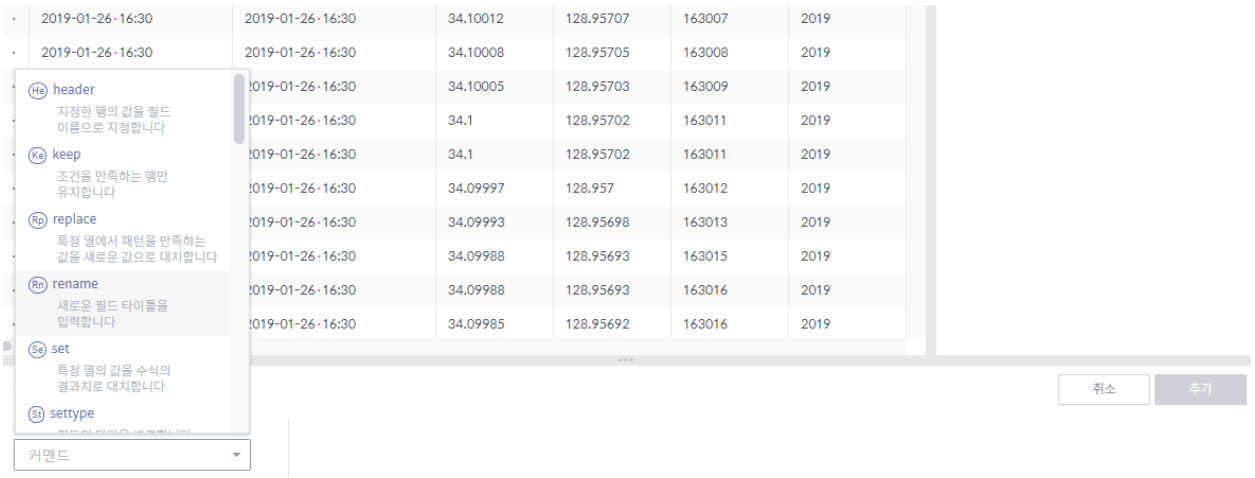

samsung_ship
937 Columns
5,052 Rows
3 Types
데이터 검색
ab column1
ab column2
column3
column4
column5
column6
colum
85 categories
85 categories
33.81 ~ 34.10
1 ~ 1
Drop
Alter
Edit
Generate
Sort
Move
Clean
Column type
Set format
Column name
2019-01-26 16:30
34.10035
34.10032
34.10025
128.95717
163002
34.10023
128.95715
163003
34.1002
128.95713
163004
34.10017
128.95712
163005
34.10012
128.95707
163007
34.10008
128.95705
163008
2019
1
컬 추가
에디터로 전환
취소
추가
커맨드
rename
컬럼 *
column3
새로운 컬럼 이름 *
새로운 컬럼 이름을 입력해 주세요
전체 컬럼 변경

- Click a distribution bar to filter the data based on the selected range (toggle).
- Click the type mismatch or null value graph to set conditions for those values.

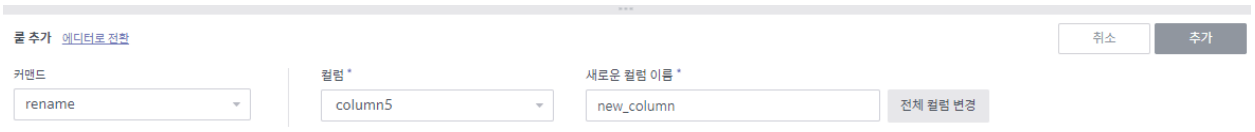


Using the command input panel

1. Select a transformation rule (command) in the command input panel.

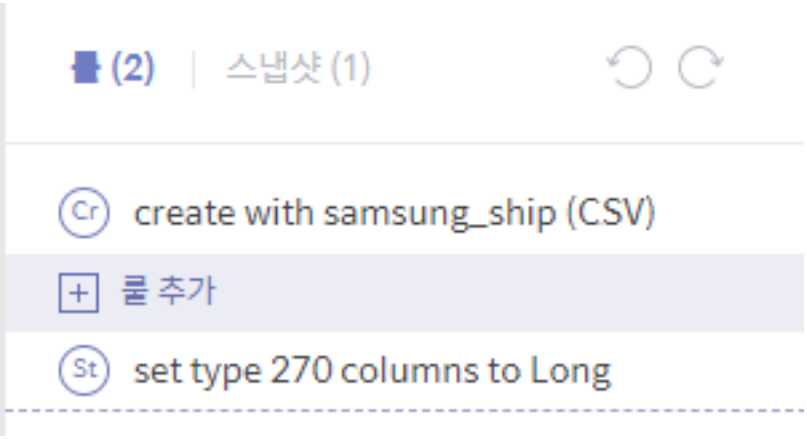


2. Add details as needed, and click the **Add** button.
   - Target columns can be selected using the input panel. You can also designate a column by clicking the column header.

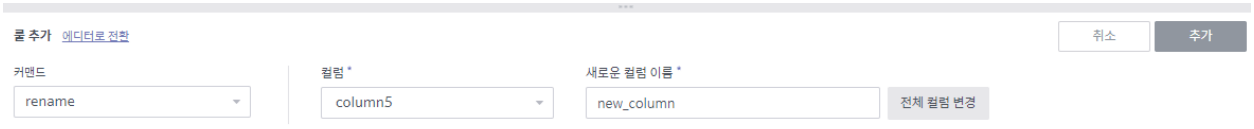


Inserting into a rule list

1. In the list of rules of the right, place the cursor over the boundary where you wish to insert a new rule. The **+ Insert rule** button appears. Press this button.



2. Select a transformation rule (command) in the command input panel. Add details as needed, and click the **Add** button.
    - When a rule is inserted in this manner, all subsequent rules are affected.
    - Rules that cannot be normally executed are displayed in red. In this case, they will revert to the results obtained in the previous step.



Edit a created rule

Editing a rule

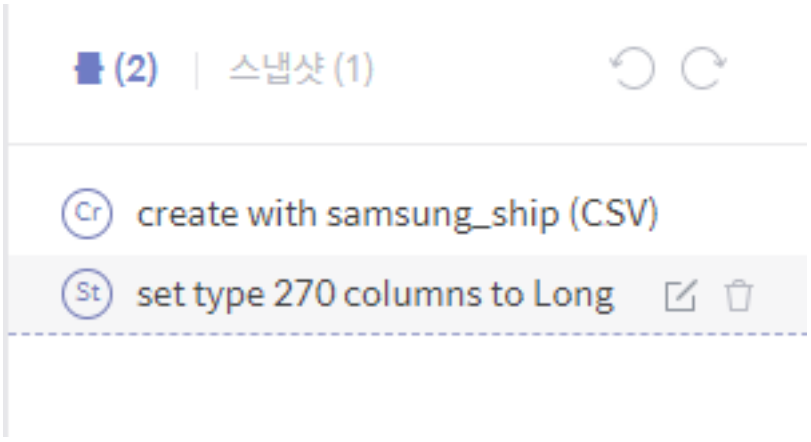
1. In the list of rules on the right, place the cursor over the rule to be edited. The button appears. Press this button.

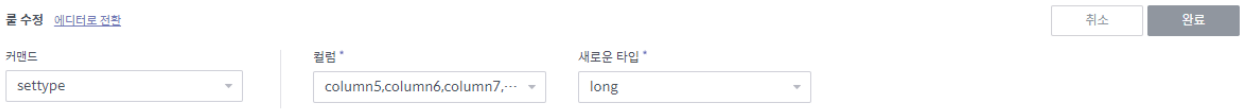
2. Edit the rule in the command input panel and press the **Done** button.
    - When a rule is edited in this manner, all subsequent rules are affected.

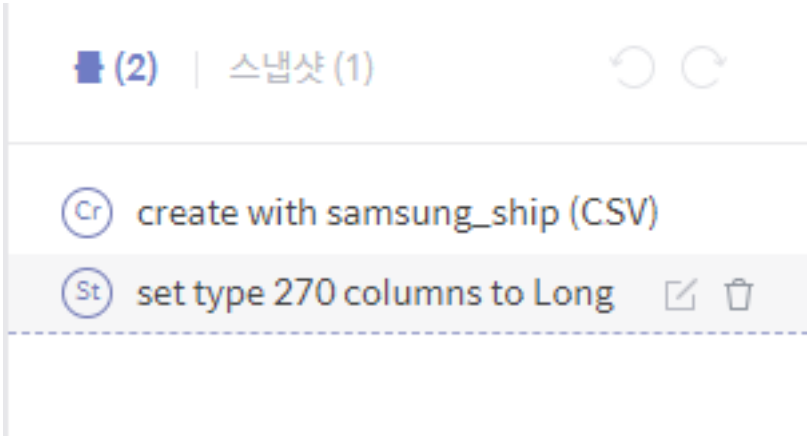
Deleting a rule

In the list of rules on the right, place the cursor over the rule to be deleted. The button appears. Press this button.

- When a rule is deleted in this manner, all subsequent rules are affected.

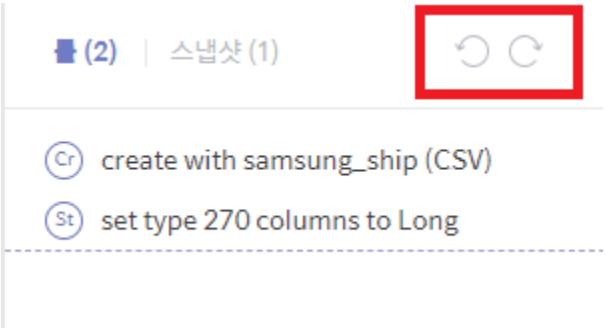

(2)
스냅샷 (1)
Cr
create with samsung_ship (CSV)
St
set type 270 columns to Long


룰 수정
에디터로 전환
취소
완료
커맨드
settype
컬럼 *
column5,column6,column7,···
새로운 타입 *
long


(2)
스냅샷 (1)
Cr
create with samsung_ship (CSV)
St
set type 270 columns to Long

Undo and redo

On the upper right of the rule list are icons to perform **undo** and **redo**.



To revert to a state before executing a command, press the button.

- The dataset reverts to the state before the last transformation (including rule creation, modification, and deletion).
- All rules that were affected also revert to their previous states.

To perform the same command again, press the button.

- Pressing is faster than following the steps to perform the same command again. It is because the transformation results are stored in memory.

8.2.3 Rule types

This section describes each rule in terms of the following.

- Name of rule
- Required arguments
- Optional arguments
- Description
- Notes

The types of rules supported in data preparation are as follows:

- *drop*
- *header*
- *settype*
- *setformat*
- *rename*
- *keep*
- *delete*
- *replace*
- *set*

- *derive*
- *split*
- *merge*
- *extract*
- *countpattern*
- *nest*
- *unnest*
- *flatten*
- *aggregate*
- *pivot*
- *unpivot*
- *join*
- *union*
- *window*

In addition to these rules, data preparation provides various expressions, thereby supporting almost every function required for general data preprocessing.

drop

Required arguments

- Column: A list of target columns

Description

- Deletes the selected columns.

header

Required arguments: Row number that contains the column name (1-base)

Description

- This rule sets the content in the designated row as the column name.
- This is useful for reading a CSV file with column names in the first row.
- Unless otherwise specified, data preparation automatically performs header. This rule may be deleted if header results are not desired, but such cases are not common.

settype

Required arguments

- Column: A list of target columns
- New type: Select one out of Long, Double, String, Boolean, and Timestamp

Optional arguments

- Set format: A format string (Joda Time) in the case of timestamp

Description

- This rule changes the type of the selected columns.
- The rule is considered successful even if the result is a type mismatch, which should be separately addressed.

setformat

Required arguments

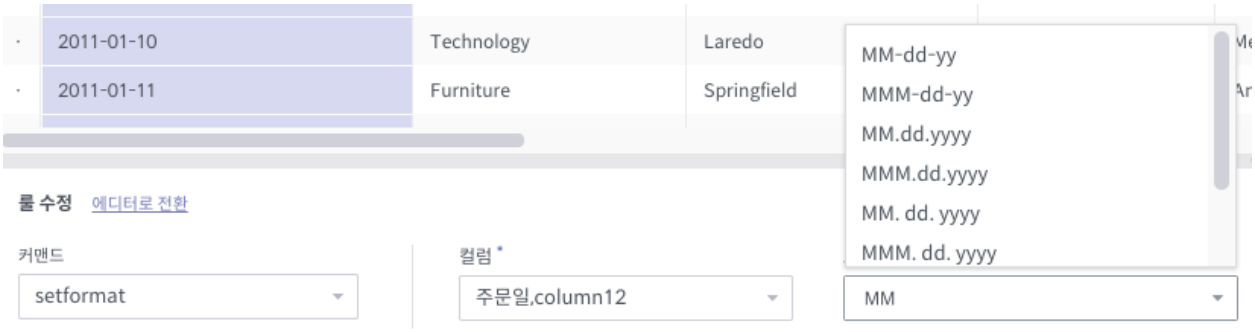
- Column: A list of target columns
- Set format: A Joda-Time format string

Description

- This rule changes the display format of a Timestamp column.
- The target column must be of the Timestamp type.

Notes

- As shown below, the format input field lists different entries depending on the input. The candidate list is narrowed as more values are entered.



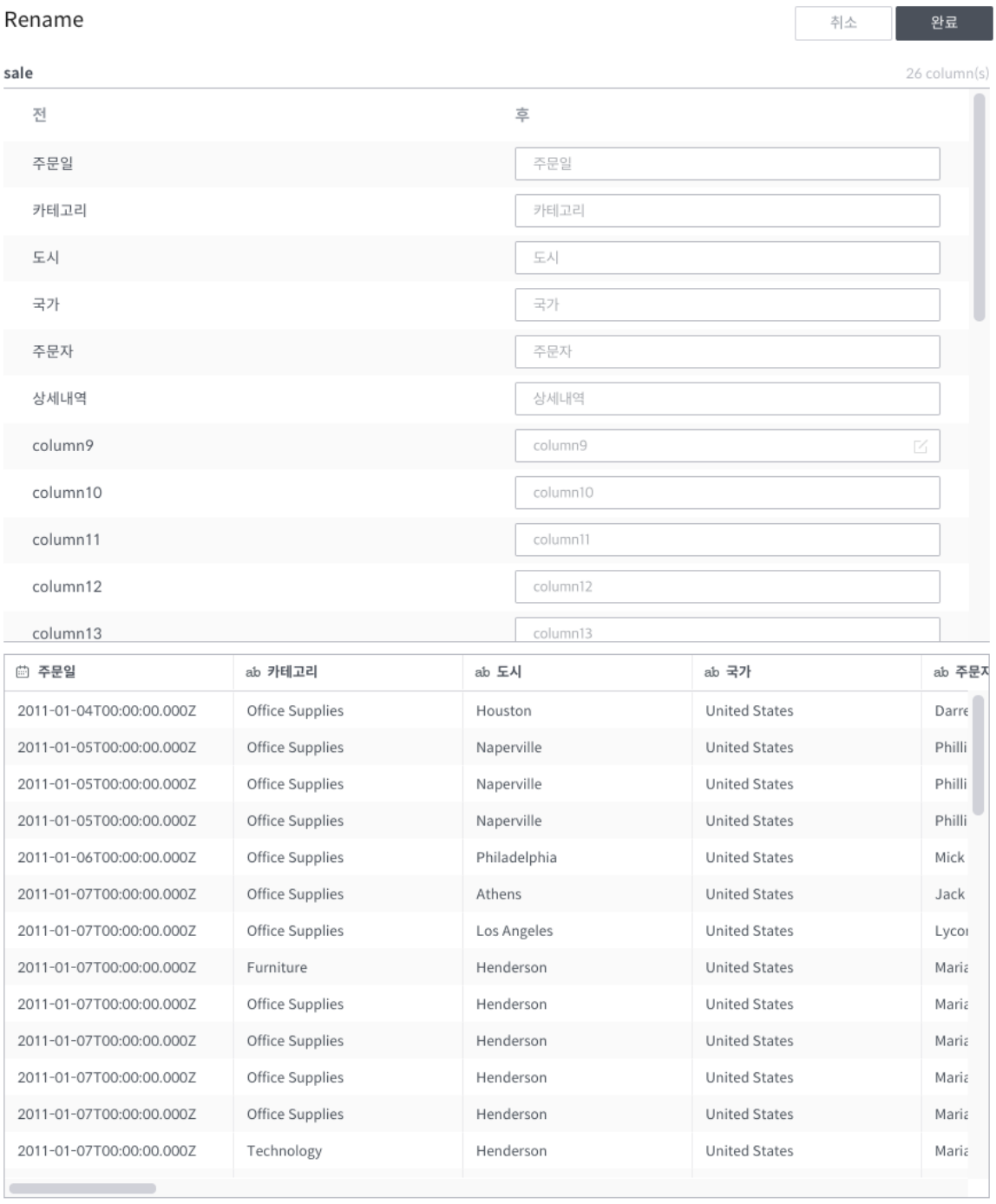
rename

Required arguments

- Column: A single target column
- New column name: New name

Description

- This rule changes the name of the selected column.
- To rename two or more columns at once, click the **Rename multiple columns** button at the bottom of the command input panel to display the following popup.


Rename
취소
완료
sale
26 column(s)
전
후
주문일
주문일
카테고리
카테고리
도시
도시
국가
국가
주문자
주문자
상세내역
상세내역
column9
column9
column10
column10
column11
column11
column12
column12
column13
column13
주문일
ab 카테고리
ab 도시
ab 국가
ab 주문자
2011-01-04T00:00:00.000Z Office Supplies Houston United States Darre
2011-01-05T00:00:00.000Z Office Supplies Naperville United States Philli
2011-01-05T00:00:00.000Z Office Supplies Naperville United States Philli
2011-01-05T00:00:00.000Z Office Supplies Naperville United States Philli
2011-01-06T00:00:00.000Z Office Supplies Philadelphia United States Mick
2011-01-07T00:00:00.000Z Office Supplies Athens United States Jack
2011-01-07T00:00:00.000Z Office Supplies Los Angeles United States Lycoı
2011-01-07T00:00:00.000Z Furniture Henderson United States Maria
2011-01-07T00:00:00.000Z Office Supplies Henderson United States Maria
2011-01-07T00:00:00.000Z Office Supplies Henderson United States Maria
2011-01-07T00:00:00.000Z Office Supplies Henderson United States Maria
2011-01-07T00:00:00.000Z Office Supplies Henderson United States Maria
2011-01-07T00:00:00.000Z Technology Henderson United States Maria

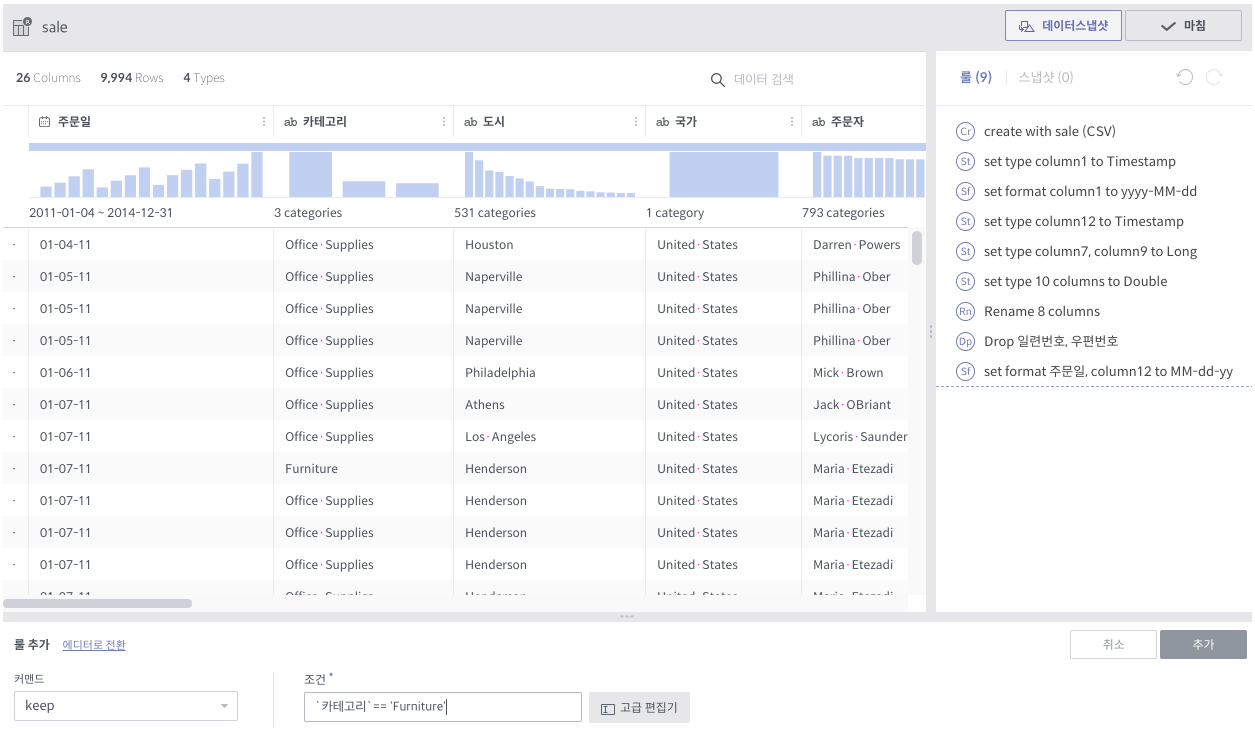
keep

Required arguments

- Condition: A conditional expression returning a Boolean value

Description

- All rows are deleted except the rows that return true for the conditional expression.



delete

Required arguments

- Condition: A conditional expression returning a Boolean value

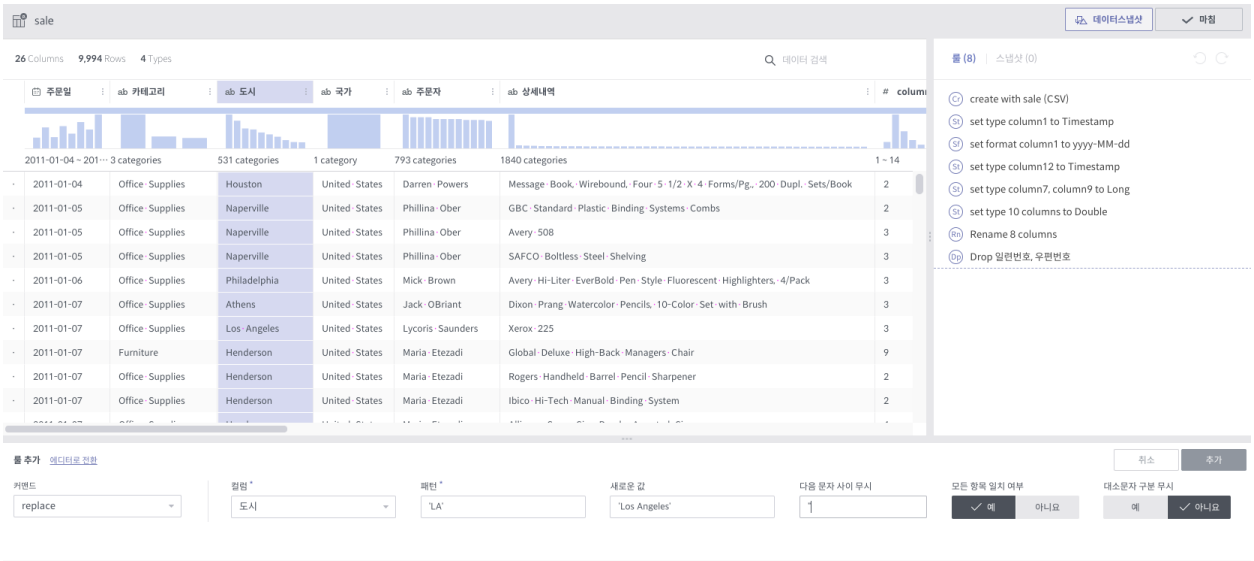
Description

- All rows that return true for the conditional expression are deleted. This is the opposite of *keep*.

replace

Required arguments

- Column: A list of target columns
- Pattern: A string pattern to be replaced



  - In the case of a constant string: Characters enclosed inside `'` (`'Houston'`, `'Naperville'`, `'Philadelphia'` etc.)
  - In the case of a regular expression: Characters enclosed inside `/` (`/[ ,_]+/`, `/\s+$/`, etc.)
- New value: A new string expression to replace the specified pattern
  - Constant string
  - Regular expression `$1_$2_$3`, etc.

Optional arguments

- Ignore between characters: Does not make any replacement for content between the characters entered here
- Match all occurrences: Whether all characters of a word must match
- Ignore case: Whether to make the strings case-insensitive

Description

- String replacement is performed for the selected columns.

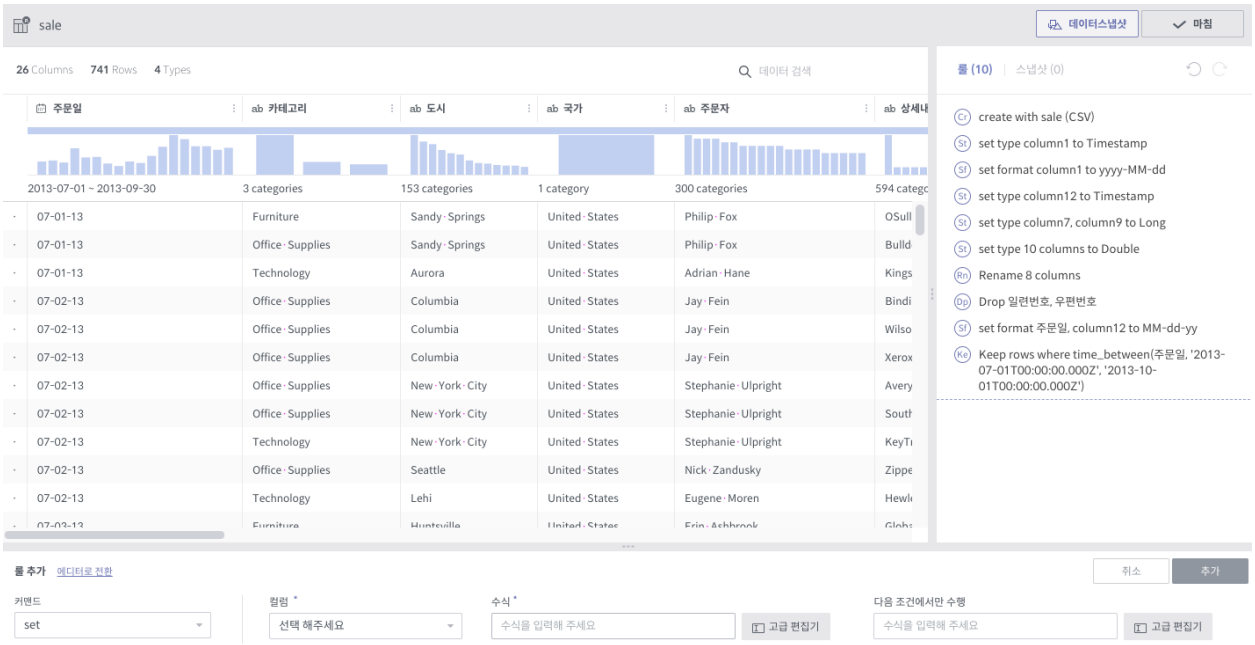
Notes

- Do not use `'` or `/` in a **new value**.
- Values from other columns are not available as **new values**. replace performs string replacement for content in the selected columns only. (cf. *set* rule)

set

Required arguments

- Column: A list of target columns
- Expression: An expression to be applied to the values of the target column. Values from other columns may be referenced. (cf. *replace* rule)
  - When multiple columns are involved, use a `$col` variable, which will be substituted by the respective target column during each conversion.



  - That is, when applying the set command on `column1` and `column2`, `$col` becomes `column1` during conversion of `column1`, and `$col` becomes `column2` during conversion of `column2`.

Optional arguments

- Use only under the following conditions
  - The set rule is applied only to rows satisfying this condition.
  - This rule may be regarded the same as the `WHERE` statement in SQL.

Description

- This rule replaces the values in the selected column with results returned by the expression.
- When using a complex expression, click the **Advanced editor** to display the popup shown below:

In the **Advanced editor**, you can edit the expression in a larger window while viewing the column list and a list of functions and their descriptions, and also run a validity check before implementing the expression.

derive

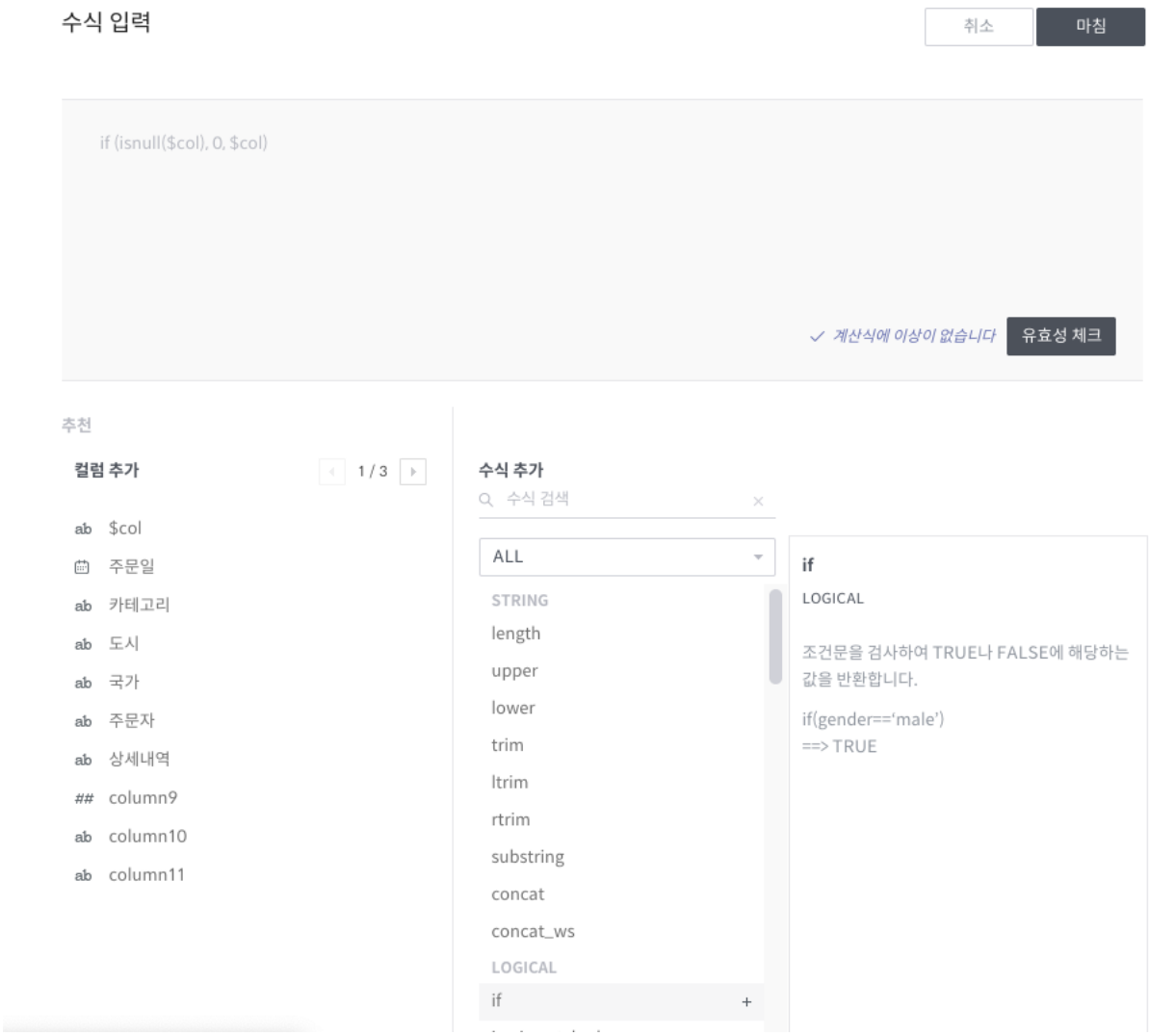
Required arguments

- Expression: An expression whose resulting values are to form a new column. Similar to the *set* rule, values from other columns may be referenced.
- New column name

Description

- While similar to the *set* rule, this rule creates a new column instead of replacing an existing one.

Notes


수식 입력
취소
마침
if (isnull($col), 0, $col)
계산식에 이상이 없습니다
유효성 체크
추천
컬럼 추가
1 / 3
ab $col
주문일
ab 카테고리
ab 도시
ab 국가
ab 주문자
ab 상세내역
column9
ab column10
ab column11
수식 추가
수식 검색
ALL
STRING
length
upper
lower
trim
ltrim
rtrim
substring
concat
concat_ws
LOGICAL
if
if
LOGICAL
조건문을 검사하여 TRUE나 FALSE에 해당하는 값을 반환합니다.
if(gender=='male')
==> TRUE

- The new column is inserted after the last existing column in the expression.

split

Required arguments

- Column: A list of target columns
- Pattern: A string expression that serves as a separator that splits the target strings. Allows a regular expression as is the case for the *replace* rule.
- Number: Number of columns to be divided into.

Description

- Each row is split by the given **Number** - 1.
- When the pattern is no longer matched, the rest columns contain a *null*.

Notes

- Note that columns are created as many as the **Number** input.

merge

Required arguments

- Column: A list of target columns
- Delimiter: A constant string with which values of different columns are concatenated.
- New column name

Description

- The target columns are merged with the **Delimiter** into a new column.

Notes

- Similar to the *replace* rule, enclosing with a `'` may be skipped. That is, strings not enclosed by `/` or `'` are automatically enclosed by `'`.

extract

Required arguments

- Column: A list of target columns
- Pattern: A string pattern to be extracted. Allows a regular expression as is the case for the *replace* rule.
- Number: Number of instances to be extracted

Optional arguments

- Ignore between characters: Does not make any replacement for content between the characters entered here
- Ignore case: Whether to make the strings case-insensitive

Description

- A new column(s) with content matching the given pattern is created.

Notes

- When there are multiple target columns, the resulting columns are inserted after each target column.

countpattern

Required arguments

- Column: A list of target columns
- Pattern: A string pattern to be detected. Allows a regular expression as is the case for the *replace* rule.

Optional arguments

- Ignore between characters: Does not make any replacement for content between the characters entered here
- Ignore case: Whether to make the strings case-insensitive

Description

- New columns are created based on the number of matches with the pattern.
- This is highly similar to *extract*. The only difference is that it counts the number of matches, rather than extracting the matched content.

Notes

- When there are multiple target columns, the resulting columns are inserted after each target column.

nest

Required arguments

- Column: A list of target columns
- Type: Map or Array
- New column name

Description

- The target columns are grouped into a new column of the given type.
- Below are examples of grouping columns into an array and map, respectively.

unnest

Required arguments

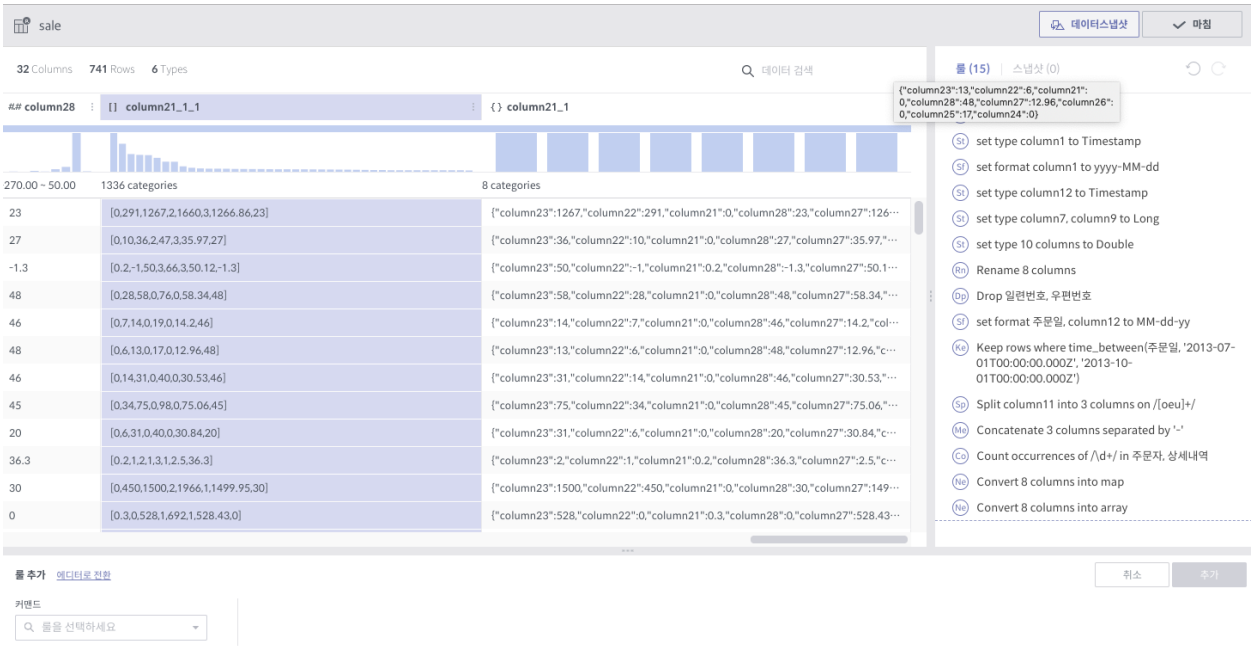
- Column: A single target column
- Select elements: *0-base* index for an array, or key value for a map

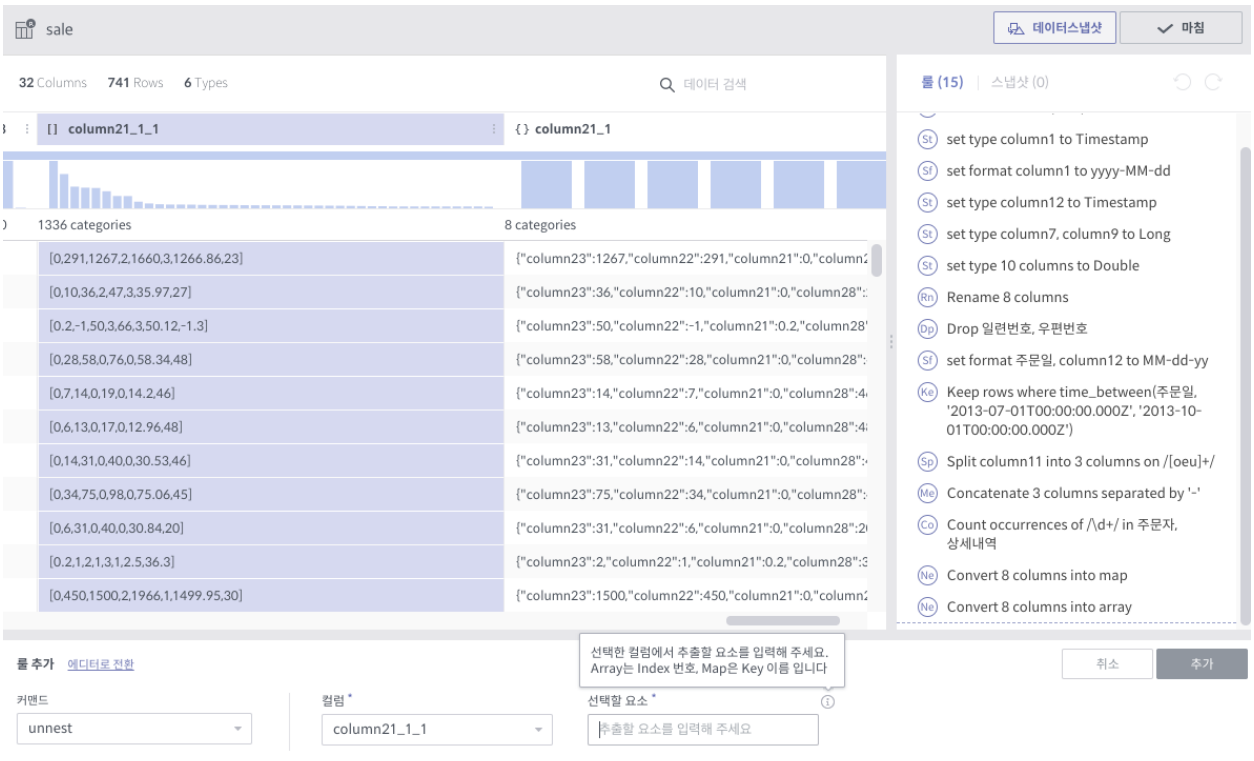
Description

- A new column is created by extracting the selected elements from an array or a map.

Notes

- The target column must be of the array or map type.


sale
데이터스냅샷
마침
32 Columns
741 Rows
6 Types
데이터 검색
column28
[] column21_1_1
{} column21_1
270.00 ~ 50.00
1336 categories
8 categories
23
27
-1.3
48
46
48
46
45
20
36.3
30
0
룰 (15)
스냅샷 (0)
set type column1 to Timestamp
set format column1 to yyyy-MM-dd
set type column12 to Timestamp
set type column7, column9 to Long
set type 10 columns to Double
Rename 8 columns
Drop 일련번호, 우편번호
set format 주문일, column12 to MM-dd-yy
Keep rows where time_between(주문일, '2013-07-01T00:00:00.000Z', '2013-10-01T00:00:00.000Z')
Split column11 into 3 columns on /[oeu]+/
Concatenate 3 columns separated by '-'
Count occurrences of /\d+/ in 주문자, 상세내역
Convert 8 columns into map
Convert 8 columns into array
룰 추가
에디터로 전환
취소
추가
커맨드
룰을 선택하세요


sale
데이터스냅샷
마침
32 Columns
741 Rows
6 Types
데이터 검색
[] column21_1_1
{} column21_1
1336 categories
8 categories
룰 (15)
스냅샷 (0)
set type column1 to Timestamp
set format column1 to yyyy-MM-dd
set type column12 to Timestamp
set type column7, column9 to Long
set type 10 columns to Double
Rename 8 columns
Drop 일련번호, 우편번호
set format 주문일, column12 to MM-dd-yy
Keep rows where time_between(주문일, '2013-07-01T00:00:00.000Z', '2013-10-01T00:00:00.000Z')
Split column11 into 3 columns on /[oeu]+/
Concatenate 3 columns separated by '-'
Count occurrences of /\d+/ in 주문자, 상세내역
Convert 8 columns into map
Convert 8 columns into array
룰 추가
에디터로 전환
선택한 컬럼에서 추출할 요소를 입력해 주세요.
Array는 Index 번호, Map은 Key 이름 입니다
취소
추가
커맨드
unnest
컬럼 *
column21_1_1
선택할 요소 *
추출할 요소를 입력해 주세요

flatten

Required arguments

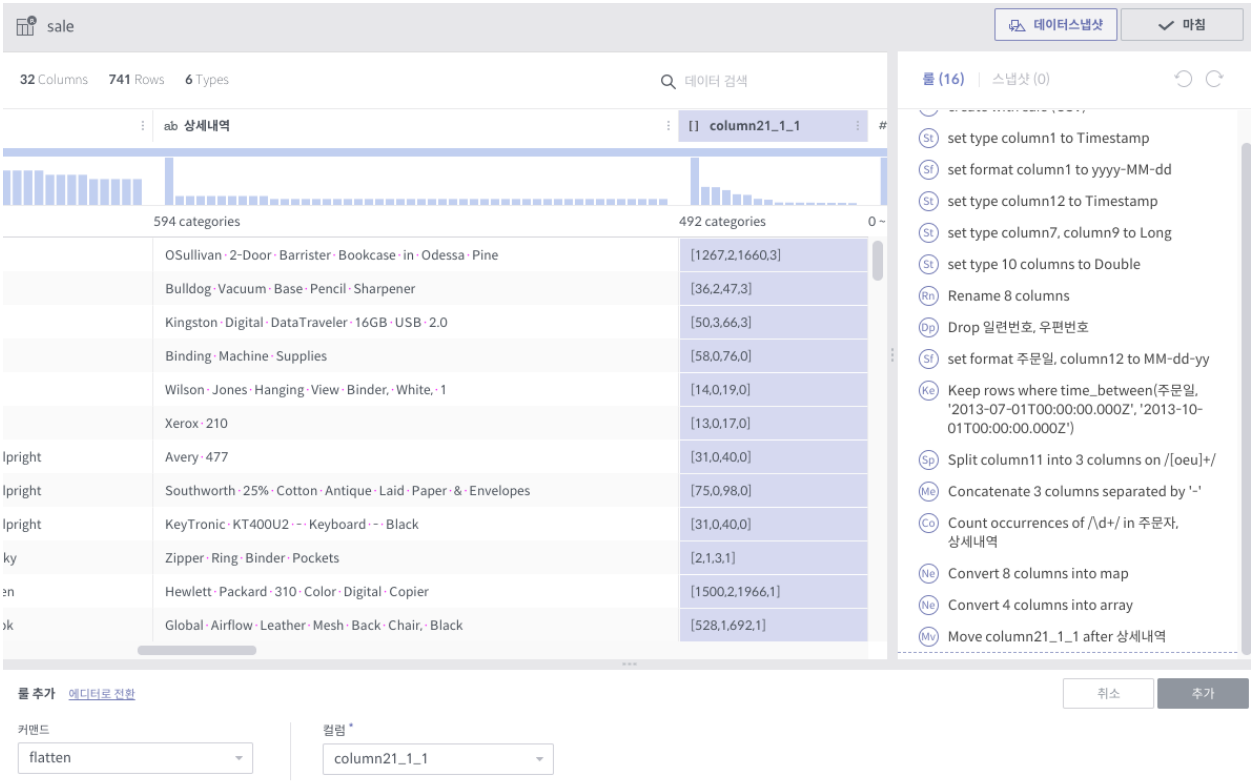
- Column: A single target column

Description

- Rows are created from elements of an array.

Notes

- The target column must be of the array type.



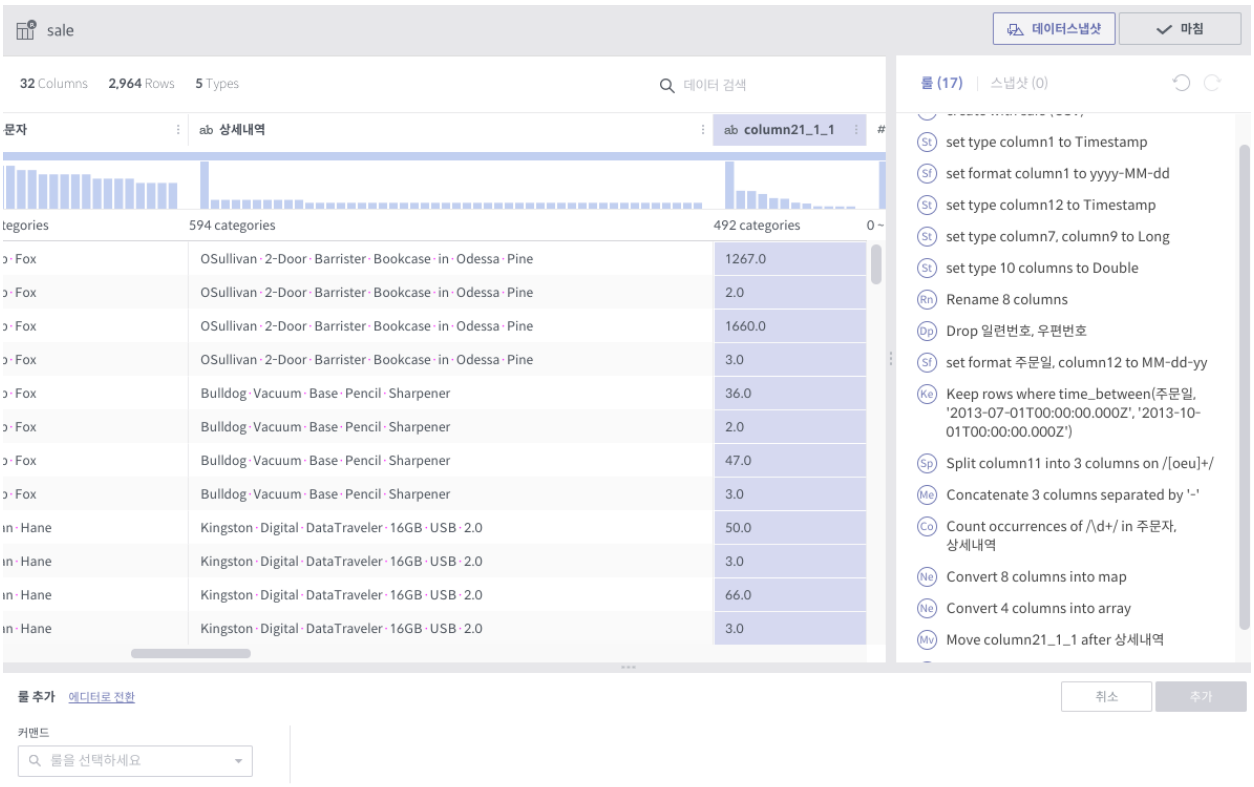
If the target array column has four elements as shown in the above example, each original row of the array results in four rows. Non-array columns result in the same columns.

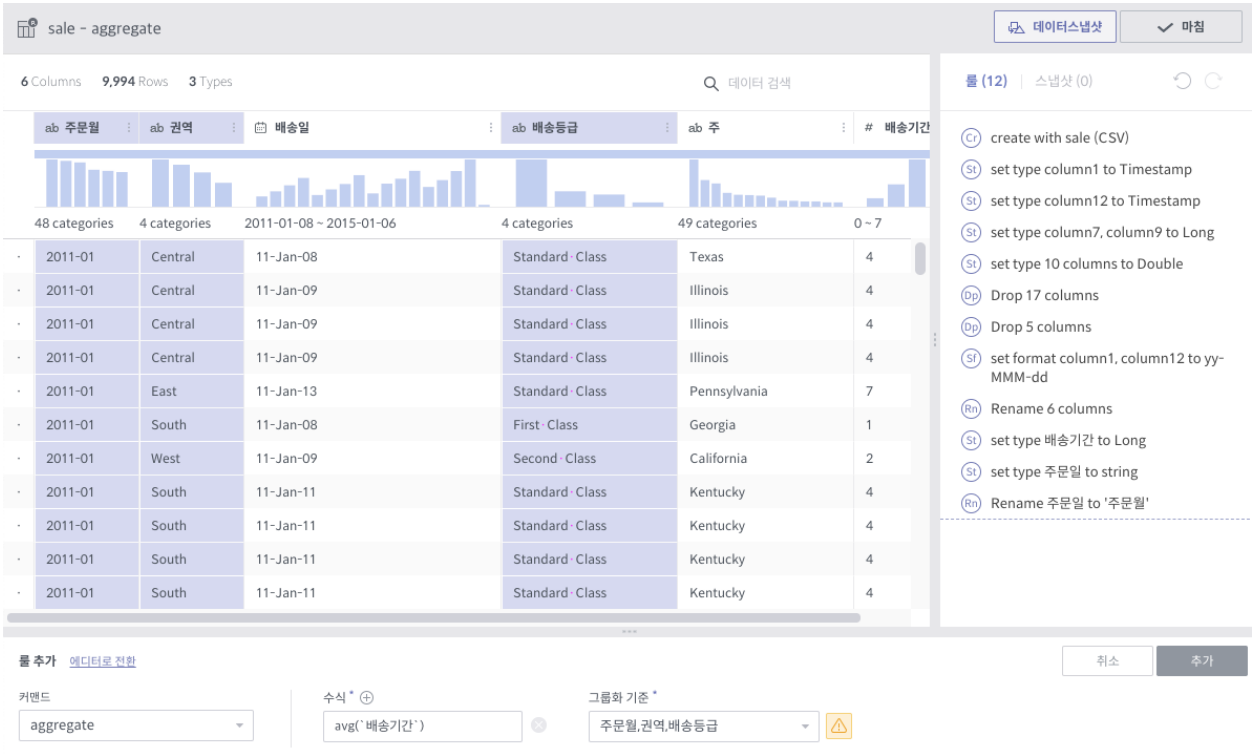
aggregate

Required arguments

- Expression: A list of aggregate functions
- Group by: A list of columns that group values by.

Description

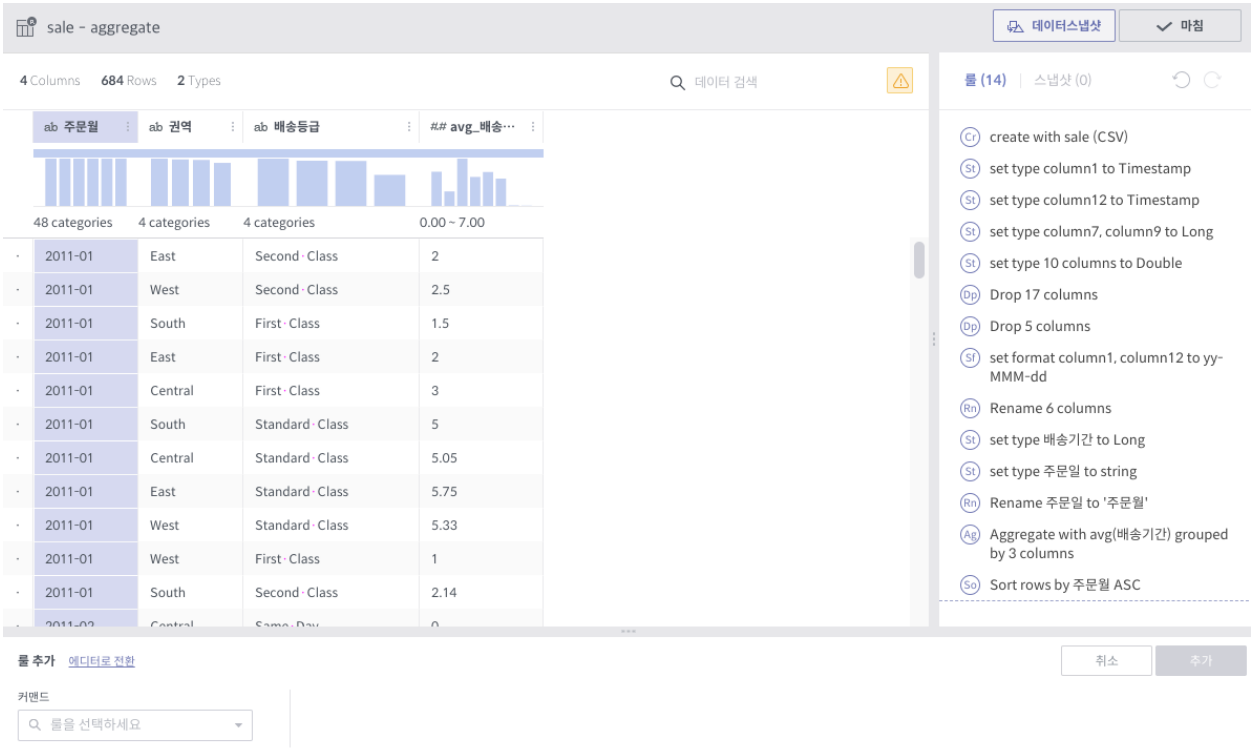

sale
데이터스냅샷
마침
32 Columns 2,964 Rows 5 Types
데이터 검색
룰 (17) | 스냅샷 (0)
문자
ab 상세내역
ab column21_1_1
594 categories
492 categories
OSullivan · 2-Door · Barrister · Bookcase · in · Odessa · Pine
Bulldog · Vacuum · Base · Pencil · Sharpener
Kingston · Digital · DataTraveler · 16GB · USB · 2.0
1267.0
2.0
1660.0
3.0
36.0
2.0
47.0
3.0
50.0
3.0
66.0
3.0
set type column1 to Timestamp
set format column1 to yyyy-MM-dd
set type column12 to Timestamp
set type column7, column9 to Long
set type 10 columns to Double
Rename 8 columns
Drop 일련번호, 우편번호
set format 주문일, column12 to MM-dd-yy
Keep rows where time_between(주문일, '2013-07-01T00:00:00.000Z', '2013-10-01T00:00:00.000Z')
Split column11 into 3 columns on /[oeu]+/
Concatenate 3 columns separated by '-'
Count occurrences of /\d+/ in 주문자, 상세내역
Convert 8 columns into map
Convert 4 columns into array
Move column21_1_1 after 상세내역
룰 추가
에디터로 전환
취소
추가
커맨드
룰을 선택하세요


sale - aggregate
데이터스냅샷
마침
6 Columns 9,994 Rows 3 Types
데이터 검색
룰 (12) | 스냅샷 (0)
ab 주문월
ab 권역
배송일
ab 배송등급
ab 주
배송기간
48 categories
4 categories
2011-01-08 ~ 2015-01-06
4 categories
49 categories
0 ~ 7
2011-01 Central 11-Jan-08 Standard · Class Texas 4
2011-01 Central 11-Jan-09 Standard · Class Illinois 4
2011-01 Central 11-Jan-09 Standard · Class Illinois 4
2011-01 Central 11-Jan-09 Standard · Class Illinois 4
2011-01 East 11-Jan-13 Standard · Class Pennsylvania 7
2011-01 South 11-Jan-08 First · Class Georgia 1
2011-01 West 11-Jan-09 Second · Class California 2
2011-01 South 11-Jan-11 Standard · Class Kentucky 4
2011-01 South 11-Jan-11 Standard · Class Kentucky 4
2011-01 South 11-Jan-11 Standard · Class Kentucky 4
2011-01 South 11-Jan-11 Standard · Class Kentucky 4
create with sale (CSV)
set type column1 to Timestamp
set type column12 to Timestamp
set type column7, column9 to Long
set type 10 columns to Double
Drop 17 columns
Drop 5 columns
set format column1, column12 to yy-MMM-dd
Rename 6 columns
set type 배송기간 to Long
set type 주문일 to string
Rename 주문일 to '주문월'
룰 추가
에디터로 전환
취소
추가
커맨드
aggregate
수식 *
avg(`배송기간`)
그룹화 기준 *
주문월,권역,배송등급

- A new column is added from the results of grouping by each combination of the elements from the GroupBy columns.
- A column is created for each expression. For example, two columns are created if average and count are designated as expressions.
- The available aggregate functions are as follows:
  - count()
  - sum(*colname*)
  - avg(*colname*)
  - min(*colname*)
  - max(*colname*)

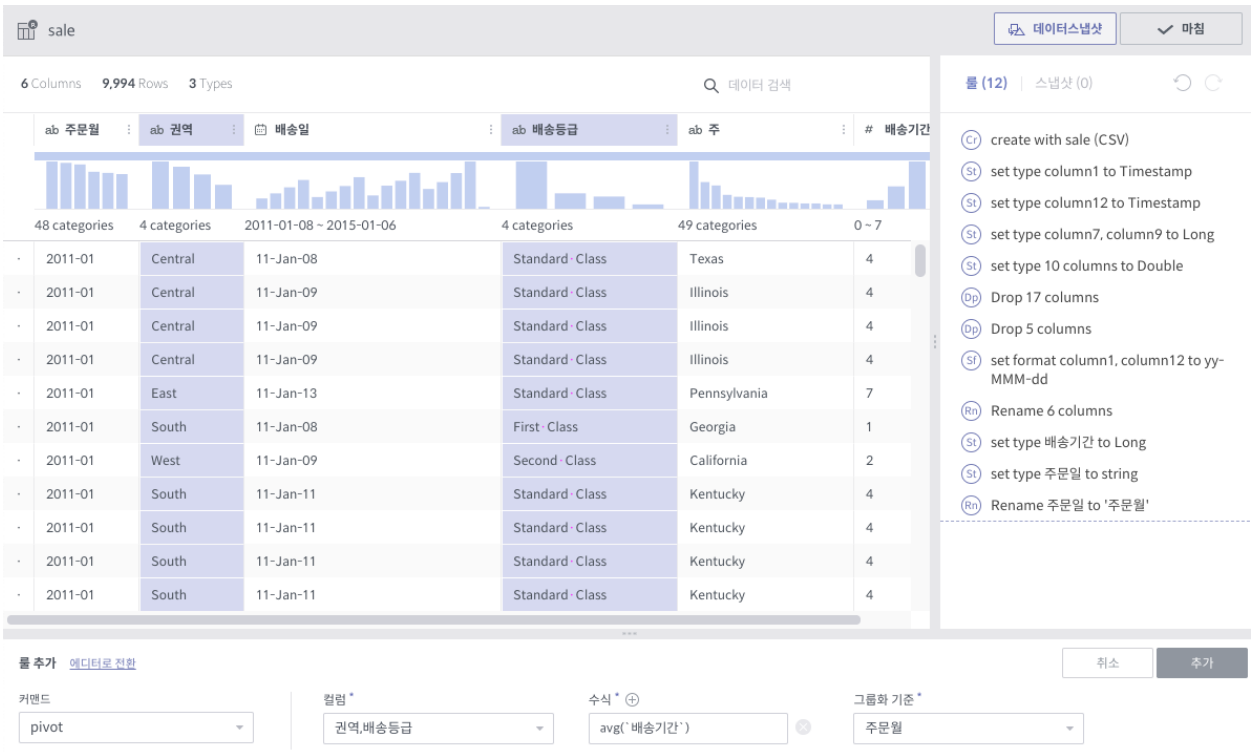
Notes

- Calculations are performed only for sampling results. Therefore, the snapshot?the results for the entire data?may be different.
- Note that `()` must be inserted when using the count function.
- count(*colname*) is currently not available.



pivot

Required arguments



- Column: A list of columns subject to pivoting
- Expression: A list of expressions whose resulting values form new columns (only aggregate functions are available)
- Group by: A list of columns that group values by.

Description

- Group By is performed for each combination of target columns and GroupBy columns. A dataset having the results as column values is created.
- A set of columns is created for each expression. For example, if average and count are designated as expressions and the values in the pivoted columns are divided into ten groups, a total of 20 columns will be created.

Notes

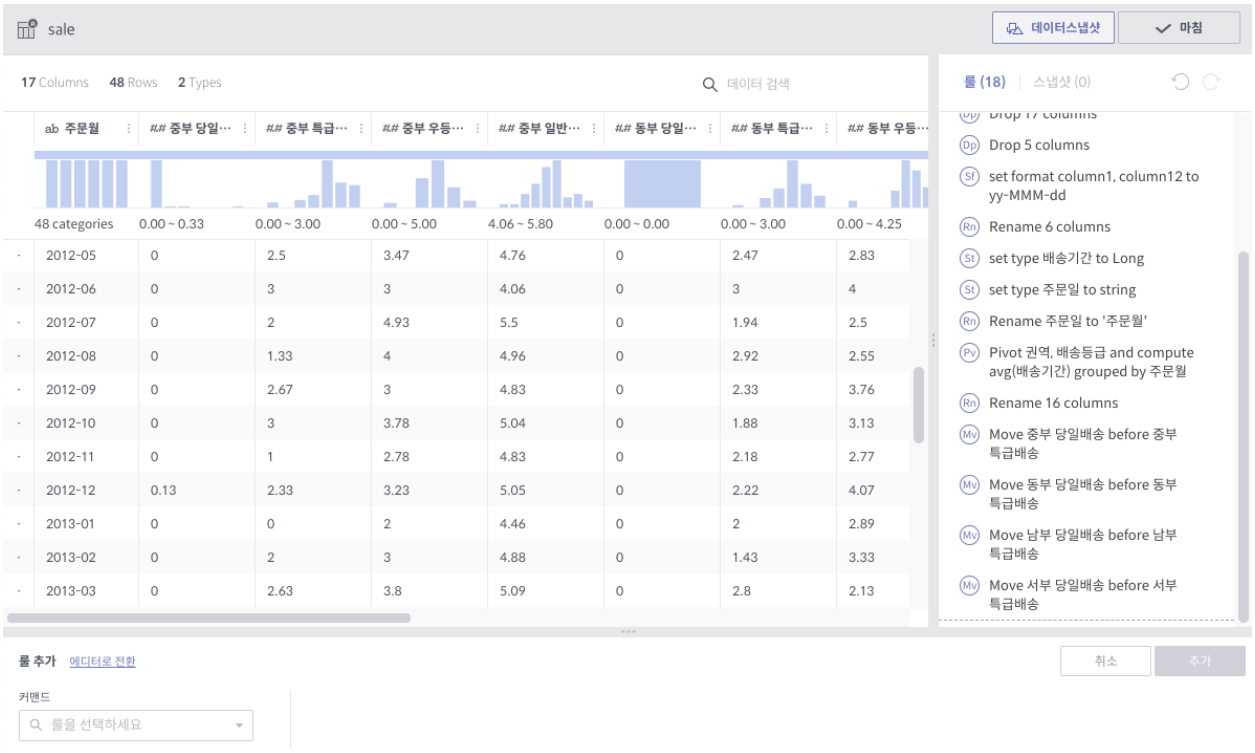
- This is used when performing GroupBy on at least two columns. (1 pivoted column, 1 GroupBy column)
- Here, **Rename multiple columns** is useful as column names tend to get longer.

unpivot

Required arguments

- Column: A list of target columns to be converted into values in new columns
- GroupEvery: Number of columns (defaults to 1)

Description


sale
데이터스냅샷
마침
17 Columns
48 Rows
2 Types
데이터 검색
룰 (18)
스냅샷 (0)
ab 주문월
#.# 중부 당일…
#.# 중부 특급…
#.# 중부 우등…
#.# 중부 일반…
#.# 동부 당일…
#.# 동부 특급…
#.# 동부 우등…
48 categories
0.00 ~ 0.33
0.00 ~ 3.00
0.00 ~ 5.00
4.06 ~ 5.80
0.00 ~ 0.00
0.00 ~ 3.00
0.00 ~ 4.25
2012-05 0 2.5 3.47 4.76 0 2.47 2.83
2012-06 0 3 3 4.06 0 3 4
2012-07 0 2 4.93 5.5 0 1.94 2.5
2012-08 0 1.33 4 4.96 0 2.92 2.55
2012-09 0 2.67 3 4.83 0 2.33 3.76
2012-10 0 3 3.78 5.04 0 1.88 3.13
2012-11 0 1 2.78 4.83 0 2.18 2.77
2012-12 0.13 2.33 3.23 5.05 0 2.22 4.07
2013-01 0 0 2 4.46 0 2 2.89
2013-02 0 2 3 4.88 0 1.43 3.33
2013-03 0 2.63 3.8 5.09 0 2.8 2.13
Drop 5 columns
set format column1, column12 to yy-MMM-dd
Rename 6 columns
set type 배송기간 to Long
set type 주문일 to string
Rename 주문일 to '주문월'
Pivot 권역, 배송등급 and compute avg(배송기간) grouped by 주문월
Rename 16 columns
Move 중부 당일배송 before 중부 특급배송
Move 동부 당일배송 before 동부 특급배송
Move 남부 당일배송 before 남부 특급배송
Move 서부 당일배송 before 서부 특급배송
룰 추가
에디터로 전환
취소
추가
커맨드
룰을 선택하세요

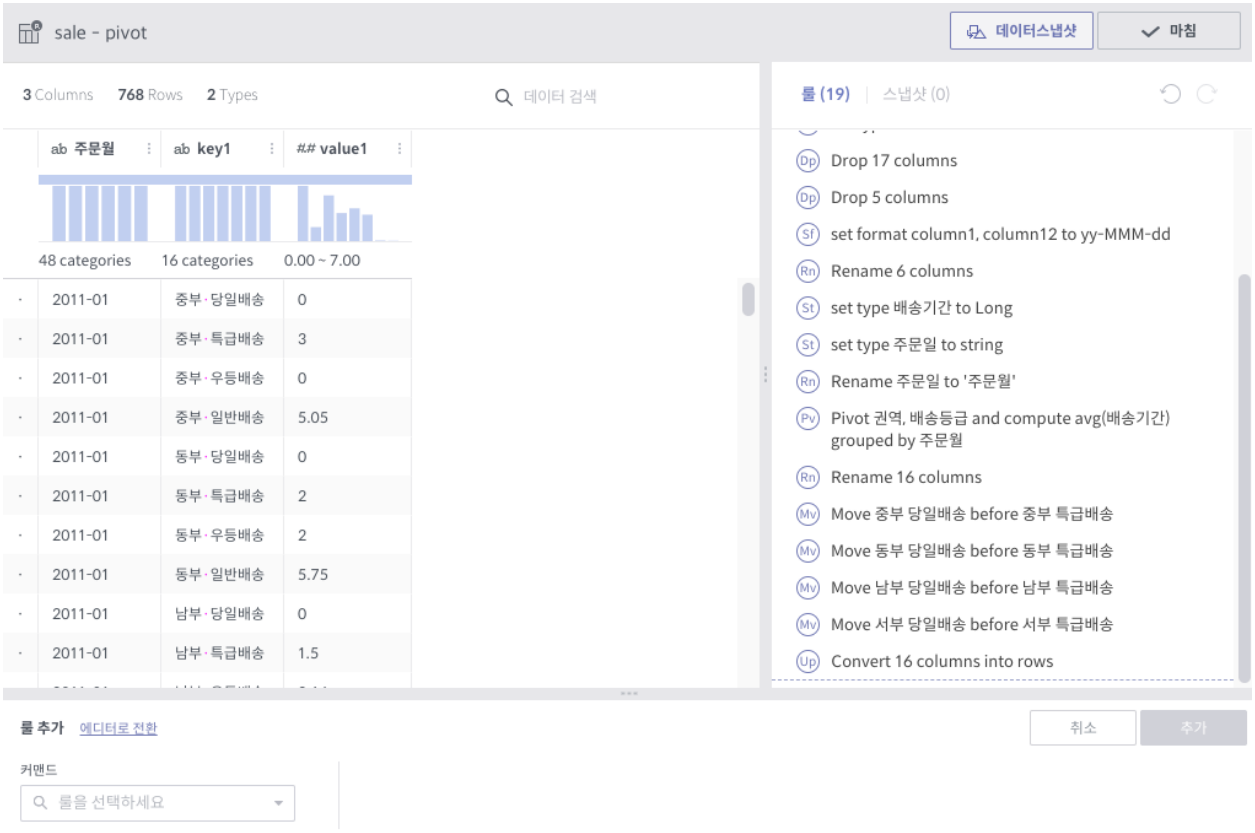

sale - pivot
데이터스냅샷
마침
17 Columns
48 Rows
2 Types
데이터 검색
룰 (18)
스냅샷 (0)
남부 당일…
#.# 남부 특급…
#.# 남부 우등…
#.# 남부 일반…
#.# 서부 당일…
#.# 서부 특급…
#.# 서부 우등…
#.# 서부 일반…
~ 0.33
0.00 ~ 3.00
0.00 ~ 5.00
4.22 ~ 7.00
0.00 ~ 1.00
0.00 ~ 3.00
0.00 ~ 4.33
4.32 ~ 6.22
1.5 2.14 5 0 1 2.5 5.33
3 0 5.2 0 1.67 3.33 4.36
2.13 2.83 4.28 0 1 4.2 4.97
2.33 3 4.65 0 1.5 3.86 5.43
2.2 3.5 5.5 0 3 4 4.84
3 2 4.55 0 3 3.13 5
3 0 5.38 0 2.5 2.33 4.95
1.71 3.25 4.57 0 2 4.14 4.67
2.67 3.33 5 0 2 3.56 4.84
2.33 2 4.93 0.33 2.6 3.69 4.32
3 4 4.97 0 2.26 2.68 5.64
Drop 5 columns
set format column1, column12 to yy-MMM-dd
Rename 6 columns
set type 배송기간 to Long
set type 주문일 to string
Rename 주문일 to '주문월'
Pivot 권역, 배송등급 and compute avg(배송기간) grouped by 주문월
Rename 16 columns
Move 중부 당일배송 before 중부 특급배송
Move 동부 당일배송 before 동부 특급배송
Move 남부 당일배송 before 남부 특급배송
Move 서부 당일배송 before 서부 특급배송
룰 추가
에디터로 전환
취소
추가
커맨드
unpivot
컬럼 *
중부 당일배송,서부 일반배송,중…
그룹 수 *
1

- Two columns are created?one contains the selected column names and the other contains their values. (If GroupEvery is set to 1)
- If GroupEvery is the same as the number of selected columns, each resulting pair of columns contains the name and values of its respective original column. Therefore, If 10 columns are unpivoted with the GroupEvery argument set to 10, for example, a total of 20 columns are created.

Notes

- Using the GroupEvery argument set to a factor of the number of columns will soon be supported.

<Where GroupEvery is set to 1>



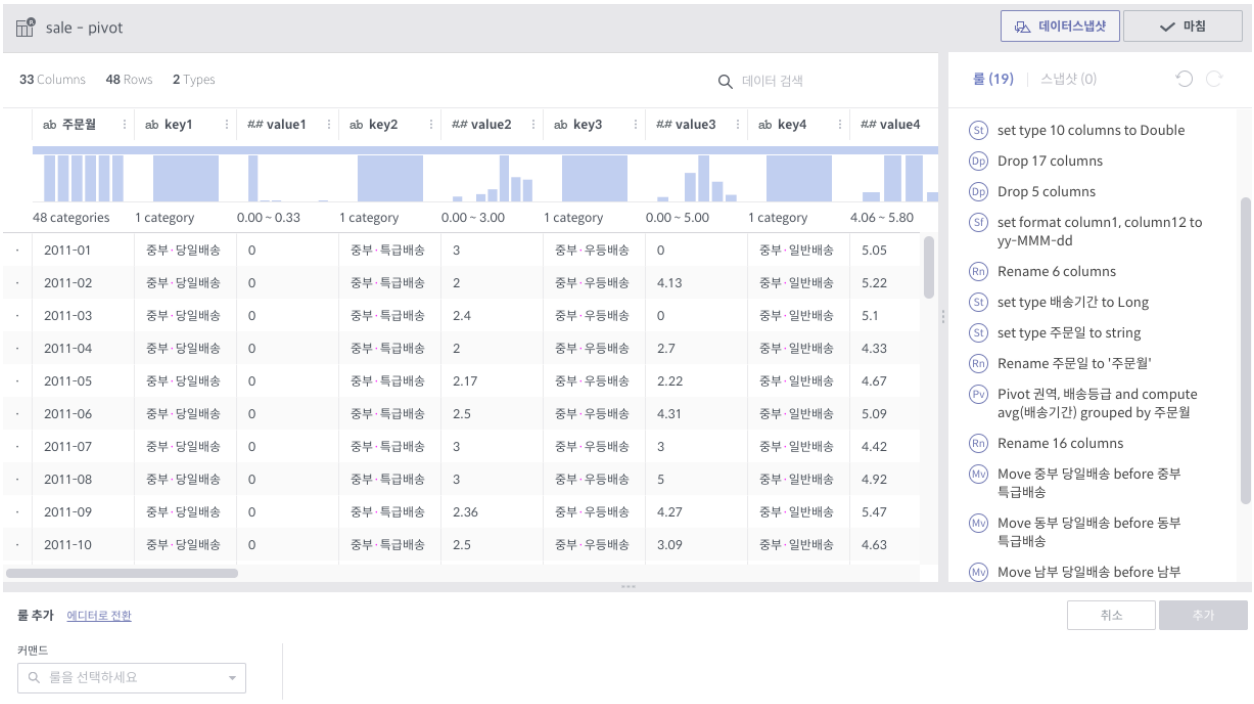
<Where GroupEvery is set to the same as the number of columns>

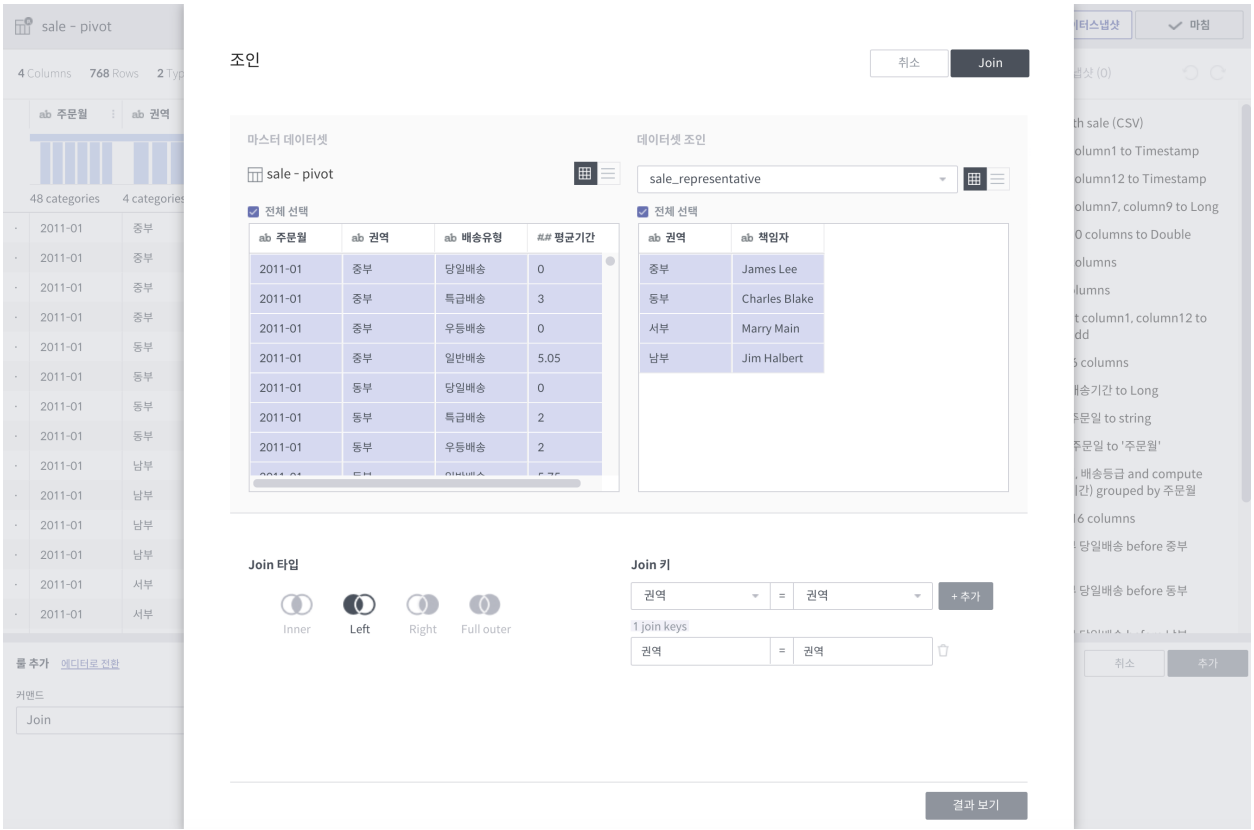
join

Unlike other rules, join has a separate popup.

Required arguments (select in a popup or enter a value)

- Dataset to join: A wrangled dataset in the same dataflow
- Columns to join (toggle)


sale - pivot
데이터스냅샷
마침
33 Columns 48 Rows 2 Types
데이터 검색
룰 (19) | 스냅샷 (0)
ab 주문월 | ab key1 | ## value1 | ab key2 | ## value2 | ab key3 | ## value3 | ab key4 | ## value4
48 categories | 1 category | 0.00 ~ 0.33 | 1 category | 0.00 ~ 3.00 | 1 category | 0.00 ~ 5.00 | 1 category | 4.06 ~ 5.80
2011-01 중부·당일배송 0 중부·특급배송 3 중부·우등배송 0 중부·일반배송 5.05
2011-02 중부·당일배송 0 중부·특급배송 2 중부·우등배송 4.13 중부·일반배송 5.22
2011-03 중부·당일배송 0 중부·특급배송 2.4 중부·우등배송 0 중부·일반배송 5.1
2011-04 중부·당일배송 0 중부·특급배송 2 중부·우등배송 2.7 중부·일반배송 4.33
2011-05 중부·당일배송 0 중부·특급배송 2.17 중부·우등배송 2.22 중부·일반배송 4.67
2011-06 중부·당일배송 0 중부·특급배송 2.5 중부·우등배송 4.31 중부·일반배송 5.09
2011-07 중부·당일배송 0 중부·특급배송 3 중부·우등배송 3 중부·일반배송 4.42
2011-08 중부·당일배송 0 중부·특급배송 3 중부·우등배송 5 중부·일반배송 4.92
2011-09 중부·당일배송 0 중부·특급배송 2.36 중부·우등배송 4.27 중부·일반배송 5.47
2011-10 중부·당일배송 0 중부·특급배송 2.5 중부·우등배송 3.09 중부·일반배송 4.63
set type 10 columns to Double
Drop 17 columns
Drop 5 columns
set format column1, column12 to yy-MMM-dd
Rename 6 columns
set type 배송기간 to Long
set type 주문일 to string
Rename 주문일 to '주문월'
Pivot 권역, 배송등급 and compute avg(배송기간) grouped by 주문월
Rename 16 columns
Move 중부 당일배송 before 중부 특급배송
Move 동부 당일배송 before 동부 특급배송
Move 남부 당일배송 before 남부
룰 추가 에디터로 전환
취소
추가
커맨드
룰을 선택하세요


sale - pivot
조인
취소
Join
마스터 데이터셋
sale - pivot
데이터셋 조인
sale_representative
전체 선택
ab 주문월 | ab 권역 | ab 배송유형 | ## 평균기간
2011-01 중부 당일배송 0
2011-01 중부 특급배송 3
2011-01 중부 우등배송 0
2011-01 중부 일반배송 5.05
2011-01 동부 당일배송 0
2011-01 동부 특급배송 2
2011-01 동부 우등배송 2
ab 권역 | ab 책임자
중부 James Lee
동부 Charles Blake
서부 Marry Main
남부 Jim Halbert
Join 타입
Inner
Left
Right
Full outer
Join 키
권역 = 권역
+ 추가
1 join keys
권역 = 권역
결과 보기

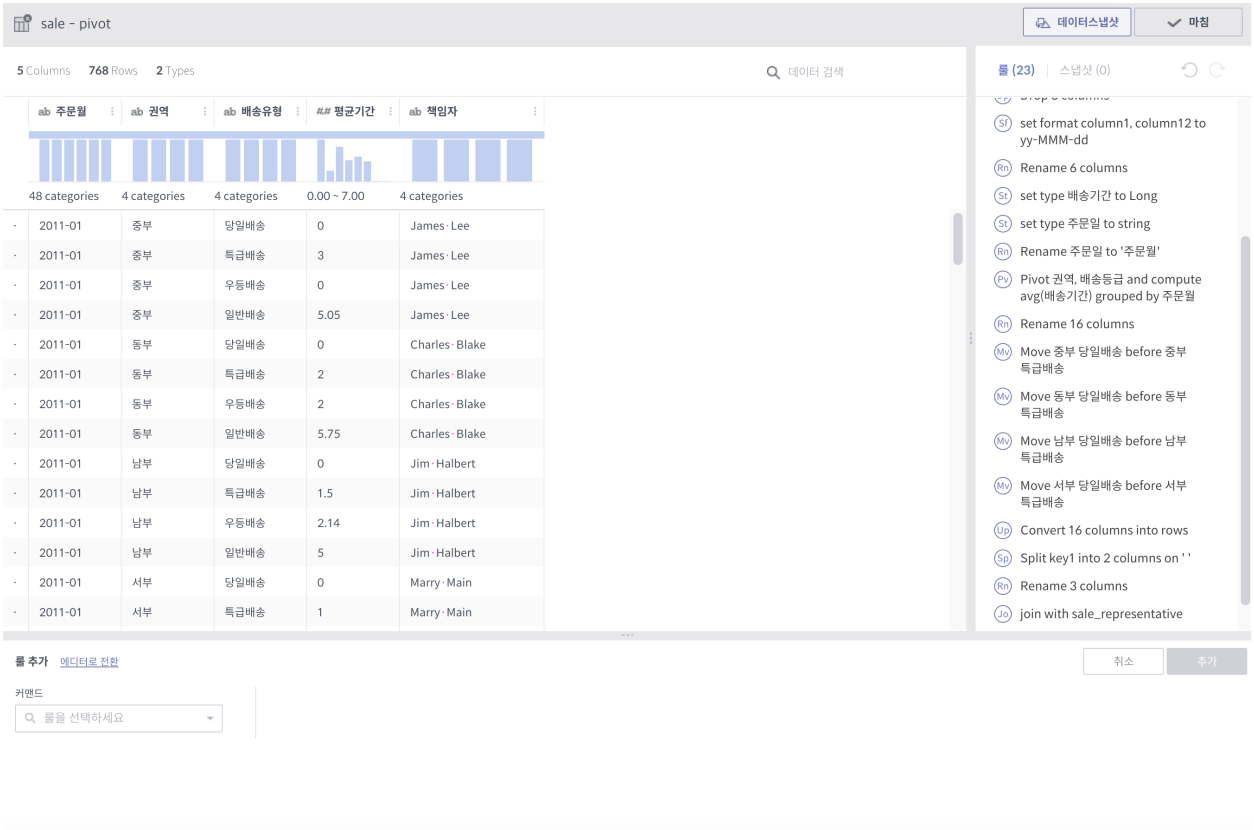
- Join keys: Multiple values may be entered
- Join type: Only inner join supported now

Description

- Joins to the target dataset to create new columns.
- This rule is the same as `join` used by a relational database.
- The results can be previewed by clicking the **Show result** button.

Notes

- The join keys must be included in the columns to join.



union

Similar to *join*, union has a separate popup.

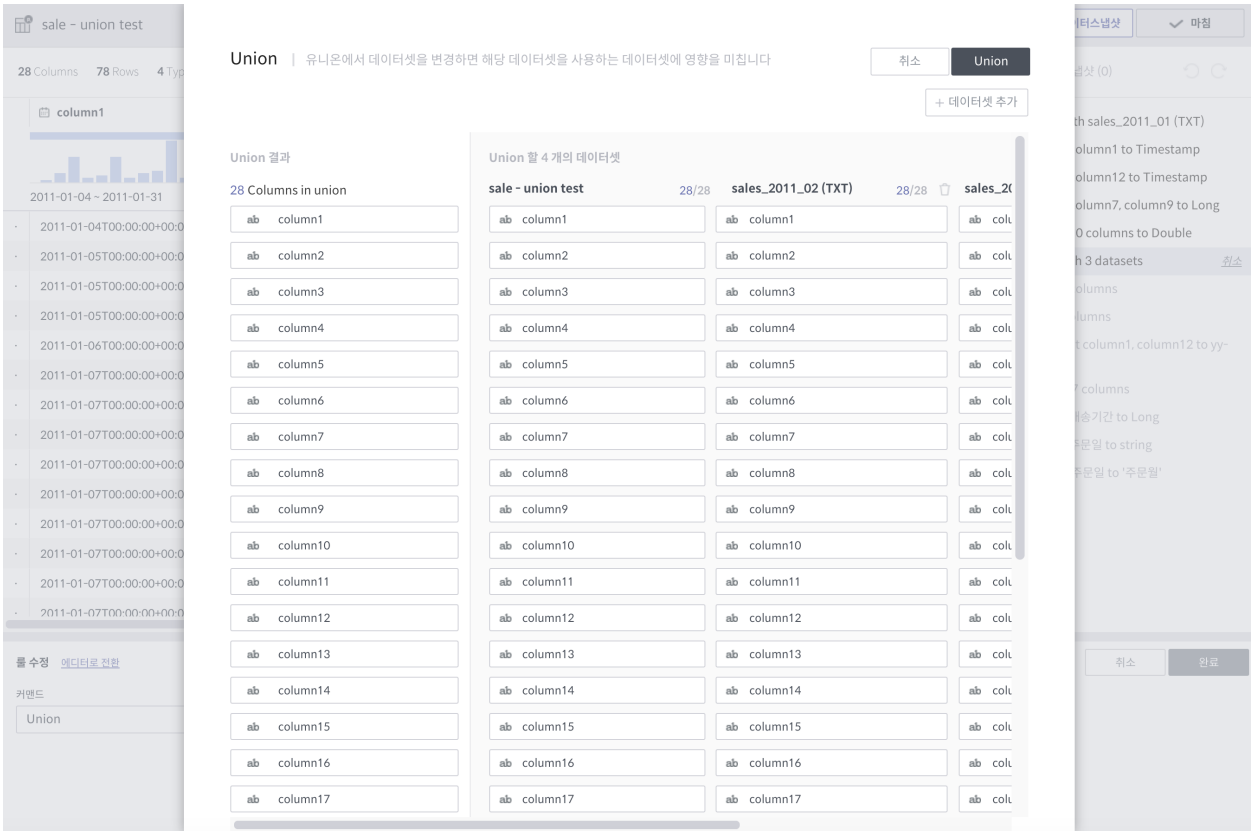
Required arguments (select in a popup)

- Datasets to union: Multiple selections allowed.

Description

- The content of the selected datasets is also processed.
- This rule is the same as `union all` used by a relational database.

Notes



- The target datasets must coincide with the dataset that unions them in terms of column name, type, and number of columns.

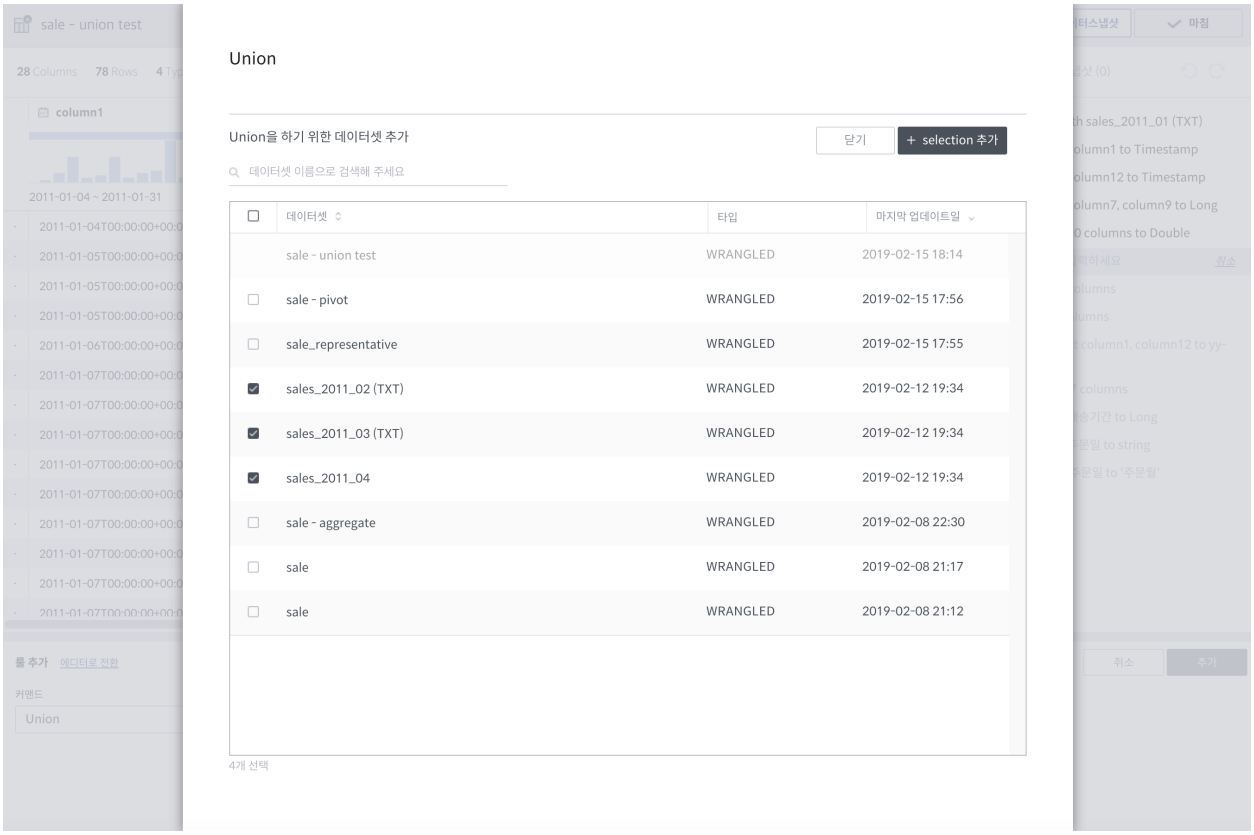
window

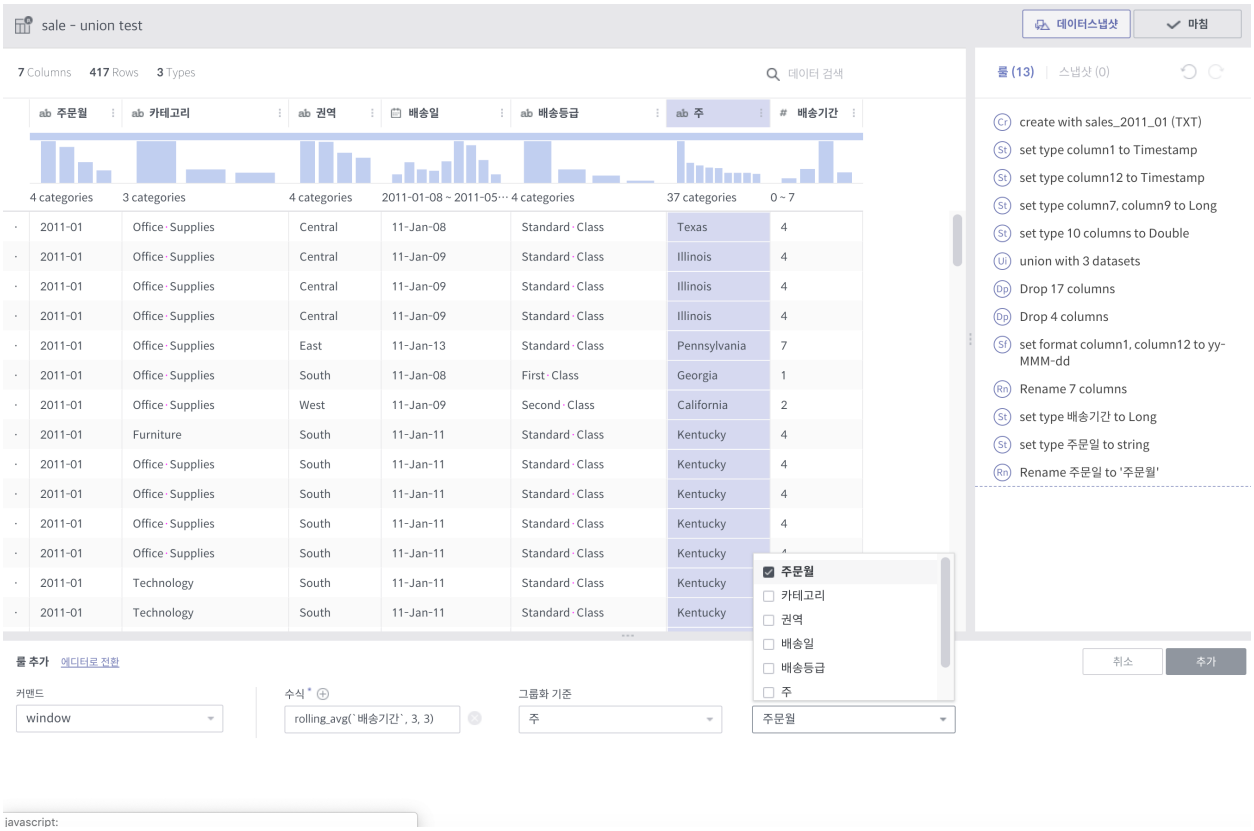
Required arguments

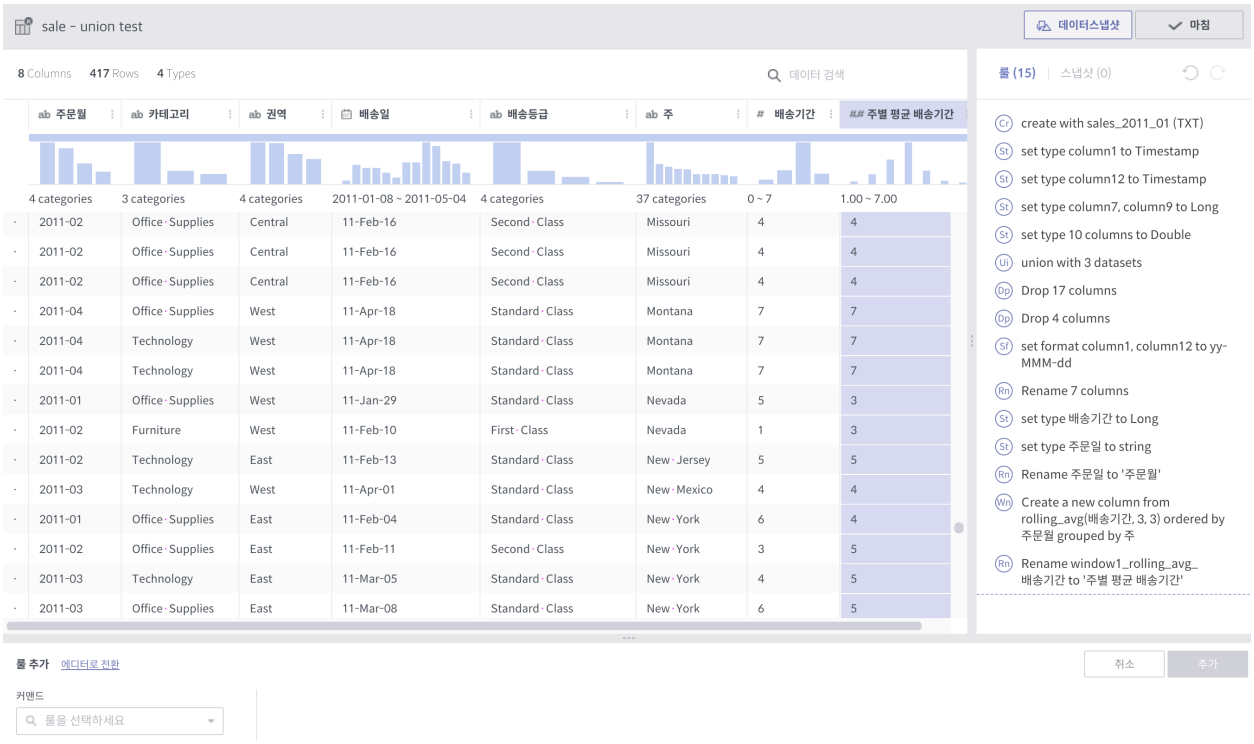
- Expression: A list of window functions
- Group by: A list of columns that group values by. Row order created within each group. If not specified, the whole data is sorted based on the Sort by setting.
- Sort by: Specifies columns by which the order of rows is determined. If not specified, data is sorted in the order of being inputted.

Description

- Column values are created by calculating with the values of the preceding and following rows.
- The rows are grouped first and then sorted within each group in the specified column order.
  - In the above example, each row value is averaged with the three preceding and following rows within the same State group.
  - If an immediately preceding row does not have the same state, earlier rows are searched.
- The currently available window functions are as follows:
  - row_number()


Union
Union을 하기 위한 데이터셋 추가
닫기
+ selection 추가
데이터셋 이름으로 검색해 주세요
데이터셋
타입
마지막 업데이트일
sale - union test WRANGLED 2019-02-15 18:14
sale - pivot WRANGLED 2019-02-15 17:56
sale_representative WRANGLED 2019-02-15 17:55
sales_2011_02 (TXT) WRANGLED 2019-02-12 19:34
sales_2011_03 (TXT) WRANGLED 2019-02-12 19:34
sales_2011_04 WRANGLED 2019-02-12 19:34
sale - aggregate WRANGLED 2019-02-08 22:30
sale WRANGLED 2019-02-08 21:17
sale WRANGLED 2019-02-08 21:12
4개 선택


sale - union test
데이터스냅샷
마침
7 Columns 417 Rows 3 Types
데이터 검색
주문월 카테고리 권역 배송일 배송등급 주 배송기간
4 categories 3 categories 4 categories 2011-01-08 ~ 2011-05… 4 categories 37 categories 0 ~ 7
2011-01 Office Supplies Central 11-Jan-08 Standard Class Texas 4
2011-01 Office Supplies Central 11-Jan-09 Standard Class Illinois 4
2011-01 Office Supplies Central 11-Jan-09 Standard Class Illinois 4
2011-01 Office Supplies Central 11-Jan-09 Standard Class Illinois 4
2011-01 Office Supplies East 11-Jan-13 Standard Class Pennsylvania 7
2011-01 Office Supplies South 11-Jan-08 First Class Georgia 1
2011-01 Office Supplies West 11-Jan-09 Second Class California 2
2011-01 Furniture South 11-Jan-11 Standard Class Kentucky 4
2011-01 Office Supplies South 11-Jan-11 Standard Class Kentucky 4
2011-01 Office Supplies South 11-Jan-11 Standard Class Kentucky 4
2011-01 Office Supplies South 11-Jan-11 Standard Class Kentucky 4
2011-01 Office Supplies South 11-Jan-11 Standard Class Kentucky 4
2011-01 Technology South 11-Jan-11 Standard Class Kentucky
2011-01 Technology South 11-Jan-11 Standard Class Kentucky
룰 (13) 스냅샷 (0)
create with sales_2011_01 (TXT)
set type column1 to Timestamp
set type column12 to Timestamp
set type column7, column9 to Long
set type 10 columns to Double
union with 3 datasets
Drop 17 columns
Drop 4 columns
set format column1, column12 to yy-MMM-dd
Rename 7 columns
set type 배송기간 to Long
set type 주문일 to string
Rename 주문일 to '주문월'
주문월
카테고리
권역
배송일
배송등급
주
룰 추가 에디터로 전환
취소
추가
커맨드
window
수식 *
rolling_avg(`배송기간`, 3, 3)
그룹화 기준
주
주문월
javascript:



  - lead(*colname*, *int*)
  - lag(*colname*, *int*)
  - rolling_sum(*colname*, *int*, *int*)
  - rolling_avg(*colname*, *int*, *int*)
- In addition to window functions, aggregate functions may be used.

Notes

- When using window functions, error messages may not be properly displayed in the event of insufficient arguments.

8.2.4 Create a data snapshot

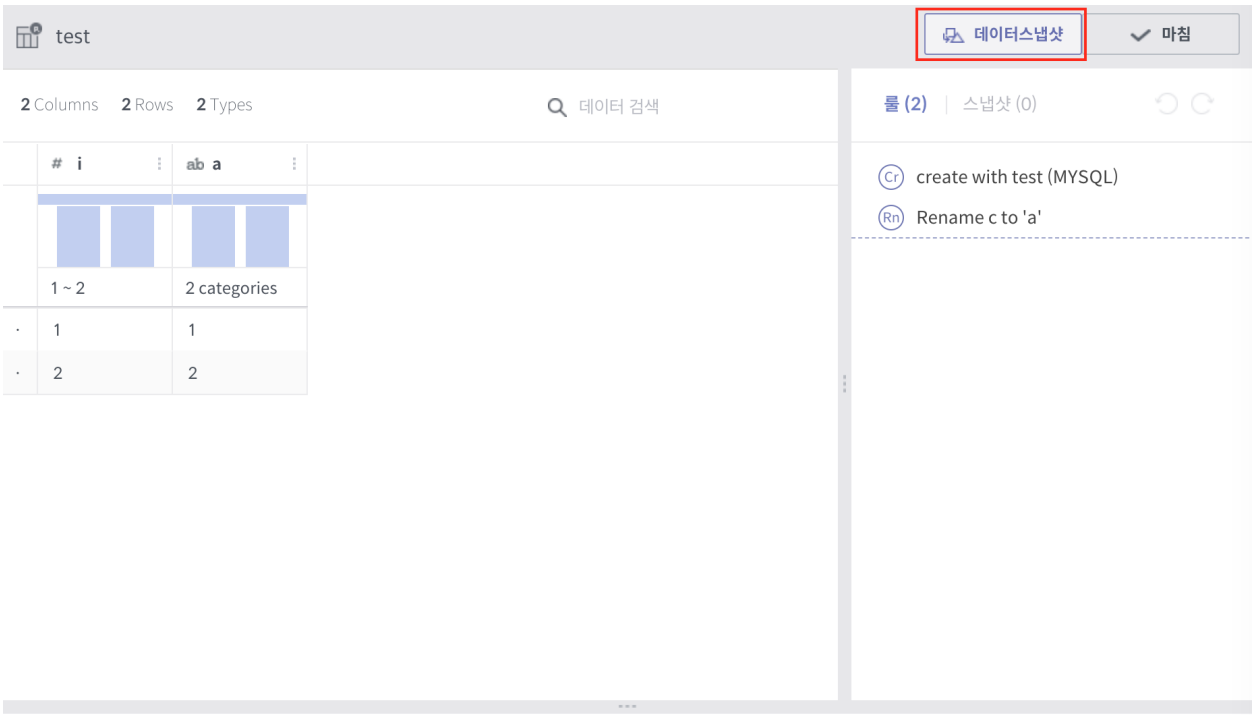
When rule editing is complete, you can create a data snapshot of the finalized dataset, which can then be downloaded to your local PC or ingested into the Metatron engine. Running the data snapshot applies the rules to the entire data, which, in the process of rule editing, applied to a sample dataset of less than 10,000 rows.

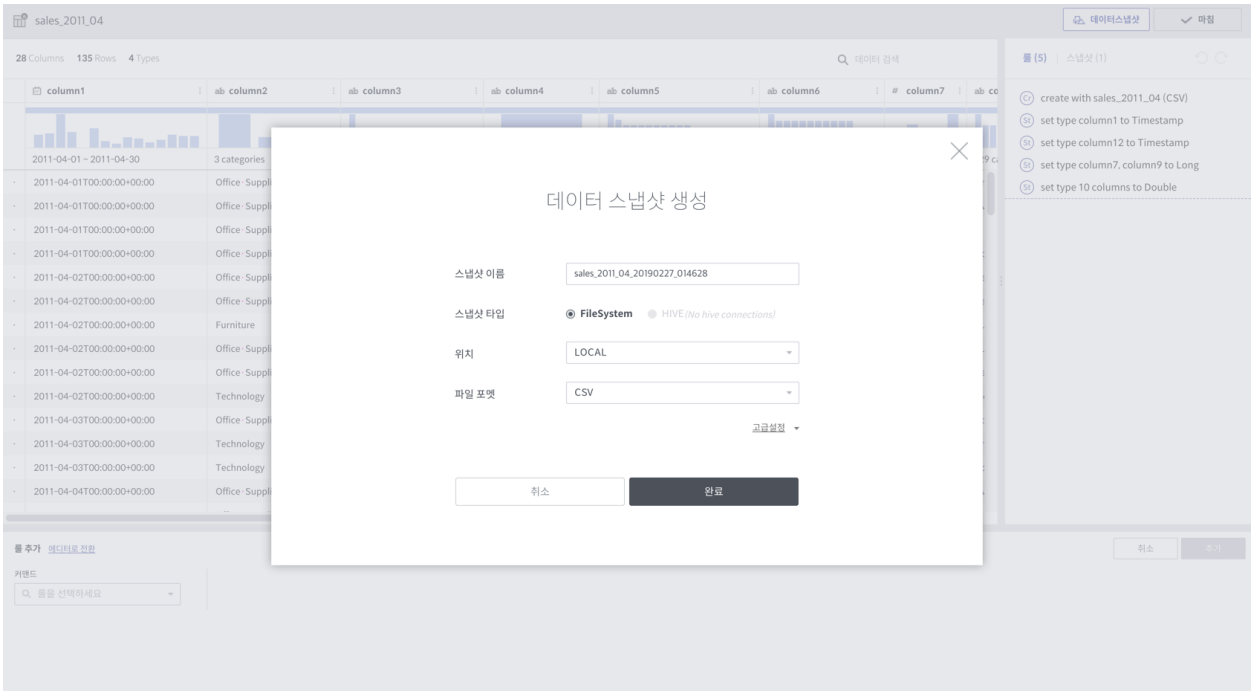
Below are instructions on creating a snapshot:

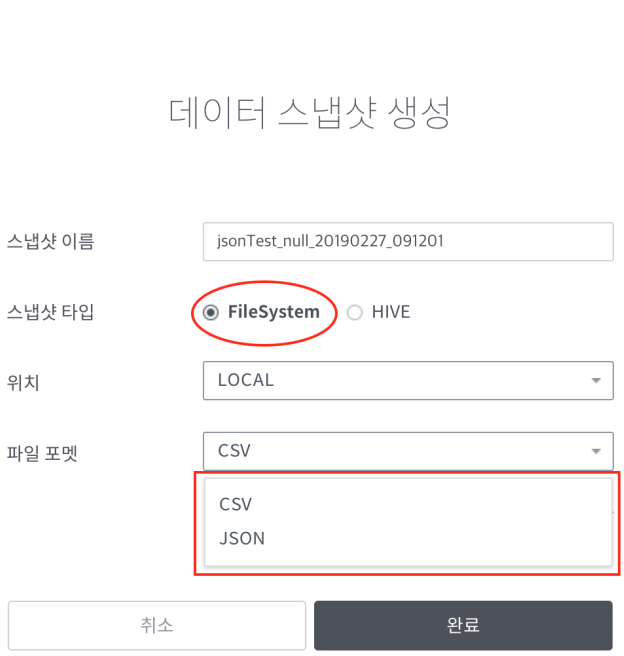
1. Click the **Data Snapshot** button on the upper right of the *Edit rules* window.

2. When a popup is displayed to set snapshot options, select either FileSystem or HIVE (STAGING_DB) under Snapshot type.

    - If FileSystem is selected as the snapshot location, the snapshot will be created as **CSV** or **JSON**.


test
데이터스냅샷
마침
2 Columns
2 Rows
2 Types
데이터 검색
i
ab a
1 ~ 2
2 categories
1
1
2
2
룰 (2)
스냅샷 (0)
create with test (MYSQL)
Rename c to 'a'
룰 추가
에디터로 전환
취소
추가
커맨드
룰을 선택하세요


sales_2011_04
데이터스냅샷
마침
데이터 스냅샷 생성
스냅샷 이름
sales_2011_04_20190227_014628
스냅샷 타입
FileSystem
HIVE
위치
LOCAL
파일 포맷
CSV
고급설정
취소
완료



- The HIVE option is available only when STAGING_DB is enabled. A snapshot is created in the table when you designate a schema name and table name.

3. When the snapshot is created, you can view the snapshot status and related information in the same window.

8.3 Use data snapshot results

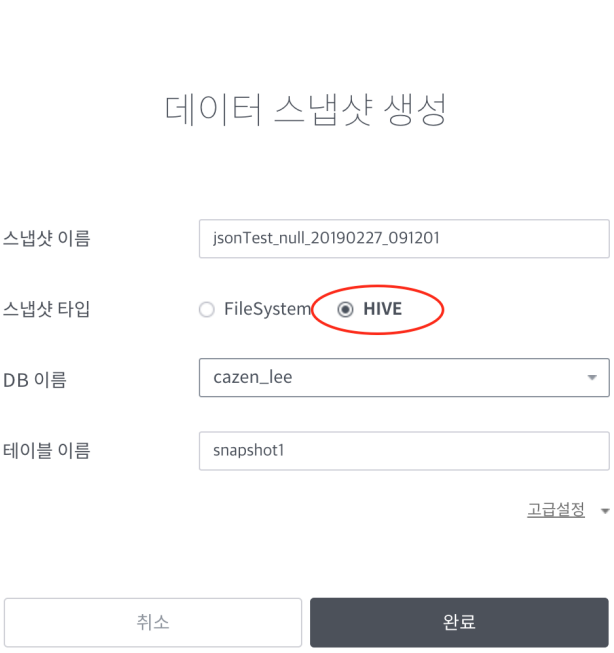
A **data snapshot** created through a dataflow can be used as follows:

- *Check the data snapshot results*
- *Ingest into the Metatron engine*
- *Download as a CSV file*

8.3.1 Check the data snapshot results

The status of snapshot creation can be classified as follows:

- **Success** = SUCCEEDED
- **Failed** = FAILED
- **Preparing** = INITIALIZING, RUNNING, WRITING, TABLE_CREATING, CANCELING



You can view the details of snapshot creation through the two paths below:

- Go to the snapshot list under **MANGEMENT > Data Preparation > Data Snapshot**.

- Click the **Snapshot (#)** tab on the right of the *Edit rules* page in **Dataflow**

In the snapshot details page, you can view details such as data validity ratio and a grid of the created snapshot, and download the results as a CSV file (*Download as CSV*).

If valid data has not been created, the snapshot details page displays an error log.

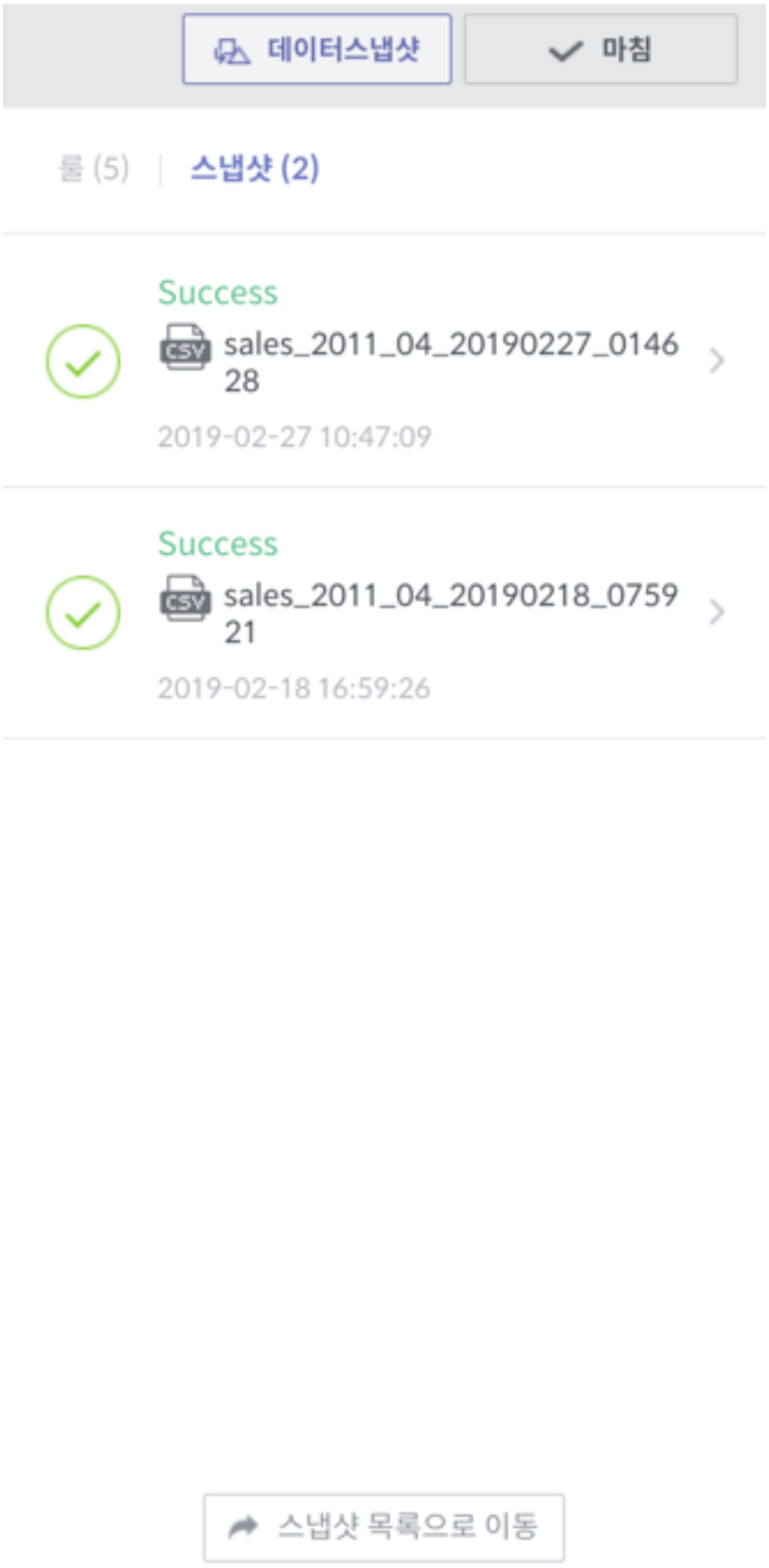
8.3.2 Ingest into the Metatron engine

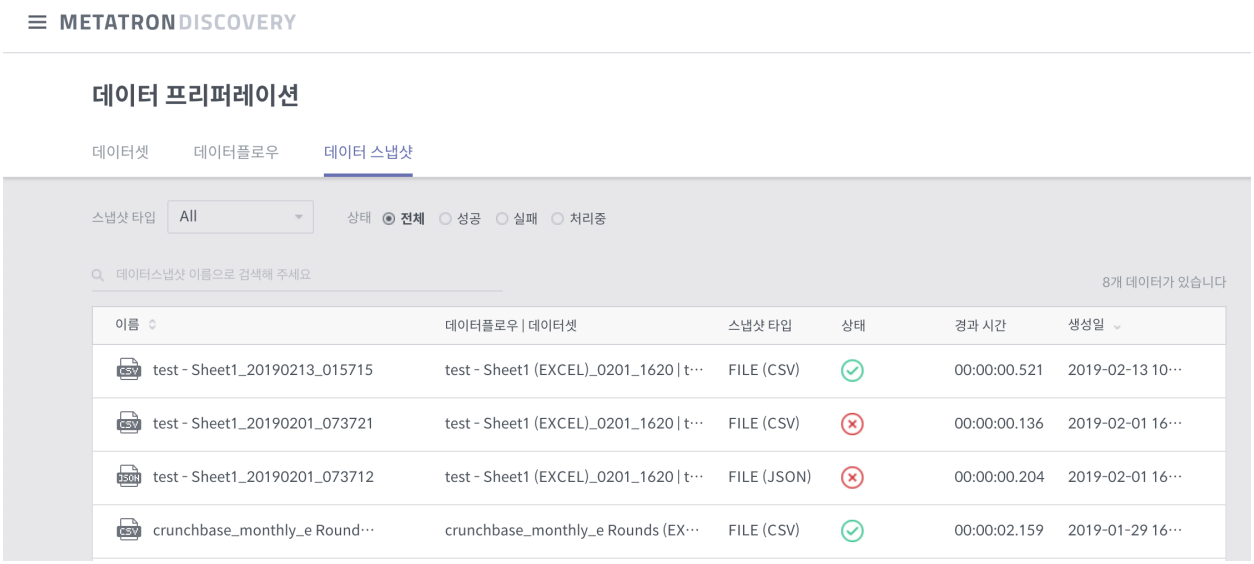
(upcoming feature)

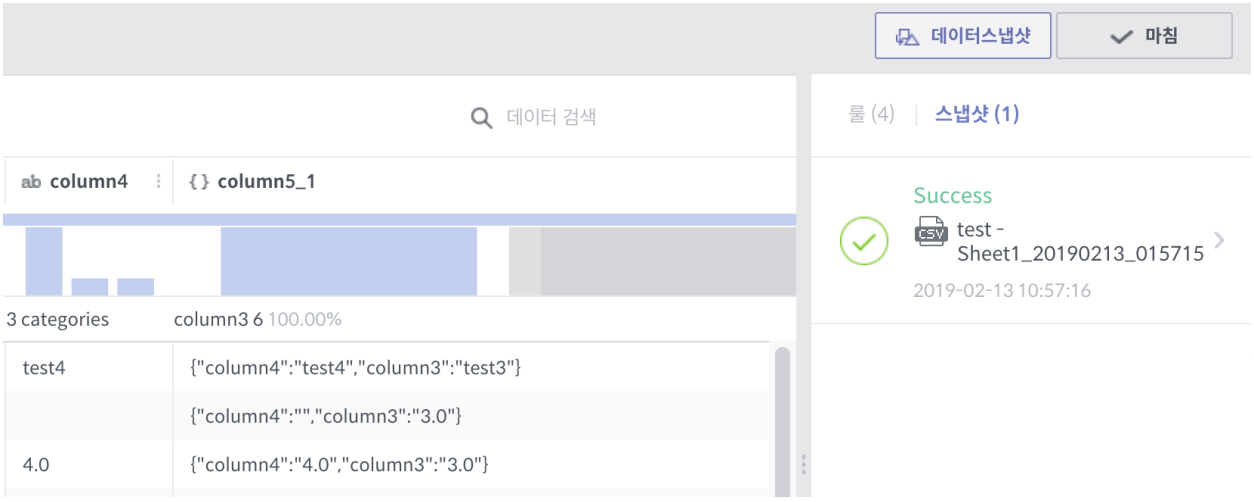
8.3.3 Download as a CSV file

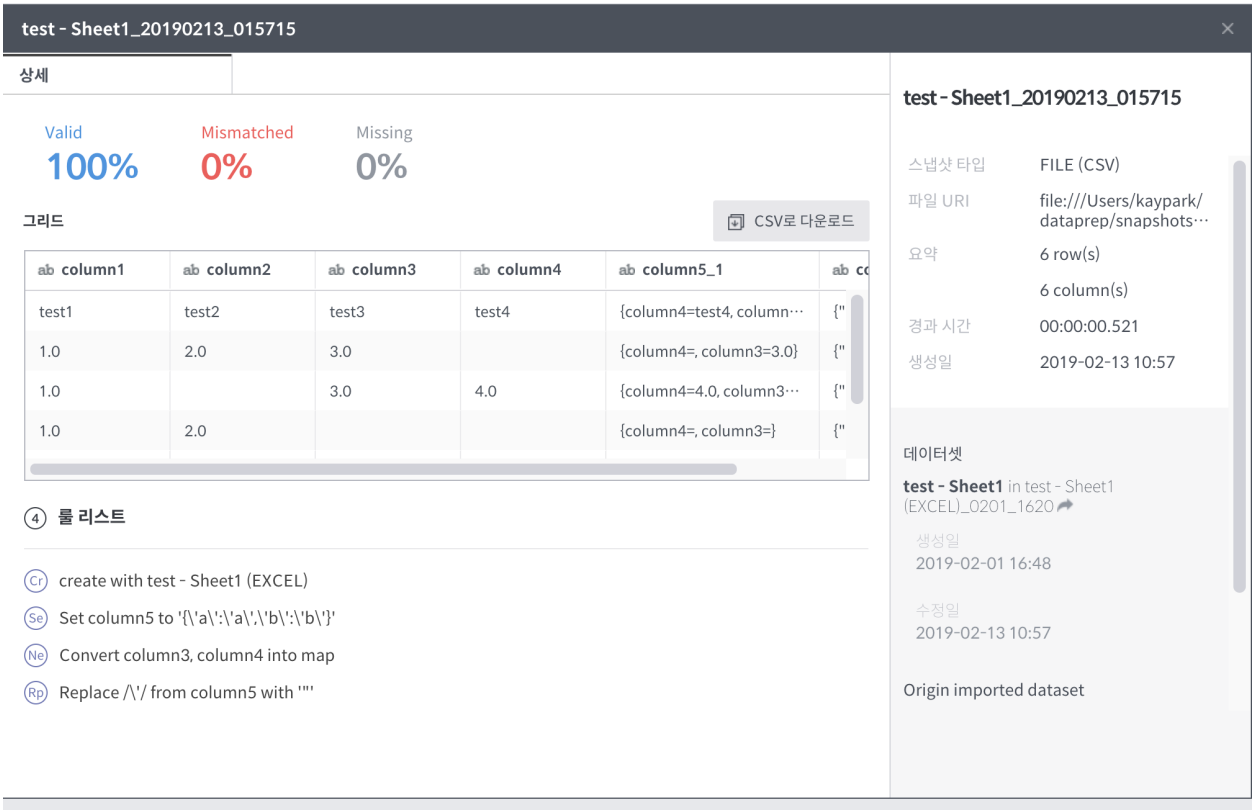
In the details page of a successfully created snapshot, the **Download as CSV** option is enabled.

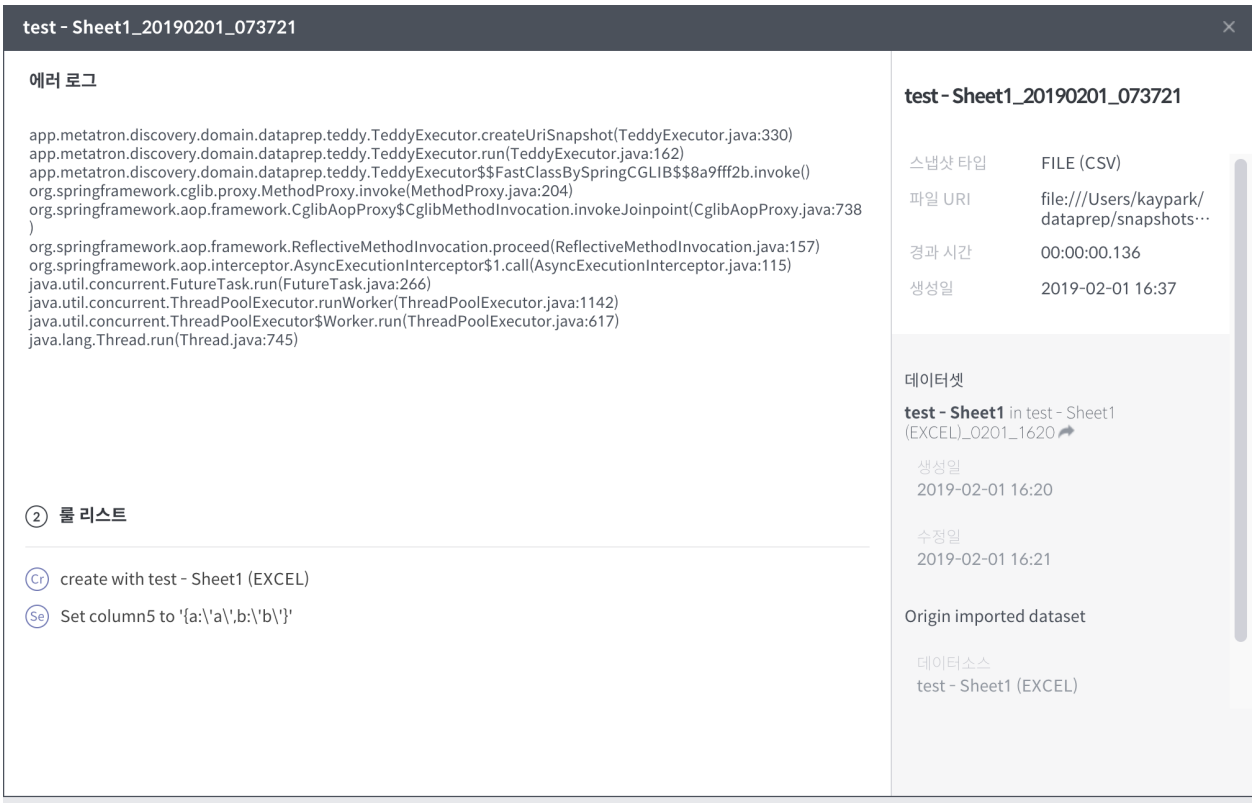
The downloaded file is a standard CSV, with each value separated by a “comma” and each row by a “new line.”

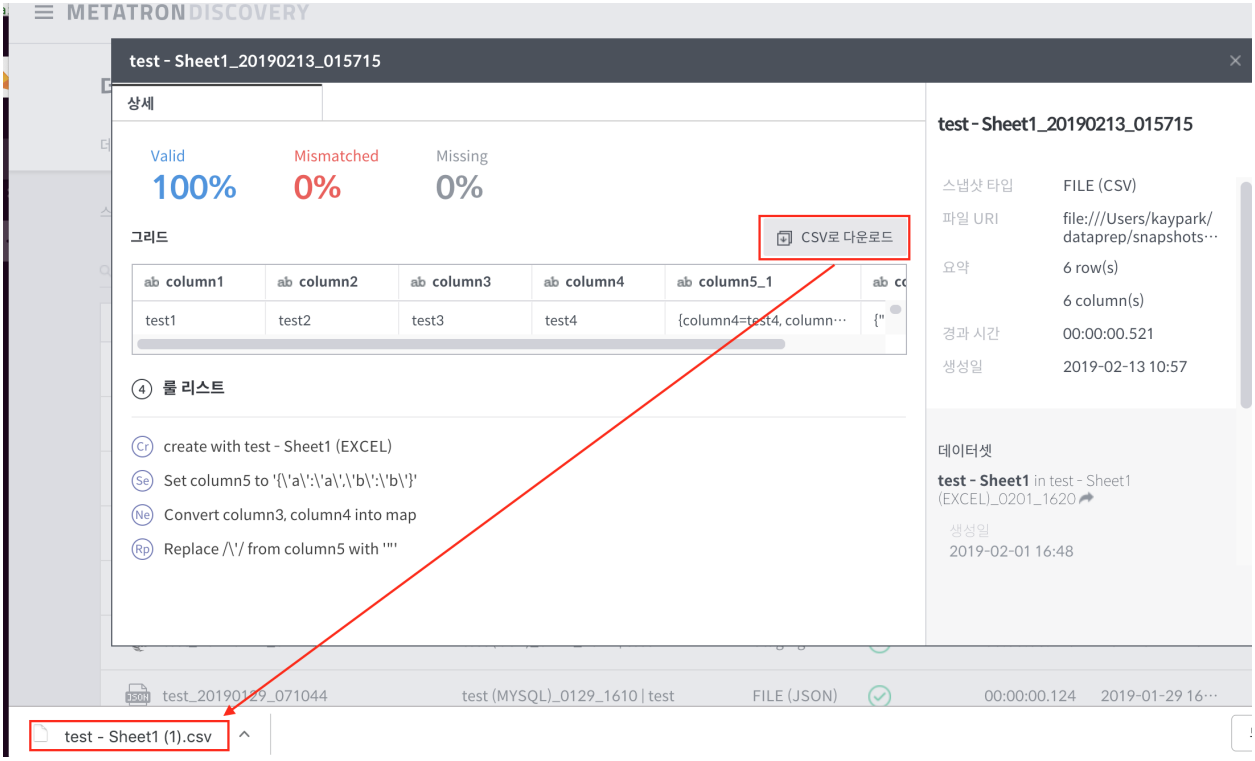

데이터스냅샷
마침
룰 (5)
스냅샷 (2)
Success
sales_2011_04_20190227_014628
2019-02-27 10:47:09
Success
sales_2011_04_20190218_075921
2019-02-18 16:59:26
스냅샷 목록으로 이동


METATRONDISCOVERY
데이터 프리퍼레이션
데이터셋
데이터플로우
데이터 스냅샷
스냅샷 타입
All
상태
전체
성공
실패
처리중
데이터스냅샷 이름으로 검색해 주세요
8개 데이터가 있습니다
이름
데이터플로우 | 데이터셋
스냅샷 타입
상태
경과 시간
생성일
test - Sheet1_20190213_015715
test - Sheet1 (EXCEL)_0201_1620 | t···
FILE (CSV)
00:00:00.521
2019-02-13 10···
test - Sheet1_20190201_073721
test - Sheet1 (EXCEL)_0201_1620 | t···
FILE (CSV)
00:00:00.136
2019-02-01 16···
test - Sheet1_20190201_073712
test - Sheet1 (EXCEL)_0201_1620 | t···
FILE (JSON)
00:00:00.204
2019-02-01 16···
crunchbase_monthly_e Round···
crunchbase_monthly_e Rounds (EX···
FILE (CSV)
00:00:02.159
2019-01-29 16···


데이터스냅샷
마침
데이터 검색
룰 (4)
스냅샷 (1)
column4
column5_1
3 categories
column3 6 100.00%
test4
{"column4":"test4","column3":"test3"}
{"column4":"","column3":"3.0"}
4.0
{"column4":"4.0","column3":"3.0"}
Success
test - Sheet1_20190213_015715
2019-02-13 10:57:16


test - Sheet1_20190213_015715
상세
Valid
100%
Mismatched
0%
Missing
0%
그리드
CSV로 다운로드
column1
column2
column3
column4
column5_1
test1
test2
test3
test4
{column4=test4, column…
1.0
2.0
3.0
{column4=, column3=3.0}
1.0
3.0
4.0
{column4=4.0, column3…
1.0
2.0
{column4=, column3=}
④ 룰 리스트
Cr create with test - Sheet1 (EXCEL)
Se Set column5 to '{\'a\':\'a\',\'b\':\'b\'}'
Ne Convert column3, column4 into map
Rp Replace /\'/ from column5 with '"'
test - Sheet1_20190213_015715
스냅샷 타입
FILE (CSV)
파일 URI
file:///Users/kaypark/dataprep/snapshots…
요약
6 row(s)
6 column(s)
경과 시간
00:00:00.521
생성일
2019-02-13 10:57
데이터셋
test - Sheet1 in test - Sheet1 (EXCEL)_0201_1620
생성일
2019-02-01 16:48
수정일
2019-02-13 10:57
Origin imported dataset


test - Sheet1_20190201_073721
에러 로그
app.metatron.discovery.domain.dataprep.teddy.TeddyExecutor.createUriSnapshot(TeddyExecutor.java:330)
app.metatron.discovery.domain.dataprep.teddy.TeddyExecutor.run(TeddyExecutor.java:162)
app.metatron.discovery.domain.dataprep.teddy.TeddyExecutor$$FastClassBySpringCGLIB$$8a9fff2b.invoke()
org.springframework.cglib.proxy.MethodProxy.invoke(MethodProxy.java:204)
org.springframework.aop.framework.CglibAopProxy$CglibMethodInvocation.invokeJoinpoint(CglibAopProxy.java:738)
org.springframework.aop.framework.ReflectiveMethodInvocation.proceed(ReflectiveMethodInvocation.java:157)
org.springframework.aop.interceptor.AsyncExecutionInterceptor$1.call(AsyncExecutionInterceptor.java:115)
java.util.concurrent.FutureTask.run(FutureTask.java:266)
java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
java.lang.Thread.run(Thread.java:745)
② 룰 리스트
Cr create with test - Sheet1 (EXCEL)
Se Set column5 to '{a:\'a\',b:\'b\'}'
test - Sheet1_20190201_073721
스냅샷 타입
FILE (CSV)
파일 URI
file:///Users/kaypark/dataprep/snapshots…
경과 시간
00:00:00.136
생성일
2019-02-01 16:37
데이터셋
test - Sheet1 in test - Sheet1 (EXCEL)_0201_1620
생성일
2019-02-01 16:20
수정일
2019-02-01 16:21
Origin imported dataset
데이터소스
test - Sheet1 (EXCEL)



test - Sheet1 (1)

column1	column2	column3	column4	column5_1	column5
test1	test2	test3	test4	{column4=test4, column3=test3}	{"a":"a","b":"b"}
1	2	3		{column4=, column3=3.0}	{"a":"a","b":"b"}
1		3	4	{column4=4.0, column3=3.0}	{"a":"a","b":"b"}
1	2			{column4=, column3=}	{"a":"a","b":"b"}
		3		{column4=, column3=3.0}	{"a":"a","b":"b"}