

尚硅谷大数据技术之 Superset

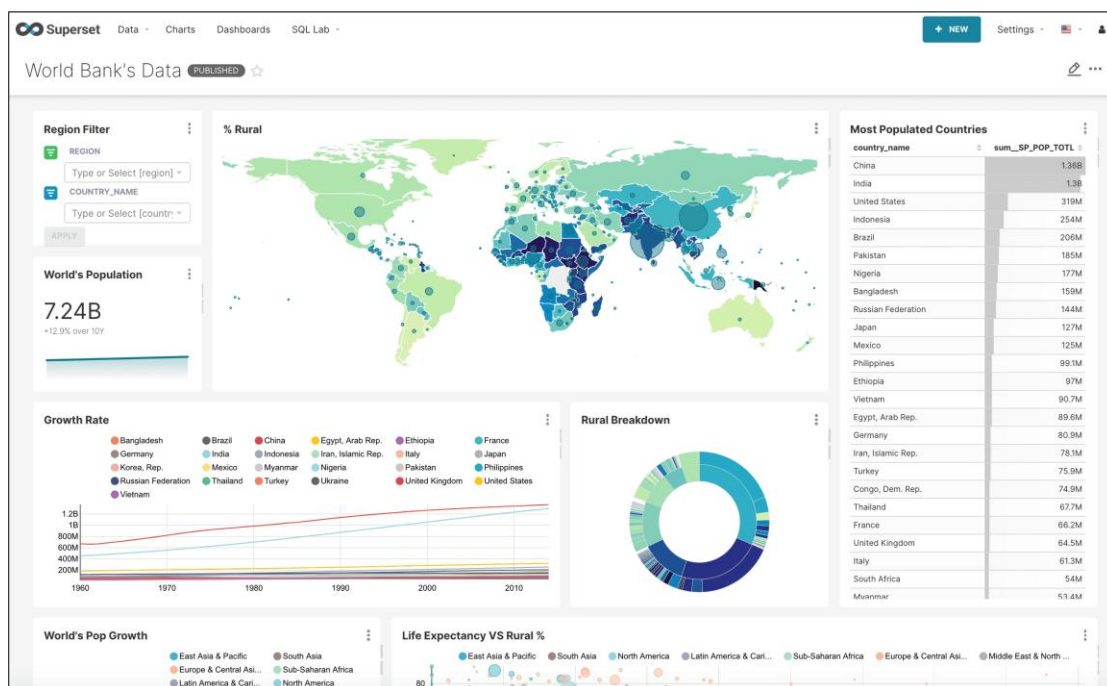
(作者：尚硅谷研究院)

版本：V4.0

第 1 章 Superset 入门

1.1 Superset 概述

Apache Superset 是一个现代的数据探索和可视化平台。它功能强大且十分易用，可对接各种数据源，包括很多现代的大数据分析引擎，拥有丰富的图表展示形式，并且支持自定义仪表盘。



1.2 环境说明

本课程使用的服务器操作系统为 CentOS 7，Superset 对接的数据源为 MySQL 数据库。

第 2 章 Superset 安装

Superset 官网地址：<http://superset.apache.org/>

2.1 安装 Python 环境

Superset 是由 Python 语言编写的 Web 应用，要求 Python3.7 的环境。

2.1.1 安装 Miniconda

conda 是一个开源的包、环境管理器，可以用于在同一个机器上安装不同 Python 版本的软件包及其依赖，并能够在不同的 Python 环境之间切换，Anaconda 包括 Conda、Python 以及一大堆安装好的工具包，比如：numpy、pandas 等，Miniconda 包括 Conda、Python。

此处，我们不需要如此多的工具包，故选择 MiniConda。

1) 下载 Miniconda (Python3 版本)

下载地址：https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh

2) 安装 Miniconda

(1) 执行以下命令进行安装，并按照提示操作，直到安装完成。

```
[atguigu@hadoop102 lib]$ bash Miniconda3-latest-Linux-x86_64.sh
```

(2) 在安装过程中，出现以下提示时，可以指定安装路径

```
Miniconda3 will now be installed into this location:
/home/atguigu/miniconda3
```

- Press ENTER to confirm the location
- Press CTRL-C to abort the installation
- Or specify a different location below

```
[/home/atguigu/miniconda3] >>> /opt/module/miniconda3
```

(3) 出现以下字样，即为安装完成

```
Thank you for installing Miniconda3!
```

3) 加载环境变量配置文件，使之生效

```
[atguigu@hadoop102 lib]$ source ~/.bashrc
```

4) 取消激活 base 环境

Miniconda 安装完成后，每次打开终端都会激活其默认的 base 环境，我们可通过以下命令，禁止激活默认 base 环境。

```
[atguigu@hadoop102 lib]$ conda config --set auto_activate_base
false
```

2.1.2 创建 Python3.7 环境

1) 配置 conda 国内镜像

```
(base) [atguigu@hadoop102 ~]$ conda config --add channels
https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/free
(base) [atguigu@hadoop102 ~]$ conda config --add channels
https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/main
```

2) 创建 Python3.7 环境

```
(base) [atguigu@hadoop102 ~]$ conda create --name superset
python=3.7
```

说明：conda 环境管理常用命令

创建环境: `conda create -n env_name python=3.7`

查看所有环境: `conda info --envs`

删除一个环境: `conda remove -n env_name --all`

3) 激活 superset 环境

```
(base) [atguigu@hadoop102 ~]$ conda activate superset
```

激活后效果如下图所示

```
(superset) [atguigu@hadoop102 ~]$
```

说明: 退出当前环境

```
(superset) [atguigu@hadoop102 ~]$ conda deactivate
```

4) 执行 python 命令查看 python 版本

```
(superset) [atguigu@hadoop102 ~]$ python
Python 3.6.10 |Anaconda, Inc.| (default, Jan 7 2020, 21:14:29)
[GCC 7.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> quit();
```

2.2 Superset 部署

2.2.1 安装依赖

安装 Superset 之前, 需安装以下所需依赖

```
(superset) [atguigu@hadoop102 ~]$ sudo yum install -y gcc gcc-
c++ libffi-devel python-devel python-pip python-wheel python-
setuptools openssl-devel cyrus-sasl-devel openldap-devel
```

2.2.2 安装 Superset

1) 安装 (更新) setuptools 和 pip

```
(superset) [atguigu@hadoop102 ~]$ pip install --upgrade
setuptools pip -i https://pypi.douban.com/simple/
```

说明: pip 是 python 的包管理工具, 可以和 centos 中的 yum 类比

2) 安装 Superset

```
(superset) [atguigu@hadoop102 ~]$ pip install apache-superset -
i https://pypi.douban.com/simple/
```

说明: -i 的作用是指定镜像, 这里选择国内镜像

注: 如果遇到网络错误导致不能下载, 可尝试更换镜像

```
(superset) [atguigu@hadoop102 ~]$ pip install apache-superset
--trusted-host https://repo.huaweicloud.com -i
https://repo.huaweicloud.com/repository/pypi/simple
```

3) 初始化 Superset 数据库

```
(superset) [atguigu@hadoop102 ~]$ superset db upgrade
```

4) 创建管理员用户

```
(superset) [atguigu@hadoop102 ~]$ export FLASK_APP=superset
(superset) [atguigu@hadoop102 ~]$ superset fab create-admin
```

说明：flask 是一个 python web 框架，Superset 使用的就是 flask

5) Superset 初始化

```
(superset) [atguigu@hadoop102 ~]$ superset init
```

2.2.3 启动 Superset

1) 安装 gunicorn

```
(superset) [atguigu@hadoop102 ~]$ pip install gunicorn -i
https://pypi.douban.com/simple/
```

说明：gunicorn 是一个 Python Web Server，可以和 java 中的 TomCat 类比

2) 启动 Superset

(1) 确保当前 conda 环境为 superset，及下图所示

```
(superset) [atguigu@hadoop102 ~]$
```

(2) 启动

```
(superset) [atguigu@hadoop102 ~]$ gunicorn --workers 5 --timeout
120 --bind hadoop102:8787 "superset.app:create_app()" --daemon
```

说明：

--workers: 指定进程个数

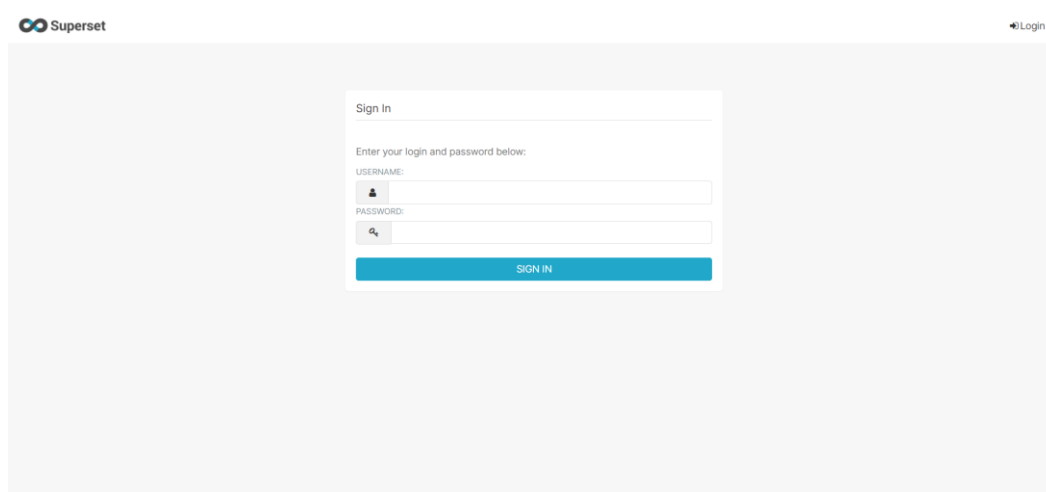
--timeout: worker 进程超时时间，超时会自动重启

--bind: 绑定本机地址，即为 Superset 访问地址

--daemon: 后台运行

(3) 登录 Superset

访问 <http://hadoop102:8787>，并使用 2.2.2 节中第 4 步创建的管理员账号进行登录。



3) 停止 superset

停掉 gunicorn 进程

```
(superset) [atguigu@hadoop102 ~]$ ps -ef | awk '/superset/ && !/awk/{print $2}' | xargs kill -9
```

退出 superset 环境

```
(superset) [atguigu@hadoop102 ~]$ conda deactivate
```

2.2.4 superset 启停脚本

1) 创建 superset.sh 文件

```
[atguigu@hadoop102 bin]$ vim superset.sh
```

内容如下

```
#!/bin/bash

superset_status(){
    result=`ps -ef | awk '/gunicorn/ && !/awk/{print $2}' | wc -l`
    if [[ $result -eq 0 ]]; then
        return 0
    else
        return 1
    fi
}

superset_start(){
    source ~/.bashrc
    superset_status >/dev/null 2>&1
    if [[ $? -eq 0 ]]; then
        conda activate superset ; gunicorn --workers 5 --
timeout      120      --bind      hadoop102:8787      --daemon
'superset.app:create_app()'
    else
        echo "superset 正在运行"
    fi
}

superset_stop(){
    superset_status >/dev/null 2>&1
    if [[ $? -eq 0 ]]; then
        echo "superset 未在运行"
    else
        ps -ef | awk '/gunicorn/ && !/awk/{print $2}' | xargs
kill -9
    fi
}

case $1 in
    start )
        echo "启动 Superset"
        superset_start
        ;;
    stop )
        echo "停止 Superset"
```

```
superset_stop
;;
restart )
    echo "重启 Superset"
    superset_stop
    superset_start
;;
status )
    superset_status >/dev/null 2>&1
    if [[ $? -eq 0 ]]; then
        echo "superset 未在运行"
    else
        echo "superset 正在运行"
    fi
fi
esac
```

2) 加执行权限

```
[atguigu@hadoop102 bin]$ chmod +x superset.sh
```

3) 测试

启动 superset

```
[atguigu@hadoop102 bin]$ superset.sh start
```

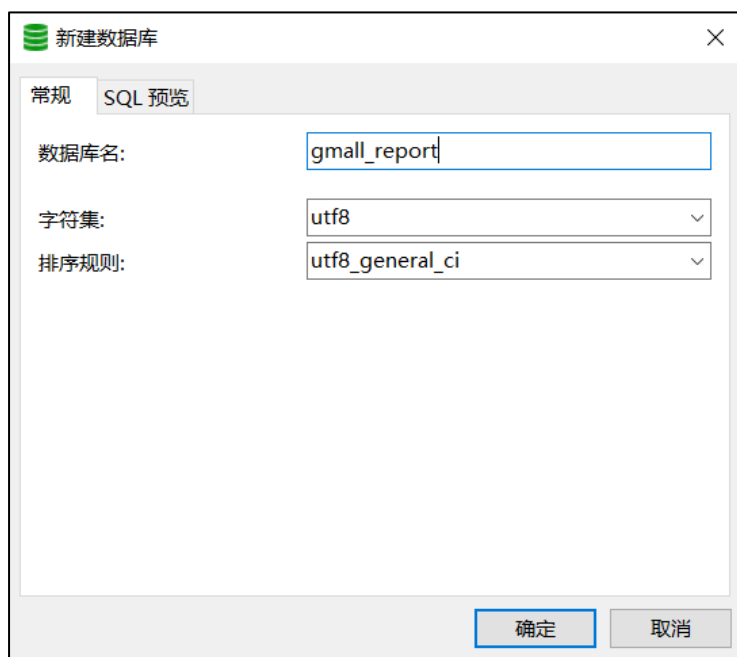
停止 superset

```
[atguigu@hadoop102 bin]$ superset.sh stop
```

第 3 章 Superset 使用

3.1 准备 MySQL 数据源

1) 创建 MySQL 数据库



新建数据库对话框，显示了数据库名、字符集和排序规则的设置。

数据库名:	字符集:	排序规则:
gmall_report	utf8	utf8_general_ci

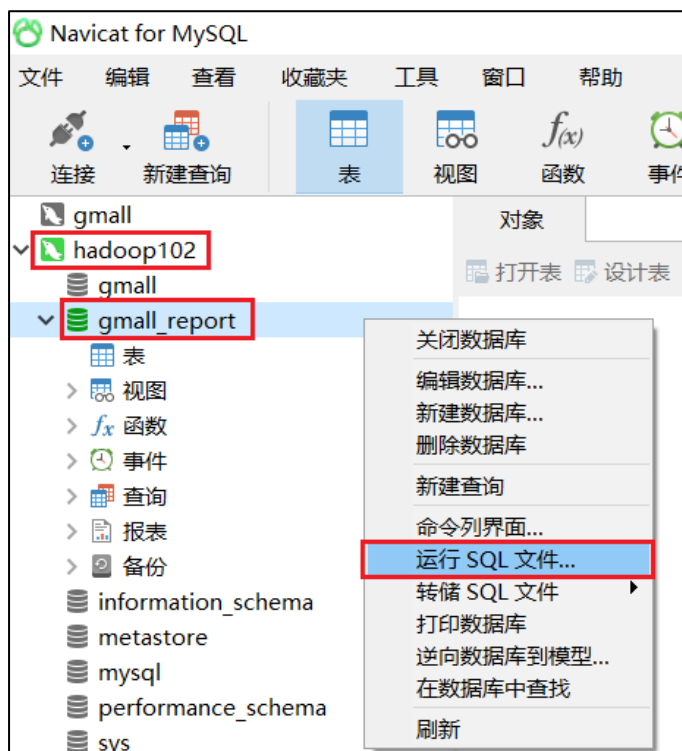
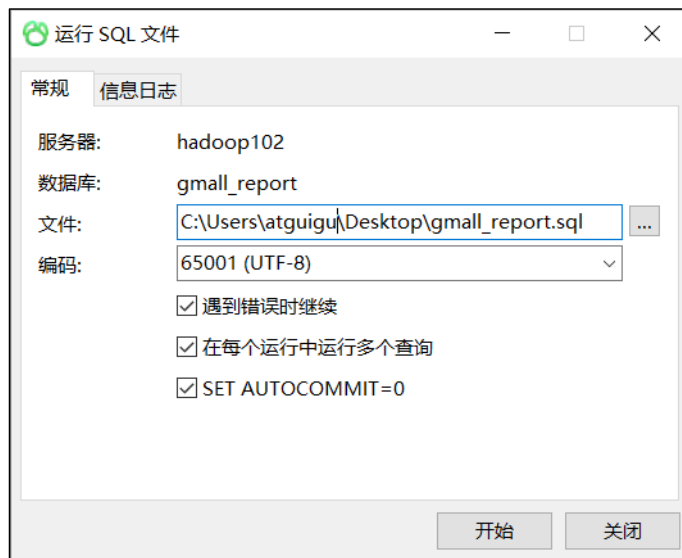
底部有“确定”和“取消”按钮。

2) 导入表结构和模拟数据

按以下步骤将 gmall_report.sql 脚本导入到刚刚创建的 gmall_report 数据库中。



gmall_report.sql



3) 查看导入结果

ads_order_by_province @gmail_report (hadoop102) - 表 - Navicat for MySQL

文件 编辑 查看 表 收藏夹 工具 窗口 帮助

连接 新建查询 表 视图 函数 事件 用户 查询 报表 备份 自动运行 模型

对象: ads_order_by_province @g...

开始事务 文本 筛选 排序 导入 导出

dt	province_id	province_name	area_code	iso_code	order_count	order_amount
2020-06-14	1	北京	110000	CN-11	13	402232.30
2020-06-14	10	福建	350000	CN-35	7	315645.65
2020-06-14	11	江西	360000	CN-36	4	114391.00
2020-06-14	12	山东	370000	CN-37	5	291400.00
2020-06-14	13	重庆	500000	CN-50	6	153022.00
2020-06-14	14	台湾	710000	CN-71	6	184981.70
2020-06-14	15	黑龙江	230000	CN-23	2	80049.00
2020-06-14	16	吉林	220000	CN-22	8	144899.00
2020-06-14	17	辽宁	210000	CN-21	2	20976.00
2020-06-14	18	陕西	610000	CN-61	3	100557.00
2020-06-14	19	甘肃	620000	CN-62	9	422565.70
2020-06-14	2	天津	120000	CN-12	11	283788.70
2020-06-14	20	青海	630000	CN-63	12	289611.65
2020-06-14	21	宁夏	640000	CN-64	5	184526.00
2020-06-14	22	新疆	650000	CN-65	9	262549.35
2020-06-14	23	河南	410000	CN-41	5	197189.30
2020-06-14	24	湖北	420000	CN-42	4	25823.35
2020-06-14	25	湖南	430000	CN-43	7	187572.35

3.2 对接 MySQL 数据源

3.2.1 安装依赖

```
(superset) [atguigu@hadoop102 ~]$ conda install mysqlclient
```

说明：对接不同的数据源，需安装不同的依赖，以下地址为官网说明

<https://superset.apache.org/docs/databases/installing-database-drivers>

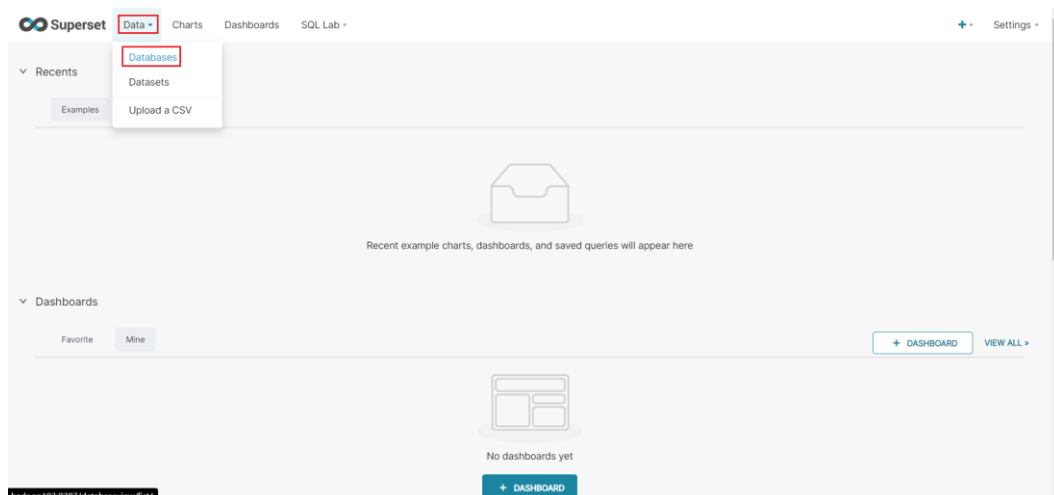
3.2.2 重启 Superset

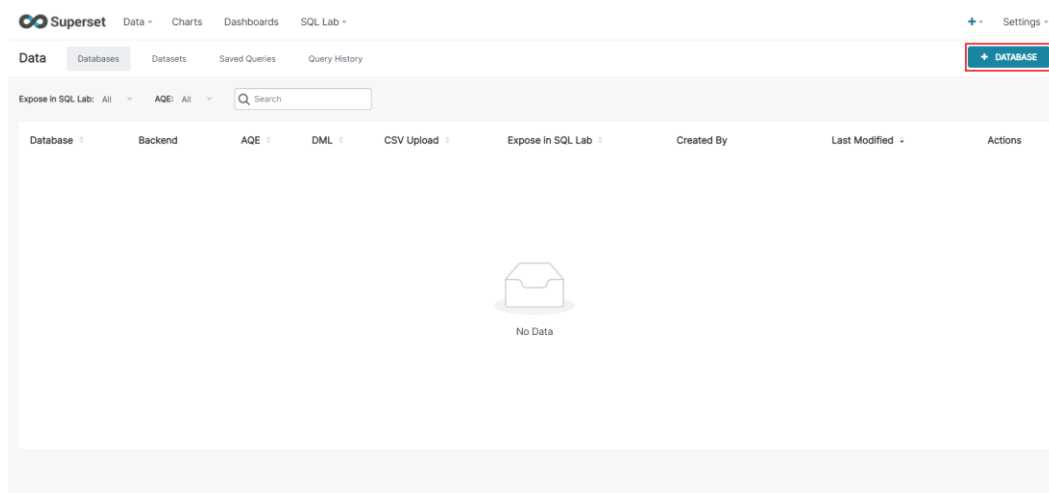
```
(superset) [atguigu@hadoop102 ~]$ superset.sh restart
```

3.2.3 数据源配置

1) Database 配置

Step1: 点击 Data/Databases

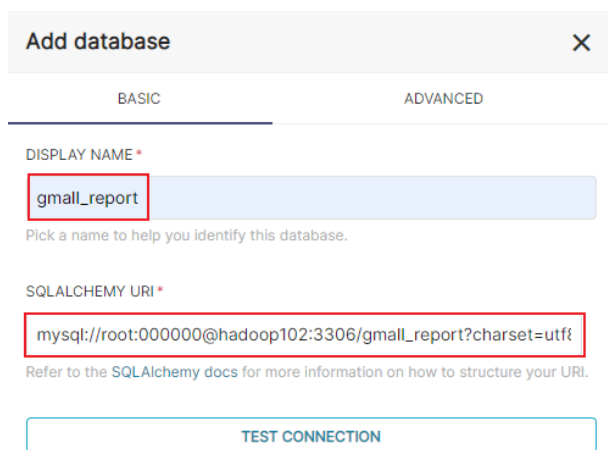


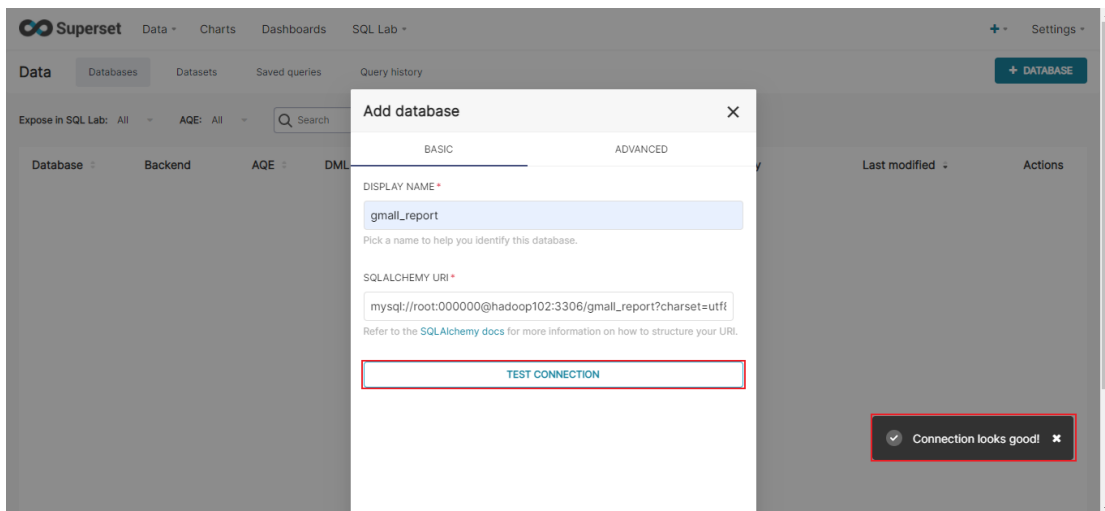
Step2: 点击+DATABASE**Step3: 点击填写 Database 及 SQL Alchemy URI**

注：SQL Alchemy URI 编写规范：`mysql://用户名:密码@主机名:端口号/数据库名称`

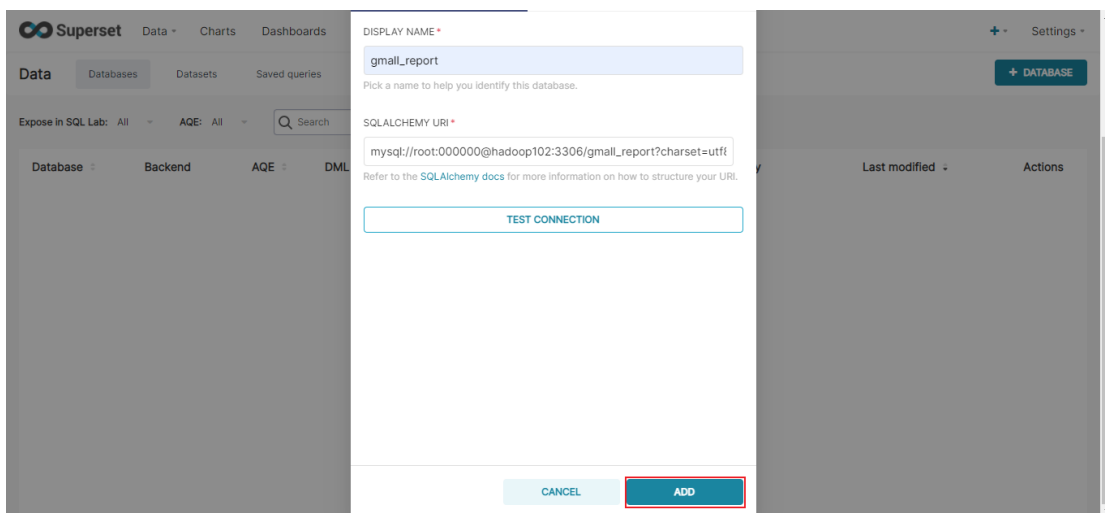
此处填写：

`mysql://root:000000@hadoop102:3306/gmall_report?charset=utf8`

A screenshot of the 'Add database' modal form in Superset. The form has two tabs: 'BASIC' and 'ADVANCED'. Under the 'BASIC' tab, there are two input fields. The first is labeled 'DISPLAY NAME *' and contains the text 'gmall_report'. The second is labeled 'SQLALCHEMY URI *' and contains the text 'mysql://root:000000@hadoop102:3306/gmall_report?charset=utf8'. Below the URI field, there is a small text reference: 'Refer to the SQLAlchemy docs for more information on how to structure your URI.' At the bottom of the form is a button labeled 'TEST CONNECTION'.**Step4: 点击 Test Connection，出现“Connection looks good!”提示即表示连接成功**

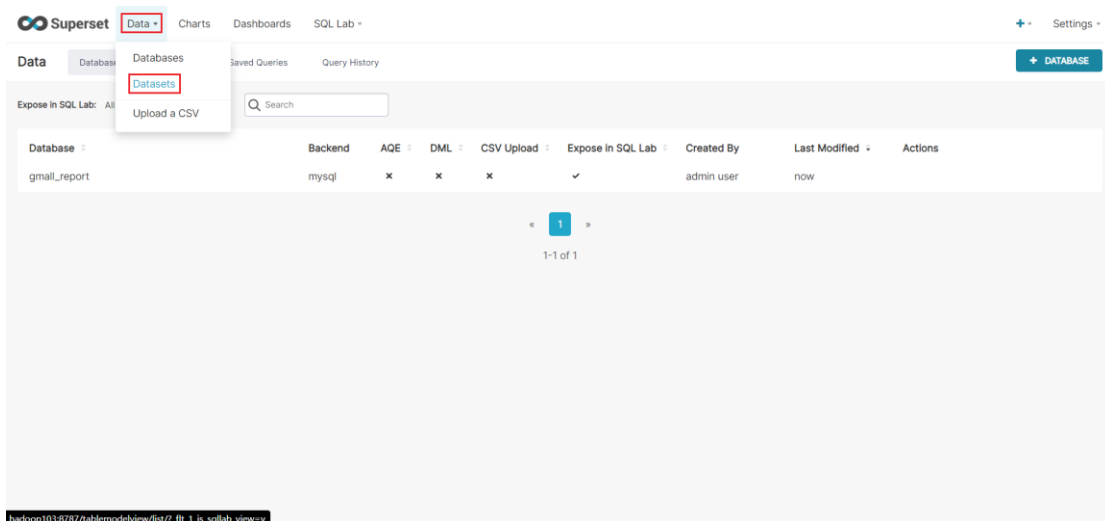


Step5: 点击 ADD

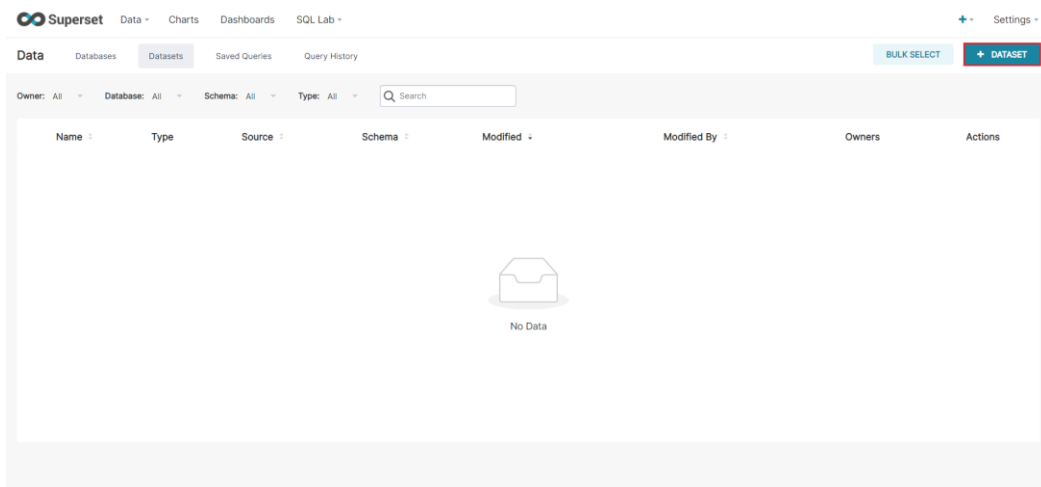


2) Table 配置

Step1: 点击 Data/Datasets



Step2: 点击 Data/ Datasets



Step3: 配置 Table

!

Add dataset

×

DATASOURCE

Database: mysql **gmall_report**

SCHEMA

Schema: **gmall_report**

SEE TABLE SCHEMA 4 IN GMALL_REPORT

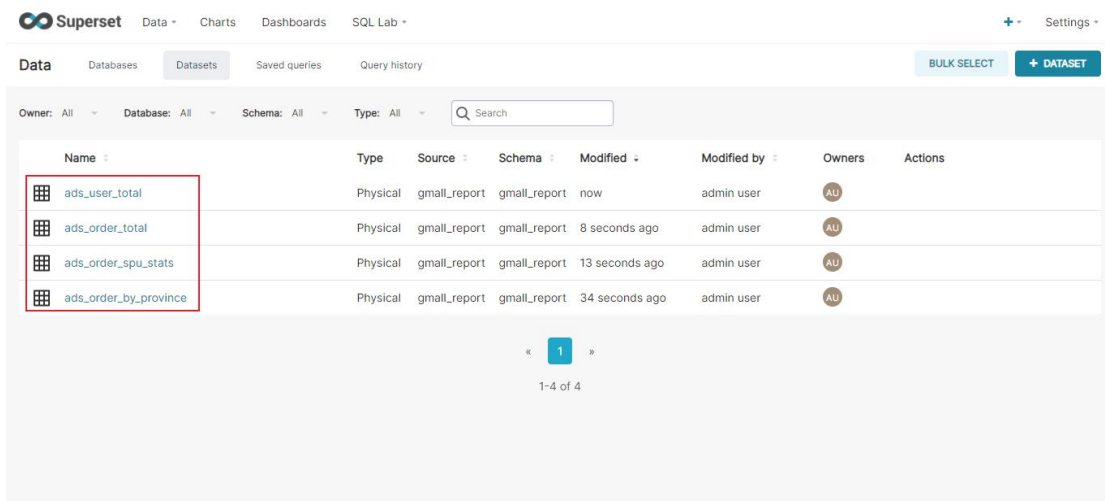
TABLE

ads_order_by_province

CANCEL

ADD

4) 所有 table 配置完毕，如下图所示



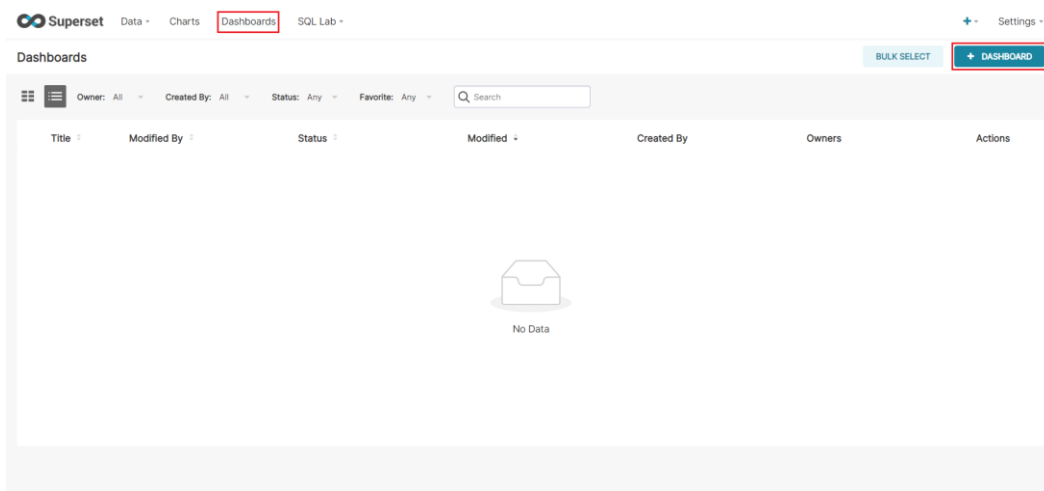
The screenshot shows the Superset 'Data' section with the 'Datasets' tab selected. A table lists four datasets, all of which are highlighted with a red border. The table has columns for Name, Type, Source, Schema, Modified, Modified by, Owners, and Actions. The datasets are: ads_user_total, ads_order_total, ads_order_spu_stats, and ads_order_by_province. All are Physical type, from gmail_report source, in the gmail_report schema, and owned by admin user.

Name	Type	Source	Schema	Modified	Modified by	Owners	Actions
ads_user_total	Physical	gmail_report	gmail_report	now	admin user	AU	
ads_order_total	Physical	gmail_report	gmail_report	8 seconds ago	admin user	AU	
ads_order_spu_stats	Physical	gmail_report	gmail_report	13 seconds ago	admin user	AU	
ads_order_by_province	Physical	gmail_report	gmail_report	34 seconds ago	admin user	AU	

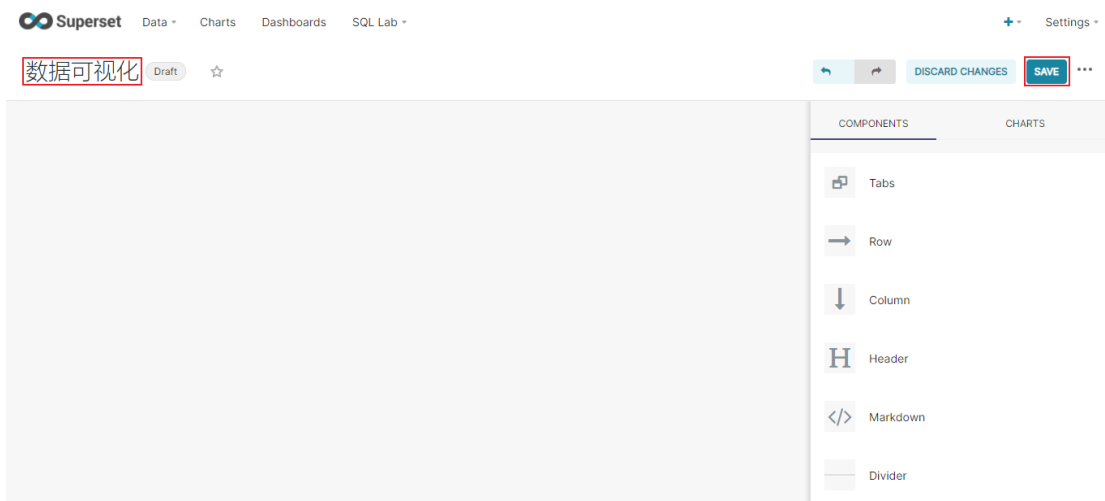
3.3 制作仪表盘

3.3.1 创建空白仪表盘

1) 点击 Dashboards/+DASHBOARDS

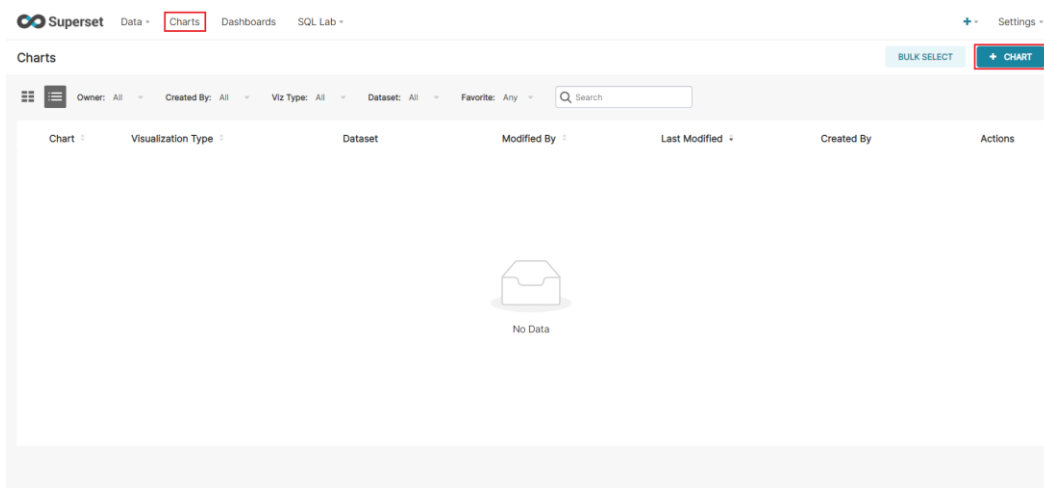


2) 命名并保存

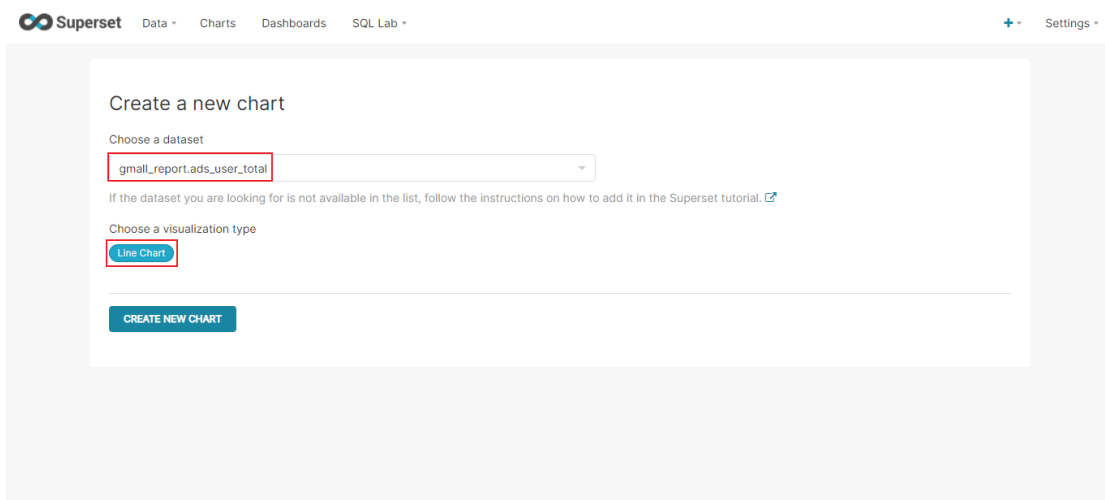


3.3.2 创建图表

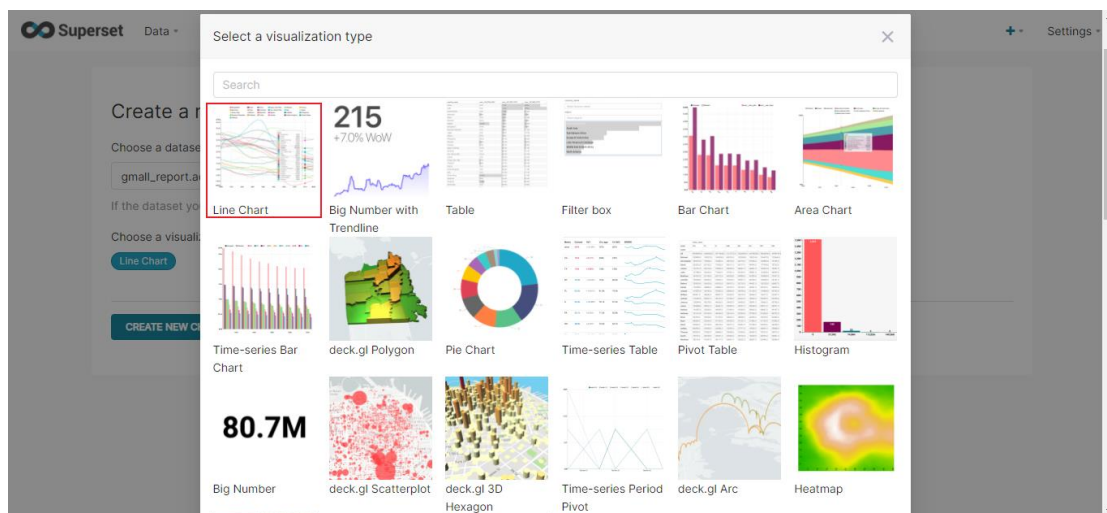
1) 点击 Charts/+CHART



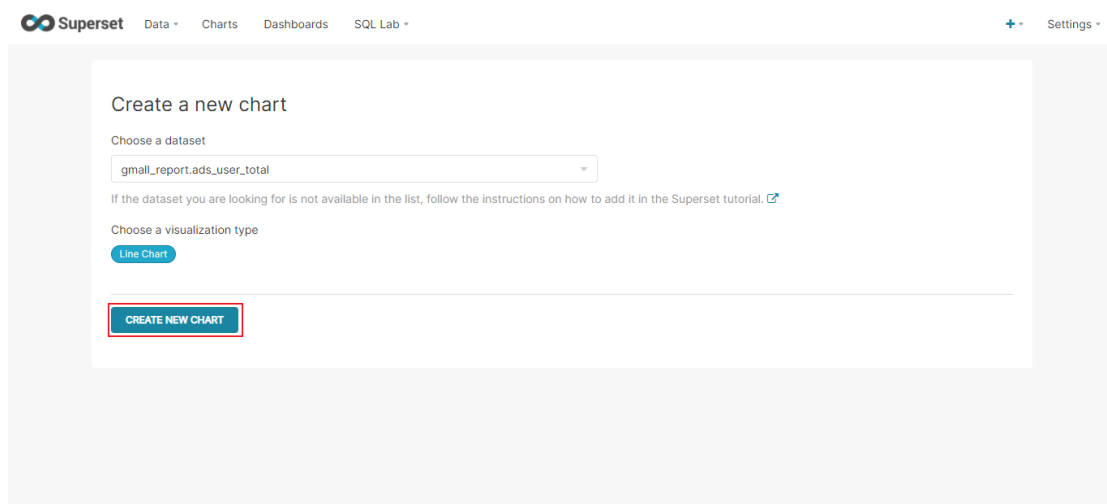
2) 选则数据源及图表类型



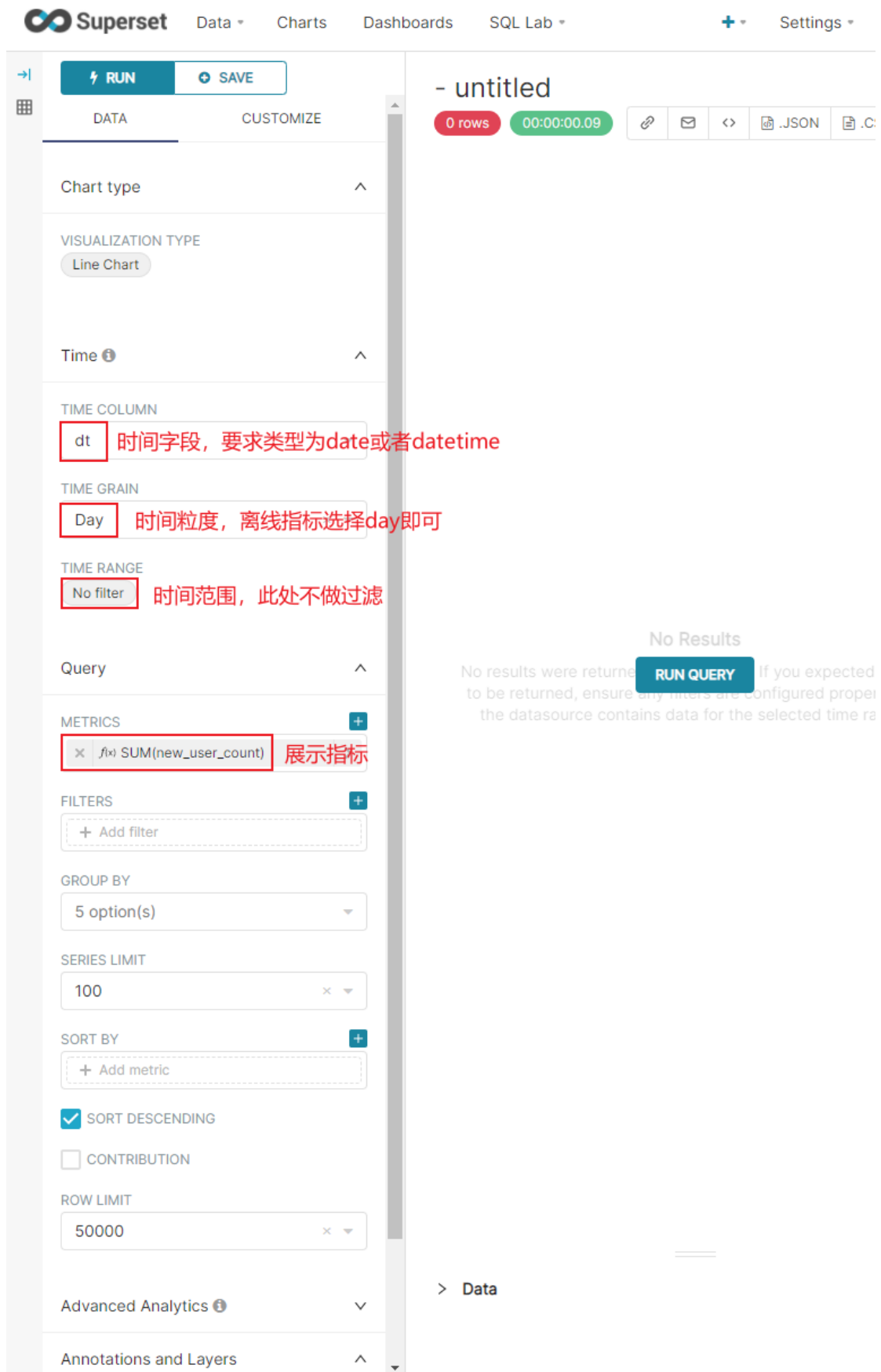
3) 选择何使的图表类型



4) 创建图表



5) 按照说明配置图表



The screenshot shows the Superset web interface. The left sidebar contains configuration sections: DATA, CUSTOMIZE, Chart type, VISUALIZATION TYPE (Line Chart), Time, TIME COLUMN (dt), TIME GRAIN (Day), TIME RANGE (No filter), Query, METRICS (SUM(new_user_count)), FILTERS, GROUP BY (5 option(s)), SERIES LIMIT (100), SORT BY (SORT DESCENDING), ROW LIMIT (50000), Advanced Analytics, and Annotations and Layers. The main panel shows a query result for 'untitled' with 0 rows and a 'RUN QUERY' button. A 'No Results' message is displayed. Red annotations highlight specific fields: 'dt' for TIME COLUMN, 'Day' for TIME GRAIN, 'No filter' for TIME RANGE, and 'SUM(new_user_count)' for METRICS.

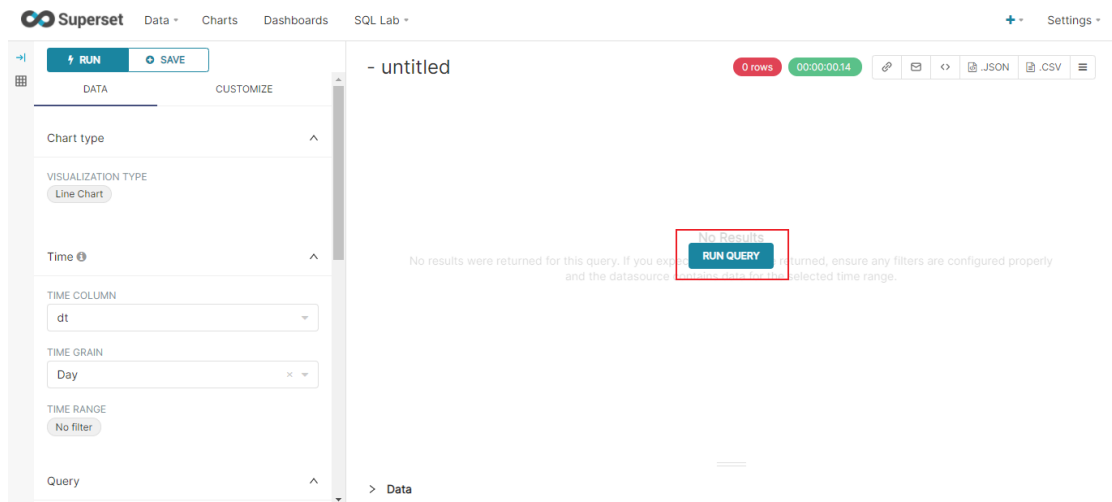
dt 时间字段，要求类型为date或者datetime

Day 时间粒度，离线指标选择day即可

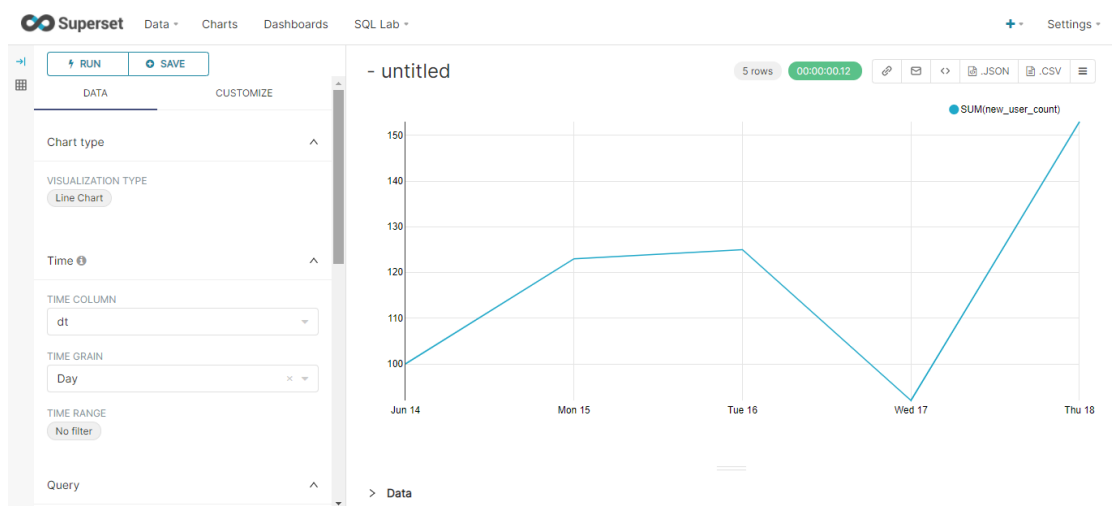
No filter 时间范围，此处不做过滤

展示指标

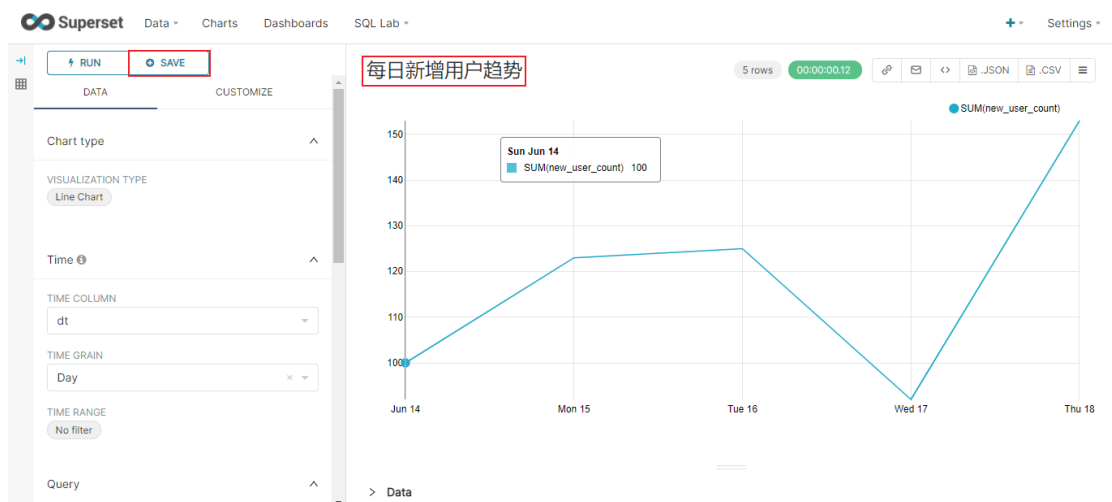
6) 点击 “Run Query”

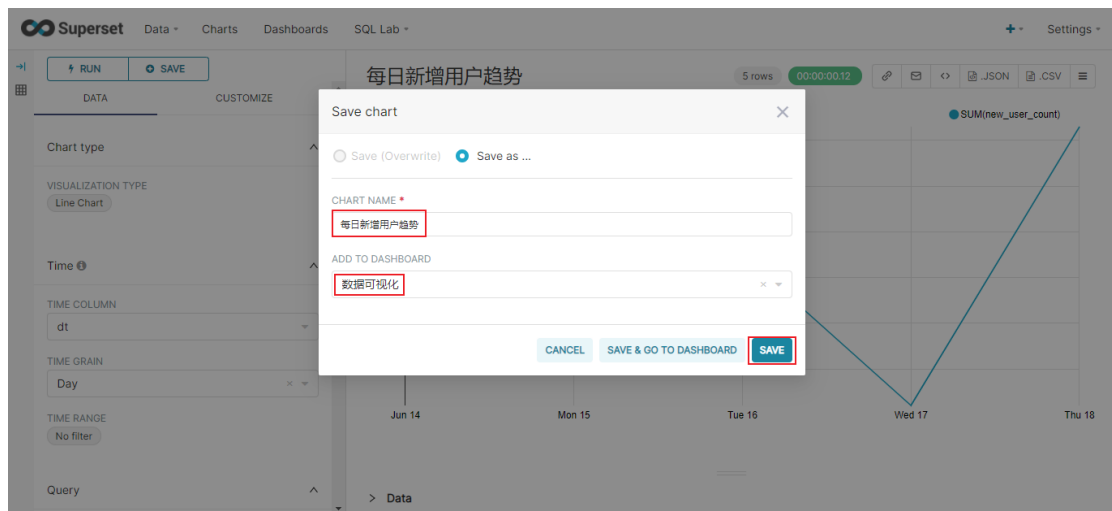


7) 如配置无误，可出现以下图表



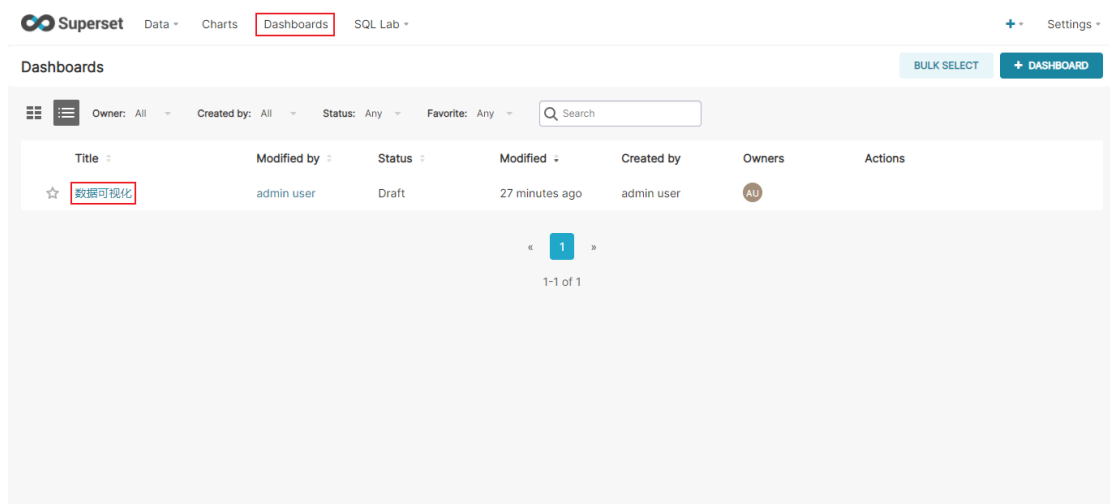
8) 命名该图表，并保存至仪表盘



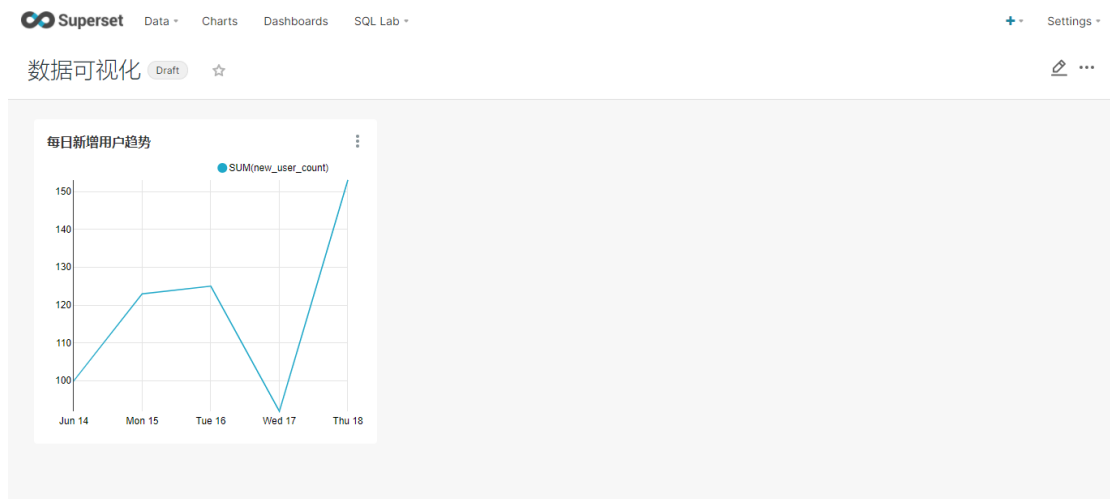


3.3.3 查看仪表盘

1) 点击“Dashboards”→“数据可视化”



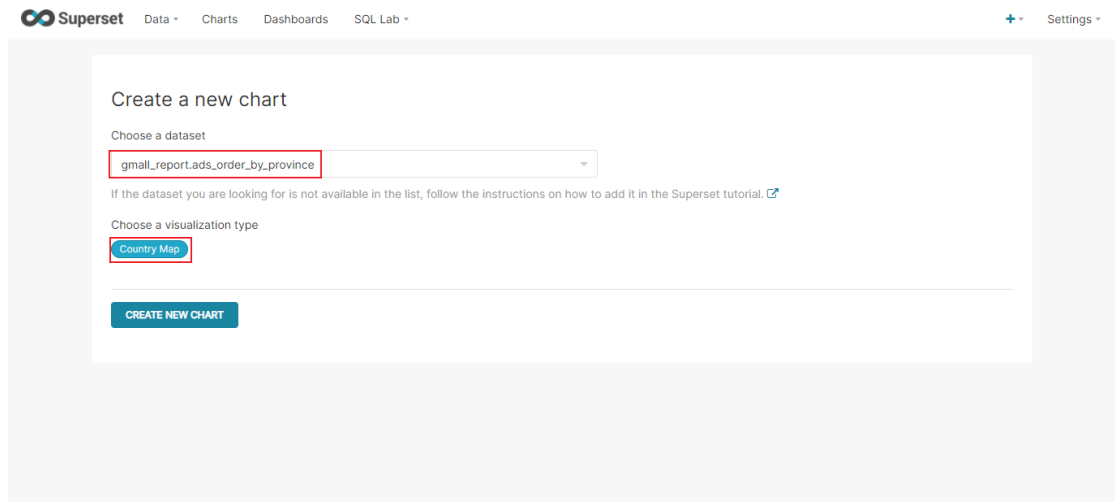
2) 查看仪表盘



第 4 章 Superset 实战

4.1 制作地图

1) 创建 Chart

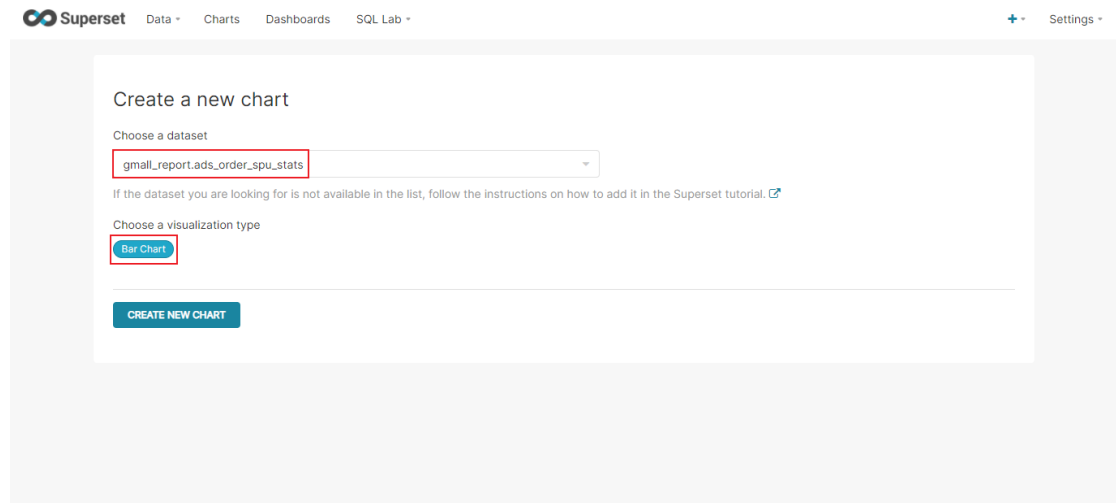


2) 配置 Chart

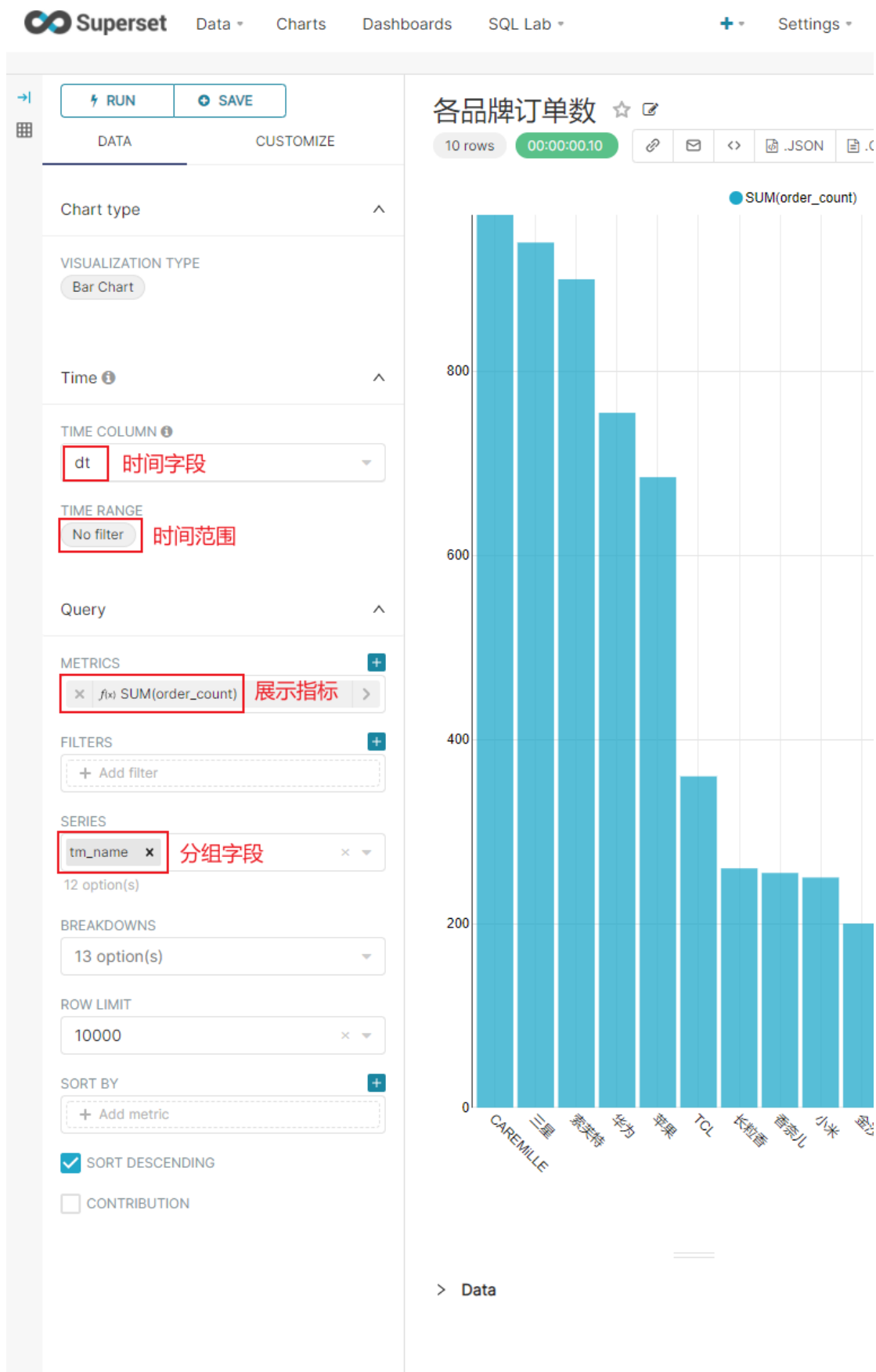


4.2 制作柱状图

1) 创建 Chart

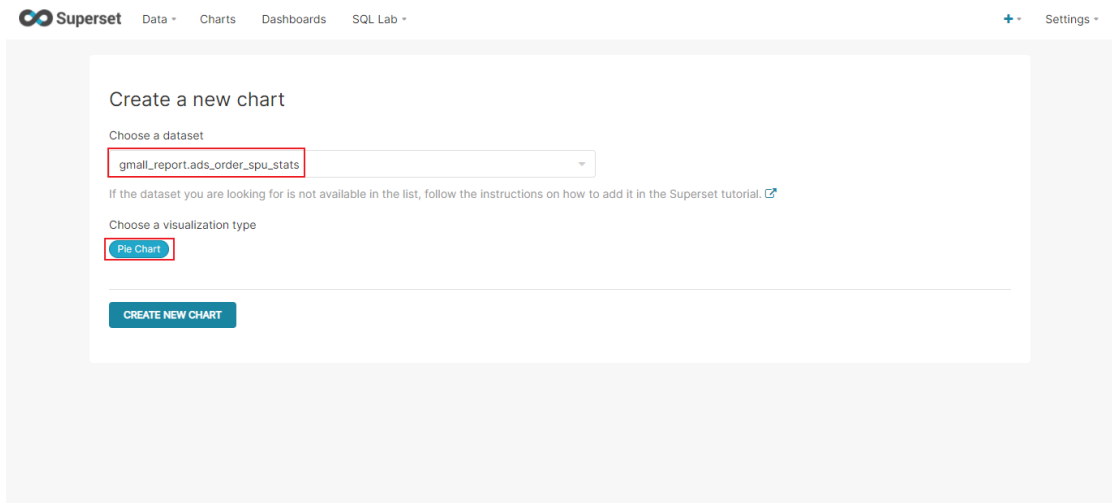


2) 配置 Chart

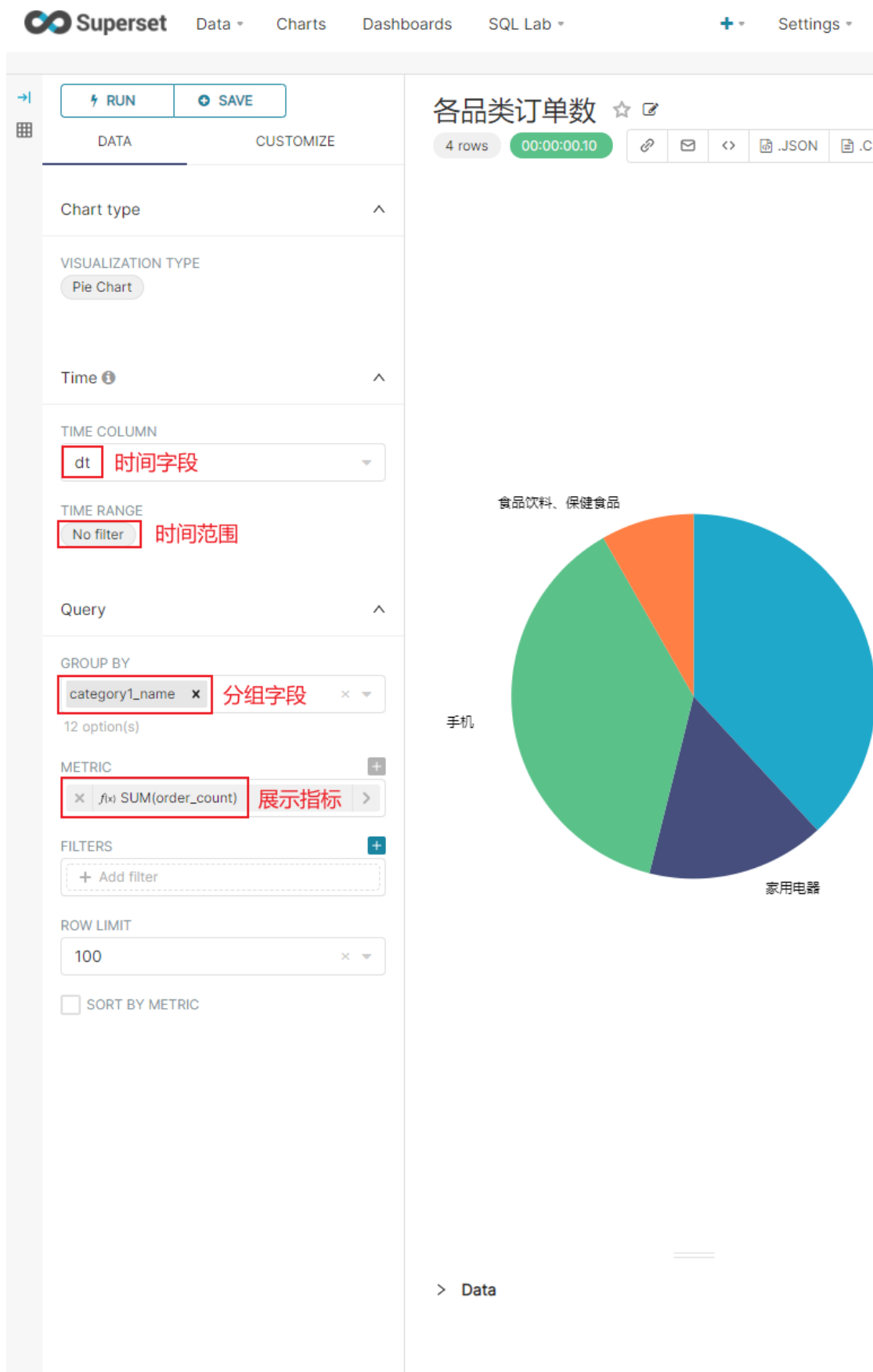


4.3 制作饼状图

1) 创建 Chart



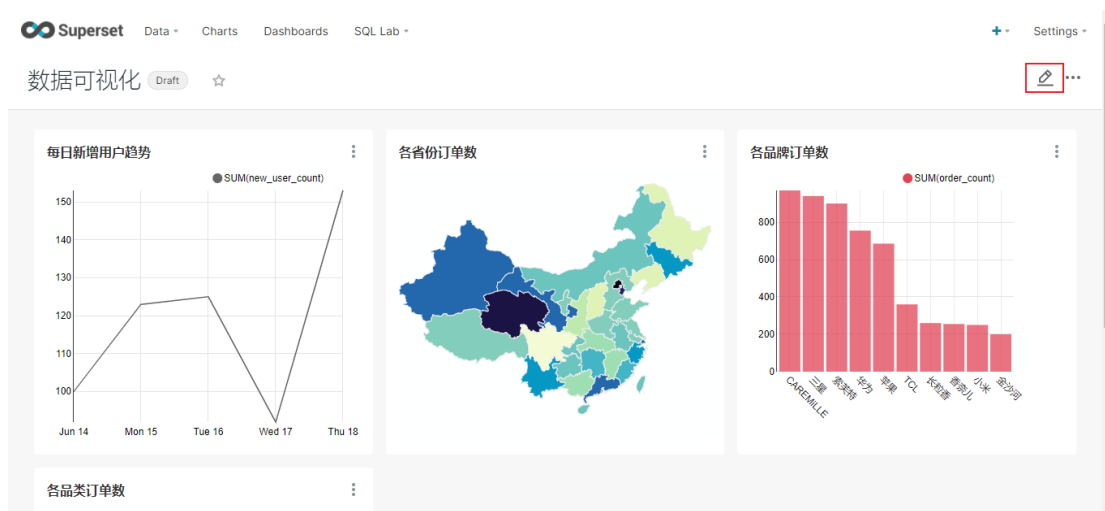
2) 配置 Chart



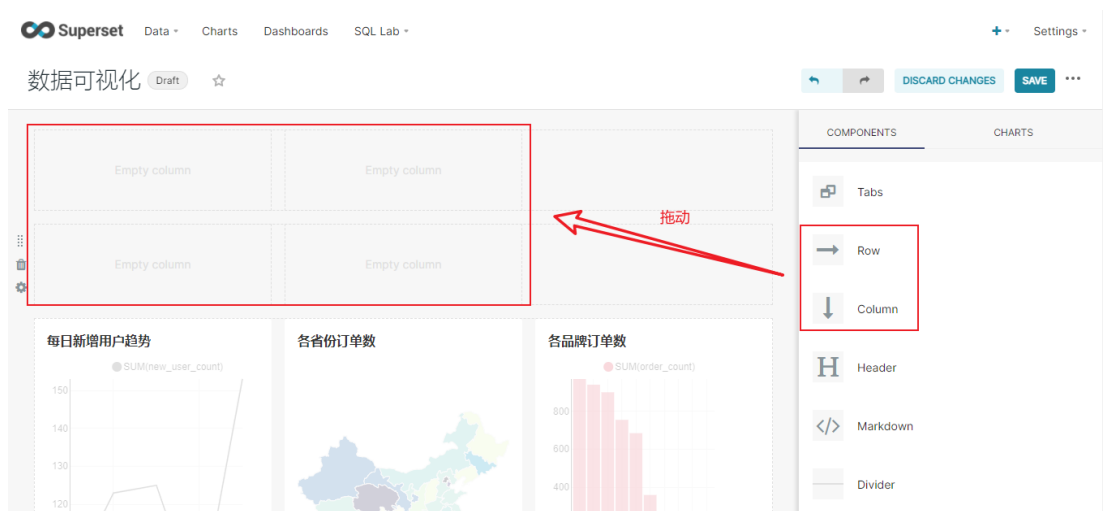
4.4 仪表盘布局

4.4.1 布局调整

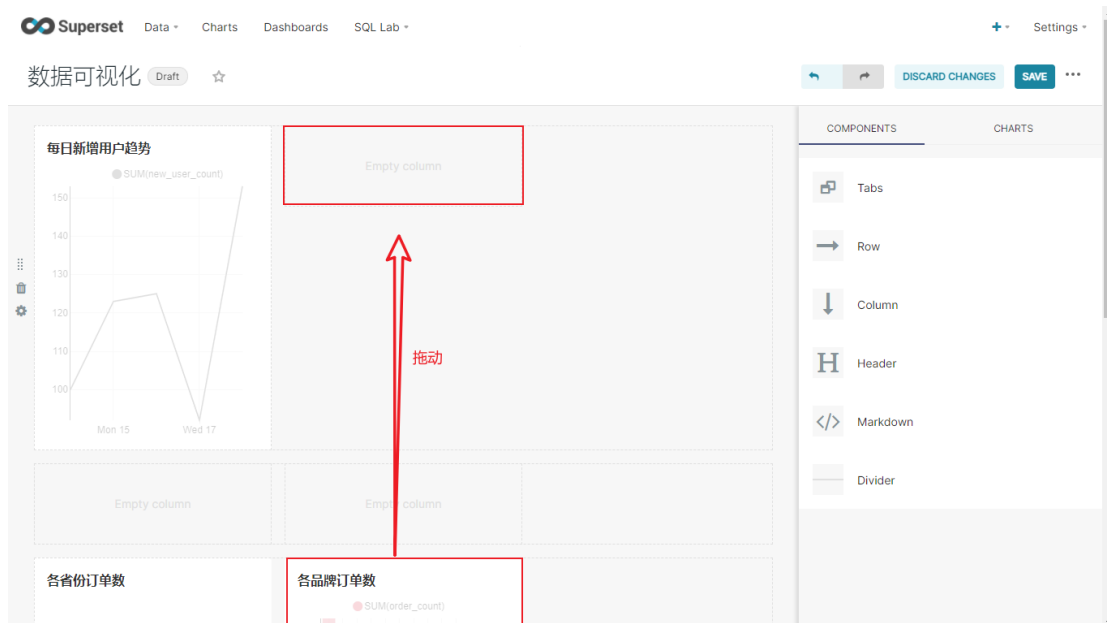
1) 点击编辑按钮



2) 使用行列组件预先布局



3) 拖动图表到指定坑位

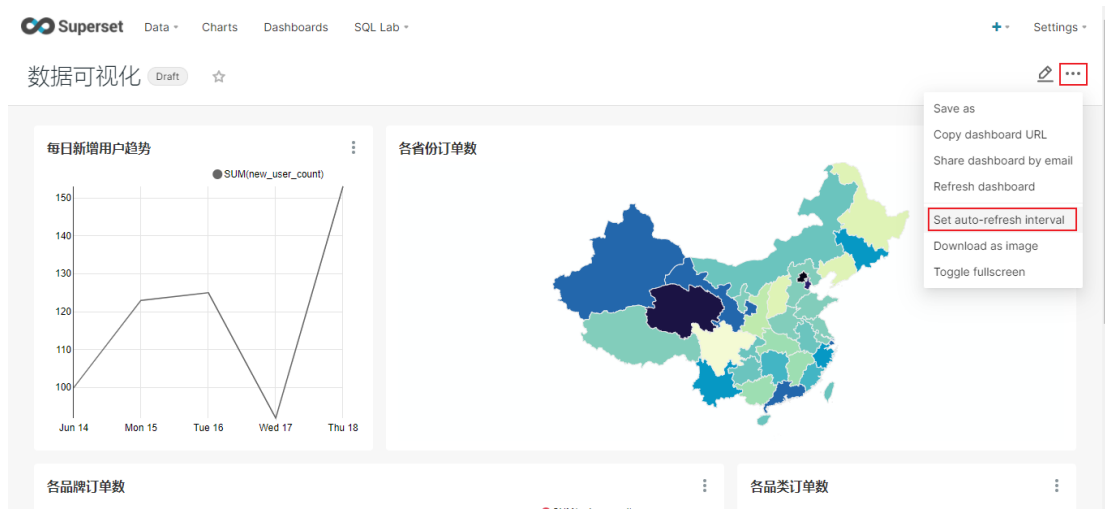


4) 最终结果

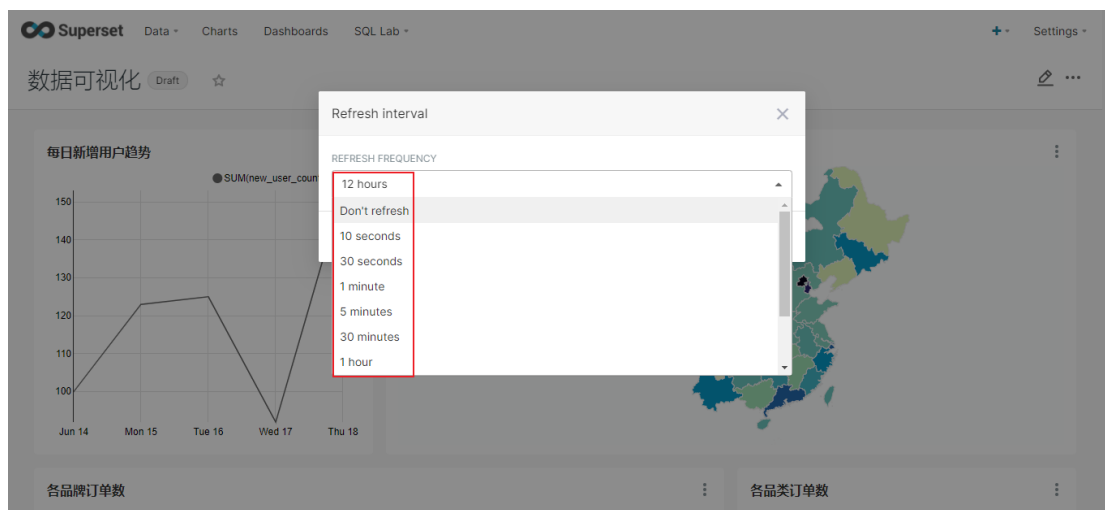


4.4.2 自动刷新

1) 点击配置按钮



2) 选择刷新时间间隔



3) 保存配置

