GREENPLUM DATABASE 腾讯云大学

# 六节课
# 快速上手Greenplum

第三节课

Greenplum生态与工具

线上直播 | 9月19日 14：00 - 15：00

GREENPLUM DATABASE®

微信技术讨论群
添加入群小助手：gp_assistant

钉钉群
钉钉客户端扫码入群

QQ群
群号：99194625

欢迎访问Greenplum中文社区：cn.greenplum.org
Github：https://github.com/greenplum-db/gpdb/

GREENPLUM DATABASE 腾讯云大学

# Greenplum 中文社区 公众号

- 技术研讨会PPT
- 活动预告
- 行业热点
- 技术干货
- Greenplum动态

GREENPLUM DATABASE 腾讯云大学

# Greenplum中文社区网站

https://cn.greenplum.org

博客 · 资料 · 文档 · 项目 · 活动

GREENPLUM DATABASE 腾讯云大学

# 六节课快速上手Greenplum

## 第三课

## Greenplum 生态与工具

辉鸿泛在CTO：阿福

# 目录

第一节 Greenplum 生态介绍

第二节 Greenplum 运维工具

第三节 Greenplum 管理工具

第四节 Greenplum BI/ETL工具

第五节 Greenplum 测试工具

# GPDB生态介绍

GREENPLUM DATABASE®

腾讯云大学

# 生态发展历程

GREENPLUM DATABASE 腾讯云大学

**2005**
- Greenplum数据库第一个版本发布
- 基于强大的开源数据库PostgreSQL

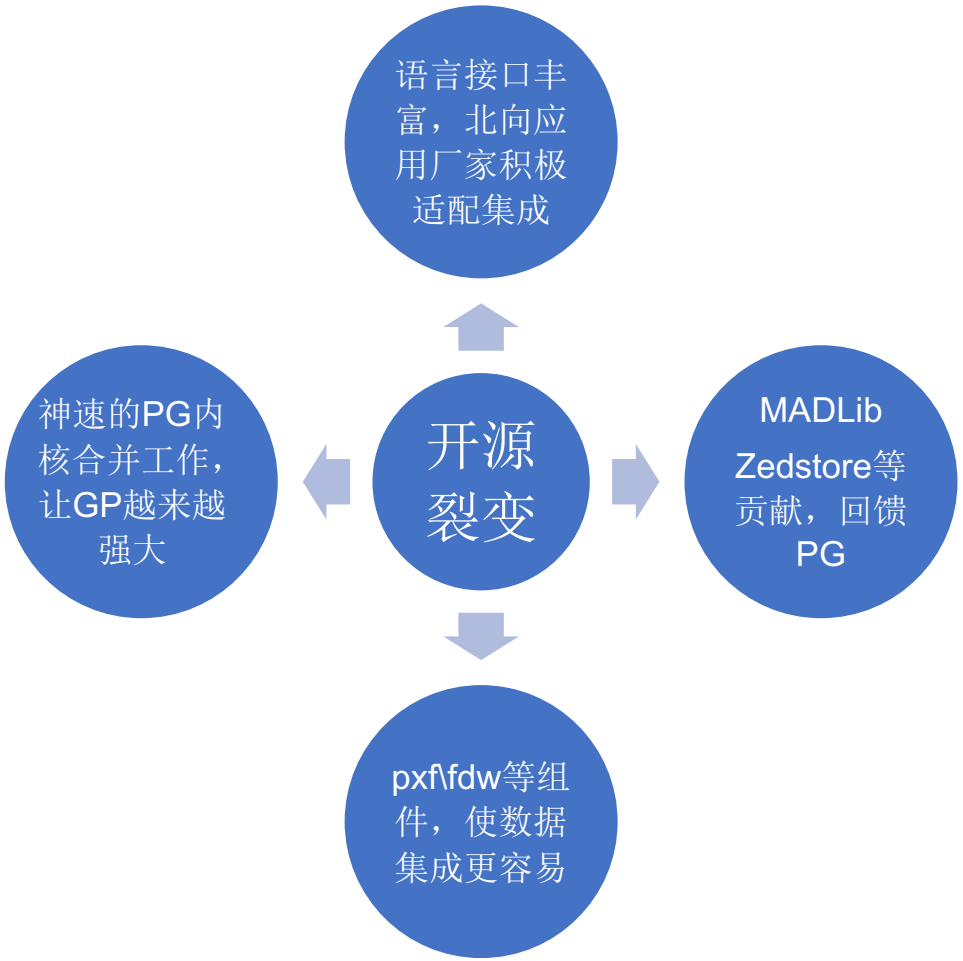
**2015**
- Greenplum开源，世界上第一款开源MPP数据库
- 开源版本基于Greenplum 4.3

**2017**
- Greenplum 5.0发布
- PostgreSQL内核由8.2升级到8.3，升级工作加速

**2018**
- Greenplum中文社区成立
- 专心运营Greenplum中国生态，服务广大爱好者

**2019**
- Greenplum 6.0发布
- PostgreSQL内核由8.3升级到9.4

**2021**
- 预计Greenplum 7.0发布
- 开源代码已经完成PostgreSQL12代码合并工作



# 商业版与开源版的异同

商业版本包含了开源版本的所有功能。此外，还包含如下内容：

- 支持 QuickLZ 压缩
- 支持如下的 data connectors：
  - Greenplum-Spark Connector
  - Greenplum-Informatica Connector
  - Greenplum-Kafka Connector（gpkafka）
  - Greenplum Stream Server（gpss）
- 支持 Data Direct ODBC/JDBC Drivers
- 支持 gpcopy
- 支持 gpcc
- 支持 gptext
- 支持原厂服务

商业版本下载地址：https://network.pivotal.io/products/pivotal-gpdb/

社区版本下载地址：https://github.com/greenplum-db/gpdb/releases

尊重商业
支持开源

# Greenplum中文社区生态

GREENPLUM DATABASE 腾讯云大学

**技术培训**
- 有偿/封闭式

**技术研讨会**
- 现场/开放式

**技术分享直播**
- 在线直播/回看

**优质文章分享**
- 使用经验沉淀

**社区实时沟通**
- 互助/共同成长

**问题反馈**
- 反馈/解决/积累

GREENPLUM DATABASE®

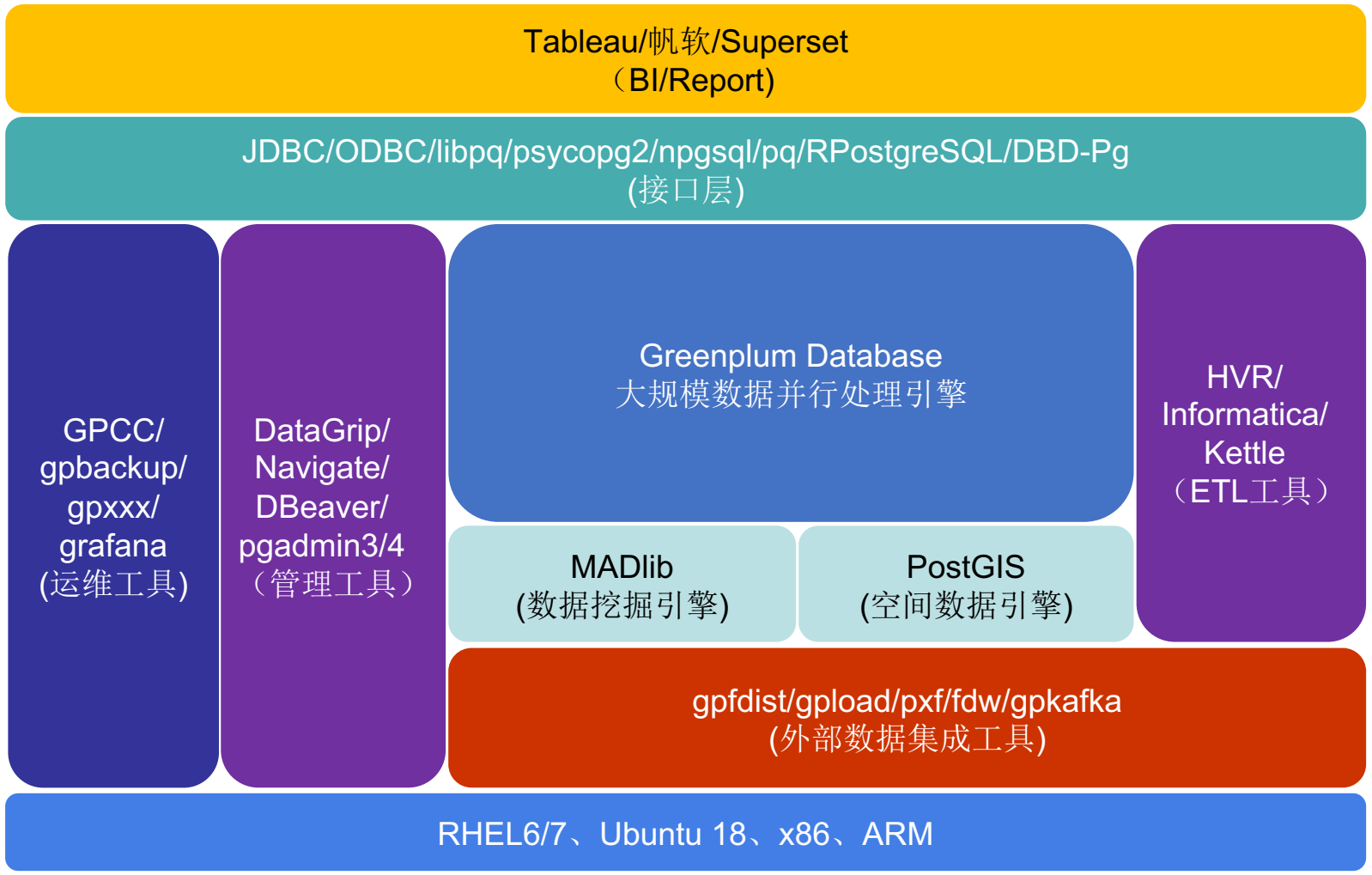
微信技术讨论群
添加入群小助手：gp_assistant

钉钉群
钉钉客户端扫码入群

QQ群
群号：99194625

欢迎访问Greenplum中文社区：cn.greenplum.org
Github：https://github.com/greenplum-db/gpdb/

# Greenplum生态软件



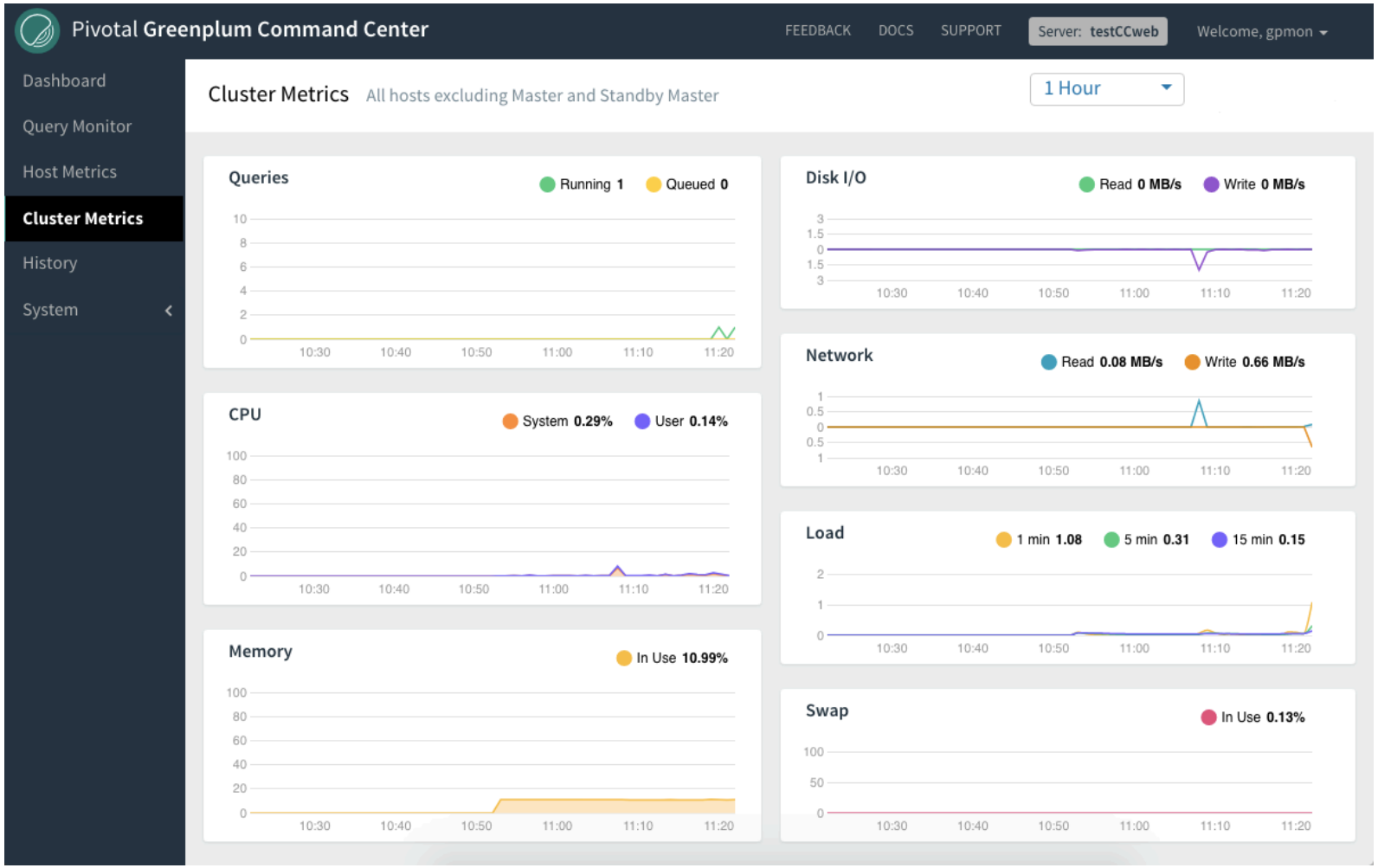

- Greenplum 生态完善。得益于对 SQL 标准的出色支持。
- 基于PostgreSQL，使其一出生便与众不同。
- 提供几乎所有语言的访问接口，像使用PG一样使用GP。
- 支持数据库内数据挖掘、支持空间数据引擎。
- 支持多种外部数据集成。

# GPDB运维工具

GREENPLUM DATABASE®

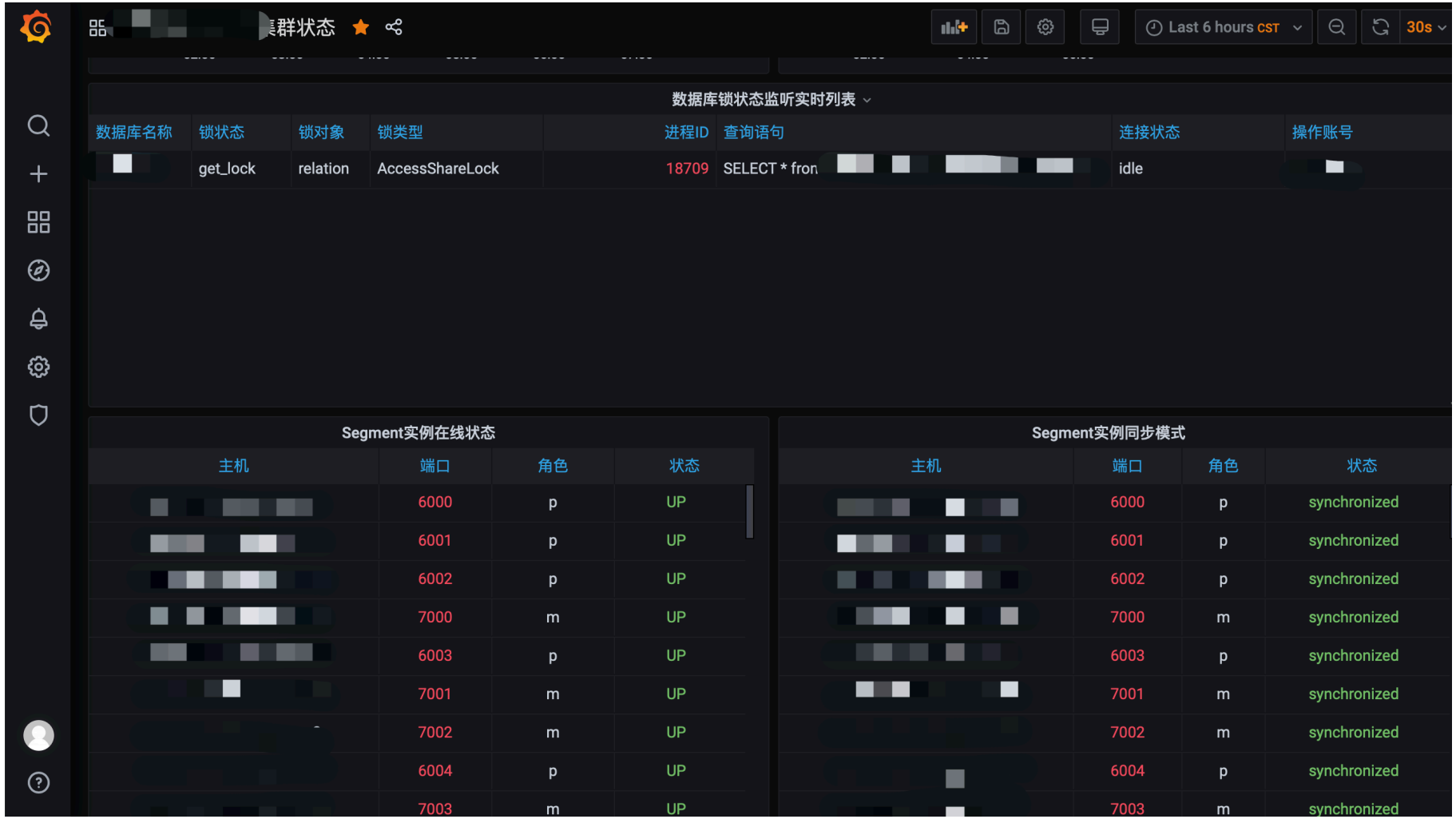
腾讯云大学

# 商业监控 - GPCC




- GPCC全称是Greenplum Command Center，属于企业版Greenplum的配套监控组件，不支持开源版本；
- 由于Pivotal开放的态度，大家可以免费下载GPDB和GPCC进行个人测试，出于对商业版权的保护，建议大家在选用企业版时，通过正规途径购买；
- GPCC提供数据库状态、机器负载情况、查询运行状态等多种参数的全面监控展示；
- 下载地址和文档地址如下：https://network.pivotal.io/products/gpdb-command-center http://gpcc.docs.pivotal.io/620/welcome.html

# 开源监控-Prometheus+Grafana




Prometheus+Grafana是目前较为主流的监控解决方案，很多大公司都基于此方案构建整体监控系统。

适配的GPOSS5/6 exporter地址如下：

https://github.com/ChrisYuan/greenplum_exporter/releases/tag/v2.0

该组件基本适配了一些常用的集群状态监控。

如有问题请留言交流

# gpxxx运维工具

**gpactivatestandby**

**作用：**用于激活一个备份主节点，使其代替原来的主节点对外提供服务
**场景：**主节点异常宕机
**常用命令：**
备节点机器上执行
gpactivatestandby -d $MASTER_DATA_DIRECTORY

**gpcheckcat**

**作用：**用于检测master和segment的catalog表并提供修复脚本
**场景：**周期性检测catalog一致性
**常用命令：**
gpcheckcat [dbname]

**gpaddmirrors**

**作用：**用于对现有集群增加镜像节点
**场景：**初始化集群不带有镜像，需要手工增加；需要指定灵活的镜像添加方式；
**常用命令：**
输出配置文件 - gpaddmirrors -o mirror_config_file
根据自定义镜像配置规则，修改配置文件
初始化镜像 - gpaddmirrors -i mirror_config_file

**gpcheckperf**

**作用：**用于在GPDB主机集群执行内存/网络/磁盘性能测试
**场景：**基础环境搭建好后，进行一遍性能验证；集群运行出现性能问题时，用该工具检测内存/网络/磁盘性能是否有所下降；
**常用命令：**
参见手册。

# gpxxx运维工具

GREENPLUM DATABASE 腾讯云大学

**gpconfig**

**作用：**用于修改集群配置参数
**场景：**修改配置参数；gpconfig修改出错集群启动失败，单独修改master配置并启动master节点重新配置集群
**常用命令：**
gpconfig -s xxx
gpconfig -c xxx -v xxx [-m xxx]

**gpinitstandby**

**作用：**用于初始化一个备份master
**场景：**；集群初始化后未增加standby master；standby master切换后，重新增加一个standby master
**常用命令：**
gpinitstandby -s host09

**gpfdist**

**作用：**用于并行数据加载
**场景：**数据批量高速入库
**常用命令：**
gpfdist -d /var/load_files -p 8081 &
配合数据库外部表一起使用

**gpexpand**

**作用：**用于扩展现有集群
**场景：**集群增加机器；集群增加节点
**常用命令：**
gpexpand -i input_file

**gpbackup/gprestore**

**作用：备份/恢复集群**
**场景：**数据备份及恢复场景
详情见：https://github.com/greenplum-db/gpbackup

# gpxxx运维工具

GREENPLUM DATABASE 腾讯云大学

**gpinitsystem**

**作用：**用于初始化集群
**场景：**集群初始化
**常用命令：**
gpinitsystem -c gpinitsystem_config \
-h hostfile_gpinitsystem --mirror-mode=spread

**gpload**

**作用：**用于并行数据加载，是对gpfdist的封装
**场景：**替换Oracle sqlloader；
**常用命令：**
gpload -f my_load.yml

**gpmovemirrors**

**作用：**用于将mirror移动到新的位置
**场景：**优化数据分布和存储；
**常用命令：**
gpmovemirrors -i move_config_file

**gprecoverseg**

**作用：**用于恢复故障节点
**场景：**集群拥有mirror，当有p或m宕机，但集群仍然可用
**常用命令：**
gprecoverseg
gprecoverseg -F

**gpssh-exkeys/gpssh/gpscp**

**作用：**设置免密登陆、批量执行命令、批量传输文件
**场景：**维护场景、集群初始化；
**常用命令：**
gpssh-exkeys -f hostfile_exkeys

**gpstart/gpstop/gpstate**

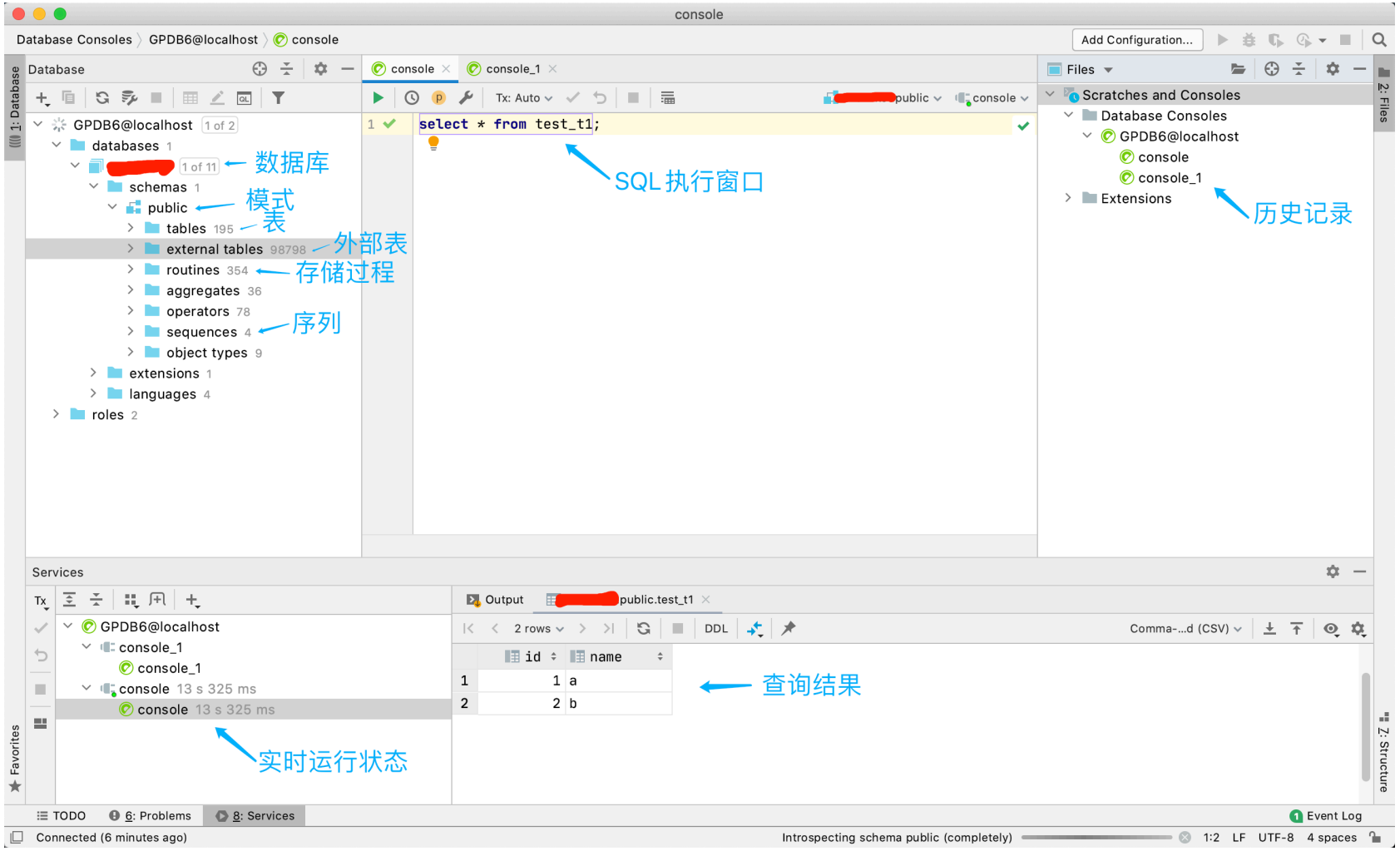
**作用：**启动集群/停止、重启、重载集群/查看集群状态
**场景：**；配置文件修改、访问入口文件修改后使生效；
**常用命令：**
gpstop -u/ gpstart / gpstop/ gpstate

# GPDB管理工具

GREENPLUM DATABASE®

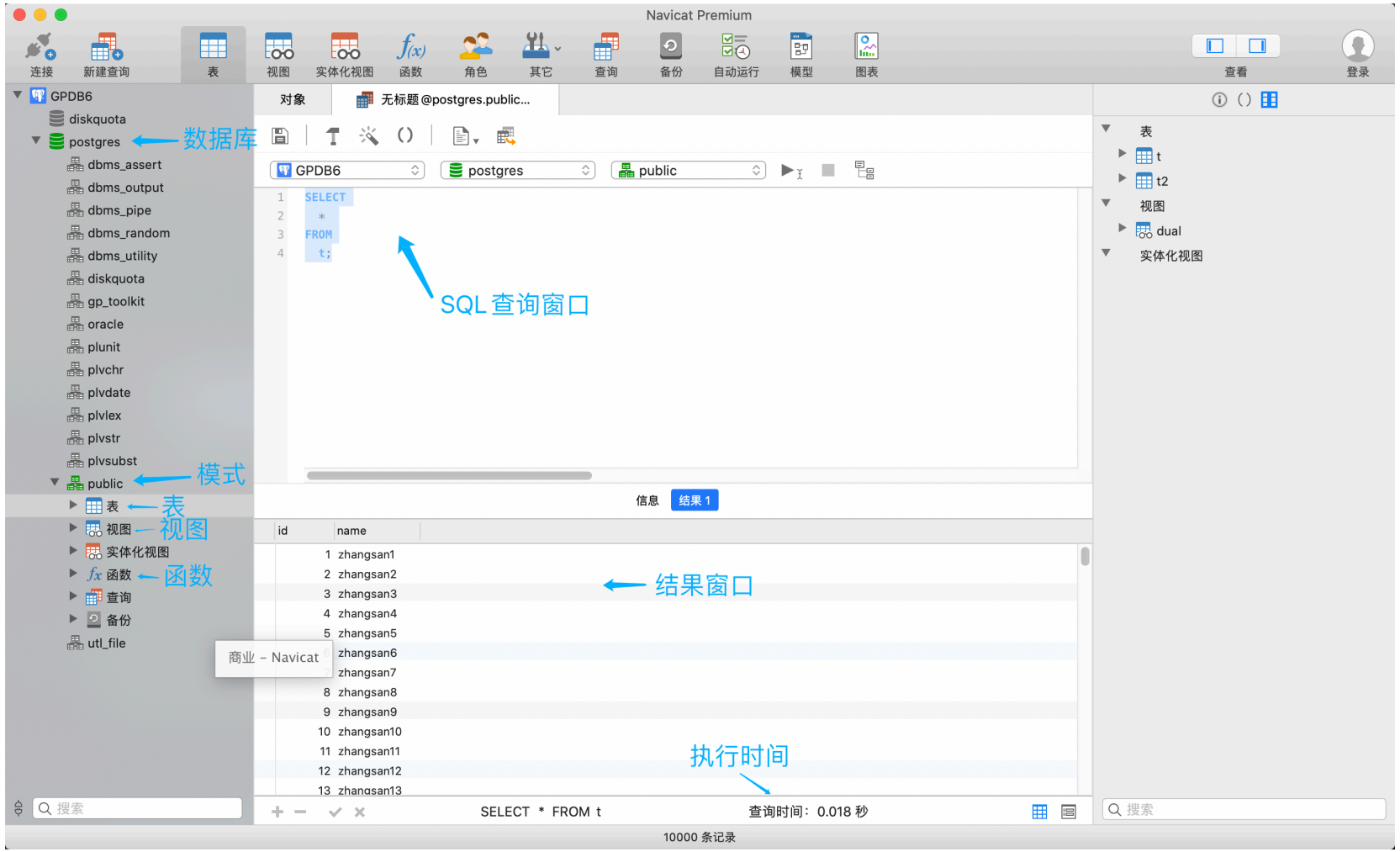
腾讯云大学

# 商业 - Datagrip/JetBrains数据库插件



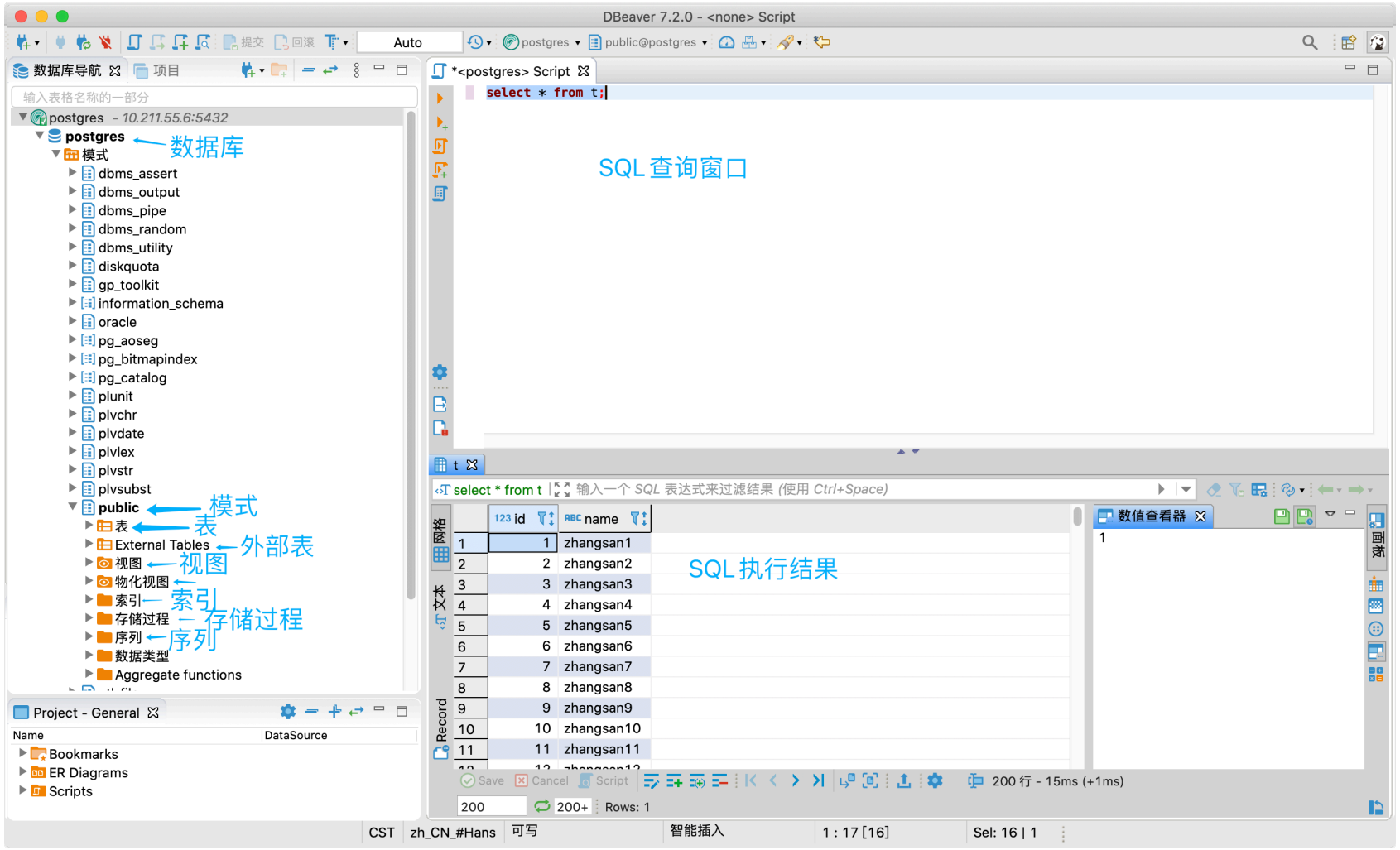

- JetBrains全家桶一直是开发者比较喜欢的工具。
- Datagrip是全家桶中一款专门用于管理数据库连接的工具；
- 支持市面上几乎所有的数据库：Oracle、MySQL、PostgreSQL、Hive、MongoDB、Clickhouse，重点是对Greenplum支持特别好；
- 支持Greenplum 5、6，支持存储过程编写、支持分区表；提供SQL美化功能、文本对比、列选择模式、字符编码识别提示、正则替换等方便功能；支持数据导出、数据库结构导出功能。
- 采用Java编写。

# 商业 - Navicat




- Navicat也是一款开发者比较喜欢的工具。
- 界面清爽，专注于数据库管理；
- 支持市面上主流数据库：Oracle、MySQL、PostgreSQL、MongoDB，连接GP需要采用PostgreSQL方式连接；
- 支持连接到Greenplum 5/6. 支持存储过程编写、不支持分区表；提供SQL美化功能、支持数据导出、数据库结构导出功能。
- 采用C++编写。

# 开源 - DBeaver




- DBeaver是一款流行的数据库管理工具；
- 源码地址：https://github.com/dbeaver/dbeaver
- 更新迭代频繁，支持市面上几乎所有的数据库：80余种；
- 支持Greenplum 5、6，支持存储过程编写、支持分区表；提供SQL美化功能；支持数据导出、数据库结构导出功能；
- 采用Java编写；

# 选型比较

| 产品名称 | 版权 | 兼容性 | 是否支持开源版本 | 函数 | 分区表 | 流畅性 | 资源占用 |
|---|---|---|---|---|---|---|---|
| **DBeaver** | 商业/开源 | GPDB5/6 | 支持 | 支持较好 | 支持 | 一般 | 较大 |
| **DataGrip** | 商业 | GPDB5/6 | 支持 | 支持很好 | 支持 | 较好 | 较大 |
| **Navicat** | 商业 | GPDB5/6 | 支持 | 支持一般 | 不支持 | 很好 | 一般 |

- 有条件的公司或个人，推荐优先选用DataGrip，如果您已经在用JetBrains家的其他工具如：IDEA，那么可以直接安装一个数据库插件即可；
- 崇尚开源免费的用户，推荐选用DBeaver，用起来功能也挺全，除了可以连接Greenplum，还可以连接几乎你所知道的所有数据库，如果不支持，您还可以自行扩展所需功能；
- Navicat、pgadmin等，只要支持PostgreSQL，都能连接上，但是存在各种不同情况的不兼容问题，只适合临时用用。

# GPDB BI/ETL

GREENPLUM DATABASE®

腾讯云大学

# 商业BI - Tableau

## 优化建议一：关掉cursor

默认情况下Tableau使用cursor模式从GPDB拉取数据。
如果提取数据量很大，并且服务器的内存足够放下所有的查询数据，可以关闭cursor的模式进行性能调优。
通过TDC文件关闭cursor模式，如下：

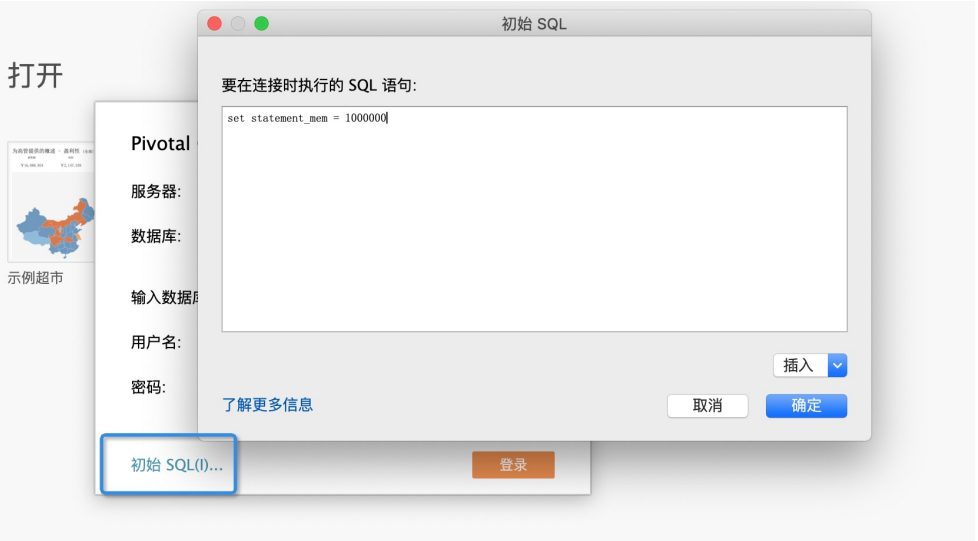
```
<?xml version='1.0' encoding='utf-8' ?>
<connection-customization class='greenplum' enabled='true' version='4.3'>
<vendor name='greenplum'/>
<driver name='greenplum'/>
<customizations>
<customization name='odbc-connect-string-extras' value='UseDeclareFetch=0' />
</customizations>
</connection-customization>
```

将该文件以tdc为后缀名，放到对应Datasources目录下，重启Tableau即可生效。
也可以修改fetch的size，让其每次fetch更多的数据：

```
<?xml version='1.0' encoding='utf-8' ?>
<connection-customization class='greenplum' enabled='true' version='4.3'>
<vendor name='greenplum'/>
<driver name='greenplum'/>
<customizations>
<customization name='odbc-connect-string-extras' value='Fetch=100000' />
</customizations>
</connection-customization>
```

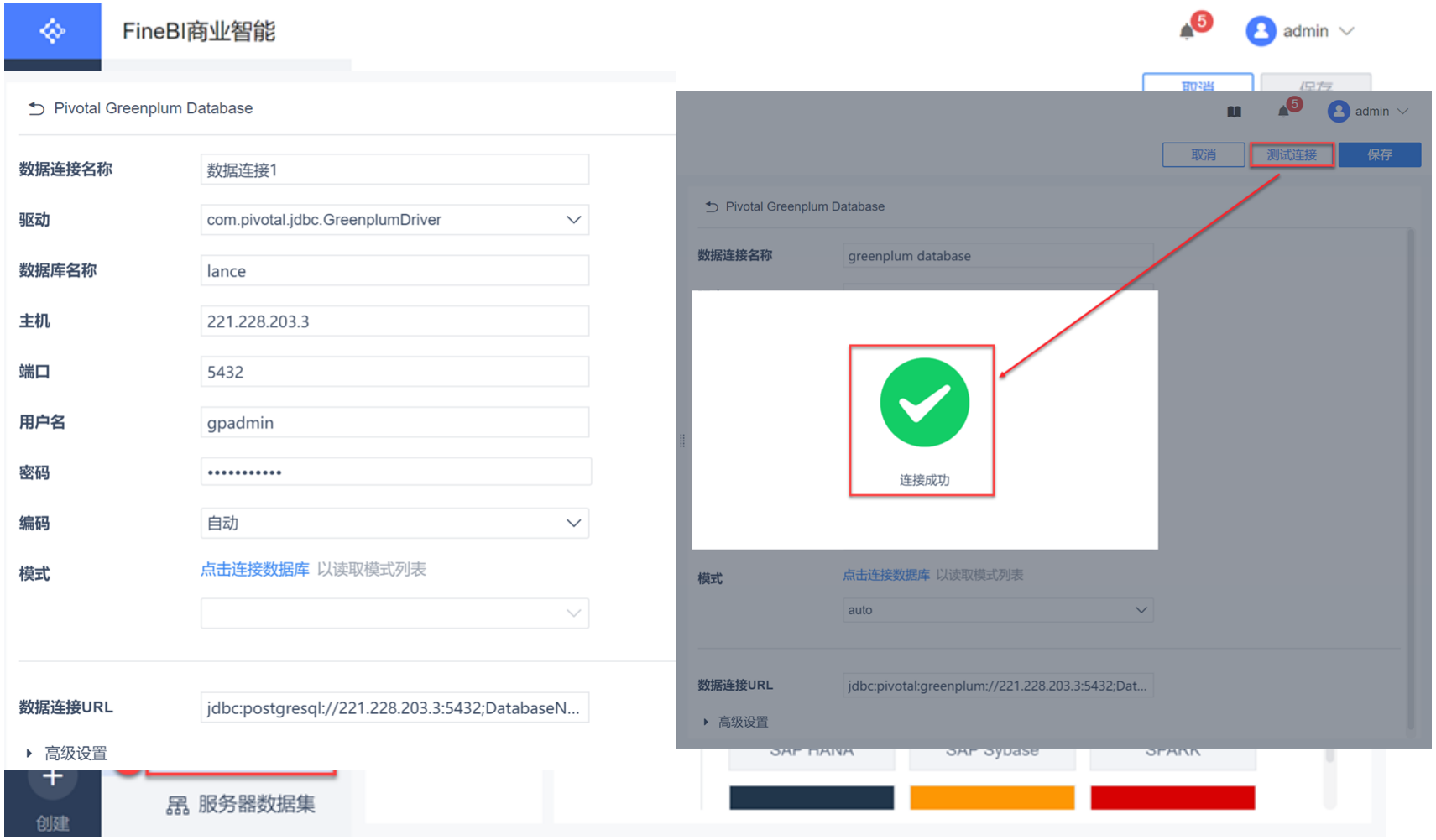
## 优化建议二：通过设置初始化SQL修改默认参数

连接建立时可以通过初始化SQL设置特定参数，例如：设置statement_mem的值



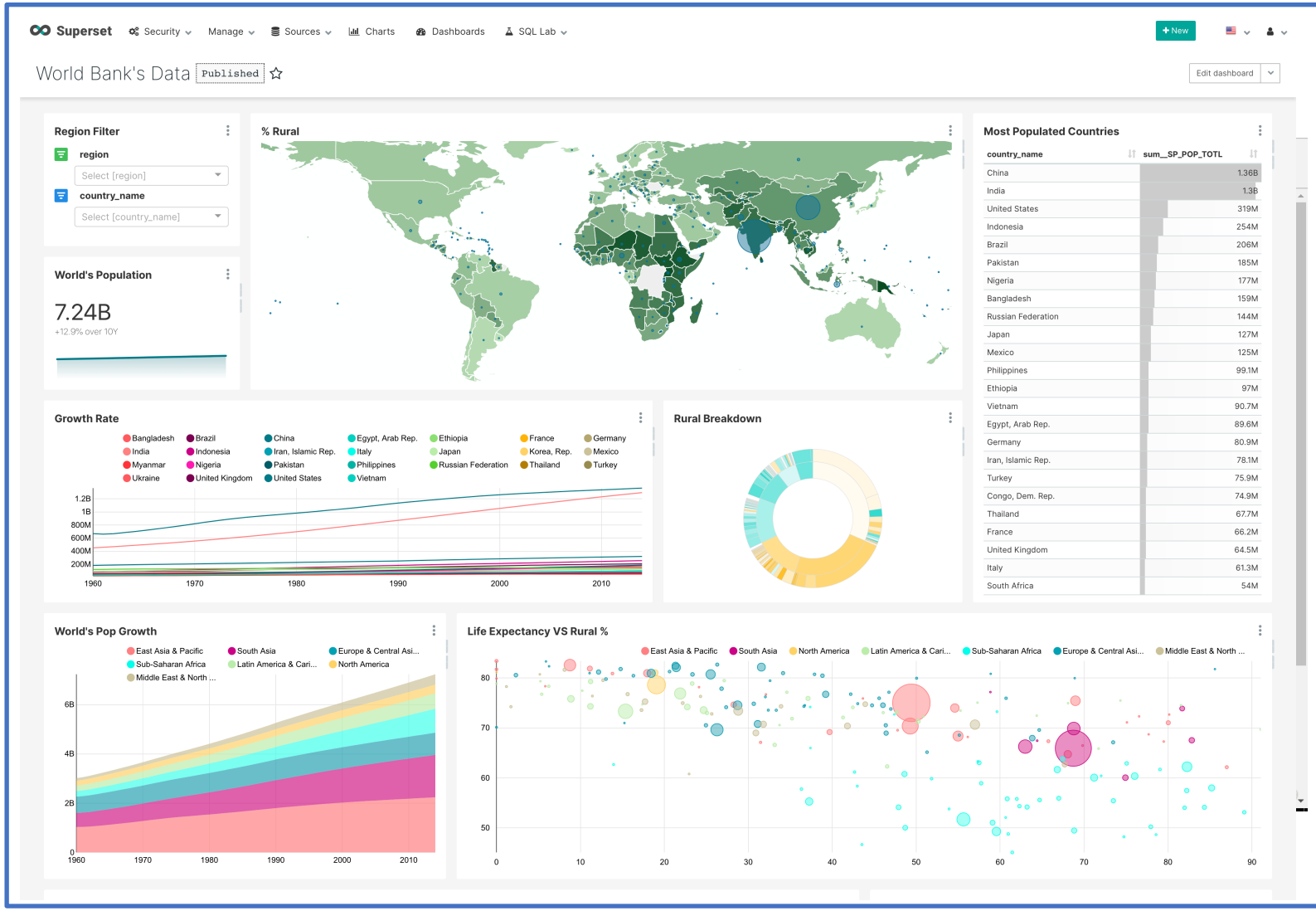

SQL后面不要带‘;’，否则执行会报错，因为Tableau会将该SQL封装执行，中间如果有分号会报语法错误。同样在自定义SQL时，SQL结尾也不能加‘;’。

# 商业BI - 帆软




- FineBi是国产商业BI报表软件中最好的一家，从各个方面来讲，都推荐大家尝试一下，可以在官网申请试用；
- 支持连接到Greenplum；
- 可以访问官方论坛获取更多连接信息：https://help.finebi.com/doc-view-289.html

# 开源BI - Superset




- Superset是爱彼迎开源的一款可视化工具；用Python编写；
- 详细信息和源代码参考：

https://superset.apache.org

https://github.com/apache/incubator-superset

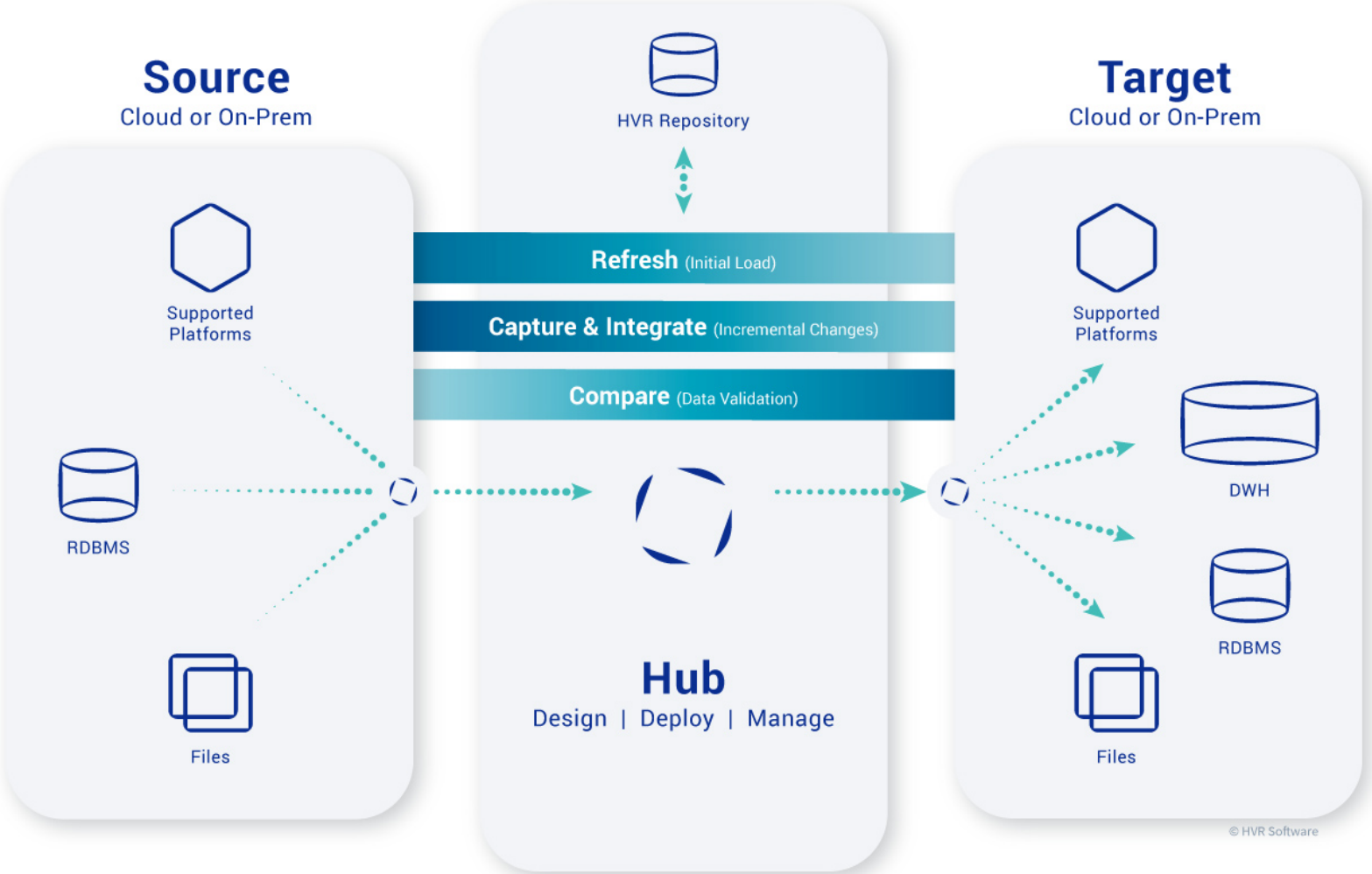
- 进入Superset后，可以通过提供的postgres连接方式连接到GPDB；然后将相应的数据库表引入到Superset，再与合适的图进行关联展示；
- 提供Docker一键式体验方式，方便大家体验功能和选型对比；

# 选型比较

| 产品名称 | 版权 | 兼容性 | 是否支持开源版本 | 可视化 | 可用性 | ETL能力 | 价格 |
|---|---|---|---|---|---|---|---|
| **Tableau** | 商业 | GPDB5/6 | 支持 | 非常好 | 大数据量变慢 | 无 | 较贵 |
| **FineBI** | 商业/个人免费 | GPDB5/6 | 支持 | 非常好 | 较好 | 有 | 灵活 |
| **Superset** | 开源 | GPDB5/6 | 支持 | 较好 | 较慢、结构复杂 | 无 | 免费 |

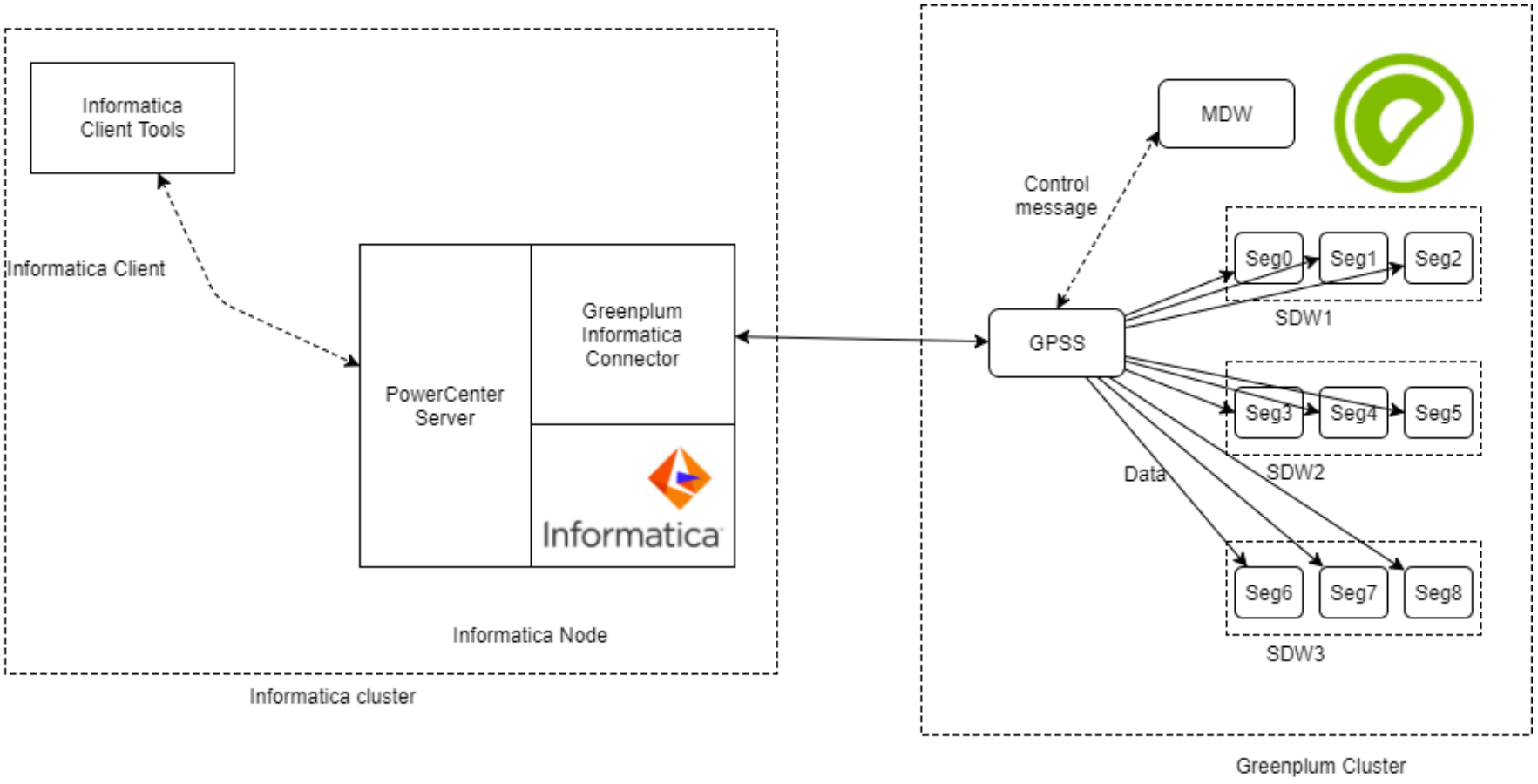
- BI工具作为辅助决策、可视化展示的工具，是面对管理层和客户的，个人建议国内用户可以优先选择使用帆软FineBI产品，它在迎合大众口味、提高用户体验方面做了很多工作。当然Tableau作为一个大品牌，也是值得信赖的；但是相较来说，我更倾向于FineBI；
- Superset算是开源工具中比较好的一个选型点了，其他的Redash、国内宜信开源的达芬奇，均可以尝试，个人感觉需要投入特别大的精力才能玩转这些开源可视化平台；
- 选型对比参考：https://blog.csdn.net/yuanziok/article/details/86467462。

# 商业ETL - HVR



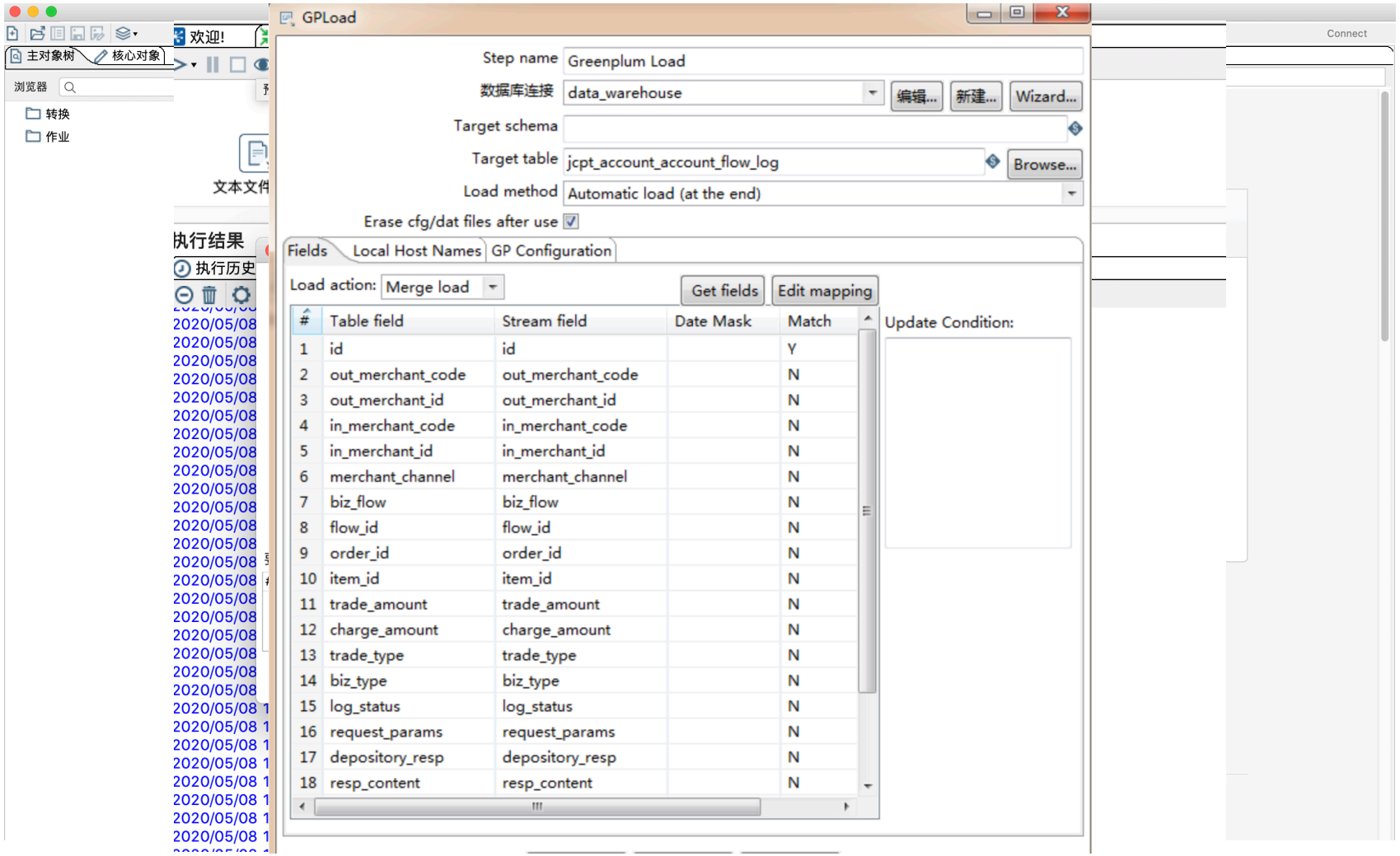

- HVR是一款集中式的数据同步工具，支持多源多目的；
- 支持GPDB 5/6数据库作为目标端数据库；
- 支持各种常用关系型数据库、文件作为数据源；支持关系型数据库、文件、NoSQL、NewSQL作为目标端；
- 支持一对一、一对多、多对一、多对多的数据传输方式；支持自定义ETL逻辑；
- 数据初始化及入库，均采用相关数据库最高效方式。

# 商业ETL - Informatica




- Informatica是一款成熟的ETL工具，在国内商业市场上占比高，易用性稳定性都很高；
- Greenplum商业版本，提供Informatica Connector，通过该连接器，可以充分结合Informatica的开发能力和Greenplum的并行处理能力；
- 针对开源版本，可以尝试采用postgres的odbc进行连接。

# 开源ETL - Kettle




- Kettle是一款老牌的开源ETL工具，深受大家喜爱；
- 开源版本，提供可视化脚本编写界面；调度可以通过开源软件kettle-scheduler进行；
- 参考网址：

https://github.com/pentaho/pentaho-kettle

https://github.com/zhaxiaodong9860/kettle-scheduler

- Kettle入库GPDB：

  insert

  copy

  gpload

# 选型比较

| 产品名称 | 版权 | 兼容性 | 是否支持开源版本 | 图形化 | 功能 | 性能 | 价格 |
|---|---|---|---|---|---|---|---|
| **Informatica** | 商业 | GPDB5/6 | 支持 | 很好 | 完整 | 高 | 较贵 |
| **HVR** | 商业 | GPDB5/6 | 支持 | 较好 | 较完整 | 很高 | 灵活 |
| **Kettle** | 开源 | GPDB5/6 | 支持 | 很好 | 完整 | 低 | 免费 |

- Informatica和HVR作为成熟的商业软件，对各种数据库都有很好的支持，相比较而言，HVR在数据高性能同步、CDC领域更胜一筹；Informatica优势则在于其复杂业务处理和图形化的优势；
- 作为开源ETL工具的Kettle，已经经过多方认可，目前绝大多数非付费场景的ETL工具，基本都是Kettle，也有很多公司通过封装Kettle，形成了自己的ETL工具产品。Kettle的劣势是比较重。

# GPDB测试工具

GREENPLUM DATABASE®

腾讯云大学

# tpc.org

- 事务处理性能委员会（Transaction Processing Performance Council），是由数10家会员公司创建的非盈利组织，它的功能是制定商务应用基准程序（Benchmark）的标准规范、性能和价格度量，并管理测试结果的发布；
- 目前有10几项相关度量标准；
- 与关系型数据库相关性比较大的有：
  - TPC-H
  - TPC-C
  - TPC-DS

# TPC-H/TPC-DS与TPC-C

- TPC-H通常用于OLAP测试。在一些客户要求的TPC-H测试场景下，客户通常会对我们提供的测试脚本进行理论验证，保证多个产品的TPC-H测试的一致性。该测试包含22个复杂查询。目前针对Greenplum的测试，我们通常采用PG社区德哥分享的脚本和逻辑进行测试，参考：

  https://github.com/digoal/gp_tpch

  https://developer.aliyun.com/article/93

  另外TPC-H测试可以采用hammerdb进行。

- TPC-DS也是用于OLAP测试的，但是测试逻辑更为复杂，一共有99个查询，有很多新兴数据库目前为止还无法满足所有99个查询的语法。目前在GPDB上的TPC-DS测试，建议参考：

  https://github.com/pivotalguru/TPC-DS 提供的测试脚本进行

- TPC-C通常用于OLTP测试，作为数据库事务处理性能的一个衡量标准。目前在GPDB上的TPC-C测试，可以采用开源测试工具BenchmarkSQL进行；

  https://sourceforge.net/projects/benchmarksql/

  编译安装及注意使用注意事项，可以参考解压目录下的使用说明文档。

# HammerDB

```
HammerDB CLI v3.2
Copyright (C) 2003-2019 Steve Shaw
Type "help" for a list of commands
The xml is well-formed, applying configuration
hammerdb>vucreate
Vuser 1 created - WAIT IDLE
Logging activated
to /tmp/hammerdb_5E4FAA4159F103E233830313.log
1 Virtual Users Created

hammerdb>vurun
RUNNING - PostgreSQL TPC-H
Vuser 1:RUNNING
Vuser 1:Executing Query 14 (1 of 22)

Vuser 1:query 14 completed in 0.786 seconds
Vuser 1:Executing Query 2 (2 of 22)
Vuser 1:query 2 completed in 4.107 seconds
Vuser 1:TPCH SCHEMA COMPLETE
Vuser 1:FINISHED SUCCESS
ALL VIRTUAL USERS COMPLETE
                    TPC-H Driver Script
```

- 获取地址：

https://www.hammerdb.com/index.html

https://github.com/TPC-Council/HammerDB

- 执行步骤：

1.安装hammerdb

2.进入控制台

3.进行配置

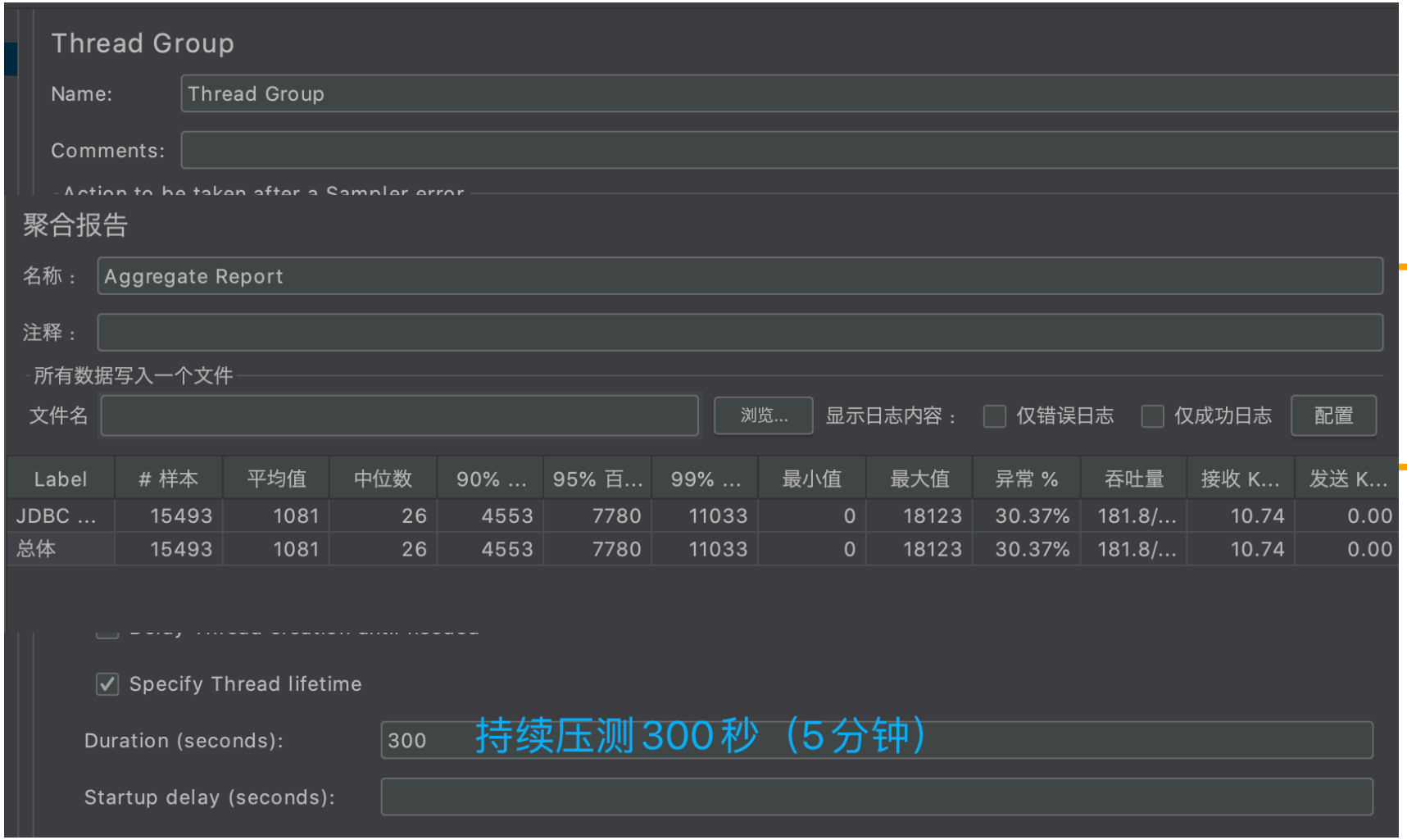
4.构建测试库

5.加载配置脚本

6.构建测试

7.运行测试

# JMeter




| Label | # 样本 | 平均值 | 中位数 | 90% ... | 95% 百... | 99% ... | 最小值 | 最大值 | 异常 % | 吞吐量 | 接收 K... | 发送 K... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| JDBC ... | 15493 | 1081 | 26 | 4553 | 7780 | 11033 | 0 | 18123 | 30.37% | 181.8/... | 10.74 | 0.00 |
| 总体 | 15493 | 1081 | 26 | 4553 | 7780 | 11033 | 0 | 18123 | 30.37% | 181.8/... | 10.74 | 0.00 |

- 获取地址：

https://jmeter.apache.org

- 执行步骤：

1.安装jmeter
2.运行jmeter
3.新建测试计划
4.新建测试线程组
5.增加驱动
6.新建jdbc连接
7.添加jdbc请求
8.添加报告
9.运行测试

# 感谢观看

Greenplum中文社区公众号

Greenplum微信技术讨论群

GREENPLUM DATABASE

腾讯云大学

GREENPLUM DATABASE 腾讯云大学

# Q&A