Greenplum Extensions 概览

王晓冉

VMware 研发工程师

©2022 VMware, Inc.

VMware Tanzu Greenplum

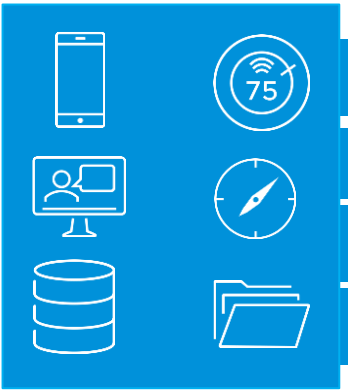
Greenplum是全球首个开源，专为分析打造，可多云部署的大数据平台。它在开源的 PG (PostgreSql) 的基础上采用 MPP 架构（Massive Parallel Processing, 海量并行处理），具有强大的大规模数据分析处理能力。

Greenplum拥有完整的产品生态，可以满足用户对例如时序数据、空间数据、全文检索等众多分析场景的需求。

GREENPLUM DATABASE

vmware ©2022 VMware, Inc.

VMware Tanzu Greenplum



ETL , Streaming

JSON , XML , Graph

IoT Data

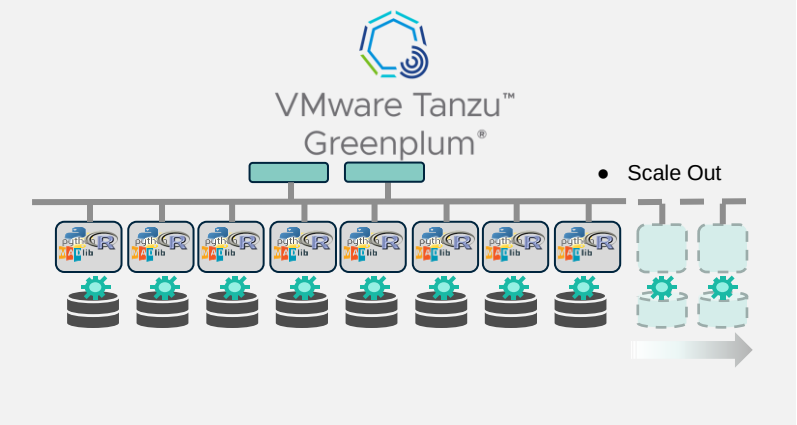
Voice, Video, Image

Geo Spatial , Time Series

And many more…

Data Sources

Structured, Semi, Un-Structured Data



All Data Types

Data Ingestion

kafka, RabbitMQ, Informatica, talend

Data Science

PostGIS, MADlib, K, SQL, jupyter, R, Java, python

SQL, R, Python, Java,C…

Data Scientists

Business Users

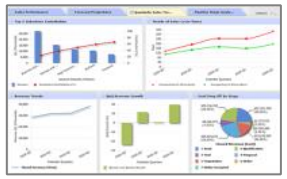
Data Analysts

IT Professionals

BI/Analytic/ML …

End-Users

Data Visualization



vmware ©2022 VMware, Inc.

Greenplum 数据库生态系统

vmware ©2022 VMware, Inc.

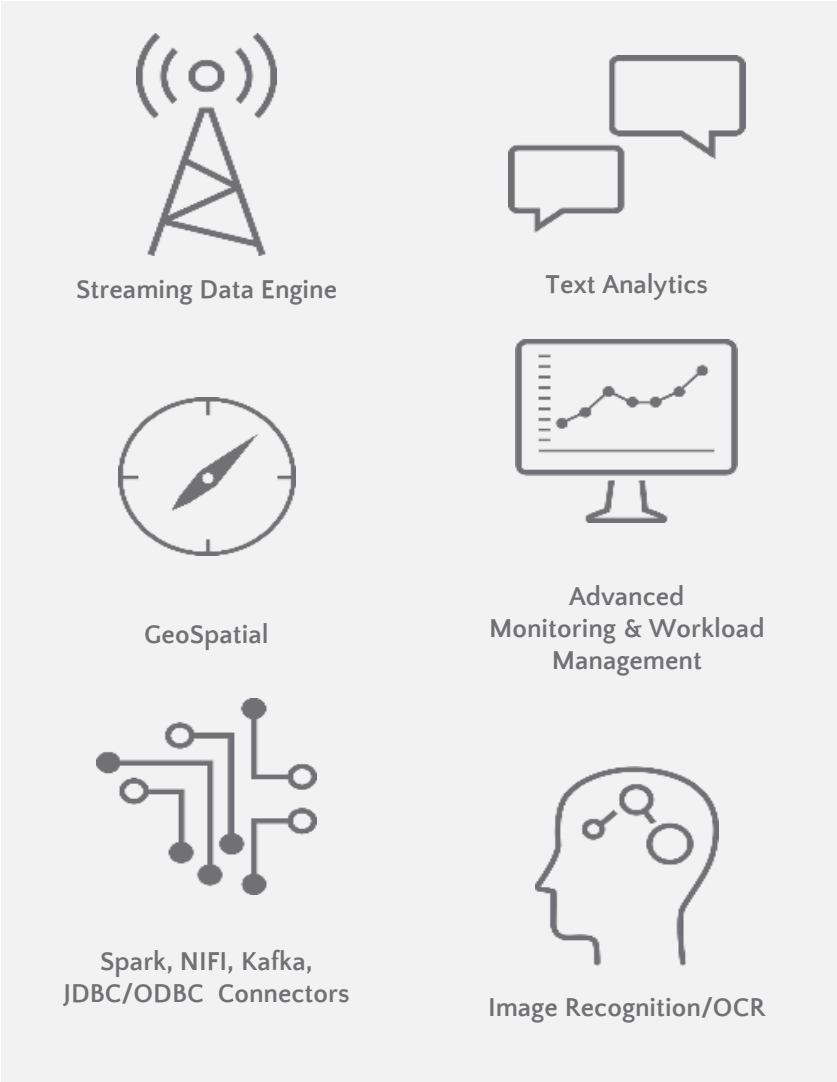
数据库扩展的意义

Greenplum 功能非常强大，为什么我们还需要数据库扩展

©2022 VMware, Inc.

为什么我们需要数据库扩展

数据库扩展的意义



- 避免核心数据库太过臃肿
- 满足用户的差异化需求
- 性能与便捷的结合
- 易于开发与维护
- 可以共享开源社区的成功
- 丰富Greenplum 生态系统

vmware ©2022 VMware, Inc.

热门Greenplum扩展介绍

- 磁盘配额管理扩展：Diskquota
- 过程化语言及容器：PL/Languages & PL/Container
- 高级文本分析扩展: GPText
- 地理分析扩展: PostGIS
- 数据库安全管理扩展： Advanced Password Check
- R语言客户端： GreenplumR
- Python语言客户端：GreenplumPython

©2022 VMware, Inc.

Greenplum 磁盘配额管理工具: Diskquota

- 监控数据表磁盘占用量
- 基于 Role/Tablespace/Schema 设置磁盘额度
- 避免写爆磁盘
- 拦截有问题的 Query

©2022 VMware, Inc.

Greenplum 磁盘配额管理工具: Diskquota

```
wxiaoran-a01:backend xiwang$ psql postgres
postgres=# create extension diskquota;
CREATE EXTENSION

  501 25871 25692   0 11:06AM ??         0:00.05 postgres:  6000, bgworker: [diskquota] - launcher
  501 25876 25590   0 11:06AM ??         0:00.04 postgres:  6002, xiwang diskquota 127.0.0.1(51880) con5 seg0 idle
  501 25877 25559   0 11:06AM ??         0:00.04 postgres:  6003, xiwang diskquota 127.0.0.1(51881) con5 seg1 idle
  501 25878 25602   0 11:06AM ??         0:00.04 postgres:  6004, xiwang diskquota 127.0.0.1(51882) con5 seg2 idle
  501 27594 25692   0 11:06AM ??         0:00.20 postgres:  6000, bgworker: [diskquota] postgres cmd2
```

©2022 VMware, Inc.

Greenplum 磁盘配额管理工具: Diskquota

```
CREATE SCHEMA s1;
SET search_path TO s1;

CREATE TABLE a(i int) DISTRIBUTED BY (i);
INSERT INTO a SELECT generate_series(1,100000);

SELECT diskquota.set_schema_quota('s1', '1 MB');
 set_schema_quota
------------------
(1 row)

INSERT INTO a SELECT generate_series(1,100);
ERROR:  schema's disk space quota exceeded with name: s1
```

©2022 VMware, Inc.

Greenplum 磁盘配额管理工具: Diskquota

```
SELECT tableid::regclass, size, segid FROM diskquota.table_size WHERE tableid::regclass::name NOT LIKE '%.%'
ORDER BY tableid;

 tableid |  size   | segid
---------+---------+-------
 a       | 3932160 |   -1
 a       | 1310720 |    2
 a       | 1310720 |    1
 a       | 1310720 |    0
```

©2022 VMware, Inc.

Greenplum 磁盘配额管理工具: Diskquota

```
CREATE SCHEMA s2;
SET search_path TO s1;
SELECT diskquota.set_schema_quota('s1', '1 MB');
 BEGIN;
CREATE TABLE b(i int);
INSERT INTO a SELECT generate_series(1,100000);
INSERT INTO a SELECT generate_series(1,100);
ERROR:  schema's disk space quota exceeded with name: s2
```

©2022 VMware, Inc.

Greenplum 过程化语言

PL/Python、PL/R、PL/Java、PL/PGSQL、PL/Perl

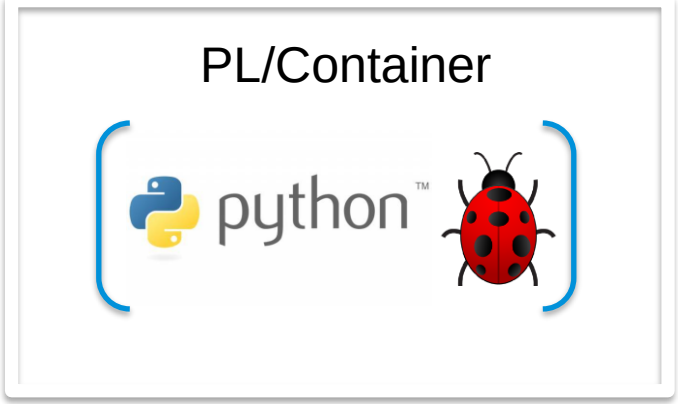
- 服务器端执行
- 基于高并发模型
- 可用过程化语言处理每一行数据
- 可借用容器增强安全性
- 支持加载三方库
- 支持用户自定义函数及自定义聚集
- 可调用开源机器学习算法

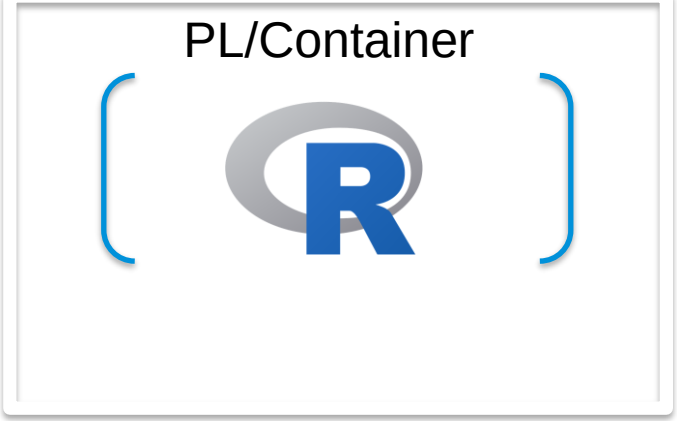
©2022 VMware, Inc.

PL/Container 简介

Greenplum 数据库 PL/Container 语言扩展（PL/Container）是一个接口，允许 Greenplum 数据库与 Docker 容器交互以在容器中执行用户定义的函数（UDF）

- 基于 Docker Container
- 高安全性/强隔离性：用户代码无法访问宿主机文件系统
- 可自定义容器资源使用上限
- 高性能：并行执行用户自定义函数
- 保持宿主机环境简洁，降低三方库依赖复杂度
- 支持GPU以助力机器学习
- 将负载发送到远程容器执行（开发中）

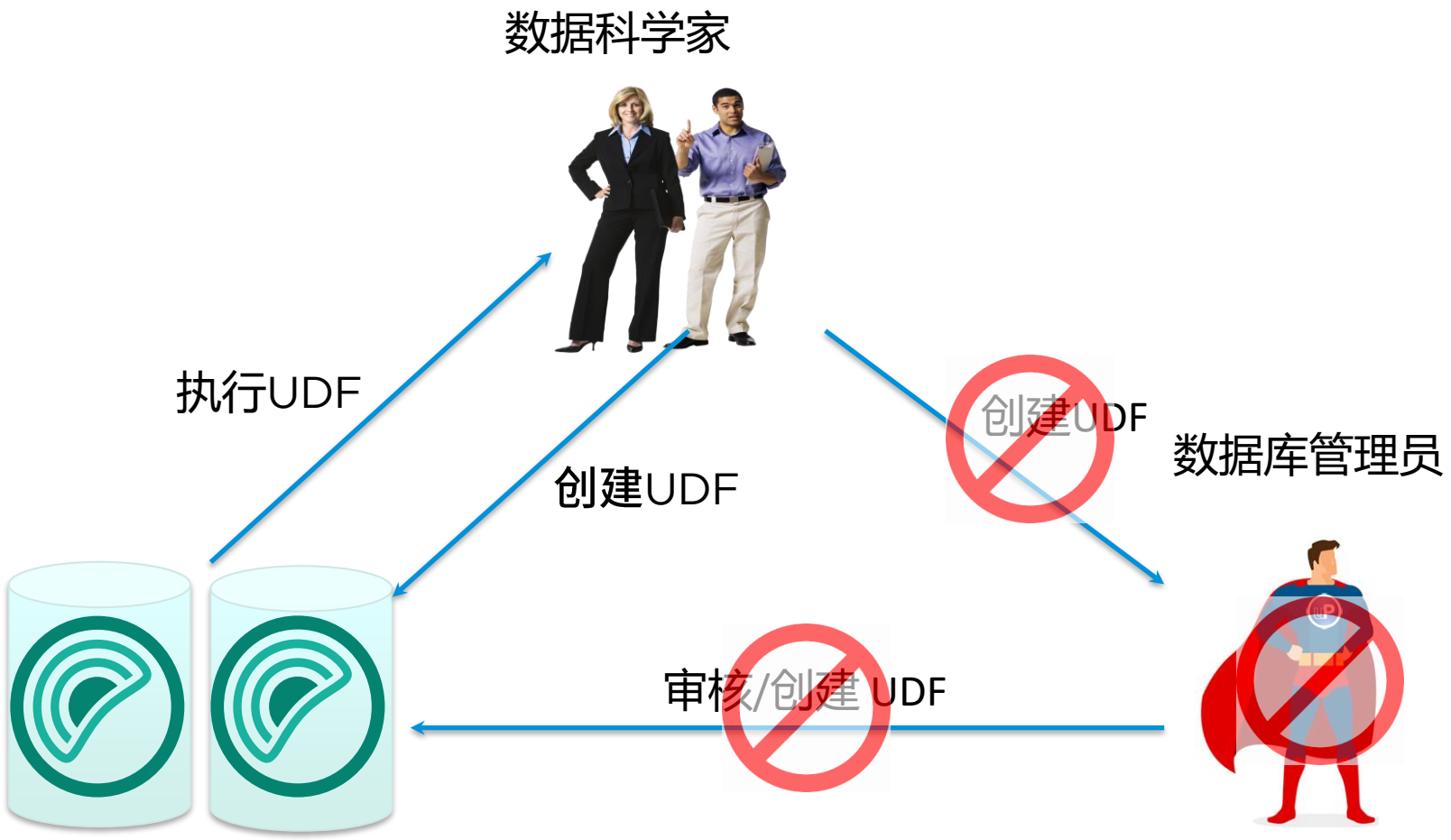




©2022 VMware, Inc.

PL/Container

Trusted



©2022 VMware, Inc.

GPText

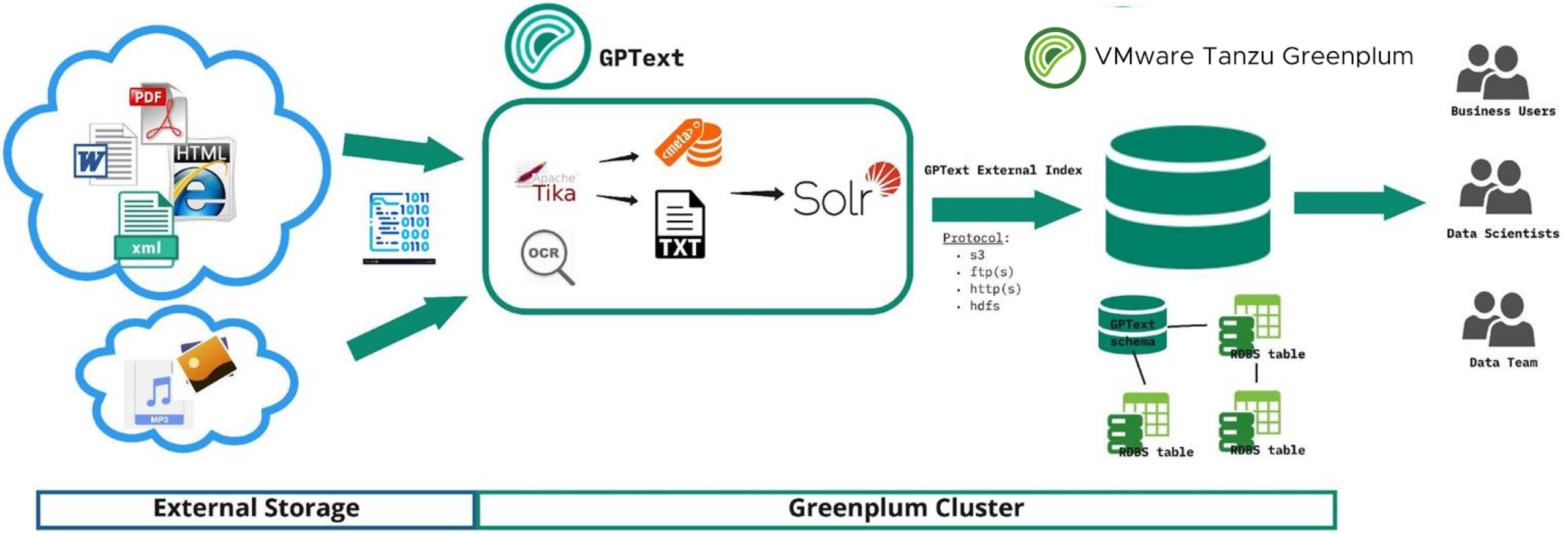
Greenplum 文本数据分析&搜索扩展

- 数据库与搜索引擎的完美结合
- 大规模高并行处理能力
- 支持常见格式的文档
- 支持多种外部数据来源
- 语义分析及自然语言处理

VMware Greenplum

©2022 VMware, Inc.

GPText 用户案例

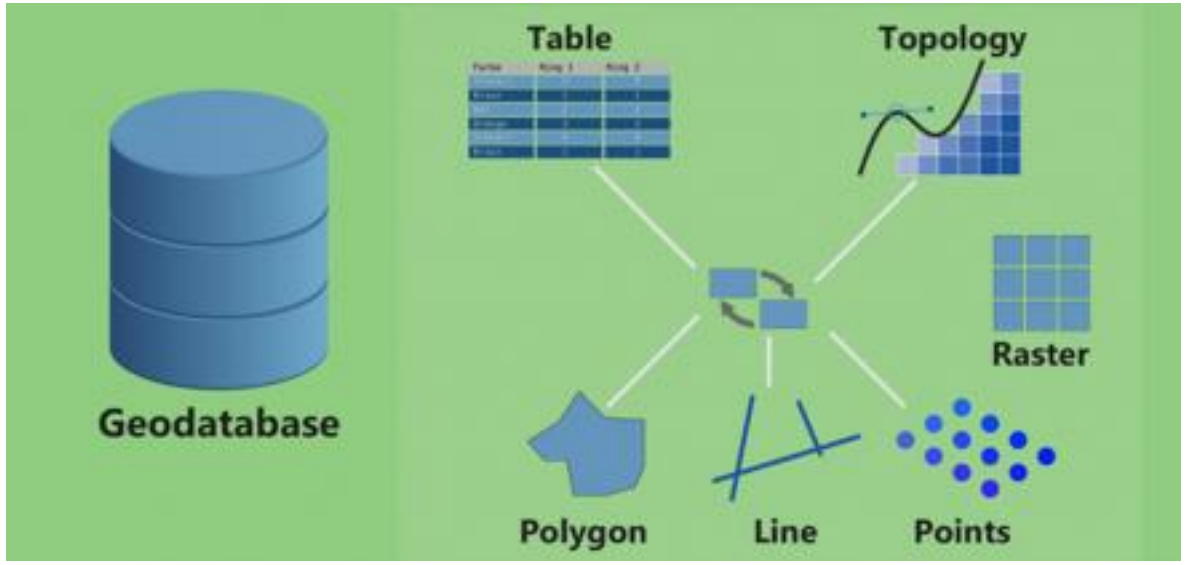


vmware ©2022 VMware, Inc.

PostGIS

Greenplum 地理数据分析扩展

- 基于PostGIS进行了企业级改进
- 支持Greenplum 分布式架构
- 支持大容量数据存储（数据存储于不同节点）
- 支持所有的空间数据类型
- 支持复杂的空间和地理位置计算
- 数据库坐标变换
- 空间聚集函数（包含、覆盖、相交等）



序号，	纬度，	经度，	时间戳
1，	115.3151，	37.2116，	20180803135530
2，	115.3153，	37.2118，	20180803135600
3，	115.3155，	37.2210，	20180803135630

…

©2022 VMware, Inc.

Greenplum Advanced Password Check

数据库密码合规化扩展

- 用户密码长度定制
- 用户密码格式及特殊字符定制
- 限制历史密码使用
- 用户自定义密码有效期
- 用户密码过期提醒
- 账号多次登录失败锁定机制

vmware ©2022 VMware, Inc.

Greenplum R

R 语言客户端扩展

- 远程执行本地R函数
- 高性能：服务器端并行执行
- 高安全性
- 不要求用户掌握 SQL 语言

- 数据无需离开database
- R代码自动转换为数据库函数

©2022 VMware, Inc.

GreenplumPython

©2022 VMware, Inc.

数据科学家的工具箱

引用的图片：

- Python logo generic，来源 https://www.python.org/community/logos/ 。Python 标志是 Python Software Foundation 的商标。
- Pandas logo，作者 Marc Garcia， 来源 https://commons.wikimedia.org/wiki/File:Pandas_logo.svg ，BSD 许可证
- Numpy logo，作者 Isabela Presedo-Floyd，来源 https://commons.wikimedia.org/wiki/File:NumPy_logo_2020.svg ， CC BY-SA 4.0 许可证
- Scikit-learn logo small， 作者 the scikit-learn developers ，来源 https://commons.wikimedia.org/wiki/File:Scikit_learn_logo_small.svg ，BSD 许可证

vmware ©2022 VMware, Inc.

GreenplumPython 的类 Pandas 接口

目的	GreenplumPython	Pandas
访问数据集	import GreenplumPython as gp db = gp.database(host="localhost", db_name="gpadmin") tb = db.table("hello_world")	import pandas as pd df = pd.read_csv("data/hello_world.csv")
提取多列数据	tb = tb[["a", "b"]]	df = df[["a", "b"]]

连接查询	tb = tb1.left_join(tb2,
	  cond=lambda t1, t2: t1["a"] == t2["x"],
	  self_columns={"a", "b"},
	  other_columns={"y"}
	)
		df = df1.join(df2,
		  left_on="a", right_on="x")

vmware® ©2022 VMware, Inc.

GreenplumPython 与 UDF

```
import greenplumpython as gp

demo = gp.database(dbname="demo").get_table("demo")

@gp.create_function
def image_zoom(image_jpeg: bytes) -> bytes:
    import io
    import numpy as np
    from PIL import Image
    from scipy import ndimage

    image = np.array(Image.open(io.BytesIO(image_jpeg)))
    zoomed_image = ndimage.zoom(image, (0.5, 0.5, 1))
    out = io.BytesIO()
    Image.fromarray(zoomed_image).save(out, format="jpeg")
    return out.getvalue()

image_zoom(demo["image_jpeg"])
```

vmware ©2022 VMware, Inc.

欢迎试用 GreenplumPython

GreenplumPython

- 完全开源
- 兼容 PostgreSQL
- 1.0 正式版会于近期发布

pip3 install --user greenplum-python

https://github.com/greenplum-db/GreenplumPython

二维码由 python-qrcode 工具（BSD许可）生成

vmware ©2022 VMware, Inc.

谢谢！

关注微信公众号

Greenplum中文社区

vmware EXPLORE 2022

©2022 VMware, Inc.