GREENPLUM DATABASE®

用Greenplum技术生态构建智慧城市

面向技术开发、大数据以及算法研究

刘昊松
Greenplum中文社区
2022 • 5

vmware®
Confidential | ©2021 VMware, Inc.

GREENPLUM DATABASE®

Agenda

智慧城市数据底座

(Greenplum的大数据平台)

时空大数据

(PostGis)

AI引擎与图计算

(Madlib/Plpython)

检索引擎

(GPText)

超融合一体化

(All in one策略)

GREENPLUM DATABASE®

智慧城市数据底座

Greenplum大数据平台

Confidential | ©2021 VMware, Inc.

GREENPLUM DATABASE

智慧城市

简单来说，智慧城市利用信息通信技术，实现城市各个方面的智能化管理和运营，包括社区、教育、交通、安防、医疗等各个领域。

智慧社区

智慧社区包括智慧家居、智慧物业等**基础设施服务**，涵盖社区内外的各种服务，以及智慧养老、智慧零售等**民生服务**。

智慧教育

引入新一代信息技术，搭建智慧教育平台，实现教育的**数字化**、**网络化**、**智能化**和**多媒体化**。

智慧交通

基于交通数据的智能分析，控制和支持交通建设管理的全过程，使系统具有**感知**、**互联**、**分析**、**预测**、**控制**等能力。

智慧安防

通过无线移动、跟踪和定位等手段建立全面的三维防护。全面整合**城市管理系统**、**环境监测系统**、**应急指挥系统**等。

智慧医疗

建立流畅的**医疗信息**体系，避免**医疗资源**两极分化，健全的医疗**监管机制**。

GREENPLUM DATABASE

数字孪生/元宇宙

业内龙头企业**51WORLD**的数字孪生案例。

www.51aes.com

1. 基于卫星正射和GIS数据自动化生成：根据卫星正拍影像和采集到的GIS数据，AI算法可以自动生成城市场景。
2. 基于OSGB数据重构生成：根据倾斜摄影/激光雷达扫描，计算机语义识别算法自动轻量化重建。
3. 基于BIM数据自动生成：基于BIM模型自动转化生成，并保留所有构件和零件目录与属性。
4. 程序化生成特定场景：基于程序算法快速生成特定场景，如输电站和输电网。
5. 基于HD Map高精地图生成：基于高精地图点云模型重构生成高精度道路场景。
6. 基于海量数字资产库构建：基于海量数字资产库特征匹配快速搭建场景。
7. 基于照片拓扑重建生成：基于照片实拍的3D拓扑重建生成模型。
8. 基于PaaS平台快速搭建：基于PaaS平台的场景二次编辑。

GREENPLUM DATABASE

关键技术的支持

1、数据底座

2、Gis等多源数据

3、计算引擎

4、3D虚拟现实

1 **基础数据类型：**

- 数值类型(int/float),货币类型(money),字符类型(text),日期时间类型(timestamp/date/time),布尔类型(boolean),枚举类型,网络地址类型,文本搜索类型,UUID类型,XML类型,JSON类型,数组类型,复合类型(结构体),二进制类型。

2 **空间地理信息数据类型：**

- Gis数据类型(Geometry),包含 点、线、多边形、多面体、复合类型(组合结构)等。

3 **视频图像数据：**

- 数组嵌套二维数组，元素为像素的RGB。

4 **图数据：**

- 知识图谱数据,三元组形式。

GREENPLUM DATABASE®

时空大数据

PostGis

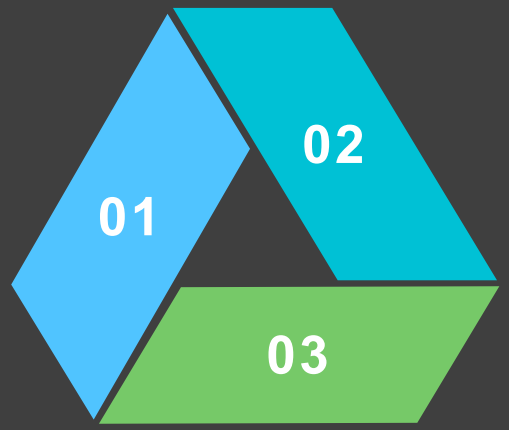
Confidential | ©2021 VMware, Inc.

GREENPLUM DATABASE

PostGis

数据存储:

地理信息数据的结构化存储，Geometry字段类型，包括点、线、多边形、多面体、复合类型等。



数据计算:

支持各种地理信息数据的计算，包括求距离、判断相交关系、计算相交区域、计算面积等各种功能。

数据展示:

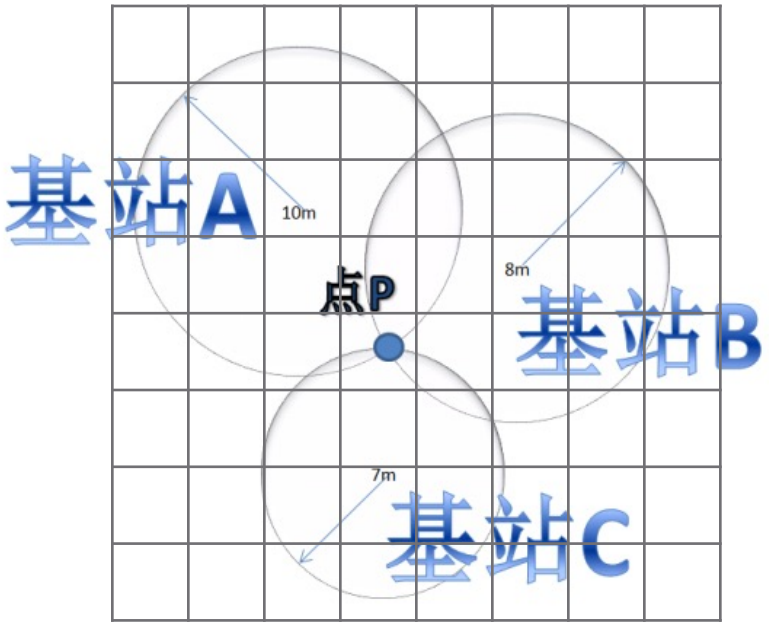
通过QGis等直连数据库，生成.shp格式的缓存文件等手段进行数据呈现与展示。

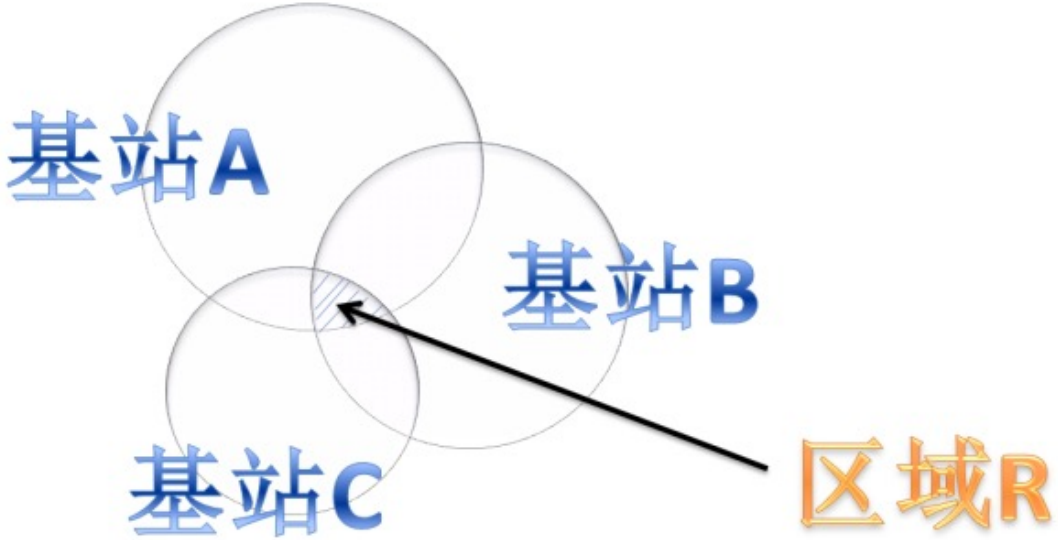
postgis-2.4.0.pdf

通信大数据

基于通信运营商的人口定位分析案例。

三点定位->栅格化定位





GREENPLUM DATABASE®

AI引擎与图计算

Madlib/PLpython

Confidential | ©2021 VMware, Inc.

GREENPLUM DATABASE

Madlib

1、机器学习

2、深度学习

3、知识图谱

主要功能模块：

- 数据类型与转换（数组与矩阵运算、稀疏向量等）
- 图算法（图计算、单源最短路径等）
- 模型评估（交叉验证、指标验证等）
- 统计（描述性统计、推断性统计、概率函数）
- 监督学习（条件随机场、回归模型、支持向量机、树方法）
- 时间序列分析
- 无监督学习（关联规则、聚类、主题模型）
- 应用函数（数据库函数、线性求解器、路径函数、PMML导出、会话化、文本分析）

KNN算法案例

小王是一家著名高尔夫俱乐部的经理。但是他被雇员数量问题搞得心情十分不好。某些天好像所有人都來玩高尔夫，以至于所有员工都忙的团团转还是应付不过来，而有些天不知道什么原因却一个人也不来，导致俱乐部为雇员数量浪费了不少资金。

- 小王的目的是通过下周天气预报寻找什么时候人们会打高尔夫，以适时调整雇员数量。因此首先他必须了解人们决定是否打球的原因。
- 在两周时间内可以得到以下记录：天气状况有晴、云和雨；华氏温度表示的气温；相对湿度百分比；是否有风。当然还有顾客是不是在这些日子光顾俱乐部。最终他得到了14行5列的数据。

GREENPLUM DATABASE

KNN算法案例

Madlib-KNN.txt

自变量				因变量
天气情况	华氏温度	相对湿度	是否有风	是否打球
sunny	85	85	FALSE	Don’t Play
sunny	80	90	TRUE	Don’t Play
overcast	83	78	FALSE	Play
rain	70	96	FALSE	Play
rain	68	80	FALSE	Play
rain	65	70	TRUE	Don’t Play
overcast	64	65	TRUE	Play
sunny	72	95	FALSE	Don’t Play
sunny	69	70	FALSE	Play
rain	75	80	FALSE	Play
sunny	75	70	TRUE	Play
overcast	72	90	TRUE	Play
overcast	81	75	FALSE	Play
rain	71	80	TRUE	Don’t Play

GREENPLUM DATABASE

PLpython

1、Python过程化语言

2、天然分布式

3、时间序列预测自动定阶

虚拟机版本：	Madlib版本：	操作工具版本：
Linux Centos 7.4 1708	Postgresql 10.13 + Madlib 1.16	Navicat Premium 11.1

PLPython脚本.txt

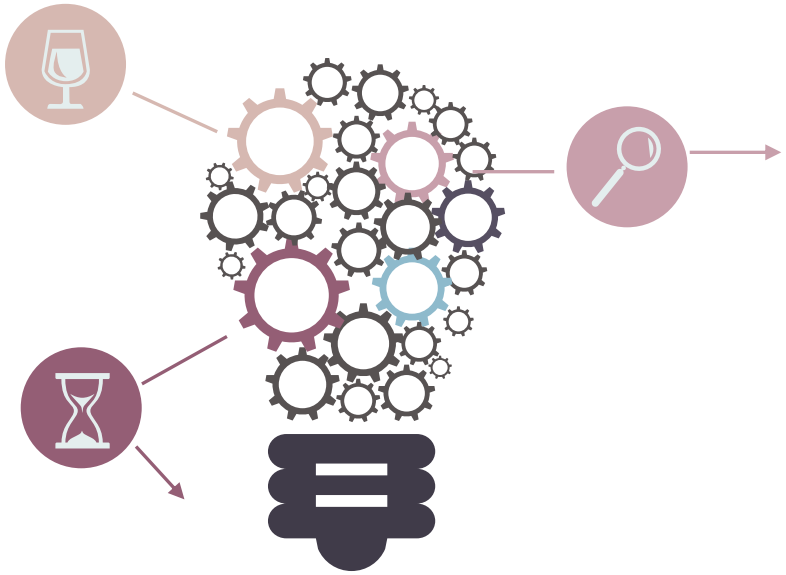
arima_pdq.txt

知识图谱

发展历史

- 知识图谱（Knowledge Graph）的概念由谷歌2012年正式提出，旨在实现更智能的搜索引擎，并且于2013年以后开始在学术界和业界普及。

- 目前，随着智能信息服务应用的不断发展，知识图谱已被广泛应用于智能搜索、智能问答、个性化推荐、情报分析、反欺诈等领域。



数据类型

- 结构化数据（Structed Data）：如关系数据库

- 半结构化数据（Semi-Structed Data）：如XML、JSON、百科

- 非结构化数据（UnStructed Data）：如图片、音频、视频、文本

知识图谱定义

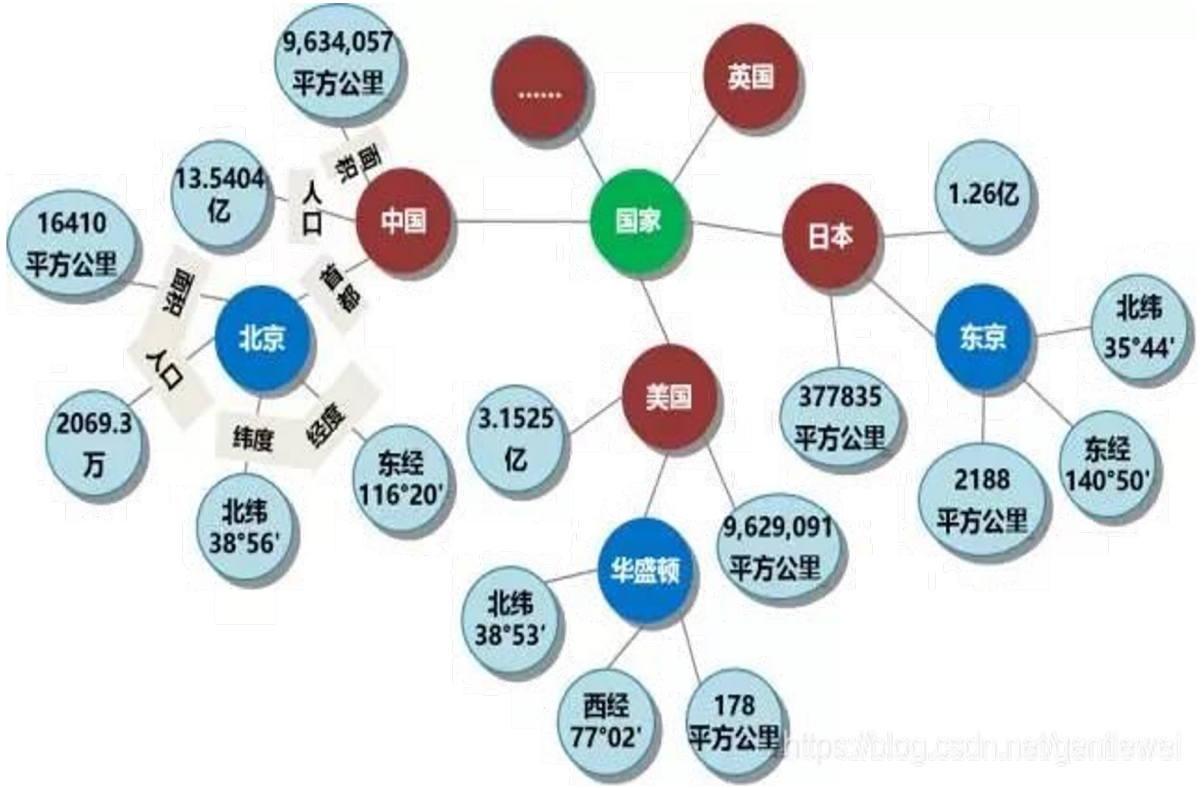
- 知识图谱是结构化的语义知识库，用于迅速描述物理世界中的概念及其相互关系。
- 知识图谱有自顶向下和自底向上两种构建方式。借助百科类网站等结构化数据源，从高质量数据中提取本体和模式信息，加入到知识库中。

GREENPLUM DATABASE

三元组

边连接在一起，这个是实体（Entity）节点称为实体

关系（Relationship）-



实体

实体指的是具有可区别性且独立存在的某种事物。实体是知识图谱中的最基本元素，不同的实体间存在不同的关系。如“中国”、“北京”、“2069.3万”等。

核心

01

02

关系

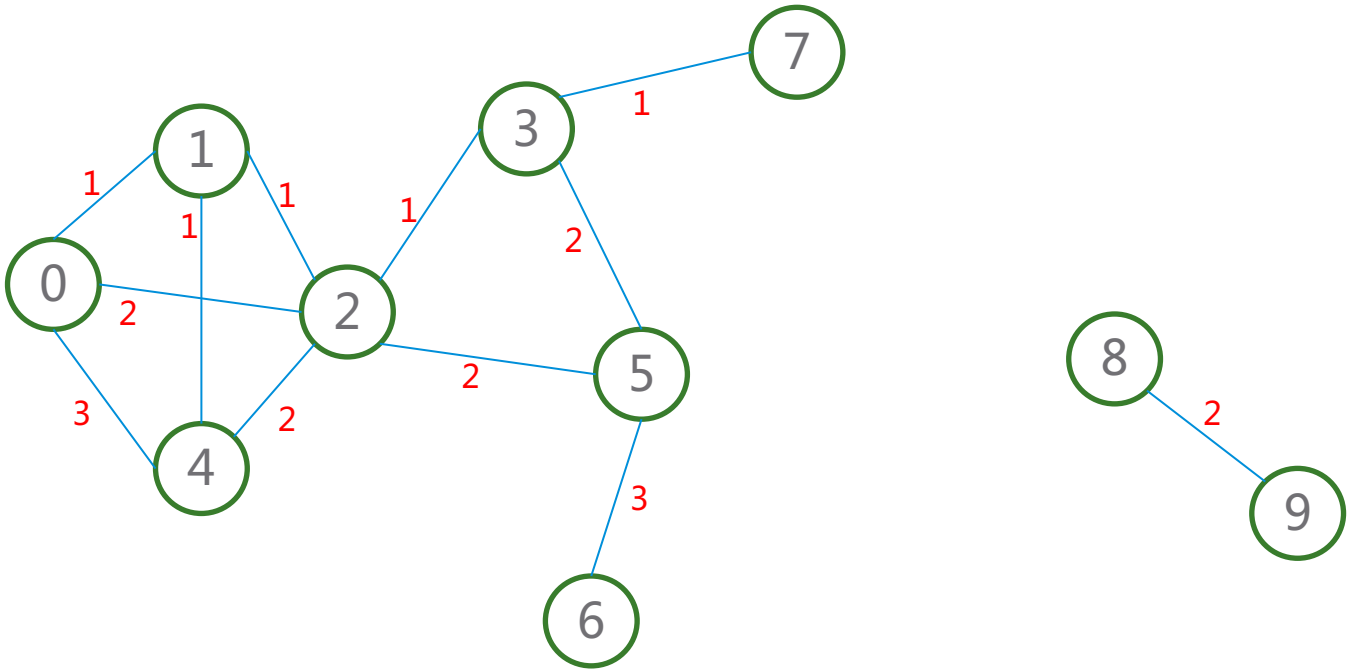
关系是连接不同的实体，指代实体之间的联系。通过关系节点把知识图谱中的节点连接起来，形成一张大图。如“人口”、“首都”、“面积”等

图计算

1、有向图与无向图

2、数据示例

3、图计算脚本



算法01：单源最短路径算法.txt

算法07：全最短路径、中心性算法.txt

GREENPLUM DATABASE®

检索引擎

GPText

Confidential | ©2021 VMware, Inc.

GREENPLUM DATABASE

文本检索

1、GPText

2、Gin索引

GPText是作为Greenplum文本分析的扩展引擎。

GPText是将Greenplum数据库集群与Apache SolrCloud相结合，对大规模的文本分析进行支持。

GREENPLUM DATABASE

GPText的特性

1. 海量文本数据处理。

2. 高速并行全文本检索。

3. 支持半结构化、结构化数据（社交媒体、XML等格式文档、Json格式数据等）。

4. 易用的SQL接口。

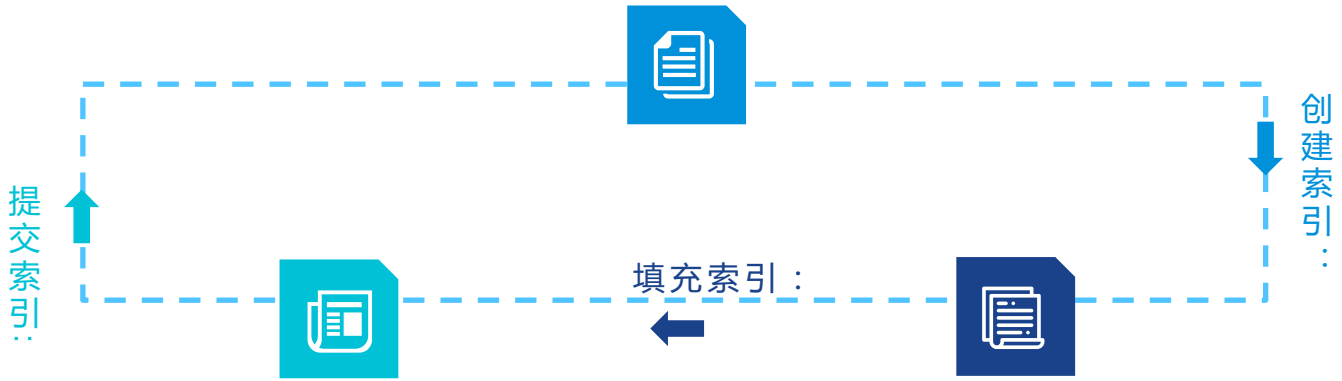
5. 复杂文本分析。

6. 内置一部分自然语言处理工具

7. 高可用（结合Zookeeper使用，内置了Solr引擎）。

GPText索引示例

- gptext.create_index(<schema_name>, <table_name>, <id_col_name>, <def_search_col_name> [, <if_check_id_uniqueness>]);
- SELECT * FROM gptext.create_index('public','people_info','id','name');



SELECT * FROM gptext.commit_index('traffic.public.people_info');

SELECT * FROM gptext.index(TABLE(SELECT * FROM public.people_info),'traffic.public.people_info');

Gin索引介绍

Gin意思是通用倒排索引。Gin索引接口常被用于多值列的检索，例如全文检索类型、数组类型。

Gin(Generalized Inverted Index) 是一个存储对(key, posting list)集合的索引结构，其中key是一个键值，而posting list 是一组出现过key的位置。

在表中的每一个属性，在建立索引时，都可能会被解析为多个键值，所以同一个元组的tid可能会出现在多个key的posting list中。

GIN索引特别适用于多值类型的元素搜索，比如支持全文搜索，数组中元素的搜索，而PG的GIN索引模块最初也是为了支持全文搜索而开发的。

Gin索引介绍

- 在Postgresql9.6及以后，Json字段可使用函数jsonb_set直接按key值批量更新。
- 现有部署Greenplum6.1.0中内置Postgresql9.4版本，无法使用一些json操作符，因此不能使用上述函数进行更新。通过编写存储过程也可实现Json格式数据直接按key批量更新、删除。
- 详见文档：
  - 《Greenplum6.1.0中JSON字段更新与删除key的存储过程.txt》
    - 参考资料：
      - https://www.cnblogs.com/fangyuan303687320/p/5509854.html
      - https://www.postgresql.org/docs/9.6/functions-json.html

Greenplum6.1.0中JSON字段更新与删除key的存储过程.txt

GREENPLUM DATABASE®

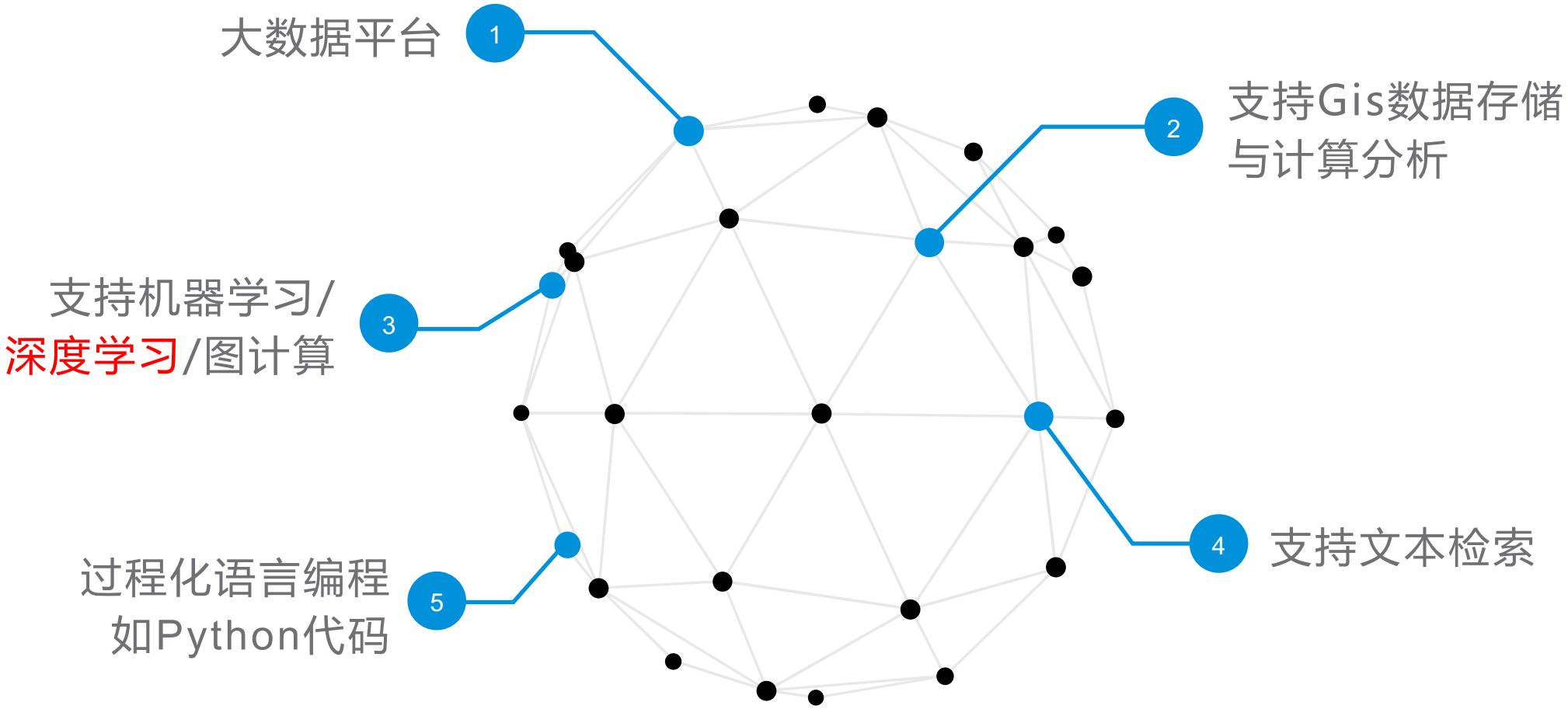
超融合一体化

All in one策略

Confidential | ©2021 VMware, Inc.

GREENPLUM DATABASE

All in one策略



GREENPLUM DATABASE®

微信技术讨论群
微信搜索添加“gp_assistant”
加入技术讨论

微信公众号
搜索添加“Greenplum中文社区”
技术干货、行业热点、活动预告

Thank You

Confidential | ©2021 VMware, Inc.