



GREENPLUM
DATABASE®

Open Source AceCon

2021

智能云边开源峰会

AI x Cloud Native x Edge Computing

人工智能 × 云原生 × 边缘计算

演讲主题

AI on Greenplum Cloud

Greenplum Cloud助力AI科学计算

杨峻峰

Greenplum内核研发工程师

强大的数据库分析平台



GREENPLUM
DATABASE®

Open Source AceCon
2021 智能云边开源峰会
AI x Cloud Native x Edge Computing
人工智能 x 云原生 x 边缘计算

一个用于分析，机器学习的开源大规模并行数据平台

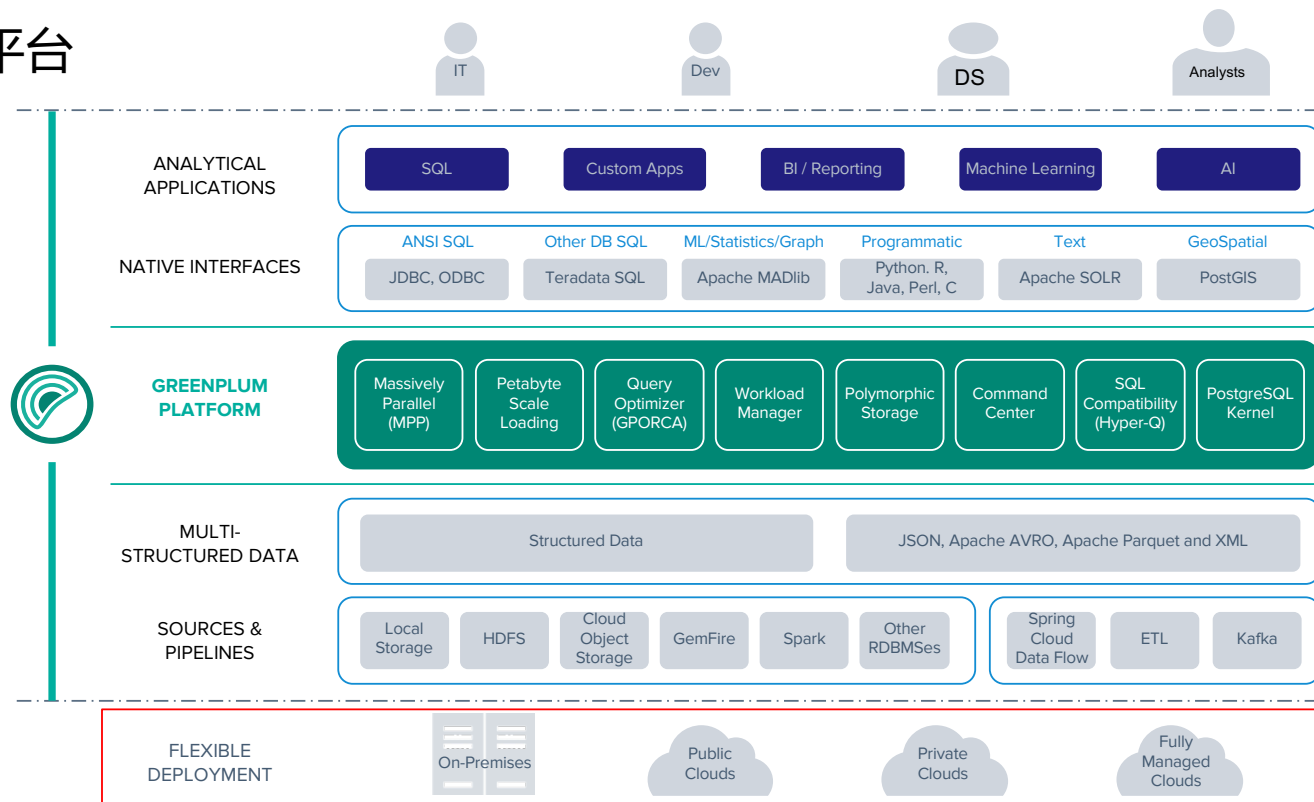
PostgreSQL

+ Massively Parallel Processing

+ Extra features

+ And addons

NEXT
GENERATION
DATA
PLATFORM



强大的数据库分析平台



GREENPLUM
DATABASE®

Open Source AceCon
2021 智能云边开源峰会
AI x Cloud Native x Edge Computing
人工智能 x 云原生 x 边缘计算



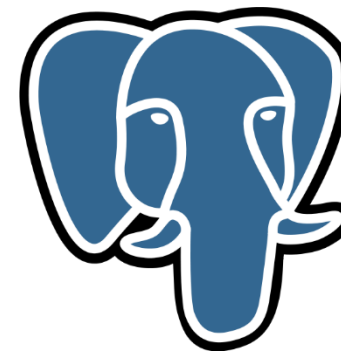
Greenplum Major Versions

- 持续继承，并与 PostgreSQL 保持一致

Greenplum V7



- Postgres v12
- 即时编译使长查询更快
- 多核处理加速
- 块范围索引 (BRIN) 类似于 zonemaps，用于跳过范围外的数据块
- 用于查询调优的多列统计信息
- Upsert 命令使ETL更容易
- 用于细粒度控制的行级安全性
- 存储过程中支持事务
- SCRAM认证更安全





GREENPLUM
DATABASE®

Open Source AceCon
2021 智能云边开源峰会
AI x Cloud Native x Edge Computing
人工智能 x 云原生 x 边缘计算

Greenplum 数据联邦 (Data Federation)

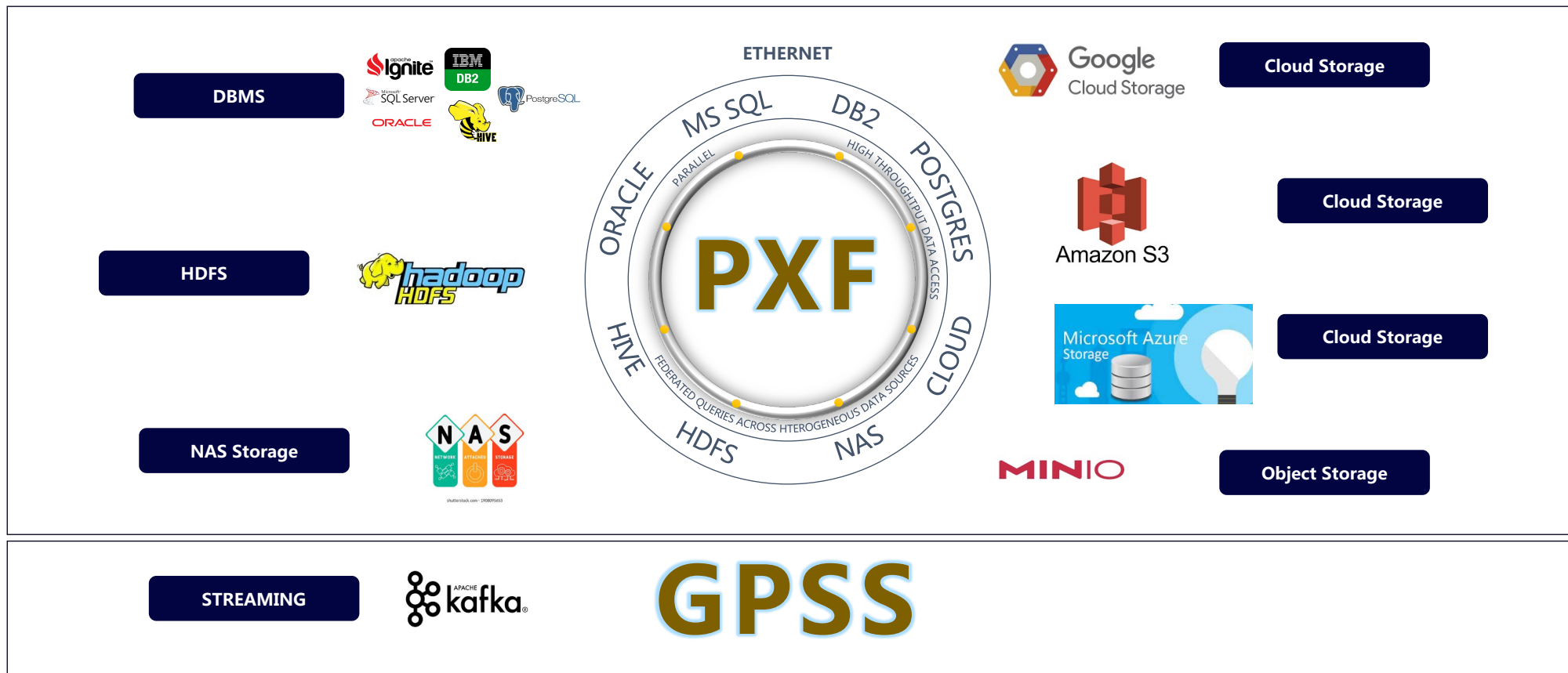
丰富的数据接口

Platform Extension Framework (PXF)

- Provides a simple interface through an external table
- Provides the best performance through parallel data loading/extraction

Greenplum Stream Server (GPSS)

- Provide Kafka Connector for streaming data data processing
- Avro, CSV, Text, JSON format support



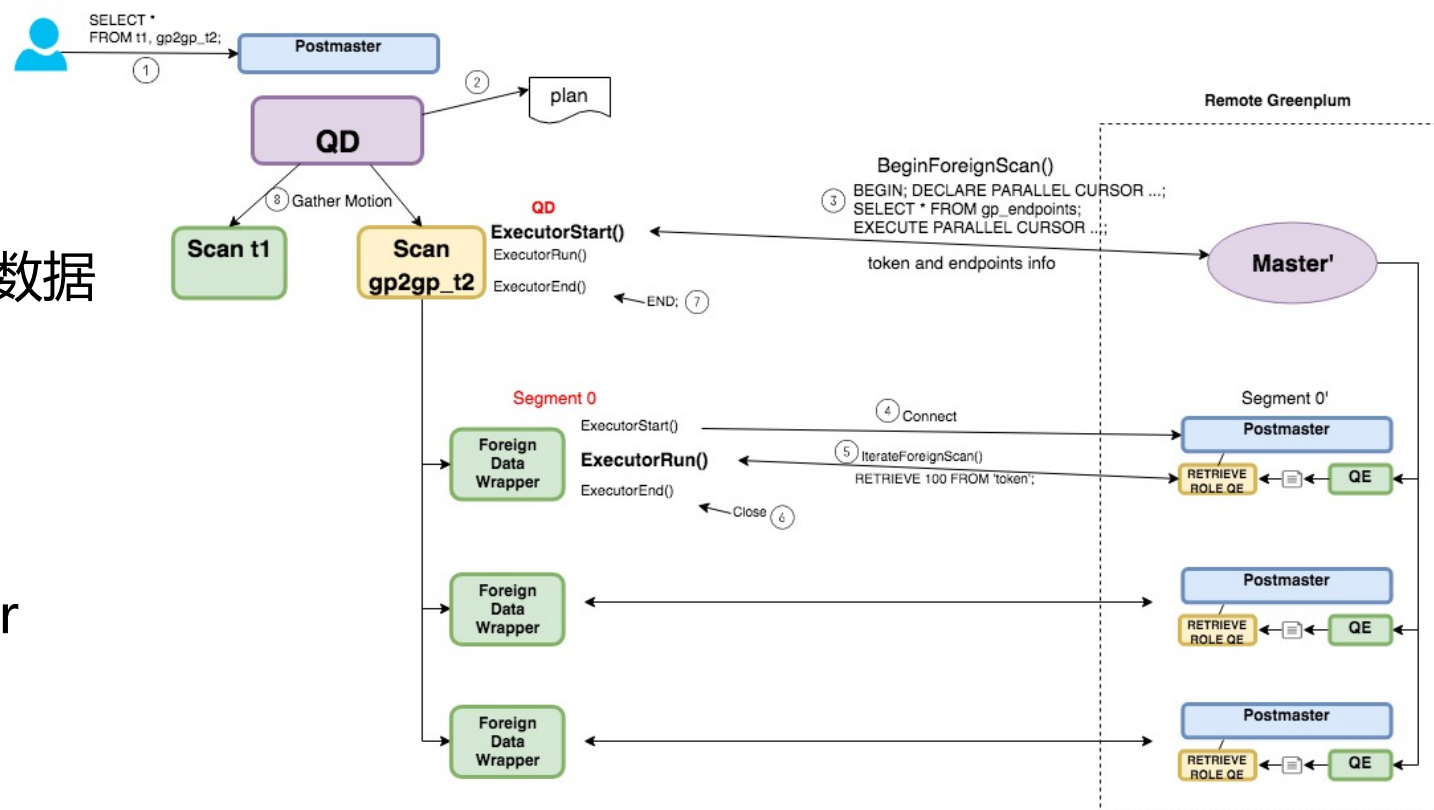
跨集群数据服务

Key Outcomes

- 避免数据复制
- 公司部门和用例之间高度隔离
- 通过创建集群到集群的连接来提高数据承载量和并发的可扩展性

Greenplum Federation

- 可感知的 Optimizer and Executor
- 可感知的 MPP Segments
- 更智能的查询优化 Joins, Pushdowns



2022 Preview



GREENPLUM
DATABASE®

Open Source AceCon
2021 智能云边开源峰会
AI x Cloud Native x Edge Computing
人工智能 × 云原生 × 边缘计算

Greenplum 与 AI

机器学习库

- 一个基于SQL的数据库内置的可扩展的机器学习库
- 强大的分析能力
- 将机器学习逻辑与数据库特定的实现细节分开
- 充分利用MPP架构处理海量数据
- Apache ASF上的顶级开源项目

MADlib 功能



Supervised Learning

- Neural Networks
- Support Vector Machines (SVM)
- Regression Models
 - Clustered Variance
 - Cox-Proportional Hazards Regression
 - Elastic Net Regularization
 - Generalized Linear Models
 - Linear Regression
 - Logistic Regression
 - Marginal Effects
 - Multinomial Regression
 - Naïve Bayes
 - Ordinal Regression
 - Robust Variance
- Tree Methods
 - Decision Tree
 - Random Forest
- Conditional Random Field (CRF)

Unsupervised Learning

- Association Rules (Apriori)
- Clustering (k-Means)
- Topic Modelling (Latent Dirichlet Allocation)

Nearest Neighbors

- k-Nearest Neighbors

Graph

- All Pairs Shortest Path (APSP)
- Breadth-First Search
- Average Path Length
- Closeness Centrality
- Graph Diameter
- In-Out Degree
- PageRank

Data Types and Transformations

- Array and Matrix Operations
- Matrix Factorization
 - Low Rank
 - Singular Value Decomposition (SVD)
- Norms and Distance Functions
- Sparse Vectors
- Principal Component Analysis (PCA)
- Categorical Variables

Statistics

- Estimators
- Variance and Covariance
- Statistical Tests
- Confidence Intervals

Model Selection

- Cross Validation
- Prediction Metrics
- Train-Test Split

Time Series Analysis

- ARIMA

Term Frequency for Text Analysis

复杂, 成熟的数据科学学习库

Sept 2017

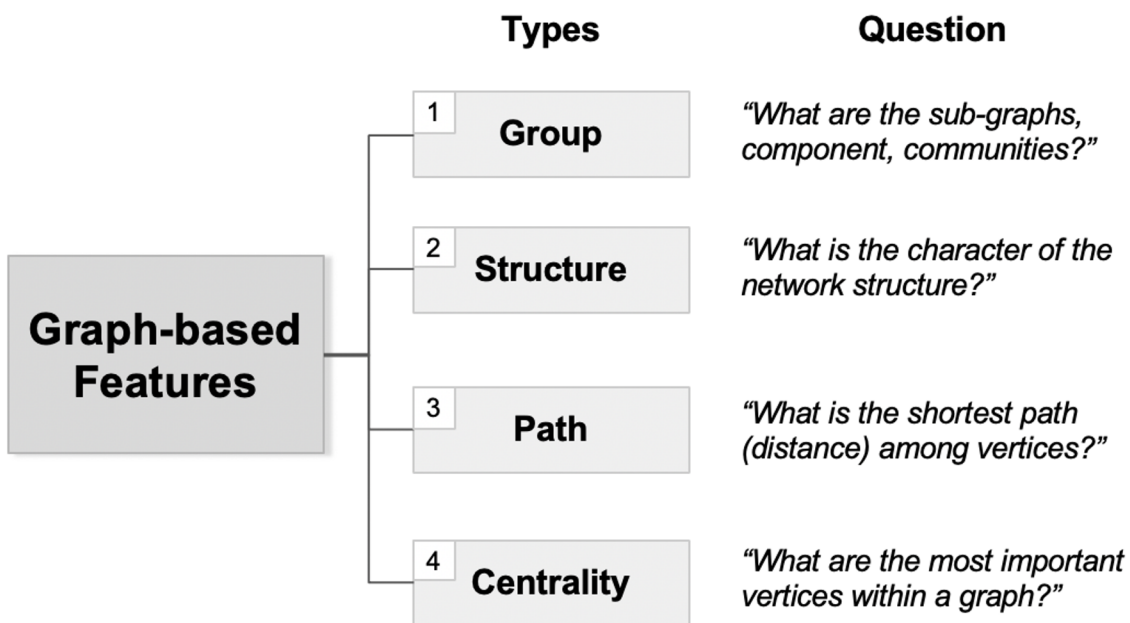
知乎 @ 行大咖啡

<https://github.com/apache/madlib>

图分析



Graph Algorithms and Measures



Social Network



* Grandjean, M. (2016)

Epidemiology



* <http://www.netminer.com/communi>

Bank Risk



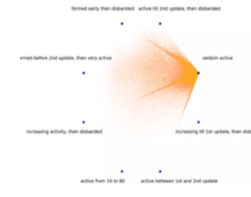
* <https://cambridge-intelligence.com>

1st Party Fraud



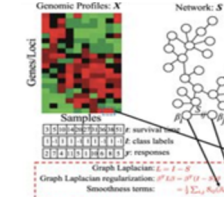
* www.infoglide.com

MMO Role-Playing Game



* www.researchgate.net

Chemistry



* <https://www.nature.com/article>

Gene



* www.researchgate.net

Manufacturing



* <https://blog.trifinance.com>

文本检索和NLP

- **Extract** data from binary or human readable formats into data that a machine can understand and operate on.

- **Index** the text data, so we can quickly search for specific text and documents.

- **Search** the text for patterns and keywords.

- **Analyze** what the text actually means.



数据分析扩展语言

- Greenplum原生
 - UDF
 - UDA
 - 支持优化器优化
- 内核逻辑支持其他语言解释器/编译器
 - R/Python/Perl/Java/C/C++/...
 - 支持加载第三方库
- 容器化和客户端组件支持
 - PL/Container
 - GreenplumR



Spark 连接

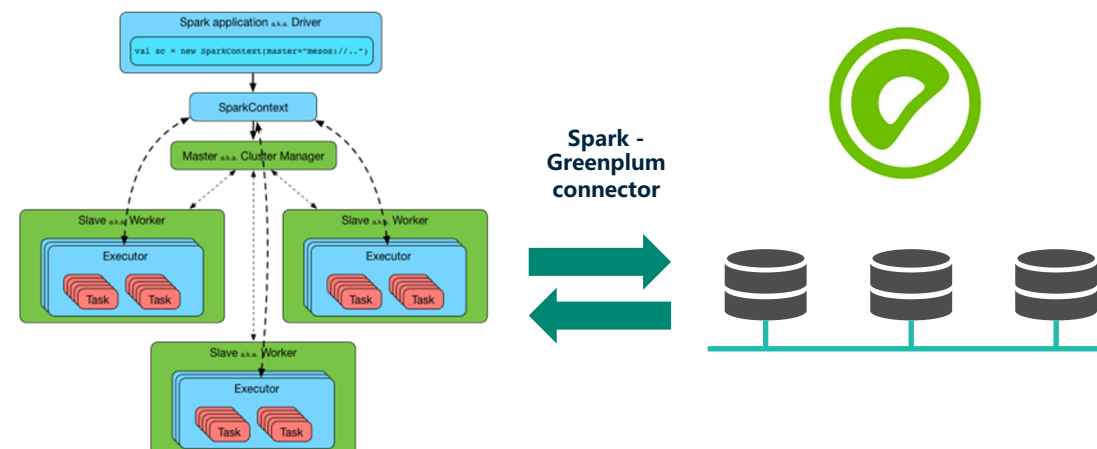


Greenplum Driver for Spark

- Spark 代码中引用 Greenplum 数据源
- Spark 和 Greenplum 都是多节点并行引擎
- 通过网络高速并行传输数据
- 从 Greenplum 读取到内存中进行 Spark 处理
- 从 Spark 写入 Greenplum 以实现数据持久化
- 无需将数据导出和复制到位于 Greenplum 之外的文件



In-memory processing



我们的课程



<https://cloud.tencent.com/edu/paths/series/GreenPlum>

Greenplum联合腾讯云大学的机器学习课程

基于 Greenplum 的机器学习算法与实践

腾讯云大学联合 GreenPlum 官方打造从机器学习算法与实战课程

阶段一

机器学习算法
与实战

课程 10

课时 43

- 课程一、基于 Greenplum 的机器学习算法与实践-机器学习的前世今生 +
- 课程二、基于 Greenplum 的机器学习算法与实践-回归算法 +
- 课程三、基于 Greenplum 的机器学习算法与实践-朴素贝叶斯分类 +
- 课程四、基于 Greenplum 的机器学习算法与实践-基于树的分类 +
- 课程五、基于 Greenplum 的机器学习算法与实践-支持向量机 +
- 课程六、基于 Greenplum 的机器学习算法与实践-非监督学习算法 +
- 课程七、基于 Greenplum 的机器学习算法与实践-神经网络与深度学习 +
- 课程八、基于 Greenplum 的机器学习算法与实践-时间序列经典算法 +
- 课程九、基于 Greenplum 的机器学习算法与实践-图算法 +
- 课程十、基于 Greenplum 的机器学习算法与实践-数据分析扩展语言 +

完成学业



**GREENPLUM
DATABASE®**

Open Source AceCon

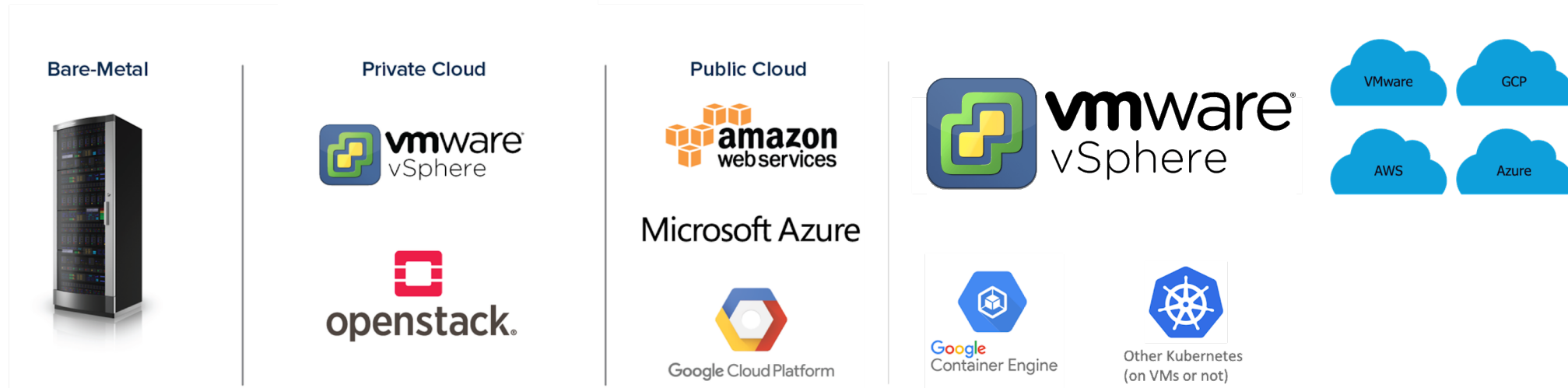
2021 智能云边开源峰会
AI x Cloud Native x Edge Computing
人工智能 x 云原生 x 边缘计算

Greenplum AS A Service

Run your analytics anywhere you need it



Infrastructure Agnostic



Open Source - 100% portable software solution

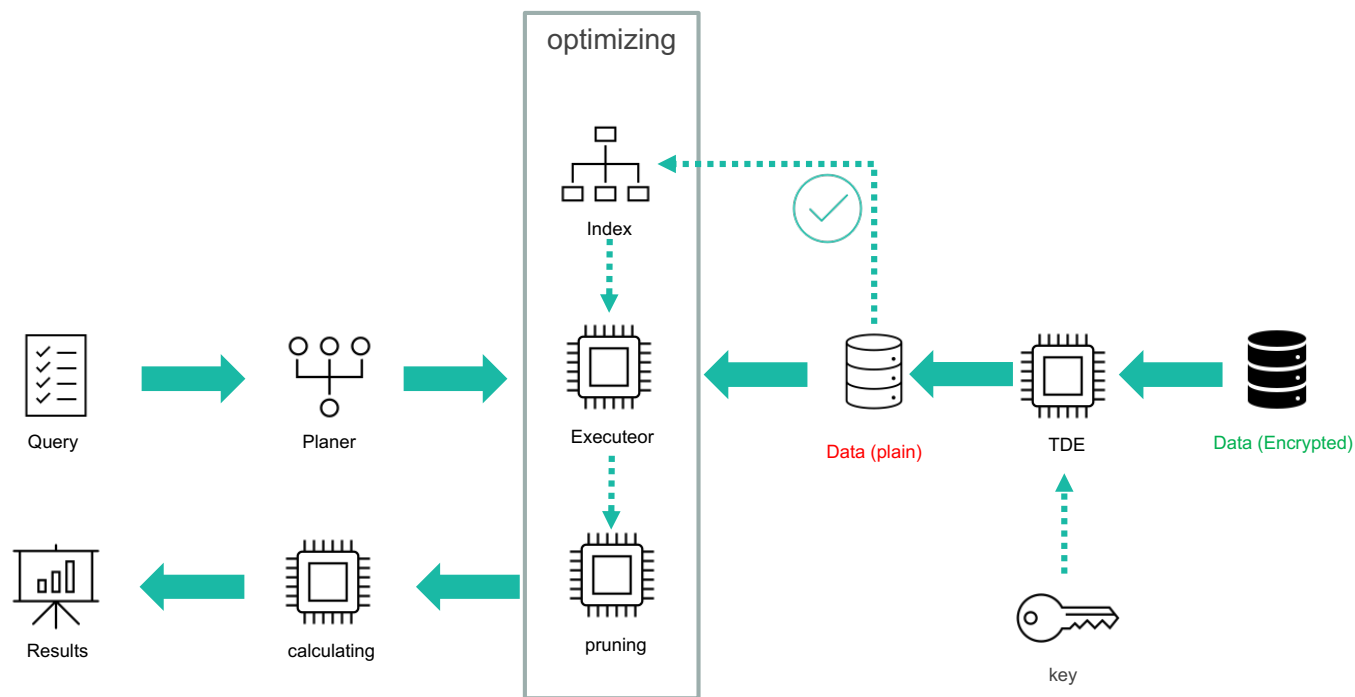
透明加密 (for cloud)

Key Outcomes

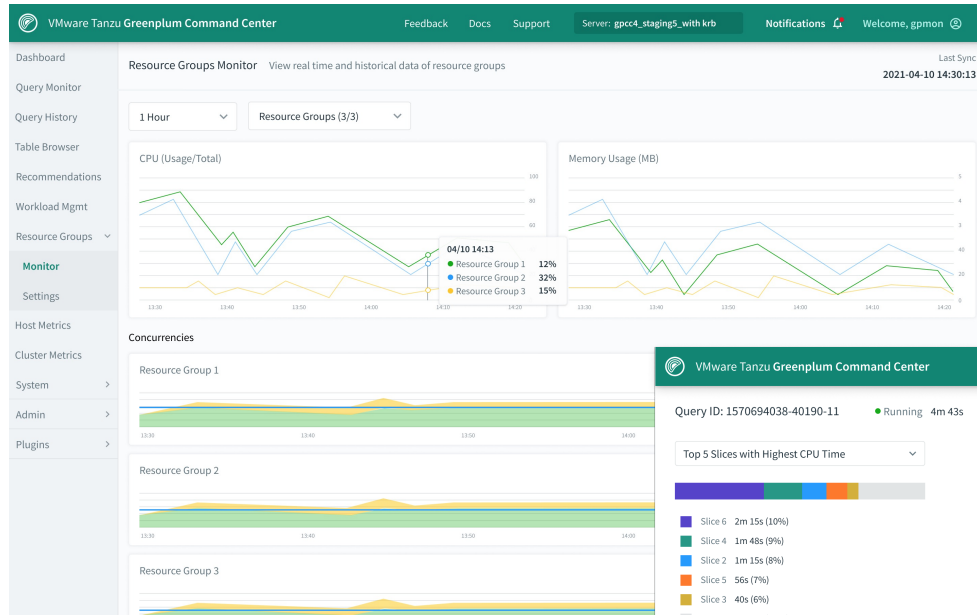
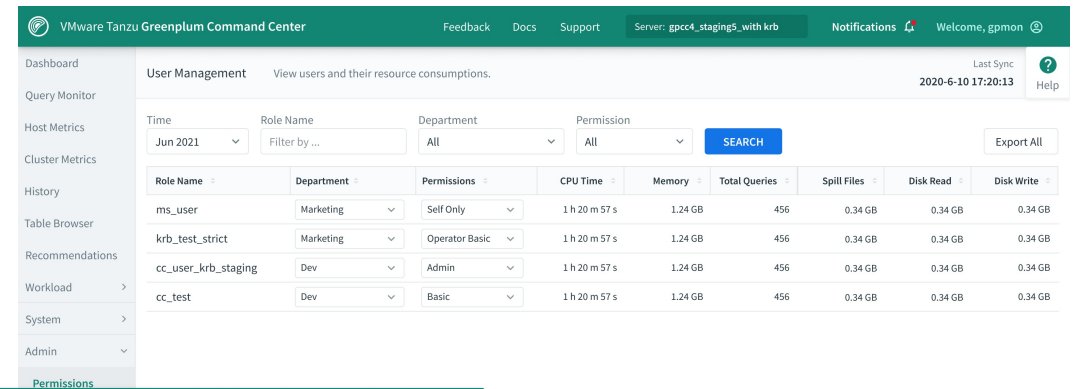
- 防止数据中心操作员访问磁盘上的数据
- 防止单个密钥丢失危及所有数据

Greenplum Transparent Data Encryption

- 磁盘上的表文件、系统文件的加密
- 主密钥将存储在外部 KMS 服务器上。
- 主要/对象密钥将存储在 GPDB Master 上
- 多级密钥层次结构
- 支持密钥轮换



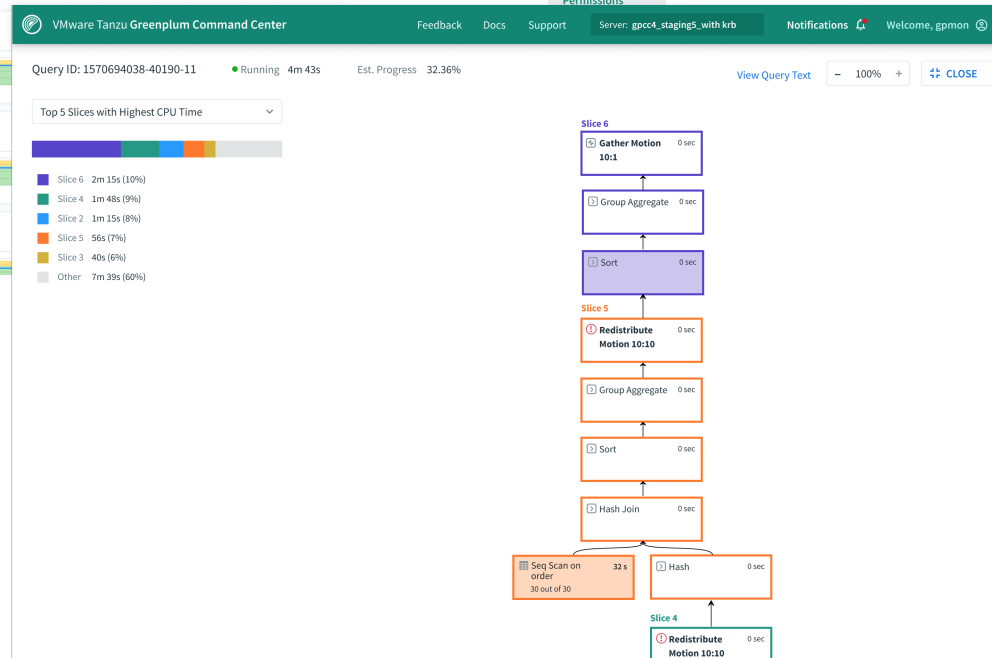
Greenplum command center

User Management View users and their resource consumptions.

Time: Jun 2021 | Role Name: Filter by ... | Department: All | Permission: All | SEARCH | Export All

Role Name	Department	Permissions	CPU Time	Memory	Total Queries	Spill Files	Disk Read	Disk Write
ms_user	Marketing	Self Only	1 h 20 m 57 s	1.24 GB	456	0.34 GB	0.34 GB	0.34 GB
krb_test_strict	Marketing	Operator Basic	1 h 20 m 57 s	1.24 GB	456	0.34 GB	0.34 GB	0.34 GB
cc_user_krb_staging	Dev	Admin	1 h 20 m 57 s	1.24 GB	456	0.34 GB	0.34 GB	0.34 GB
cc_test	Dev	Basic	1 h 20 m 57 s	1.24 GB	456	0.34 GB	0.34 GB	0.34 GB



CPU Time	Memory	Total Queries	Spill Files	Disk Read	Disk Write
145 h 45 m	783.45 GB	123,456,789	783.45 GB	78.45 GB	78.45 GB

Greenplum On vSphere



Open Source AceCon
2021 智能云边开源峰会
AI x Cloud Native x Edge Computing
人工智能 x 云原生 x 边缘计算

提高自动化程度并减少运维

- Greenplum vSphere OVA Deployment
- 从 VMware 提供的 OVA 文件自动部署
- 使用OVA提供的功能来完成Day 1 and Day 2的操作部署
- 通过VMWare提供的下载保证安全和快速patch应用
- 减少对虚拟机的登录配置需求





VMWARE GREENPLUM
AS A SERVICE



GREENPLUM
DATABASE®

Open Source AceCon
2021 智能云边开源峰会
AI x Cloud Native x Edge Computing
人工智能 x 云原生 x 边缘计算

在AWS中的VMware Cloud中部署

由VMware负责管理和运维

服务的修复和升级托管给VMWare Cloud

安全的数据存储和传输

备份托管

在线快照

提供增强的Greenplum Command Center

内置所有扩展组件

存储计算分离



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



GREENPLUM
DATABASE®

Open Source AceCon

2021 智能云边开源峰会

AI x Cloud Native x Edge Computing

人工智能 × 云原生 × 边缘计算

Thank You