GemFire Greenplum Integration 应用场景分析

闫刚
Pivotal大中华区大数据资深架构师

2016年11月25日

Pivotal

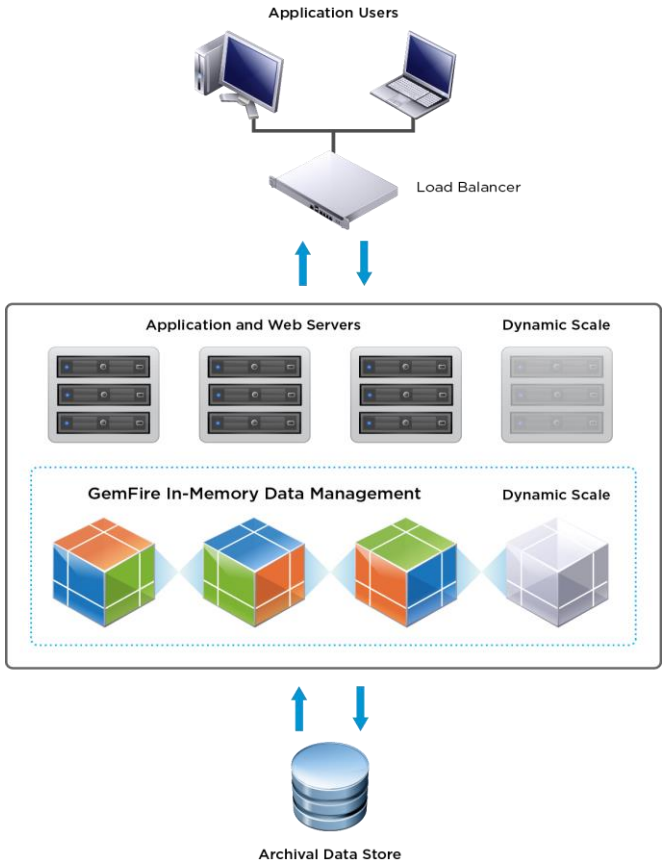
目录

- Pivotal GemFire 介绍
- GemFire Greenplum 整合

Pivotal GemFire 功能

Pivotal GemFire 是一个基于内存、具有横向扩展能力、高性能计算的分布式数据管理平台（IMDG）。
具有内存处理、数据分割（Data-partition）和并行计算（Map-Reduce)能力
可以使传统应用的性能得到数倍到几百倍的提升

- 所有计算操作都在内存，极大提高性能
- 把数据移动到中间层，更靠近于使用它的地方
- 集群支持在线热伸缩，易于适应用户数量波动大的场合
- 多层次的数据备份机制，根据项目需要提供不同级别的高可用性
- 支持跨地域分布，便于多活数据中心建设



Pivotal

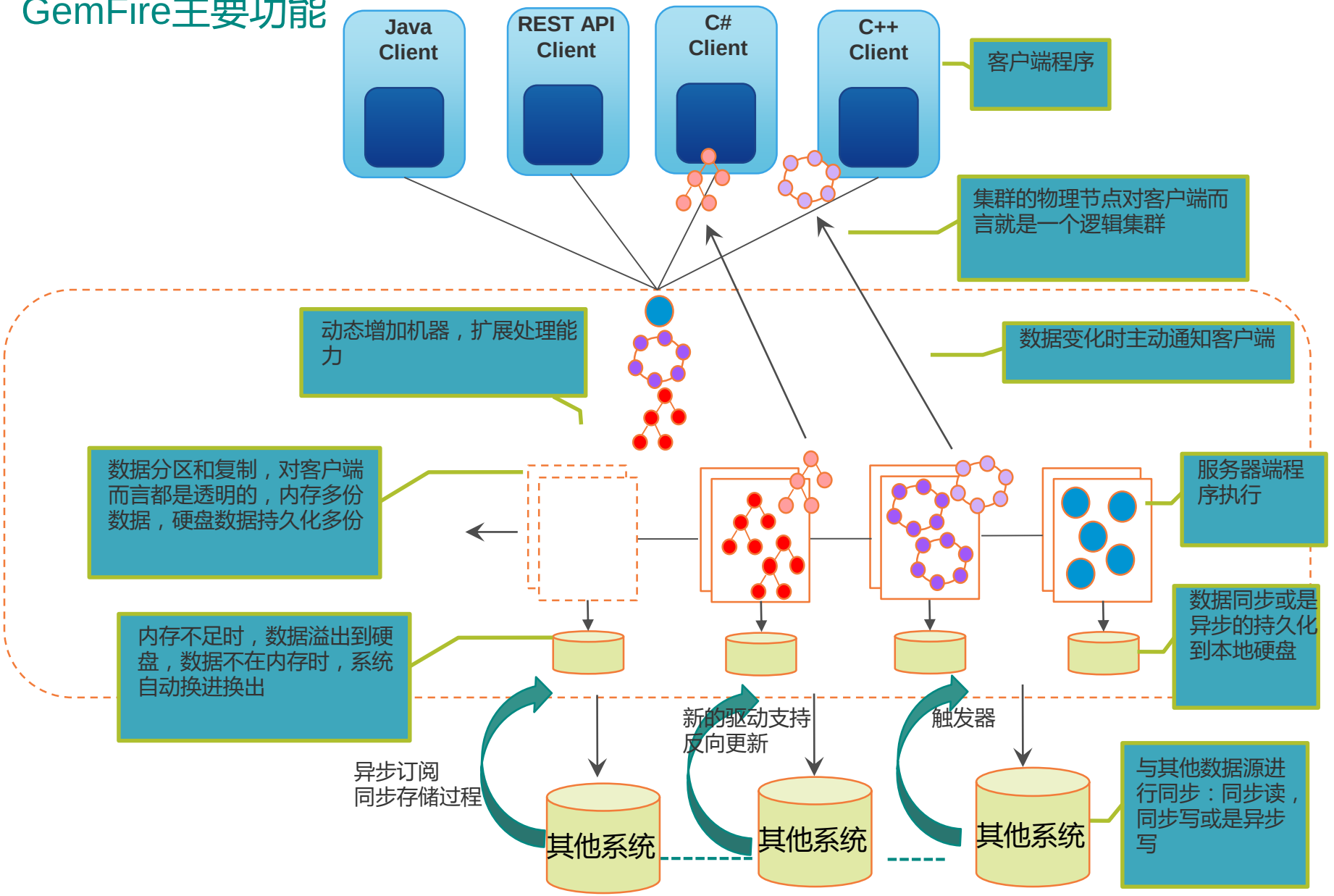
不同设备访问延迟

Numbers Everyone Should Know

L1 cache reference	0.5 ns
Branch mispredict	5 ns
L2 cache reference	7 ns
Mutex lock/unlock	25 ns
Main memory reference	100 ns
Compress 1K bytes with Zippy	3,000 ns
Send 2K bytes over 1 Gbps network	20,000 ns
Read 1 MB sequentially from memory	250,000 ns
Round trip within same datacenter	500,000 ns
Disk seek	10,000,000 ns
Read 1 MB sequentially from disk	20,000,000 ns
Send packet CA->Netherlands->CA	150,000,000 ns

Pivotal

GemFire主要功能



Pivotal

Open Source GemFire

- 2015/4 – acceptance to Apache
- 2016/11 成为 Apache TLP
- 拥抱开源社区
  - 在线 meetups
  - 更多的示例，交流分享和培训
- 参考链接
  - http://geode.apache.org/
  - https://github.com/apache/incubator-geode

APACHE GEODE

Performance is key. *Consistency* is a must.

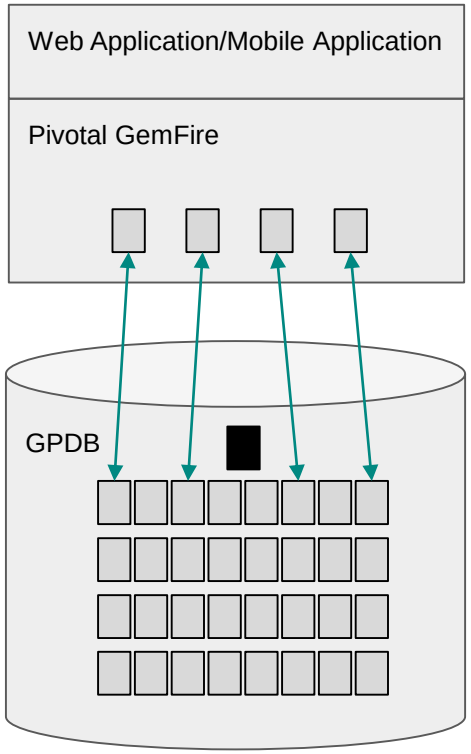
Providing low latency, high concurrency data management solutions since 2002.

目录

- Pivotal GemFire 介绍
- GemFire Greenplum 整合

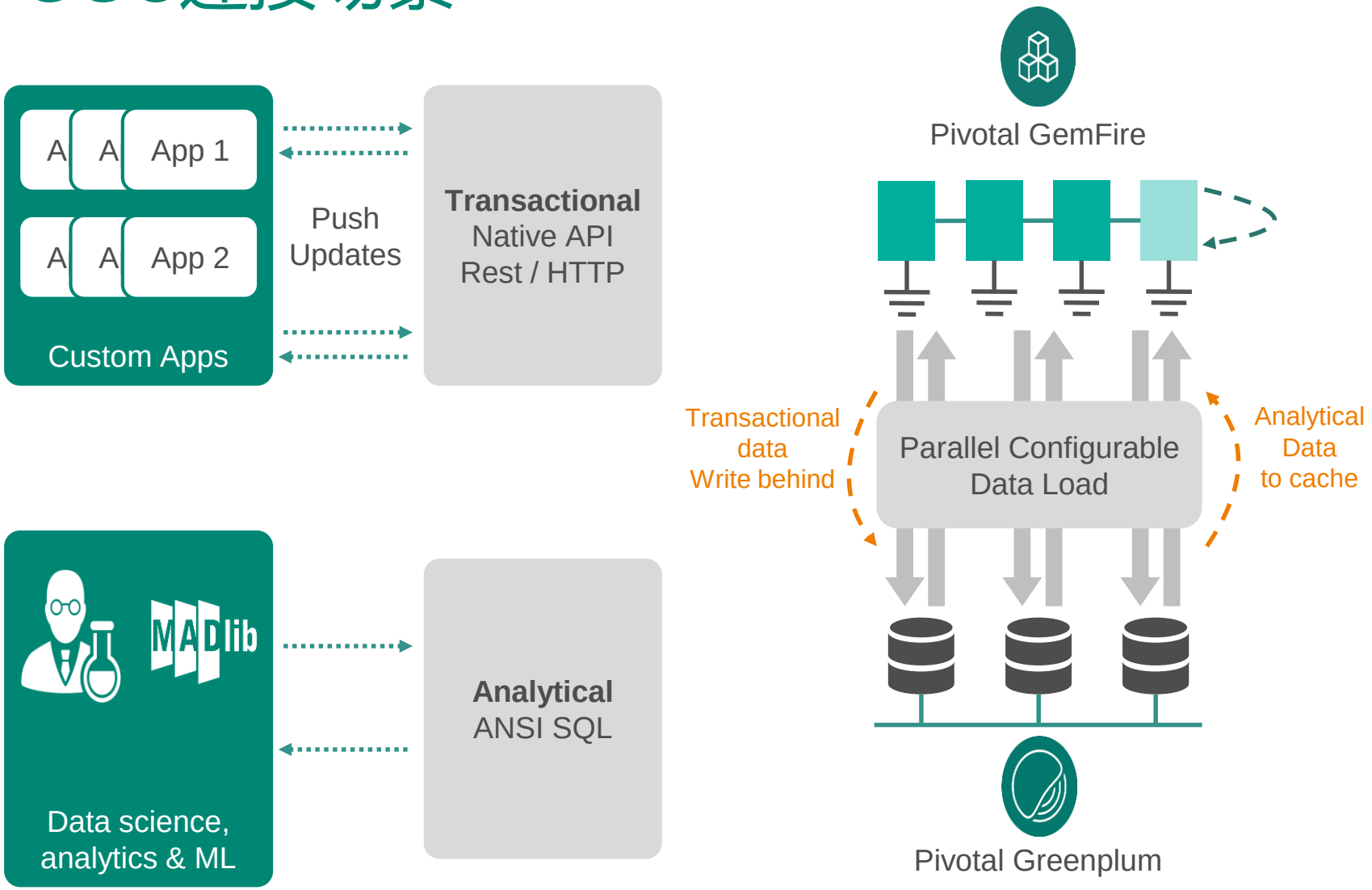
GemFire Greenplum Connector(GGC)功能

- 支持以表为单位进行数据导入导出操作，简单易操作
- 通过在GemFire数据节点和Greenplum Segments节点之间，建立直接连接的方式，大幅提高了数据加载的性能
- 通过配置的方式，在大幅减少了定制化开发的工作量的同时， 提高了整个系统的稳定性



Pivotal

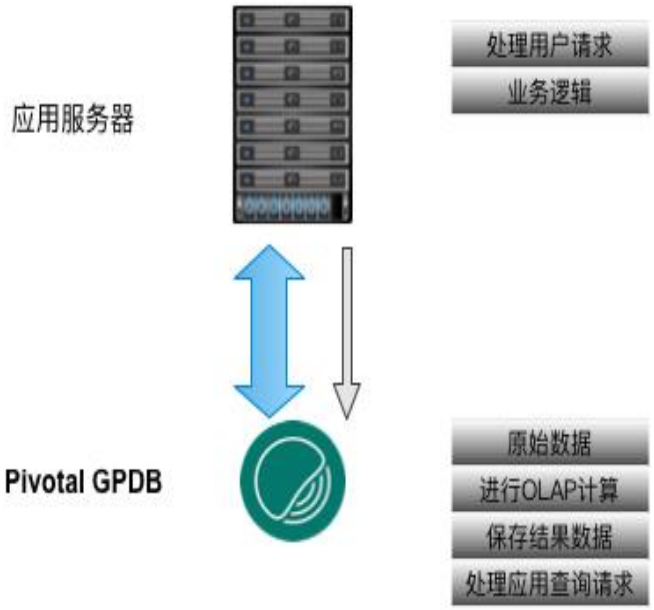
GGC连接场景



Pivotal

传统GPDB 处理OLAP场景

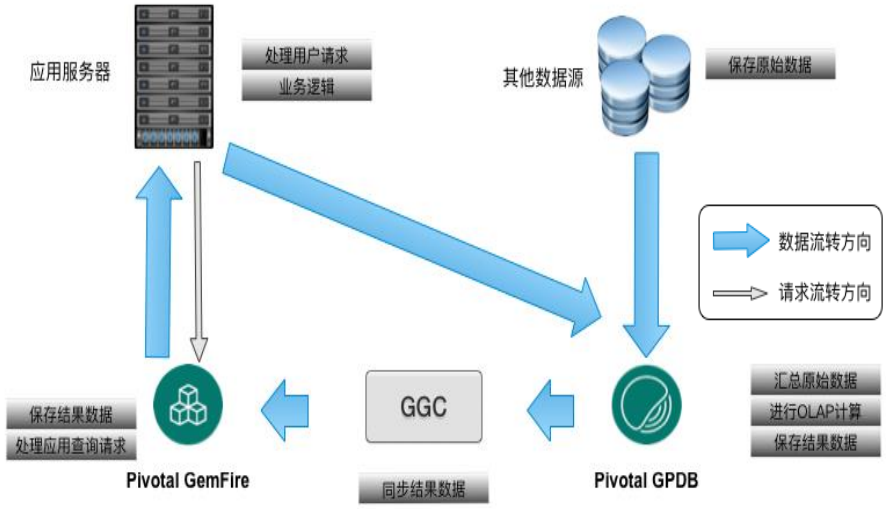
- 设计思路：
  - Greenplum保存原始数据和结果数据
  - Greenplum进行OLAP计算
  - Greenplum接收数据更新请求
  - Greenplum处理应用查询请求
- 待优化点：
  - Greenplum的小批量写性能相对较差
  - 对于高并发支持相对较弱（在100KTPS+数量级别）



Pivotal

适用GGC场景1 -- T+1 OLAP场景

- 优点：
  - 使用Greenplum分析数据，开发工作量比较低
  - GemFire向应用展示分析结果，可以承载高并发量，大幅降低响应时间
- 场景：
  - 已部署GPDB，需要对分析结果的查询进行加速
  - 数据量比较大，对于OLAP分析耗时无严格要求
  - 高并发报表查询，T+1数据分析

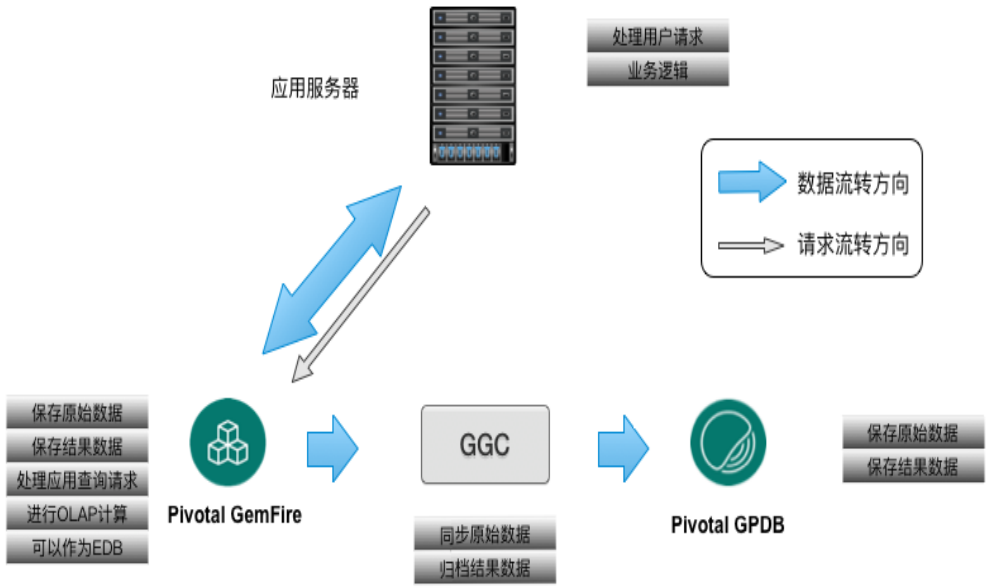


Pivotal

适用GGC场景2 -- 准实时OLAP场景

- 优点：
  - 可以提供更快的分析速度，更快的数据导入速度
  - 以GemFire为核心构建企业数据总线，整体架构清晰

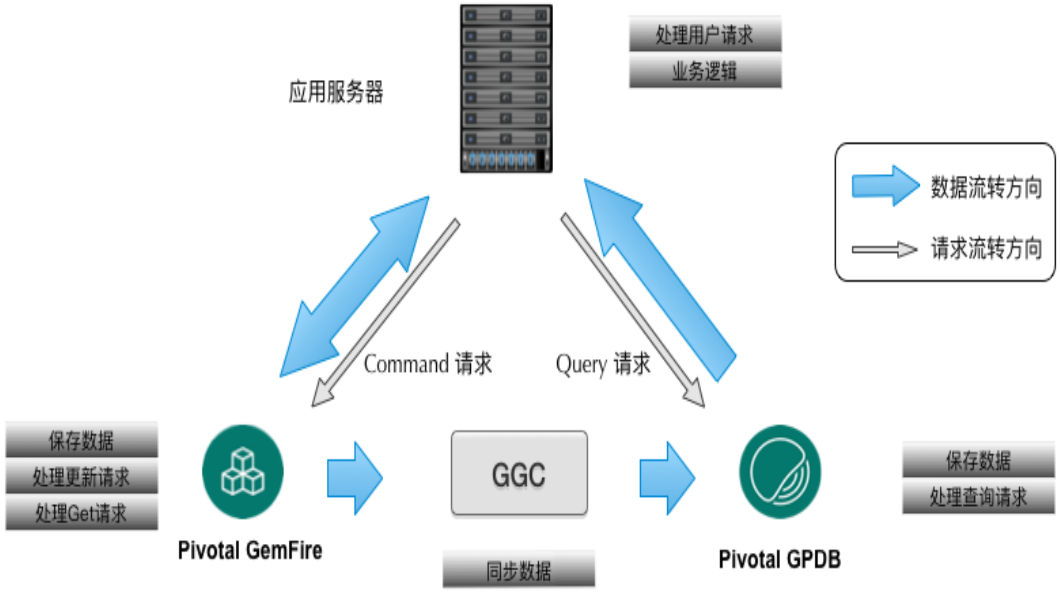
- 应用方向：
  - 数据量在TB级，对OLAP分析耗时有非常严格的要求，准实时报表
  - 新建系统，特别是并发请求波动很大，需要在线热伸缩要求
  - 数据总线，打破各项目或各系统数据隔离的需要



Pivotal

适用GGC场景3 -- 在线查询场景 CQRS

- 优点：
  - 请求分离：使用GemFire接收前端应用的CUD请求和Get请求；使用Greenplum接收前端应用SQL查询请求；
  - 高吞吐量，低响应时间
  - 使用标准SQL
- 应用方向：
  - 支持访问量大、高性能、可伸缩且允许最终一致性的互联网站点，或者互联网金融应用



Pivotal

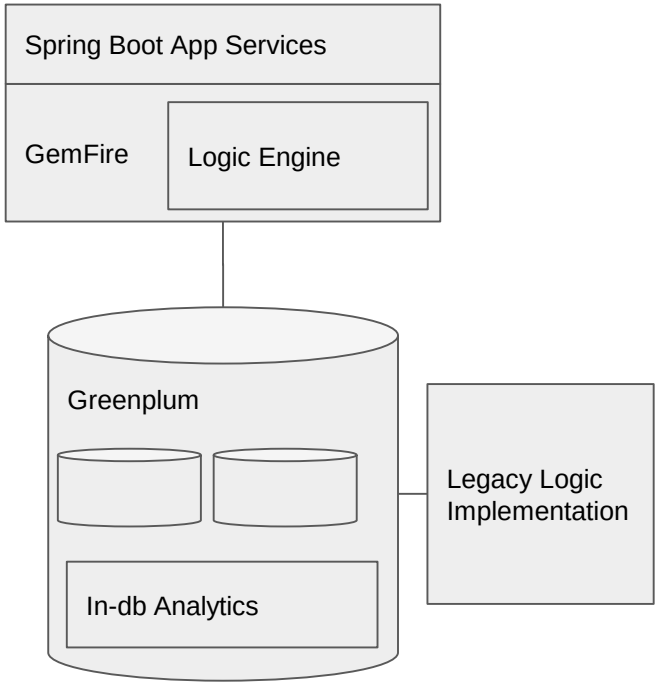
参考案例 – 反欺诈系统

准备
- 汇集数据
- 格式转换 (ETL)

评分
- 模型评估

初评
- 业务逻辑评估
- 鉴权

响应
- 支付
- 停止支付

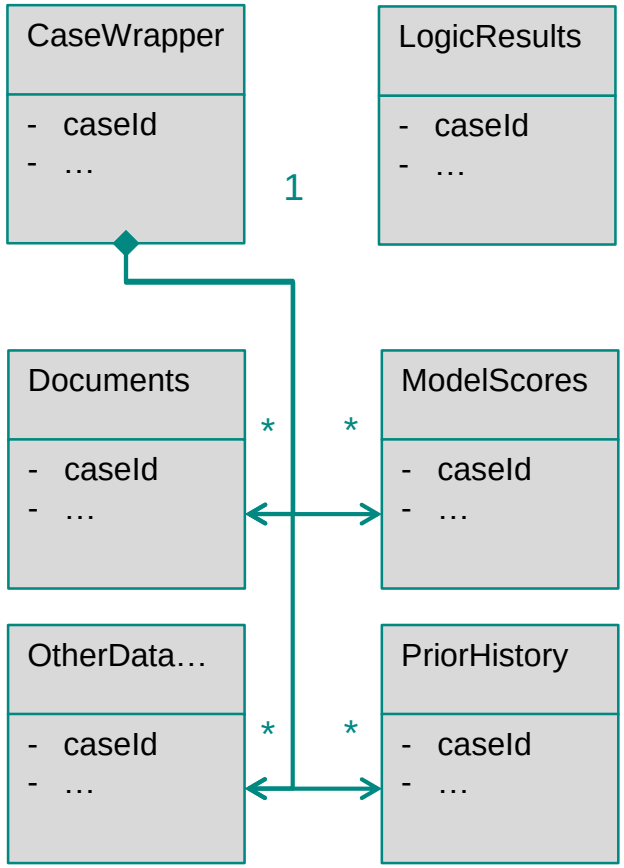


Pivotal

Pivotal

附录：Case Study G2C Configuration Details

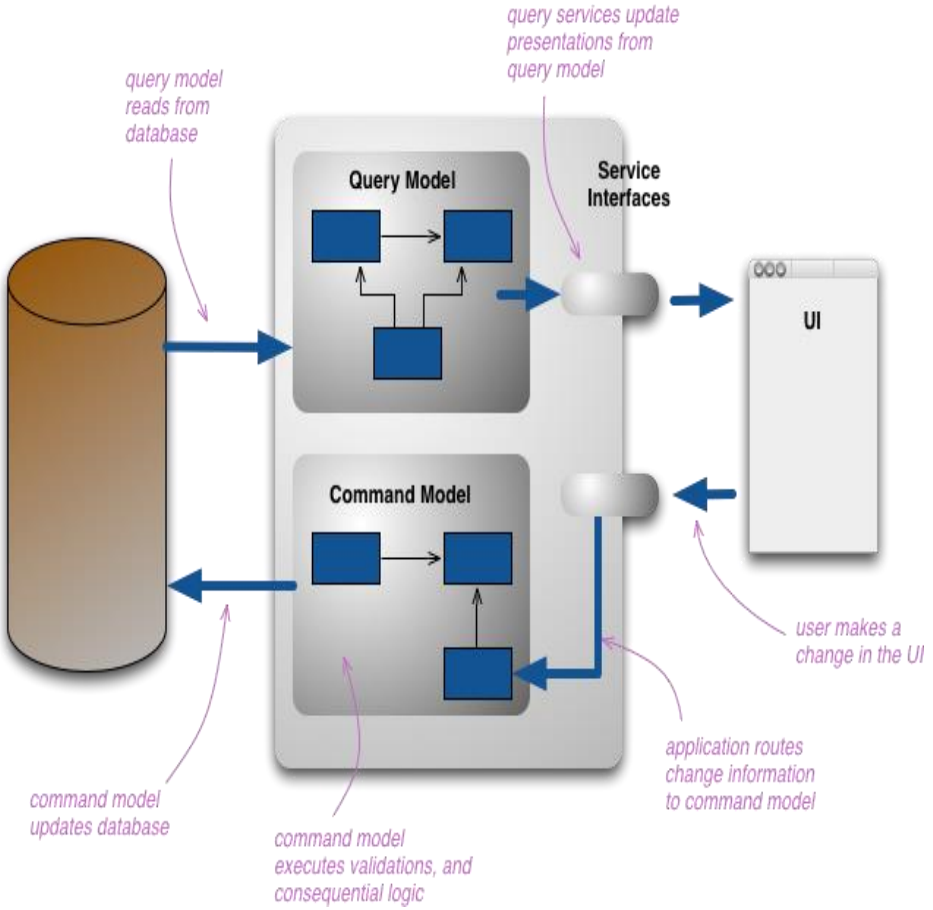
- Existing required domain objects
- Multiple many-to-one groupings
- Wide tables / objects (500+ fields)
- Data Collocation configured on caseId
- Source tables wrapped in views



Pivotal

附录：CQRS

- Greg Young，Martin Fowler
- Domain-Driven Design
- Query, command seperate



Pivotal