唯品会海量数据实时OLAP分析实践

唯品会大数据 谢麟炯

唯品会
全球精选 下品特卖

谢麟炯

唯品会大数据平台高级技术架构经理，主要负责大数据自助多维分析平台，离线数据开发平台及分析引擎团队的开发和管理工作，加入唯品会以来还曾负责流量基础数据的采集和数据仓库建设以及移动流量分析等数据产品的工作。

目录
CONTENTS

01 海量数据实时OLAP场景的困境

02 唯品会大数据实时OLAP升级过程

03 唯品会在开源计算引擎上所做的改进

04 OLAP方案升级方向

海量数据实时OLAP的困境

- 数据量迅速膨胀，传统的RDBMS已无法满足存储需求

海量数据实时OLAP的困境

- OLAP查询速度变慢
- ETL数据处理效率降低

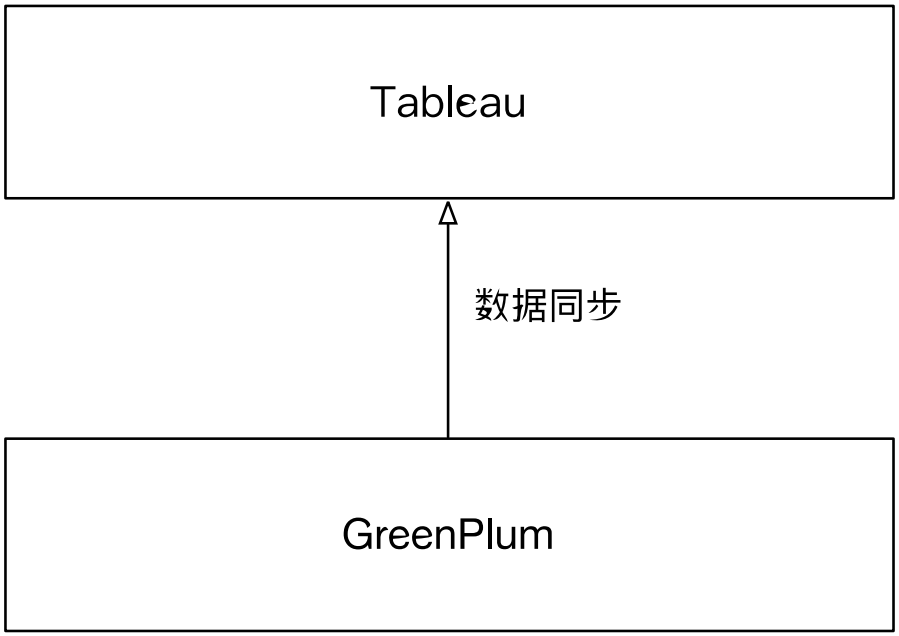
海量数据实时OLAP的困境

- 新业务开发周期变缓
- 旧业务更新周期拖延

快快快

业务部门的需求和现状的冲突急需解决

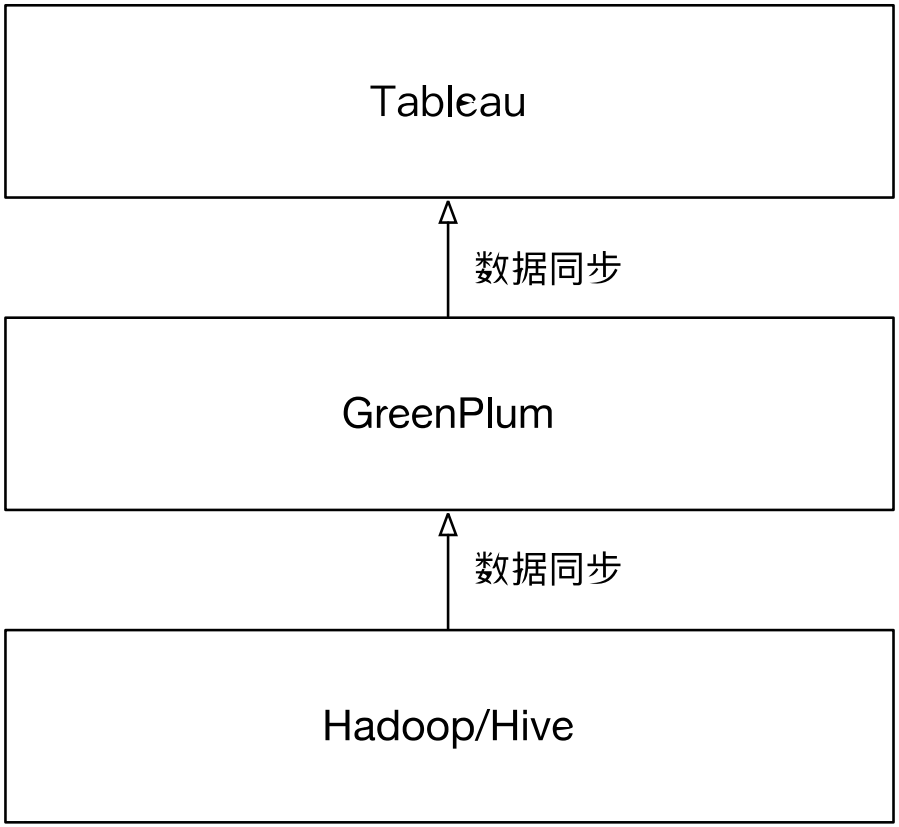
唯品会大数据实时OLAP升级过程-第 0 阶段



- 优点
  - 使用商业敏捷BI工具，快速满足OLAP报表需求
  - GreenPlum MPP数据库作为数据仓库的存储、计算介质
  - 数据仓库和OLAP分析混用同一个数据库实现。

- 缺点：
  - 海量日志数据的接入使得存储和计算资源迅速枯竭，但RDBMS数据库水平扩展能力有限

唯品会大数据实时OLAP升级过程-第 1 阶段



- 优点：
  - 根据服务提供的目标不同，分拆数据仓库和OLAP分析库，Hadoop/Hive负责数据仓库部分，GreenPlum专注负责OLAP部分

- 缺点：
  - 数据需要在两个不同的DB之间同步，数据冗余且可能存在不一致性
  - BI工具没有进化，但用户的需求正在不断进化

我们需要一个新的BI方案

足够灵活

- 维度和指标可以任意组合
- 可以快速得到数据，不需要长时间的等待

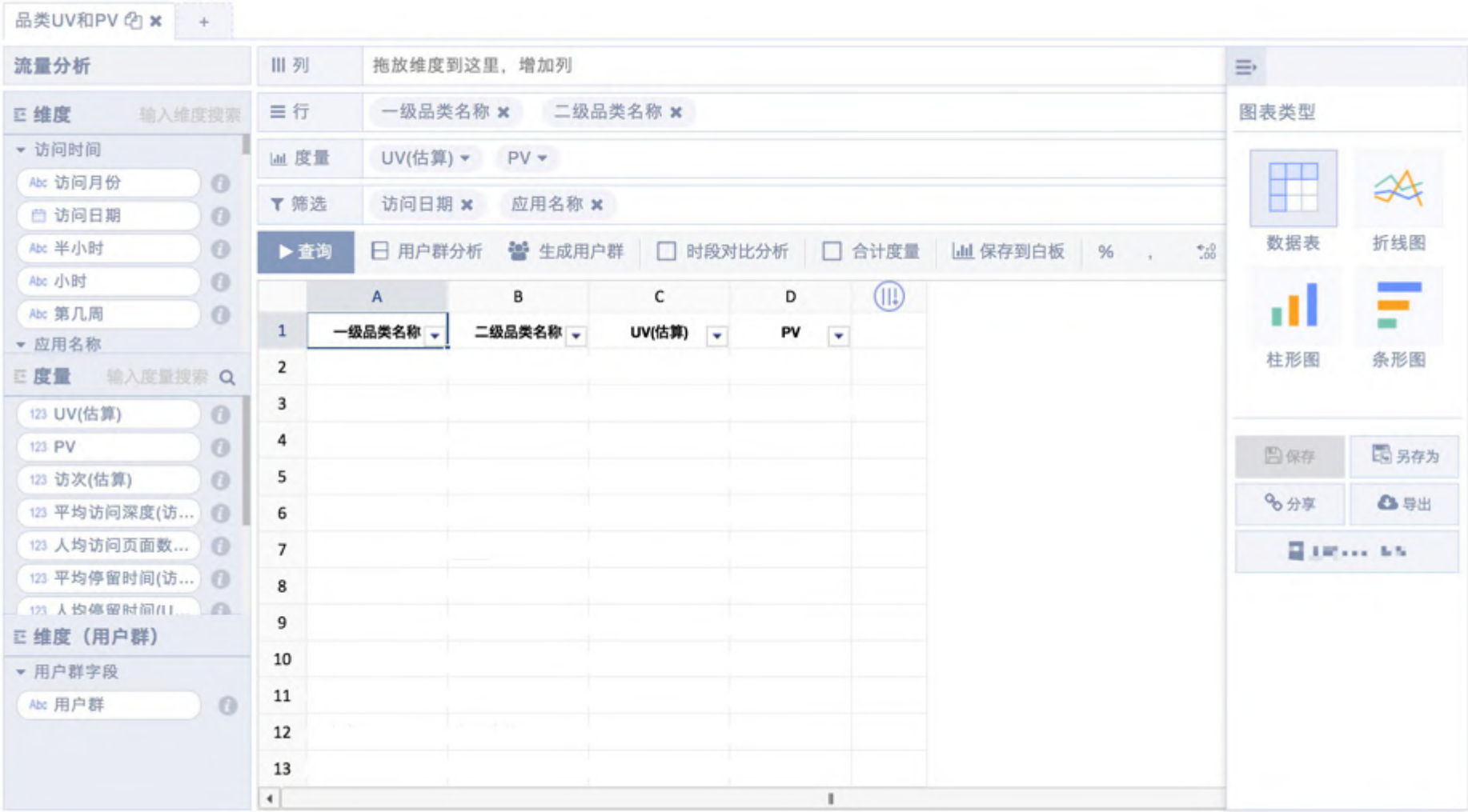
门槛要低

- 入门简单，业务人员做OLAP分析不能像Tableau一样学习曲线长
- 如果能用语言描述自己的需求来做OLAP分析会比用SQL来描述更好

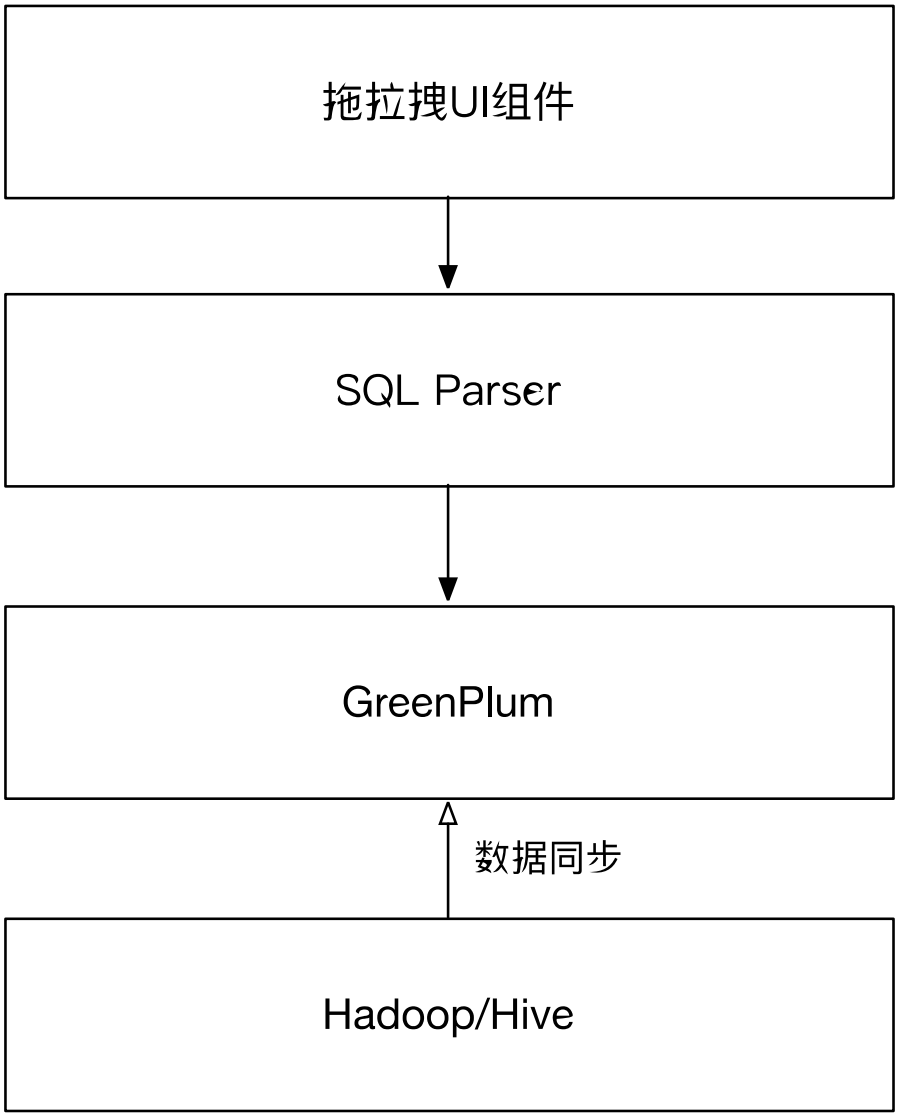
开发周期短

- 业务人员厌倦了任何数据都需要提需求->排期->实施->变更->排期->实施的更长时间的等待

唯品会大数据实时OLAP升级过程-第 2 阶段



唯品会大数据实时OLAP升级过程-第 2 阶段



- 优点
  - 业务人员通过拖拽指标维度确定组合及业务含义
  - 由SQL Parser将用户对数据的描述转化为SQL并进行查询
  - 数据模型一次建立，多次复用

- 缺点：
  - OLAP数据是通过预先定义模型设计开发准备的，用户可自定义查询组合但数据的范围比较有限

我们需要一个新的OLAP计算引擎

无预计算模型

- 没有传统意义上的数据汇总层，OLAP分析要从DW层明细数据上直接计算
- 对于缓慢变化维或迟到维等维度变化不需要手工刷新数据

速度足够快

- 数据平均要在10秒内返回
- 支持Join来满足维度建模的模型设计

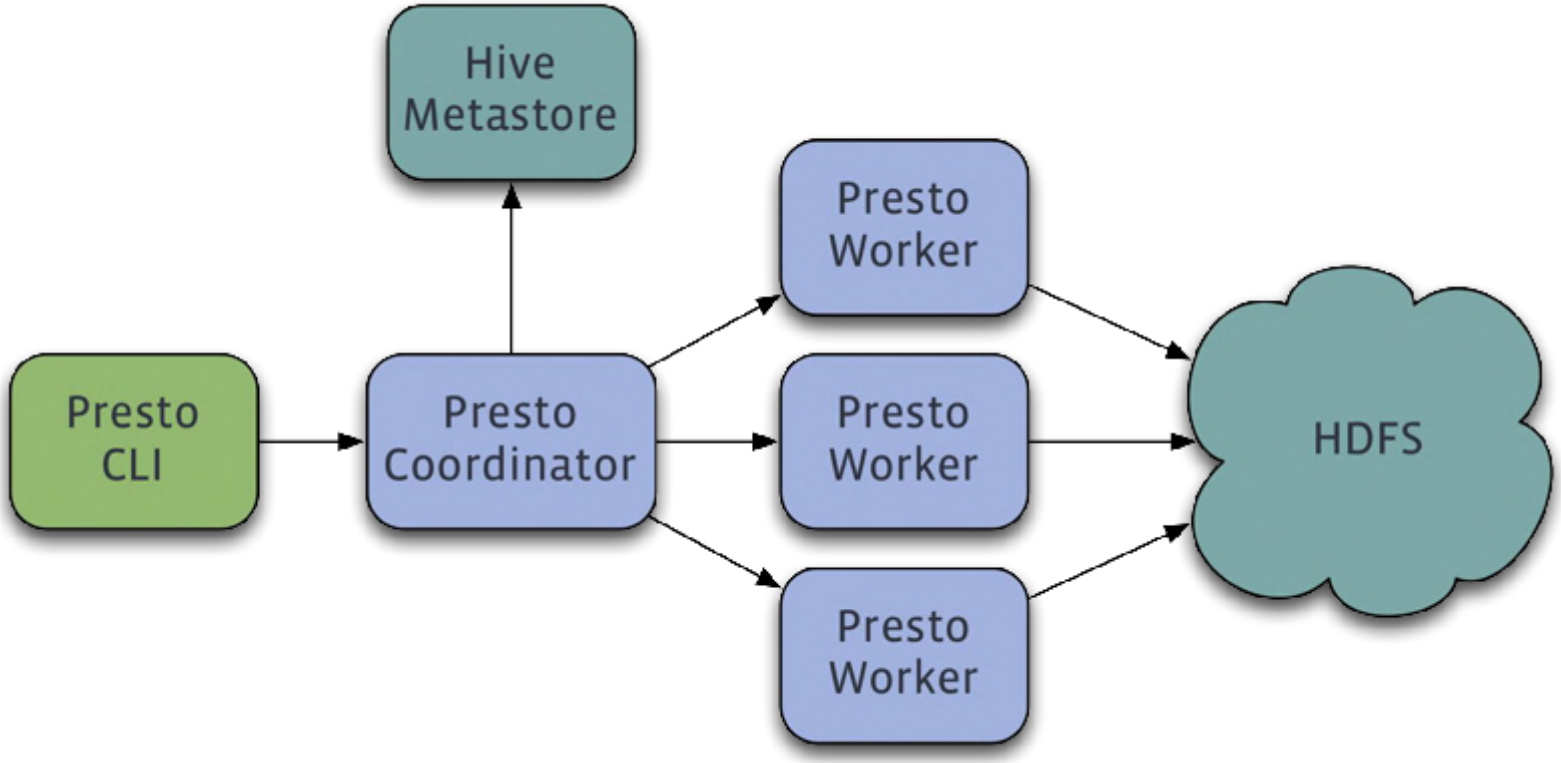
支持横向扩展

- 计算能力可通过计算节点扩容获得提高
- 数据不需要在两个Db之间同步减少冗余和不一致性发生的可能性

我们需要一个新的OLAP计算引擎

目标	解决方案
查询速度要快	1. SQL内高并发优于顺序读 2. 纯内存计算优于内存+磁盘的计算
数据模型是维度建模	1. 方案必须支持Join
数据不需要同步	1. 数据要存储在HDFS 2. 计算引擎要能够感知到Hive的Metadata
可通过扩容提高计算能力	1. 计算引擎要无状态 2. 计算节点之间互相无依赖 3. 存储和计算分离
方案成熟稳定	1. 在大型互联网公司有成功经验

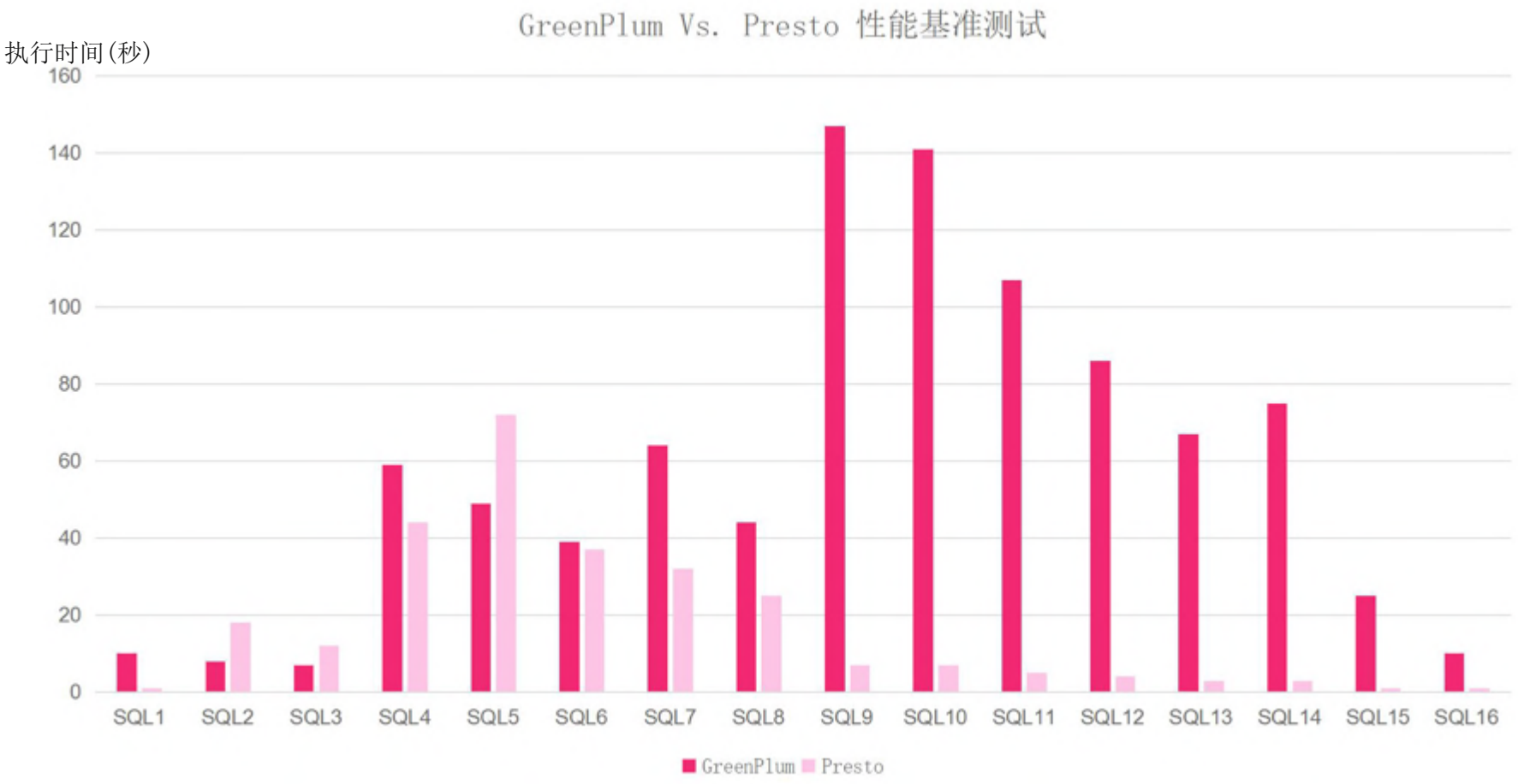
Presto - Facebook贡献的开源MPP OLAP引擎



为什么是Presto

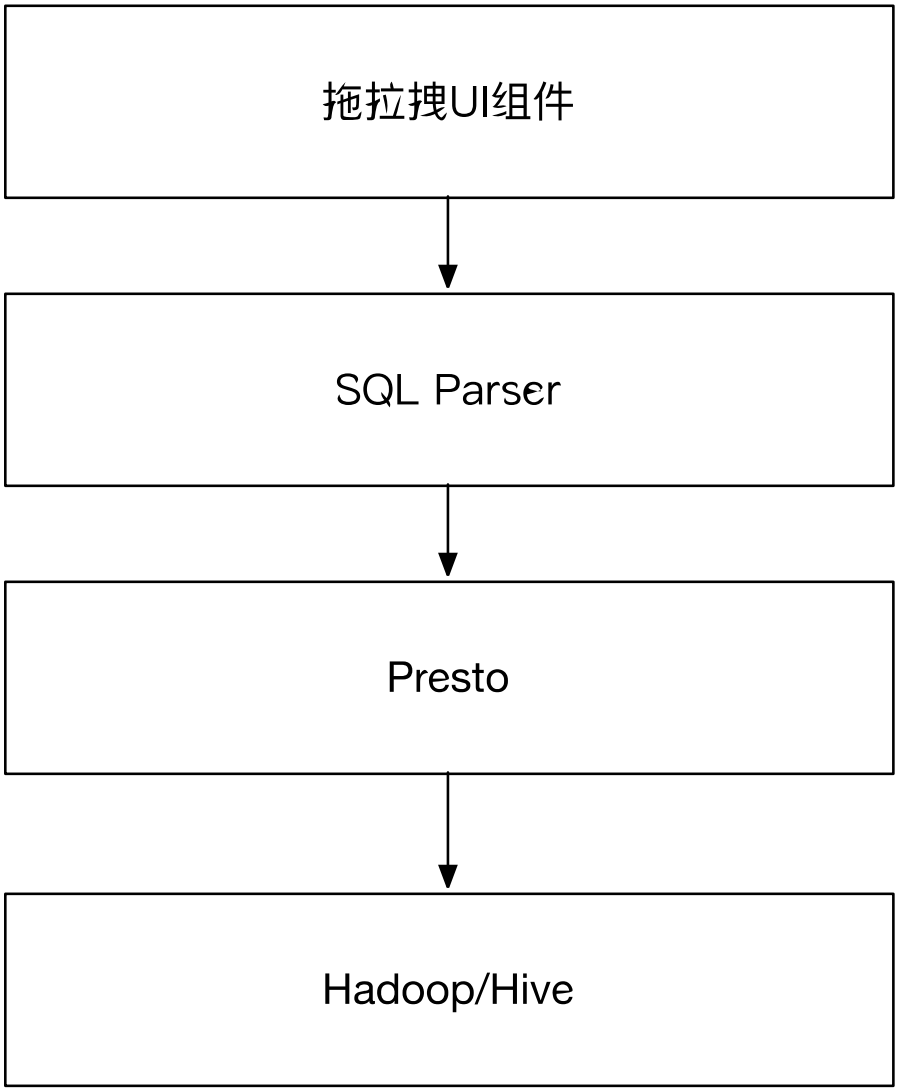
解决方案	Presto
SQL内高并发查询	☑
纯内存计算	☑
支持Join	☑
数据存储在HDFS	☑
感知Hive的Metadata	☑
计算引擎要无状态	☑
计算节点之间互相无依赖	☑
成熟方案	☑

Presto性能测试



- 整体而言，相同节点数的Presto比Greenplum性能提升70%

唯品会大数据实时OLAP升级过程-第 3 阶段



- 优点
  - Presto是Facebook开源的MPP计算引擎，直接读取Hive数据无需单独存储，同时计算能力强，横向扩展能力强
  - 中位数5秒（平均15秒）内返回所有类型的OLAP查询

- 缺点：
  - 每次OLAP Ad-hoc查询都需要从DW层拉取数据查询，当用户反复发起相同条件的查询时，资源存在浪费的情况

最快的计算方法是不计算

缓存复用

- 完全相同条件的查询结果不用走Presto，应该直接得到上次查询的结果

缓存二次加工

- 如果条件不完全相同，但新查询的条件是旧查询条件的子集，也可以不使用Presto查询，通过二次加工缓存结果复用缓存

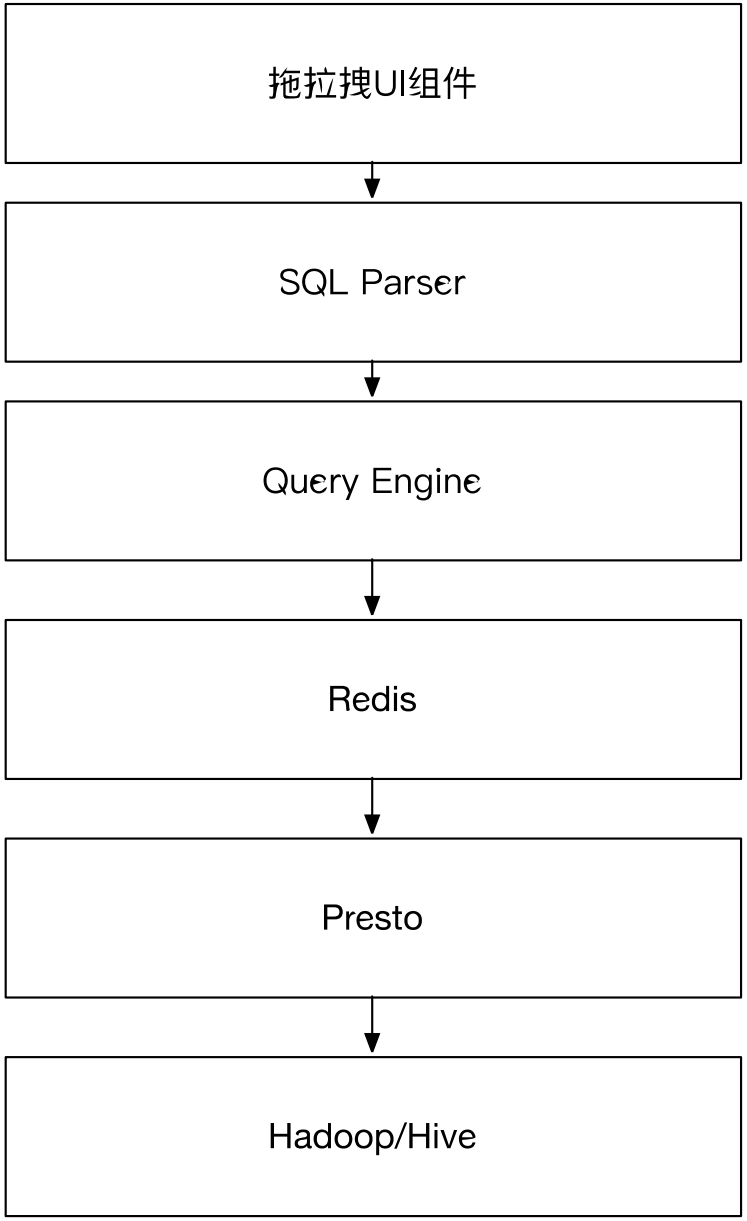
生命周期管理

- 缓存的生命周期可由缓存自己根据源数据的生命周期进行失效管理

Redis - 成熟的缓存方案

目标	解决方案
SQL和结果集要可以1对1匹配	Redis key-value存储符合需求
结果可以快速匹配	内存存储，Key匹配快速
生命周期管理	Redis提供API，可二次开发与ETL调度系统联动自动清理
复用缓存结果	Redis不支持，但可以自己二次开发

唯品会大数据实时OLAP升级过程-第 3.5 阶段



- 优点
  - 缓存平均命中率15%，最高命中率60%，减少了不必要的Presto查询，提高了Presto集群的服务能力
  - 命中缓存的数据返回速度从中位数5秒提升至0.5秒内

- 缺点：
  - 查询Pattern第一次出现的时候，Presto查询的时候还是相对比较慢

空间换时间—OLAP分析的另一条途径

空间换时间

- 如果OLAP ad-hoc查询的时候从已经计算好的结果中直接提取，速度会大幅提升

要支持维度建模

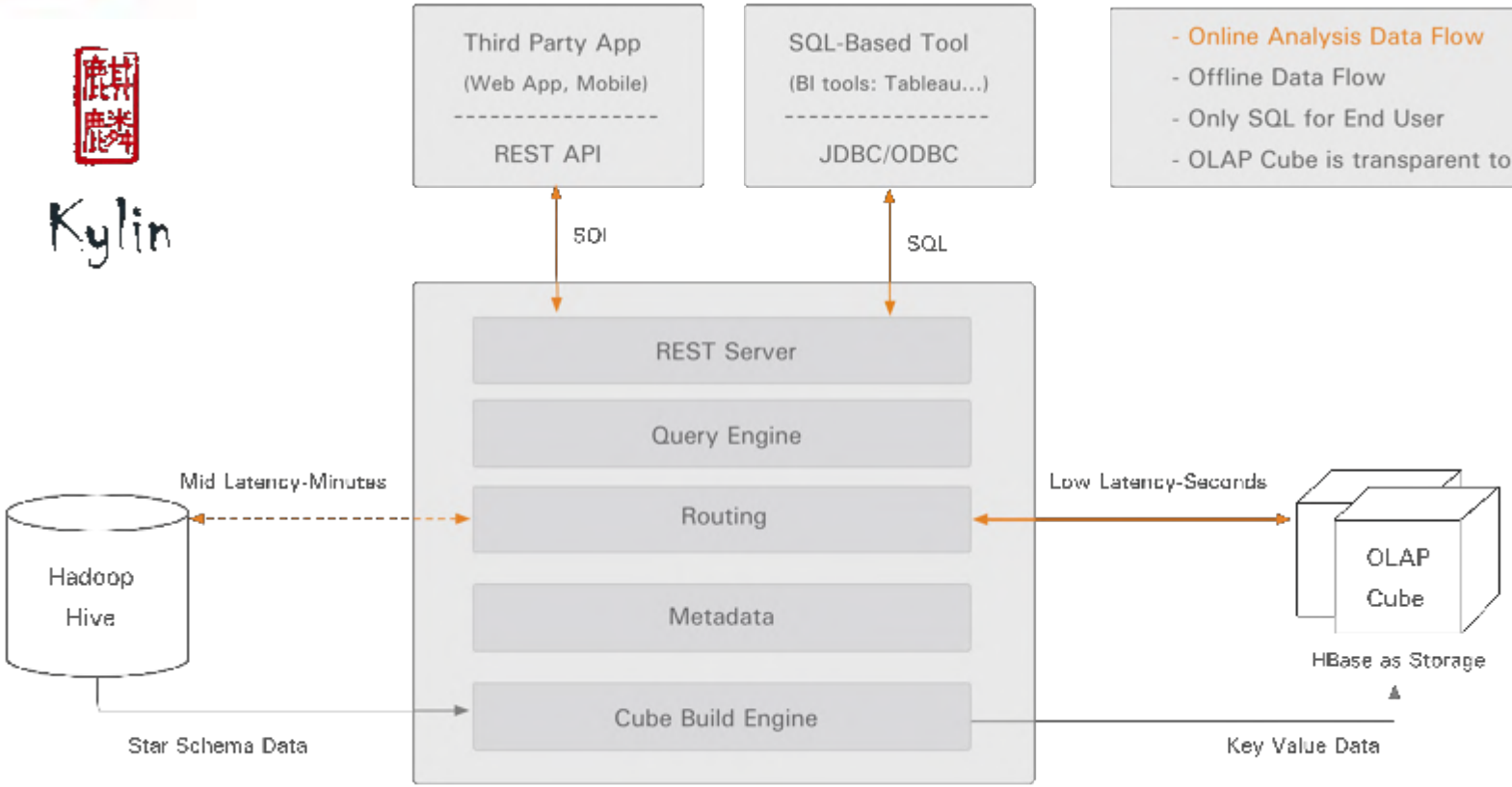
- 维度建模中事实表和维度表的Join在新的引擎中也需要支持

数据管理简单

- 预计算的过程和结果同步不需要二次开发支持
- 由于预先设计的CUBE存在过度设计的维度和指标没人查询的情况，要可以灵活的删除冗余

Kylin - eBay贡献的开源MOLAP引擎

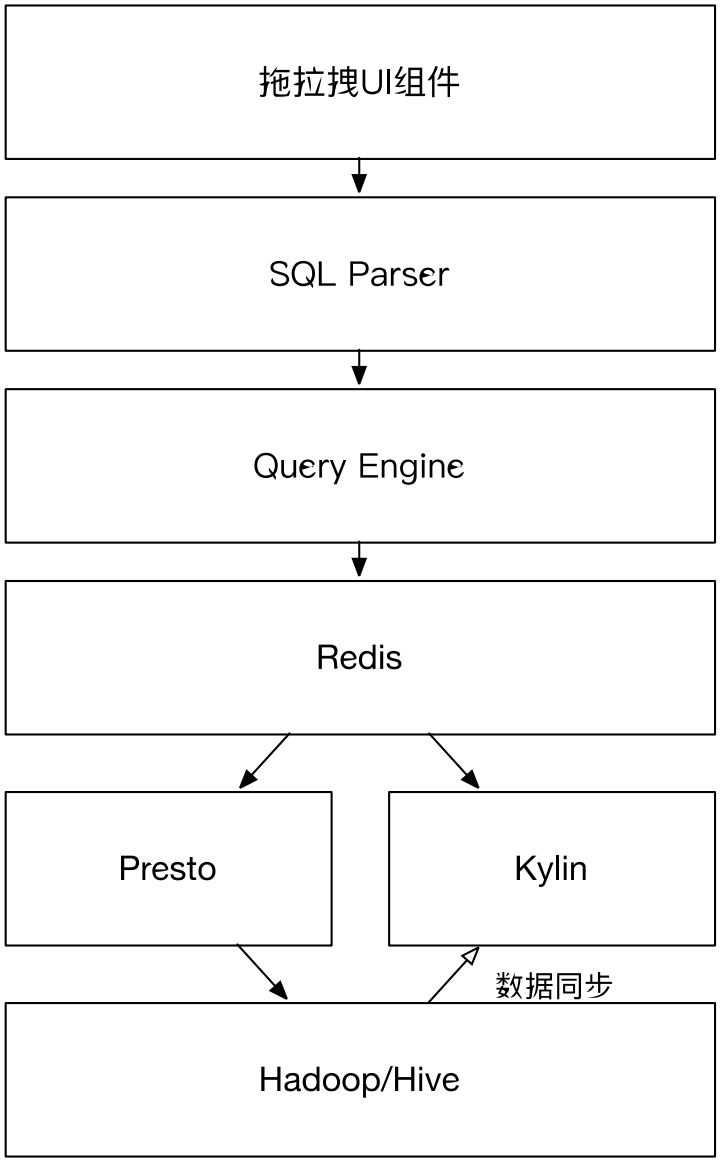
Kylin



为什么是Kylin

目标	解决方案
空间换时间	Kylin仿照Oracle CUBE开发，通过预定义/计算CUBE的方式保证Ad-hoc速度
支持维度建模的模型	Kylin支持在定义模型时的Join
数据同步管理简单	1. 预计算调度和数据管理均由Kylin自行管理 2. 有开放的API用于与ETL调度系统打通 3. 有简单的Web Console用户配置和管理CUBE
方案成熟稳定	在大型互联网公司有成功经验

唯品会大数据实时OLAP升级过程-第 4 阶段



- 优点
  - 核心数据使用Kylin预建CUBE，提高常用数据首次查询的响应速度，90%的查询在5秒内，中位数响应时间1.2秒
  - Kylin查询平均覆盖率15%，最高25%，大幅提升了这些核心查询的性能，同时针对流量数据的预汇总，降低了Presto从RawData汇总的资源消耗

- 缺点：
  - 数据管理成本的提高，通过在开源基础上二次开发Cuboid监控和裁剪降低

OLAP分析的技术进化是一个迷宫而不是金字塔

适合自己业务场景的方案才是好方案

唯品会在开源计算引擎上所做的改进

Presto

- 提升查询性能
- 减少计算量
- 提高实时性

Kylin

- 优化维表查找
- 提升CUBE利用率

Presto上的改进

- 提升查询性能
  - 新增Hint语法，可在SQL Join级别动态设置策略，通过编译时让join在replica和distribute两者之间设置，提高Join效率
  - 监控告警Join数据倾斜，通过减少数据倾斜提高执行效率
  - 增加多集群Load Balance，可平衡不同集群间计算量
  - 经过改造，Presto的查询实时性大幅提升

Presto上的改进

- 减少计算量
  - 新增Kylin Connector，通过CUBE探嗅自动匹配SQL子查询中可以命中Kylin CUBE的部分，从Kylin提取数据后做进一步的计算，降低查询计算量
  - 经过改造，Presto升级为Hybird OLAP引擎，同时支持ROLAP和MOLAP两种模式

Presto上的改进

- 提高实时性
  - 重写Kafka Connector，支持热更新Kafka中Topic、Message 和表/列的映射定义
  - 支持Kafka按offset读取数据，支持PB格式，提高Kafka数据源的读取效率
  - 经过改造，Presto不仅是离线OLAP引擎，准实时数据处理的能力也得到提高

Kylin上的改进

- 优化维表查找
  - 通过引入Presto解决Kylin亿级维表实时Lookup OOM的问题，通过Presto查询替换了原有复杂的维表映射值查找机制
  - 经过改造，唯品会版的Kylin相比开源版本极大的扩展了对业务场景的支持程度

Kylin上的改进

- 提升CUBE利用率
  - 开发CUBE Advisor，通过统计分析总结合适的维度和指标组合辅助开发选择判断新建CUBE的策略，减少冗余和经验判断上的误差
  - 提供CUBE命中率监控，形成CUBE新建、使用到总结升级的闭环
  - 经此改造，CUBE命中率大幅提高，减少了资源的浪费提升了响应速度

OLAP方案升级方向

Presto

- 探索用RowID级别的索引的存储格式替换现有RowGroup级别索引的ORC File，在数据Scan级别尽可能精确的获取所需的数据，减少数据量
- 提高OLAP查询的并发度，应对大量用户并发OLAP分析场景

Kylin

- 尝试Streaming Cubing，Kylin OLAP分析从离线数据向实时数据迁移
- 探索Lamda Cubing，实现实时离线CUBE无缝融合

下一代方案

- 更强的实时离线融合能力
- 更强原始数据和汇总数据的融合
- 更强的数据处理能力

让大数据成为唯品会的增长引擎

加入我们！

感谢聆听

THANKS!

唯品会