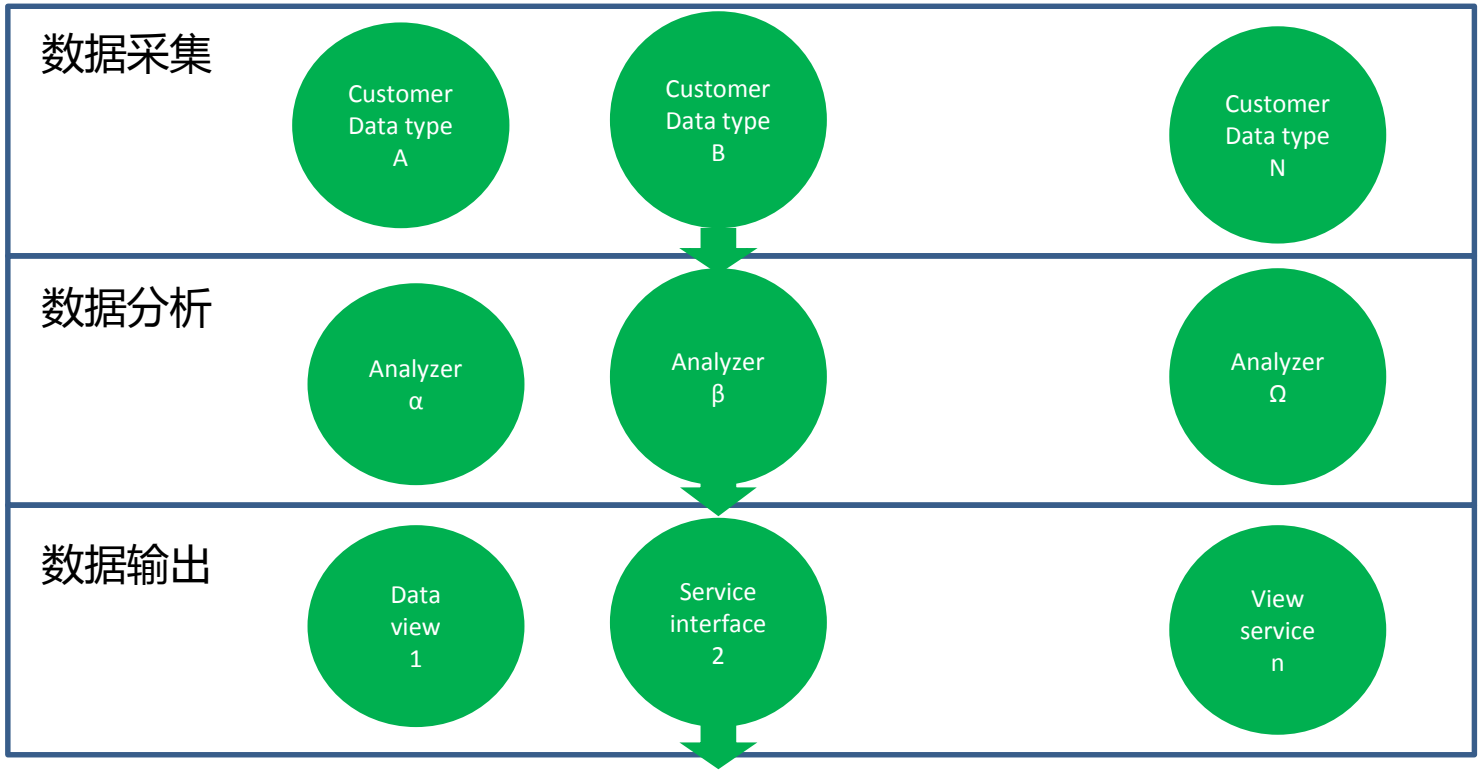
为什么我们需要GREENPLUM

北京博雅立方科技有限公司
J.W.

2016/10/15

问题

系统


数据采集
Customer Data type A
Customer Data type B
Customer Data type N
数据分析
Analyzer α
Analyzer β
Analyzer Ω
数据输出
Data view 1
Service interface 2
View service n

要求

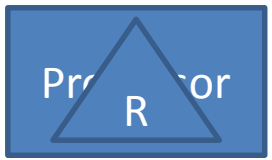
- 自定义的源，自定义的格式，任意的扩展
- 数据密度相对高
- 分析工具与手段众多，数据强类型
- 服务的响应时间敏感

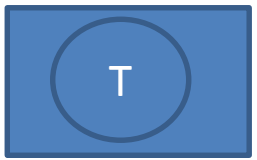
困惑

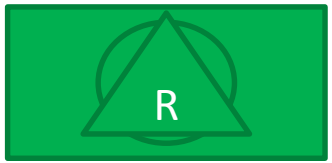
- 数据的存储
  - Hbase，redis，mangodb…?
- 数据的分析
  - Hadoop，spark，R…?
- 这些还不够用
- 不可预先处理的需求和实时响应的矛盾

数据存储器与分析器

- Hbase,redis,mangodb不具备数据分析能力或者很弱
- Hadoop，spark不适合实时计算
- 大数据量的实时计算与响应，必须要求数据存储与演算一体化
- 数据的存储系统必须具备分析计算能力，包括强数据类型，运算操作符，处理数据的逻辑关联（笛卡儿积、集合运算），支持复杂处理过程（函数，批处理，事件响应...）
- 关系型数据库——数据的计算器







POSTGRES

多样的数据类型及处理能力

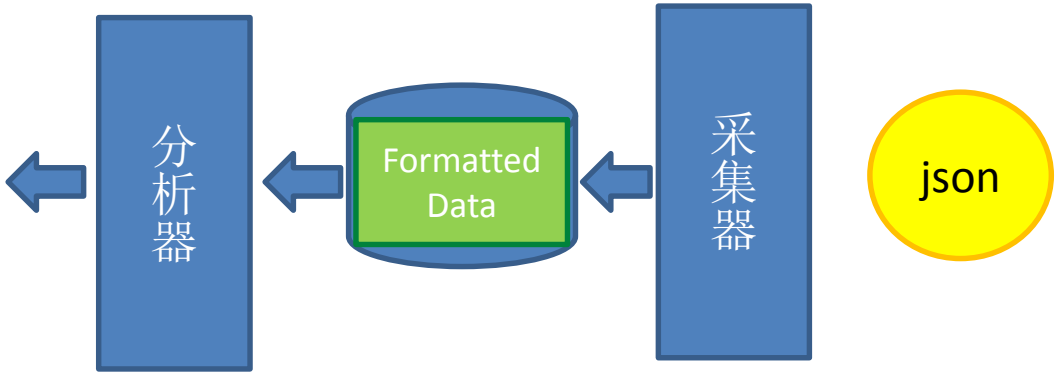
- 支持数据类型的多样性，针对不同类型的数据进行处理
- 如地理和空间位置信息，格式化数据信息（网址、邮箱、ip），数组...
- 自定义数据类型及其处理函数
- 复合数据类型
  - **CREATE TYPE** item **AS** (
    name text,
    supplier_id integer,
    price numeric
    );
- 可扩展数据类型的框架
  - **struct pg_usertype_t{unsigned char data[LEN];};**

    Datum usertype_in(PG_FUNCTION_ARGS)

    Datum usertype_out(PG_FUNCTION_ARGS) ...

自定义数据格式与强类型的矛盾

- 数据schema的可定制，决定了数据流的不确定性。
- 数据在存储器上即可进行业务逻辑处理，决定了数据流必须能转化为强类型。
- 数据采集模块与业务逻辑无关，不宜知悉具体的数据类型。
- 采集存储的数据需长期保持完整的原始信息以备后用，并容易根据不同的业务需要分解和转化成各种格式化数据表（schema）
- ——Json　type



以及其它

- 大数据量下处理能力的稳定性。
- 弹性灵活的集群扩展能力。
- 强大、丰富的工具集和可扩展的演算和处理能力（hyperloglog）
- 开放的环境，易于定制和扩展以及高质量社区支持。
- 。。。

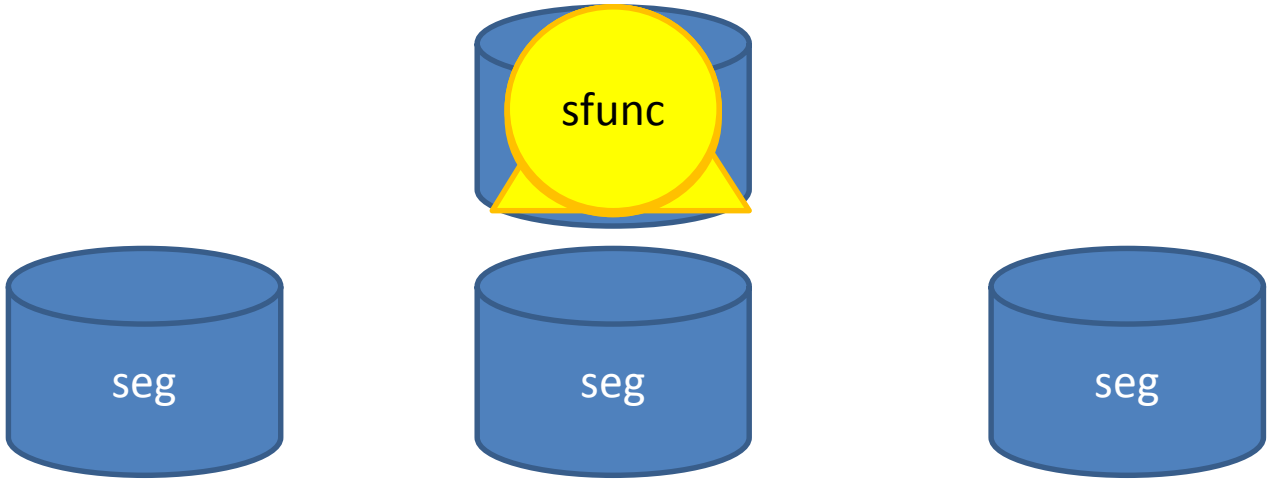
GREENPLUM

现实困境

- Postgres仍然不够，数据量与响应时间存在两难
- 传统数据库存在无法逾越的边界——cpu的能力
- 单处理器的能力难以大幅提升，瓶颈无法突破，计算能力的扩展方向转向多核心
- 并行计算成为当前提升计算能力的主要发展方向
- 传统数据库的任一演算局限于单一进程（线程）之中，计算能力具有天然的边界
- 信息时代来临太快，全社会的数字化发展的速度远远超过cpu发展的速度
- 众多数据使用者面临的问题，已经超越数据库的边界，数据库技术不作革新，将无法使用
- 就我们面临的问题而言，典型计算达到响应时间要求的，数据量不宜超过million，现实的问题规模为billion

Greenplum数据仓库

- 继承postgres的优良特性，容易吸收postgres生态中的有用资源
- 克服传统数据库的物理边界，并行计算的透明化(简化开发)
- 面向大数据量的体系架构设计(segment、列式存储)
- 方便灵活，容易定制和扩展的分布与聚合



局限

- 设计、使用、部署和维护具有难度，专业性更强，须对数据敏感，技能要求也高
- 依然存在边界，网络、io等瓶颈因素，决定其无法无限制的扩展
- 并非所有的计算问题都能转化为可并行计算的
- 所以，数据仓库与hadoop等数据分析系统并非竞争，而是各司其职，相互补足的关系

为什么选择Greenplum

- 中小企业的计算规模，已经在接近甚至超出传统数据库的能力，要求具备更强大的实时计算能力的数据仓库系统，是未来信息化的共同的呼声
- 其他的数据仓库为何不是可选项？
- 数据的核心地位，存储与计算的一体化，论开放与自主的重要性

2016 The Computing Conference

THANKS

2016/10/15

J.W.