QDBase
高性能离线跑批
方案与案例

跑批困境

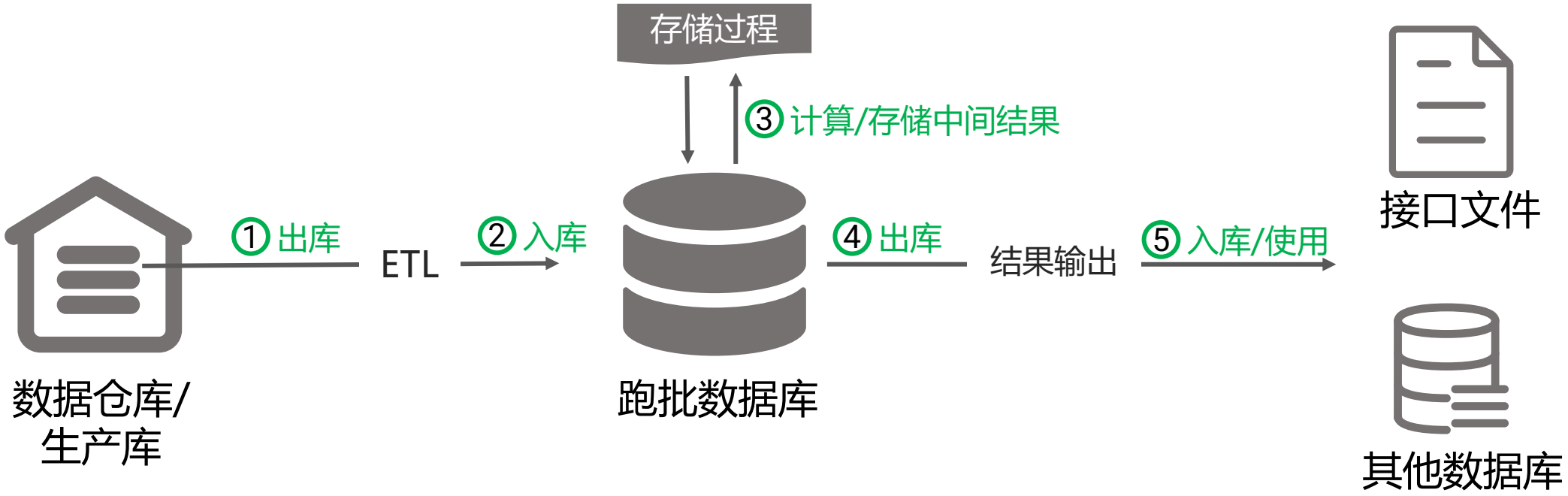
随着数据量不断增大，跑批任务越来越多，跑批压力随之增大

跑批太慢，跑不完影响业务

月底年底尤其提心吊胆

依赖存储过程，想扩容却无能为力

传统跑批过程



数据仓库/生产库

Oralce, Teradata、Hadoop等

跑批数据库

Oracle、DB2、Mysql等

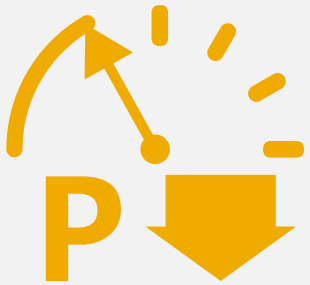
接口文件/其他数据库

文本文件：其他应用
BI数据库、其他应用数据库

跑批问题主要原因分析

关系数据库出入库太慢

数据库的存储和计算能力封闭，数据只有入库才能计算
数据进出要做的检查和处理过多，大量数据导入导出十分耗时



存储过程性能差

SQL语法限制，很多高效算法不支持
复杂多步计算涉及中间结果落地，增加IO开销
数据库游标性能差，且不支持并行无法提速

跑批重度依赖关系数据库！

因为只有数据库具备足够的计算能力，跑得慢也不得不用

如何破局？

想要解决数据库跑批面临的这些问题需要从以下两方面着手

开放性

需要足够的开放性可以计算库外文件数据，避免出入库开销

还可以直接对接多种数据源混合计算

强计算

拥有足够丰富的计算函数，可以处理跑批过程中的任意复杂计算，且实现相对简单

需要高计算性能保障跑批效率

分布式数据库跑批?

能否使用分布式数据库增加节点来提升跑批效率呢?

答案是，经常**并不能**!

存储过程支持弱

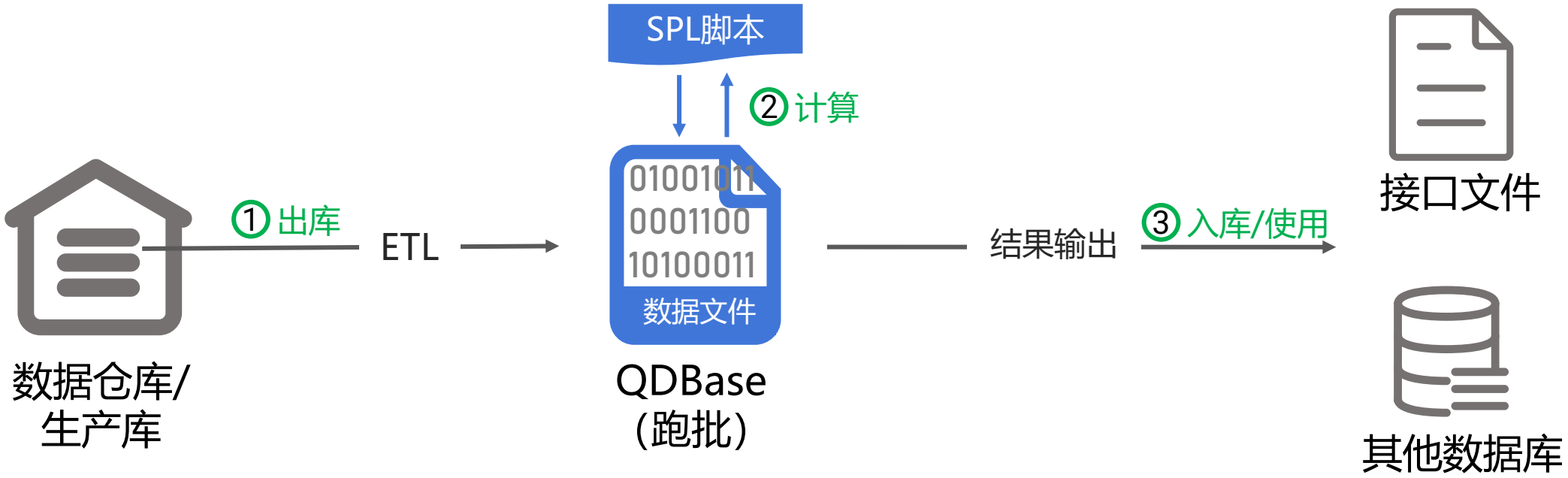
跑批逻辑复杂，往往需要成千上万行的存储过程实现，而分布式数据库对存储过程的支持程度很弱，无法胜任

中间结果使用问题

中间临时结果落地被不同节点使用会引发大量跨网络读写导致性能不可控，且无法通过数据冗余解决

因此，目前大多数跑批任务仍然使用单体数据库完成

QDBase跑批方案



数据仓库/生产库

Oralce, Teradata、Hadoop等

QDBase

直接使用文件存储并计算

接口文件/其他数据库

文本文件：其他应用
BI数据库、其他应用数据库

QDBase开放性-文件计算

文件存储的优势

存取效率更高

相对数据库的约束和接口限制导致的低效，文件具备更高的IO性能

使用更加灵活

文件更利于拆分存储，也更容易实施分段并行计算

管理更加方便

文件系统的树状目录更方便数据管理

QDBase文件计算能力

节约时间

QDBase可以直接基于数据文件完成跑批计算，节省了数据出入库时间

高效存储

提供了私有数据格式使用更高效，进一步缩短跑批时间

QDBase开放性-多源计算

QDBase提供多种数据源接口；支持多源混合计算
跑批涉及多源数据时可以直接使用，无需入库效率更高

混合计算

QDBase强计算

丰富运算类库

QDBase具备完善的计算能力，封装了大量计算类库可以直接使用

分组、循环、排序、过滤、集合计算、有序集合、并行计算、...

计算库

高效敏捷语法

QDBase提供了专有SPL敏捷语法，相对SQL/存储过程实现跑批计算逻辑更加简洁，代码更短

实际应用中代码量可以减少数倍

SPL

QDBase高性能

高性能存储

高效数据存储格式

集文件和组表两种私有格式，提供压缩、列存、索引等机制

有序存储

有序存储提高压缩率和定位性能

倍增分段

倍增分段方式支持任意数量并行

……

高性能算法

在高性能存储的基础上（高性能计算离不开存储），提供诸多高效算法提升计算性能

遍历复用

仅对大表遍历一次，就可以同时完成多种计算，降低IO成本

延迟游标

在一个游标上定义多个计算步骤，有效减少中间结果落地

多路游标

实施并行计算，充分利用多CPU计算能力提升计算性能

……

保险业历史保单关联业务跑批性能优化

历史保单关联业务跑批计算任务

每天新增保单2万条，一个月60万

任务：查找新增保单对应的历史保单

车险三年历史保单：2亿条数据

历史保单关联业务跑批的困难

跑批计算规则复杂

判断是否同一辆车，有三种方式；
还要判断是否贷款车、交强险

数据量大

2亿条中找几十万、百万

跑批时间过长

30天新增保单跑2小时
90天新增保单跑不出结果

跑批计算规则复杂


强制保险标志
2017
交强险
贷款车
90天
终保日期
1800行
存储过程
同一辆车：
车架号相同
同一辆车：
牌照号、种类相同
同一辆车：
VIN码相同
不可能完成
的任务
计算规则调整？要重算一年的新增保单！！

实测：QDBase表现优秀

30天新增保单提速6.5倍

500格

代码量减至30%

注：QDBase没有在历史保单上建索引，要用遍历计算；数据库建有索引，在短期保单计算时会略有优势。业主方仅关注中长期保单计算，因此没有再对此场景做进一步优化

解析：QDBase为什么快

数据有序存放

按保险单号有序存放数据
保单表和明细表有序分段关联
中间结果有序，无需重建索引

采用压缩列存

保险单表70字段，十几个参与计算
保险单明细表56字段，不到十个参与计算
文件压缩10倍左右，有效减少磁盘读取量

解析：QDBase支持更快的算法

数据库存储过程

1、新增保单关联保单表、明细表过滤
2、结果1关联明细表（VIN码方式）
3、结果2过滤，关联保单表
…
15、结果1关联明细表（车架号方式）
16、结果15过滤，关联保单表
…
24、结果1关联明细表（车牌号方式）
25、结果24过滤，关联保单表
…

多次大表关联计算

QDBase SPL

1、有序分段关联
保单表、明细表
2、循环遍历分段结果1
3、新增保单3种方式
关联分段结果1
4、结束循环2
…

仅一次分段关联计算

高性能计算是个手艺活

关系数据库(SQL)无法实现许多优化算法和存储方案，QDBase(SPL)则无此障碍