

基于Flink的异构海量数据涌传输系统

技术创新 变革未来

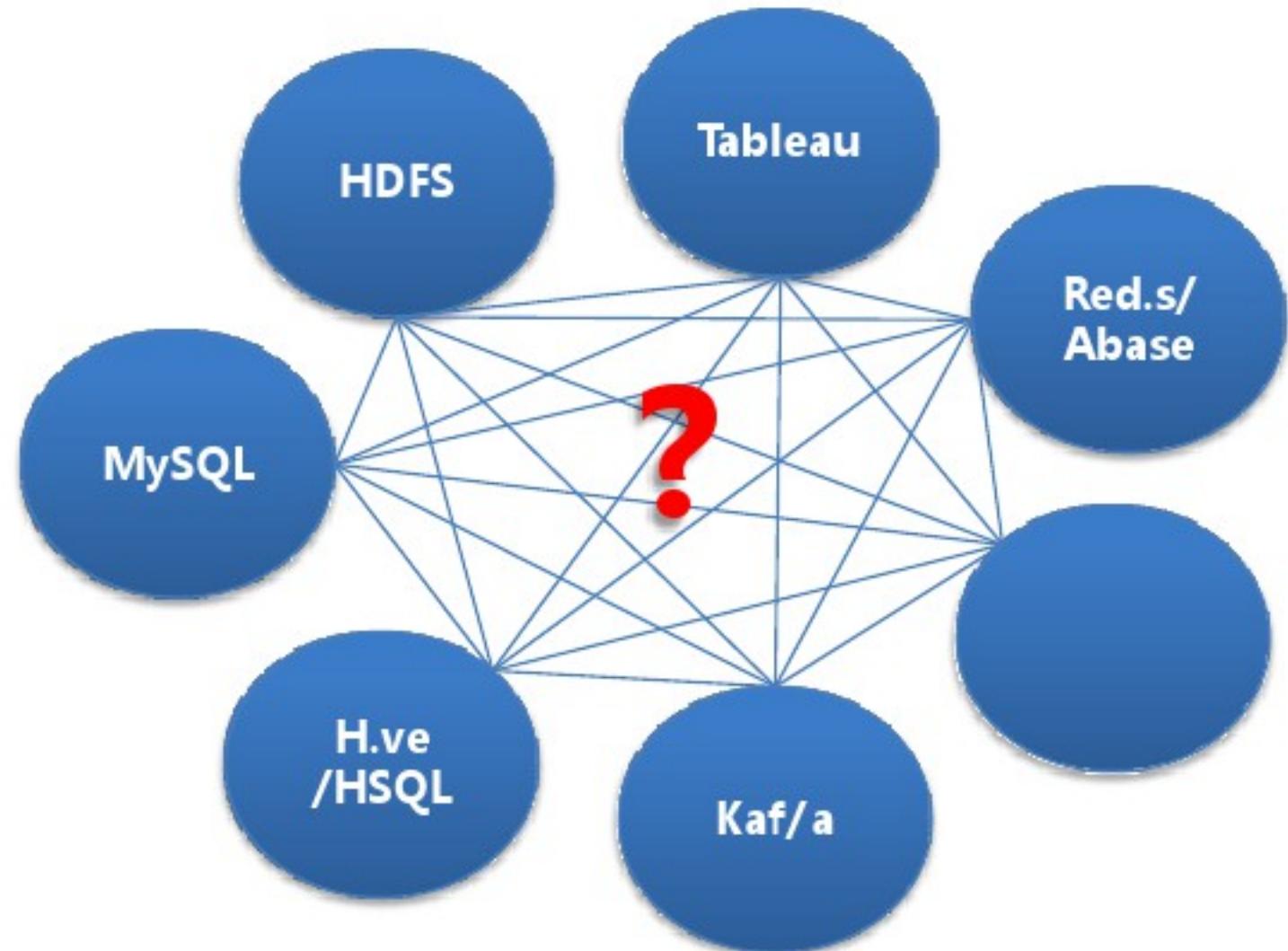
Agenda



挑战

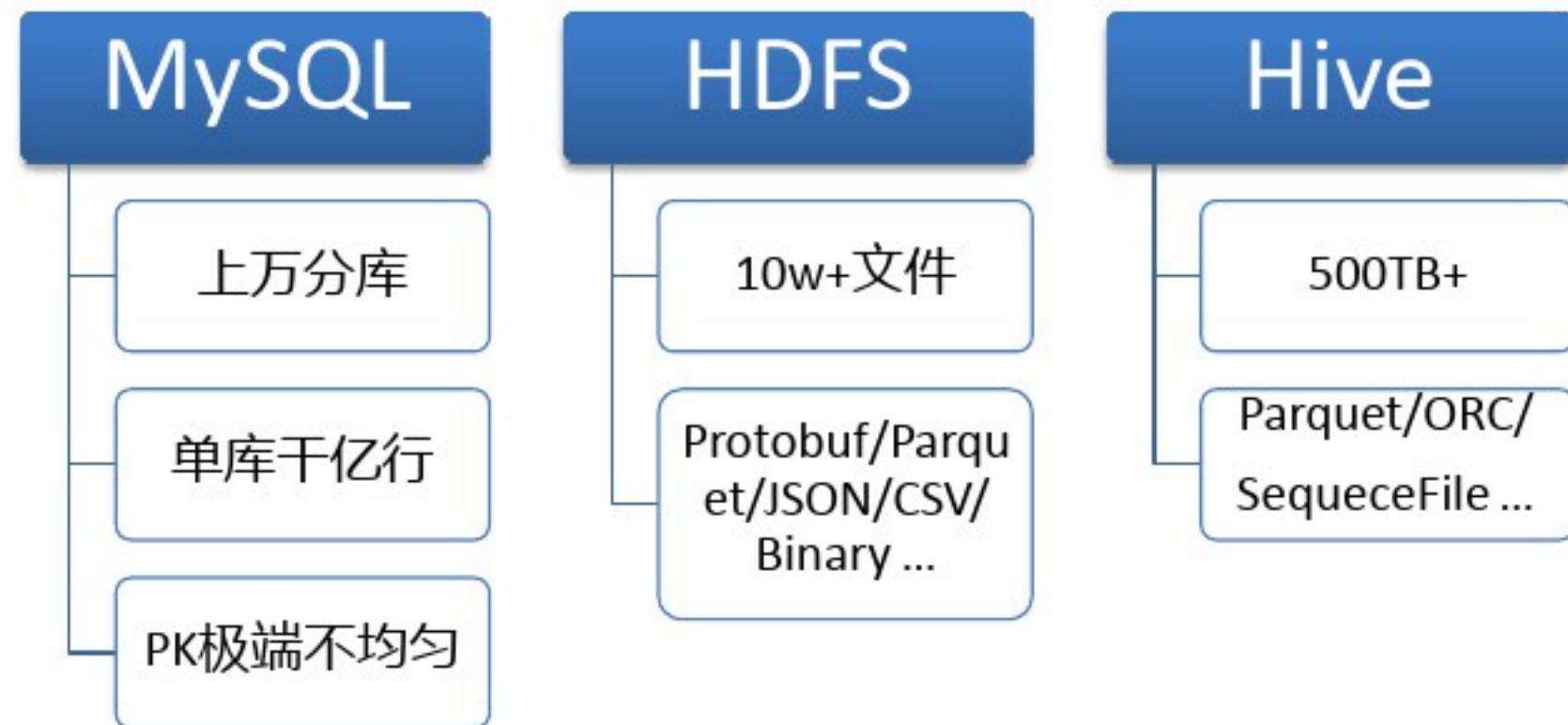
- 数据源种类多
- 传输方式多
 - 全量 / 增量 / 流式
- 频繁有新数据源出现

对架构的可扩展性要求高



挑战

对于单/务，数据L大，实现细节繁多。SLA(性能 / 稳定性等) 要A高。
。



传统架构



- 技术栈零散,M * N模式
- 缺乏统一的数据管理和功能支持框架
- 开发和运维成本高,性能和稳定性保证难度大

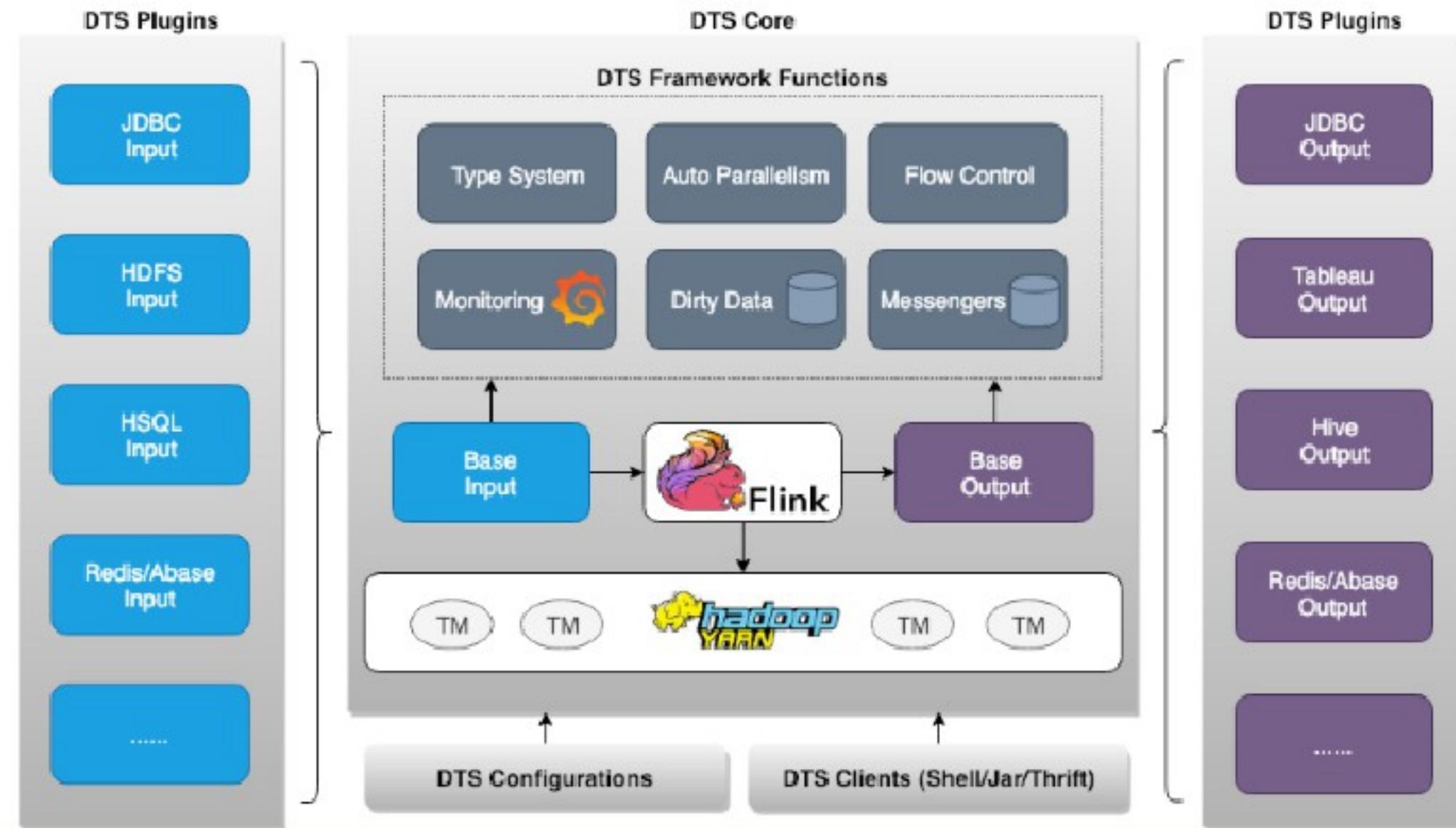
现有开源方案

- DataX (开源版本)
 - 类型系统完善，数据源类型支持较完整
 - 单机模式，横向扩展性不足
 - 不支持流式传输
- Sqoop
 - 关系型DB到Hadoop导入较为成熟
 - 新增数据源较为困难
 - 基于MR，性能非最优
- 缺乏性能和架构可扩展性强，支持流批统一传输的框架。

Agenda

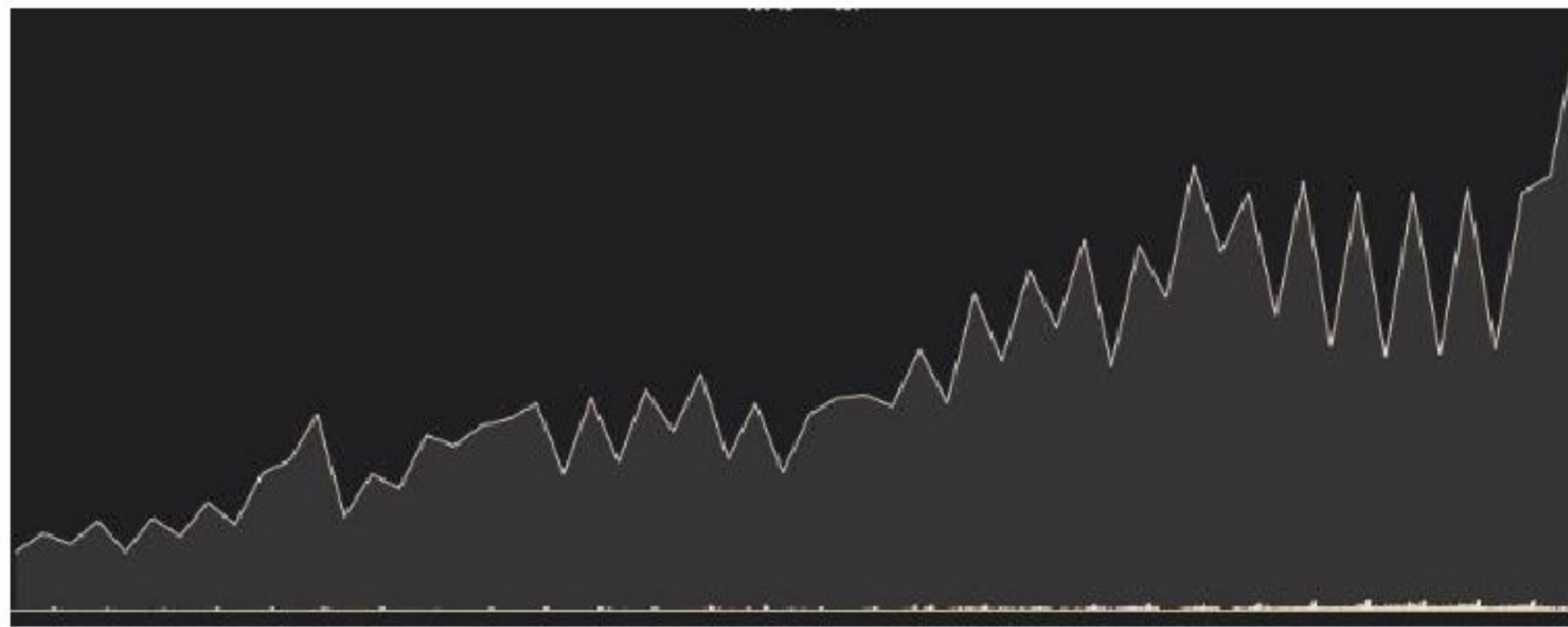


基于Flink的统一传输架构

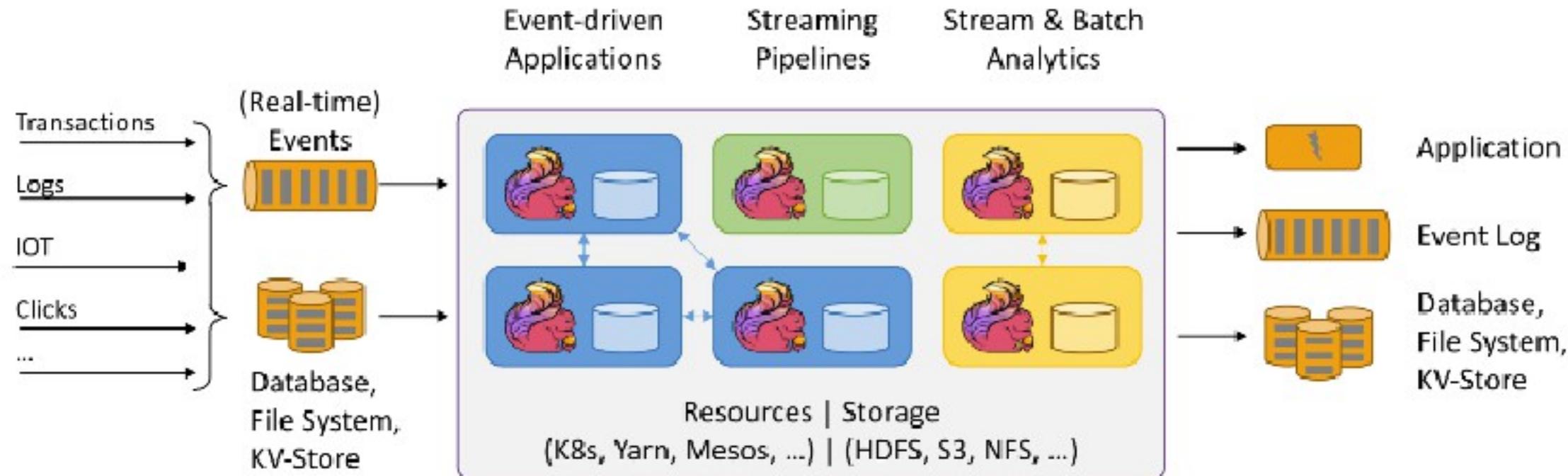


线上业务支撑

日均任务数 2w , 日均传输数据万亿级别



Flink Overview



- **Stream & Batch**
- **Operation Focus**
- **Exactly Once**
- **Scalability**
- **Layered API**
- **High Performance**

Plugin Framework

- 技术栈统一
- 所有数据对应独立的 Input/Output插件,M+N模式
- 框架层面提供统一基础功能

分布式计算

Flink

类型系统

自动并行度

流量控制

脏数据处理

小流量N试

Type System

- 所有输入源数据类型先转换为DTS类型，再统一转换为输出源数据类型。
- 对于每种类型，分别有对应的序列化/反序列化器，用于分布式数据传输。

基础类型

- BoolColumn
- BytesColumn
- DateColumn
- DoubleColumn
- LongColumn
- StringColumn

复合类型

- ListColumn<T>
- MapColumn<K, V>

Auto Parallelism

并发度决定了Job运行需要的计算资源，速度和对数据源的压力。

- 并发度过大
 - 可能导致计算资源浪费，或对数据源压力过大（例如线上MySQL）
- 并发度过小
 - 可能传输速度过慢以至超时（例如HDFS）
- Flink启动时需要预先指定并发度
 - 对Flink框架加以改进

Auto Parallelism

对Flink架增加了Job预处理 / 后处理流程，根据以下指标计算输入 / 输出最佳并发度

输入原/片数

输入原总行数

输入原总大小

输入原可承压

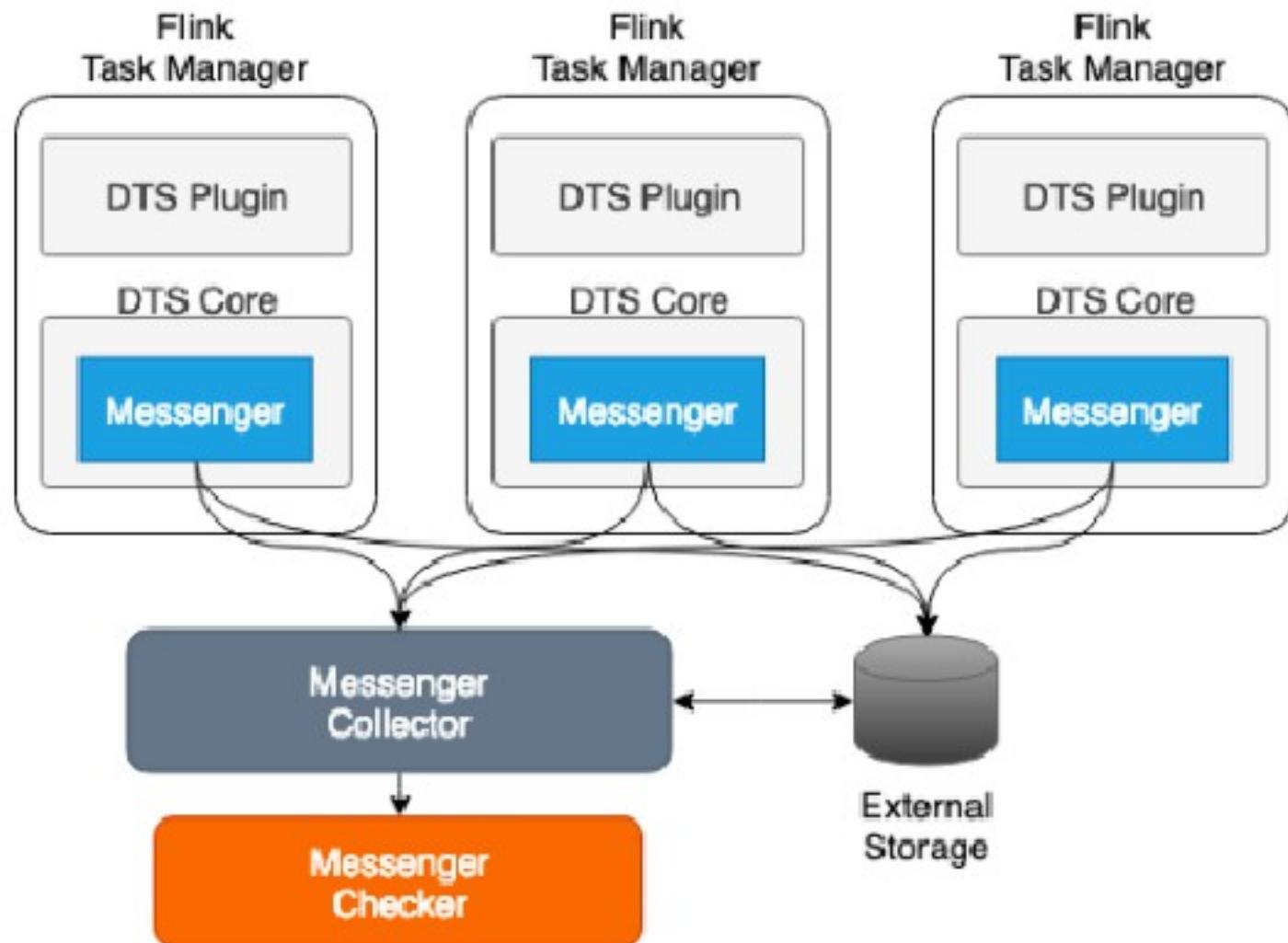
输出原可承压

用户配置并发

Messenger Framework

在框架层面解决 Flink 多节点间 数据收集的问题。

- 任务运行指标收集
- 脏数据收集
- 支持多种目标数据洒



Messenger - Metrics

收集Job运行过程中各类Metrics：

- Total/Success Splits Number
- Success Row Number
- Fail Row Number (with Exception)
- Success/Fail Bytes Number
- Success/Fail Byte Rate
- ...

支持写入各类目标源：

- Memory/Job Log
- Kafka
- Flink Metrics
- ByteDance Metrics
- ...

Messenger - Dirty Data

用于收集和处理脏数据，写入中间数据库/Kafka等

脏数据收集

- 少量任务运行日志预览
- 增量脏数据写入
- 全量Log日志查询等

任务质量控制

- 控制任务失败的脏数据阈值
- 支持绝对行数阈值
- 支持相对百分比阈值

实现优化

为了保证尽可能高效和稳定的传输，实现上也做了许多优化。

- 读取MySQL支持多种分片策略
 - 解决了表ID分布不均匀线上慢查询的问题
- 写入MySQL/Tableau支持动态分区
 - 类似Hive动态分区，目前基于分布式锁实现
- Flink内序列化时增加字节压缩
 - 有效解决AKKA传输大小限制和内存不足等问题
- ...

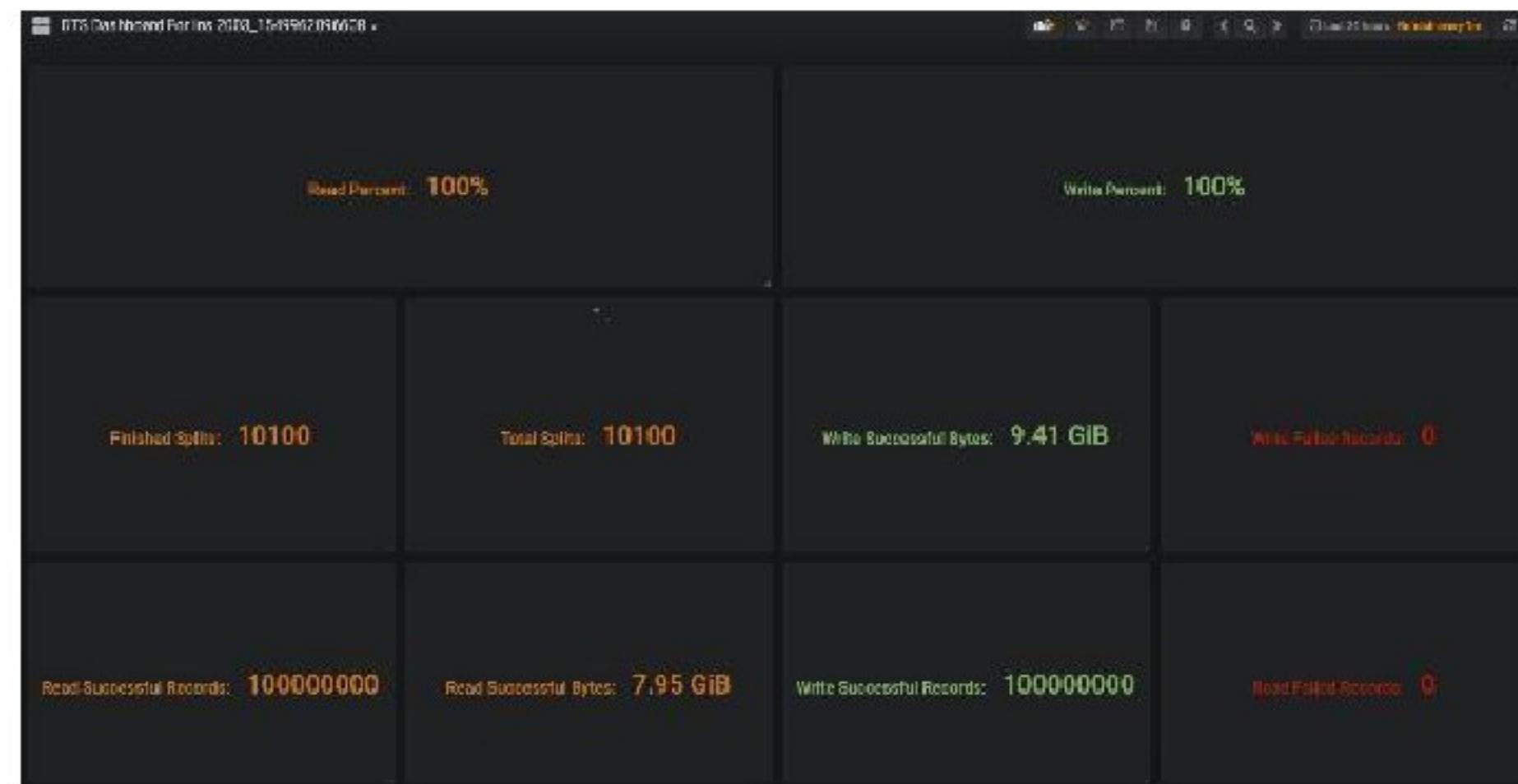
Flink相关改进

- [FLINK-10941] Slots prematurely released which still contain unconsumed data
 - 已修复并贡献回社区
- [Backport to 1.5][FLINK-9455][RM] Add support for multi task slot TaskExecutors
- [Backport to 1.5] [FLINK-10848] Flink's Yarn ResourceManager can allocate too many excess containers
- [WIP] Batch failover strategy
- [WIP] Flink on Yarn Job Leak Improvement

Monitor & Alarm

单Job监控看板

- 总读取 / 写入进/
- 总分片数
- 当前-成分片数
- 当前读取 / 写入行数
- 当前读取 / 写入大小
- 失败读取 / 写入行数

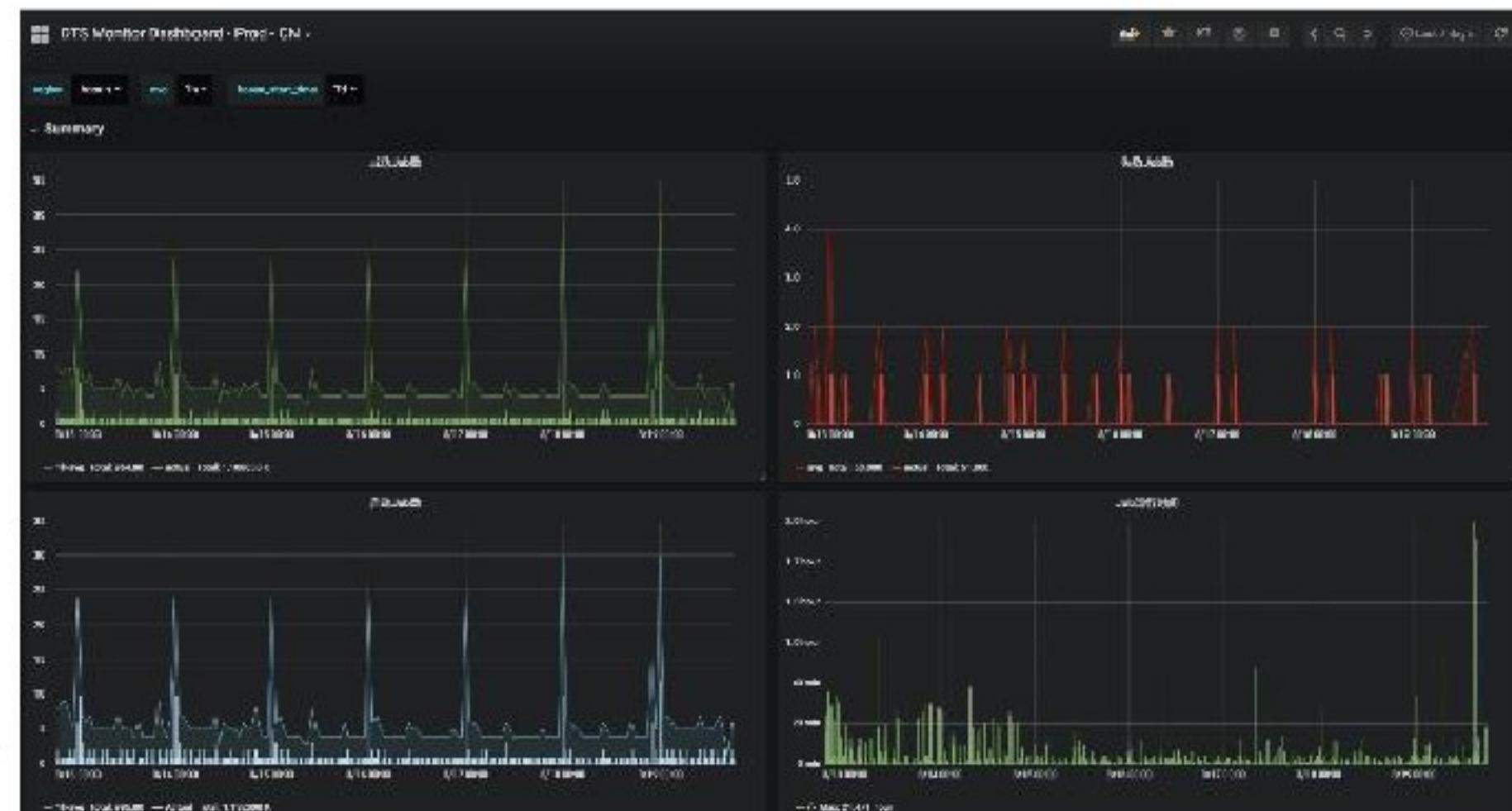


Monitor & Alarm

系统整-监控看板

- 已启动Job数量
- 正在运行Job数量
- 已成功Job数量
- 已失败Job数量
- Job运行时间

另有独立的Yarn Job轮询服务进行兜底，防止Job Leak



Agenda



未来规划

- 全量传输
 - 支持更多数据源，例如ES，ClickHouse等
- 增量传输
 - 实现MySQL基于bin log的自助式增量导入
- 流式导入
 - Kafka/RabbitMQ Exactly Once传输
- Flink框架改进
 - External Shuffle Service/Speculative Execution etc.
- 服务化
 - OpenAPI自助式接入



THANKS