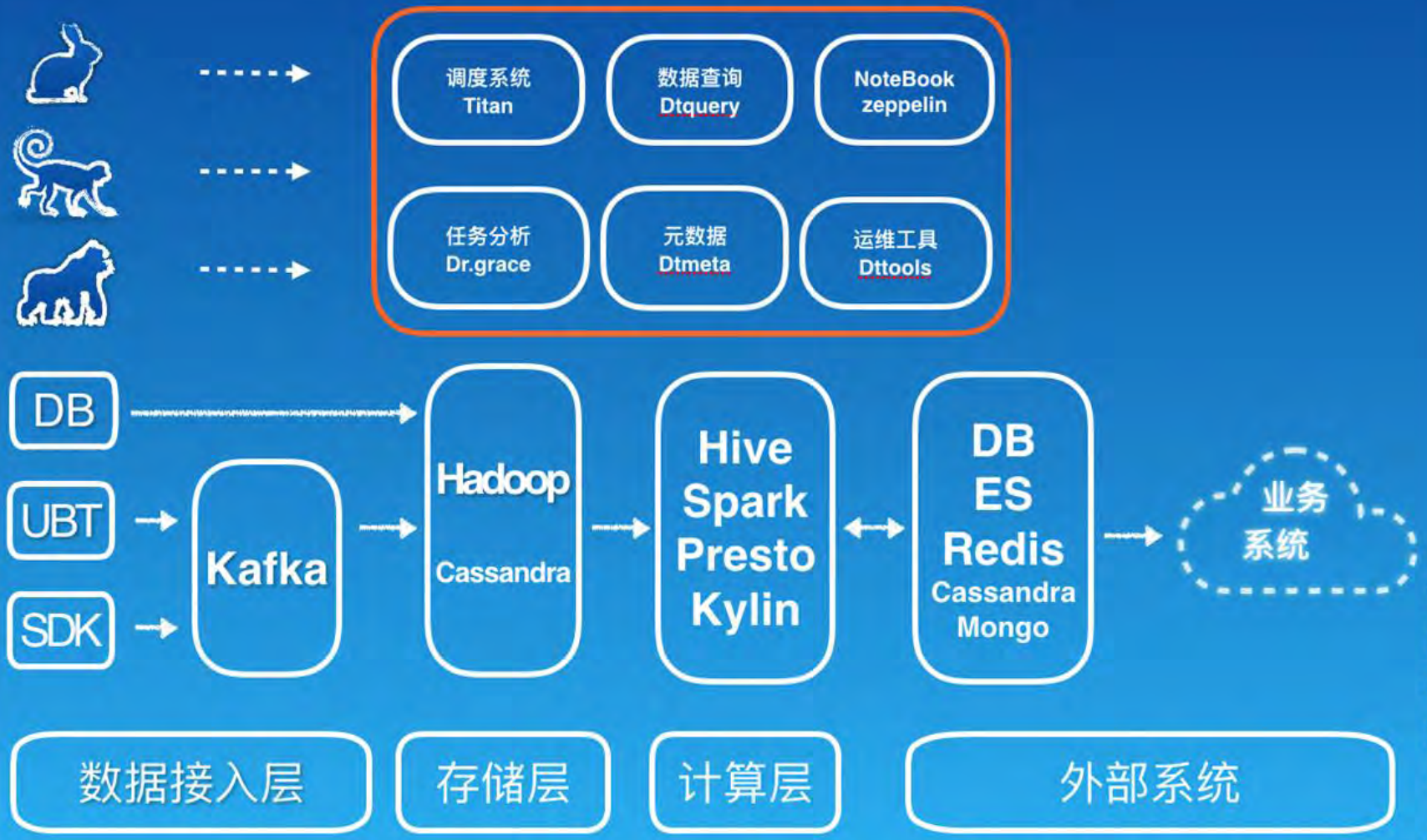
离线平台规模

- 数据增量80TB/天
- Hadoop集群规模1000-1500节点
- 调度任务2万+
- 作业数10W+
- 计算吞吐量3PB/天

离线平台架构


调度系统
Titan
数据查询
Dtquery
NoteBook
zeppelin
任务分析
Dr.grace
元数据
Dtmeta
运维工具
Dttools
DB
UBT
SDK
Kafka
Hadoop
Cassandra
Hive
Spark
Presto
Kylin
DB
ES
Redis
Cassandra
Mongo
业务
系统
数据接入层
存储层
计算层
外部系统

目录

- 平台概况
- 服务治理
- 提速增效
- 数据化运营

平台存储挑战

- Namenode
  - 数据量和文件数飞速增长，存储告急
  - 文件数增多导致频繁Young GC和RPC出现瓶颈
  - 如何实现升级不停服务
- 元数据治理
  - 建表没有规范
  - 只有数据没有表
  - 数据存放目录随意
  - 临时数据不清理

Namenode治理

- 合并小文件
  - Hive/Spark集成自动合并小文件
  - 多种文件快速合并工具
- 不停服务升级
  - Service RPC和Client RPC端口分离
  - 动态刷新datanode超时时间
- 锁优化（ DOING ）
  - 锁开销Metrics
  - 非公平锁策略：适合读多写少场景
  - 锁拆分

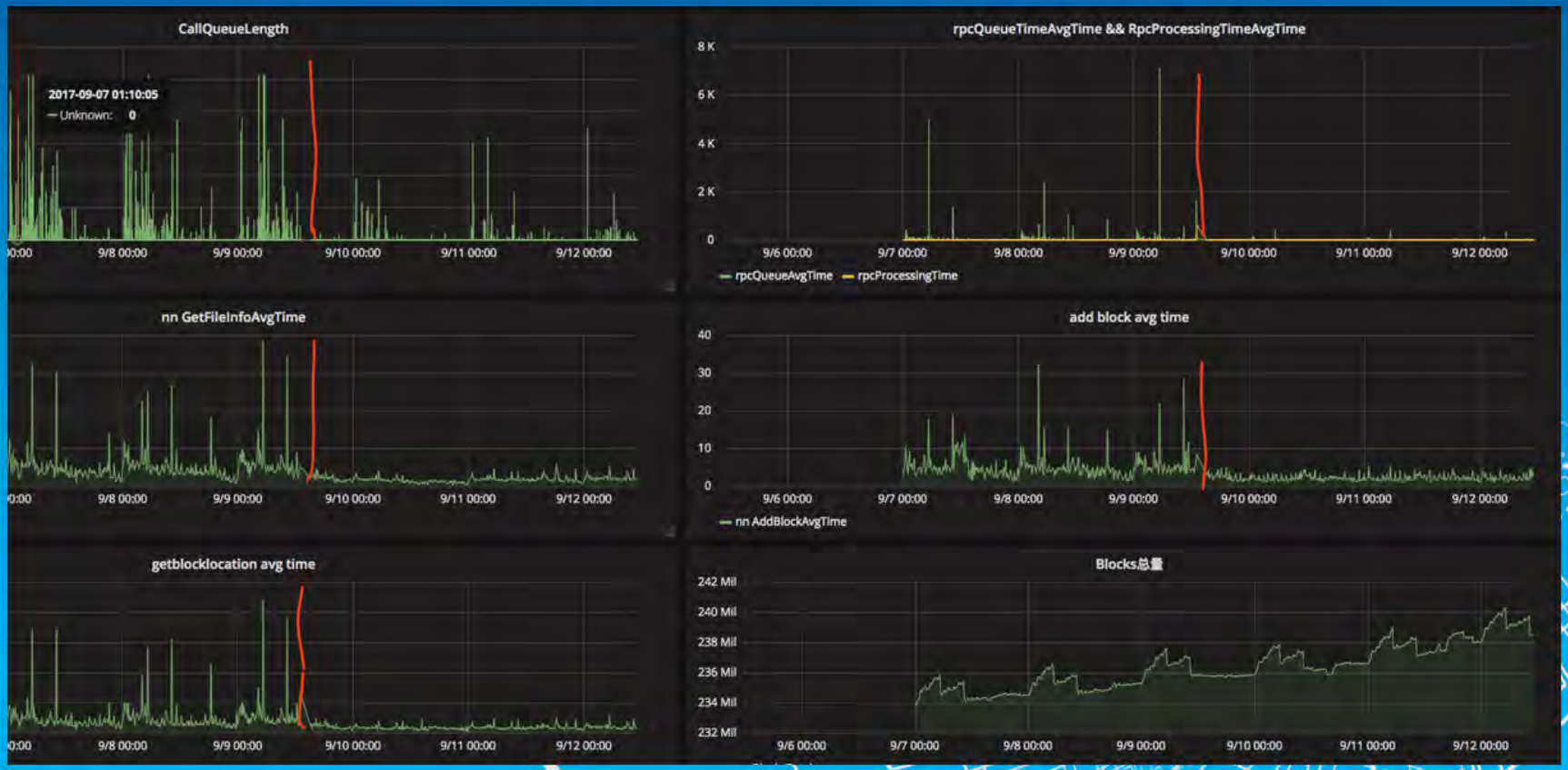
Spark多种文件快速合并：
https://github.com/eleme-datainfra/spark/tree/ESPARK-116

Namenode JVM调优

- CMS调优关键点
  - Young GC的频率和时间
  - CMS GC的Initial mark和Remark时间
  - 防止Concurrent mode failure和Promotion failure
- 对应措施
  - -Xmn30g -XX:MaxTenuringThreshold=3
  - 升级JDK1.8 -XX:+CMSScavengeBeforeRemark
  - -Xmx180g -XX:CMSInitiatingOccupancyFraction=85

Namenode治理效果

高峰期间YoungGC频率降低到17s一次
吞吐率提升到 99.59%



存储容量治理

- 数据生命周期
- 数据热度反向推动清理
- ORC/Parquet文件格式推广

元数据治理

- 建表收口，只能在元数据系统建表
- 临时表只能建在temp库下
- 统一数据表存放位置
- 所有数据都要有表

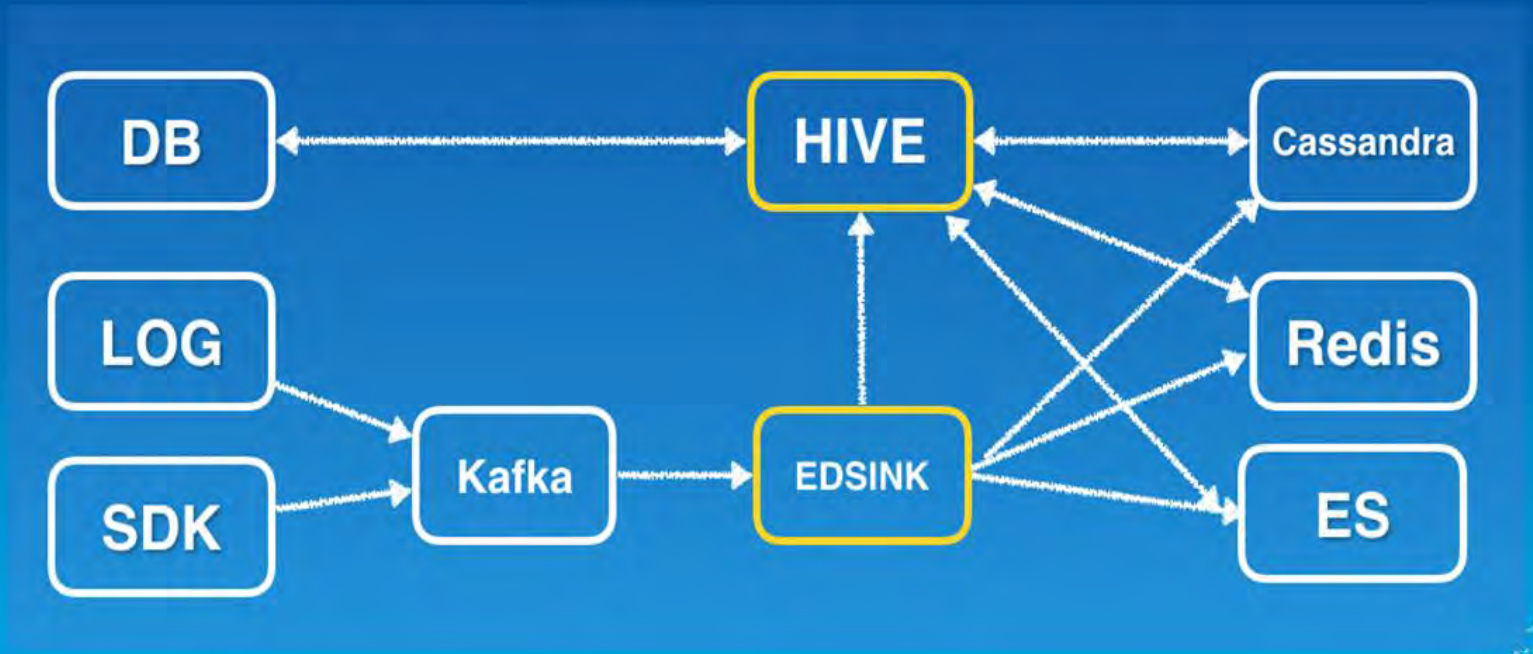
目录

- 平台概况
- 服务治理
- 提速增效
- 数据化运营

提效增速

- 日益增长的计算需求与落后的生产计算工具之间的矛盾
- 数据流通效率
- 计算效率
- 开发效率

数据高速公路



Spark推广

- Hive ETL切换SparkSQL
- Adhoc接入Spark ThriftServer
- Carbondata
- Streaming SQL
- Zeppelin + Livy

SparkSQL切换

- HiveSQL成功率50% -> 92.5%
- 测试平均加速6倍
- MapJoin深入优化
  - 支持分区表
  - 根据表原始大小来选择
- 自动化数据质量校验

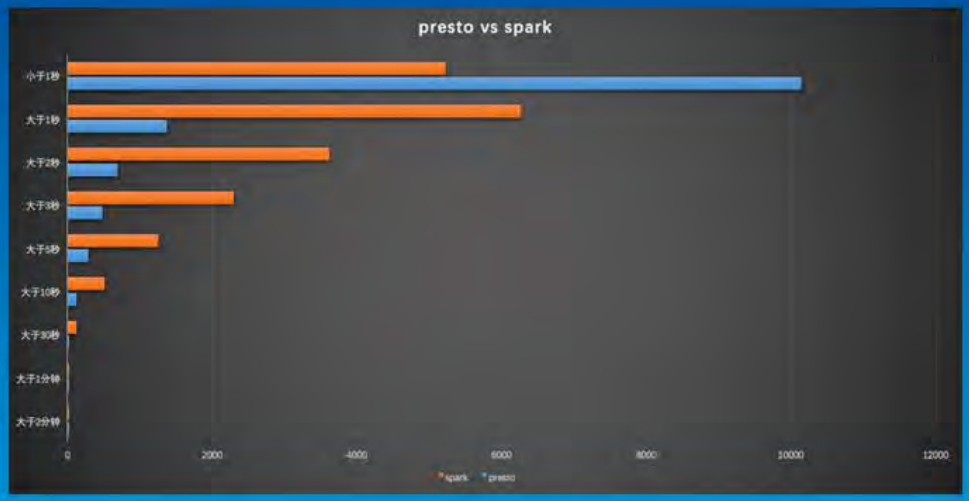
更多详细信息：http://dwz.cn/elemespark

Spark ThriftServer增强

- 动态Executor模式无法正常回收Executor
- Task级别多用户数据权限功能
- FileSystem Cache bug引发OOM
- SQL作业无法取消
- SQL作业进度打印

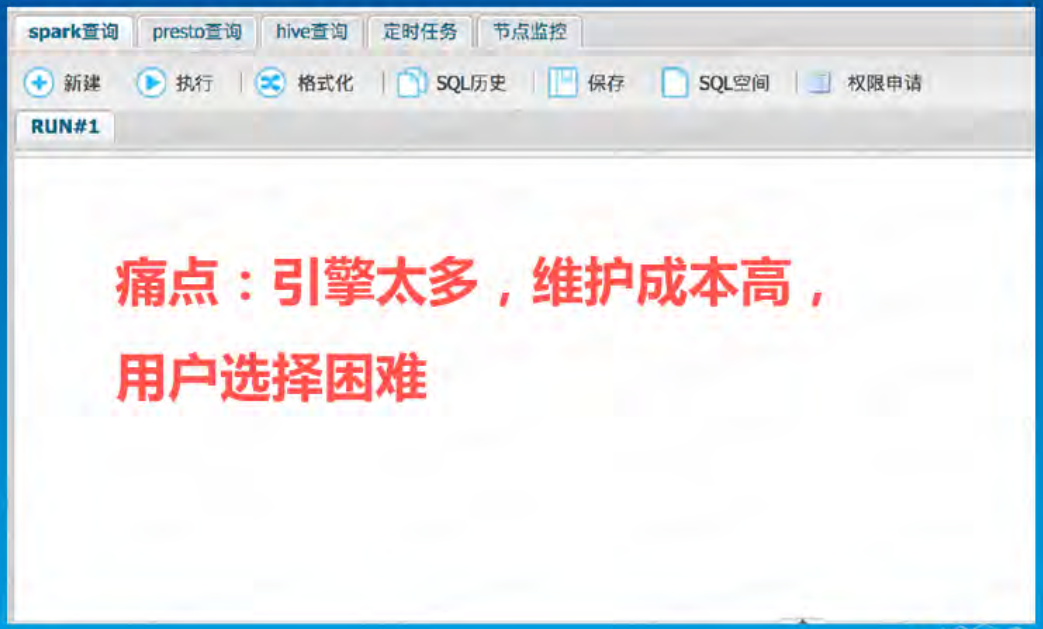
Presto

- 优势
  - 内存MPP引擎
  - Pipeline DAG执行计划
  - 调度延迟低
- 劣势
  - 大查询容易失败
  - 方言与HiveSQL有差别
- 应用业务：餐厅报表
- 每日查询量10W+

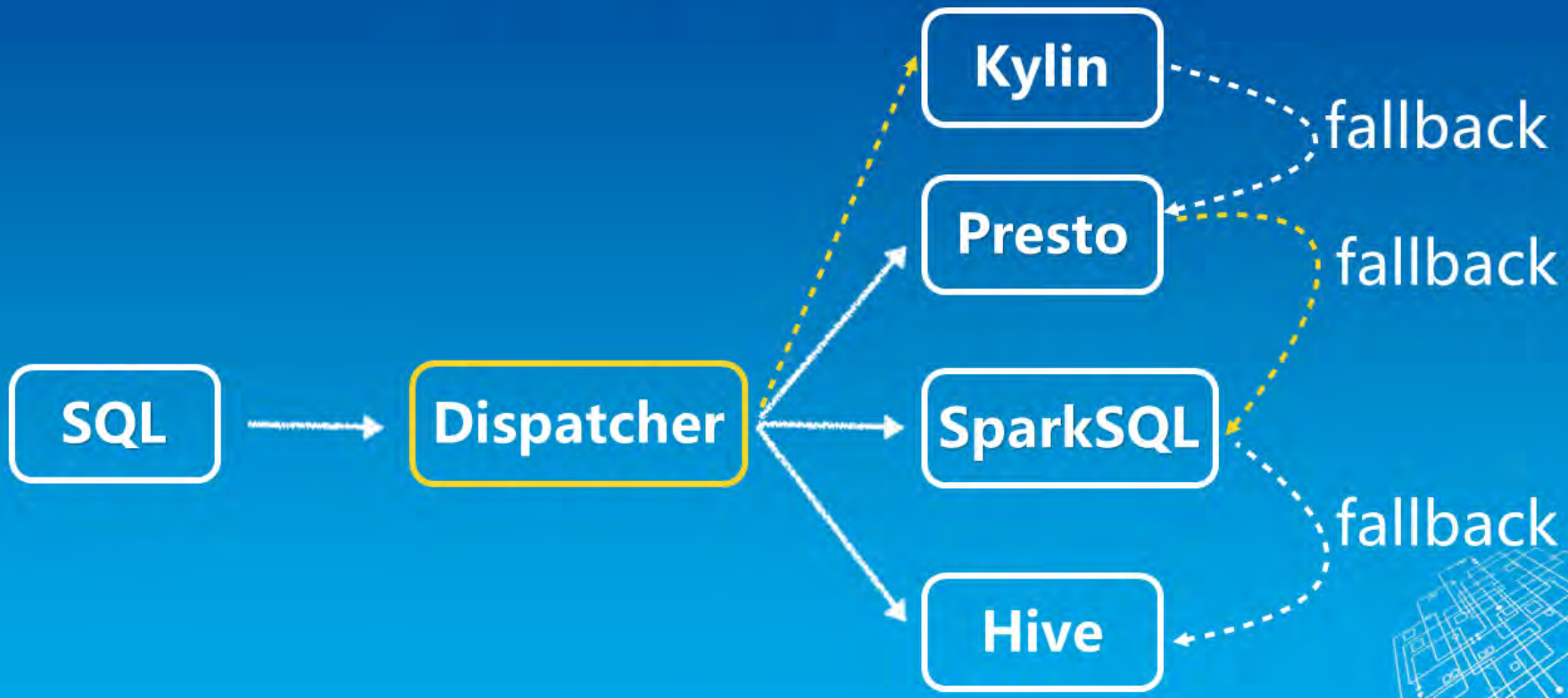


多种引擎的烦恼

- Adhoc Query查询量
  - SparkSQL 4500+
  - Hive 1700+
  - Presto 1000+
- 使用场景
  - Presto：中小型查询
  - SparkSQL：大中型查询
  - Hive：兜底
- 还有刚上线的Kylin...



统一路由引擎

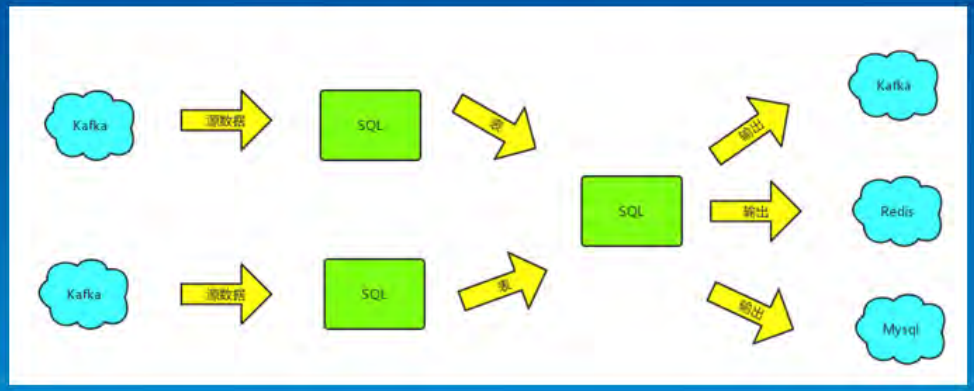

Kylin
fallback
Presto
fallback
SQL
Dispatcher
SparkSQL
fallback
Hive

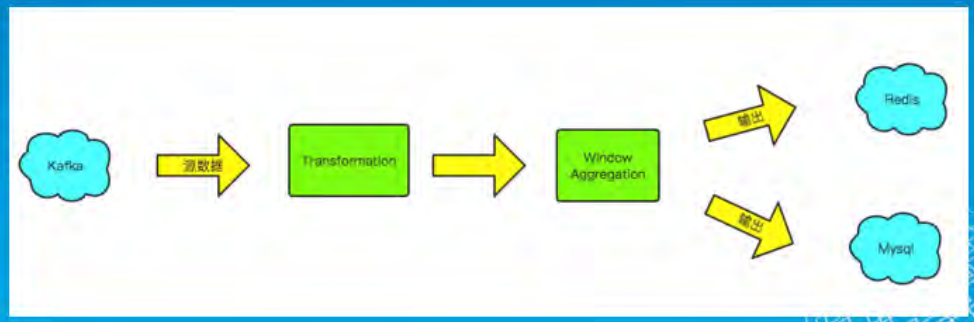
Carbondata

- 读取数据量是影响查询速度关键
- Hive integration(doing)
- 测试数据：3列索引(总共126列 42亿行数据)
- 数据导入时间是parquet的3倍
- 存储大小是parquet的1.5倍
- 命中索引的批量查询比parquet快一倍以上
- 详单查询2秒

Streaming SQL

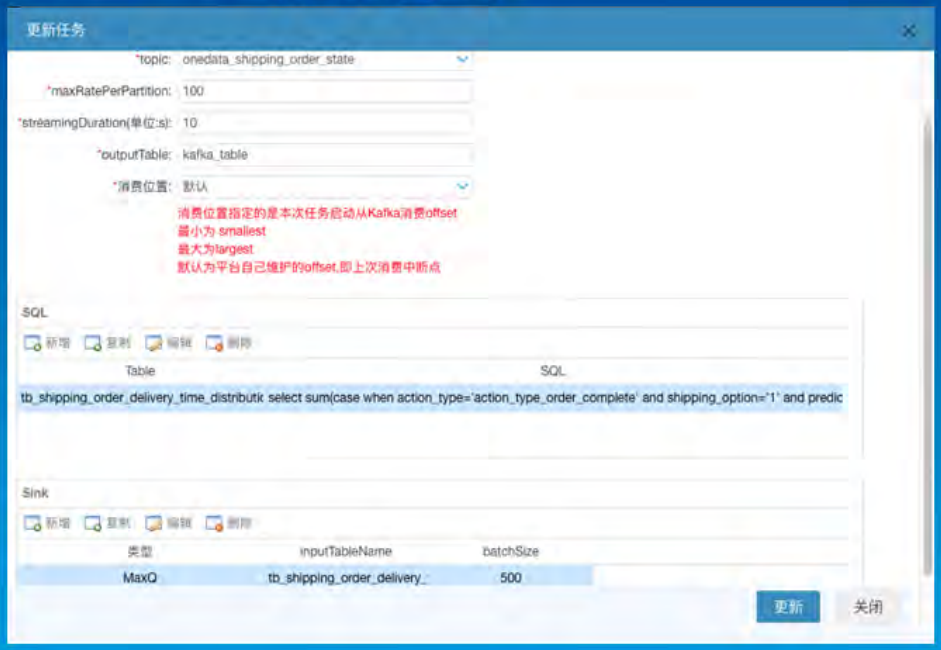
- 支持两种场景
  - 无状态ETL场景
  - Structured Streaming

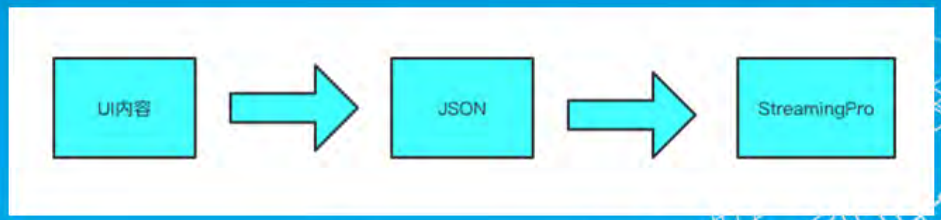




无状态ETL场景

- Topic管理
- Metrics接入influxdb
- 异常Task日志接入ES
- 自动注册UDF
- 支持多种SINK





Structured Streaming SQL

```
SELECT action, WINDOW(time, "10 minutes"), COUNT(*)
FROM events
GROUP BY action, WINDOW(time, "10 minutes")
```

```
val windowedCounts = actions
  .withWatermark("time", "10 minutes")
  .groupBy(
    $"action"
    window($" time", "10 minutes", "5 minutes"),
  )
  .count()
```

目录

- 平台概况
- 服务治理
- 提速增效
- 数据化运营