2023 DataFunCon

OPPO大数据诊断平台设计与实践

演讲人：戴巍 - OPPO - 数据平台架构师

Contents
目录

1 背景

2 技术方案

3 实践效果

4 总结与规划

01 背景

背景 / OPPO大数据现状

数据量：1EB+
系统组件：20+

离线任务：百万
实时任务：数千

数据分析师/开发师
1000+

数据开发人员水平参差不齐，问题排查难

任务链路长，组件众多，运维复杂

僵尸任务和不合理任务治理难度大

背景 / 业界产品对比

开源产品 Dr. Elephant 分析

01 **LinkedIn**

由LinkedIn开发并开源。旨在提高开发人员效率和增加集群任务调试的高效性

02 **支持多种计算引擎**

支持多个计算引擎框架性能诊断：Spark、Tez、MapReduce、TonY等

03 **支持多种调度框架**

集成多个调度器框架如：Azkaban、Airflow、Oozie等

04 **分析报告**

统计历史作业和工作流的性能指标
Job级别工作流对比

01 **新版本兼容性不好**

支持Spark, Hadoop系统版本比较低，对于新版本Spark, Hadoop兼容性不好

02 **诊断指标少**

支持的Spark相关指标仅4个

03 **诊断手段少**

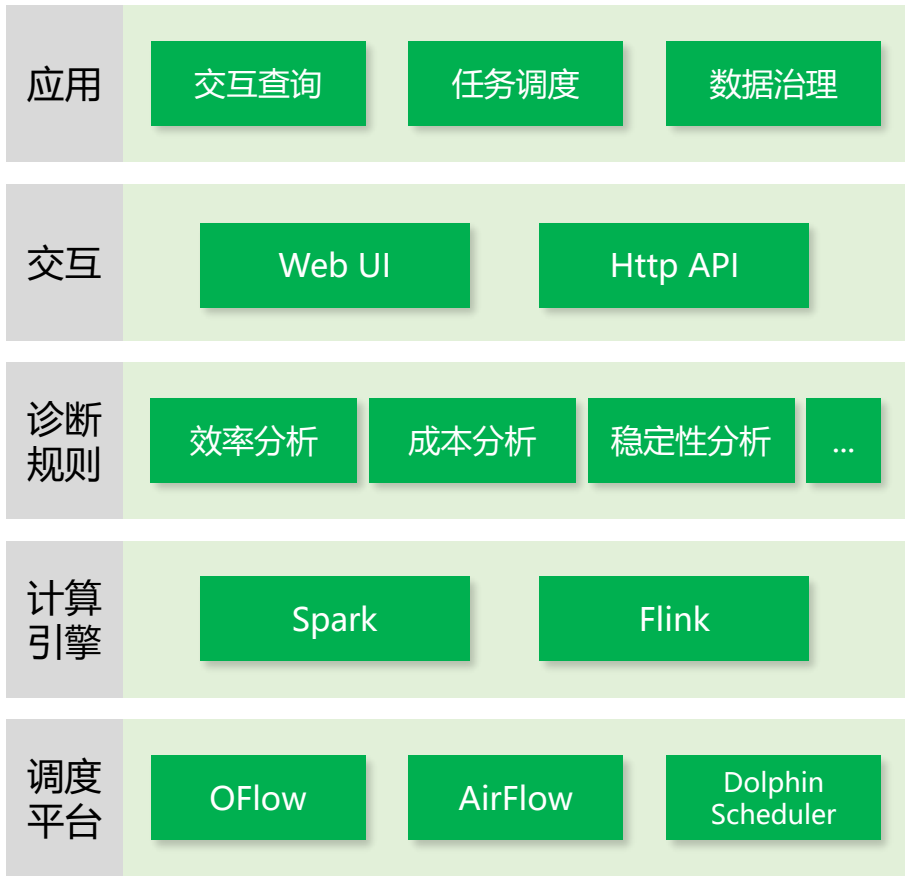
不支持日志级别问题诊断
不支持异常资源的管理

04 **稳定性风险**

对Spark History服务接口频繁调用影响History服务的稳定性

02 技术方案

技术方案 / 平台特性



1. 非侵入式，即时诊断，无需修改已有的调度平台，即可体验诊断效果
2. 支持 OPPO 自研调度平台及多种主流调度平台，如 DolphinScheduler、Airflow 等，进行工作流层异常诊断
3. 支持多版本 Flink、Spark、Hadoop 任务诊断
4. 支持 40+ 离线和实时场景异常类型判定，并在不断丰富
5. 支持自定义规则编写和异常阈值调整，可自行根据场景调整

技术方案 / 系统架构

外部系统适配层

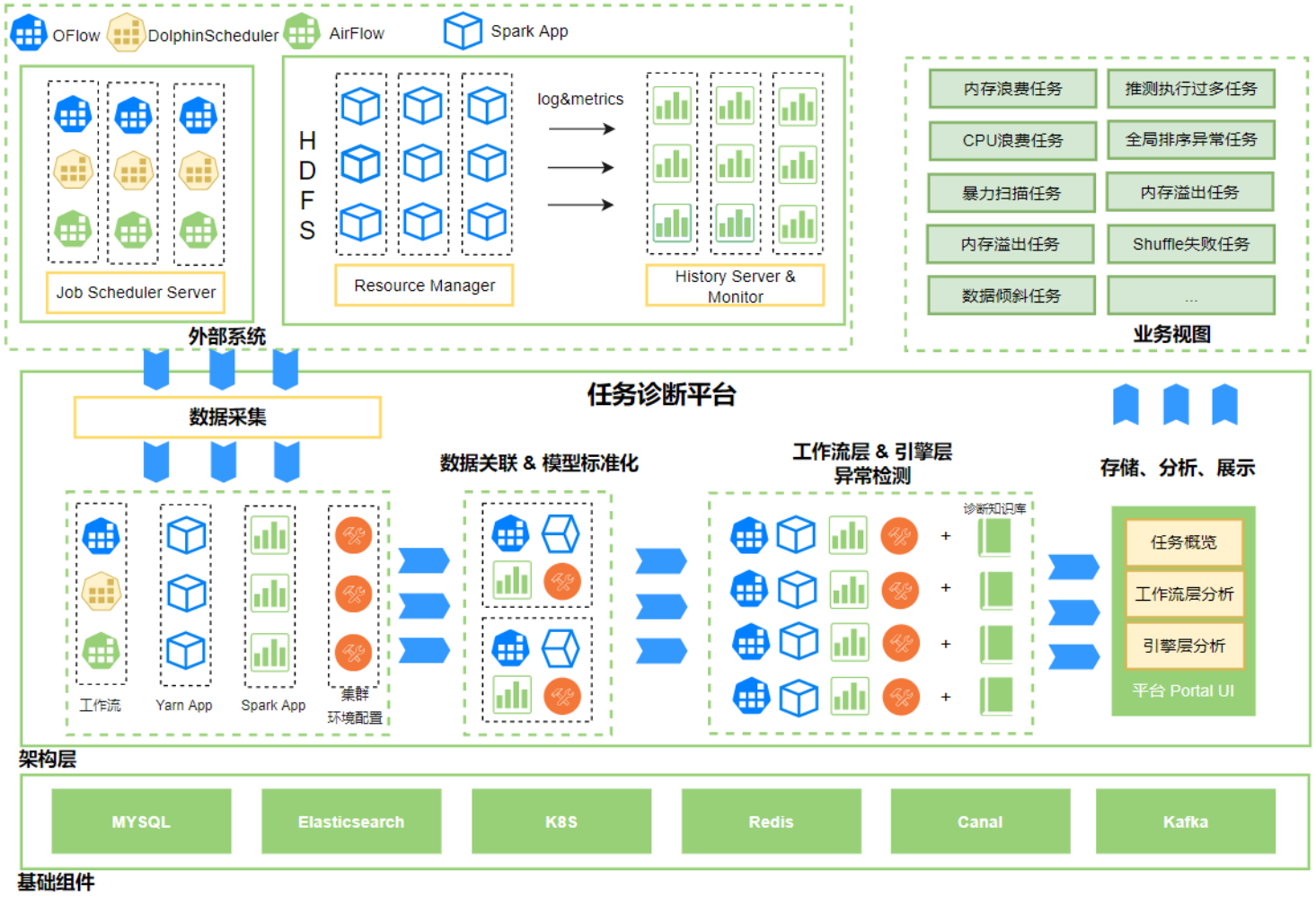
调度器、Yarn、HistoryServer、HDFS等系统，同步元数据、集群状态、运行环境状态、日志等到诊断系统分析

诊断架构层

包括数据采集、元数据关联&模型标准化、异常检测、诊断Portal模块

基础组件层

包括MySQL、 ElasticSearch、Kafka、Redis等组件



技术方案 / 流程阶段

1 数据采集阶段

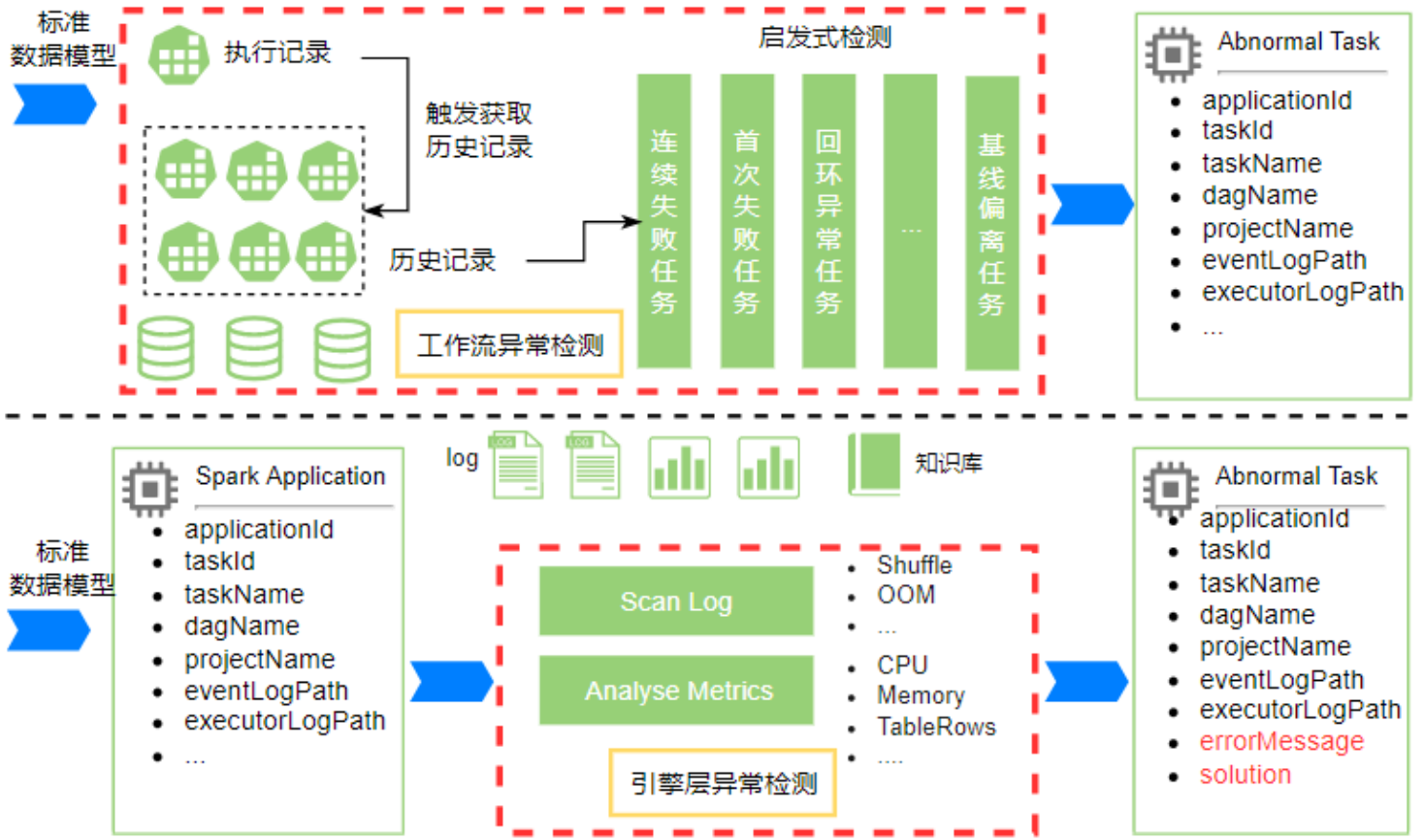
同步调度系统用户、DAG、执行记录等工作流元数据；同步Yarn ResourceManager、Spark HistoryServer App、Flink Job元数据等

2 数据关联&模型标准化阶段

将分步采集的工作流执行记录、Spark App、Yarn App、 Flink Job、集群运行环境配置等数据基于Workflow进行关联，形成标准数据模型

3 工作流层&引擎层异常检测阶段

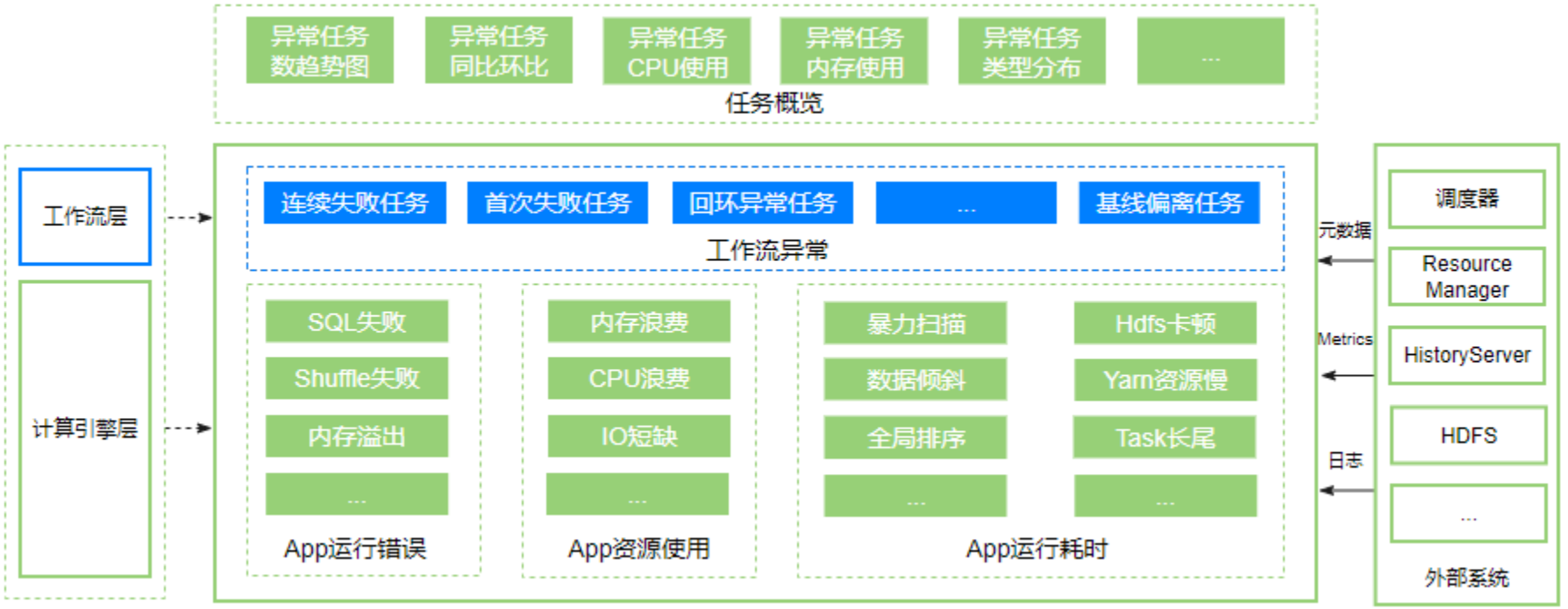
进行Workflow异常检测流程，加载知识库到标准模型，通过启发式规则，对指标数据、日志进行异常挖掘，结合集群状态及运行时状态，分析工作流层、引擎层异常结果



技术方案 / 流程阶段

4 业务视图

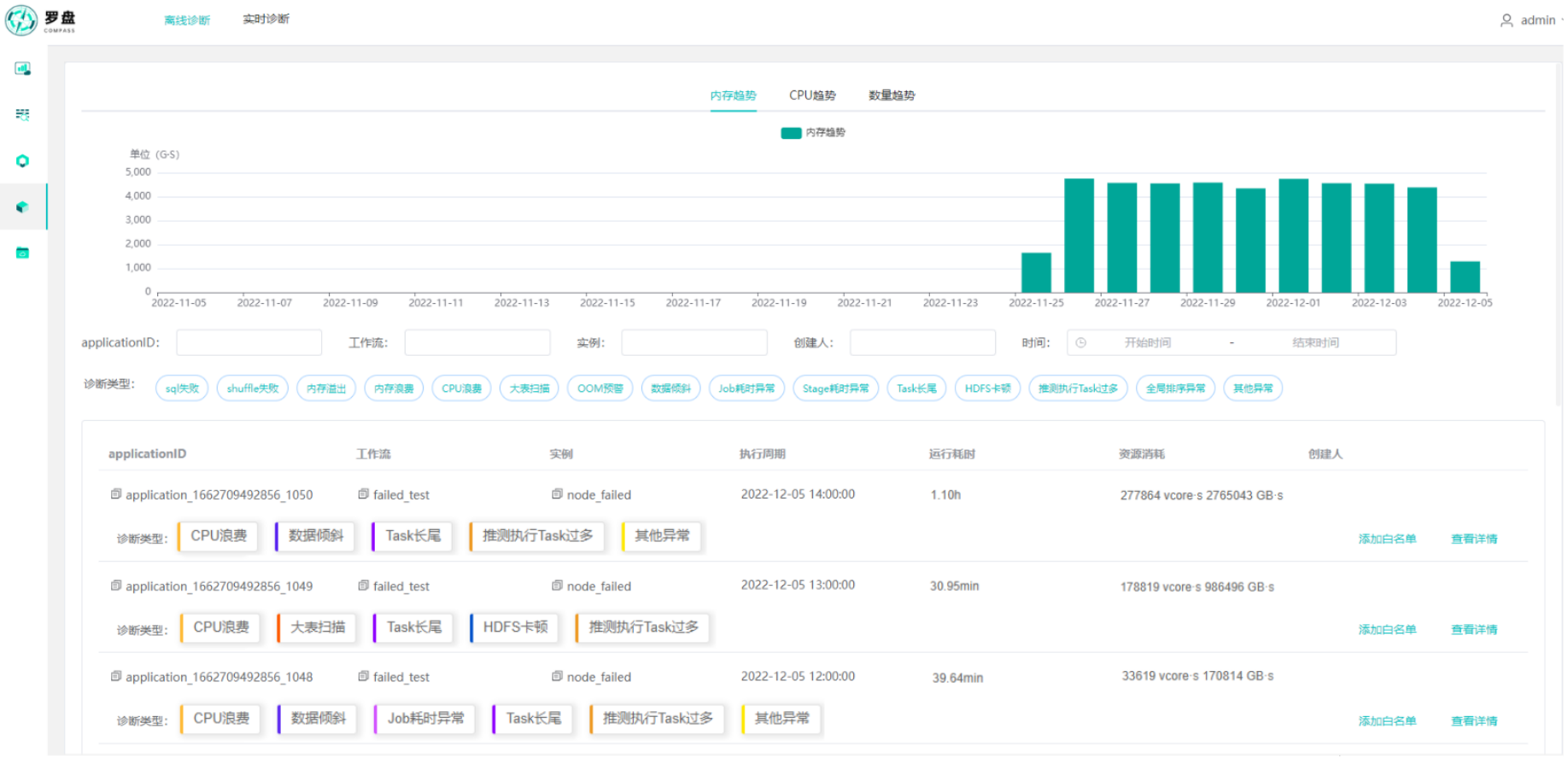
存储、分析数据，提供给用户任务概览、工作流层任务诊断、引擎层作业Application诊断，工作流层展示调度器执行任务引发的异常，如任务失败、回环任务、基线偏离任务等问题，计算引擎层展示Spark、Flink作业执行引发的耗时、资源使用、运行时问题



03 实践效果

实践效果 / 交互设计

用户可以一眼看到关心的任务问题所在，并能给出指导性处理建议



统一

简洁

直观

实践效果 / 诊断类型丰富

针对 **离线**、**实时** 任务的健康度诊断

支持 **40+** 场景异常类型判定

效率分析

长尾Task分析
HDFS卡顿分析
推测执行过多分析
全局排序异常分析
...

稳定性分析

全表扫描问题
数据倾斜分析
Shuffle失败分析
内存溢出
...

实时作业分析

作业TM空跑
作业并行度不足
反压算子诊断
慢算子诊断
...

成本分析

CPU浪费分析
内存浪费分析
长期失败分析
耗时分析
...

实践效果 / 效率分析案例

长尾Task分析

长尾任务是由于作业运行过程中，一个Task或多个Task单元执行时间过长，拖延整个任务运行时间

从时间角度计算，执行时间过长原因在于Task读取数据量多或者数据读取慢。如果读取数据过多，那么将出现数据倾斜，按数据倾斜方式处理；如果读取数据过慢，那么可能是HDFS集群节点负载高或网络丢包问题等

实践效果 / 成本分析案例

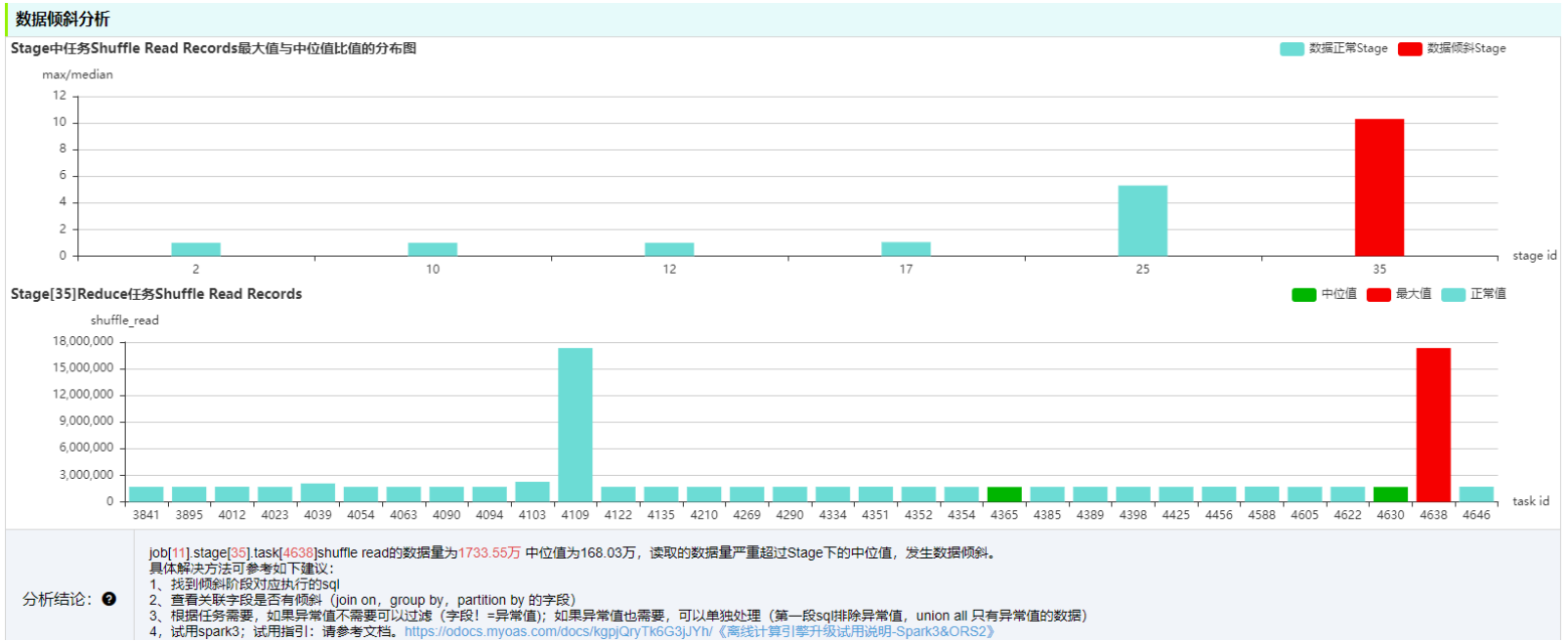
CPU浪费分析

Spark Driver/Executor cores参数配置不合理导致CPU空闲浪费

通过Spark Application采集指标，分析Spark Driver、Spark Executor执行过程中的CPU的运行时间（单位: vcore·second）占比，如果空闲时间超过一定的比例，判定为浪费，用户根据比例降低启用CPU数量

实践效果 / 稳定性分析案例

数据倾斜分析



数据倾斜是Task计算过程中Key分布不均造成的，个别Key的数据特别多，超出计算节点的计算能力。会导致任务内存溢出、计算资源利用率低、作业执行时间超出预期

解决数据倾斜常用方式有：

- 增大并行度spark.sql.shuffle.partitions，使得数据再次分配到不同Task；
- 过滤异常值的数据，过多冗余值也会导致数据倾斜；
- SQL中group by或者RDD的reduceByKey添加key的随机数打散Map, Reduce两个阶段数据，最后在Reduce阶段将随机数去掉；
- 表Join关联时，可以使用Broadcast方式广播小表数据，避免shuffle, 就不会发生数据倾斜；

实践效果 / 稳定性分析案例

SQL常见问题分析

SQL执行过程中没权限、表不存在、语法错误等

根据SQL失败特征从指标数据或者日志提取，用户根据问题去申请相应权限、创建表或者修正语法问题

实践效果 / 实时分析案例

资源利用率分析

Flink参数设置不合理导致资源浪费

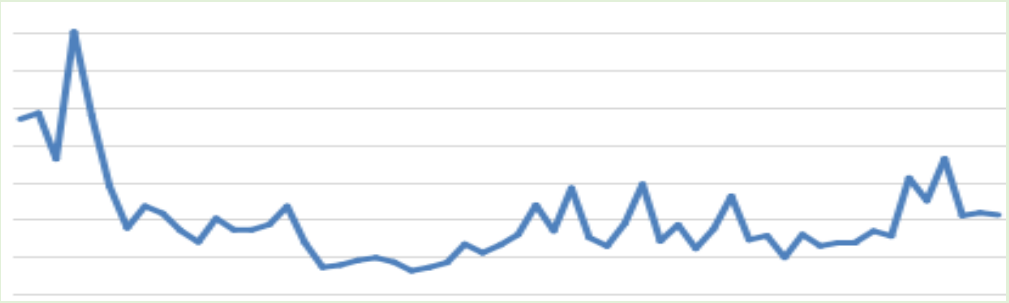
根据Flink作业运行时上报的指标，计算判断CPU、内存利用率是否过低，并给出建议参数调整值

区域	作业名称	集群	诊断结果	诊断来源
CHINA			(点击复制) 作业的tm最大归一化cpu利用率低(1.07%) tm数量在100以内,调整jm内存到1024MB tm数量:12->2 任务并行度:12 单任务节点核数:1 Slots:1->6 单主节点内存大小:4096MB->1024MB Memory:2048MB->6144MB	每日定时诊断
CHINA				每日定时诊断
CHINA			缩减资源 作业的tm最大归一化cpu利用率低(1...	每日定时诊断
CHINA			缩减资源 作业的tm最大归一化cpu利用率低(1...	每日定时诊断
CHINA			缩减资源 作业的tm最大归一化cpu利用率低(1...	每日定时诊断
CHINA			扩容资源 作业部分tm峰值归一化cpu利用率...	每日定时诊断

实践效果 / 降本增效

异常任务、不合理任务分析 → 成本口径统计 → 数据治理

通过长期推进治理，可以看出成本趋势，用户聚焦的任务问题得以改善



04 总结与规划

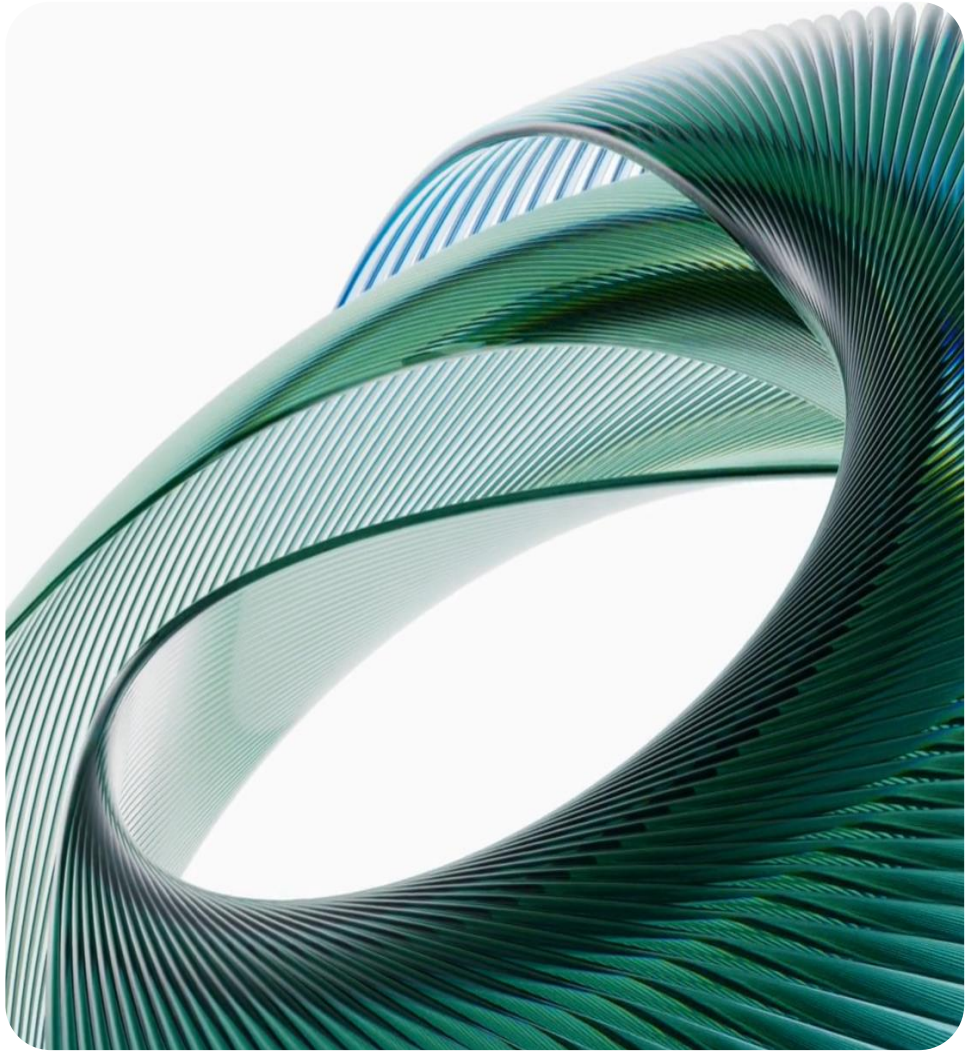
总结与规划

OPPO大数据诊断平台主要围绕 调度引擎 和 计算引擎 两方面进行智能化定位分析，为用户快速处理优化任务，为企业降本增效

技术方面采用非入侵方案对接其他系统，保证了其他系统的安全性

系统架构基于启发式规则定位和分析问题方式，但知识库比较依赖人员经验，计划引入数据挖掘算法扩大检测范围，智能化诊断

支持Spark、Flink任务问题诊断，除OPPO自研调研平台外，还支持DolphinScheduler、Airflow等开源调度平台



总结与规划

罗盘 Compass

https://github.com/cubefs/compass

回馈社区

- 为了回馈开源社区，并希望更多人参与进来，共同解决任务诊断的痛点和难题，我们现已将该项目开源：罗盘（Compass）

版本特性

- 支持多种主流调度平台，例如 DolphinScheduler、Airflow 等
- 支持多版本 Spark、Hadoop 2.x 和 3.x 任务日志诊断和解析
- 支持引擎层异常诊断，包含数据倾斜、大表扫描、内存浪费等 14 种异常类型
- 支持各种日志匹配规则编写和异常阈值调整，可自行根据实际场景优化

2023 DataFunCon

— THANKS —

感谢您的观看

演讲人：戴巍 - OPPO - 数据平台架构师