饿了么

饿了么大数据离线平台架构

主讲人：翟玉勇
时间：2017.06.24

饿了么

大数据离线平台现状

饿了么

Agenda

大数据离线平台现状

面临的挑战

架构设计和技术选型

平台工具链

离线平台的一些想法

离线平台规模

增量(不考虑副本) 100TB/day

集群规模 1000-1500 nodes, x10 expanding 表数据 90K表 400报表

调度任务 20K+

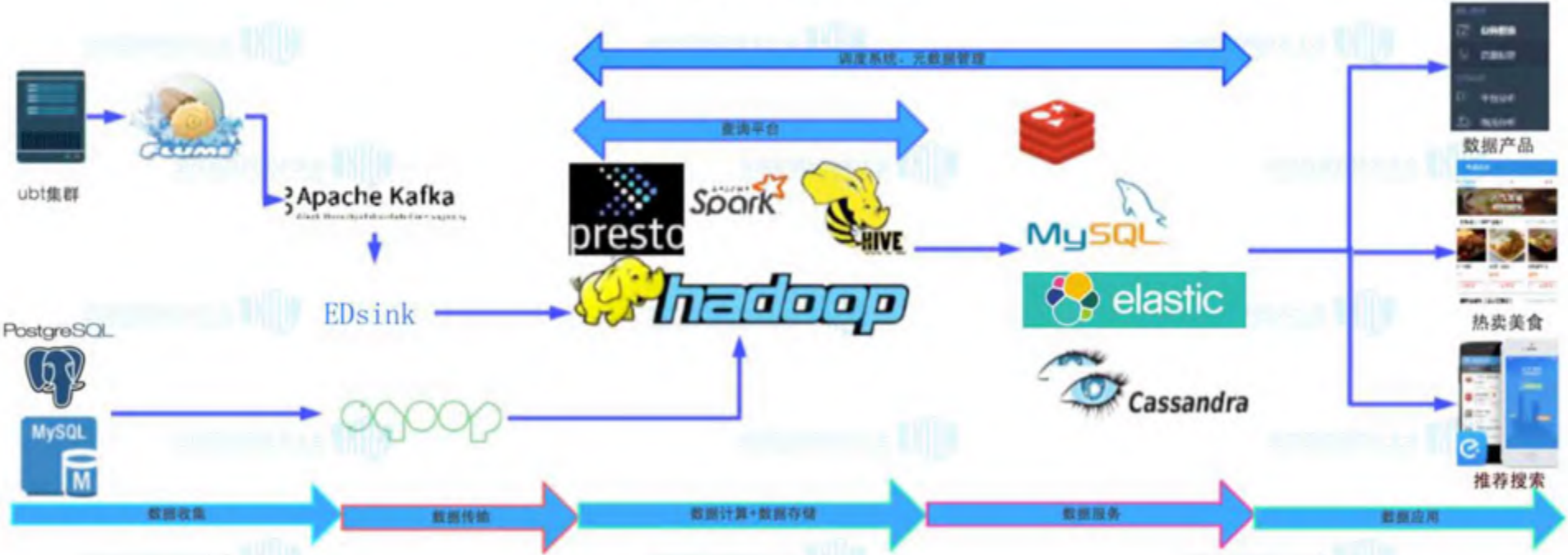
任务数 10W+ mapreduce/spark

计算数据吞吐 3PB/day

2015年5月
团队成立

2年

饿了么离线集群架构图


调度系统：元数据管理
查询平台
ubt集群
Apache Kafka
EDsink
PostgreSQL
MySQL
presto
Spark
HIVE
hadoop
MySQL
elastic
Cassandra
数据产品
热卖美食
推荐搜索
数据收集
数据传输
数据计算+数据存储
数据服务
数据应用

应用支撑

BU ETL

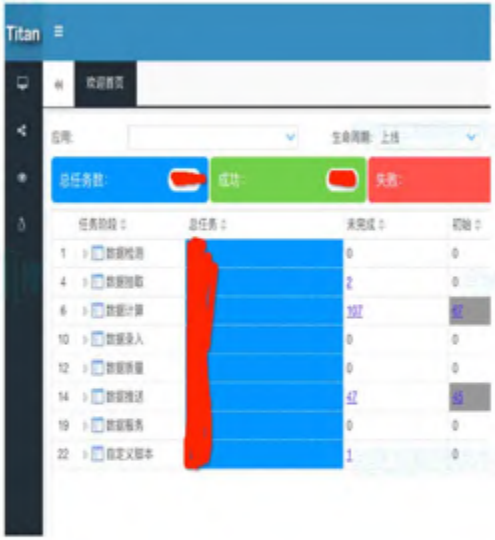
报表支撑

数据接口

用户画像

历史订单

数据接口



数据来源

交易数据

应用数据

流量数据

饿了么

面临的挑战

人少活多 积累不足

历史的技术债

应对不断的业务需求

饿了么

架构设计和技术选型

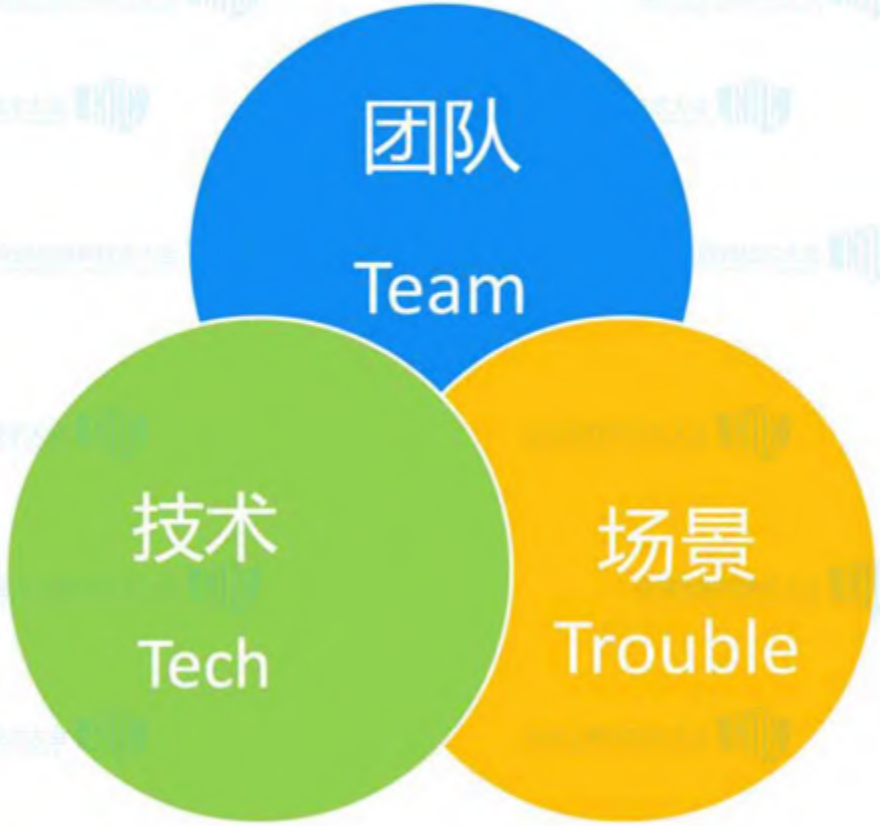
架构设计和技术选型

3T

Trouble：解决什么问题

Tech：哪些合适的技术，生态和社区状态

Team：熟悉程度/学习成本/使用成本/运维成本



数据收集 Flume vs EDisk(storm)

场景
数据从kafka流到hdfs

团队
storm相对于flume经验多

技术
运维成本 edisk
使用成本 edisk

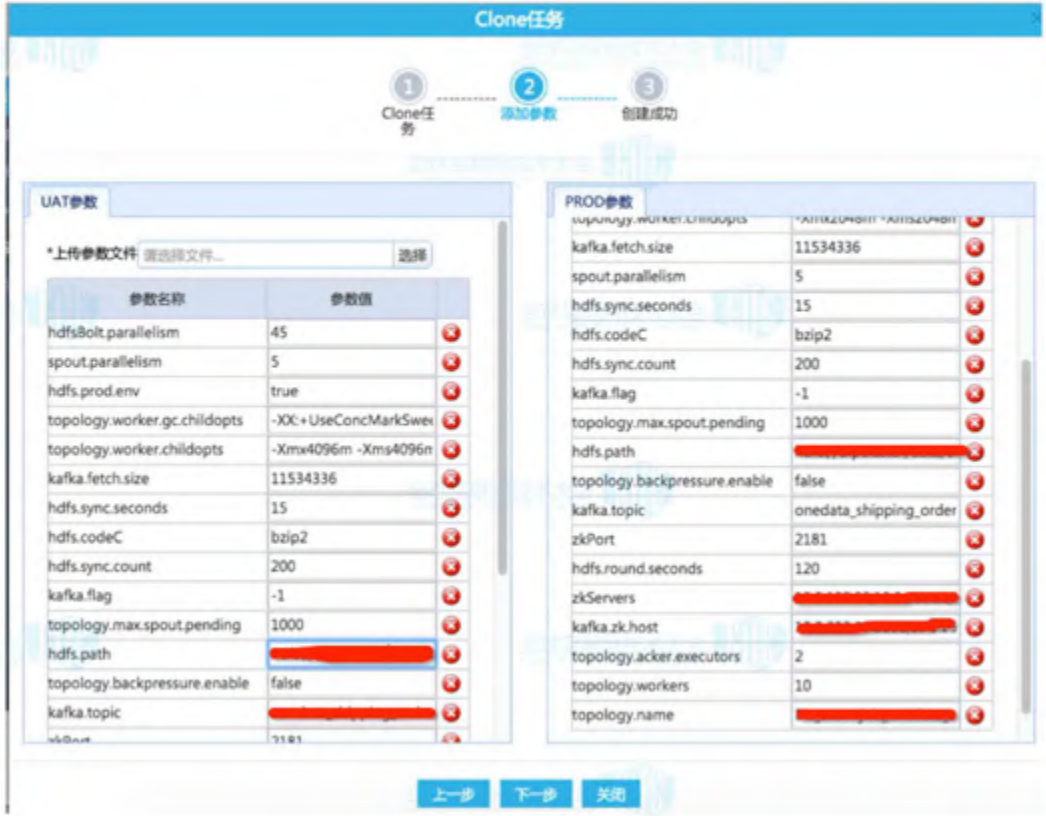
EDSink(storm)

自助抽取数据到hive

不同的SLA等级

降级/熔断策略

EDSink(storm)


Clone任务
1 Clone任务
2 添加参数
3 创建成功
UAT参数
*上传参数文件
选择
参数名称
参数值
hdfsBolt.parallelism 45
spout.parallelism 5
hdfs.prod.env true
topology.worker.gc.childopts -XX:+UseConcMarkSwe
topology.worker.childopts -Xmx4096m -Xms4096m
kafka.fetch.size 11534336
hdfs.sync.seconds 15
hdfs.codeC bzip2
hdfs.sync.count 200
kafka.flag -1
topology.max.spout.pending 1000
hdfs.path
topology.backpressure.enable false
kafka.topic
PROD参数
kafka.fetch.size 11534336
spout.parallelism 5
hdfs.sync.seconds 15
hdfs.codeC bzip2
hdfs.sync.count 200
kafka.flag -1
topology.max.spout.pending 1000
hdfs.path
topology.backpressure.enable false
kafka.topic onedata_shipping_order
zkPort 2181
hdfs.round.seconds 120
zkServers
kafka.zk.host
topology.acker.executors 2
topology.workers 10
topology.name
上一步
下一步
关闭

数据计算 Spark vs Hive

场景
- ETL计算

团队
- hive略熟

技术
- 运维成本 hive
- 使用成本 spark
- 稳定性 hive
- 速度 spark

数据落地 Hbase vs Cassandra

场景
海量存储/批量更新/K-V(object)访问

团队
不熟悉 学习成本差不多

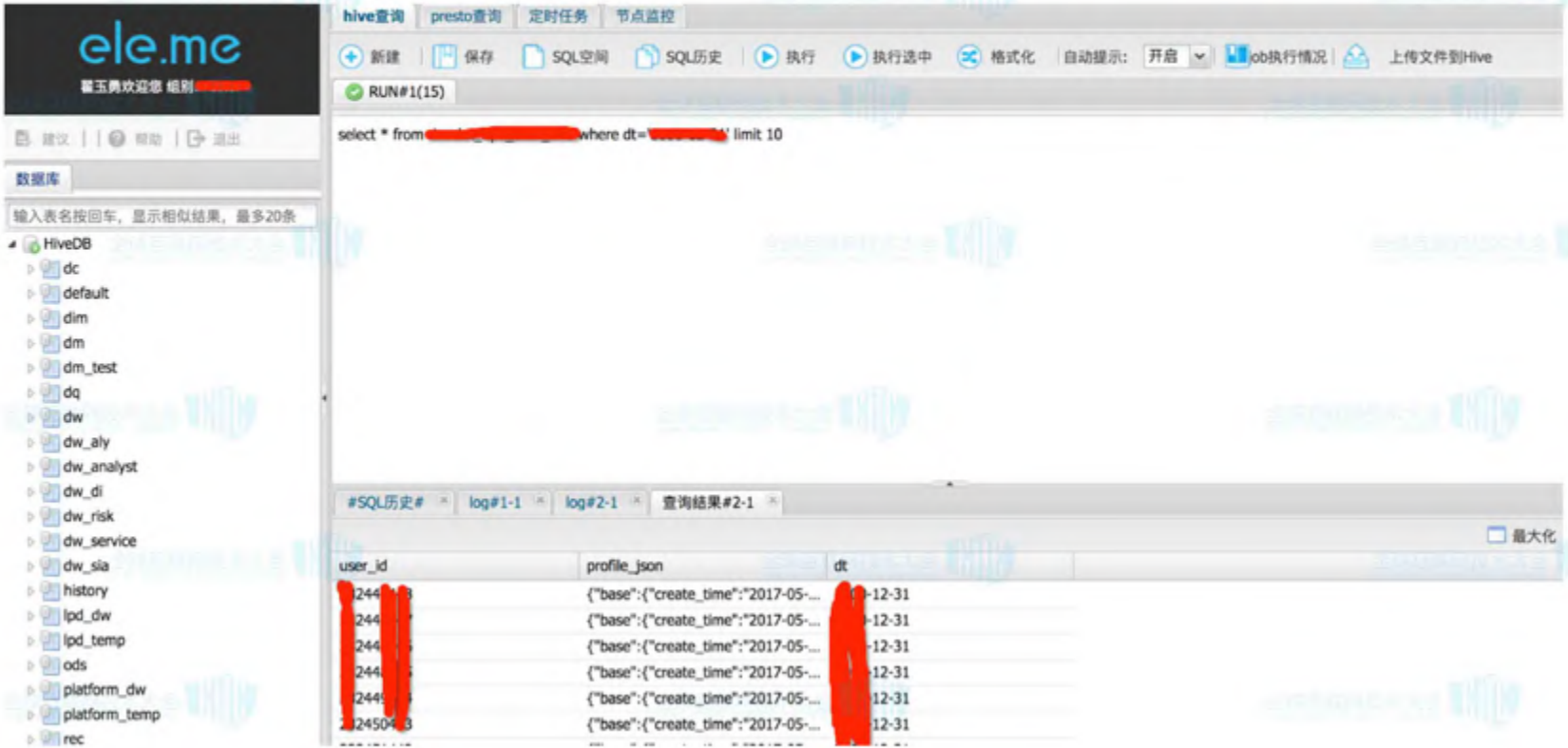
技术
社区方面 国内Hbase相对于Cassandra
运维成本 Cassandra较低
使用成本 Cassandra使用方便
解决问题 Cassandra天然支持多机房策略

饿了么

平台工具链


报表
数据接口
Ad-hoc
数据分析
数据开发
调度引擎
流计算任务管理
基础设施管理
元数据管理
权限管理

Adhoc平台


ele.me
hive查询
presto查询
定时任务
节点监控
新建
保存
SQL空间
SQL历史
执行
执行选中
格式化
自动提示:
开启
job执行情况
上传文件到Hive
RUN#1(15)
select * from [redacted] where dt=[redacted] limit 10
数据库
输入表名按回车，显示相似结果，最多20条
HiveDB
dc
default
dim
dm
dm_test
dq
dw
dw_aly
dw_analyst
dw_di
dw_risk
dw_service
dw_sla
history
lpd_dw
lpd_temp
ods
platform_dw
platform_temp
rec
#SQL历史#
log#1-1
log#2-1
查询结果#2-1
最大化
user_id
profile_json
dt
{"base":{"create_time":"2017-05-...
-12-31

Adhoc平台

多引擎支持 hive/presto/spark

统一权限验证

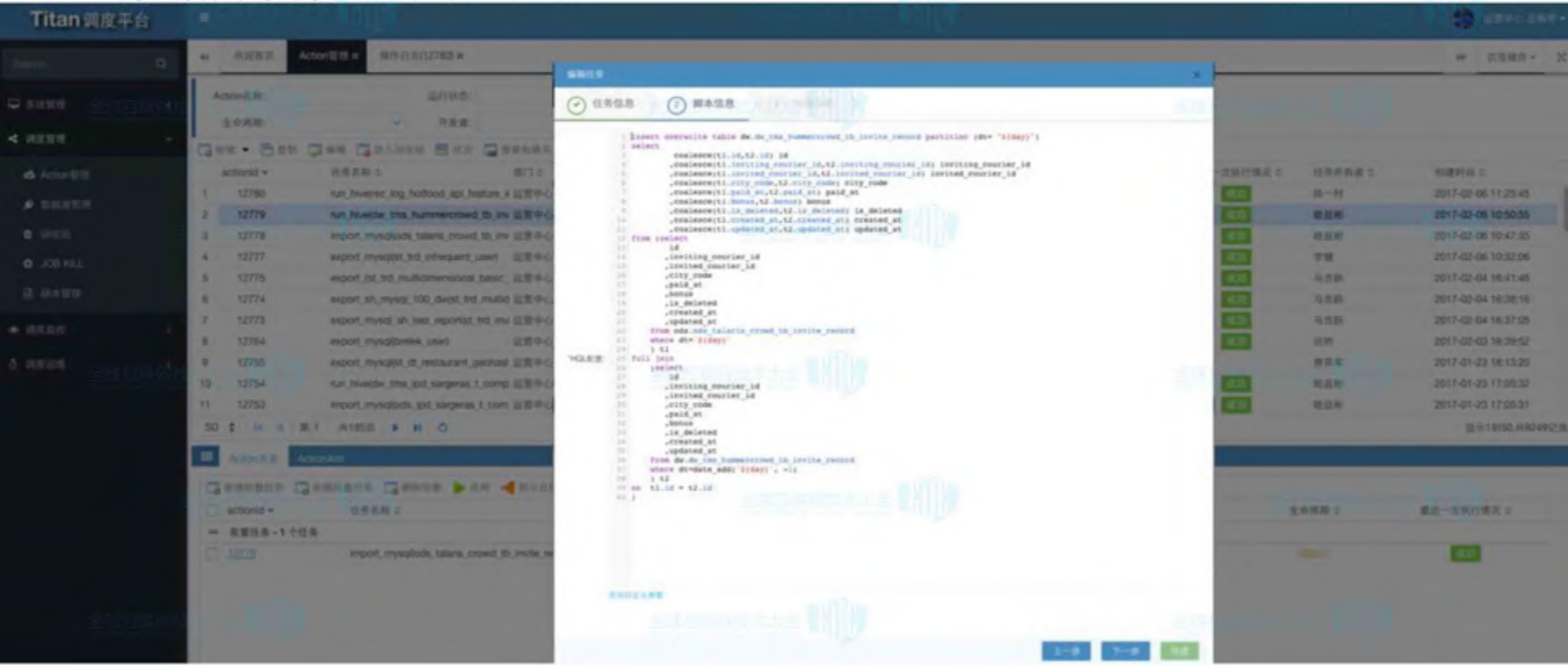
即席/定时

运行状态

历史记录

资源隔离

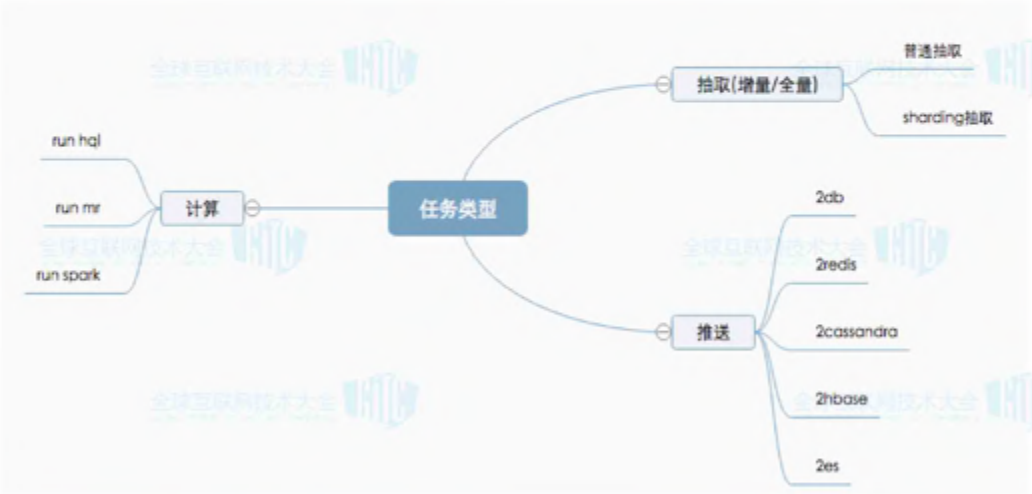
数据开发管理平台-ETL



数据开发管理平台-ETL

任务类型

1. 数据抽取
2. 数据计算
3. 数据推送



数据开发管理平台-ETL

底层工具支持

1. hql并发支持和并发限制
2. hive2redis/hive2cassandra/hive2es等等
3. hive2db merge优化
4. sharding抽取
4. spark list和map类型udf兼容
5. 多引擎一键切换

元数据管理


Horus
业务指标监控
监控看板
指标配置
数据仓库管理
Hive 表管理
码表管理
数据抽取
工单管理
请输入表名
搜索
高级搜索
创建新表
hive表名
数据库名
一级主题
表热度
表容量(G)
生命周期
所有者
创建时间
dw 交易 3984891 433593.79 三个月 2014-12-15 18:23:41
dw 基础流量 244514 299537.25 一个月 2016-11-16 16:13:31
dw_aly 大数据 74084 241319.87 半年 2016-10-31 13:48:43
dw 基础流量 2159232 230560.72 半年 2015-06-25 11:49:53
dw 物流 160311 206932.1 三个月 2016-04-26 16:52:37
dm 交易 26211247 183677.28 三个月 2017-01-23 18:21:06
dw 交易 1496845 156679.28 三个月 2014-11-18 15:30:48
dw 基础流量 2185338 155635.78 半年 2015-12-24 12:24:56
dc 基础流量 68232 148178.87 半年 2015-10-24 11:06:55
dw 交易 11957669 139078.92 三个月 2016-03-04 14:14:03
dw_aly 大数据 28212 136523.42 半年 2017-01-06 12:22:23
platform_dw 平台研发中心 2059825 126603.2 一年 2016-11-24 16:38:20
50 条/页
1 2 3 4 5 6 ··· 279
前往 1 页

元数据管理

- 统一表管理
- 表生命周期控制
- 重要数据表一键备份
- 表热度管理和监控
- 表容量管理

用户自助任务分析


Dr.Grace
首页
今日分析
搜索
错误查询
帮助
集群
ETL集群
应用ID (精确匹配)
job_1496358728688_1238564
用户名 (精确匹配)
User
队列名
action_sid (调度任务sid)
actionSid
query_id (dtquery查询id)
queryId
严重性
错误项
[root.bigdata.etl.hourlyetl.veryhigh][[redacted]] [Hive] job_1496358728688_1238564
2017-06-18 17:14:06
4936_18294047:st_tms_report_dashboard_monitor_hour.sql:s1q1[redacted]:create temp.temp_li...
Jobtracker: [redacted]jobhistory/job/job_1496358728688_1238564
等待启动时间
Mapper数据倾斜
MapperGC
Mapper时间
Mapper速度
Mapper溢出
Mapper内存
Reducer的数据倾斜
ReducerGC
Reducer时间
Reducer内存
ShuffleSort
0.222 GB Hours
19.26 %
0:00:38
0:00:05
21.05 %
[Explain]
等待启动时间
Severity: success
Start Time 1497777208000
Submit Time 1497777203000
Wait Time 5 sec
Mapper数据倾斜
Severity: danger [Explain]
Group A 47 tasks @ 20 MB avg
Group B 1 tasks @ 1 GB avg
Number of tasks 48

用户自助任务分析

任务严重等级

任务分析指标

1. 等待启动时间
2. map 数据倾斜/GC时间/速度/内存
3. reduce数据倾斜/GC时间/速度/内存
4. shuffle

推荐优化参数

用户自助任务分析

主题

性能指标
资源使用
资源浪费
运行时间
等待时间

Mapred任务
Mapper数据倾斜
Mapper GC
Mapper时间
Mapper速度
Mapper溢出
Mapper内存
Reducer数据倾斜
Reducer GC
Reducer时间
Reducer内存
Shuffle & Sort
异常信息

Reducer的数据倾斜

分析显示reducer task是否有数据倾斜.

该分析结果显示两组光谱，其中第一组与第二组相比具有显着较少的输入数据.

例子

Reducer Data Skew

Severity: Critical

Number of tasks	999
Group A	875 tasks @@ 28 MB avg
Group B	124 tasks @@ 2 GB avg

建议

通常倾斜是由于key导致(比如group by的key,join的key),需要找出倾斜键并进行sql优化.

Spark在线交互式数据分析平台

Zeppelin Notebook Job Search your Notes

Hi Head default

```
import org.apache.spark.ml.feature.PCA
import org.apache.spark.ml.linalg.Vectors

val data = Array(
  Vectors.sparse(5, Seq((1, 1.0), (3, 7.0))),
  Vectors.dense(2.0, 0.0, 3.0, 4.0, 5.0),
  Vectors.dense(4.0, 0.0, 0.0, 6.0, 7.0)
)
val df = spark.createDataFrame(data.map(Tuple1.apply)).toDF("features")
val pca = new PCA().setInputCol("features").setOutputCol("pcaFeatures").setK(3).fit(df)
val pcaDF = pca.transform(df)
val result = pcaDF.select("pcaFeatures")
result.show()
```

FINISHED

```
+--------------------+
|         pcaFeatures|
+--------------------+
|[1.64857282308838...|
|[-4.6451043317815...|
|[-6.4288805356764...|
+--------------------+
```

Took 3 sec. Last updated by [redacted] at May 19 2017, 11:31:44 AM.

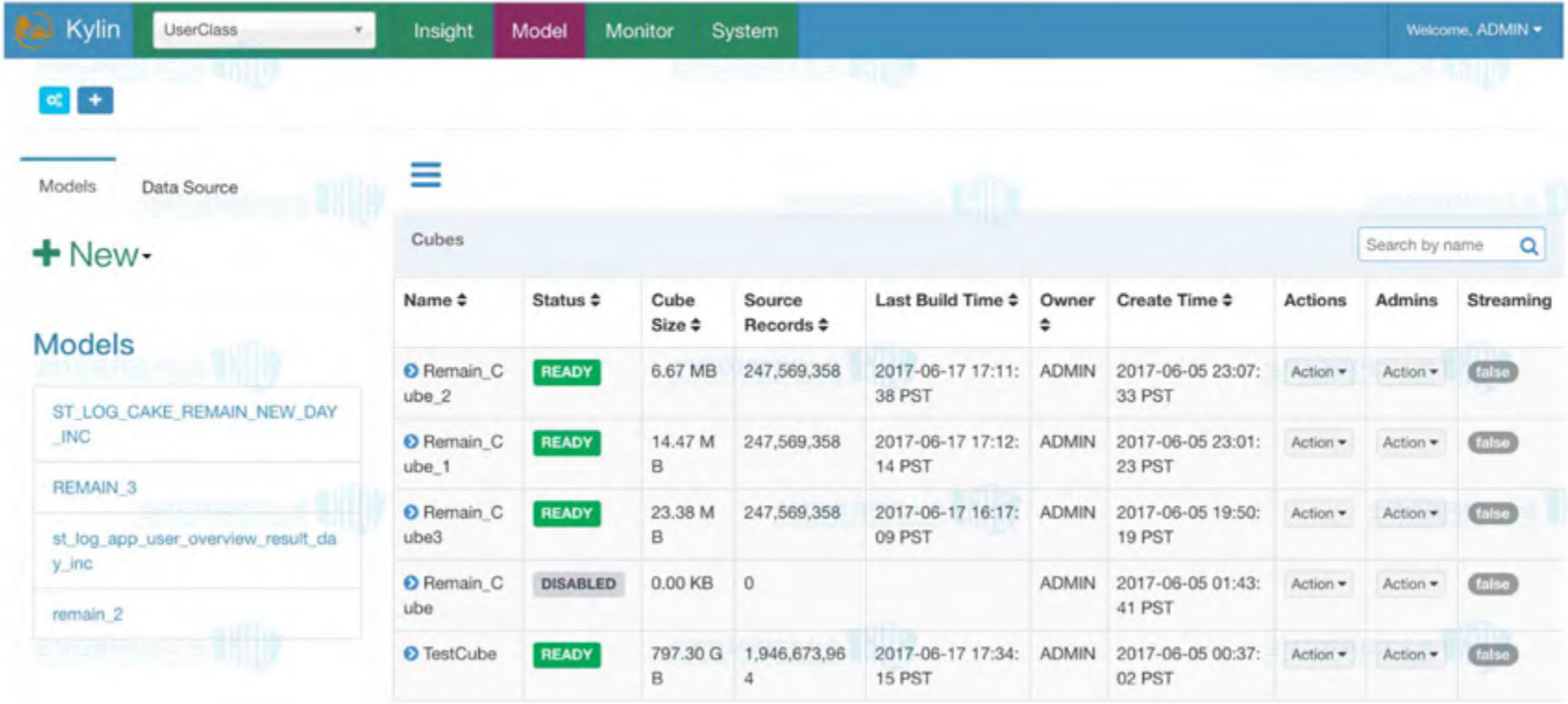
Spark在线交互式数据分析平台

统一的权限控制

资源管控

方便分析师试错

Kylin多维分析


Kylin
UserClass
Insight
Model
Monitor
System
Welcome, ADMIN
Models
Data Source
New
Models
ST_LOG_CAKE_REMAIN_NEW_DAY_INC
REMAIN_3
st_log_app_user_overview_result_day_inc
remain_2
Cubes
Search by name
Name
Status
Cube Size
Source Records
Last Build Time
Owner
Create Time
Actions
Admins
Streaming
Remain_Cube_2
READY
6.67 MB
247,569,358
2017-06-17 17:11:38 PST
ADMIN
2017-06-05 23:07:33 PST
Action
Action
false
Remain_Cube_1
READY
14.47 MB
247,569,358
2017-06-17 17:12:14 PST
ADMIN
2017-06-05 23:01:23 PST
Action
Action
false
Remain_Cube3
READY
23.38 MB
247,569,358
2017-06-17 16:17:09 PST
ADMIN
2017-06-05 19:50:19 PST
Action
Action
false
Remain_Cube
DISABLED
0.00 KB
0
ADMIN
2017-06-05 01:43:41 PST
Action
Action
false
TestCube
READY
797.30 GB
1,946,673,964
2017-06-17 17:34:15 PST
ADMIN
2017-06-05 00:37:02 PST
Action
Action
false

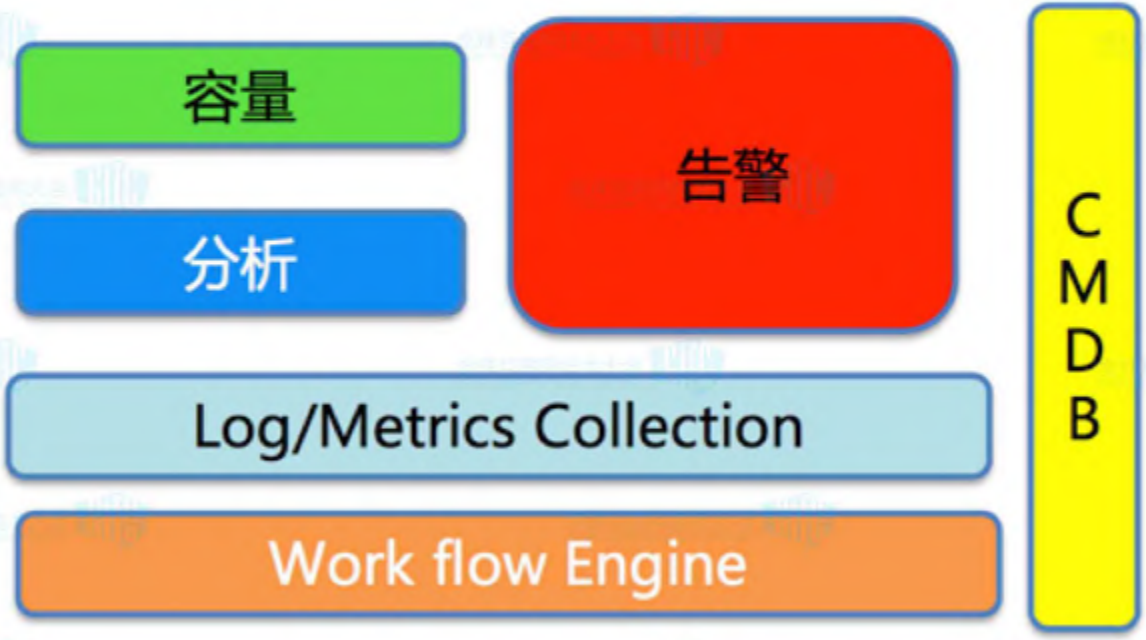
Kylin多维分析

预计算

多维展示、秒级响应

spark/hive 构建cube

基础设施管理平台



基础设施管理平台

Cmdb 平台自动化的基础

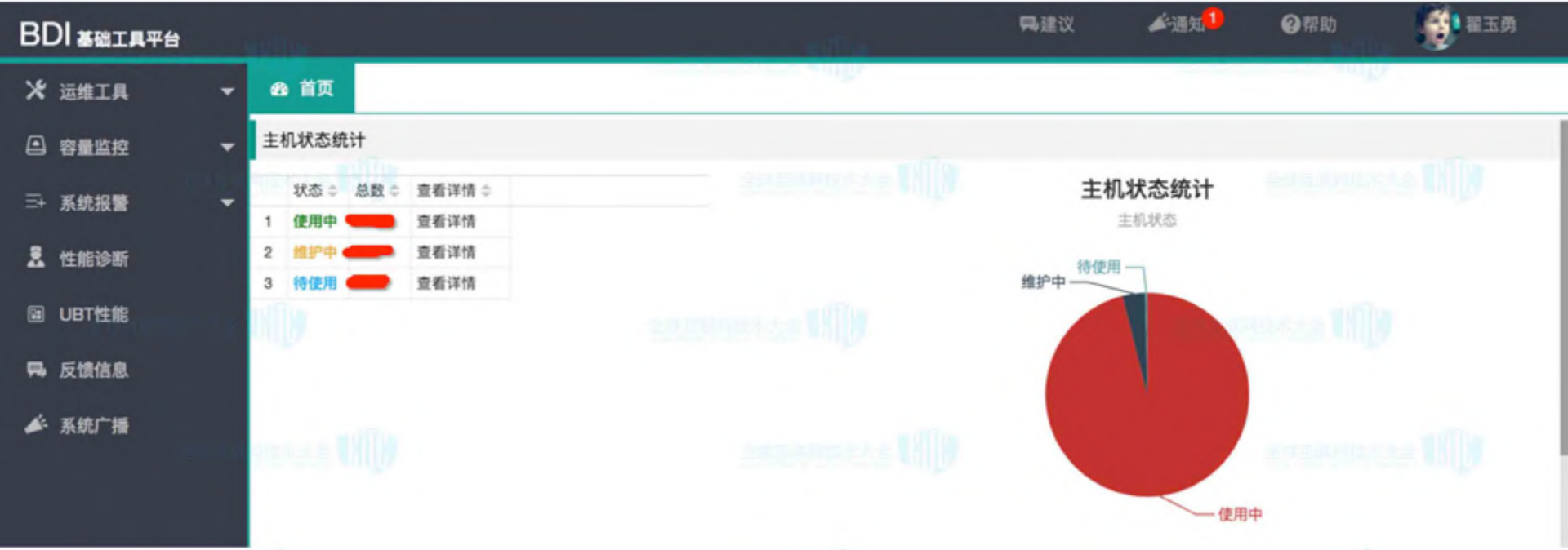
Ops 重复的操作固化到workflow中

Alter 告警的管理和分析

Capacity 访问热度，空间增量/速，计算增量

Analysis 统一的任务/链路性能分析

基础设施管理平台


BDI 基础工具平台
建议
通知 1
帮助
翟玉勇
运维工具
容量监控
系统报警
性能诊断
UBT性能
反馈信息
系统广播
首页
主机状态统计
状态
总数
查看详情
1 使用中 查看详情
2 维护中 查看详情
3 待使用 查看详情
主机状态统计
主机状态
待使用
维护中
使用中

生产HDFS分BU容量

风控部门

平台研发部门

阿里云

物流部门

分析团队

饿了么

离线平台一些想法

平台稳定性

回滚方案 回滚方案 回滚方案

double check double check

灰度 灰度 灰度

停服时间估算

智能化运维

报警关联分析

ETL链自动化分析

一键操作

平台化支撑

资源审计

资源治理

饿了么

THANKS！