



饿了么大数据离线平台架构

主讲人：翟玉勇

时间：2017.06.24



大数据离线平台现状





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

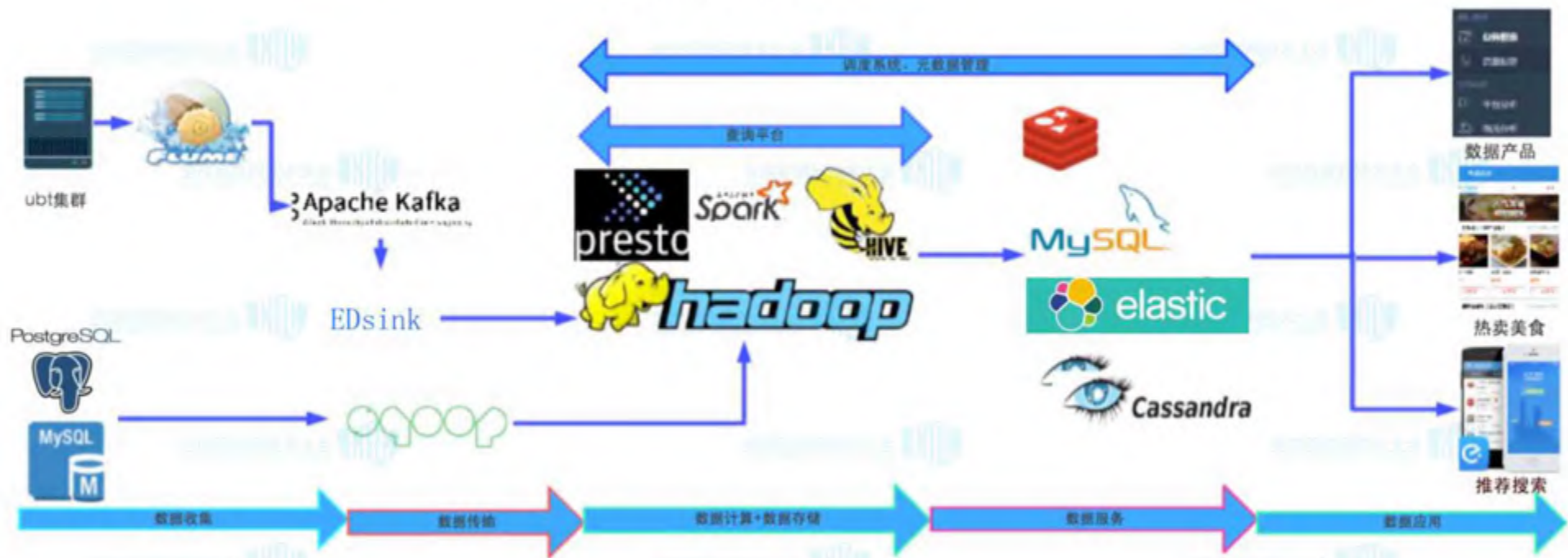
计算数据吞吐 3PB/day

2015年5月
团队成立

2年



饿了么离线集群架构图



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

The image displays two screenshots from the Titan system. The left screenshot shows a dashboard with a table of tasks and a bar chart of device usage. The right screenshot shows a mobile app interface with a list of orders.

任务名称	总任务数	成功	失败
1	数据同步	0	0
4	数据抽取	2	0
6	数据计算	13	0
10	数据导入	0	0
12	数据清理	0	0
14	数据推送	12	0
19	数据服务	0	0
22	自定义脚本	1	0

产品	数量
iPhone 8	1
iPhone 8 Plus	1
iPhone 6S	1
iPhone 6S Plus	1
iPhone 5S	1
Other	1
iPhone 5	1
Mi 4/2	1
OPPO R9s	1
Redmi Note 3	1

商家名称	订单时间	商品	金额	状态
二十四道煲仔饭	2016-09-17 10:17	红烧大腩煲仔饭等2件商品	¥42.00	订单已完成
农家小菜 (中厨路店)	2016-09-17 14:35	红烧鲫鱼等6件商品	¥59.00	订单已完成
老大烧烤大排档	2016-09-16 19:38	早的鸡14件商品	¥34.00	订单已完成
养生粥	2016-09-16 18:18	皮蛋瘦肉粥等6件商品	¥33.00	订单已完成
川湘小厨	2016-09-16 21:11	番茄炒蛋等4件商品	¥26.00	订单已完成

数据来源

交易数据

应用数据

流量数据

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

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)

饿了么·实时计算平台

任务管理 集群管理 任务版本管理 集群权限分配 binlog管理页面

登录 退出

名称: hdfs

实时计算任务列表

名称	项目名称	状态	集群U1	集群	日志查询	任务类型	任务操作	系统操作
1	bigDataRealtime	已上线	Elu_stormjob		Log Link	storm	发布 更新 管理 详情 删除 克隆 设置	
2	bigDataRealtime	已上线	Elu_stormjob		Log Link	storm	发布 更新 管理 详情 删除 克隆 设置	
3	bigDataRealtime	已上线	Elu_stormjob		Log Link	storm	发布 更新 管理 详情 删除 克隆 设置	
4	bigDataRealtime	已上线	Elu_stormjob		Log Link	storm	发布 更新 管理 详情 删除 克隆 设置	
5	bigDataRealtime	已上线	Elu_stormjob		Log Link	storm	发布 更新 管理 详情 删除 克隆 设置	
6	bigDataRealtime	已上线	Elu_stormjob		Log Link	storm	发布 更新 管理 详情 删除 克隆 设置	
7	bigDataRealtime	已上线	Elu_stormjob		Log Link	storm	发布 更新 管理 详情 删除 克隆 设置	
8	bigDataRealtime	已上线	Elu_stormjob		Log Link	storm	发布 更新 管理 详情 删除 克隆 设置	
9	bigDataRealtime	已上线	Elu_stormjob		Log Link	storm	发布 更新 管理 详情 删除 克隆 设置	
10	bigDataRealtime	已上线	Elu_stormjob		Log Link	storm	发布 更新 管理 详情 删除 克隆 设置	
11	bigDataRealtime	已上线	Elu_stormjob		Log Link	storm	发布 更新 管理 详情 删除 克隆 设置	
12	bigDataRealtime	已上线	Elu_stormjob		Log Link	storm	发布 更新 管理 详情 删除 克隆 设置	
13	bigDataRealtime	已上线	Elu_stormjob		Log Link	storm	发布 更新 管理 详情 删除 克隆 设置	
14	bigDataRealtime	已上线	Elu_stormjob		Log Link	storm	发布 更新 管理 详情 删除 克隆 设置	

50 第 1 共 1 页 批量发布 批量重启 批量Kill

显示1到14,共14条

运行状况

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 请选择文件... 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	[REDACTED]	⌵
topology.backpressure.enable	false	⌵
kafka.topic	[REDACTED]	⌵
zkPort	2181	⌵

PROD参数

supernogy.worker.childopts	*ATTI2UPH0IT*ATTI2UPH0IT	⌵
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	[REDACTED]	⌵
topology.backpressure.enable	false	⌵
kafka.topic	onedata_shipping_order	⌵
zkPort	2181	⌵
hdfs.round.seconds	120	⌵
zkServers	[REDACTED]	⌵
kafka.zk.host	[REDACTED]	⌵
topology.acker.executors	2	⌵
topology.workers	10	⌵
topology.name	[REDACTED]	⌵

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

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

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

Titan 调度平台
运维中心 王梅平

Action 列表

任务ID	任务名称	运行状态
12778	run_hivejar_log_hadoop_spl_feature	运行中心
12779	run_hivejar_tm_humercrowd_to_m	运行中心
12778	import_mysql2ds_talark_crowd_to_m	运行中心
12777	export_mysql2td_spl_sequent_user	运行中心
12776	export_bt_bd_multidimensional_base	运行中心
12774	export_bt_mysql_100_dept_bd_multid	运行中心
12773	export_mysql_bp_3es_report_bt_m	运行中心
12784	export_mysql2bmk_user	运行中心
12785	export_mysql2td_restaurant_gethad	运行中心
12784	run_hivejar_tm_spl_sargeras_1_comp	运行中心
12785	import_mysql2ds_spl_sargeras_1_com	运行中心

编辑任务

任务信息 | 脚本信息

```

1 1 select overwrite table dw_dw_tm_humercrowd_to_talark_reward partition (dt = '${date}')
2 select
3   *
4   ,coalesce(m11.is_deleted,0) is_deleted
5   ,coalesce(m11.inviting_cursor_id,m12.inviting_cursor_id) inviting_cursor_id
6   ,coalesce(m11.invited_cursor_id,m12.invited_cursor_id) invited_cursor_id
7   ,coalesce(m11.city_code,m12.city_code) city_code
8   ,coalesce(m11.paid_at,m12.paid_at) paid_at
9   ,coalesce(m11.bonus,m12.bonus) bonus
10  ,coalesce(m11.is_deleted,m12.is_deleted) is_deleted
11  ,coalesce(m11.updated_at,m12.updated_at) updated_at
12 from (select
13   id
14   ,inviting_cursor_id
15   ,invited_cursor_id
16   ,city_code
17   ,paid_at
18   ,bonus
19   ,is_deleted
20   ,updated_at
21 from ods_ods_talark_crowd_to_talark_reward
22 where dt='${date}')
23 t1
24
25 Full join
26 select
27   id
28   ,inviting_cursor_id
29   ,invited_cursor_id
30   ,city_code
31   ,paid_at
32   ,bonus
33   ,is_deleted
34   ,updated_at
35 from dw_dw_tm_humercrowd_to_talark_reward
36 where dt='${date}' = t1
37 t2
38 on t1.id = t2.id
39
40
                
```

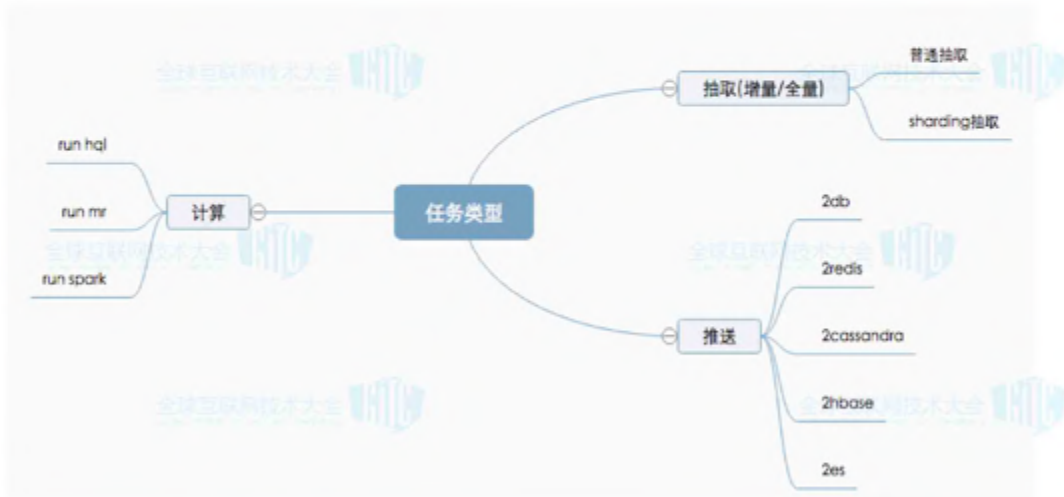
任务执行记录

任务名称	任务所属用户	创建时间
第一村		2017-02-06 11:25:43
数据源		2017-02-06 10:50:55
数据源		2017-02-06 10:47:35
李健		2017-02-06 10:32:06
马志凯		2017-02-04 16:41:46
马志凯		2017-02-04 16:38:16
马志凯		2017-02-04 16:37:05
吕然		2017-02-03 16:29:52
曹春平		2017-01-23 16:13:25
魏益新		2017-01-23 17:05:37
魏益新		2017-01-23 17:05:31

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

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
[REDACTED]	dw	交易	3984891	433593.79	三个月	[REDACTED]	2014-12-15 18:23:41
[REDACTED]	dw	基础流量	244514	299537.25	一个月	[REDACTED]	2016-11-16 16:13:31
[REDACTED]	dw_aly	大数据	74084	241319.87	半年	[REDACTED]	2016-10-31 13:48:43
[REDACTED]	dw	基础流量	2159232	230560.72	半年	[REDACTED]	2015-06-25 11:49:53
[REDACTED]	dw	物流	160311	206932.1	三个月	[REDACTED]	2016-04-26 16:52:37
[REDACTED]	dm	交易	26211247	183677.28	三个月	[REDACTED]	2017-01-23 18:21:06
[REDACTED]	dw	交易	1496845	156679.26	三个月	[REDACTED]	2014-11-18 15:30:48
[REDACTED]	dw	基础流量	2185338	155635.76	半年	[REDACTED]	2015-12-24 12:24:56
[REDACTED]	dc	基础流量	68232	148178.87	半年	[REDACTED]	2015-10-24 11:06:55
[REDACTED]	dw	交易	11957669	139078.92	三个月	[REDACTED]	2016-03-04 14:14:03
[REDACTED]	dw_aly	大数据	28212	136523.42	半年	[REDACTED]	2017-01-06 12:22:23
[REDACTED]	platform_dw	平台研发中心	2059825	126603.2	一年	[REDACTED]	2016-11-24 16:38:20

共 [REDACTED] 条 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著较少的输入数据.

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



Notebook - Job

Search your Notes



[redacted]

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_daily_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

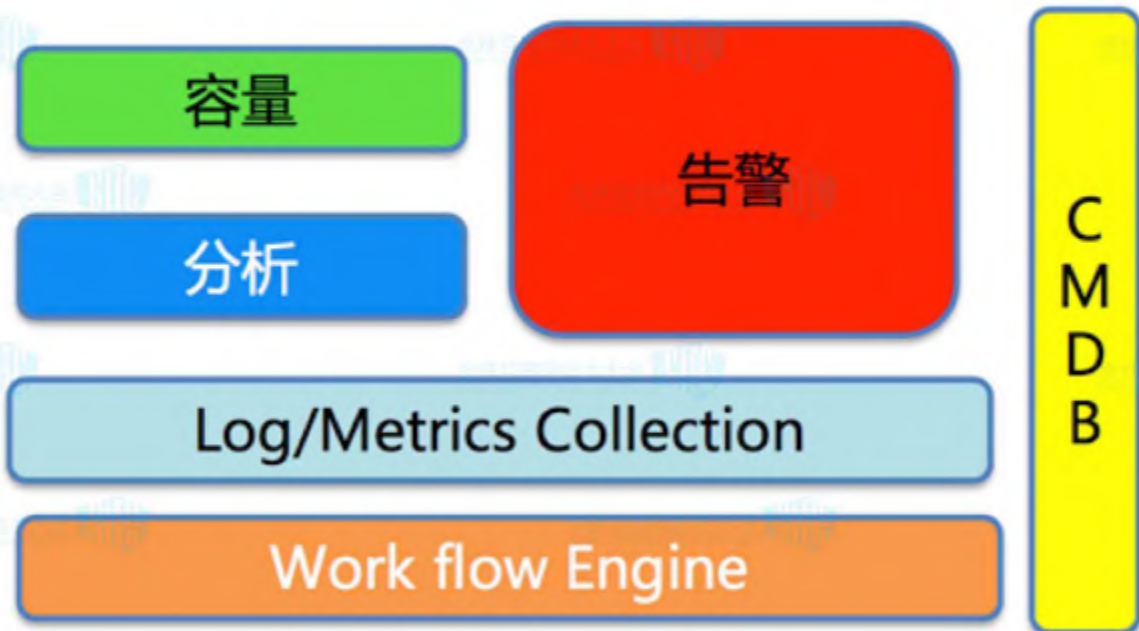
Kylin多维分析

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

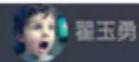
Analysis 统一的任务/链路性能分析

基础设施管理平台

提建议

通知 1

帮助



翟玉勇

BDI 基础工具平台

运维工具

首页

容量监控

主机状态统计

系统报警

性能诊断

UBT性能

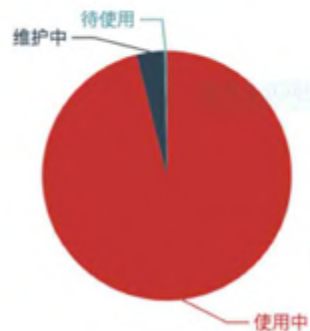
反馈信息

系统广播

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

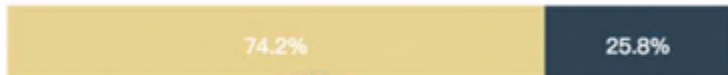
主机状态



生产HDFS分BU容量

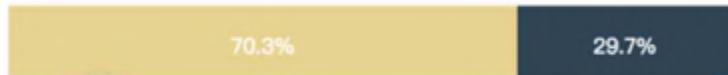
风控部门

已用 剩余



平台研发部门

已用 剩余



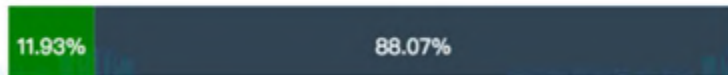
阿里云

已用 剩余



物流部门

已用 剩余



分析团队

已用 剩余

已用 剩余

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

停机时间估算

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



THANKS !

