

基于Hive/ES 金融大数据指标系统

沈百军 2016年10月 @QCon



促进软件开发领域知识与创新的传播



关注InfoQ官方信息
及时获取QCon软件开发者
大会演讲视频信息

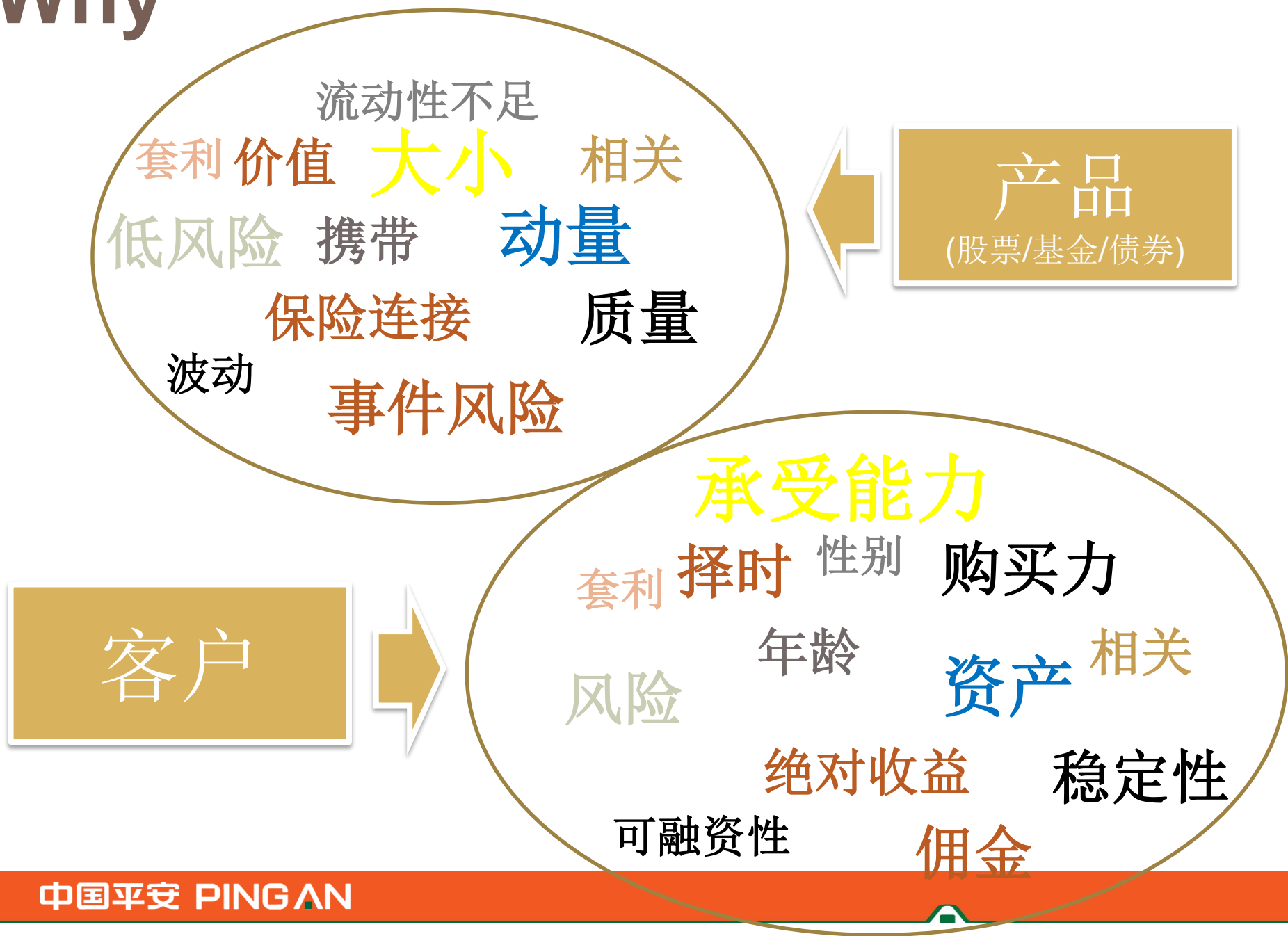


[北京站] 2016年12月2日-3日
咨询热线: 010-89880682



[北京站] 2017年4月16日-18日
咨询热线: 010-64738142

Why



Why-目标要求

- 超级标签系统，标签可无限制扩展，所有标签可联合Filter
- 超级**Cube**，可无限制扩展维度和度量值
- 毫秒级别的多维分析工具，用户体验好，查询快
- 水平扩展，必须是分布式，最好开源
- 架构部署简单，维护成本低
- 支持Sql查询，入门快

其他系统比较

	ElasticSearch	Kylin	Spark	HBase	SSAS Cognos
聚合度量值 排序	√	X	√	X	√
维度&度量值 扩展性	10W~50W	32个维度	10W	1维	有限个
条件明细筛 选	√	X	√	X	X
水平扩展型	√	√	√	√	X
易用性	√	√	√	√	X
性能和用户 体验	毫秒	毫秒	10~60s	毫秒	2~10s
高维聚合查 询	X	√	√	√	X
数据实时性	流式写入	流式写入	流式写入	流式写入	X

方案



Hive



ElasticSearch

Schema



系统介绍

什么是指标系统

条件筛选 & 多维统计分析

指标定义和维护

➤ 什么是指标系统（3W）

WHO

- 对象：给谁打指标，必须有一个主题对象（如客户、股票、视频等）

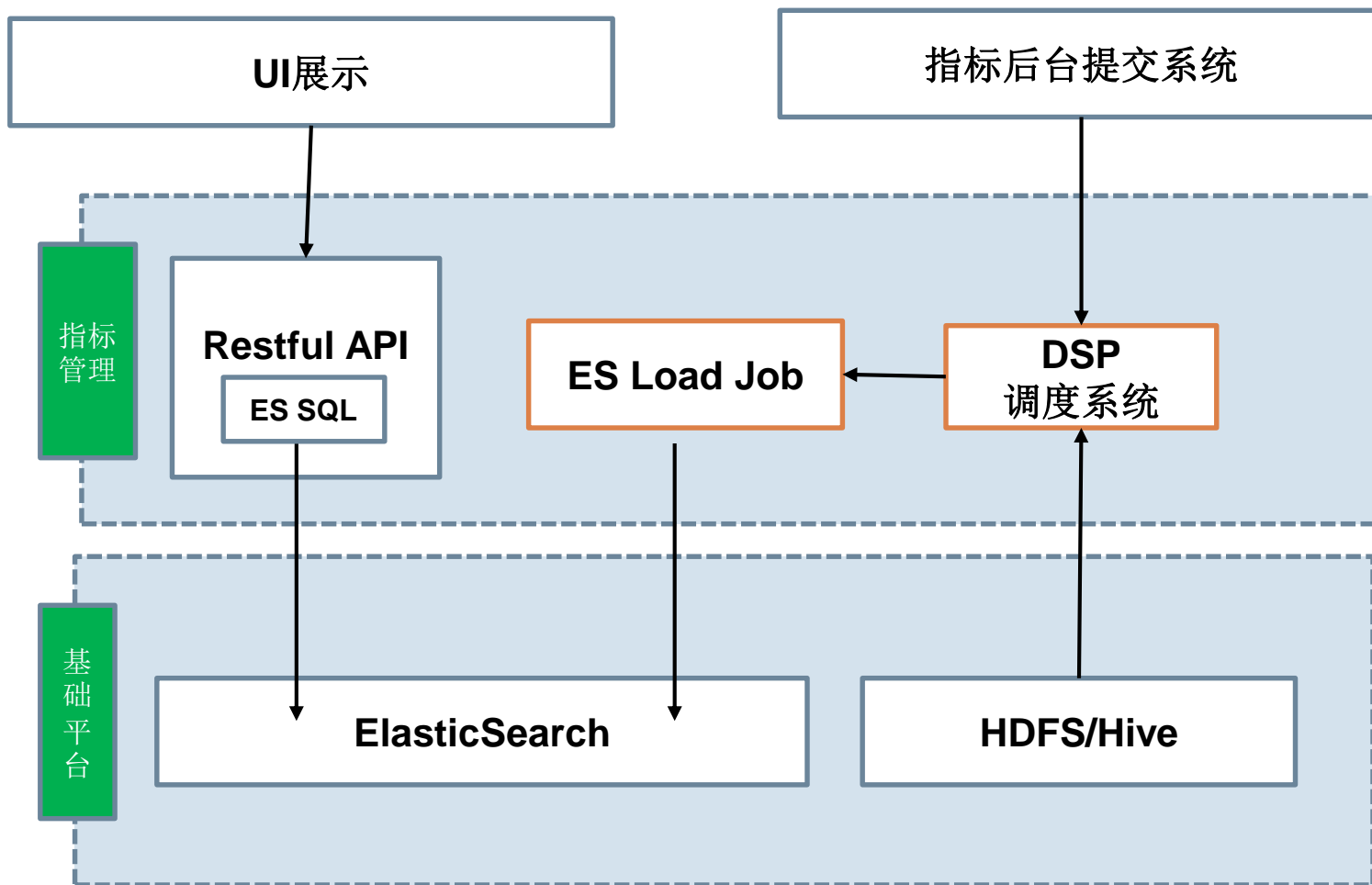
WHAT

- 给对象的各个属性，标签，度量值等

HOW

- 精准化营销
- 多维统计分析
- 千人千面

指标整体架构图



条件筛选和多维分析 (Filter & Cube)

搜索

基本筛选

出生日期

费用比率

风险等级

风险评级日期

风险属性评分

柜台开户

婚姻状况

开户日期

考核渠道

考核渠道来源

考核营业部

考核营业部区域

客户类型

客户状态

客户资产

指标

搜索模板

AUM考核

标题重命名

标准五开

客户代码

AUM(不含两融)昨日

开户日期

性别

手机号

条件搜索区域

导出excel

导出excel

清单

汇总

保存数据模板

提取数据

恢复默认查询

生产hive查询脚本

50

记录/页

显示第 1 至 50 项记录, 共 161,821 项

操作

高级

开户日期	资产分层	性别	渠道归属	AUM(不含两融)昨日	风险等级	AUM(含两融)昨日
汇总	汇总	汇总	汇总	汇总	汇总	汇总
20160919	大众客户 (5万以下)	男性	O2O_线下渠道		保守型	
20160919	大众客户 (5万以下)	男性	O2O_线下渠道		稳健型	
20160919	大众客户 (5万以下)	男性	O2O_线下渠道		积极型	
20160919	大众客户 (5万以下)	男性	O2O_线下渠道		进取型	
20160919	大众客户 (5万以下)	男性	O2O_线下渠道		安全型	

指标系统元数据定义

- 指标模型管理
- 指标管理
- 维度管理
- 其他辅助设置

指标系统

模型管理

维度管理

指标管理

类目管理

索引管理

高维度指标管理

别名管理

指标列表

客户总资产汇总

指标名称

查询

新增表达式指标

新增行为指标

新增指标

指标字段	指标标题	类目	类型	指标类型	IT口径	负责人	JIRA号	是否启用	操作
omm_inout_amt	理财净转入（净入金）	资产动态	double	普通				启用	禁用 编辑 删除 授权
omm_out_amt	理财总转出（出金）	资产动态	double	普通				启用	禁用 编辑 删除 授权
cash_transfer_inout_amt	保证金总净转入（入金）	资产动态	double	普通				启用	禁用 编辑 删除 授权
cash_inout_amt	总净转入（入金）	资产动态	double	普通				启用	禁用 编辑 删除 授权
cash_out_amt	总转出（出金）	资产动态	double	普通				启用	禁用 编辑 删除 授权
cash_in_amt	总转入（入金）	资产动态	double	普通				启用	禁用 编辑 删除 授权
omm_in_amt	理财总转入（入金）	资产动态	double	普通				启用	禁用 编辑 删除 授权
all_aum	AUM(不含两融)	资产	double	普通				启用	禁用 编辑 删除 授权
omm_card_out_amt	理财银行卡赎回金额（出金）	资产动态	double	普通				启用	禁用 编辑 删除 授权
omm_card_in_amt	理财银行卡购买金额（入金）	资产动态	double	普通				启用	禁用 编辑 删除 授权

<<

1

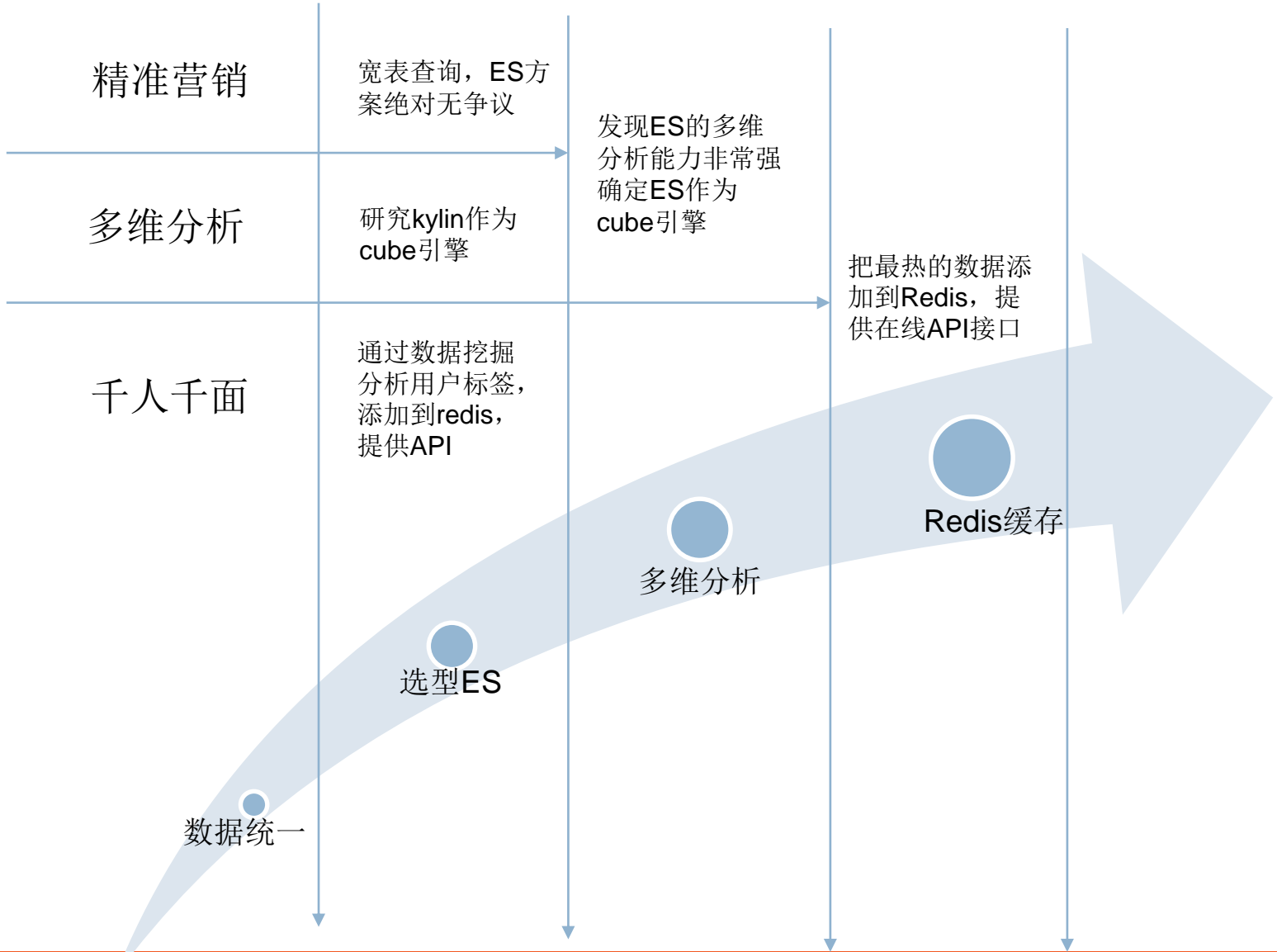
2

3

4

>>

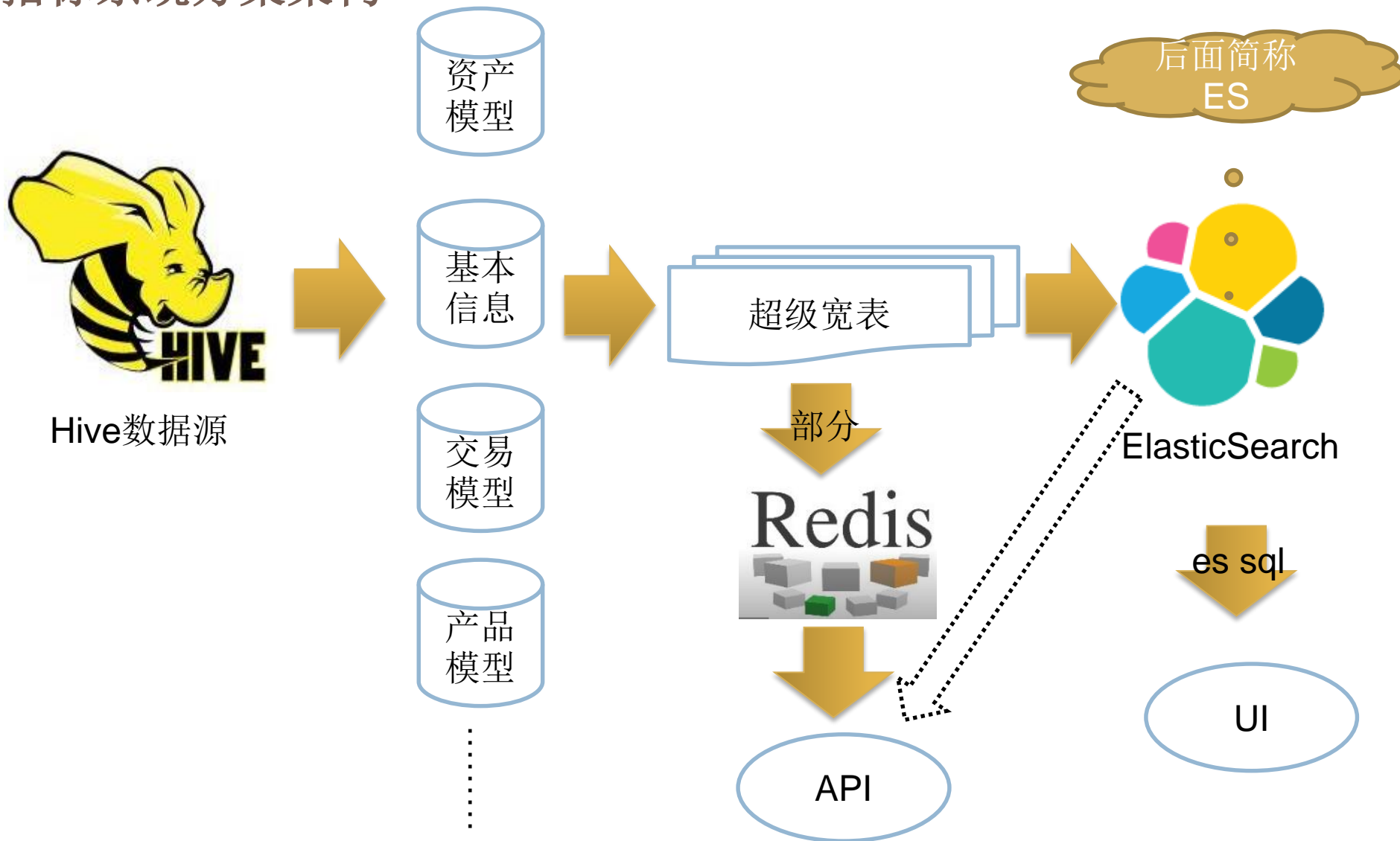
开发实践 ---- 摸石头过河



系统技术架构

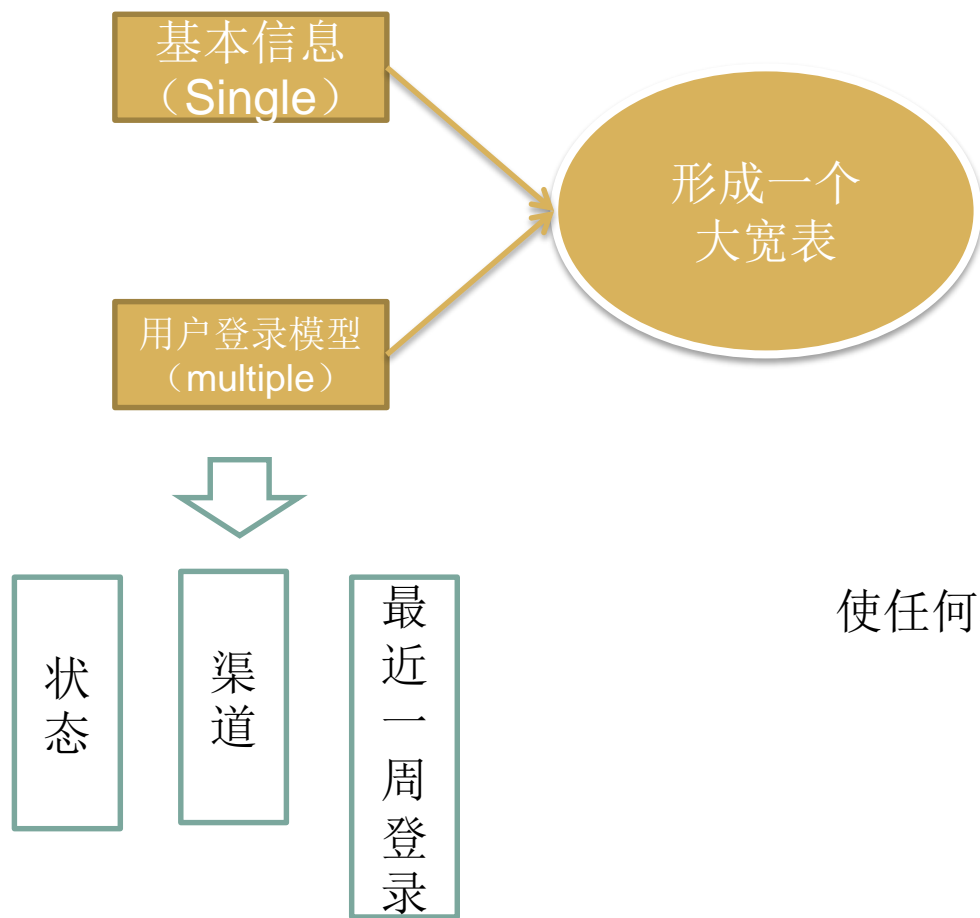
- 整体架构
- 指标元数据设计
 - 模型设计
 - 指标设计

指标系统方案架构



指标元数据设计---- 模型设计

局限性：大宽表，每个用户只能一条记录



使任何模型，每个对象都变为一条记录

指标元数据设计---- 模型维度设计

- 模型设计

*模型名称

USER_LOGIN

*模型标题

用户登录

模型描述

用户登录

*Hive脚本

select cust_code,channel login_channel,APP_VERSION
LOGIN_APP_VERSION,COALESCE(if(login_time='null' or
isnull(login_time),null,login_time),dt) as login_time,cast(1
BASE.UDS_B_I_USER_LOGIN where dt=\${start|yyyyMMdc

☐ 是否客户唯一主键

☐ 是否跳过非交易日

☒ 是否启用

*计算类型

☐ 手动计算 ☒ 日计算 ☐ 周计算 ☐ 月计算

预览

返回

- 维度设计

*字段

用户登录

login_channel

*名称

login_channel

*标题

登录渠道

*类型

string

描述

*字典类型

LOGIN_CHNL

保存

返回

指标元数据设计---- 指标设计

- 指标设计

*字段

用户登录

login_count

*名称

login_count

*标题

登陆次数

*数据类型

double

☒ 是否需要数字聚合

描述

*类目

行为

新增类目

*敏感等级

☐ 高敏 ☐ 中敏 ☒ 低敏 ☐ 不敏感

选择扩展维度

选择维度

增加

已选扩展维度

登录渠道

Sum ☒ 全选

☒ 自然年 ☒ 自然半年 ☒ 自然季度 ☒ 自然月 ☒ 自然周

☒ 最近一年 ☒ 最近半年 ☒ 最近一季度 ☒ 最近一个月 ☒ 最近一周

Avg ☐ 全选

☐ 自然年 ☐ 自然半年 ☐ 自然季度 ☐ 自然月 ☐ 自然周

☐ 最近一年 ☐ 最近半年 ☐ 最近一季度 ☐ 最近一个月 ☐ 最近一周

TAvg ☐ 全选

☐ 自然年 ☐ 自然半年 ☐ 自然季度 ☐ 自然月 ☐ 自然周

☐ 最近一年 ☐ 最近半年 ☐ 最近一季度 ☐ 最近一个月 ☐ 最近一周

☐ 最近5个交易日 ☐ 最近10个交易日 ☐ 最近20个交易日

- 扩展指标
 - **Hive**里面是**Map**字段
 - 设置维度后，每个渠道一个**key**
 - 存放在**Map**，**ES**是一个字段
- 添加指标时
可以选择多个字段

— 登陆次数

+

— 登录渠道

+

登录渠道 指标选择

×

登录渠道：

APP-APP

▼

扩展指标

APP-APP

#other-其他

PA18-PA18

I-通达信

MACS-MACS

OMM-OMM

REST-REST

添加

扩展指标

昨日

▼

昨日

自然年（累计值）

自然半年（累计值）

自然季度（累计值）

自然月（累计值）

最近一周（累计值）

最近一年（累计值）

最近半年（累计值）

最近一季（累计值）

最近一个月（累计值）

自然周（累计值）

添加

指标元数据设计----其他设置

➤ 表达式指标

- 用户可以根据自己的逻辑添加指标

➤ 模板定义

- 根据设计的查询分享给其他用户

➤ 索引设置

- 索引是有限资源

➤ 指标别名设置

多项聚合

Use group by (fieldName),(fieldName, fieldName)

```
SELECT * FROM account GROUP BY (gender, state,  
age),(state),(age)
```

分区间聚合

put fieldName followed by your ranges

```
SELECT COUNT(age) FROM bank GROUP BY range(age,  
20,25,30,35,40)
```

地理查询

```
GEO_DISTANCE_RANGE(center,'1m','1km',100.5,0.50001)
```

```
GEO_DISTANCE(center,'1km',100.5,0.5)
```

优化和填坑

➤ Elasticsearch优化和填坑

- Elasticsearch优化点
- 填坑大行动

➤ 核心计算调度

ElasticSearch原理

- ✓ 时间复杂度: $O(N) \rightarrow O(\log N) \rightarrow O(1)$
- ✓ 文档(Document) 单词(Word) 倒排索引(Inverted Index)
- ✓ Bool filter & And/Or/Not filters(Bitset & non-bitset)
- ✓ Filter execution order
- ✓ Instance & Shard

	doc1	doc2	doc3	doc4	doc5	doc6	doc7	doc8
gender-男		✓		✓		✓		
Level-高净值	✓	✓	✓		✓	✓	✓	✓

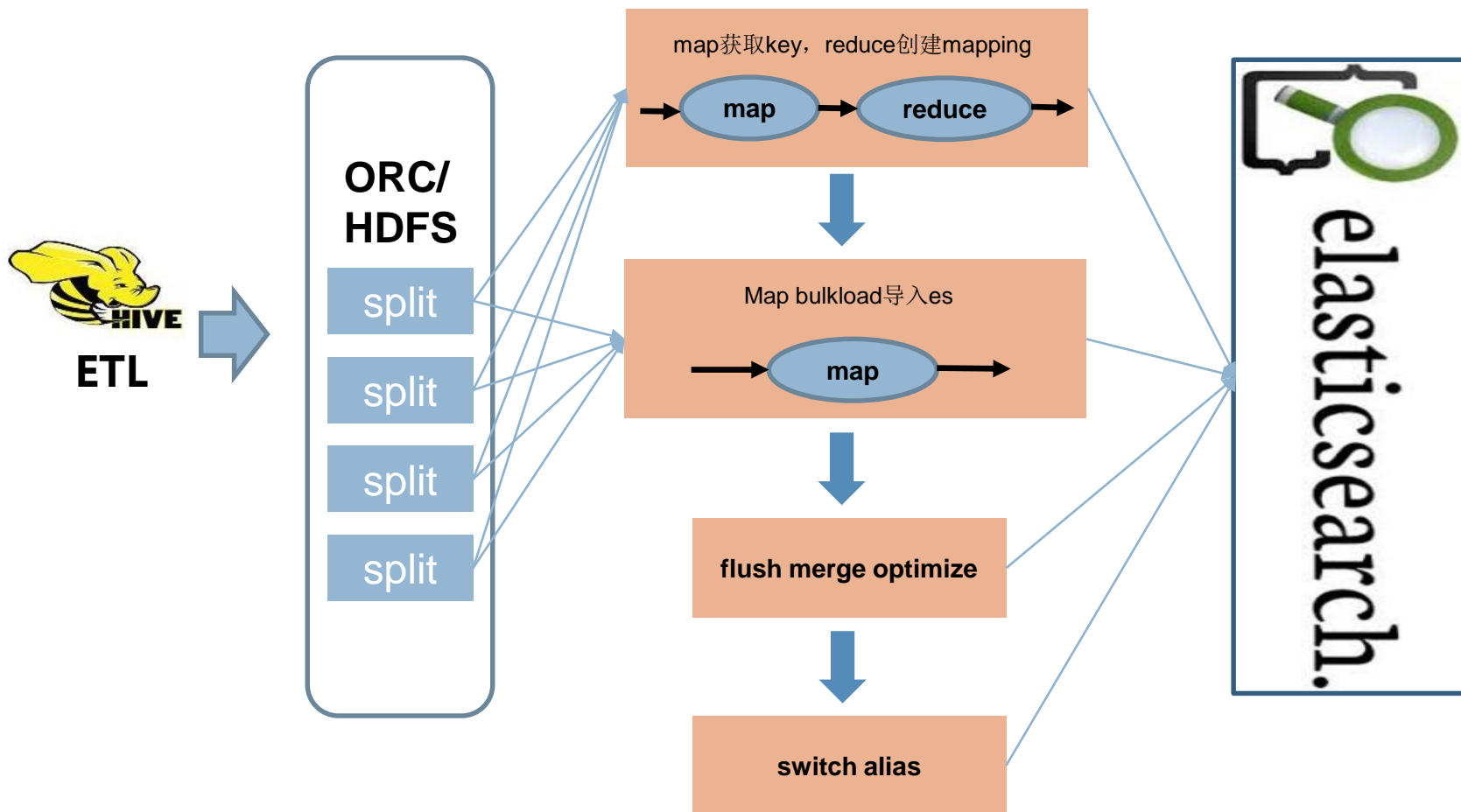


bitset - and		✓				✓		
--------------	--	---	--	--	--	---	--	--

ElasticSearch--blukload

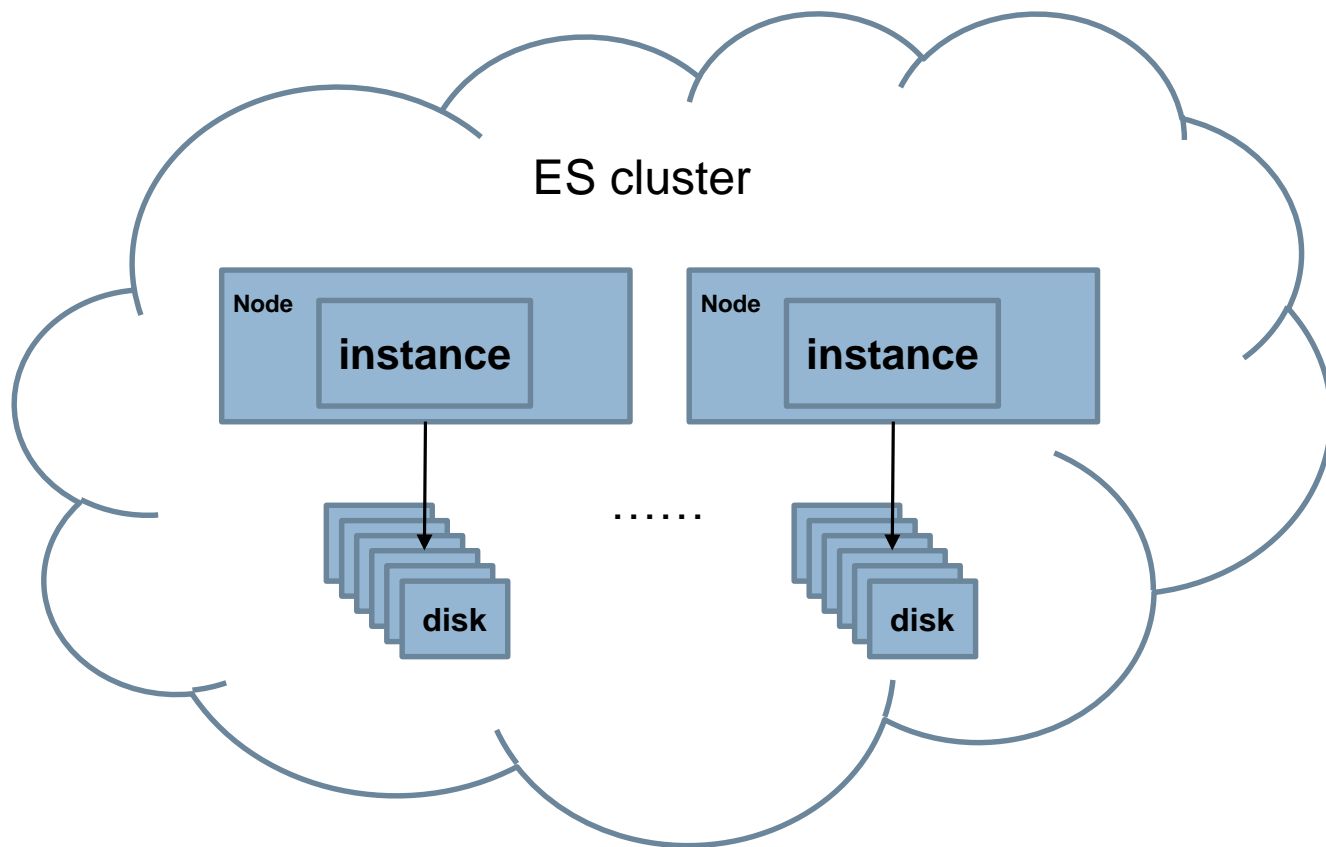
11万字段，两万个索引字段，每天要7点前导入es，核心将ORCFile里的数据导入ES

虽然ES可以自动创建mapping，但创建的效率非常低，所以希望能一次性的把mapping创建



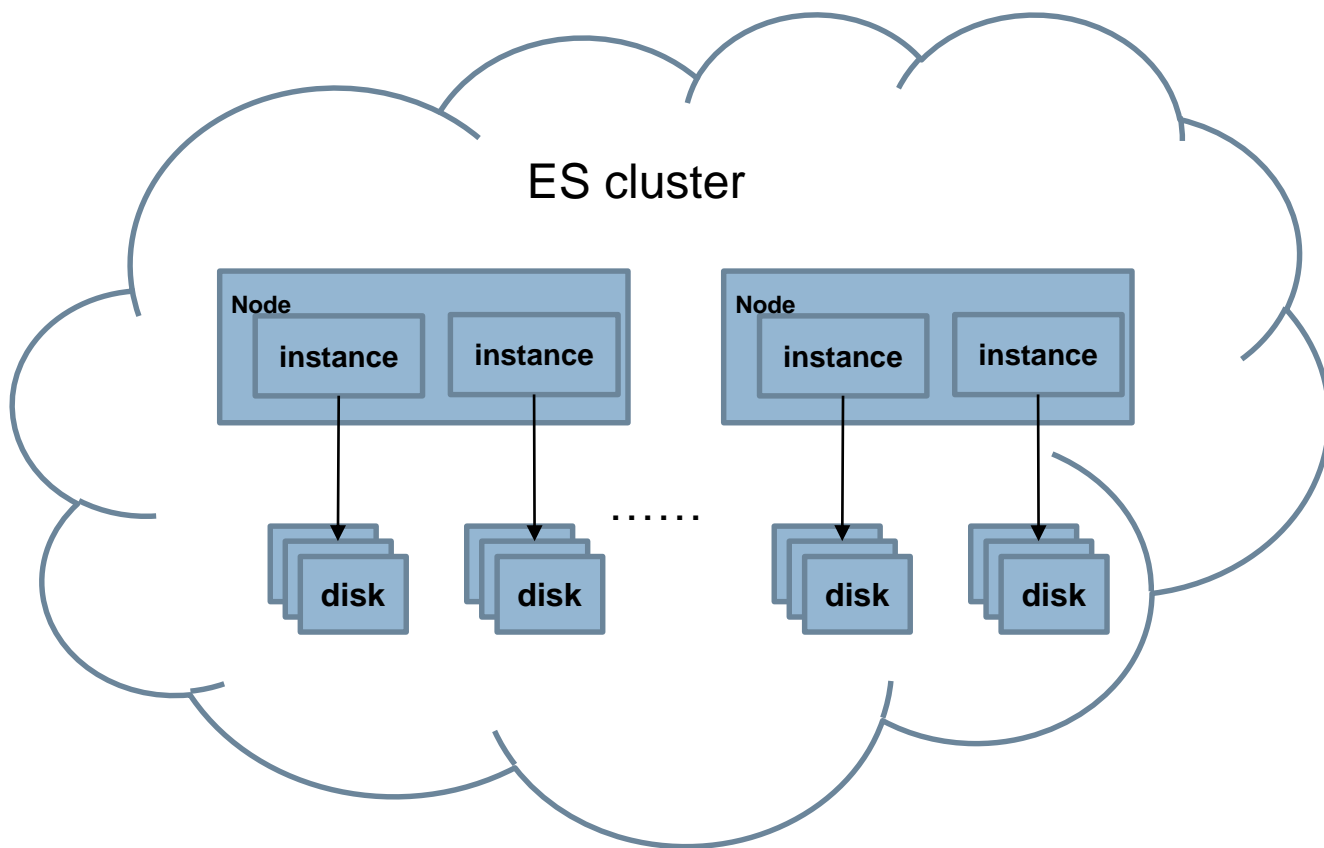
ElasticSearch优化—IO优化

- 一台机器多个实例，硬盘裸盘
- 多硬盘提高IO，实例之间硬盘不共享
- SHARD个数等于硬盘个数，最大化利用硬盘IO



ElasticSearch优化—IO优化

- 一台机子多个实例，硬盘为裸盘
- 多硬盘提高IO，实例之间硬盘不共享
- SHARD个数等于硬盘个数，最大化利用硬盘IO



ElasticSearch优化—IO优化

➤ Bulkload 的瓶颈在IO

- 购买32块SSD硬盘和96块普通硬盘比较
(以下数据为1000w记录, 11万字段, 1w索引, 索引500G)

	32块SSD硬盘	96块普通硬盘
Import time	80分钟	90分钟
Merge time	20分钟	30分钟
Hard Disk Load	40%	80%

➤ 结论: 时间上面没有明显提高, 但load提高一倍

- 进一步优化方法, 购买更多SSD硬盘
- 增加机器数量
- 减少索引个数, 优化索引

ElasticSearch优化

- 尝试G1 GC，提高吞吐能力
- 一台机器多个实例
- 多硬盘提高IO，实例之间硬盘不共享
- SHARD个数等于硬盘个数，最大化利用硬盘IO
- bootstrap.mlockall设置true，防止swap
- 使用bulk load，加大threadpool.bulk.queue_size=1024 避免数据丢失
 - Map=90
- Index settings disable _all
- 使用index alias代替index name

填坑：ElasticSearch TopN—慎用terms size

ElasticSearch terms size聚合的时候，如果维度基数大于size，聚合结果求TopN可能是近似值

shard A	shard B	shard C					
A 30	B 12	E 15	求top3	shard A	shard B	shard C	结果
B 25	C 10	B 10		A 30	B 12	E 15	
C 4	D 8	D 8		B 25	C 10	B 10	
D 3	E 7	C 6		C 4	D 8	D 8	
E 2	A 2	A 1					
							B 47
							A 30
							D 16



正确结果：

B	47
A	33
E	24



结论：

- terms size & shard size 必须超过维度元素个数

填坑：禁止高维组合查询 & distinct 不精确

➤ 高维查询

- 当高维组合查询时，es的heap & load 就会增高，会拒绝服务
- 高维在业务应用中其实不多，无业务含义，应该禁用（kylin 对高维支持很好）

➤ Distinct 使用 Hyper loglog 算法，有一定误差

➤ 原因:

如果terms字段为null值，group by无法进行统计显示

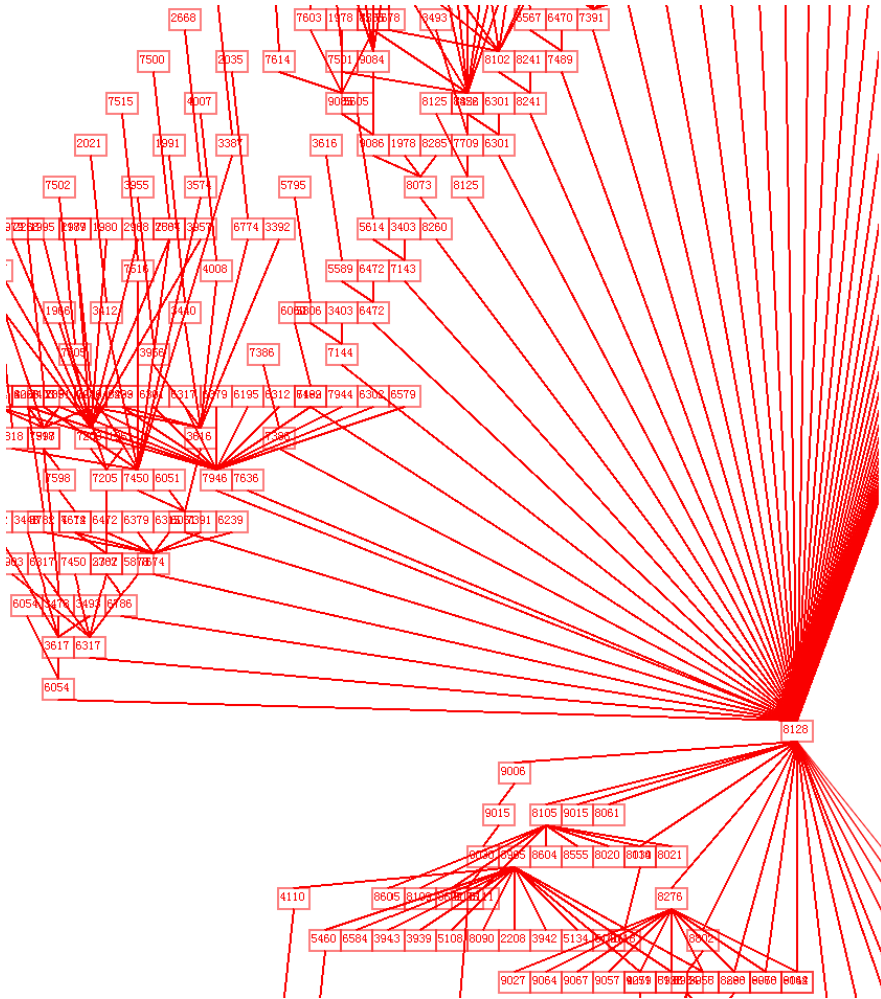
➤ ElasticSearch-sql: pull-267

对terms聚合添加missing和order by功能

未来发展

- 优化ElasticSearch bulkload，缩短导入时间
- 对象记录唯一性的缺陷，支持多记录和跨索引的Join
- 指标系统抽象化，可以任何对象；目前是客户指标系统
- 能够和多维分析UI工具打通
- 支持实时数据查询和分析
- 有望推向社区，开源

核心计算调度（主题外补充）



➤ 自动依赖

✓ 通过Sql分析，形成一个有向无环图

名称

UI_INDEX_MODULE_COLLECT2

标题

指标系统汇总job_NEW

选择分类

无分类

描述

指标系统汇总job_NEW

计算周期

☒ 月计算 ☐ 周计算 ☐ 日计算

启动时机

☐ 收市后(15:30 PM)
☐ 清算后(支持清算标志的库有 FISL,KGDB,MIS,OCRM,ODS,OTC)
☐ SQL:SELECT 1 FROM DUAL WHERE ...有数据则启动
☐ 定时: HH:mm

高级选项

☐ 删除数据: 全部(DB)
☐ 跳过条件: 14:00 - 下午14点以后跳过或SELECT 1 FROM DUAL WHERE ...有数据则跳过或NT非交易日跳过
☒ 自动检测依赖

监控策略

☐ 如果今天09:00 前还未启动, 则给我和输入十个以内的UM账号以空格分割,不要包含@发邮件
☐ 如果今天08:30 前还未完成, 则给我和输入十个以内的UM账号以空格分割,不要包含@发邮件
☐ 如果今天导出条数> 100 条, 则给我和输入十个以内的UM账号以空格分割,不要包含@发邮件
☐ 如果今天导出条数与平均数相比浮动超过30 %, 则给我和输入十个以内的UM账号以空格分割,不要包含@发邮件
☐ 如果今天运行时间与平均时间相比浮动超过30 %, 则给我和输入十个以内的UM账号以空格分割,不要包含@发邮件
☐ 如果运行这个SQL有结果[SPData(SELECT 1 from dual(SQL请不要包含逗号)], 则给我和输入十个以内的UM账号以空格分割,不要包含@发邮件

存储方式

FACT(BASE->FACT) FACT表, 是报表来源

选择数据库

HIVE

标签

ES任务

输入SQL

```
1 --ES_INDEX=custom
2 --ES_TYPE=JOB
3 --ES_PK=cust_code
4 --ES_AUTO_DT=Y
5 --HIVE_CONFIG=set
6 --reduce.map.output.compress.codec=org.apache.hadoop.io.compress.SnappyCodec
7 --ES_SKIP_AUTO=DSPMT|SELECT 1 FROM DUAL WHERE ${setart|yyyyMMdd}|=${now-1day|yyyyMMdd}
8 with module_TRADE_MOVVT as (
9 select *
10 from fact.ui_index_module_TRADE_MOVVT
```

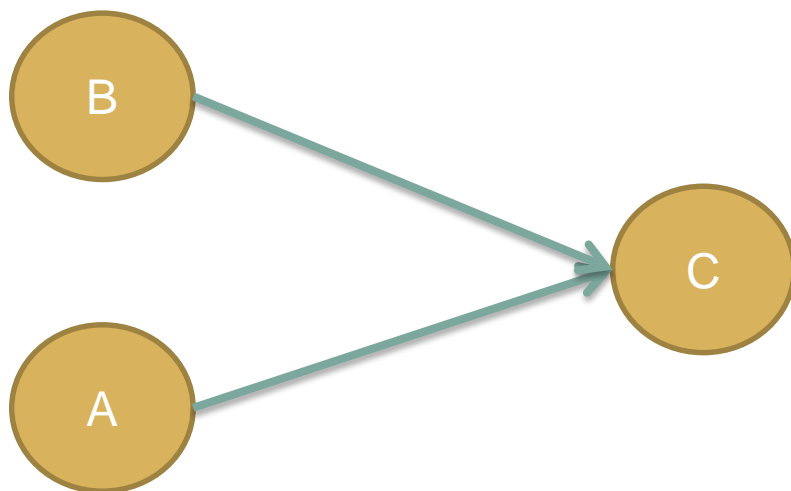

核心计算调度----sql dependent

- 平安95% 的job使用 sql 表示
- Sql 语义分析Job 依赖

A: select f1 from table1

B: select f2 from table2

C: select a.f1 + b.f2 from a join b



谢谢

