clickhouse数仓应用实践

演讲人：朱元

日期： 2019-10-20

目录
CONTENTS

01

现状背景

现状

即席查询性能差

数据链路长

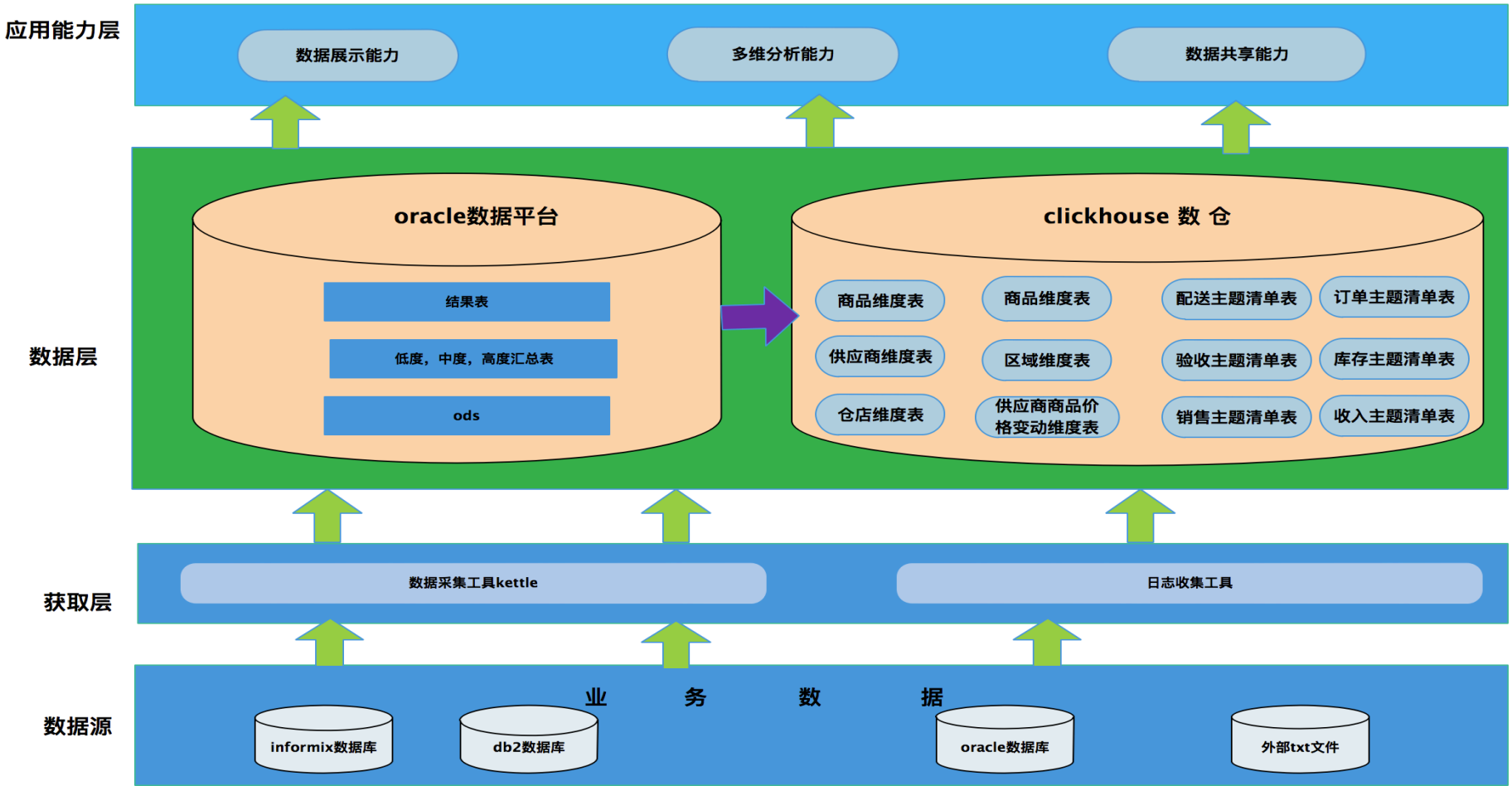
数据压缩率低

需求响应慢

02

应 用 实 践

数据架构

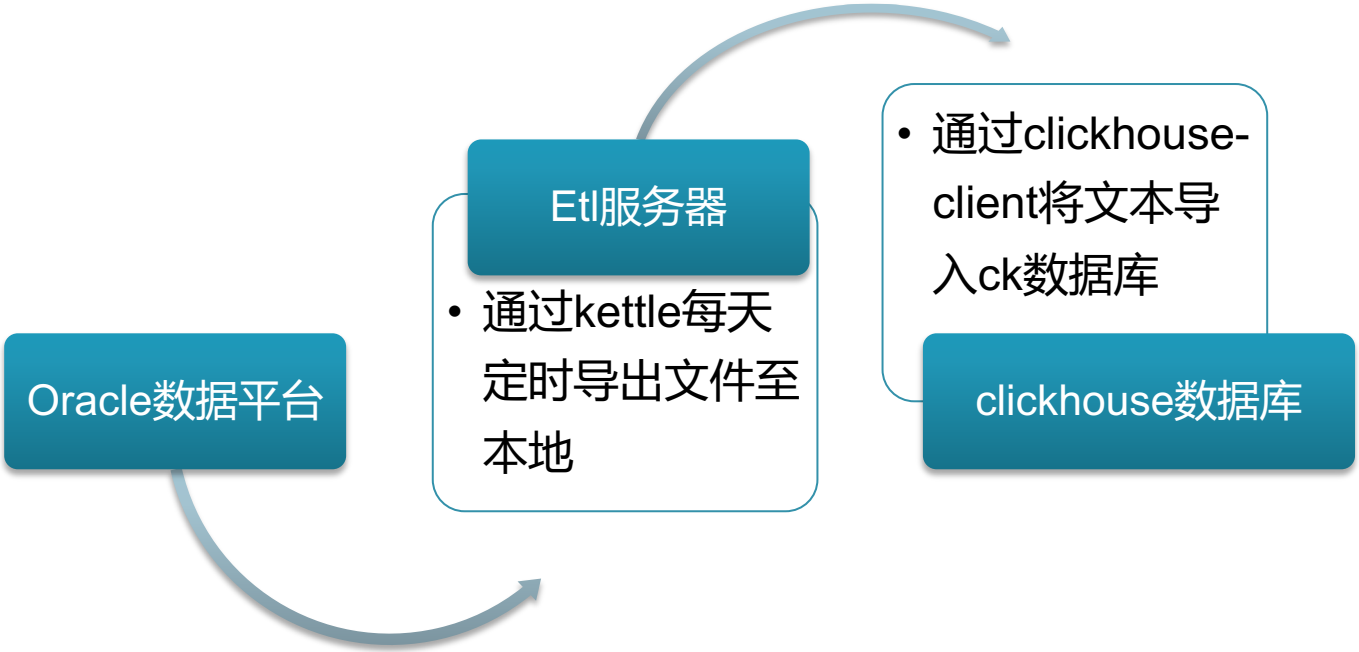

应用能力层
数据展示能力
多维分析能力
数据共享能力
数据层
oracle数据平台
结果表
低度，中度，高度汇总表
ods
clickhouse 数 仓
商品维度表
商品维度表
配送主题清单表
订单主题清单表
供应商维度表
区域维度表
验收主题清单表
库存主题清单表
仓店维度表
供应商商品价格变动维度表
销售主题清单表
收入主题清单表
获取层
数据采集工具kettle
日志收集工具
数据源
业 务 数 据
informix数据库
db2数据库
oracle数据库
外部txt文件

数据同步ck

1，基于公司对数据要求为T+1

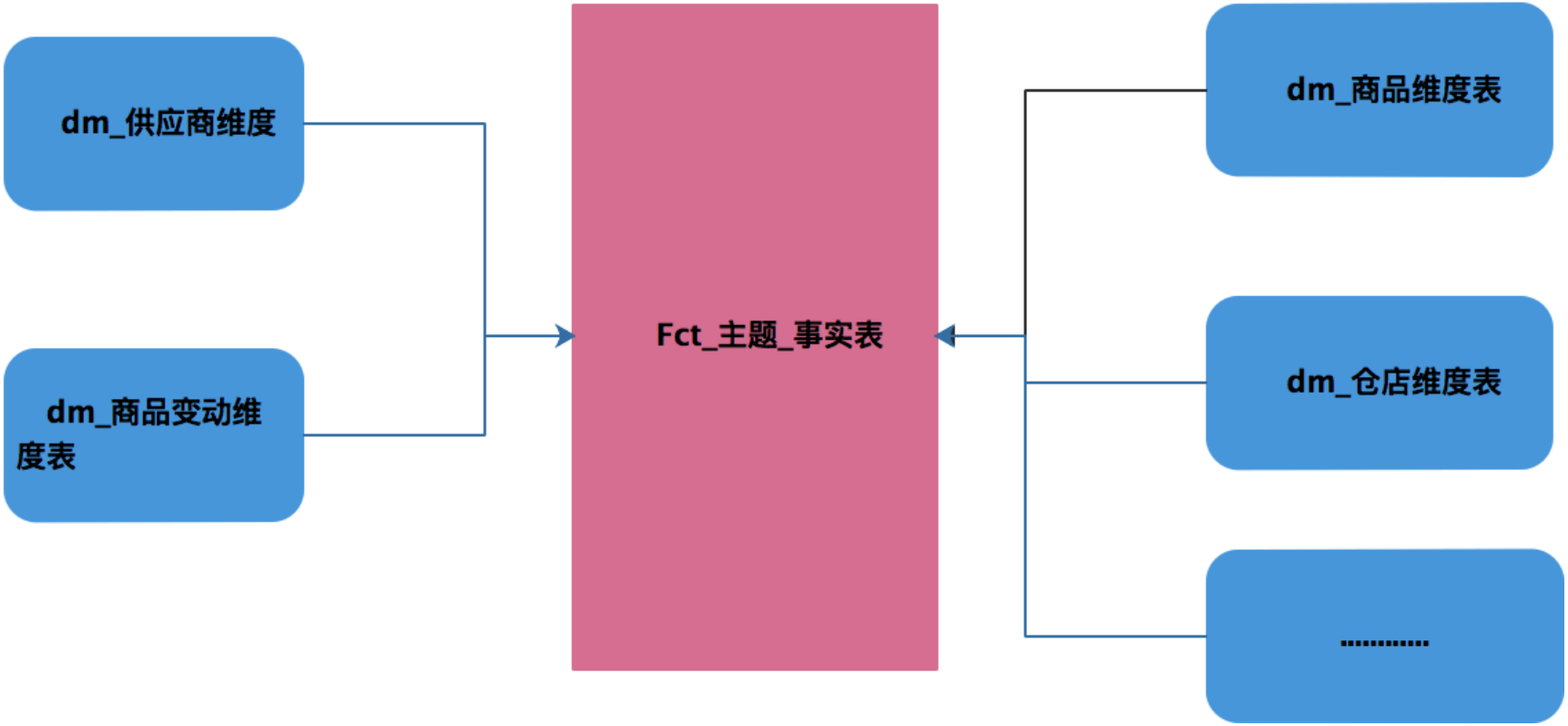
2. 基于现有开发人员水平及成本

因此采用可视化同步工具kettle. 先将oracle数据平台维度信息以及相关主题清单数据同步至clichouse数据仓库



数仓建设

ck数仓数据模型采用星型模型搭建



数仓建设–维度表

一般维度表数据量不大. 目前采用的是引擎Log+字典表(dictionary)

```
create table dw_hr.dw_shop_info
(
  shopid String,
  shopcode String,
  shopname String,
  address String,
  buid UInt8,
  inbuid UInt8,
  shopformid UInt8,
  shofortype UInt8,
  shopstatus UInt8,
  buname String,
  categorytreeid UInt8,
  payshopid String,
  regionid UInt64,
  provinces_regions String,
  propertype String,
  dctype UInt8,
  prov_name String,
  city_name String
)
engine = Log
```

```
<dictionary>
 <name>dw_shop</name>
 <source>
  <clickhouse>
   <host>192.168.1.103</host>
   <port>9000</port>
   <user>default</user>
   <password>1q2w3e</password>
   <db>dw_hr</db>
   <table>dw_shop_info</table>
  </clickhouse>
 </source>
 <lifetime>
  <min>200</min>
  <max>260</max>
 </lifetime>
 <layout>
   <complex_key_hashed/>
 </layout>
 <structure>
  <key>
    <attribute>
     <name>shopid</name>
     <type>String</type>
   </attribute>
  </key>
  <attribute>
    <name>buid</name>
    <type>UInt8</type>
    <null_value>0</null_value>
  </attribute>
  <attribute>
    <name>buname</name>
    <type>String</type>
    <null_value>未知</null_value>
  </attribute>
  <attribute>
    <name>categorytreeid</name>
    <type>UInt8</type>
    <null_value>0</null_value>
  </attribute>
  <attribute>
    <name>shopname</name>
```

数仓建设–主题事实清单表

主题事实清单表采用引擎MergeTree. 同步策略: 每日从oracle数据平台增量同步到ck数仓.

```
create table dw_hr.fct_rpt_dc_shop_vender_day
(
  stat_year UInt16,
  stat_month UInt32,
  stat_day Date,
  stat_day_str String default formatDateTime(stat_day, '%F'),
  buid UInt8,
  dc_id String,
  venderid String,
  dctype UInt8,
  shop_id String,
  shopformid UInt8,
  logistics UInt8,
  datasource UInt8,
  rpt_qty Decimal(18,4),
  rpt_boxes Decimal(18,4),
  rpt_cost Decimal(18,4),

)
engine = MergeTree PARTITION BY toYYYYMM(stat_day) ORDER BY (stat_day, dc_id) SETTINGS index_granularity = 8192
;
```

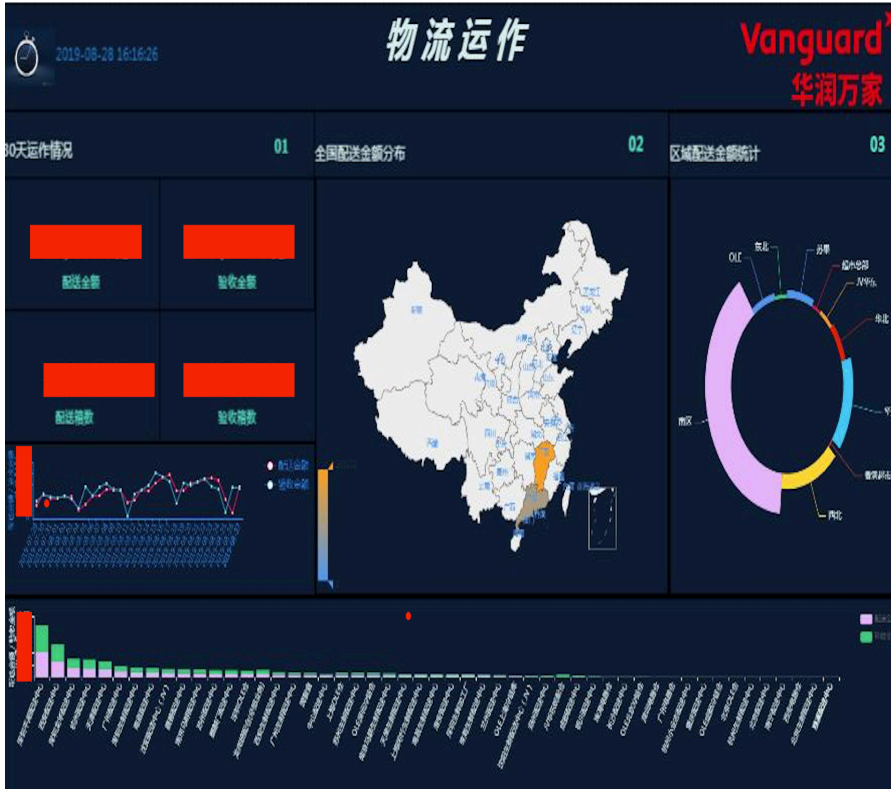
数仓建设–对外数据

目前对外开放是主题事实清单表+维度表 封装成一个视图,类似如下

```
create or replace view vw_fct_rpt_dc_shop_sku_vender_day as
select
  stat_year ,
  stat_month ,
  stat_day ,
  stat_day_str ,
  buid ,
  dictGet('dw_buinfo', 'buname', toUInt64(buid)) AS buid_name,
  dc_id,
  dictGet('dw_shop', 'shopname', tuple(dc_id))    AS dc_shop_name,
  dictGet('dw_shop', 'provinces_regions', tuple(dc_id))  AS dc_shop_region,
  dictGet('dw_shop', 'prov_name', tuple(dc_id))  AS dc_shop_prov,
  dictGet('dw_shop', 'city_name', tuple(dc_id)) as dc_shop_city,
  dctype ,
  shop_id ,
  dictGet('dw_shop', 'shopname', tuple(shop_id))    AS order_shop_name,
  dictGet('dw_shop', 'provinces_regions', tuple(shop_id))  AS order_shop_region,
  dictGet('dw_shop', 'prov_name', tuple(shop_id))  AS order_shop_prov,
  dictGet('dw_shop', 'city_name', tuple(shop_id)) as order_shop_city,
```

数据 展示+多维分析

采用开源报表系统davinci 地址: https://github.com/edp963/davinci



03

实践遇到的问题

1. Memory limit (for query) exceeded

解决：通过在users.xml 配置
max_bytes_before_external_sort
max_bytes_before_external_group_by

2. 用户并发量一上来,负载太高

解决：目前是在中间加redis缓存