美团网
meituan.com

O2O数据仓库实践

应对快速变化的业务模型

xuzhang@meituan.com

目录

- 美团数据业务现状和要求
- 演进中出现的问题
- 统一模型管理方案和工具
- 经验总结

美团的数据业务场景

- 团购行业低毛利,要求高效率
- 业务复杂. 订单、财务、供应链、客服等近20个业务部门,关系复杂需求多
- 业务变化快. 行业高速发展,形式不断更新

数据业务规模 (截至2014.4)

美团网 meituan.com

- 每日新增数据量: 5T(压缩后)
- ETL任务: 1800
- 每日Hadoop Job 数: 20000+
- 报表数量: 985
- 支撑RD和分析师: 460+

对数据仓库的要求

美团网 meituan.com

- 快速建模: 全面地覆盖所有业务
- 业务模型灵活变更: 可行且响应周期短
- 方便使用: 每个人都能使用数据资源

目录

- 美团数据业务现状和要求
- 演进中出现的问题
- 统一模型管理方案和工具
- 经验总结

开放数据生产

业务需求庞大,
数据团队疲于应付数据接入和建模工作

↓

构建数据开放平台,业务方RD自行生产数据

↓

数据团队人员Review ETL

快速粗放式增长之痛

美团网
meituan.com

- 指标管理混乱: 重复定义, 口径不一致, 变更困难
- 元数据过于松散: 找不到数据, 业务知识零散

快速粗放式增长之痛

美团网
meituan.com

- 依赖关系过于复杂: 层次过深, 指标来源不明
- 规范执行困难: 人工审核,建模方法论难以贯彻,精力耗费在业务无关的细节上

目录

- 美团数据业务现状和要求
- 演进中出现的问题
- 统一模型管理方案和工具
- 经验总结

基本思路

美团网 meituan.com

- 形式化模型定义
- 简化处理层级
- 模型驱动数据生产

形式化模型定义

美团网
meituan.com

- 精确表达业务逻辑
- 一处定义,各处复用
- 模型即文档

总线图方法

美团网 meituan.com

	浏览	注册验证	登录	下单	支付	消费	评价	退款	...
日期	X	X	X	X	X	X	X	X	
城市			X	X	X	X	X		
团购项目				X	X	X	X	X	
商家门店						X	X		
用户		X	X	X	X	X	X	X	
流量来源	X	X	X	X	X				
...									

多维模型视角

- 按事实角度(纵向): 星型多维
- 按主题组织(横向): 维度层级组合合并

模型元素

- 实体/事实: 名称,物理表名,属性/度量列表
- 属性: 名称,对应字段/表达式,属性字典
- 度量: 聚合方法
- 层级关系: 各层级属性

模型元素:实体/事实

```
<Entity name="User" caption="用户" primaryAttribute="User" creationalFact="Signup">
  <Table name="user" schema="dim"></Table>
  <Attribute name="User" caption="用户" column="userid" dataType="Integer"></Attribute>
  <Attribute name="Email" caption="用户邮箱" column="email" dataType="String"></Attribute>
  …
</Entity>
```

```
<Fact name="Feedback" caption="评价">
  <Table name="feedback" schema="fact"></Table>
  <EntityAttribute name="Deal"  caption="项目" column="deal_id" entity="Deal"></
EntityAttribute>
  …
  <Measure name="Feedback User Count" caption="评价用户数" column="user_id"
aggregator="COUNT-DISTINCT" dataType="Integer"></Measure>
  …
</Fact>
```

模型元素:属性

美团网 meituan.com

```
<Attribute name="Coupon Type" caption="团购券类型" column="coupontype"
nameColumn="coupontypename" dataType="Integer">
  <Dictionary>
    <AttributeItem value="0" name="NORMAL" caption="美团券"></AttributeItem>
    <AttributeItem value="1" name="DELIVERY" caption="物流"></AttributeItem>
    <AttributeItem value="2" name="OTHERS" caption="第三方优惠码"></AttributeItem>
    <AttributeItem value="3" name="WDMSMS" caption="翼码二维码"></AttributeItem>
    <AttributeItem value="4" name="LOTTERY" caption="抽奖单"></AttributeItem>
    <AttributeItem value="5" name="YGLZMMS" caption="阳光绿洲二维码"></AttributeItem>
  </Dictionary>
</Attribute>


<Attribute name="Is Open" caption="是否是已上线项目" dataType="Integer">
  <Expression>
    <![CDATA[
      IF($Attributes('Status') IN (32, 64) AND $Attributes('Deal Attribute') & 8 = 0, 1, 0)
    ]]>
  </Expression>
</Attribute>
```

模型元素:度量值

<Measure name="Revenue" caption="交易额" column="amount" aggregator="SUM" dataType=“Decimal”></Measure>

```
<Measure name="Gross Profit" caption="毛收入" aggregator="SUM" dataType="Decimal">
  <Expression>
    <![CDATA[ $Measures('Revenue')- $Measures('Order Buy Price')]]>
  </Expression>
</Measure>
```

模型元素:层级关系

美团网 meituan.com

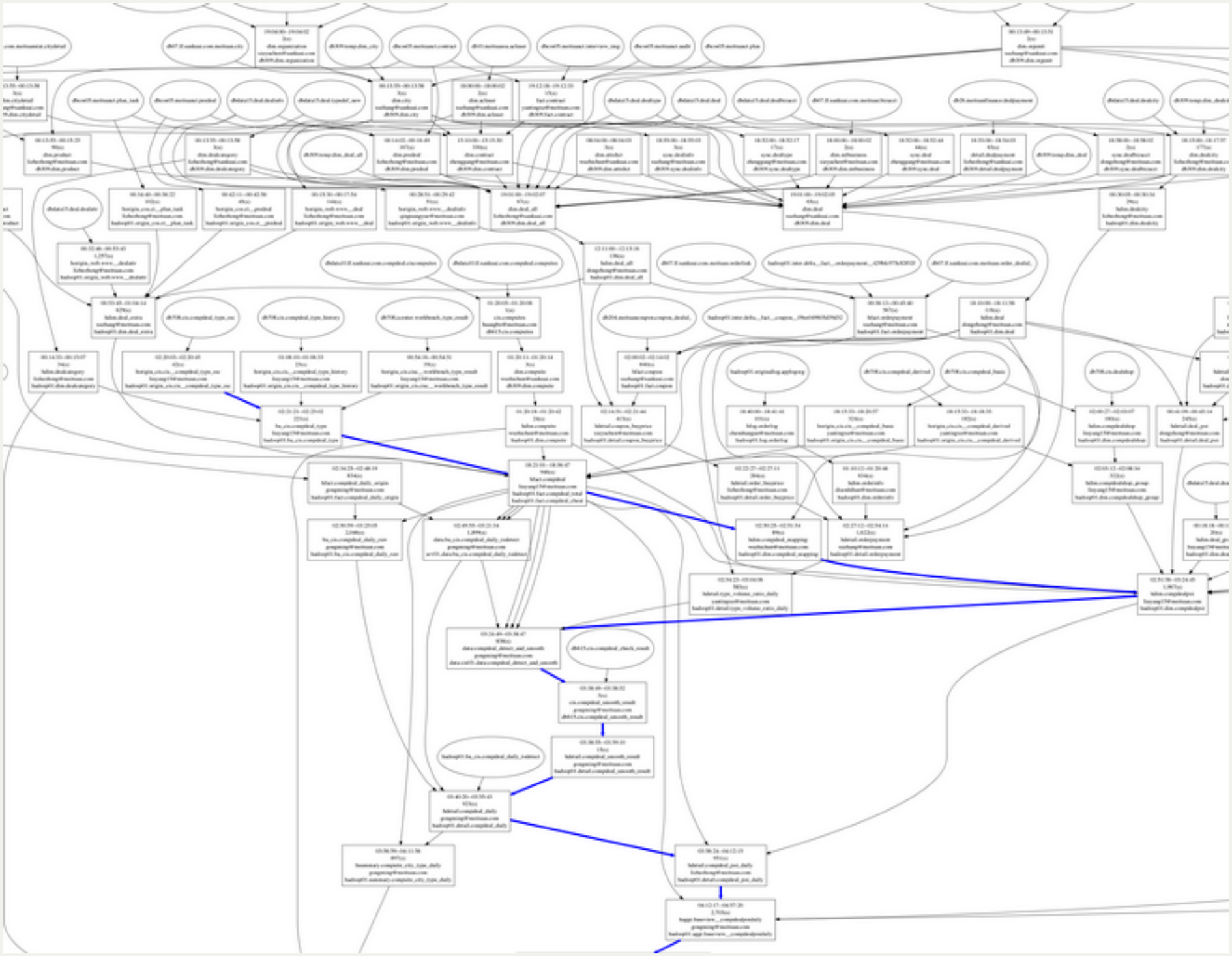
```
<Hierarchy name="Category" caption="品类" allMemberName="All Categories">
  <Level attribute="Deal"></Level>
  <Level attribute="Type"></Level>
  <Level attribute="Category"></Level>
  <Level attribute="Class"></Level>
</Hierarchy>

<Hierarchy name="Calendar" caption="自然时间周期" allMemberName="All Periods">
  <Level attribute="Day"></Level>
  <Level attribute="Month"></Level>
  <Level attribute="Quarter"></Level>
  <Level attribute="Year"></Level>
</Hierarchy>

<Hierarchy name="City Rank" Caption="城市级别" allMemberName="All City Rank">
  <Level attribute="Organization Unit"></Level>
  <Level attribute="City"></Level>
  <Level attribute="City Rank"></Level>
</Hierarchy>
```

```
▶<Attribute name="Deal" caption="项目" column="dealid" nameColumn="title" dataType="Integer">...</Attribute>
 <EntityAttribute name="DealPayment" caption="项目支付方式" entity="DealPayment" column="dealid" dataType="Integer" partial="True"/>
 <EntityAttribute name="Contract City" caption="项目签单城市" entity="City" column="cityid" nameColumn="cityname" dataType="Integer"/>
 <EntityAttribute name="Organization Unit" caption="组织单元" entity="Organization Unit" column="org_unit_id" dataType="Integer"/>
 <EntityAttribute name="Begin Date" caption="项目上线日期" entity="Date" column="begindate" referencedColumn="Date" dataType="Date"/>
 <Attribute name="BD" caption="签单BD" column="bdid" dataType="Integer"/>
 <EntityAttribute name="End Date" caption="项目下线日期" entity="Date" column="enddate" referencedColumn="Date" dataType="Date"/>
 <Attribute name="Buy Price" caption="项目进价" column="buyprice" dataType="Decimal"/>
 <Attribute name="Price" caption="项目售价" column="price" dataType="Decimal"/>
 <Attribute name="Coupon Period" caption="美团券有效期" description="Coupon Validity Period" column="couponperiod" dataType="Integer"/>
 <Attribute name="Adfee" caption="广告费" column="adfee" dataType="Decimal"/>
 <Attribute name="Type" caption="三级品类" column="typeid" nameColumn="typename" dataType="Integer"/>
 <Attribute name="Category" caption="二级品类" column="categoryid" nameColumn="categoryname" dataType="Integer"/>
 <Attribute name="Class" caption="一级品类" column="classid" nameColumn="classname" dataType="Integer"/>
 <Attribute name="Contract" caption="合同" column="contract_id" nameColumn="contract_id" dataType="String"/>
 <Attribute name="Main Contract" caption="主合同" column="main_contract_id" nameColumn="main_contract_id" dataType="String"/>
▼<Attribute name="Coupon Type" caption="团购券类型" column="coupontype" nameColumn="coupontypename" dataType="Integer">
  ▼<Dictionary>
     <AttributeItem value="0" name="NORMAL" caption="美团券"/>
     <AttributeItem value="1" name="DELIVERY" caption="物流"/>
     <AttributeItem value="2" name="OTHERS" caption="第三方优惠码"/>
     <AttributeItem value="3" name="WDMSMS" caption="翼码二维码"/>
     <AttributeItem value="4" name="LOTTERY" caption="抽奖单"/>
     <AttributeItem value="5" name="YGLZMMS" caption="阳光绿洲二维码"/>
   </Dictionary>
  ▶<!--...-->
 </Attribute>
 <Attribute name="Status" caption="项目状态" column="status" dataType="Integer"/>
 <Attribute name="Deal Attribute" caption="项目属性" column="attributekey" dataType="BigInteger"/>
▼<Attribute name="Is Open" caption="是否是已上线项目" column="status" dataType="Integer">
  ▼<Expression>
    ▼<![CDATA[
       IF($Attributes('Status') IN (32, 64) AND $Attributes('Deal Attribute') & 8 = 0, 1, 0)
     ]]>
   </Expression>
 </Attribute>
▼<Attribute name="Is Normal Coupon Type" caption="是否是美团券类型" dataType="Boolean">
  ▼<Expression>
     <![CDATA[ IF($Attributes('Coupon Type') = 0, 1, 0) ]]>
   </Expression>
 </Attribute>
 <Attribute name="Is Local Service" caption="本地服务" column="islocalservice" dataType="Boolean"/>
 <Attribute name="Is Multi City" caption="多城市" column="ismulticity" dataType="Boolean"/>
 <Attribute name="Is All City" caption="全国单" column="isallcity" dataType="Boolean"/>
 <Attribute name="Is Money Card" caption="是否代金券" column="ismoneycard" dataType="Boolean"/>
 <Attribute name="Coupon Begin Time" caption="代金券开始时间" column="couponbegintime" dataType="String"/>
```

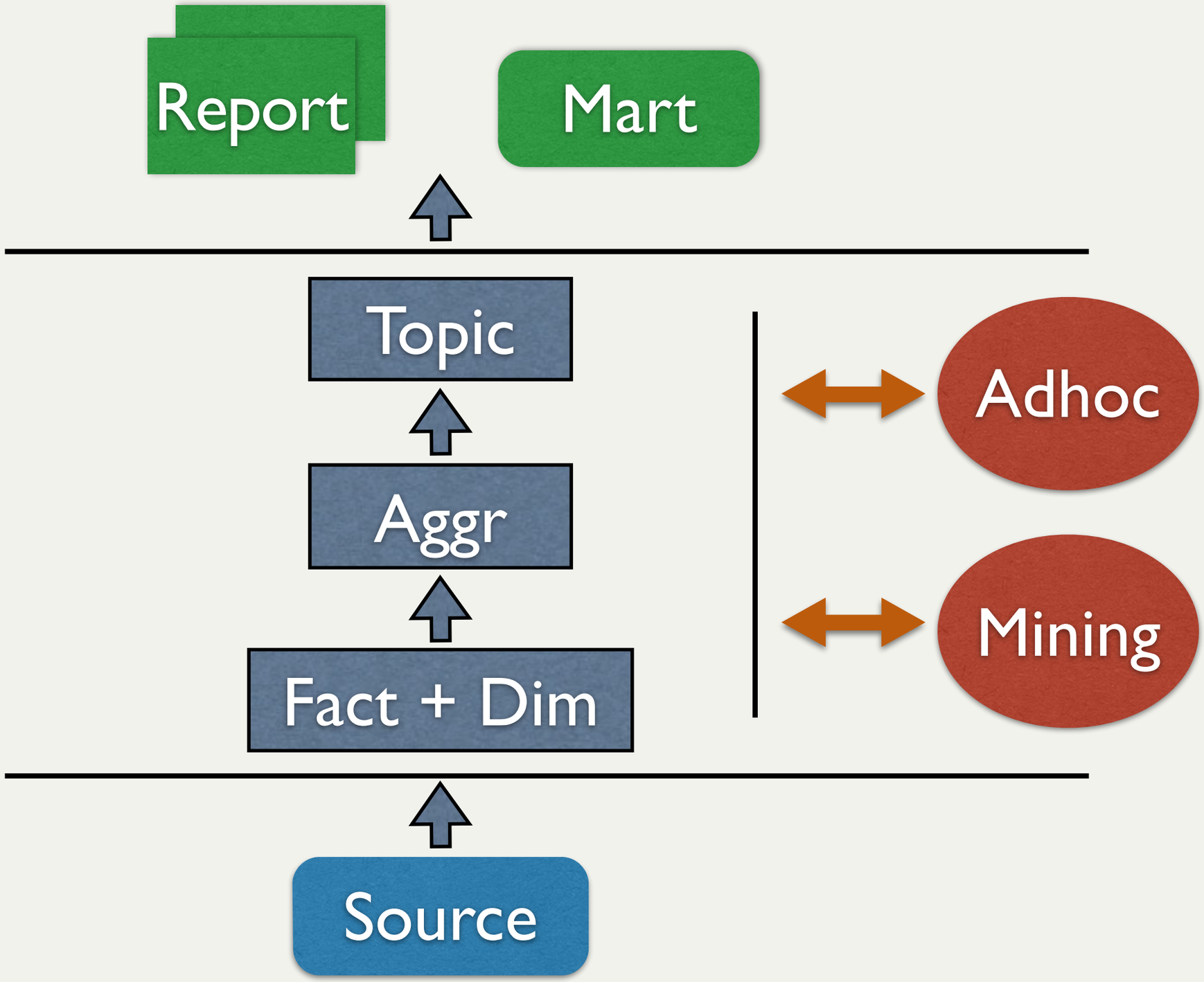
属性

名称	中文名	数据类型	引用实体
DealPayment	项目支付方式	Integer	DealPayment
Contract City	项目签单城市	Integer	City
Organization Unit	组织单元	Integer	Organization Unit
Begin Date	项目上线日期	Date	Date
End Date	项目下线日期	Date	Date
Deal	项目	Integer	
BD	签单BD	Integer	
Buy Price	项目进价	Decimal	
Price	项目售价	Decimal	
Coupon Period	美团券有效期	Integer	
Adfee	广告费	Decimal	
Type	三级品类	Integer	
Category	二级品类	Integer	
Class	一级品类	Integer	
Contract	合同	String	
Main Contract	主合同	String	
Coupon Type	团购券类型	Integer	
Status	项目状态	Integer	
Deal Attribute	项目属性	BigInteger	
Is Open	是否是已上线项目	Integer	
Is Normal Coupon Type	是否是美团券类型	Boolean	
Is Local Service	本地服务	Boolean	

简化处理层级:之前



简化处理层级:目标



模型驱动自动化工作流

美团网 meituan.com

- 自生成聚合表流程
- 自动创建主题表流程
- 自动部署上线并注册任务调度

驱动数据生产:聚合计划

美团网 meituan.com

- 选定事实
- 指定维度层级组合
- 生成Hive SQL (Multi Groupby)

驱动数据生产:聚合计划

美团网 meituan.com

```
<Aggregate fact="PageVisit">
    <LevelTuple>
        <Level entity="Date" level="Day"></Level>
    </LevelTuple>
    <LevelTuple>
        <Level entity="City" level="City"></Level>
        <Level entity="Date" level="Day"></Level>
    </LevelTuple>
    <LevelTuple>
        <Level entity="Deal" hierarchy="Category" level="Class"></Level>
        <Level entity="Date" level="Day"></Level>
    </LevelTuple>
    <LevelTuple>
        <Level entity="City" level="City"></Level>
        <Level entity="Date" level="Day"></Level>
        <Level entity="Deal" level="Deal"></Level>
    </LevelTuple>
</Aggregate>
```

驱动数据生产:聚合流程

美团网 meituan.com

```
FROM (
    SELECT
        `pagevisit`.`time` `pagevisit__time`
        `pagevisit`.`_c_ga_uuid` `pagevisit__uv`
        ...

    FROM `log`.`blog` `pagevisit`
    JOIN `dim`.`date` `dim_date` ON `dim_date`.`datekey` = `pagevisit`.`dt`
    JOIN `dim`.`city` `city` ON ...
    WHERE `pagevisit`.`dt` = '$now.datekey'
) `_inter`

INSERT OVERWRITE TABLE `pagevisit__date__calendar__day` PARTITION(`day`)
SELECT
    COUNT(`pagevisit__pv`) `pv`
    , COUNT(DISTINCT `pagevisit__uv`) `uv`
    ...
    , `dim_date__day`
GROUP BY `dim_date__day`

INSERT OVERWRITE TABLE `pagevisit__city__city_rank__city__date__calendar__day`
PARTITION(`day`)
SELECT ...
GROUP BY `dim_date__day`,`city__city`

INSERT OVERWRITE TABLE ...
```

驱动数据生产:主题定义

美团网 meituan.com

- 选定维度层级组合
- 指定事实列表
- 生成Hive SQL (FULL OUTER JOIN)

驱动数据生产:主题定义

美团网 meituan.com

```
<Topic name="Deal Day" caption="项目按天汇总">
    <TopicDimension>
        <Dimension name="Deal" entity="Deal" hierarchy="Category" level="Deal">
        <Dimension name="Date" entity="Date" hierarchy="Calender" level="Day"><
    </TopicDimension>

    <MeasureGroup>
        <Cube fact="OrderPayment"></Cube>
        <Cube fact="Consume"></Cube>
        <Cube fact="Refund"></Cube>
        <Cube fact="Feedback"></Cube>
        <Cube fact="Lottery"></Cube>
        <Cube fact="LotteryCode"></Cube>
        <Cube fact="DealCityShow"></Cube>
        <ExternalTable table="summary.dealuv_enddate"> ... </ExternalTable>
    </MeasureGroup>
</Topic>
```

自动注册调度

美团网 meituan.com

*文件名 haggr.pagevisit.sql

上游任务 +

*启动命令 /home/sankuai/bin/sw run -t

*增量重导 date # 增量重导参数配置手册

备注 Created by AggrGen base on ModelBuilder 0.4

+ 新增trigger

按天调度 # 1 按天调度 crontab任务 weekly队列 daytime队列 trihour队列 hour队列 monthly队列 halfday队列 删除

最早开始时间

参数 --delta 1

优先级 1

运行超时时长 7200 s

提交Review 预览

模型元数据即文档

美团网 meituan.com

属性

名称	中文名	数据类型	引用实体	备注
Deal	竞对项目	Integer	Comp Deal	
Date	日期	Integer	Date	
Poi	门店	Integer	Poi	
Comp Site	竞对站点	Integer		
Poi City	竞对商户所在城市	Integer		
Is Focus	是否重点城市	Boolean		

度量

名称	中文名	字段	数据类型	备注
Comp Poi Revenue	竞对销售额	revenue	Decimal	
Comp Poi Raw Revenue	竞对未脱水销售额	revenue_raw	Decimal	
Comp Poi Volume	竞对销量		Decimal	
Comp Poi Online Deal Count	竞对在线项目数	dealid	Integer	

模型元数据即文档

美团网 meituan.com

事实

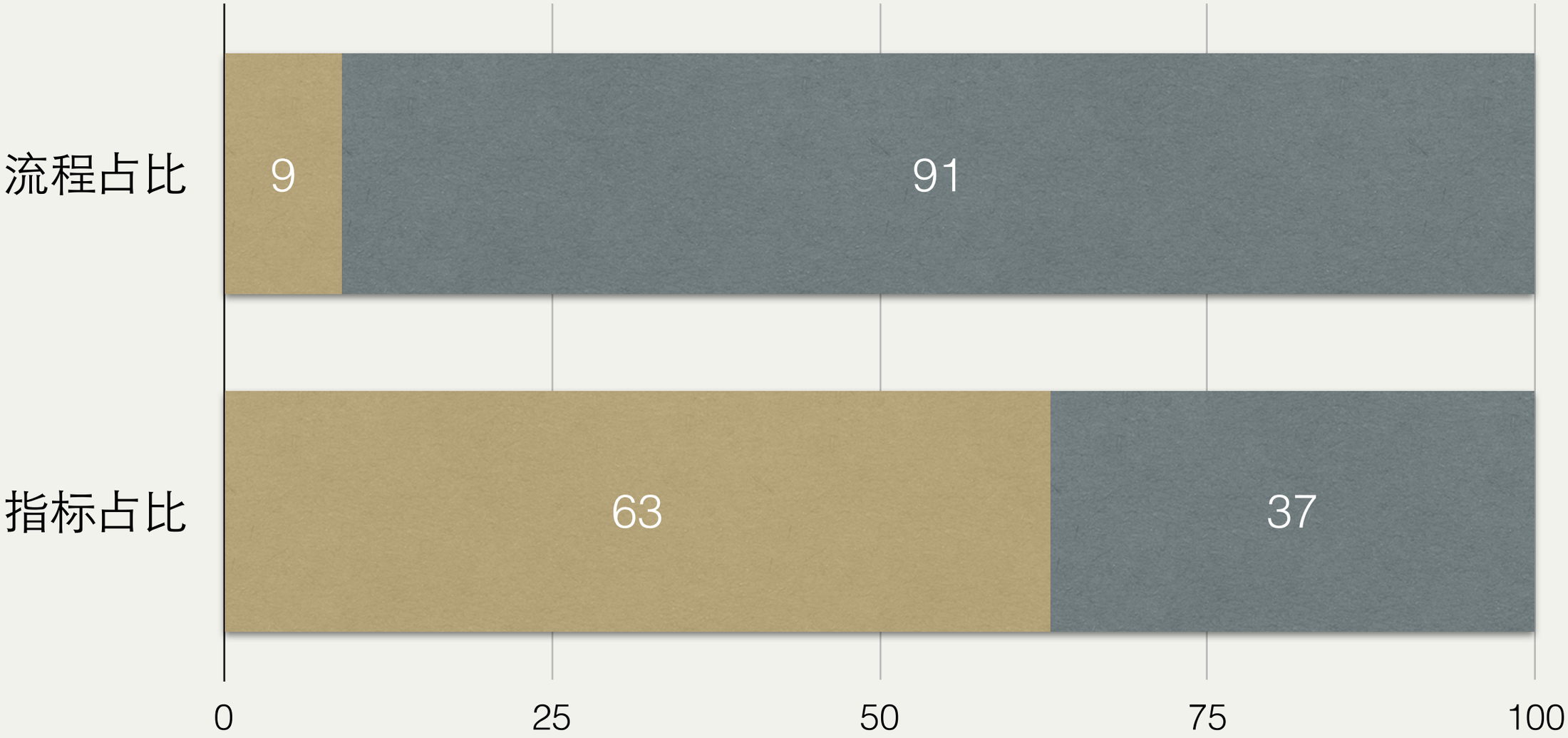
EmployeeStatus	员工每日在职状态
BDVisit	BD拜访
ContractSigning	合同签订
ContractPlan	上单计划
DealOpen	项目上线
DealCityShow	项目城市每日展示
DealPaymentTicketPay	项目支付日志
DealPaymentTicketPayLog	付款记录日志
Signup	用户注册
UserVerify	用户验证
OrderPlacement	下单
OrderPayment	订单支付
Consume	消费
CreditLog	用户资金流水
Refund	退款

效果

美团网
meituan.com

- 关注于业务建模, 工具执行规范和生产
- 业务知识和指标查找变得容易
- 减少混乱的指标定义带来的沟通工作
- 快速响应模型变更
- 便于快速铺开到新的产品线 (电影,酒店,外卖...)

效果



只解决了问题的80%

- 只满足常规需求
- 融合剩余的20%

目录

美团网
meituan.com

- 美团数据业务现状和要求
- 演进中出现的问题
- 统一模型管理方案和工具
- 经验总结

经验总结

- 元数据管理的形式: 文档 < wiki < 字典系统 < 形式化模型
- 对开放数据生产的态度: 平衡自治和统一
- 自动化,自助化,平台化

和美团一起高速成长

美团网
meituan.com

- 数据开发工程师
- 后台研发工程师
- Hadoop平台高级工程师
- 实时计算高级开发工程师

xuzhang@meituan.com

美团网

meituan.com

谢谢！