CHINA DATA ANALYST SUMMIT CDA 数据分析师 www.cda.cn

CDA 数据分析师 CERTIFIED DATA ANALYST IT大咖说 知识分享平台

大数据下的企业数据仓库建设

演讲人：代立冬

易观大数据平台负责人，近8年专注于企业大数据平台构建及优化，数据仓库建设。
之前曾在多个数据公司担任负责人及架构师等职务，熟悉零售及统计分析业务

跨界互联 数聚未来

第四届中国数据分析师行业峰会

CHINA DATA ANALYST SUMMIT

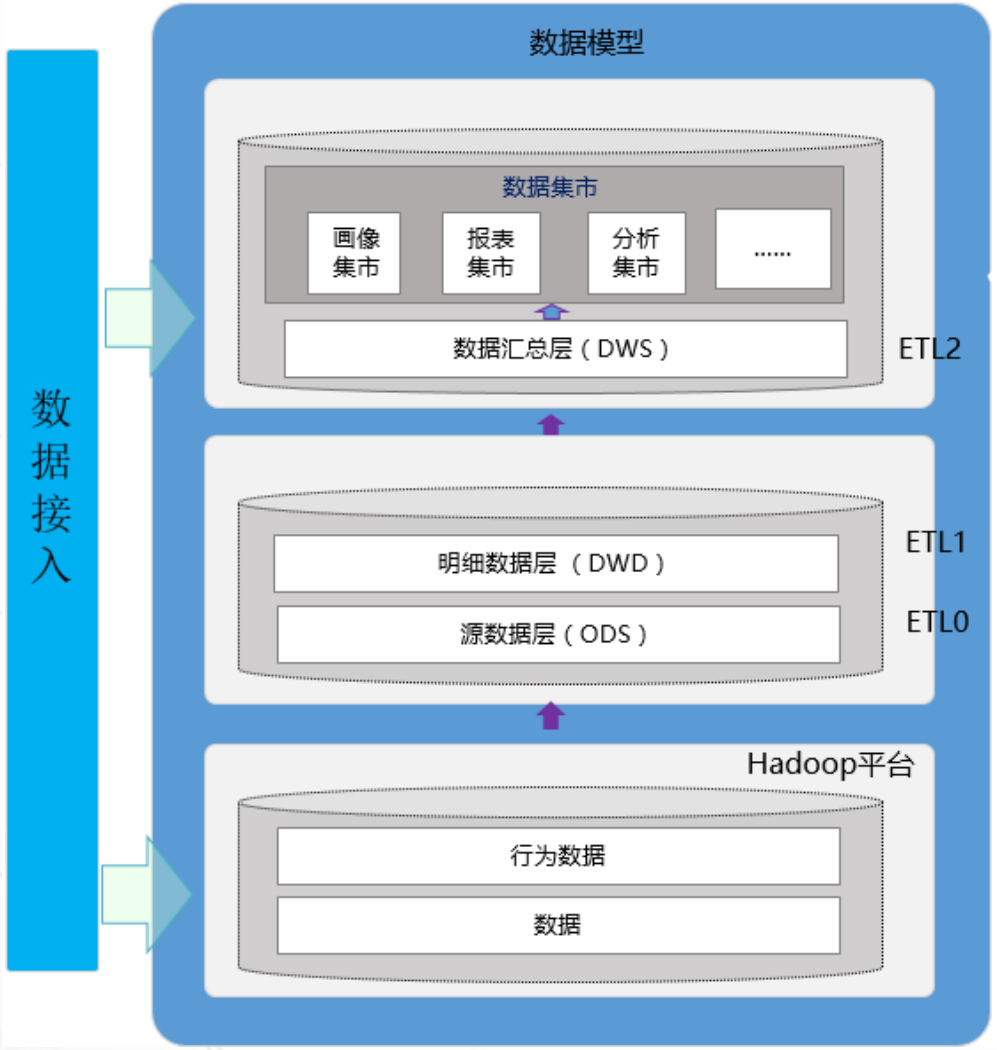
北京 中国大饭店 2017.07

为何要建数据仓库

- 避免底层业务变动对上层需求影响过大，不必改一次业务需求就重新从头处理数据
- 屏蔽底层复杂的业务逻辑，清晰数据结构并尽可能简单、完整的在接口层呈现业务数据，一句话总结就是使得业务人员使用起来更简单
- 数据来源和去向可追溯，即数据血缘关系，主要用于快速定位问题
- 减少重复开发，开发通用的中间层数据，减少很多重复的计算

那么如何做到上述要点呢？ 通常的办法是根据业务建立一套合理的数据分层模型

数据仓库整体架构



① 源数据导入ETL，进行数据的清洗、转换和入库。

② 基础数据加载到数据仓库ODS，规划保存7天

③ 清洗、转换后的ODS加载到DWD数据库永久保存，在分布式数据库内完成明细数据和轻度汇总数据加工生成

④ ODS数据，如消费数据经轻度汇总到DWD层，做长久保存

⑤ DWD数据分析处理在Hadoop平台处理到DWS层，产生的结果加载到DWS数据库

⑥ 业务应用通过数据访问接口获取所需求数据。

数据仓库—STG层

STG层是源数据层，也有伙伴叫它原始数据层、接口层、缓冲层等名字，不管什么名字，都是用来表示原始数据在数据平台的落地的。原始数据层就是数据接收端接到的数据，数据结构和客户端发送上来的保持一致。

- 数据结构

  数据结构与客户端上传时保持一致，存储使用parquet文件格式，方便后续MR处理的数据读取

- 存储周期

  STG层需保留全部数据

- 表命名规范

  [层次]_[主题] [_表内容]_[分表规则]

数据仓库—ODS层

ODS层是数据操作层，全称是Operational Data Store，只用于原始数据在数据平台的落地。这些数据从数据结构、数据之间的逻辑关系上都与原始数据层基本保持一致。在源数据装入这一层时，要进行诸如业务字段提取或去掉不用字段、脏数据处理等等

- 数据结构

  数据结构与原始数据层基本保持一致

- 存储周期

  ODS层默认保留近30天的数据

- 表命名规范

  [层次]_[主题] [_表内容]_[分表规则]

数据仓库—DWD层

DWD层是数据明细层，全称是Data Warehouse Detail,用于源系统数据在数据平台中的永久存储。它用以支撑DWS层和DM层无法覆盖的需求，比如像用户通话详单类业务需求。该层的数据模型不建议开给不懂技术的业务人员直接使用。这一层主要解决一些数据质量问题和数据的完整度问题。比如商场的会员信息来自于不同表，某些会员的数据可能不完整等等问题，我们可以在这一层做一些屏蔽

- 数据结构

  数据结构与源系统保持一致

- 存储周期

  保留历史至今所有的数据

- 命名规范

  dwd.dwd_业务描述+_时间粒度

  举例：dwd.dwd_sales_d　　销售详情表

数据仓库—DWS层

DWS层是数据汇总层，全称是 Data Warehouse Service，主要包含两类汇总表：一是细粒度宽表、二是粗粒度汇总表。我们拿商场售卖举例。那么包含基于订单、会员、商品、店铺等实体的细粒度宽表和基于维度组合(会员日进场汇总、会员日消费汇总、商场销售日汇总、店铺销售日汇总等)的粗粒度汇总表。这层是对外开放的，用以支撑绝大部分的业务需求。汇总层是为简化源系统复杂的逻辑关系以及质量问题等，这层使得业务结构容易理解，各个层面的工程师容易上手。dws层的汇总数据目标是能满足**80%的业务计算**

- 数据结构

  宽表：以业务实体进行展开，将与业务有关的相关字段和属性进行预关联、预处理和预计算，对业务实体进行拉伸形成宽表。汇总表：维度组合形成的汇总表

- 存储周期

  原则上保留历史至今全部的数据

- 命名规范

  所有数据汇总层的表都放在DWS下面

  dws.dws_业务描述+_时间粒度+_sum

  举例：dws.dws_trans_d_sum　交易日汇总表

数据仓库—DWS层

序号	目标表字段中文名	目标表字段英文名	数据类型	数据类别	主键/分区	源表名称	源表字段中文名称	源表字段英文名称	加工逻辑
1	会员ID	memberId	string	会员	主键	会员	ID	symbol_mid	
2	消费总额	sumShoppingSales	decimal	订单		订单	消费金额	sales	每天每人消费汇总
3	消费总笔数	sumShoppingCount	int	订单		订单	订单ID	orderId	每个订单算一笔
4	消费总次数	sumShoppingNumber	int	订单		订单	订单ID	orderId,memberId	每天每个人多笔订单算一次
5	消费总数量	sumShoppingQuantity	int	订单		订单	消费数量	saleQuantity	每天每人消费数量汇总
6	客单价	perShoppingTicket	decimal	订单		订单	消费金额	sales	消费总额/消费总笔数
7	统计日期	statisticsDate	date	订单	分区	订单	销售日期	saleDate	
8	统计时间	statisticsTime	timestamp	订单		订单	销售时间	saleTime	
9	年	year	int	订单		订单	销售日期	saleDate	当日是哪年
10	季	quarter	int	订单		订单	销售日期	saleDate	当日是哪季
11	月	month	int	订单		订单	销售日期	saleDate	当日是哪月
12	周	week	int	订单		订单	销售日期	saleDate	当日是哪周，一年52周

数据仓库—DIM层

DIM层这一层很简单，主要存储公共的信息数据，比如国家代码和国家名、地理位置等信息就存在DIM层表中对外开放，用于DWD、DWS和APP层的数据维度关联

- 数据结构
  维表，以国家ID等字段为主键。
- 存储周期
  按需存储，一般会保留历史至今所有的数据。
- 命名规范
  所有维度表都放在DIM下面。
  dim.dim+_业务描述
  举例：dim.dim_Year时间维表

数据仓库—集市层

DM层是数据集市层，用于BI、多维分析、推荐营销、标签、数据挖掘模型和其它数据服务。对外开放，为所有数据产品和数据出口提供数据支持。简称DM，以某个应用为出发点而建设的局部DW，为什么这么说，DM只关心自己需要的数据。不会全盘考虑企业整体的数据架构和应用，每个应用都有自己的DM。所以DM可以基于仓库建设也可以独立建设。集市层是按照业务主题、分主题构建出来的、面向特定部门或人员的数据集合，该层次的数据模型会开放给业务人员使用，进行数据挖掘及业务分析。这样业务人员利用工具或手工写出简单的SQL，将统计数据提取出来进行分析。

- 数据结构

  星型表，事实表+维表

- 存储周期

  按需存储，一般会保留历史至今所有的数据

数据仓库处理流程

层与层之间的流转是一个ETL过程,ETL过程分成3个步骤，E、T、L分别代表抽取、转换和装载。其实ETL过程就是数据流动的过程，从不同的数据层流向目标数据层。我们来看看数据模型这几层的ETL的说明

- ETL0

主要说明数据从哪里来到哪里去，以及对应的操作。
数据处理方向说明：源数据层 --> ODS层
处理方式：以Hive外部表形式，加载源数据到ODS层，主要明确了数据对应的实体、实体的属性、属性的类型等；此部分加工逻辑和频率为初始全量，日常增量。
该ETL过程需要考虑一定的数据清洗，比如异常字段的处理、字段命名规范化、时间字段统一等等问题

- ETL1

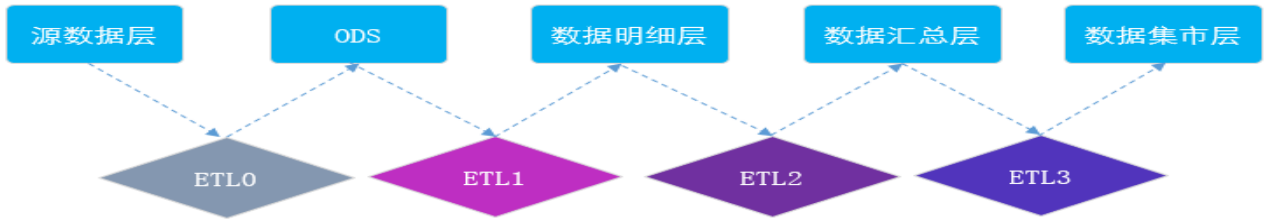
数据处理方向说明：ODS层 --> DWD层
处理方式：按业务关注点确定的主题域划分实体，确定实体间的关联关系，将ODS层的实体映射为DWD层的实体。此部分加工逻辑和频率为初始全量，日常增量

- ETL2

数据处理方向说明：DWD层 --> DWS层
处理方式：根据数据挖掘、数据分析等团队的实际需求，总结通用性的宽表、轻度汇总表、维表。此部分只处理简单也ETL处理，加工逻辑和频率为初始全量，日常增量或Merge

- ETL3

数据处理方向说明：DWS层 --> DM层
处理方式：根据实际业务需求进行更进一步或复杂的ETL处理



数据仓库建设的价值

✓高效的数据组织形式，方便维护

面向主题的特性决定了大数据分层模型拥有高效的数据组织形式，更加完整的数据体系，清晰的数据分类和分层机制。数据经过清洗和过滤，使原始数据不再杂乱无章，有效提高数据获取、统计和分析的效率

✓时间价值

分层模型的构建将大大缩短数据获取的时间，所有需要的数据都可以从中直接获取，一旦从各类数据源到分层模型的环节构建完成，每天各种数据就通过自动任务调度的方式流入到各个层中去，从而数据获取的效率会有一个很大的提升。从实际应用场景来看，使用分层也是可以大大提高数据的查询效率，尤其对于海量数据的关联查询和复杂查询

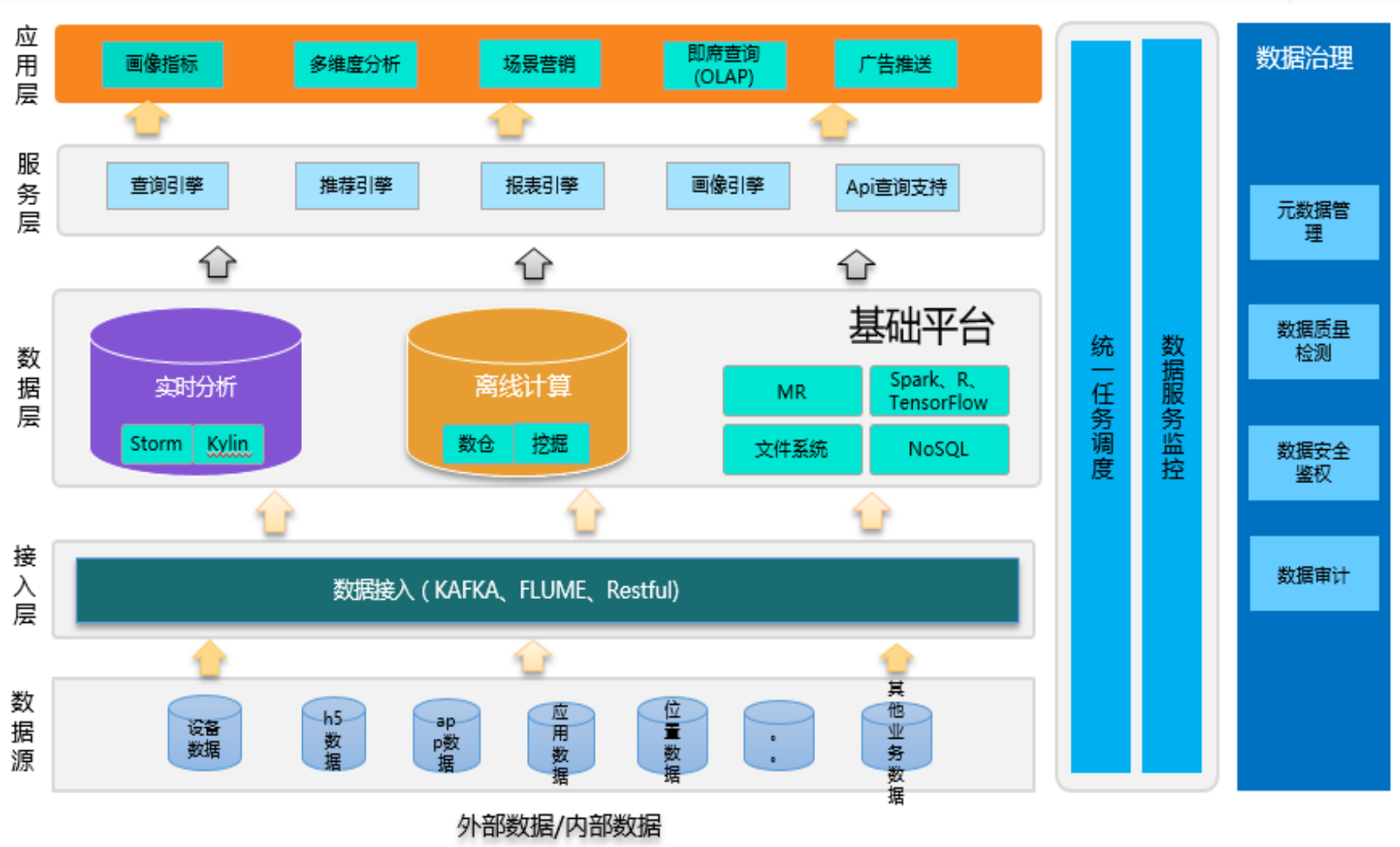
✓集成价值

分层模型中的数据是所有数据的集合，对于应用来说，实现各种不同数据的关联并使多维分析更加方便，为从多角度多层次地数据分析和决策制定提供的可能

✓回溯历史数据

记录历史是分层模型的典型能力，分层模型能够还原历史时间点上的各种状态，以便于能更好的回溯、分析历史，形成一个连续时间行为的分析，更好的预测未来的行为

数据平台整体架构


应用层
画像指标
多维度分析
场景营销
即席查询(OLAP)
广告推送
服务层
查询引擎
推荐引擎
报表引擎
画像引擎
Api查询支持
数据层
基础平台
实时分析
Storm
Kylin
离线计算
数仓
挖掘
MR
Spark、R、TensorFlow
文件系统
NoSQL
接入层
数据接入（KAFKA、FLUME、Restful）
数据源
设备数据
h5数据
app数据
应用数据
位置数据
。。
其他业务数据
外部数据/内部数据
统一任务调度
数据服务监控
数据治理
元数据管理
数据质量检测
数据安全鉴权
数据审计

Q/A

CHINA DATA ANALYST SUMMIT

CDA 数据分析师 www.cda.cn

THANKS

跨界互联 数聚未来

第四届中国数据分析师行业峰会

CHINA DATA ANALYST SUMMIT