链滴

基于大量数据复杂分析需求的解决方案

作者：jenphyjohn

原文链接：https://ld246.com/article/1577555835165

来源网站：链滴

背景

随着客户的统计需求日趋复杂，以及数据量的日渐增大，我们在应用程序内使用复杂SQL进行统计计的效率变得极低，执行时间超出了正常可以接受的范围。因此需要一个新的解决方案，可以满足复杂经常变化的统计需求。

ETL

ETL是英文Extract-Transform-Load 的缩写，用来描述将数据从来源端经过抽取（extract）、转换（transform）、加载（load）至目的端的过程。

我们整个系统的实现过程可以以ETL为模型，进行相关的设计及开发。

数据中心

我们姑且把我们的系统理解为一个数据中心，关于数据中心，这里有三个名词

- **ODS（Operational Data Store）**

可操作数据仓库，有如下特点

1. 在业务系统和数据仓库之间形成一个隔离，ODS直接存放从业务系统抽取过来的数据，这些数从结构和数据上与业务系统保持一致，降低了数据抽取的复杂性。

2. 转移一部分业务系统的细节查询功能，因为ODS存放的数据与业务系统相同，原来有业务系统

生的报表，现在可以从ODS中产生了。

3. ODS数据只能增加不能修改，而且数据都是业务系统原样拷贝，所以可能存在数据冲突的可能解决办法是为每一条数据增加一个时间版本来区分相同的数据。

- **DW（Data Warehouse)**

数据仓库，把ODS的数据进行处理、清洗，并转存到数据仓库中并长期保存，提供所有类型数据支持战略集合，是一个包含所有主题的通用的集合。

- **DM（Data Mart)**

数据集市，可以理解为对数据仓库的数据进行进一步加工，并提供给各级应用。

相关工具

列举一些常用的ETL相关工具：

数据抽取/同步：

- kafka
- flume
- sqoop

数据清洗

- hive
- pig
- storm
- spark

数据存储

- Hadoop
- hbase
- ES
- Redis

业务分析及选型

目前的需求是对历史数据进行离线分析，并且数据来源为关系型数据库MySQL，所以选择如下方案：

1. 使用 sqoop 进行数据同步，把有可能用到的表直接灌到ODS层，ODS使用 Hadoop（HDFS）行数据存储，以供后续的数据清洗。

2. 使用 hive 对ODS中的数据进行查询、清洗和计算，并输出到DW，DW的存储仍使用MySQL。

3. 回归到Web业务层，进行相关需求的开发，这里可以把不同的需求理解为不同的DM，进行数据的视化展示或导出。