数据仓库和OLAP傻傻分不清

Original 虞大胆 虞大胆的叽叽喳喳 5/9

Scan to Follow

大数据领域体系非常庞大，最近自己在了解数仓部分，做些记录。

首先解释OLTP和OLAP的概念，作为开发对OLTP比较了解，操作对象是数据库，也称为OLTP数据库（比如Mysql），主要用于CRUD操作，讲求高并发、低延时，一般作为业务数据使用。

而OLAP则是联机分析处理，做数据分析用的，比如进行数据聚合操作，它操作的数据源比较大，对性能要求相对较低。操作对象是数仓。有的时候OLAP也等同数仓。

数仓一般是多维模型模型，数据分层，ETL处理。它的数据源来源很多，格式也很多，比如结构化的数据，非结构化的数据。

对于ETL处理，需要对业务的理解非常透，比如MySQL是作为业务使用的，比如商品业务可能有很多类型的表，而到数仓后，可能会重新建模，比如分为维度表和事实表。

现在我们面临两个问题，第一就是ETL机制非常弱，基本上是原样将MySQL库导入到数仓；第二业务库变更后，需要重新构建，对于业务数据库的理解总是落后的。

那数仓有什么用呢，可以进行交互式查询，数据分析，数据挖掘，BI报表。

根据不同的理解，数仓也有很多的分类，比如：

1：根据建模分为MOLAP，ROLAP，HOLAP

MOLAP需要进行预计算，将可能的查询结果存储起来，适合分析比较稳定的场景，Kylin是这个领域的解决方案。

ROLAP是目前的主流，基于关系模型，构建在多维数据模型上，一般通过SQL就能查询。

2：对于ROLAP，有两种解决方案，一种是宽表模型，比如现在比较流行的clock-house；另外就是多表组合模型，比如Presto。

3：从实时性分，分为实时数仓和离线数仓，本文主要理解离线数仓，也叫批处理，就是数据是提前准备好的，比如Hadoop就是解决这类问题的。

4：对于OLAP来说，处理的数据是非常大的，为了加快处理，有两种解决方案：并行处理（比如 Hadoop 的Mapreduce，Spark，或者MPP架构的Presto），另外就是预计算（比如Kylin）。

那具体如何选型呢?

1：我们用的是比较常规的Hadoop，HDFS作为分布式存储，Mapreduce作为并行计算框架，但HDFS只是存储，没有结构化的概念，那怎么做数仓呢?

使用Hive解决了两个问题，首先它存储表结构元数据，其次Hive查询中的sql自动变为MR并行任务，MR从元数据中读取信息，然后去HDFS中读取数据，最后进行运算。

一般情况下这属于离线数仓，HDFS存储的是T-1的全量数据（不支持数据增删改查，只能整个文件覆盖），使用sqoop工具将MySQL导入到HDFS中。

2：MPP on Hadoop 的解决方案

由于MR操作HDFS的中间结果还是在磁盘，所以运算还是很慢的。

Presto是基于MPP架构，充分利用各个节点的cpu能力，中间结果放入内存，减少磁盘消耗。

比如Presto作为SQL执行引擎，本身不存储数据，它可以直接调用MySQL进行运算。

也可以调用Hive，读取元数据，然后操作HDFS的数据，进行并行运算。

有了Hive，有了Presto，结合可视化的BI工具，就能产生数据报表，进行数据分析和挖掘。

最后简单说下BI，有个公式：

BI平台=数据仓库+OLAP服务/报表。

People who liked this content also liked

召回和NLP在推荐系统中的作用
虞大胆的叽叽喳喳

女子地铁淫秽直播！网友直呼“恶心” 检察机关已批捕
扫黄打非

洪水中，外卖员深夜接单救人
真实故事计划