一种基于Hive的点评网站数据仓库的设计方法

宋江帆[1]　赵跃龙[1]　吴志攀[1,2]
(1.华南理工大学 计算机科学与工程学院，广东 广州 510006；
2.惠州学院 计算机科学系，广东 惠州 516007)

问题及解决方案

点评网站在O2O行业经过了十多年的耕耘后，已经积累了大量的评论数据、商户数据和用户数据等数据信息，而且每个月仍以百万级的数据量在递增。

问题：

存储：高性能服务器、Oracle　　----扩展性差、代价高

查询：面向业务，复杂查询耗时　　----查询效率低

解决方案：

存储-→Hadoop集群　　查询-→数据仓库

Hive=Hadoop集群+数据仓库

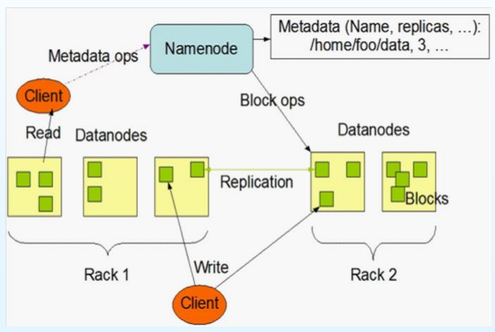
数据仓库

数据仓库是一种结构化的数据存储形式，其存储的数据具有集成性、面向主题、企业范围跨度、历史性以及稳定性的特点,数据仓库的目的就是用于对分析型数据进行检索。从它的定义中就可以很直观的得出它的如下特点：

- ●结构化数据存储
- ●集成性
- ●面向主题
- ●企业性
- ●历史性
- ●稳定性

HDFS

HDFS(Hadoop Distributed File System, Hadoop分布式文件系统)作为Hadoop的核心技术之一，是分布式计算中数据存储管理的基础。HDFS实际上是一个集群的主/从(Master/Slave)分布式的存储体系结构，一个HDFS集群拥有着一个NameNode(命名节点)和多个DataNode(数据节点)。



HDFS文件系统架构

Hive

Hive是一个建立在Hadoop上的数据仓库平台,能够应对大规模的数据集(如点评网站的数据)，它的设计目标是使Hadoop上的数据操作与传统的SQL语言相结合.

1)提供类似于SQL的查询语言HiveQL，可以执行查询、变化数据等操作；

2)通过解析将HiveQL语句在底层被转换为相应的Map/Reduce程序以便在Hadoop上执行；

ETL

Extraction-Transformation-Loading的英文简称，用来描述将数据从数据源经过抽取、转换、加载至数据仓库的过程，包含三个主要任务:

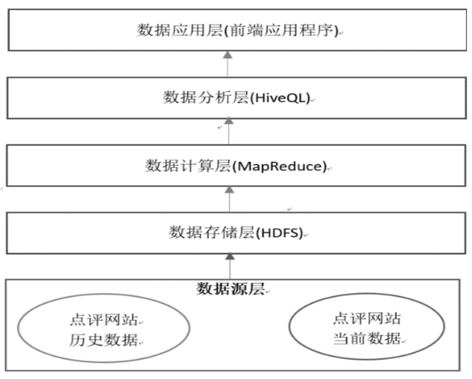
1）从操作型数据源中抽取可用于分析的有用数据;

2）转换数据使其满足数据仓库预先定义好的模型，同时利用数据清洗过程来保证转换后的数据质量;

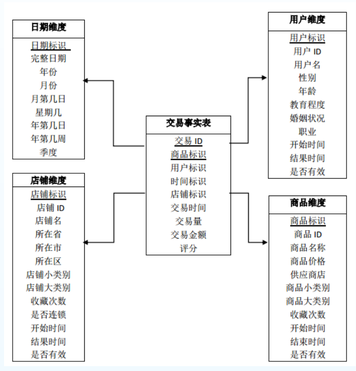
3）将转换后的高质量数据加载到目标数据仓库中。

一个成功的数据仓库背后必然有一个成功的ETL过程！

具体设计



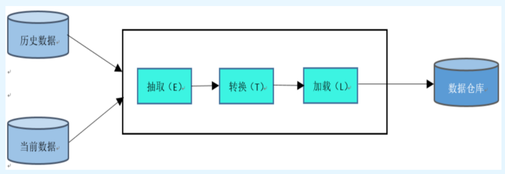
基于Hive的点评网站的数据仓库架构



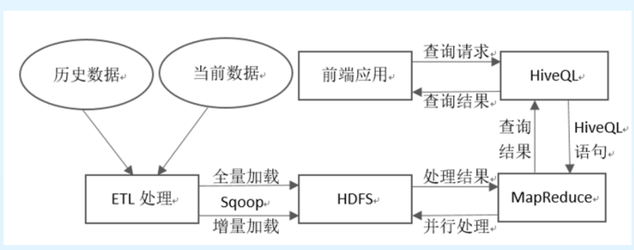
点评网站数据仓库模型

下图是针对维度表缓慢变化的解决方案。

用户标识	用户 ID	用户名	教育程度	……	开始时间	结束时间	是否有效
1	1001	coder	college	…	2010.09.01	2014.07.01	否
2	1002	dragon	college	…	2012.05.13	…	是
3	1001	coder	master	…	2014.07.02	…	是



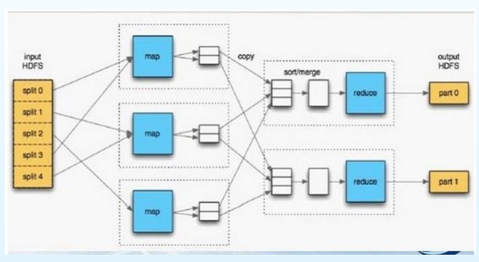
ETL体系结构设计



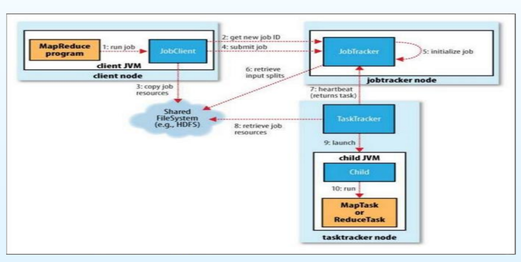
基于Hive的点评网站的操作过程

Map/Reduce

Map/Reduce是一种处理海量数据的并行编程模型和计算框架，用于对大规模数据集(通常大于1TB)的并行计算。Map/Reduce的核心部分就是map和reduce函数，map负责将一个大任务分解成多个小任务，reduce负责把分解后的多任务处理的结果汇总起来。在Hadoop中，用于执行Map/Reduce任务的机器角色有两个:JobTracker和TaskTracker。

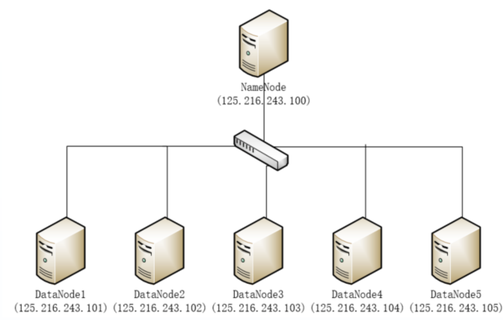


MapReduce计算模型

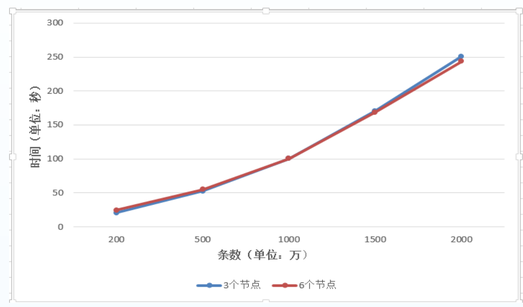


MapReduce执行过程

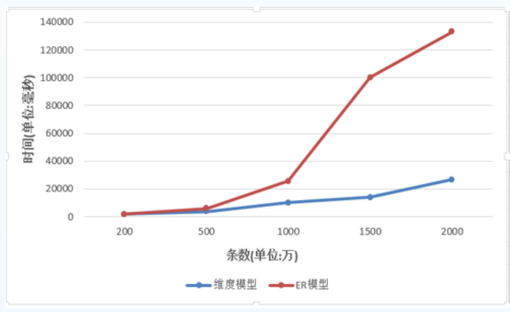
实验分析



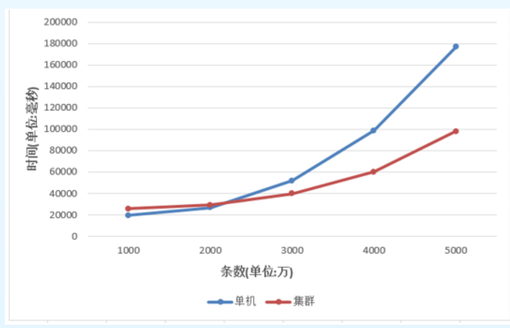
Hadooop集群部署



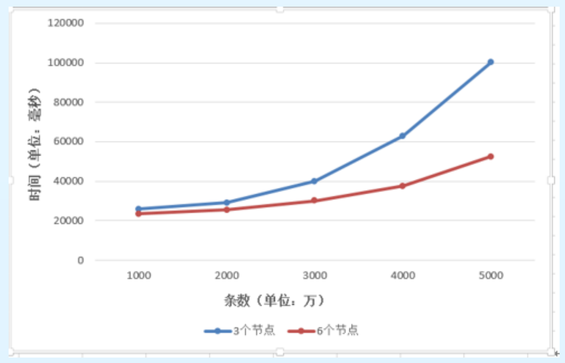
系统I/O性能测试



维度模型和ER模型查询时间对比



维度模型单机与集群查询时间对比



不同大小集群查询时间对比

总结与展望

总结：随着近些年的不断发展，点评网站已经积累了海量的数据，为了解决传统数据库在应对点评网站海量数据查询分析请求时效率低下的问题，本文提出了一种基于Hive的点评网站的数据仓库的构成方法。并且通过实验说明了在海量数据的查询效率方面，本文提出的数据仓库维度模型优于一般传统数据库的ER模型，而且集群优于单机，另外由于Hive本身已经集成了维度模型和集群的优势，所以实际上也说明了本文提出的方案的可行性。

展望：考虑到Hive的查询任务本质上仍是由Map/Reduce来执行的，而Map/Reduce在计算方面还存在一定的延时等问题，计算能力也有待进一步的提升，所以本文后续的研究重点是将点评网站的数据仓库架设在Spark等新一代分布式计算平台之上来进一步提高计算能力和查询效率。