3200/28937 4237/32233 21565

《基于Hadoop平台的大数据迁移与查询方法研究及应用》

1. 当代社会，伴随着信息技术的快速发展，全球时时刻刻都在产生数据，面对如此大量的数据，各领域、各企业对“大数据”的处理需求不断的提高，而如何科学的管理这些数据、如何快速查询或者更新数据信息成为一个亟待解决的问题。
2. 关系型数据库以交互式访问方式处理数据集、处理数据大小一般为GB级，其结构为静态模式、非线性线性横向扩展，更适合持续更新的数据集。而随着数据量越来越大、数据库中的数据类型不断的变化、各种非结构化数据逐渐成为了需要存储和处理的大数据的重要组成部分，仅仅使用传统数据库或数据分析软件(如SAS\SPAA)已经不能很好地解决大数据带来的问题，无法满足经济有效的存储、分析和访问数据。而在众多解决大数据量和处理效率问题上，Apache基金会设计实现的Hadoop在分布式计算和大量数据处理方面脱颖而出。
3. Hadoop平台以用于存储数据的文件系统HDFS和处理大数据集的并行计算框架MapReduce两部分为底层基础，此外，还提出了一些高层的抽象工具，如按列存储的数据仓库Hive，按列存储的数据库HBase，用以检索大数据集的Pig，以及在DBMS与HDFS之间进行数据传输的工具Sqoop等。
4. Hadoop平台相对于关系数据库具有用向外扩展架构代替向上扩展、处理类型为key/value的形式而非传统表形式、用MapReduce框架编程实现SQL语言查询、处理状态可为离线批量处理四个方面的特点。
5. Hadoop并行处理框架核心HDFS和MapReduce在集群上分别实现了大数据的分布式存储和并行处理。Hadoop通过HDFS为用户提供具有高容错性和高伸缩性的海量数据的分布式存储，并通过MapReduce为用户提供逻辑简单、底层透明的并行处理框架。
6. Hadoop是近年来IT界研究热点之一，大数据存储及处理是Hadoop的技术研究的主要两个方面。
7. DBInputFormat为Hadoop新版本提供的数据迁移工具，可以让Hadoop与多种关系数据库之间进行数据导入导出。它链接数据库的方式为主流的JDBC方式，让程序员更为方便的进行数据链接操作。DBInputFormat实现数据交换的原理。
8. Sqoop是一种使关系数据库和Hadoop之间数据能够相互转换的命令行应用程序。
9. 数据迁移是数据在存储类型、格式或不同机器（computer）之间相互转换的过程。
10. 通常情况下，执行数据迁移编程方式都是自动化的过程，这样既可避免重复而单一的劳动，又可以提高生产效率。
11. 程序化的数据迁移可能涉及多个阶段，且至少会包括两点：一是数据抽取，这是指数据从旧系统中读取的过程，二是数据装载，是指数据被写入到新系统中的过程。
12. 数据迁移需要保证迁移成本低于原系统的维护成本，还需要降低大量数据因结构变化导致的数据错误或丢失的风险；数据迁移必须保证既不影响原系统的正常运行又能够在迁移之后提高系统运行效率。
13. 本文综合考虑这些因素，对数据迁移工具进行对比，分析其迁移的具体实现；然后结合水路规费征稽系统（本次研究的课题来源），设计一个有针对性的数据迁移方案。
14. 本文主要讨论的理论体系为Hadoop的核心，基于Hadoop平台的数据仓库Hive、数据库HBase，可用于数据库和HDFS之间高效数据交换的工具。
15. 在此基础上，需要进一步对Hadoop数据交换原理、Hadoop作业调度器及MapReduce作业并行工作机制进行研究，进而将原系统关系数据库中的数据迁移到Hadoop平台中。
16. 本文首先从现有的数据管理系统中获取源数据，然后进行数据迁移，重点研究Hadoop平台下结构化数据迁移方案的设计与实现。由于MapReduce可以实现更优的并发性、体现优化数据转换的能力，因此，针对结构化迁移提出基于MapReduce架构的迁移方案。