<<Hadoop云计算实战>>

图书基本信息

书名：<<Hadoop云计算实战>>

13位ISBN编号：9787302296737

10位ISBN编号：7302296731

出版时间：2012-10

出版时间：清华大学出版社

作者：周 品

页数：411

字数：612000

<<Hadoop云计算实战>>

内容概要

《hadoop云计算实战》全面介绍了云计算的基本概念、google（谷歌）云计算的关键技术，以及hadoop云计算的相关配套项目及其实战，包括hadoop的hdfs、mapreduce、hbase、hive、pig、cassandra、chukwa及zookeeper等配套项目的实现机制、用法及应用。

《hadoop云计算实战》可作为高等院校本科生和研究生的教材，也可作为广大科研人员、学者、工程技术人员的参考用书。

<<Hadoop云计算实战>>

书籍目录

章节摘录

版权页： 插图： 3.收集器 Chukwa的收集器弥补了Hadoop集群不利于存储大量小文件的缺点。
收集器先是把收集到的小文件数据进行部分合并，然后写入集群，大幅减少了Chukwa产生的HDFS文件数量。
具体来说，通过HTTP数据被传送到收集器，每个收集器接收来自数百台主机的数据，并将所有数据写入到一个Sink文件中，MapReduce作业定期将Sink中记录的信息整合为日志收集文件。
Sink文件是一个由连续的Chunks组成的Hadoop序列文件，其是由大量的数据块和描述每一个数据块来源和格式的元数据组成的。
在收集数据期间，收集器会定期关闭Sink文件，更改文件名（便于保存及整理），重新创建一个新的文件，新文件仍被命名为“ Sink ”，接着再用新的Sink文件存储收集的信息，这就是所谓的“ 文件循环 ”。
收集器位于数据源和数据存储间，其屏蔽了HDFS文件系统的一些细节，方便于使用HDFS。
在某种意义上，收集器缓解了大量低速率数据源和文件系统间“ 步调 ”不协调的矛盾，优化了少量高速率数据源的写入。
为了防止收集器出现单点，Chukwa允许设置多台收集器，代理可以从收集器列表中随机地选择一个收集器传输数据。
当某个收集器失败或繁忙时，就选择其他收集器，以免影响代理的正常工作。
随机选择的节点使收集器的载入可能会极不均匀。
在实际应用中，收集器的任务负载很轻的情况很少出现；为了防止过载，系统设置了代理重试限制机制，如果数据写入收集器失败，收集器把待写入数据标记为“ 坏 ”数据，在重新写入数据前代理需要等待一段配置时间。
在实际应用中，多收集器的负载几乎是平均的，从而实现了负载的均衡化。
4.MapReduce作业 收集器顺序写入数据文件，方便于快速获取数据和稳定存储，但是，不便于数据分析和查找特征数据。
因此，Chukwa利用MapReduce作业实现数据分析和处理。
在MapReduce阶段，Chukwa提供了复用和存档任务两种内置的作业类型。
（1）demux作业 demux作业负责对数据的分类、排序和去重。
由收集器写入集群中的数据，都有自己的类型。
demux作业在执行过程中，通过数据类型和配置文件中指定的数据处理类，执行相应的数据分析工作，一般是把非结构化的数据结构化，抽取其中的数据属性。
由于demux的本质是一个MapReduce作业，所以用户可以根据自己的需求制定自己的demux作业，进行各种复杂的逻辑分析。
Chukwa提供的demux接口可以用Java语言来方便地扩展。

编辑推荐

《Hadoop云计算实战》可作为高等院校本科生和研究生的教材，也可作为广大科研人员、学者、工程技术人员的参考用书。

<<Hadoop云计算实战>>

版权说明