文章编号：2096-1472(2016)-11-20-05

基于HADOOP集群的数据采集和清洗

刘 晨[1]，焦合军[2]
（1.71320部队，河南 开封 475000
2.河南工程学院计算机学院，河南 郑州 451191）

摘 要：互联网的到来，使计算机行业蓬勃发展，各公司的业务数据也都到达P级别的数据量。本文结合Hadoop框架的中Hive和Hbase，对各个模块进行了详细的描述，重点分析了集群搭建步骤，及如何对集群的数据进行采集和清洗，并通过建立表来存储分析结果。

关键词：海量数据；Hadoop；hive；数据采集；数据清洗
中图分类号：TP311 **文献标识码**：A

Data Acquisition and Data Cleaning Based on the Hadoop Cluster

LIU Chen[1],JIAO Hejun[2]
(1.*Unit* 71320,*Kaifeng* 475000,*China*;
2.*School of Computer Science*,*Henan University of Engineering*,*Zhengzhou* 451191,*China*)

Abstract: With the flourishing development of computer industry,the business data in enterprises has reached level-P. Based on Hive and Hbase in the Hadoop framework,this paper elaborates on each module and analyzes the process of cluster construction, data acquisition,data cleaning and table construction to store analysis results.

Keywords:mass data;Hadoop;hive;data acquisition;data cleaning

1 引言(Introduction)

Hadoop框架本身大多是用Java编程语言编写的，一些本地代码是使用C语言编写的，命令行实用程序写成shell脚本，同时随着不同公司的工作需要，随之产生了许多不同的版本，极大的丰富了Hadoop的内容，如同后续出现的Hive、Zookeeper。截至2013年，已经有超过一半的世界500强企业采用Hadoop。Hadoop也在技术上被世界所认同。随着技术的发展和革新，全球大多的企业都对Hadoop青眼有加。

目前为止，Hadoop已经涉及了全球一半以上的数据处理的工作，是当下最为实用数据处理平台，研究Hadoop会使得海量数据的处理变得异常轻松。Hadoop在现代社会的应用已经涉及了通信、电子商务、军事领域和互联网。

通过使用Hadoop，旅游公司通过Hadoop和Hive可以迅速的帮助游客筛选理想的旅游地点和酒店住宿等功能，并能够分析出中短期时间内旅游热门的趋势。而一些网络公司，如Facebook、百度等，也运用Hadoop来处理用户的状态更新，日志生成，并根据用户的喜好分析并推送相应的应用和产品。甚至军方也在应用Hadoop，譬如美国军方就应用Hadoop的Digital Reasning来梳理来自于情报部门的大量非结构化文本数据，并从这些分析报告中寻找出可能危害国家及人民安全的文件。最常用的也是我们普通人时时刻刻都在运用的搜素引擎，搜索引擎通过网络爬虫和建立索引来搜罗网上的信息，而这两项技术就依靠Hadoop平台，将网页上的内容爬取到字节的本地服务器上，然而百度的爬取量是非常巨大的，并且为了保证数据的新鲜度，百度需要时时刻刻的向不同的网站发送爬取的请求，所以就会拥有成千上万个爬虫程序同时爬取数据，这是一个非常大的挑战，运用Hadoop平台可以将爬取的数据高效的存储起来，当然这也是一个非常大的工程。爬取之后的数据存放在本地的服务器上，但是用户这个时候并不能通过这些数据查到东西，在查询之前，百度还需要将这些数据一一建立索引，就如同从字典中查汉字，字典中的偏旁部首就是索引，除了一些像的、和等字不需要之外，大部分的数据都需要一一建立索引，每种格式的文档都要有一种相对应的解析程序，以此来规避一些奇怪的符号，从而提取出数据中有用的信息。而索引的生成需要的就是高效的执行速度，所以需要运行在足够多个机器上，在每个机器上同时进行扫描输入数据和内存更新索引的操作，随着数据的增多，这些索引的合并操作是呈线性增长的，基于这个原因，Hadoop项目下的MapReduce就体现出了它的价值，MapReduce[1]是一个应用广泛的分布式计算框架，MapReduce会将一个比较大的任务分割成诸多小任务，并将这些小任务分发给多个Mapper程序，也就是将任务分割之后

布置在成千上万台机器上同时运算，以此来提高效率[2]。而Hadoop在这一方面展现出了极大的优势。

大数据要经过清洗、分析、建模，以及可视化后体现出其潜在的价值。但是，由于网民数量的不断提升、社交网络的繁荣和业务应用的多样化，单个文件(如日志文件)变得越来越大，文件的存储成本和硬盘的读取速度越来越显得捉襟见肘。与此同时，政府、保险公司和银行等内部存在海量的不规则、非结构化的数据；只能将这些数据采集并清洗为有条理的数据，才能提高企业决策支撑能力，以及政府的决策服务水平，使其发挥应有的作用。

2　国内外研究现状(The research status at home and abroad)

2.1　国外文献研究

Apache Hadoop项目正式启动以支持MapReduce和HDFS的独立发展。Yahoo建立了一个300个节点的Hadoop研究集群，逐渐研究集群增加到600个节点。1.0.0版本出现，标志着Hadoop已经初具生产规模。

有人在ACM数字图书馆中以Hadoop为主题检索获得218篇论文，文献统计表明，对Hadoop进行研究的美国高校较多。企业中主要有雅虎等大公司，其他机构还有一些数据分析公司、研究所和技术协会等，具体统计见表1。

表1 对Hadoop进行研究的机构

Tab.1 The research institutions in Hadoop

研究类型	雅虎	微软	Google	IBM	Facebook	HP	易趣	AT&T	新浪	合计
自主研究	19	9	4	3	2	2	0	0	0	39
合作研究	12	5	5	4	3	2	2	1	1	35

从研究内容上分析总体上可以分为理论和应用两大方面。理论研究主要是性能优化和任务调度，应用研究主要是数据分析和数据查询，结果如表2所示。

表2 对Hadoop的研究方向

Tab.2 Study of Hadoop

理论研究	比较研究	任务调度	性能优化	功能扩展	合计
数量(篇)	9	26	36	21	92
比例(%)	9.78	28.26	39.31	100	22.83

2.2　国内研究现状

国内引用大数据技术最早的要数淘宝和百度了。2012年后，淘宝拥有2—3个集群，单一集群在3000节点以上。正是因为Hadoop可以部署在要求不高的节点之上，大量减少成本，大数据技术才能发展如此迅速。而支付宝的集群规模也达到了700台，使用Hbase，将个人的个人消费记录，以key-value键值对存储。

反观百度，从2008年到2013年，仅5年的时间，百度搭建的规模从300台机器，扩展到很大的规模。主要进行日志的存储和统计；网页数据的分析和挖掘；在线数据的反馈，及时得到在线广告点击情况。

总的来说，中国在大数据方面的研究要落后于国外好几年，在技术反面也是没有什么自己独创的技术，都是在学习国外的技术，在人才方面，人才匮乏，处于培养阶段，有经验的人特别少，大部分人还处于学习阶段，从国外的书籍，文献中学习Hadoop技术。在数据交易方面，2014年2月20日，国内首个面向数据交易的产业组织——中关村大数据交易产业联盟成立。同时成立的中关村数海大数据交易平台是国内首个重点面向大数据的数据交易服务平台，目前有1203家数据提供商。

2015年4月14日，全国首家以大数据命名的交易所，即贵阳大数据交易所正式挂牌成立，并在当日成功完成了首笔数据交易。值得注意的是，贵阳大数据交易所交易的并不是底层数据，而是基于底层数据，通过数据的清洗、分析、建模、可视化出来的结果。而采取这一过程的目的，就是为了解决数据交易和使用过程中保护隐私及数据所有权的问题。

3　环境搭建(Environment building)

安装好虚拟机，操作系统版本centerOS-6.5，选用的是桥接模式。我的Hadoop集群一共是三个节点，主节点IP是192.168.15.148，分节点分别是192.168.15.161和192.168.15.228。由于条件有限，我是在一台电脑上安装三个虚拟机，另外一台电脑通过SecureCRT，WinSCP等软件对虚拟机进行操作。WinSCP可以与虚拟机连接，然后给Linux系统传输数据，而SecureCRT连接虚拟机，可以进入控制台执行Linux命令，完成搭建。

3.1　jdk的安装

验证网络和设置静态IP，在安装jdk之前，需要确定虚拟机是否能够与物理机器ping通。我们可以给Linux系统分别配置一个静态ip地址，在有网络的时候系统会给我们自动分配一个ip地址，配置静态ip，命令是vi/etc/sysconfig/network-scripts/ifcfg-eth0，内容如图1所示。

图1 静态ip的配置内容

Fig.1 Static IP configuration

由BOOTPROTO=dhcp变为none，此时改为静态，并设定静态IP地址。BOOTPROTO是用来配置静态化的。之后，可以执行命令service network restart重启网络。这样就可以通过查看重启后的ip是不是自己指定的静态ip地址。

配置hosts文件，分别在虚拟机上配置hosts文件，方便集群搭建，然后在三台虚拟机上分别执行ping namenode、datanode1、datanode2，看是否ping成功。

卸载原有java版本，安装sun公司的jdk。因为Hadoop需要使用到jdk中的编译工具，所以我们为了更好的运行Hadoop，需要安装jdk。使用Centos系统安装好虚拟机，一定是有已经安装好的open jdk。把已有的java版本全都删除。

发送文件到linux中，使用的软件是WinSCP有JDK和Hadoop，传输过程如图2所示。

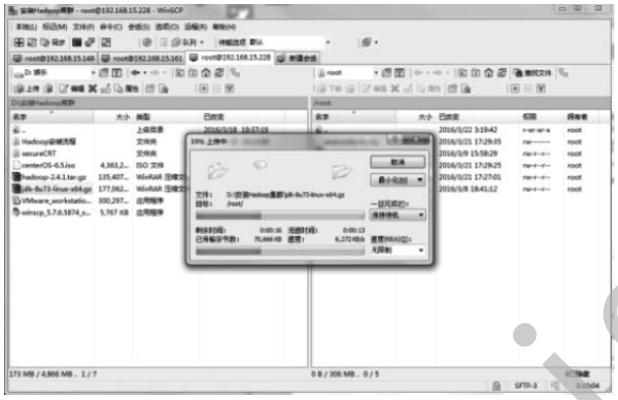

图2 WinSCP传输文件到linux

Fig.2 WinSCP transfer files to linux

安装jdk，传送完文件之后，使用命令行解压压缩文件。解压jdk的命令是gzip-d jdk-8u73-linux-x64.gz得到一个jdk1.8.0_73的文件夹，配置环境jdk变量，配置文件内容如图3所示。

图3 配置jdk的环境变量

Fig.3 The JDK environment variables configure

添加完毕保存退出，执行source/etc/profile，使配置的环境变量生效。执行java-version，将出现java1.8的版本，即安装成功。

3.2 ssh的安装

ssh会产生一对公钥和密钥，客户端访问服务端的时候，服务端会发送一个公钥给客户端，通过产生的密钥来对传输过来的数据进行解密，如果能够解密，说明可以相互通信。具体过程如图4所示。

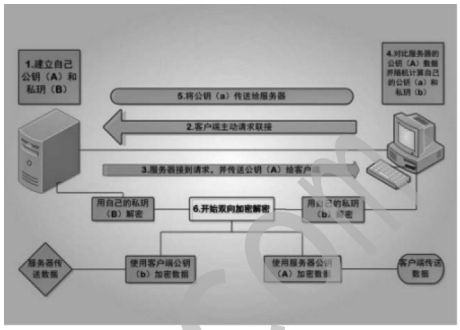

图4 ssh免密码登录的原理

Fig.4 The principle of free ssh password

设置免密码登录，执行ssh-keygen-t rsa产生密钥，会产生两个文件id_rsa(密钥)和id_rsa.pub(公钥)。将id_rsa_pub的内容复制到对方的authorized_keys文件下。验证ssh是否成功：ssh localhost成功就会登录到本地服务器上。

3.3 Hadoop集群搭建

解压缩Hadoop安装包，解压命令是tar-zxvf Hadoop-2.6.4.tar.gz，解压到/root目录下。配置Hadoop的环境变量，并使其生效。修改Hadoop的配置文件，添加jdk的环境变量。修改Hadoop-env.sh文件，执行命令vi Hadoop-env.sh，把JAVA_HOME的路径改为当前路径/root/jdk1.8.0_73。修改yarn-env.sh文件，编辑yarn-env.sh文件，在文件中加上export JAVA_HOME=/root/jdk1.8.0_73。

修改core.site.xml文件、hdfs.site.xml文件、mapred-site.xml文件和yarn-site.xml，在文件中添加相关内容。

拷贝profile到子节点，主节点上执行scp/etc/profile root@datanode1:/etc/，scp/etc/profile root@datanode2:/etc/然后在两个子节点上分别使新的profile生效。

完成配置之后，主节点namenode格式化，分节点不要格式化，并且主节点只能格式化一次。权限不够的话修改文件权限，-R代表的是整个目录下的东西都赋予权限。

执行格式化，如果出现了Successfully formatted代表格式化成功了。然后可以主节点namenode上在sbin目录下启动start-all.sh文件。查看是否安装成功，输入jps命令，

namenode上会出现四个进程，如图5所示。

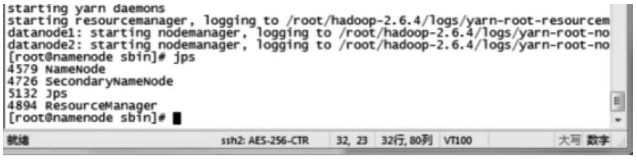

图5 成功启动namenode节点的进程列表

Fig.5 The namenode node process list

3.4 Hive集群搭建

拷贝安装文件到linux中/root/apache-hive-2.0.0-bin.tar.gz，执行解压命令tar-zxvf apache-hive-2.0.0-bin.tar.gz。

配置环境变量使其生效，编辑内容如下：

#hive export HIVE_HOME=/root/apache-hive-2.0.0-bin

export PATH=$PATH:$HIVE_HOME/bin

在hdfs中新建目录tmp和/user/hive/warehouse三级目录。分别给目录赋予权限。将mysql的驱动jar包拷入hive的lib目录下面，进入hive的conf目录下面，复制hive-default.xml.template并命名为hive-site.xml。使用schematool初始化schema，生成一个hive数据库，运行schematool-initSchema-dbType mysql，此时可以测试hive是否安装成功，执行hive命令，使用jps查看会有一个RunJar的进程，证明hive搭建成功。

3.5 Sqoop与MySQL搭建

3.5.1 Sqoop搭建

解压sqoop文件，将sqoop-1.4.6.bin__Hadoop-2.0.4-alpha.tar.gz拷贝到linux，然后解压tar-zxvf sqoop-1.4.6.bin__Hadoop-2.0.4-alpha.tar.gz。

修改sqoop的配置文件，找到sqoop-1.4.6.bin__Hadoop-2.0.4-alpha/conf目录中的sqoop-env-template.sh文件，执行命令cp sqoop-env-template.sh sqoop-env.sh，复制sqoop-env-template.sh的内容改名为sqoop-env.sh，修改sqoop-env.sh的内容，修改的内容如下：

#Set path to where bin/Hadoop is available

export HADOOP_COMMON_HOME=/root/Hadoop-2.6.4

#Set path to where Hadoop-*-core.jar is available

export HADOOP_MAPRED_HOME=/root/Hadoop-2.6.4

#set the path to where bin/hbase is available

export HBASE_HOME=/root/hbase-1.2.1

#Set the path to where bin/hive is available

export HIVE_HOME=/root/apache-hive-2.0.0-bin

#Set the path for where zookeper config dir is

export ZOOCFGDIR=/root/zookeeper-3.4.8

修改环境变量。执行命令vi/etc/profile，编辑profile文件内容，配置sqoop的环境变量。配置内容如下：

#sqoop

export SQOOP_HOME=/root/sqoop-1.4.6.bin__Hadoop-2.0.4-alpha

export PATH=$PATH:$SQOOP_HOME/bin

为了使配置的环境变量保存后能够生效，执行source/etc/profile

添加jar包，sqoop是将数据库中的数据导入导出到其他数据库，这里是将hive中的数据库的数据导出到mysql数据库中，即把mysql-connector-java-5.1.32-bin.jar放入sqoop的lib目录中。

3.5.2 MySQL搭建

检查mysql是否已安装，执行命令rpm-qa|grep-i mysql检查是否已安装Linux系统，结果发现在安装Linux系统的时候，会自动安装mysql-libs-5.1.71-1.el6.x86_64。

删除已安装的mysql执行yum-y remove mysql-libs*命令会将mysql名字中含有mysql-libs的mysql全部删除。

解压mysql的文件tar xvf MySQL-5.5.49-1.linux2.6.x86_64.rpm-bundle.tar命令将压缩包解压成多个文件，安装MySQL-server、MySQL-devel、MySQL-client。安装的MySQL就可以满足需求。下面的是安装命令：

rpm-ivh MySQL-server-5.5.49-1.linux2.6.x86_64.rpm

rpm-ivh MySQL-devel-5.5.49-1.linux2.6.x86_64.rpm

rpm-ivh MySQL-client-5.5.49-1.linux2.6.x86_64.rpm

修改mysql登录密码，首次安装时，默认密码为空，可以使用mysqladmin-u root password mysql命令修改root密码，其中mysql就是你自己设置的mysql新密码，测试是否设置成功登录mysql，执行mysql-u root-p mysql看是否能够登录mysql。Rpm压缩包安装的MySQL是不会自动安装/etc/

my.cnf文件的，需要执行命令cp/usr/share/mysql/my-huge.cnf/etc/my.cnf。

mysql默认是不可以远程访问，使用下面的sql语句设置远程访问Grant all privileges on *.* to 'root' @' %' with grant option。只是这样并不能是权限生效，可以执行flush privileges，使权限生效。

4 数据采集和清洗(Data acquisition and data cleaning)

对采集到的数据，用MapReduce对数据进行清洗。Java代码是建立在maven项目下，所以可以直接进入到项目所在地址，使用cmd命令执行mvn package将项目打成jar包。要想在hive中执行这个jar包，我们可以创建一个新方法，方法名为log_date_paser，执行的命令是create function log_date_paser as'com.hiveudf.sample.LogDateParser'；把jar加入到指定目录，即可调用log_date_paser方法清理日志数据。把原始处理清洗后，放到hdfs的/hmbbs_cleaned目录下。

4.1 执行过程

MapReduce的执行过程如图6所示。

图6 Java代码实现数据清洗

Fig.6 Java code for data cleaning

4.2 执行结果

MapReduce的执行结果如图7所示。

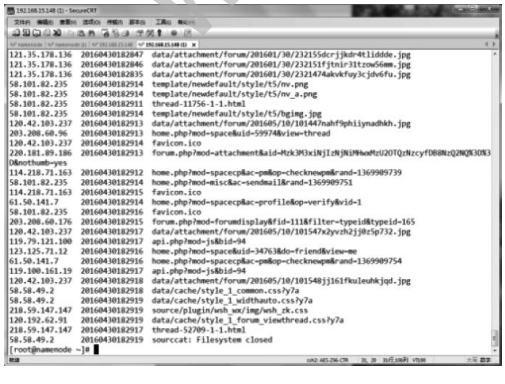

图7 MapReduce进行数据清洗的执行结果

Fig.7 The results of data cleaning

5 结论(Conclusion)

本文设计了基于Hadoop的数据清洗程序，执行结果显示执行成功。对于一个分析系统来说，Hive可以用于对清洗后的数据进行分析，脚本语句是create external table hmbbs(ip string,atime string,url string) partitioned by (logdate string) row format delimited fields terminated by'\t' location'/hmbbs_cleaned';创建一个外部分区表，从hmbbs_cleaned目录中得到数据，该表的字段指定的有ip,atime,url。由hive的数据库中导出到mysql的hmbbs数据库中，借助于Sqoop使用命令select*from hmbbs_logs_stat;可以将汇总数据给用户展示出来。

参考文献(References)

[1] Benjamin Harvey,Soo-Yeon Ji.Cloud-Scale Genomic Signals Processing Classification Analysis for Gene Expression Microarray Data[J].Proceeding of 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society,Chicago,2014(8):1843-1846.

[2] XIONG Wei,et al.A Self-Adaptation Approach Based on Predictive Control for SaaS[J].Chinese Journal of Computers,2016,39(2):364-376.

[3] Lvchen Zhao,et al.A New Method for Fetal Movement Detection Using an Intelligent T-Shirt Embedded Physiological Sensors[J].Proceeding of IEEE 16th International Conference on Communication Technology,Hangzhou,2015(10):563-567.

[4] 曾金梁.分布式日志分析系统的设计与实现[D].北京:北京邮电大学,2014.

[5] 郑启龙,等.基于MapReduce模型的并行科学计算[J].微电子学与计算机,2009(08):43-48.

[6] 刘永增,张晓景,李先毅.基于Hadoop/Hive的web日志分析系统的设计[J].广西大学学报(自然科学版),2011(1):64-66.

[7] J.Locke.An Introduction to the Internet Networking Environment and SIMNET/DIS[J].Technical Science Dept,Naval Postgraduate School.2003(1):24-29.

作者简介：

刘 晨(1988-)，男，本科，助理工程师.研究领域：信息安全，大数据分析.

焦合军(1981-)，男，博士，讲师.研究领域：云计算.