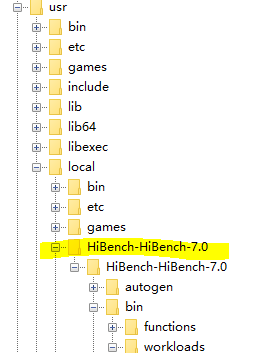
安装hibench

1. 下载安装包

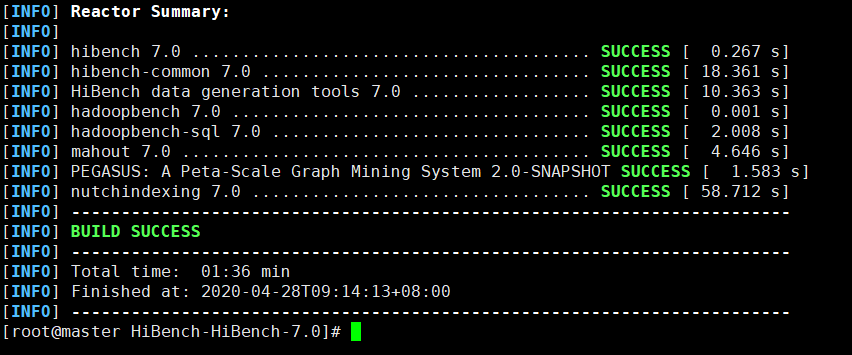
直接运行git clone <https://github.com/intel-hadoop/HiBench.git>下载安装包，也可在Windows系统下进入https://github.com/intel-hadoop/HiBench下载zip包，利用xftp移动到Linux系统，解压到/usr/local目录下。



路径：/usr/local/HiBench-HiBench-7.0/HiBench-HiBench-7.0

1. 安装Hadoobench框架

在该路径/usr/local/HiBench-HiBench-7.0/HiBench-HiBench-7.0下运行一下内容安装Hadoopbench

mvn -Phadoopbench -Dspark=2.1 -Dscala=2.11 clean package

安装成功！

1. 运行Hadoopbench

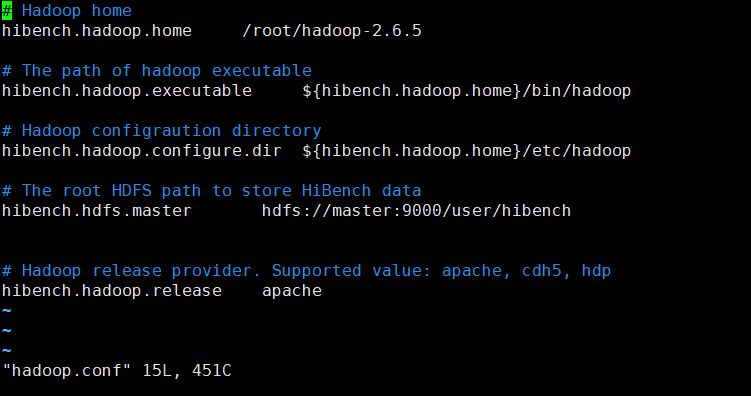
3.1 修改配置文件

路径：cd /usr/local/HiBench-HiBench-7.0/HiBench-HiBench-7.0/conf

运行：vim hadoop.conf

hibench.hadoop.home是hadoop的安装路径

hibench.hdfs.master可根据hadoop-2.6.5/etc/hadoop/core-site.xml查询到本机端口为matser:9000



* 1. 执行测试脚本，开始运行

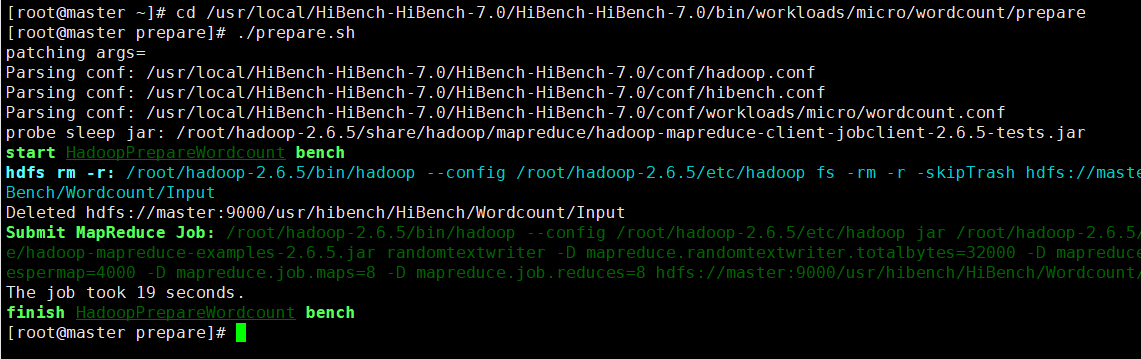
在bin/run\_all.sh 该脚本为测试所有的测试基准模块(将运行所有在conf/benchmarks.lst和conf/frameworks.lst中的workloads)；

以word count为例

1. 在master中，开启分布式系统start-all.sh
2. 生成测试数据

进入该路径：/usr/local/HiBench-HiBench-7.0/HiBench-HiBench-7.0/bin/workloads/micro/wordcount/prepare

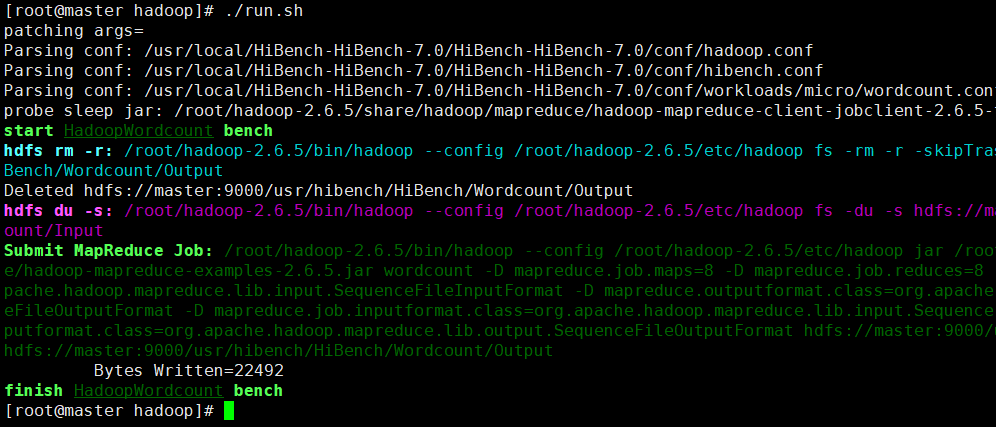
运行脚本：./prepare.sh



1. 运行word count的测试例子

进入该路径：/usr/local/HiBench-HiBench-7.0/HiBench-HiBench-7.0/bin/workloads/micro/wordcount/hadoop

运行脚本：./run.sh



1. 查看测试报告

安装bc工具用来生成report文件，输入yum install bc，（已经安装过，不用重复安装）

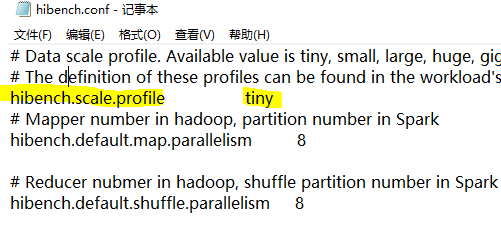
进入路径：/usr/local/HiBench-HiBench-7.0/HiBench-HiBench-7.0/report

运行：vim hibench.report

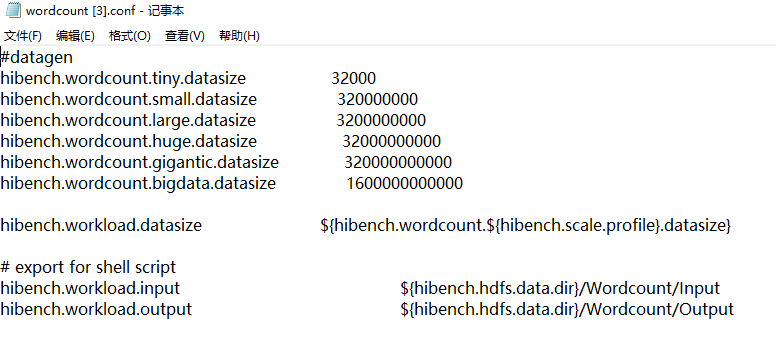


1. （可选）配置输入数据规模：

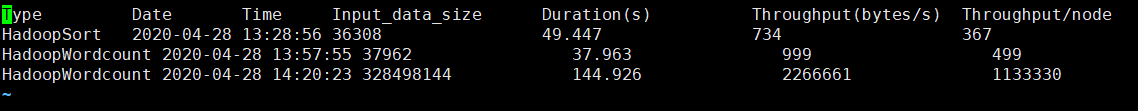
进入/usr/local/HiBench-HiBench-7.0/HiBench-HiBench-7.0/conf文件中，编辑hibench.conf文件，如下图所示：



可以配置tiny, small, large, huge, gigantic and bigdata。值的定义可以在对应的测试实例的配置文件中找到，例如Wordcount的在/usr/local/HiBench-HiBench-7.0/HiBench-HiBench-7.0/conf/workloads/micro/wordcount.conf中



我们将数据规模从tiny变为small，重复2、3、4步，比较两次测试情况。



Type Date Time Input\_data\_size Duration(s) Throughput(bytes/s) Throughput/node

HadoopSort 2020-04-28 13:28:56 36308 49.447 734 367

HadoopWordcount 2020-04-28 13:57:55 37962 37.963 999 499

HadoopWordcount 2020-04-28 14:20:23 328498144 144.926 2266661 1133330

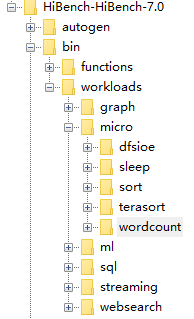
Duration：持续时间

Troughtput: 吞吐量

Troughtput/node: 吞吐量/节点

1. Hibench测试分类

在bin/workloads目录下：对应着不同的大数据测试点，比如sql、ml、graph、streaming等。

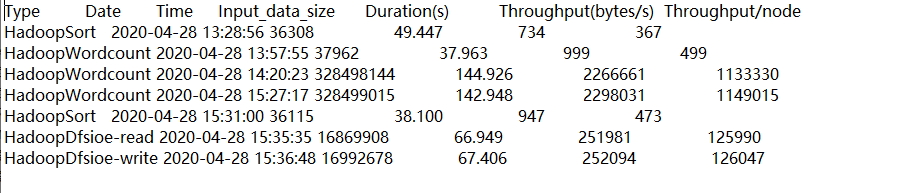


进入这些目录后：有对应不同的测试小项目，根据大数据平台需要，对其进行相应的测试。

HiBench共计19个测试方向，可大致分为6个测试类别：分别是micro，ml（机器学习），sql，graph，websearch和streaming。

* 1. Micro Benchmarks:

1. 排序（sort）  
   此工作负载对其文本输入数据进行排序，该数据是使用RandomTextWriter生成的。
2. 词频统计（wordcount）  
   此工作负载计算输入数据中每个单词的出现次数，这些单词是使用RandomTextWriter生成的。它代表了一种典型的MapReduce作业。
3. TeraSort（terasort）  
   TeraSort是由Jim Gray创建的标准基准测试。其输入数据由Hadoop TeraGen示例程序生成。
4. 休眠（sleep）  
   此工作负载在每个任务中休眠几秒钟以测试框架的调度性能。
5. 增强型DFSIO（dfsioe）  
   增强型DFSIO通过生成大量同时执行写入和读取的任务来测试Hadoop集群的HDFS吞吐量。它测量每个映射任务的平均I / O速率，每个映射任务的平均吞吐量以及HDFS集群的聚合吞吐量。注意：此基准测试没有Spark相应的实现



* 1. Machine Learning:

贝叶斯、K均值聚类、逻辑回归…

* 1. SQL

扫描（Scan），连接（Join），聚合（Aggregate）

* 1. Websearch benchmarks

网页排名、Nutch索引

参考：https://www.cnblogs.com/ratels/p/10978146.html