# 第三周报告　分布式HDFS搭建 MapReduce测试 wordcount

柳杰　2017-5-1

## 一、azure上分布式文件系统的搭建

**新建虚拟网路和虚拟机**

## 创建虚拟机

虚拟机配置如下表

| 名称（主机名） | 用户名 | 资源组 |
|---|---|---|
| master | hadoop | HADOOP_NET |
| slave01 | hadoop | HADOOP_NET |
| slave02 | hadoop | HADOOP_NET |
| Slave03 | hadoop | HADOOP_NET |
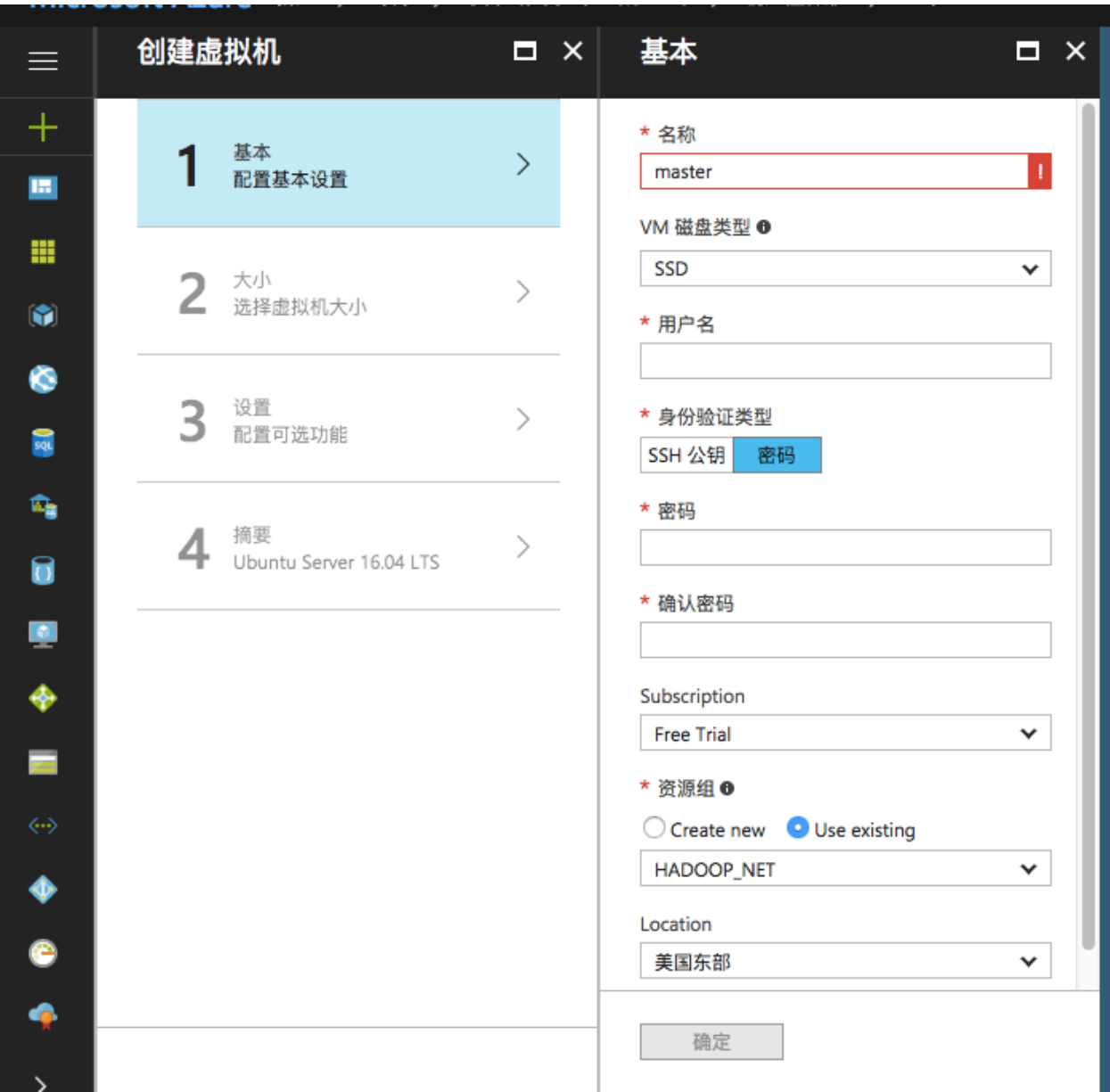


## 配置每个节点的/etc/hosts文件

在每个节点上

```
sudo vim /etc/hosts
```

添加如下内容

```
127.0.0.1 localhost

10.1.0.4 master

10.1.0.5 slave01

10.1.0.6 slave02

10.1.0.7 slav
```

## 配置master无密码登录slave节点

用

> ssh  localhost

在每台虚拟机上，测试能否连接到本地localhost，如果连不通，则需要安装openssh-server服务

> sudo apt-get install openssh-server

接着生成每台机器的公钥

```
ssh-keygen -t rsa -P ""

cat$HOME/.ssh/id_ras.pub >> $HOME/.ssh/authorized_keys
```

在保证了4台虚拟主机都能连接到本地localhost之后，还需要让master免密码登录slave01、slave02、slave03主机,在master上执行下面的命令，将master的id_rsa.pub传给三台slave虚拟主机

```
scp ~/.ssh/id_rsa.pub hadoop@slave01:/home/hadoop/

scp ~/.ssh/id_rsa.pub hadoop@slave02:/home/hadoop/

scp ~/.ssh/id_rsa.pub hadoop@slave03:/home/hadoop/
```

在slave01,slave02,slave03上将master公钥添加到各自的节点上，在slave01,slave02,slave03上执行如下的命令

```
cat ~/id_rsa.pub >> ~/.ssh/authorized_keys

rm ~/id_rsa.pub
```

在master上测试时候能免密码登录到slave

```
ssh slave01

ssh slave02

ssh slave03
```

## 安装JDK

执行命令

```
sudo apt-get install openjdk-8-jre
```

测试

```
java -version
```

在/etc/profle中配置环境变量

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export JRE_HOME=$JAVA_HOME/jre
export CLASSPATH=$JAVA_HOME/lib:$JRE_HOME/lib:$CLASSPATH
export PATH=$JAVA_HOME/bin:$JRE_HOME/bin:$PATH
```

## 安装Hadoop

[下载并解压hadoop安装包](#)

```
sudo tar -zxf hadoop-2.7.2.tar.gz -C /usr/local

cd /usr/local

sudo mv ./hadoop-2.7.2/ ./hadoop

sudo chown -R hadoop ./hadoop
```

修改master主机修改Hadoop的如下配置文件，这些配置文件都位于/usr/local/hadoop/etc/hadoop目录下

修改slaves

这里把DataNode的主机名写入该文件，每行一个。让master主机节点仅仅作为NameNode使用

```
vim slaves
slave01
slave02
slave03
```

配置core-site.xml

```xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
<name>hadoop.tmp.dir</name>
<value>file:/usr/local/hadoop/tmp</value>
<description>Abase for other temporary directories.</description>
</property>
<property>
<name>fs.defaultFS</name>
<value>hdfs://master:9000</value>
</property>

</configuration>
```

配置hdfs-site.xml

```xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
<name>dfs.replication</name>
<value>3</value>
</property>
<property>
<name>dfs.namenode.name.dir</name>
<value>file:/home/hadoop/tmp/dfs/name</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>file:/home/hadoop/tmp/dfs/data</value>
</property>

</configuration>
```

配置mapped-site.xml

```xml
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>

<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>

</configuration>
```

配置yarn-site.xml

```xml
<?xml version="1.0"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->
<configuration>

<!-- Site specific YARN configuration properties -->
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
<property>
<name>yarn.resourcemanager.hostname</name>
<value>master</value>
</property>

</configuration>
```

配置好后，将master节点上的/usr/local/hadoop 文件加复制到各个节点上。

```
cd /usr/local

tar -zcf ~/hadoop.master.tar.gz ./hadoop    #打包文件夹

cd ~



scp ./hadoop.master.tar.gz slave01:/home/hadoop

scp ./hadoop.master.tar.gz slave02:/home/hadoop

scp ./hadoop.master.tar.gz slave03:/home/hadoop



在slave01,slave02,slave03节点上执行



sudo tar -zxf ~/hadoop.master.tar.gz -C /usr/local

sudo chown -R hadoop /usr/local/hadoop
```

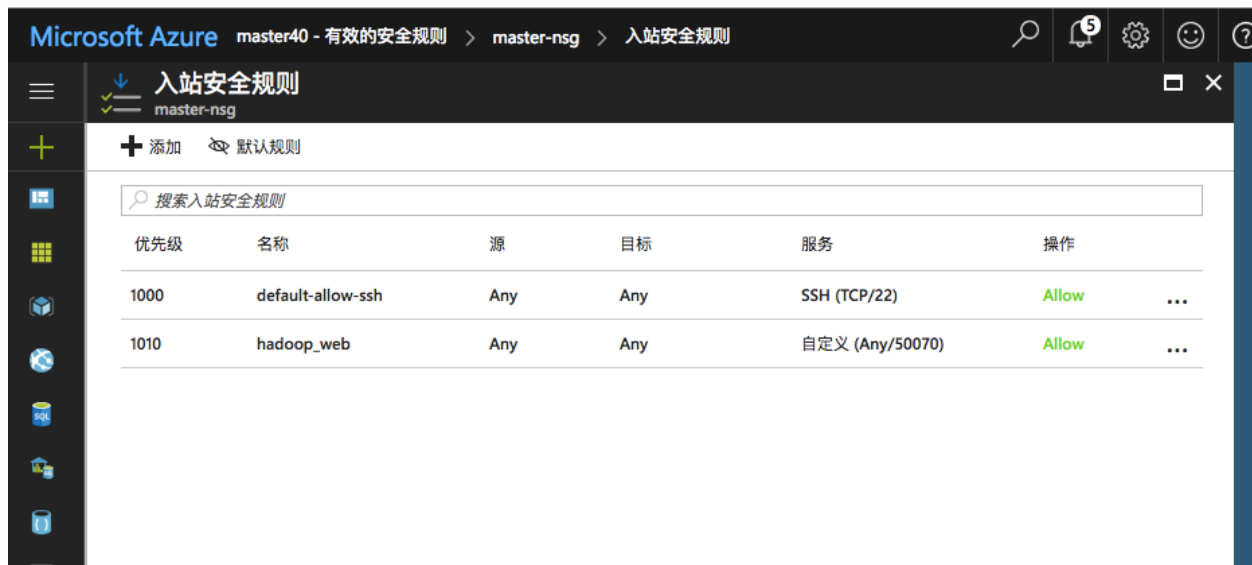## 启动Hadoop集群

登录master节点

```
cd /usr/local/hadoop

bin/hdfs namenode -format

sbin/start-all.sh



jps    #分别在各个节点查看运行状态
```
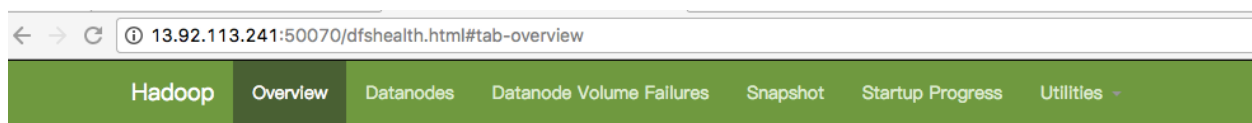
## 开启Hadoop集群的web监控页面

开放master50070端口访问限制，让外网可以web端查看文件系统

访问master：50070，可以通过web页面对hadoop集群监控，以及简单管理



# 二、MapReduce测试wordcount

登录master节点

在HDFS上创建存放测试文件的文件夹

```
bin/hdfs dfs -mkdir /testwordcount
```

上传测试文件到HDFS

```
bin/hdfs dfs -put LICENSE.txt /testwordcount
```

到hadoop项目根目录

```
hadoop share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.2.jar wordcount
/testwordcount /testwordcount_output
```

得到测试性能输出结果

```
17/05/01 15:07:33 INFO client.RMProxy: Connecting to ResourceManager at
master/10.1.0.4:8032
17/05/01 15:07:34 INFO input.FileInputFormat: Total input paths to process :
1
17/05/01 15:07:34 INFO mapreduce.JobSubmitter: number of splits:1
17/05/01 15:07:34 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1493384096713_0003
17/05/01 15:07:34 INFO impl.YarnClientImpl: Submitted application
application_1493384096713_0003
17/05/01 15:07:34 INFO mapreduce.Job: The url to track the job:
http://master:8088/proxy/application_1493384096713_0003/
17/05/01 15:07:34 INFO mapreduce.Job: Running job: job_1493384096713_0003
17/05/01 15:07:40 INFO mapreduce.Job: Job job_1493384096713_0003 running in
uber mode : false
17/05/01 15:07:40 INFO mapreduce.Job:  map 0% reduce 0%
17/05/01 15:07:44 INFO mapreduce.Job:  map 100% reduce 0%
17/05/01 15:07:49 INFO mapreduce.Job:  map 100% reduce 100%
17/05/01 15:07:49 INFO mapreduce.Job: Job job_1493384096713_0003 completed
successfully
17/05/01 15:07:49 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=10992
                FILE: Number of bytes written=256797
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=15538
                HDFS: Number of bytes written=8006
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=2162
                Total time spent by all reduces in occupied slots (ms)=1961
                Total time spent by all map tasks (ms)=2162
                Total time spent by all reduce tasks (ms)=1961
```

```
                Total vcore-milliseconds taken by all map tasks=2162
                Total vcore-milliseconds taken by all reduce tasks=1961
                Total megabyte-milliseconds taken by all map tasks=2213888
                Total megabyte-milliseconds taken by all reduce tasks=2008064
        Map-Reduce Framework
                Map input records=289
                Map output records=2157
                Map output bytes=22735
                Map output materialized bytes=10992
                Input split bytes=109
                Combine input records=2157
                Combine output records=755
                Reduce input groups=755
                Reduce shuffle bytes=10992
                Reduce input records=755
                Reduce output records=755
                Spilled Records=1510
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=96
                CPU time spent (ms)=1080
                Physical memory (bytes) snapshot=435122176
                Virtual memory (bytes) snapshot=3792809984
                Total committed heap usage (bytes)=328728576
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=15429
        File Output Format Counters
                Bytes Written=8006
```

测试结果输出

```
"AS      4
"Contribution"  1
"Contributor"   1
"Derivative     1
"Legal  1
"License"       1
"License");     1
"Licensor"      1
"NOTICE"        1
"Not    1
```

```
"Object"        1
"Source"        1
"Work"  1
"You"   1
"Your") 1
"[]"    1
"control"       1
"printed        1
"submitted"     1
(50%)   1
(C)     1
(Don't  1
(INCLUDING      2
(INCLUDING,     2
(a)     1
(an     1
(and    1
(b)     1
(c)     2
(d)     1
(except 1
(http://www.one-lab.org)        1
(http://www.opensource.org/licenses/bsd-license.php)    1
(i)     1
(ii)    1
(iii)   1
(including      3
(or     3
(such   1
(the    1
*       34
*/      3
-       7
/*      1
/**     2
034819  1
1       1
1.      1
2-Clause        1
2.      1
2.0     1
2.0,    1
2004    1
2005,   1
2008,2009,2010  1
2011-2014,      1
3.      1
4.      1
5.      1
```

```
6.      1
7.      1
8.      1
9       1
9.      1
:       3
A       3
ADVISED 2
AND     11
ANY     10
APACHE  1
APPENDIX:       1
ARE     2
ARISING 2
Accepting       1
Additional      1
All     2
Apache  5
Appendix        1
BASIS,  2
BE      2
BSD     1
BSD-style       1
BUSINESS        2
BUT     4
BY      2
CAUSED  2
CONDITIONS      4
CONSEQUENTIAL   2
CONTRACT,       2
CONTRIBUTORS    4
COPYRIGHT       4
CRC     1
Catholique      1
Collet. 1
Commission      1
Contribution    3
Contribution(s) 3
Contribution."  1
Contributions)  1
Contributions.  2
Contributor     8
Contributor,    1
Copyright       5
DAMAGE. 2
DAMAGES 2
DATA,   2
DIRECT, 2
DISCLAIMED.     2
```

```
DISTRIBUTION    1
Definitions.    1
Derivative      17
Disclaimer      1
END     1
EVEN    2
EVENT   2
EXEMPLARY,      2
EXPRESS 2
Entity  3
Entity" 1
European        1
FITNESS 3
FOR     6
Fast    1
File    1
For     6
GOODS   2
Grant   2
HADOOP  1
HOLDERS 2
HOWEVER 2
Hadoop  1
Header  1
How     1
However,        1
IF      2
IMPLIED 4
IN      6
INCIDENTAL,     2
INCLUDING,      2
INDIRECT,       2
INTERRUPTION)   2
IS      2
IS"     4
If      2
In      1
Institute       1
January 1
KIND,   2
LIABILITY,      4
LIABLE  2
LICENSE 1
LIMITED 4
LOSS    2
LZ      1
LZ4     3
Legal   3
Liability.      2
```

```
License 10
License,         6
License.        11
License;         1
Licensed         1
Licensor         8
Licensor,        1
Limitation       1
Louvain 1
MERCHANTABILITY 2
MERCHANTABILITY,          1
Massachusetts    1
NEGLIGENCE       2
NO      2
NON-INFRINGEMENT,         1
NOT     4
NOTICE  5
Neither 1
Notwithstanding 1
OF      19
ON      2
OR      18
OTHERWISE)       2
OUT     2
OWNER   2
Object  4
OneLab  1
PARTICULAR       3
POSSIBILITY      2
PROCUREMENT      2
PROFITS;         2
PROVIDED         2
PURPOSE 2
PURPOSE.         1
Patent  1
REPRODUCTION,    1
Redistribution  2
Redistribution. 1
Redistributions 4
SERVICES;        2
SHALL   2
SOFTWARE         2
SOFTWARE,        2
SPECIAL,         2
STRICT  2
SUBCOMPONENTS:   1
SUBSTITUTE       2
SUCH    2
Sections         1
```

```
See       1
Source    8
Subject   2
Submission        1
TERMS     2
THE       10
THEORY    2
THIS      4
TITLE,    1
TO,       4
TORT      2
Technology.       1
The       3
This      1
To        1
Trademarks.       1
UCL       1
USE       2
USE,      3
University        1
Unless    3
Use       1
Version   2
WARRANTIES        4
WARRANTIES,       2
WAY       2
WHETHER   2
WITHOUT   2
Warranty          1
Warranty.         1
We        1
While     1
Work      20
Work,     4
Work.     1
Works     12
Works"    1
Works,    2
Works;    3
Yann      1
You       24
Your      9
[name     1
[yyyy]    1
a         21
above     4
above,    1
acceptance        1
accepting         2
```

```
act       1
acting   1
acts)    1
add       2
addendum         1
additional       4
additions        1
advised 1
against 1
against,         1
agree    1
agreed   3
agreement        1
algorithm        1
all       3
alleging         1
alone    1
along    1
alongside        1
also     1
an        6
and       51
and/or   3
annotations,     1
any       28
appear. 1
applicable       3
applies 1
apply    2
appropriate      1
appropriateness 1
archives.        1
are       10
arising 1
as        15
asserted         1
associated       1
assume   1
at        3
attach   1
attached         1
attribution      4
author   1
authorized       2
authorship,      2
authorship.      1
available        1
based    1
be        7
```

```
been    2
behalf  5
below). 1
beneficial      1
binary  4
bind    1
boilerplate     1
brackets        1
brackets!)      1
but     5
by      21
by,     3
calculation     1
can     2
cannot  1
carry   1
cause   2
changed 1
character       1
charge  1
choose  1
claims  2
class   1
classes:        1
code    5
code,   2
combination     1
comment 1
commercial      1
common  1
communication   3
compiled        1
compliance      1
complies        1
compression     1
computer        1
conditions      14
conditions.     1
conditions:     1
configuration   1
consequential   1
consistent      1
conspicuously   1
constitutes     1
construed       1
contact 1
contained       1
contains        1
content 1
```

```
contents        1
contract        2
contract,       1
contributors    1
contributory    1
control 2
control,        1
controlled      1
conversions     1
copies  1
copy    3
copyright       15
copyright,      1
counterclaim    1
cross-claim     1
customary       1
damages 3
damages,        1
damages.        1
date    1
de      1
defend, 1
defined 1
definition,     2
deliberate      1
derived 2
describing      1
description     1
designated      1
determining     1
different       1
direct  2
direct, 1
direction       1
disclaimer      2
disclaimer.     2
discussing      1
display 1
display,        1
distribute      3
distribute,     2
distributed     3
distribution    3
distribution,   1
distribution.   2
do      3
document.       1
documentation   3
documentation,  2
```

```
does     1
each     4
easier   1
editorial        1
either   2
elaborations,    1
electronic       1
electronic,      1
enclosed         2
endorse 1
entities         1
entity   3
entity, 1
entity. 2
even     1
event    1
example 1
except   2
excluding        3
executed         1
exercise         1
exercising       1
explicitly       1
express 2
failure 1
fee      1
fields  1
fifty    1
file     6
file,    1
file.    2
filed.   1
files    1
files.   1
files;   1
following        10
for      19
for,     1
form     10
form,    4
form.    1
format. 1
forms,  2
forum   1
found    1
from     4
from)    1
from,    1
generated        2
```

```
give      1
goodwill,         1
governed          1
governing         1
grant     1
granted 2
granting          1
grants  2
grossly 1
harmless          1
has       2
have      2
hereby  2
herein  1
hold      1
http://code.google.com/p/lz4/     1
http://www.apache.org/licenses/ 1
http://www.apache.org/licenses/LICENSE-2.0        1
https://groups.google.com/forum/#!forum/lz4c      1
identification    1
identifying       1
if        4
implementation  1
implied,          1
implied.          1
import, 1
improving         1
in        31
inability         1
incidental,       1
include 3
included          2
includes          1
including         5
including,        1
inclusion         2
incorporated      2
incurred          1
indemnify,        1
indemnity,        1
indicated         1
indirect,         2
individual        3
information.      1
informational    1
infringed         1
infringement,     1
institute         1
intentionally     2
```

```
interfaces      1
irrevocable     2
is      10
issue   1
its     4
language        1
law     3
lawsuit)        1
least   1
legal   1
liability       2
liability.      1
liable  1
licensable      1
license 7
licenses        1
licenses.       1
limitation,     1
limitations     1
limited 4
link    1
list    4
lists,  1
litigation      2
loss    1
losses),        1
made    1
made,   1
mailing 1
make,   1
making  1
malfunction,    1
managed 1
management      1
marked  1
marks,  1
materials       2
may     10
mean    10
means   2
mechanical      1
media   1
medium, 1
meet    1
merely  1
met:    2
modification,   2
modifications   3
modifications,  3
```

```
modified        1
modify  2
modifying       1
more    1
must    8
name    2
name)   1
names   2
names,  1
native  1
necessarily     1
negligence),    1
negligent       1
no      2
no-charge,      2
non-exclusive,  2
nor     1
normally        1
not     11
nothing 1
notice  2
notice, 5
notices 9
object  1
obligations     1
obligations,    1
obtain  1
of      75
of,     3
offer   1
offer,  1
on      11
one     1
only    4
or      65
or,     1
org.apache.hadoop.util.bloom.*  1
origin  1
original        2
other   9
otherwise       3
otherwise,      3
out     1
outstanding     1
own     4
owner   4
owner.  1
owner]  1
ownership       2
```

```
page"    1
part     4
patent   5
patent,  1
percent  1
perform,          1
permission        1
permission.       1
permissions       3
permitted         2
perpetual,        2
pertain  2
places:  1
portions          1
possibility       1
power,   1
preferred         1
prepare  1
prior    1
product  1
products          1
project  2
prominent         1
promote  1
provide  1
provided          9
provides          2
public   1
publicly          2
purpose  2
purposes          4
readable          1
reason   1
reasonable        1
received          1
recipients        1
recommend         1
redistributing  2
regarding         1
remain   1
replaced          1
repository        1
represent,        1
representatives,          1
reproduce         3
reproduce,        1
reproducing       1
reproduction,   3
required          4
```

```
reserved.        2
responsibility,  1
responsible      1
result  1
resulting        1
retain  2
retain, 1
revisions,       1
rights  3
risks   1
royalty-free,    2
same    1
section)         1
sell,   2
sent    1
separable        1
separate         2
service 1
shall   15
shares, 1
should  1
slicing-by-8     1
software         3
sole    1
solely  1
source  9
source, 1
special,         1
specific         2
src/main/native/src/org/apache/hadoop/io/compress/lz4/{lz4.h,lz4.c,lz4hc.h,lz
4hc.c}, 1
src/main/native/src/org/apache/hadoop/util:      1
state   1
stated  2
statement        1
stating 1
stoppage,        1
subcomponents    2
subject 1
sublicense,      1
submit  1
submitted        2
submitted.       1
subsequently     1
such    17
supersede        1
support,         1
syntax  1
systems 1
```

```
systems,        1
terminate       1
terms   8
terms.  1
text    4
that    25
the     122
their   2
then    2
theory, 1
thereof 1
thereof,        2
thereof.        1
these   1
third-party     2
this    22
those   3
through 1
to      41
tort    1
tracking        1
trade   1
trademark,      1
trademarks,     1
transfer        1
transformation  1
translation     1
types.  1
under   10
union   1
unless  1
use     8
use,    4
used    1
using   1
verbal, 1
version 1
warranties      1
warranty        1
warranty,       1
was     1
where   1
wherever        1
whether 4
which   2
whole,  2
whom    1
with    16
within  8
```

```
without 6
work    5
work,   2
work.   1
works   1
worldwide,      2
writing 1
writing,        3
written 2
you     2
your    4
```

# 三、下周计划

**学习并搭建分布式redis**

**学习并大家分布式数据库HBase**