

# MapReduce 课程设计数据说明

2020.6

HDFS 路径: /MP\_Data/

```
(base) [root@master001 MP_Data]# hdfs dfs -ls /MP_Data/
Found 4 items
-rw-r--r--   3 root supergroup      15184 2020-06-04 00:02 /MP_Data/Stop_words.txt
drwxr-xr-x   - root supergroup         0 2020-06-04 00:06 /MP_Data/task1
drwxr-xr-x   - root supergroup         0 2020-06-04 00:07 /MP_Data/task2
drwxr-xr-x   - root supergroup         0 2020-06-04 00:39 /MP_Data/task3
```

Task1:/MP\_Data/task1/JN1\_LOG

```
(base) [root@master001 MP_Data]# hdfs dfs -ls /MP_Data/task1/JN1_LOG
Found 15 items
-rw-r--r--   3 root supergroup 970797265 2020-06-04 00:03 /MP_Data/task1/JN1_LOG/2015-09-08.log
-rw-r--r--   3 root supergroup 985767567 2020-06-04 00:03 /MP_Data/task1/JN1_LOG/2015-09-09.log
-rw-r--r--   3 root supergroup 894144556 2020-06-04 00:03 /MP_Data/task1/JN1_LOG/2015-09-10.log
-rw-r--r--   3 root supergroup 943822968 2020-06-04 00:03 /MP_Data/task1/JN1_LOG/2015-09-11.log
-rw-r--r--   3 root supergroup 845005268 2020-06-04 00:03 /MP_Data/task1/JN1_LOG/2015-09-12.log
-rw-r--r--   3 root supergroup 784354256 2020-06-04 00:04 /MP_Data/task1/JN1_LOG/2015-09-13.log
-rw-r--r--   3 root supergroup 917563007 2020-06-04 00:04 /MP_Data/task1/JN1_LOG/2015-09-14.log
-rw-r--r--   3 root supergroup 718757505 2020-06-04 00:04 /MP_Data/task1/JN1_LOG/2015-09-15.log
-rw-r--r--   3 root supergroup 758602804 2020-06-04 00:04 /MP_Data/task1/JN1_LOG/2015-09-16.log
-rw-r--r--   3 root supergroup 739751283 2020-06-04 00:05 /MP_Data/task1/JN1_LOG/2015-09-17.log
-rw-r--r--   3 root supergroup 716302112 2020-06-04 00:05 /MP_Data/task1/JN1_LOG/2015-09-18.log
-rw-r--r--   3 root supergroup 677571206 2020-06-04 00:05 /MP_Data/task1/JN1_LOG/2015-09-19.log
-rw-r--r--   3 root supergroup 641661626 2020-06-04 00:06 /MP_Data/task1/JN1_LOG/2015-09-20.log
-rw-r--r--   3 root supergroup 716627831 2020-06-04 00:06 /MP_Data/task1/JN1_LOG/2015-09-21.log
-rw-r--r--   3 root supergroup 741494732 2020-06-04 00:06 /MP_Data/task1/JN1_LOG/2015-09-22.log
```

Task2:/MP\_Data/task2/

```
(base) [root@master001 MP_Data]# hdfs dfs -ls /MP_Data/task2/
Found 2 items
-rw-r--r--   3 root supergroup      13061 2020-06-04 00:06 /MP_Data/task2/People_List_unique.txt
drwxr-xr-x   - root supergroup         0 2020-06-04 00:51 /MP_Data/task2/wuxia_novels
(base) [root@master001 MP_Data]# hdfs dfs -ls /MP_Data/task2/wuxia_novels
Found 218 items
-rw-r--r--   3 root supergroup 1428907 2020-06-04 00:06 /MP_Data/task2/wuxia_novels/卧龙生01.镖旗.txt.segmented
-rw-r--r--   3 root supergroup 1868226 2020-06-04 00:06 /MP_Data/task2/wuxia_novels/卧龙生02.春秋笔.txt.segmented
-rw-r--r--   3 root supergroup 2462011 2020-06-04 00:06 /MP_Data/task2/wuxia_novels/卧龙生03.翠袖玉环.txt.segmented
-rw-r--r--   3 root supergroup 975882 2020-06-04 00:06 /MP_Data/task2/wuxia_novels/卧龙生04.地狱门.txt.segmented
-rw-r--r--   3 root supergroup 1199167 2020-06-04 00:06 /MP_Data/task2/wuxia_novels/卧龙生05.飞花逐月.txt.segmented
-rw-r--r--   3 root supergroup 2100512 2020-06-04 00:06 /MP_Data/task2/wuxia_novels/卧龙生06.飞铃.txt.segmented
-rw-r--r--   3 root supergroup 2421644 2020-06-04 00:06 /MP_Data/task2/wuxia_novels/卧龙生07.飞燕惊龙.txt.segmented
-rw-r--r--   3 root supergroup 2614016 2020-06-04 00:06 /MP_Data/task2/wuxia_novels/卧龙生08.风尘侠隐.txt.segmented
-rw-r--r--   3 root supergroup 1825850 2020-06-04 00:06 /MP_Data/task2/wuxia_novels/卧龙生09.风雨燕归来.txt.segmented
-rw-r--r--   3 root supergroup 1304751 2020-06-04 00:06 /MP_Data/task2/wuxia_novels/卧龙生10.黑白剑.txt.segmented
-rw-r--r--   3 root supergroup 995450 2020-06-04 00:06 /MP_Data/task2/wuxia_novels/卧龙生11.黑白双娇.txt.segmented
-rw-r--r--   3 root supergroup 1544201 2020-06-04 00:06 /MP_Data/task2/wuxia_novels/卧龙生12.剑气洞彻九重天.txt.segmented
-rw-r--r--   3 root supergroup 1385236 2020-06-04 00:06 /MP_Data/task2/wuxia_novels/卧龙生13.剑无痕.txt.segmented
-rw-r--r--   3 root supergroup 1041321 2020-06-04 00:06 /MP_Data/task2/wuxia_novels/卧龙生14.剑仙列传.txt.segmented
-rw-r--r--   3 root supergroup 2294923 2020-06-04 00:06 /MP_Data/task2/wuxia_novels/卧龙生15.绛雪玄霜.txt.segmented
-rw-r--r--   3 root supergroup 2278826 2020-06-04 00:06 /MP_Data/task2/wuxia_novels/卧龙生16.金笔点龙记.txt.segmented
-rw-r--r--   3 root supergroup 1584683 2020-06-04 00:06 /MP_Data/task2/wuxia_novels/卧龙生17.金凤剪.txt.segmented
-rw-r--r--   3 root supergroup 2303736 2020-06-04 00:06 /MP_Data/task2/wuxia_novels/卧龙生18.金剑雕翎.txt.segmented
-rw-r--r--   3 root supergroup 2618881 2020-06-04 00:06 /MP_Data/task2/wuxia_novels/卧龙生19.惊鸿一剑震江湖.txt.segmented
-rw-r--r--   3 root supergroup 1015204 2020-06-04 00:06 /MP_Data/task2/wuxia_novels/卧龙生20.梦幻之刀.txt.segmented
-rw-r--r--   3 root supergroup 1054373 2020-06-04 00:06 /MP_Data/task2/wuxia_novels/卧龙生21.妙绝天香.txt.segmented
```

### Task3: /MP\_Data/task3/news20/20\_newsgroup

```
(base) [root@master001 MP_Data]# hdfs dfs -ls /MP_Data/task3/news20/20_newsgroup
Found 20 items
drwxr-xr-x - root supergroup 0 2020-06-04 00:08 /MP_Data/task3/news20/20_newsgroup/alt.atheism
drwxr-xr-x - root supergroup 0 2020-06-04 00:10 /MP_Data/task3/news20/20_newsgroup/comp.graphics
drwxr-xr-x - root supergroup 0 2020-06-04 00:11 /MP_Data/task3/news20/20_newsgroup/comp.os.ms-windows.misc
drwxr-xr-x - root supergroup 0 2020-06-04 00:13 /MP_Data/task3/news20/20_newsgroup/comp.sys.ibm.pc.hardware
drwxr-xr-x - root supergroup 0 2020-06-04 00:15 /MP_Data/task3/news20/20_newsgroup/comp.sys.mac.hardware
drwxr-xr-x - root supergroup 0 2020-06-04 00:16 /MP_Data/task3/news20/20_newsgroup/comp.windows.x
drwxr-xr-x - root supergroup 0 2020-06-04 00:18 /MP_Data/task3/news20/20_newsgroup/misc.forsale
drwxr-xr-x - root supergroup 0 2020-06-04 00:20 /MP_Data/task3/news20/20_newsgroup/rec.autos
drwxr-xr-x - root supergroup 0 2020-06-04 00:21 /MP_Data/task3/news20/20_newsgroup/rec.motorcycles
drwxr-xr-x - root supergroup 0 2020-06-04 00:23 /MP_Data/task3/news20/20_newsgroup/rec.sport.baseball
drwxr-xr-x - root supergroup 0 2020-06-04 00:25 /MP_Data/task3/news20/20_newsgroup/rec.sport.hockey
drwxr-xr-x - root supergroup 0 2020-06-04 00:26 /MP_Data/task3/news20/20_newsgroup/sci.crypt
drwxr-xr-x - root supergroup 0 2020-06-04 00:28 /MP_Data/task3/news20/20_newsgroup/sci.electronics
drwxr-xr-x - root supergroup 0 2020-06-04 00:29 /MP_Data/task3/news20/20_newsgroup/sci.med
drwxr-xr-x - root supergroup 0 2020-06-04 00:31 /MP_Data/task3/news20/20_newsgroup/sci.space
drwxr-xr-x - root supergroup 0 2020-06-04 00:33 /MP_Data/task3/news20/20_newsgroup/soc.religion.christian
drwxr-xr-x - root supergroup 0 2020-06-04 00:34 /MP_Data/task3/news20/20_newsgroup/talk.politics.guns
drwxr-xr-x - root supergroup 0 2020-06-04 00:36 /MP_Data/task3/news20/20_newsgroup/talk.politics.mideast
drwxr-xr-x - root supergroup 0 2020-06-04 00:38 /MP_Data/task3/news20/20_newsgroup/talk.politics.misc
drwxr-xr-x - root supergroup 0 2020-06-04 00:39 /MP_Data/task3/news20/20_newsgroup/talk.religion.misc
```

集群本地路径(备份): /home/MP\_Data