暑期实训学习报告

1813012　柳鑫

一．学习内容

了解 MapReduce 以及 Hadoop 分布式文件系统（HDFS）

二．学习结果

① MapReduce

MapReduce 是一种编程模型，用于大规模数据集（大于 1TB）的并行运算。概念"Map（映射）"和"Reduce（归约）"，是它们的主要思想，都是从函数式编程语言里借来的，还有从矢量编程语言里借来的特性。它极大地方便了编程人员在不会分布式并行编程的情况下，将自己的程序运行在分布式系统上。 当前的软件实现是指定一个 Map（映射）函数，用来把一组键值对映射成一组新的键值对，指定并发的 Reduce（归约）函数，用来保证所有映射的键值对中的每一个共享相同的键组。

② Hadoop 分布式文件系统（HDFS）

Hadoop 分布式文件系统(HDFS)是指被设计成适合运行在通用硬件(commodity hardware)上的分布式文件系统（Distributed File System）。它和现有的分布式文件系统有很多共同点。但同时，它和其他的分布式文件系统的区别也是很明显的。HDFS 是一个高度容错性的系统，适合部署在廉价的机器上。HDFS 能提供高吞吐量的数据访问，非常适合大规模数据集上的应用。HDFS 放宽了一部分 POSIX 约束，来实现流式读取文件系统数据的目的。HDFS 在最开始是作为 Apache Nutch 搜索引擎项目的基础架构而开发的。HDFS 是 Apache Hadoop Core 项目的一部分。

HDFS 有着高容错性（fault-tolerant）的特点，并且设计用来部署在低廉的（low-cost）硬件上。而且它提供高吞吐量（high throughput）来访问应用程序的数据，适合那些有着超大数据集（large data set）的应用程序。HDFS 放宽了（relax）POSIX 的要求（requirements）这样可以实现流的形式访问（streaming access）文件系统中的数据。

```
[hadoop@localhost ~]$ hdfs dfs -mkdir /input        # 在HDFS 根目录下创建 input 目录
[hadoop@localhost ~]$ hdfs dfs -mkdir /output       # 在HDFS 根目录下创建 output 目录
[hadoop@localhost ~]$ hdfs dfs -ls /                # 查看HDFS 根目录下文件列表
Found 4 items
drwxr-xr-x   - hadoop supergroup          0 2020-02-13 15:01 /input
drwxr-xr-x   - hadoop supergroup          0 2020-02-13 14:57 /output
drwx------   - hadoop supergroup          0 2020-02-13 14:49 /tmp
drwxr-xr-x   - hadoop supergroup          0 2020-02-13 14:55 /user
# 将本地主机上的 hadoop 文件夹上传到 HDFS 的 input 目录
[hadoop@localhost ~]$ hdfs dfs -put /usr/local/hadoop-3.2.1/etc/hadoop /input
[hadoop@localhost ~]$ hdfs dfs -ls /input/hadoop    # 查看HDFS 中/input/hadoop 目录下文件列表
Found 32 items
-rw-r--r--   1 hadoop supergroup       8260 2020-02-13 15:13 /input/hadoop/capacity-scheduler.xml
-rw-r--r--   1 hadoop supergroup       1335 2020-02-13 15:13 /input/hadoop/configuration.xsl
# 此处省略其余 30 个文件信息
# 使用 hadoop 提供的mapreduce 示例程序，统计上述 32 个文件中匹配'dfs[a-z.]+'的字符串出现的次数
[hadoop@localhost ~]$ hadoop jar /usr/local/hadoop-3.2.1/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.2.1.jar grep /input/hadoop /output/demo 'dfs[a-z.]+'
[hadoop@localhost ~]$ hdfs dfs -ls /output/demo   # 查看结果存放目录/output/demo 下的文件列表
Found 2 items
-rw-r--r--   1 hadoop supergroup          0 2020-02-13 15:19 /output/demo/_SUCCESS
-rw-r--r--   1 hadoop supergroup        331 2020-02-13 15:19 /output/demo/part-r-00000
# 你可以将结果文件目录从 HDFS 拷贝到本地主机的 /tmp/ 目录下
[hadoop@localhost ~]$ hdfs dfs -get /output/demo /tmp/
[hadoop@localhost ~]$ ls /tmp/demo/
part-r-00000  _SUCCESS
[hadoop@localhost ~]$ cat /tmp/demo/part-r-00000
5 dfs.audit.logger
3 dfs.logger
3 dfs.server.namenode.
2 dfs.audit.log.maxbackupindex
2 dfs.audit.log.maxfilesize
2 dfs.sh
1 dfshealth.html
1 dfsadmin
1 dfs.replication
1 dfs.namenode.servicerpc
1 dfs.namenode.rpc
1 dfs.namenode.name.dir
```

```
1 dfs.log
1 dfs.http.address
1 dfs.datanode.data.dir
1 dfs.namenode.ec.policies.max.cellsize
# 或者直接查看HDFS中mapreduce任务的统计结果
[hadoop@localhost ~]$ hdfs dfs -cat /output/demo/part-r-00000
# mapreduce任务执行完毕后的的统计结果内容如下:
5 dfs.audit.logger
3 dfs.logger
3 dfs.server.namenode.
2 dfs.audit.log.maxbackupindex
2 dfs.audit.log.maxfilesize
2 dfs.sh
1 dfshealth.html
1 dfsadmin
1 dfs.replication
1 dfs.namenode.servicerpc
1 dfs.namenode.rpc
1 dfs.namenode.name.dir
1 dfs.log
1 dfs.http.address
1 dfs.datanode.data.dir
1 dfs.namenode.ec.policies.max.cellsize
```