setup Hadoop 2.7.1 single node cluster on CentOS

Configure Hadoop Pseudo-Distributed Mode

Hadoop可以在单节点上以所谓的伪分布式模式运行，此时每一个Hadoop守护进程都作为一个独立的Java进程运行。

Cd/usr/local/hadoop

接着开启 NaneNode 和 DataNode 守护进程：

./sbin/start-dfs.sh

成功启动后，可以访问 Web 界面 http://localhost:50070 查看 NameNode 和 Datanode 信息，还可以在线查看 HDFS 中的文件。

浏览NameNode和JobTracker的网络接口，它们的地址默认为：

* NameNode - <http://localhost:50070/>
* JobTracker - <http://localhost:50030/>

运行程序时，输出目录不能存在

运行 Hadoop 程序时，为了防止覆盖结果，程序指定的输出目录（如 output）不能存在，否则会提示错误，因此运行前需要先删除输出目录。在实际开发应用程序时，可考虑在程序中加上如下代码，能在每次运行时自动删除输出目录，避免繁琐的命令行操作：

要使用 HDFS，首先需要在 HDFS 中创建用户目录：

./bin/hdfs dfs -mkdir -p /user/hadoop

Note that you can also run Spark locally (possibly on multiple cores) without any special setup by just passing local[N] as the master URL, where N is the number of parallel threads you want.