1 系统配置

JDK8-191, hadoop-2.9.2, Maven, spark-2.4.0, scala-2.11, python-3.6.7
配置步骤： System:
(1). 安装 JDK8，安装包：jdk-8u191-linux-x64.tar.gz, 安装目录：/usr/lib/jdk/jdk1.8
(2). 安装 Hadoop2.9.2，安装包：hadoop-2.9.2.tar.gz, 安装目录：/usr/local/Hadoop
(3). 安装 sshd, sh:> sudo apt-get install openssh-server, Hadoop、ssh 配置详见：[1]
(4). intellij setup, package: ideaIU-2018.3.tar.gz
(5). scala-2.11.12, package: scala-2.11.12-bin-....tgz
将/etc/profile 文件中加入以下内容：

```
 /etc/profile
# JDK1.8
export JAVA_HOME=/usr/lib/jdk/jdk1.8
export JRE_HOME=${JAVA_HOME}/jre
export CLASSPATH=.:${JAVA_HOME}/lib:${JRE_HOME}/lib
export PATH=.:${JAVA_HOME}/bin:$PATH
# hadoop2.9.2
export HADOOP_HOME=/usr/local/hadoop
export CLASSPATH=$($HADOOP_HOME/bin/hadoop classpath):$CLASSPATH
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
export PATH=.:${JAVA_HOME}/bin:${HADOOP_HOME}/bin:$PATH
# intellij
export IDEA_HOME=/usr/local/intellij/bin
# scala 2.12.7
export SCALA_HOME=/usr/local/scala/scala-2.11.12
export PATH=$PATH:$SCALA_HOME/bin
# spark
export SPARK_HOME=/usr/local/spark
export PATH=${SPARK_HOME}/bin:$PATH
```

2 项目配置

Project: spark-scala
shell
sudo mkdir IdeaProjects
将 spark-scala 项目移入 IdeaProjects，运行 intellij，在 pom.xml 文件中运行 maven reimport， 会出现 Scala 内核无法找到，但 spark 的文件均配置好的情况，此时，请将项目附带文件下的 scalalibrary 文件夹中的 modules、scala-library/2.11、scala-compiler/2.11、scala-reflect/2.11 文

件夹复 制到 Ubuntu 系统文件夹下的~/.m2/repository/org/scala-lang/下的 modules、scala-library、scalacompiler 、scala-reflect 文件夹下，若 intellij 中 pom.xml 无错误 import 问题，则配置成功。

数据文件位于项目文件夹下的 sogou.500w 文件夹下，未考出，以防占用空间过多。

参考文献 References

[1] Hadoop setup. https://www.cnblogs.com/87hbteo/p/7606012.html.