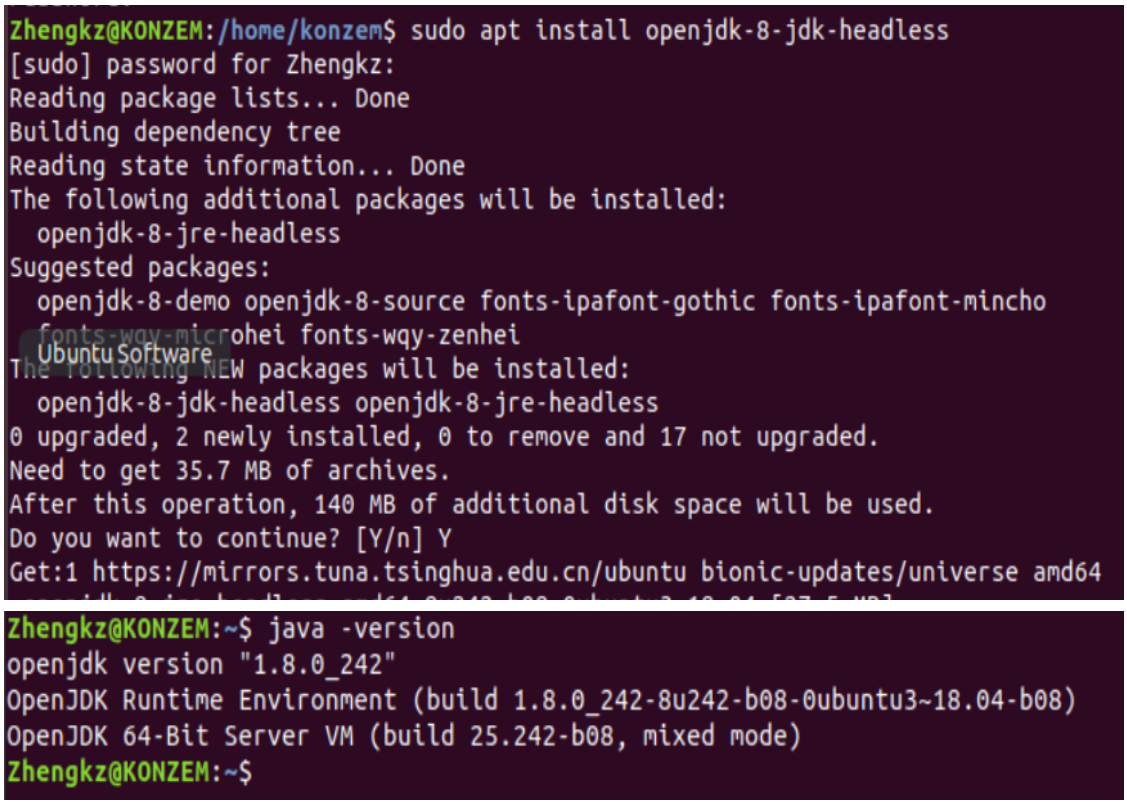
学院：数据科学与计算机学院　　专业：计算机科学与技术
姓名：郑康泽　　学号：17341213

云计算项目实践

Spark安装部署与基础实践

一. 实验过程截图汇总

1. Java安装



2. Spark下载及测试

```
Zhengkz@KONZEM:/home/konzem$ sudo wget http://mirror.bit.edu.cn/apache/spark/sp
ark-2.4.5/spark-2.4.5-bin-hadoop2.7.tgz
--2020-03-19 14:46:10--  http://mirror.bit.edu.cn/apache/spark/spark-2.4.5/spar
k-2.4.5-bin-hadoop2.7.tgz
Resolving mirror.bit.edu.cn (mirror.bit.edu.cn)... 219.143.204.117, 202.204.80.
77, 2001:da8:204:1205::22
Connecting to mirror.bit.edu.cn (mirror.bit.edu.cn)|219.143.204.117|:80... conn
ected.
HTTP request sent, awaiting response... 200 OK
Length: 232530699 (222M) [application/octet-stream]
Saving to: ‘spark-2.4.5-bin-hadoop2.7.tgz’

spark-2.4.5-bin-had 100%[==================>] 221.76M   881KB/s    in 4m 55s

2020-03-19 14:51:05 (770 KB/s) - ‘spark-2.4.5-bin-hadoop2.7.tgz’ saved [2325306
99/232530699]
```

```
20/03/19 14:59:57 INFO DAGScheduler: ResultStage 0 (reduce at SparkPi.scala:38)
 finished in 1.823 s
20/03/19 14:59:57 INFO DAGScheduler: Job 0 finished: reduce at SparkPi.scala:38
, took 2.044309 s
Pi is roughly 3.141255141255141
20/03/19 14:59:57 INFO SparkUI: Stopped Spark web UI at http://192.168.87.129:4
040
20/03/19 14:59:57 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEn
dpoint stopped!
20/03/19 14:59:57 INFO MemoryStore: MemoryStore cleared
20/03/19 14:59:57 INFO BlockManager: BlockManager stopped
20/03/19 14:59:57 INFO BlockManagerMaster: BlockManagerMaster stopped
20/03/19 14:59:57 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint:
 OutputCommitCoordinator stopped!
20/03/19 14:59:57 INFO SparkContext: Successfully stopped SparkContext
20/03/19 14:59:57 INFO ShutdownHookManager: Shutdown hook called
20/03/19 14:59:57 INFO ShutdownHookManager: Deleting directory /tmp/spark-e19ad
b63-99b5-469c-945a-09f0adf87f98
20/03/19 14:59:57 INFO ShutdownHookManager: Deleting directory /tmp/spark-9a110
9e2-45c8-4856-82dc-8a3f1257a80b
Show Applications
Zhengkz@KONZEM:/home/konzem/spark$
```

3. 计算Top N

```
Zhengkz@KONZEM:/home/konzem/spark$ ./bin/spark-shell
20/03/19 15:28:33 WARN Utils: Your hostname, KONZEM resolves to a loopback addr
ess: 127.0.1.1; using 192.168.87.129 instead (on interface ens33)
20/03/19 15:28:33 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
 address
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/h
ome/konzem/spark/jars/spark-unsafe_2.11-2.4.5.jar) to method java.nio.Bits.unal
igned()
WARNING: Please consider reporting this to the maintainers of org.apache.spark.
unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflec
tive access operations
WARNING: All illegal access operations will be denied in a future release
20/03/19 15:28:34 WARN NativeCodeLoader: Unable to load native-hadoop library f
or your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLev
el(newLevel).
Spark context Web UI available at http://192.168.87.129:4040
Spark context available as 'sc' (master = local[*], app id = local-158460292870
3).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
     \ \/ _ \/ _ `/ __/  '_/
```

```
scala> val textFile = sc.textFile("/home/Zhengkz/number.txt")
textFile: org.apache.spark.rdd.RDD[String] = /home/Zhengkz/number.txt MapPartit
ionsRDD[1] at textFile at <console>:24

scala> val nums = textFile.flatMap(line => line.split(" "))
nums: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[2] at flatMap at <con
sole>:25

scala> val nums_map= nums.map(x => (x.toInt, null))
nums_map: org.apache.spark.rdd.RDD[(Int, Null)] = MapPartitionsRDD[3] at map at
 <console>:25

scala> val sorted_nums_map = nums_map.sortByKey(false)
sorted_nums_map: org.apache.spark.rdd.RDD[(Int, Null)] = ShuffledRDD[4] at sort
ByKey at <console>:25

scala> val sorted_nums = sorted_nums_map.map(_._1)
sorted_nums: org.apache.spark.rdd.RDD[Int] = MapPartitionsRDD[5] at map at <con
sole>:25

scala> val top_10 = sorted_nums.take(10)
[Stage 0:>                                                          (0 + 1) / 1

top_10: Array[Int] = Array(100, 99, 98, 97, 96, 95, 94, 93, 92, 91)

scala> top_10.foreach(println)
100
99
98
```

4. word count

```
scala> val textFile = sc.textFile("README.md")
textFile: org.apache.spark.rdd.RDD[String] = README.md MapPartitionsRDD[11] at
textFile at <console>:24

scala> val words = textFile.flatMap(line => line.split(" "))
words: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[12] at flatMap at <c
onsole>:25

scala> val ones = words.map(w => (w, 1))
ones: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[13] at map at
<console>:25

scala> val counts = ones.reduceByKey(_ + _)
counts: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[14] at reduceByKe
y at <console>:25

scala> counts.foreach(println)
(package,1)
(For,3)
(Programs,1)
(processing.,1)
(Because,1)
(The,1)
(page](http://spark.apache.org/documentation.html).,1)
(cluster.,1)
(its,1)
([run,1)
```

5. RDD编程实践

```
scala> val rdd = sc.makeRDD(1 to 5, 3)
rdd: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[15] at makeRDD at <c
onsole>:24

scala> rdd.collect
res2: Array[Int] = Array(1, 2, 3, 4, 5)

scala> rdd.map(_ * 2).collect
res3: Array[Int] = Array(2, 4, 6, 8, 10)

scala> val rdd1 = rdd.flatMap(x => (1 to x))
rdd1: org.apache.spark.rdd.RDD[Int] = MapPartitionsRDD[17] at flatMap at <conso
le>:25

scala> rdd.collect
res4: Array[Int] = Array(1, 2, 3, 4, 5)

scala> rdd1.collect
res5: Array[Int] = Array(1, 1, 2, 1, 2, 3, 1, 2, 3, 4, 1, 2, 3, 4, 5)
```

```
scala> rdd.union(rdd1).collect
res6: Array[Int] = Array(1, 2, 3, 4, 5, 1, 1, 2, 1, 2, 3, 1, 2, 3, 4, 1, 2, 3,
4, 5)

scala> rdd.intersection(rdd1).collect
res7: Array[Int] = Array(3, 4, 1, 5, 2)

scala> rdd.first
res8: Int = 1

scala> rdd.count
res9: Long = 5

scala> rdd.reduce(_ + _)
res10: Int = 15

scala> rdd.reduce(_ * _)
res11: Int = 120

scala> rdd.top(3)
res12: Array[Int] = Array(5, 4, 3)

scala>
```

二. 遇到的问题

一开始执行完`java -version`命令后，输出如下：

```
Zhengkz@KONZEM:~$ java -version
openjdk version "11.0.6" 2020-01-14
OpenJDK Runtime Environment (build 11.0.6+10-post-Ubuntu-1ubuntu118.04.1)
OpenJDK 64-Bit Server VM (build 11.0.6+10-post-Ubuntu-1ubuntu118.04.1, mixed mo
de, sharing)
```

并不是上面第一部分那个结果，即openjdk的版本不是"1.8.0_162"，我也没有在意这个版本问题。然后在计算Top N部分时，一直报错，报的错误是`java.lang.IllegalArgumentException: Unsupported class file major version 55`。在网上搜索，没有什么收获，只知道报这个错是因为jdk版本不匹配，然后我发现了我的版本与实验所需的版本不同，但我确实也是下载了"1.8.0_162"版本的jdk，所以只能是因为默认的版本是"11.0.6"。通过搜索引擎，找到`update-alternatives --config java`

命令可以切换不同版本的java，这条命令的输出如图：

```
Zhengkz@KONZEM:~$ sudo update-alternatives --config java
There are 2 choices for the alternative java (providing /usr/bin/java).

  Selection    Path                                            Priority   Statu
s
------------------------------------------------------------
* 0            /usr/lib/jvm/java-11-openjdk-amd64/bin/java      1111      auto
mode
  1            /usr/lib/jvm/java-11-openjdk-amd64/bin/java      1111      manua
l mode
  2            /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/java   1081      manua
l mode

Press <enter> to keep the current choice[*], or type selection number: 2
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/java to pr
ovide /usr/bin/java (java) in manual mode
```

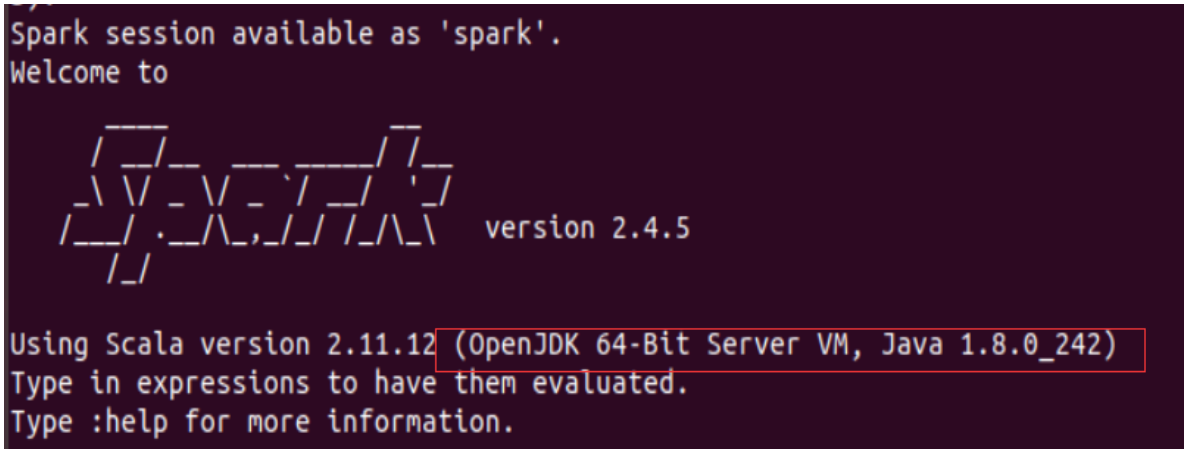
输入2切换到所需的版本，然后利用命令`java -version`查看java版本，果然版本就对了。但是最后还是报错了，然后我发现了一个问题，如图：

```
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.4.5
      /_/

Using Scala version 2.11.12 (OpenJDK 64-Bit Server VM, Java 11.0.6)
Type in expressions to have them evaluated.
Type :help for more information.
```

说明我现在还是用的是"11.0.6"版本的jdk，通过回顾实验一的实验步骤，发现java的配置还包括环境变量，所以通过修改环境变量`JAVA_HOME`：

```
# ~/.bashrc: executed by bash(1) for non-login shells.
# see /usr/share/doc/bash/examples/startup-files (in the package bash-doc)
# for examples

#export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/jre
```

成功解决了这个版本不匹配的问题：


Spark session available as 'spark'.
Welcome to
version 2.4.5
Using Scala version 2.11.12 (OpenJDK 64-Bit Server VM, Java 1.8.0_242)
Type in expressions to have them evaluated.
Type :help for more information.