本地开发环境

2020年11月18日

9:11

按住 ctrl 单机鼠标查看源码

计算机生成了可选文字:
from pyspark import 
SparkConf 
from pyspark import 
SparkContext 
、 参 、 ' 佇 息 
#@ 建 S arkConf: 讠 几 置 、 ark 
SparkConf() ． setMaster( " local [ 2 ] ' ） ． setAppName( " spark-demo' ） 
conf 
#@ arkContext 
SparkContext( 
-conf) 
conf 
S C 
# 业 夂 逻 车 。 
data 
[ 2 4 ] 
s c .parallelize(data) 
distData 
print(distData ． collect()) 
sc.stop() 

<http://spark.apache.org/docs/2.4.0/rdd-programming-guide.html>

设置基本参数， PYTHONPATH SPARK\_HOME 2个zip包

使用local进行本地测试

Pycharm：

设置基本python环境先

再：

计算机生成了可选文字:
〈 stul _ ...Xpy_spark stulrdd firstXhi.py 
Edit Configurations... 
Save 'hi' Configuration 

修改

Environment variables:

自行添加spark解压目录下的python路径以及spark路径

【PYTHONUNBUFFERED=1;PYTHONPATH=G:\spark-2.4.6-bin-hadoop2.6\python;SPARK\_HOME=G:\spark-2.4.6-bin-hadoop2.6】

计算机生成了可选文字:
Environment Variables 
User environment variables: 
Name 
PYTHONUNBUFFERED 
G:Xspark-2.4.6-bin-hadoop2.6Xpython 
PYTHONPATH 
PARK HOME 
Value 
1 
G:Xspark-2.4.6-bin 
〔 I Include syste m environment variables: 
-hadoop2.d 
0 
Name 
ALLUSERSPROFILE 
ANDROID S D K HOME 
APPDATA 
CLASSPATH 
COMPUTERNAME 
Value 
C:XProgramData 
C:XAndroid 
C:WsersWihanXAppDataXRoaming 
.;G:X 丿 dk8Xbin;G:X 丿 dk8XIibXdt.jar;G:X 丿 d... 
MSI 
C_•AProaram FilesXC_ommon Files 

添加如下两个包：

计算机生成了可选文字:
Appearance & Behavior 
Keymap 
． Ed ito r 
． Ve rsiO n Control 
' Project: Py spark Stu 
P roject Interpreter 
Project Structure 
· Build, Execution, Deployment 
． Languages & Frameworks 
． Tools 
Project ： Py spark Stu 》 Project Structure 
Mark as: ． Sources 
- Excluded 
V ． G:xwarxpy_spark_stu 
For current project 
+ Ad d Content Root 
Stu 
． rd d fi rst Select content root directory 
Content root is a directory containing all files related to this module 
Hide path 
G:Xspark-2.4.6-bin-hadoop2.6XpythonXlibXpyspark.zip 
> ． examples 
> ． kubernetes 
> ． licenses 
． python 
> py4j-0.10.7-src.zip 
> ． pyspark 
> ． test coverage 
> ． test_support 
Svstem Volume Information 
Exclude files: 
Use ： to 
symbol 
0 
Drag and drop a file i nto the space a bove to quickly locate it in the tree 
OK 
Cancel 

放到集群环境中时 ，按照官网建议，不要硬编码设置参数

计算机生成了可选文字:
· 1 Pspark 
2 Pspa. 
from 
pyspark 
import 
SparkConf 
from 
pyspark 
import 
SparkContext 
# 创 建 SparkConf ： 设 置 的 是 Spark 相 关 的 参 数 信 息 
conf 
#仓 刂 建 SparkContext 
SparkContext(conf—conf) 
S C 
# 业 务 逻 辑 
data 
1 2 3 4 
distData 
s c ． parallelize(data) 
print(distData. collect()) 
# 习 惯 
s c .stop() 

使用spark-submit进行提交

[root@Pspark bin]# ./spark-submit --master local[2] --name spark-11 /home/script/spark-11.18.py

计算机生成了可选文字:
20 / 11 / 17 
20 / 11 / 17 
20 / 11 / 17 
20 / 11 / 17 
20 / 11 / 17 
1 2 3 
20 / 11 / 17 
20 / 11 / 17 
20 / 11 / 17 
20 / 11 / 17 
20 / 11 / 17 
20 / 11 / 17 
20 / 11 / 17 
20 / 11 / 17 
20 / 11 / 17 
20 / 11 / 17 
21 ： 36 ： 24 
21 ： 36 ： 24 
21 ： 36 ： 24 
21 ： 36 ： 24 
21 ： 36 ： 24 
4 
21 ： 36 ： 24 
21 ： 36 ： 24 
21 ： 36 ： 24 
21 ： 36 ： 24 
21 ： 36 ： 24 
21 ： 36 ： 24 
21 ： 36 ： 24 
21 ： 36 ： 24 
21 ： 36 ： 24 
21 ： 36 ： 24 
INFO 
INFO 
INFO 
INFO 
INFO 
INFO 
INFO 
INFO 
INFO 
INFO 
INFO 
INFO 
INFO 
INFO 
INFO 
INFO 
Executor: Finished task 0 ． 0 in stage 0 ． 0 (TID 0 ). 727 bytes result sent t 0 driver 
TaskSetManager: Finished task 0 ． 0 in 
stage 0 ． 0 (TID 0 ） in 180 ms on localhost (executor driver) 
TaskSetManager: Finished task 1 ． 0 in 
stage 0 ． 0 (TID 1 ） in 159 ms on localhost (executor driver) 
TaskScheduIerImpI: Removed TaskSet 0 ． 0 ， whose tasks have a11 completed ， from p001 
DAGScheduler: ResultStage 0 (collect 
at /home/script/spark-Il.18.py:13) finished in 0 ． 491 s 
DAGScheduIer: Job 0 finished: collect at /home/script/spark-Il.18.py:13, took 0 ． 567803 s 
SparkUI ： Stopped Spark we b UI at http://Pspark:4040 
MapOutputTrackerMasterEndpoint ： MapOutputTrackerMasterEndpoint stopped! 
MemoryStore ： MemoryStore cleared 
BlockManager: BlockManager stopped 
BlockManagerMaster: BlockManagerMaster stopped 
OutputCommitCoordinator$OutputCommitCoordinatorEndpoint ： OutputCommitCoordinator stopped! 
SparkContext ： Successfully stopped SparkContext 
ShutdownH00kManager: Shutdown hook called 
ShutdownHookManager: Deleting directory /tmp/spark-f34213a7-fa1c-45c2-ae32-d7c11e85a4f1 
ShutdownH00kManager: Deleting directory /tmp/spark-32cb65ff-baf4-4490-a088-8d1c8981da21/pyspark 
-aG 
228C7 一 7897 一 4945 一 be5c -8af1def57dec 