**Hue之Spark使用**

目录

[1. Spark简介 1](#_Toc430767376)

[2. Hue中使用Spark 1](#_Toc430767377)

[2.1. Spark & Python 2](#_Toc430767378)

[2.2. Spark & Scala 3](#_Toc430767379)

[2.3. Spark & Impala 4](#_Toc430767380)

[2.4. Spark & Hive 4](#_Toc430767381)

[3. 注意事项 5](#_Toc430767382)

# Spark简介

Spark 是基于 map reduce 算法实现的分布式计算，拥有 Hadoop MapReduce 所具有的优点 ; 但不同于 MapReduce 的是 Job 中间输出和结果可以保存在内存中，从而不再需要读写 HDFS ，因此 Spark 能更好地适用于数据挖掘与机器学习等需要迭代的 map reduce 的算。

# Hue中使用Spark

UDH安装过程中勾选Spark服务和Hue服务，安装完成后Hue服务界面会内嵌Spark服务。切换到hue的主页从快速链接进入到hue的web端，点击Query Editor选择Spark进入该服务的操作界面：

****

可以看到Spark的页面为：

****

Spark提供多种类型的工作方式，包括支持Scala、Python、SQL（hive，impala）、Text。可从下图的下拉菜单中选择您需要的工作方式，进行编程实现相应的功能：

****

下面的章节将介绍多种类型的语言在Spark中运行的具体案例。

## Spark & Python

在下拉菜单中选择Python，点击右侧“加号”按钮，则页面出现一个Python的编辑框，如下图所示：

****

在编辑框中输入相应的Spark程序，下图的例子是一个利用Spark python通过扔飞镖估算Pi的值的计算 （spark 适用密集型计算）。

点击左侧的“运行”按钮，即可运行：

****

运行结果将显示在代码编辑框下方的空白处。

## Spark & Scala

 在下拉菜单中选择Scala，点击右侧“加号”按钮，则页面出现一个Scala的编辑框，添加相应代码：

 该代码统计hdfs文件里单词个数，具体如下图所示：



 执行结果为：



## Spark & Impala

 同时，Spark可通过调用hive/impal api来完成增、删、改、查等操作：

 在下拉菜单中选择Impala，点击右侧“加号”按钮，则页面出现一个Impala的编辑框，添加相应代码：



 可如Impala页面一样在编辑框下方查看执行结果。

## Spark & Hive

在下拉菜单中选择Hive，点击右侧“加号”按钮，则页面出现一个Hive的编辑框，添加相应代码：



 可如Hive页面一样在编辑框下方查看执行结果。

# 注意事项

多次添加python或scala 编辑框会报如下错误：

---FAILED SelectChannelConnector@0.0.0.0:4040: java.net.BindException: 地址已在使用

解决方案：

1. 建议每个打开一个即可；
2. 手动杀掉其中一个进程（kill -9 进程号）

