

原

Spark_Bench使用文档

2018年02月03日 22:27:30

xfg0218

阅读数：845

更多

版权声明：本文为博主原创文章，未经博主允许不得转载。 <https://blog.csdn.net/xfg0218/article/details/79250019>

Spark_Bench使用文档

1. 什么是Spark-Bench

SparkBench是Spark的基准测试组件(集成了很多spark支持的经典测试案例)。它大致包含四种不同类型的测试案例，包括机器学习，图形处理，流处理以及查询。Spark-Bench所选择的测试案例可以，在不同的工作负载情况下测试出系统瓶颈;目前，我们大致涵盖了CPU，内存和Shuffle以及IO密集型工作负载（如HDFS读写）。它还包括一个数据生成器，允许用户生成任意大小的输入数据。

2. 为什么是Spark-Bench

虽然Spark已经快速发展，但spark社区缺乏为Spark量身定做的基准测试组件。这种组件的目的是帮助用户了解不同系统设计之间的优缺点，同时指导如何进行Spark配置优化集群配置。特别是SparkBench有四个主要的用例。

用例1、它可以对Spark系统优化进行定量比较，如缓存策略优化、内存管理优化，调度策略优化的定量比较。研究开发人员可以使用SparkBench来全面评估、比较优化过后的性能差异。

用例2、它为不同平台和硬件集群设置（如Google云和Amazon云）提供定量比较。

用例3、它为集群大小和配置提供指导。它还有助于确定瓶颈资源，并最大限度地减少资源争用的影响。

用例4、它允许深入研究Spark系统在各种方面的性能影响，包括工作负载表征，参数影响研究，Spark系统的可扩展性和容错行为。

机器学习测试案例：逻辑回归、支持向量机、矩阵分解

图计算测试案例：PageRank、SVD++、三角计数（Triangle Count）

SQL查询测试案例：Hive、RDDRelation

流处理测试案例：Twitter Tag、Page View

其他测试案例：Kmeans，线性回归，决策树,最短路径, 标签传播, 连通图，强连通图

3. 如何开始

3.1 系统环境配置

安装JDK，Hadoop，spark运行环境

1、要运行新版本的SparkBench, 在安装SparkBench之前要先下载并安装wikixmlj，进入到一个目录后直接执行clone命令下载

```
git clone https://github.com/synhershko/wikixmlj.git
```

```
cd wikixmlj
```

```
mvn package -Dmaven.test.skip=true
```

```
mvn install -Dmaven.test.skip=true
```

注：上面两条mvn命令原本是mvn package install命令转换而来，直接运行打包和安装命令会报在test过程出错，所以我们要设置跳过测试打包和安装，运行成功后就会在w生成wikixmlj-1.0-SNAPSHOT.jar文件，同时会安装到maven 仓库中。

2、下载 SparkBench基准测试组件：

<https://github.com/SparkTC/spark-bench/tree/legacy>

解压后，首先运行bin目录下的build-all.sh文件去build整个SparkBench工程，在这个过程中主要是根据工程的pom.xml文件去网上下载SparkBench工程依赖的所有jar包。

```
./SPARK_BENCH_HOME/bin/build-all.sh
```

3、可以通过修改SPARK_BENCH_HOME的conf目录下的env.sh对配置Spark-Bench环境

```
SPARK_HOME=/home/hadoop/Spark-1.6.0-bin-hadoop2.6.0
```

```
HADOOP_HOME=/home/hadoop/hadoop-2.6.0
```

```
SPARK_MASTER=spark://master:7077
```

```
HDFS_MASTER=hdfs://master:9000/
```

3.2 SparkBench配置

可以通过修改SPARK_BENCH_HOME的conf目录下的env.sh对配置Spark-Bench环境,保证下列环境变量一定要设置。

```
SPARK_HOME=/home/hadoop/Spark-1.6.0-bin-hadoop2.6.0
```

```
HADOOP_HOME=/home/hadoop/hadoop-2.6.0
```

```
SPARK_MASTER=spark://master:7077
```

```
HDFS_MASTER=hdfs://master:9000/
```

你的浏览器目前处于缩放状态，页面可能会出现错位现象，建议100%大小显示。

Local mode: (本地模式)

```
DATA_HDFS="file:///home/whoami/SparkBench" SPARK_MASTER=local[2] MC_List=""
```

3.3 执行

Scala版本执行方式：

直接进入相应案例的目录的bin目录下，先自动生成测试数据，然后再运行

```
<SPARK_BENCH_HOME>/<Workload>/bin/gen_data.sh
```

```
<SPARK_BENCH_HOME>/<Workload>/bin/run.sh
```

Java版本执行方式：

```
<SparkBench_Root>/<Workload>/bin/gen_data_java.sh
```

```
<SparkBench_Root>/<Workload>/bin/run_java.sh
```

在运行SQL查询案例时应该注意的：

在运行SQL查询案例时，默认是运行其中的RDDRelation案例，如果要想运行其中的Hive案例可以执行下面代码：

```
<SPARK_BENCH_HOME>/SQL/bin/run.sh hive
```

在运行Streaming案例时应该注意的：

在运行流数据处理案例时，例如TwitterTag，Streaming逻辑回归，

首先在一个终端中执行<SPARK_BENCH_HOME>/Streaming/bin/gen_data.sh

然后再另一个终端中执行<SPARK_BENCH_HOME>/Streaming/bin/run.sh

而且在执行脚本时必须指定你要运行案例名字的参数，如下：

```
<SPARK_BENCH_HOME>/Streaming/bin/gen_data.sh TwitterPopularTags
```

```
<SPARK_BENCH_HOME>/Streaming/bin/run.sh TwitterPopularTags
```

当然你也可以在Streaming/conf/env.sh 配置文件中指定你要运行的子案例的名称，通过修改 subApp= TwitterPopularTags。

3.4 如何查看结果

可以直接去<SPARK_BENCH_HOME>/report 目录下去查看最后的结果

4. 高级配置

4.1 配置你要运行的案例

<SPARK_BENCH_HOME>/bin/run-all.sh 可以运行所有在<SPARK_BENCH_HOME>/bin/applications.lst中指定的所有案例，applications.lst中每一行都指定一个要运行的案例

当然你也可以单独运行每一个案例，在每一个案例的根目录下都有三个文件：

```
<Workload>/bin/config.sh 案例配置文件
```

```
<Workload>/bin/gen_data.sh 案例测试数据生成文件
```

```
<Workload>/bin/run.sh 案例运行文件
```

4.2 Apache Spark运行配置

Spark运行时的各种配置可以在配置文件中指定，如指定下面的参数：

```
spark.executors.memory
```

```
Executor memory,
```

```
standalone or YARN mode
```

spark.driver.memory

Driver memory,

standalone or YARN mode

spark.rdd.cache

-----以下是具体配置-----

#配置附图

1.进入到conf文件配置env.sh

cd /opt/xintongyuan/spark-bench-legacy/conf

vim env.sh(修改下面画红线部分为现有集群环境)

SPARK_HOME The Spark installation location

HADOOP_HOME The HADOOP installation location

SPARK_MASTER Spark master

HDFS_MASTER HDFS master

Local mode:

DATA_HDFS="file:///home/whoami/SparkBench" SPARK_MASTER=local[2] MC_List=""

下面是调整spark执行参数部分（默认如下：）

SPARK_EXECUTOR_MEMORY=1g

SPARK_DRIVER_MEMORY=2g

SPARK_EXECUTOR_INSTANCES=4

SPARK_EXECUTOR_CORES=1

Storage levels, see <http://spark.apache.org/docs/latest/api/java/org/apache/spark/api/java/StorageLevels.html>

- STORAGE_LEVEL, set MEMORY_AND_DISK, MEMORY_AND_DISK_SER, MEMORY_ONLY, MEMORY_ONLY_SER, or DISK_ONLY

STORAGE_LEVEL=MEMORY_AND_DISK

for data generation

NUM_OF_PARTITIONS=2

for running

NUM_TRIALS=1

备注：如果集群开启了kerberos认证，则所有节点均需要登录同一用户，否则报错，参考下面issue

<http://community.cloudera.com/t5/Advanced-Analytics-Apache-Spark/Issue-on-running-spark-application-in-Yarn-cluster-mode/m-p/50719>

SparkBench运行环境搭建 - xiaowei_582648206

按照官方文档进行SparkBench系统环境配置的时候会出现几个错误，在这里——解决



想对作者说点什么

安装spark-bench - haoxiaoyan的专栏

👁 1417

git clone https://github.com/SparkTC/spark-bench.git cd spark-bench/ mvn package install [root@dat...

来自： [haoxiaoyan的专栏](#)

干货分享：SparkBench--Spark平台的基准性能测试 - 数控小J 对大数据的探索与见解

👁 3178

SparkBench的测试项目覆盖了Spark支持的四种最主流的应用类型，即机器学习、图计算、SQL查询和流数据计算...

来自： [数控小J 对大数据的探索...](#)

Spark组件的benchmark - sinat_18497785的博客

👁 2554

Spark组件的benchmark 一、 Benchmark 简单介绍 基准测试（ benchmark ），主要指的是，实现对一类测试对象...

来自： [sinat_18497785的博客](#)

一个退役操盘手肺腑之言，写给无数正在亏钱的散户

协奥·熿燚

你的浏览器目前处于缩放状态，页面可能会出现错位现象，建议100%大小显示。

hadoop 和spark的基准测试(1) - r8t7y6的博客

654

Hadoop 2.8.0 基准测试 1.查看jar包命令 2.建立乱序100M数据 3.排序 4.删除文件 1. 执行: hadoop jar ../share/had...

来自： r8t7y6的博客

下载

Spark-Bench使用文档 - xiaowei_582648206

你的浏览器目前处于缩放状态，页面可能会出现错位现象，建议100%大小显示。

SparkBench是为Spark量身定做的基准测试组件(集成了很多spark支持的经典测试案例)。 它大致包含四种不同类型的测试案例，包括机器

Spark Streaming介绍与基本执行过程 - cache007的专栏

2330

Spark Streaming作为Spark上的四大子框架之一，肩负着实时流计算的重大责任 而相对于另外一个当下十分流行的...

来自： cache007的专栏

标签传播算法 (Label Propagation) 及Python实现 - jiandanjinxin的专栏

1453

半监督学习 (Semi-supervised learning) 发挥作用的场合是：你的数据有一些有label，一些没有。而且一般是绝大...

来自： jiandanjinxin

社区发现算法之标签传播 (LPA) - 6丁儿的猫

来自： 6丁儿的猫

标签传播算法 (LPA) 的做法比较简单：第一步: 为所有节点指定一个唯一的标签；第二步: 逐轮刷新所有节点的标...

大连女孩揭秘海参行业潜规则，常吃海参的人一定得知道！

恒利源商贸 · 熾燚

spark-基准测试 - qq_34969081的博客

123

背景 因成本影响，公司想从高价格的阿里云转到价格较低的金山云上，让我们做一下对金山云上自带的spark_on_y...

来自： qq_34969081的博客

文章热词 机器学习 机器学习课程 机器学习教程 深度学习视频教程 深度学习学习

相关热词 bootstrap 使用 文档 样式类 c++test独立版使用文档 c++11文档 c# 文档 bootstrapswitch 文档 python3.6教程文档 图灵学院python文档

浅析HiBench之SparkBench (单节点) 配置 - don_chiang709的专栏

350

一 前言: 1. 术语： Hadoop 版本: Version 2.7.1 HiBench 版本： Version 6.0 Spark 版本： Version 2.1.0 2. 本篇讲个...

来自： don_chiang709的专栏



liukuan73

关注

487篇文章



郑清

关注

302篇文章



wangqi0079

关注

193篇文章

spark开发环境——本地安装spark2.x及启动 - 闻曦的博客

696

利用python开发spark项目需要在本地安装spark 一 本地安装 1.下载http://spark.apache.org/downloads.html 选择适...

来自： 闻曦的博客

win7 eclipse 调试spark - JackLi31742的博客

224

还没测试成功 http://blog.csdn.net/fly_leopard/article/details/51250443 http://jingyan.baidu.com/article/...

来自： JackLi31742的博客

sparkbenchjar包资源

spark基本性能测试，结合binglia使用，挺好用的，欢迎大家下载。

青羊大道某老板一年败光千万家产，却在短短几个月赚到7位数！

彦亿投资 · 熾燚

浅析HiBench之SparkBench (集群) 配置 - don_chiang709的专栏

94

一、前言: 1. 术语： Hadoop 版本: Version 2.7.1 HiBench 版本： Version 6.0 Spark 版本： Version 2.1.0 Scala 版...

来自： don_chiang709的专栏

Spark2.0.1伪分布式安装配置 - santiago-02-01

4920

前言.Spark简介和hadoop的区别 Spark 是一种与 Hadoop 相似的开源集群计算环境，但是两者之间还存在一些不同...

来自： santiago-02-01

Python海量数据处理之_Hadoop&Spark - 谢彦的技术博客

1871

本篇将介绍Hadoop+Spark的安装配置及如何用Python调用Spark。

来自： 谢彦的技术博客

Spark Java程序案例入门 - 小强签名设计 的博客

9690

spark 安装模式： local(本地模式)： 常用于本地开发测试，本地还分为local单线程和local-cluster多线程 standalone(...

来自： 小强签名设计 的博客

Spark伪分布式安装 (不依赖hadoop) - wangmm0218

8546

上传包： 解压 并重命名： 进入spark100目录： 修改配置： Cd conf 启动： 出错： ...

下载

jira中文使用文档

06-10

jira中文使用文档，挺简洁的大家可以看一下，哈哈。

使用Java API文档 - yqj2065的博客

Java API文档

如何使用官方Android开发文档 - boatImpish

如何使用官方Android开发文档 pre-conditions VPN + AndroidMethod https://developer.android.com 这个是根目录...

417

来自： boatImpish

Java-API文档的使用 - xjx09190的博客

java API使用

534

来自： xjx09190

老股民酒后无意说漏：20年炒股 坚持只看1指标

集升商贸 · 熾燚

FastJson使用文档 - Kiven 每天都需要有所获，每天都需要进步.....

672

1、主要的使用入口 Fastjson API入口类是com.alibaba.fastjson.JSON，常用的序列化操作都可以在JSON类上的静... 来自： Kiven 每天都需要有所获...

学习笔记：利用markdown写readme文档（Udacity学城） - mini猿要成长QAQ

757

周末逛知乎等资讯平台的时候，无意间了解到Udacity学城有一期关于利用markdown撰写readme文档的教程，想到... 来自： mini猿要成长QAQ

Azkaban使用文档 - qq_28549905的博客

154

Azkaban使用文档 1.Azkaban简介 Azkaban是由Linkedin开源的一个批量 workflow 任务调度器。用于在一个 workflow 内... 来自： qq_28549905的博客

“您只能在 HTML 输出中使用 document.write。如果您在文档加载后使用该方法，会覆盖整个文档。” - Li...

5166

提示：您只能在 HTML 输出中使用 document.write。如果您在文档加载后使用该方法，会覆盖整个文档。今天开始... 来自： LittleLawson的博客

您只能在 HTML 输出流中使用 document.write。如果您在文档已加载后使用它（比如在函数中），会覆...

3395

html的输出流（可以是字节流，也可以是字符流，把整个流读取完，才算流结束。）数据载体只是一个数据载体 Ja... 来自： 金莲你没事开啥窗

十大猎头公司

百度广告

vs 文档使用 - xmmdbk的博客

258

vs文档使用查找函数时，函数分为几大类型，函数都可以在这几大类型中的createxxxx中去查看，单独搜索时，查... 来自： xmmdbk的博客

最全的vi使用文档 - yan753124的博客

216

转自： http://blog.csdn.net/donahue_idz/article/details/17139361 曾经使用了两年多的Vim，手册也翻过一遍。虽然... 来自： yan753124的博客

面向对象_如何使用JDK提供的帮助文档 - 辐_射的博客

903

1:打开帮助文档 2:点击显示,找到,索引,看到输入框 3:知道你要找谁?以Scanner举例 4:在输入框内输入Scanner,然后... 来自： 辐_射的博客

绝不要使用在文档加载之后使用 document.write()。这会覆盖该文档 - CferZ的代码之路

4785

W3school-javascript第一部分，关于使用document.write()写入HTML文档流。教程后面有一行提示：您只能在 HTM... 来自： CferZ的代码之路

下载 RT-LAB培训资料 - qq_27526715

12-25

RT-LAB使用文档，RT-LAB使用文档，RT-LAB使用文档，RT-LAB使用文档，RT-LAB使用文档

自己创建网站

百度广告

下载 git 使用文档

12-05

git 使用文档，git 使用文档git 使用文档git 使用文档git 使用文档



xfg0218

关注

原创	粉丝	喜欢	评论
361	155	47	108

等级：	博客 5	访问：	42万+
积分：	7392	排名：	4445
勋章：	恒		



大数据可视化



- 最新文章
- Sqlldr把文本文件导入到ORACLE中

常用随机数生成

superset使用实例

superset安装

Greenplum 对JSON的支持

个人分类	
GreenPlum	13篇
java	25篇
java注解说明使用	1篇
java优化	2篇
大数据书籍	30篇
展开	

归档	
2018年11月	2篇
2018年10月	9篇
2018年9月	41篇
2018年8月	7篇
2018年7月	5篇
展开	

- 热门文章
- 大数据面试题

阅读量：41296

两年工作经验java面试题精炼汇总

阅读量：18165

常用的邮箱服务器（SMTP、POP3）地址、端口

阅读量：12947

Logstash 配置总结

阅读量：12691

Redis之Pipeline使用注意事项

阅读量：11282

你的浏览器目前处于缩放状态，页面可能会出现错位现象，建议100%大小显示。

最新评论

大数据资料数据集

qq_17641711 : 全部取消了啊，求更新

phoenix 把CSV格式的数据...

qq_27252133 : [reply]qq_21577617[/reply] 这个我也没有解决，推荐使用另外一种方式：pl...

phoenix 把CSV格式的数据...

qq_21577617 : [reply]qq_27252133[/reply] 层主，client.RpcRetryingC...


Logstash 配置总结

qq_39570637 : xiexie 有帮助 很全面


phoenix 把CSV格式的数据...

qq_27252133 : [reply]xfg0218[/reply] 也就是你博文中的：17/04/08 00:06:08...

Greenplum



武汉 车展



联系我们



微信客服



QQ客服

 QQ客服

 kefu@csdn.net

 客服论坛

 400-660-0108

工作时间 8:00-22:00

关于我们 招聘 广告服务 网站地图

 百度提供站内搜索 京ICP证09002463号

©1999-2018 江苏乐知网络技术有限公司

江苏知之为计算机有限公司 北京创新乐知信息技术有限公司版权所有

网络110报警服务 经营性网站备案信息

北京互联网违法和不良信息举报中心

中国互联网举报中心

你的浏览器目前处于缩放状态，页面可能会出现错位现象，建议100%大小显示。