python hadoop的应用_hadoop python api

原创

weixin_39914938 于 2020-12-18 12:46:03 发布 178 收藏 1
文章标签： python hadoop的应用

[雪峰磁针石博客]大数据Hadoop工具python教程4-mrjob

mrjob是由Yelp创建的Python MapReduce库，它封装了Hadoop流，允许MapReduce应用程序以更加Pythonic的方式编写。 mrjob用纯Python编写多步MapReduce作业。使用mrjob编写的MapReduce作业可以在本地测试，在Hadoop集群上运行，或...

文章

python人工智能命理

2019-01-28

1413浏览量

《Spark与Hadoop大数据分析》——导读

前　　言

本书讲解了Apache Spark和Hadoop的基础知识，以及如何通过简单的方式将它们与最常用的工具和技术集成在一起。所有Spark组件(Spark Core、Spark SQL、DataFrame、Dataset、Conventional Streaming、Structured S...

文章

华章计算机

2017-09-01

909浏览量

Hadoop概念学习系列之Hadoop的文件系统(十六)

Hadoop整合了众多文件系统，它首先提供了一个高层的文件系统抽象类org.apache.hadoop.fs.FileSystem，这个抽象类展示了一个分布式文件系统，并有几个具体实现。

如下表所示。

Hadovp提供了许多文件系统的接口，用户可使用URI方案选取合适的文件系统来实...

文章

技术小哥哥

2017-11-14

1075浏览量

《Hadoop实战第2版》——3.4节Hadoop流

3.4　Hadoop流 Hadoop流提供了一个API，允许用户使用任何脚本语言写Map函数或Reduce函数。Hadoop流的关键是，它使用UNIX标准流作为程序与Hadoop之间的接口。因此，任何程序只要可以从标准输入流中读取数据并且可以写入数据到标准输出流，那么就可以通过Hadoop流使用其...

文章

华章计算机

2017-08-01

885浏览量

Hadoop学习资源集合

Hadoop是一个由Apache基金会所开发的开源分布式系统基础架构。用户可以在不了解分布式底层细节的情况下，开发分布式程序，充分利用集群的威力进行高速运算和存储。Hadoop得以在大数据处理应用中广泛应用得益于其自身在数据提取、变形和加载(ETL)方面上的天然优势。Hadoop的分布式架构，将大...

文章

readygo

2016-05-18

36399浏览量

[喵咪大数据]HUE大数据管理工具

日常的大数据使用中经常是在服务器命名行中进行操作,可视化功能仅仅依靠着各个组件自带的网页进行,那么有没有一个可以结合大家能在一个网页上的管理工具呢?答案是肯定的,今天就和大家一起来探索大数据管理工具HUE的庐山真面目.

附上:

喵了个咪的博客:w-blog.cn

1.环境准备

编译依赖

wget ...

文章

喵了个咪_

2020-08-11

137浏览量

用python写MapReduce函数——以WordCount为例

阅读目录

1. Python MapReduce 代码

2. 在Hadoop上运行python代码

3. 利用python的迭代器和生成器优化Mapper 和 Reducer代码

4. 参考

尽管Hadoop框架是用java写的，但是Hadoop程序不限于java，可以用pyth...

文章

技术mix呢

2017-10-18

1077浏览量

Hadoop大数据平台实战(02)：HBase vs. Hive vs. Impala 对比

Hadoop大数据平台中非常重要的三个技术：HBase vs. Hive vs. Impala。他们之间的关系和区别。

Apache™Hadoop是目前最流行的开源大数据平台，核心组件使用Java语言开发。

Apache Hadoop软件库是一个框架，允许使用简单的编程模型跨计算机集群分布式处理大...

文章

徐雷frank

2019-04-06

1386浏览量

[雪峰磁针石博客]大数据Hadoop工具python教程9-Luigi工作流

管理Hadoop作业的官方工作流程调度程序是Apache Oozie。与许多其他Hadoop产品一样，Oozie是用Java编写的，是基于服务器的Web应用程序，它运行执行Hadoop MapReduce和Pig的工作流作业。 Oozie工作流是在XML文档中指定的控制依赖性指导非循环图(DAG)...

文章

python人工智能命理

2019-01-28

1294浏览量

如何在 Apache Flink 中使用 Python API？

作者：孙金城(金竹)整理：韩非

本文根据 Apache Flink 系列直播课程整理而成，由 Apache Flink PMC，阿里巴巴高级技术专家 孙金城 分享。重点为大家介绍 Flink Python API 的现状及未来规划，主要内容包括：Apache Flink Python API 的...

文章

阿里云实时计算Flink

2019-09-09

2978浏览量

Apache Flink 1.9.0 为什么将支持 Python API？

作者：孙金城(金竹)

本文目录：1.最流行的编程语言2.互联网最火热的领域2.1大数据时代，数据量与日俱增2.2数据的价值来源于数据分析2.3数据价值最大化，时效性3.阿尔法与人工智能4.总结

众所周知，Apache Flink(以下简称 Flink)的 Runtime 是用 Java 编写的，而...

文章

Ververica

2019-08-01

1784浏览量

Apache Flink 1.9.0 为什么将支持 Python API？

作者：孙金城(金竹)

本文目录：1.最流行的编程语言2.互联网最火热的领域2.1大数据时代，数据量与日俱增2.2数据的价值来源于数据分析2.3数据价值最大化，时效性3.阿尔法与人工智能4.总结

众所周知，Apache Flink(以下简称 Flink)的 Runtime 是用 Java 编写的，而...

文章

阿里云实时计算Flink

2019-08-03

2690浏览量

Hue安装配置实践

Hue是一个开源的Apache Hadoop UI系统，最早是由Cloudera Desktop演化而来，由Cloudera贡献给开源社区，它是基于Python Web框架Django实现的。通过使用Hue我们可以在浏览器端的Web控制台上与Hadoop集群进行交互来分析处理数据，例如操作HDFS...

文章

shiyanjuncn

2016-04-13

5111浏览量

我为什么说 Python 是大数据全栈式开发语言

前段时间，ThoughtWorks在深圳举办一次社区活动上，有一个演讲主题叫做“Fullstack JavaScript”，是关于用JavaScript进行前端、服务器端，甚至数据库(MongoDB)开发，一个Web应用开发人员，只需要学会一门语言，就可以实现整个应用。

受此启发，我发现Pyth...

文章

小旋风柴进

2017-05-02

2085浏览量

来！PyFlink 作业的多种部署模式

关于 PyFlink 的博客我们曾介绍过 PyFlink 的功能开发，比如，如何使用各种算子(Join/Window/AGG etc.)，如何使用各种 Connector(Kafka, CSV, Socket etc.)，还有一些实际的案例。这些都停留在开发阶段，一旦开发完成，我们就面临激动人心的...

文章

阿里云实时计算Flink

2020-01-20

2324浏览量

来！PyFlink 作业的多种部署模式

关于 PyFlink 的博客我们曾介绍过 PyFlink 的功能开发，比如，如何使用各种算子(Join/Window/AGG etc.)，如何使用各种 Connector(Kafka, CSV, Socket etc.)，还有一些实际的案例。这些都停留在开发阶段，一旦开发完成，我们就面临激动人心的...

文章

阿里云实时计算Flink

2020-01-20

880浏览量

Spark-python-快速开始

1. 概览

这篇文章主要是关于Spark的快速熟悉和使用，我们使用Python和Spark的shell接口来操作Spark。 Spark shell使得我们可以很简单的学习Spark的Api，同时也是一个强大数据分析交互的工具。

2. Spark shell

我们使用Python版本的Spark...

文章

陈国林

2016-09-11

1075浏览量

手把手教你入门Hadoop(附代码&资源)

GETINDATA公司创始人兼大数据顾问彼得亚·雷克鲁斯基(Piotr Krewski)和GETINDATA公司首席执行官兼创始人亚当·卡瓦(Adam Kawa)

目录

内容简介设计理念HADOOP组件HDFSYARNYARN 应用程序监控 YARN 应用程序用HADOOP处理数据 HADOOP 的...

文章

技术小能手

2018-05-02

2845浏览量

Spark 概念学习系列之Spark的优点(八)

Spark的一站式解决方案，非常之具有吸引力，毕竟啊，任何公司都想用统一的平台去处理遇到的问题，减少开发和维护的人力成本和部署平台的物力成本。

当然，Spark并没有以牺牲性能为代价。相反，在性能方面，Spark具有很大的优势。

Spark凭借以下的优点在众多的大数据分...

文章

技术小哥哥

2017-11-02

1000浏览量

Spark 概念学习系列之Spark的优点(八)

Spark的一站式解决方案，非常之具有吸引力，毕竟啊，任何公司都想用统一的平台去处理遇到的问题，减少开发和维护的人力成本和部署平台的物力成本。

当然，Spark并没有以牺牲性能为代价。相反，在性能方面，Spark具有很大的优势。

Spark凭借以下的优点在众多的大数...

文章

技术小哥哥

2017-11-14

937浏览量

《Hadoop海量数据处理：技术详解与项目实战》一3.3 如何访问HDFS

本节书摘来异步社区《Hadoop海量数据处理：技术详解与项目实战》一书中的第3章，第3.3节，作者： 范东来责编： 杨海玲，更多章节内容可以访问云栖社区“异步社区”公众号查看。

3.3 如何访问HDFS

Hadoop海量数据处理：技术详解与项目实战HDFS提供给HDFS客户端访问的方式多种多样，...

文章

异步社区

2017-05-02

2887浏览量

零基础大数据学习框架

大数据开发最核心的课程就是Hadoop框架，几乎可以说Hadoop就是大数据开发。这个框架就类似于Java应用开发的SSH/SSM框架，都是Apache基金会或者其他Java开源社区团体的能人牛人开发的贡献给大家使用的一种开源Java框架。科多大数据大数据来带你看看。

Java语言是王道就是这个道...

文章

游客j3pqckwdg637c

2019-05-31

780浏览量

《Spark与Hadoop大数据分析》——3.2　学习Spark的核心概念

本节书摘来自华章计算机《Spark与Hadoop大数据分析》一书中的第3章，第3.2节,作者：文卡特·安卡姆(Venkat Ankam) 更多章节内容可以访问云栖社区“华章计算机”公众号查看。

3.2　学习Spark的核心概念

在本节，我们要了解 Spark 的核心概念。Spark 提供的主要抽象...

文章

华章计算机

2017-07-03

2271浏览量

《Spark与Hadoop大数据分析》——

本节书摘来自华章计算机《Spark与Hadoop大数据分析》一书中的第2章，第2.2节,作者：文卡特·安卡姆(Venkat Ankam) 更多章节内容可以访问云栖社区“华章计算机”公众号查看。

2.2　Apache Spark概述

Hadoop和MR已有10年历史，已经被证明是高性能处理海量数据的...

文章

华章计算机

2017-07-03

3385浏览量

开源大数据技术专场(下午):Databricks、Intel、阿里、梨视频的技术实践

开源大数据技术专场下午场在阿里技术专家封神的主持下开始，参与分享的嘉宾有Spark Commiter、来自Databriks的范文臣，HDFS committer、Intel 研发经理郑锴，逸晗网络科技大数据平台负责人杨智，Intel技术专家毛玮，以及阿里云技术专家木艮。

Databricks...

文章

百遇

2016-10-16

7152浏览量

MaxCompute 2.0 生态开放之路及最新发展

文章转自yizhuo

MaxCompute(原ODPS)是阿里云自主研发的分布式大数据处理系统。长久以来，这套阿里自研的系统为阿里内部服务，有自己的类型系统，配套工具以及 SDK 和编程接口。但是随着公共领域对 MaxCompute 的需求越来越强烈，我们也在尽自己所能，使 MaxCompute ...

文章

隐林

2016-10-11

6542浏览量

Splunk Hunk 6.1：面向Hadoop和NoSQL

文章讲的是Splunk Hunk 6.1：面向Hadoop和NoSQL，日前，领先的实时运维智能软件供应商Splunk Inc. (NASDAQ：SPLK)宣布推出面向Hadoop与NoSQL Data Stores的6.1版Hunk: Splunk Analytics for Hadoop and...

文章

青衫无名

2017-09-01

1040浏览量

和封神一起“深挖”Spark

2016云栖大会·北京峰会于8月9号在国家会议中心拉开帷幕，在云栖社区开发者技术专场中，来自阿里云技术专家曹龙(封神)为在场的听众带来《Deep dive into Spark》精彩分享。

关于分享者

曹龙，花名封神，专注在大数据领域，6年分布式引擎研发经验。先后研发上万台Hadoop、ODPS集...

文章

云学习小组

2016-08-24

9789浏览量

13个最流行机器学习框架 帮你解决网络安全机器学习的困难问题

在过去的一年中， 机器学习 发展得热火朝天，已成为主流。机器学习的“空降”并非仅仅由廉价的云环境以及 日益强大的GPU硬件驱动 ，同时也受到了开源框架的蓬勃发展的影响。这些开源框架用于提取机器学习中最困难部分，使机器学习可供广泛开发者使用。

用机器学习解决网络安全问题 开源机器学习框架能助力

《 ...

文章

晚来风急

2017-09-01

7282浏览量

解析Cloudera Manager内部结构、功能包括配置文件、目录位置等

1. 相关目录

/var/log/cloudera-scm-installer : 安装日志目录。

/var/log/* : 相关日志文件(相关服务的及CM的)。

/usr/share/cmf/ : 程序安装目录。

/usr/lib64/cmf/ : Agent程序代码。

/v...

文章

cloudcoder

2016-05-13

3387浏览量