

博客 (//blog.csdn.net/)

学院 (//edu.cs/d/wwet).csd亦載t/http://download.csdn.net)

GitChat (http://gitbook.cn/?ref=csdn)

论坛 (http://bbs.csdn.net)

Spark SQL 函数操作

2016年12月11日 13:04:50 原创

标签:函数 (http://so.csdn.net/so/search/s.do?q=函数&t=blog) /

spark (http://so.csdn.net/so/search/s.do?q=spark&t=blog) /

api (http://so.csdn.net/so/searrqh/s.do?q=api&t=blog) /

sql (http://so.csdn.net/so/search/s.do?q=sql&t=blog) /

数据 (http://so.csdn.net/so/search/s.do?q=数据&t=blog)

3461

Spark 内置函数

使用Spark SQL中的内置函数对数据进行分析,Spark SQL API不同的是,DataFrame中的内置函数操作的结果是 返回一个Column对象,而DataFrame天生就是" A distributed collection of data organized into named columns.",这就为数据的复杂分析建立了坚实的基础并提供了极大的方便性,例如说,我们在操作DataFrame的 方法中可以随时调用内置函数进行业务需要的处理,这之于我们构建附件的业务逻辑而言是可以极大的减少不必须 的时间消耗(基于上就是实际模型的映射),让我们聚焦在数据分析上,这对于提高工程师的生产力而言是非常有 价值的Spark 1.5.x开始提供了大量的内置函数,

还有max、mean、min、sum、avg、explode、size、sort_array、day、to_date、abs、acros、asin、atan 总体上而言内置函数包含了五大基本类型:

- 聚合函数,例如countDistinct、sumDistinct等;
- 集合函数,例如sort array、explode等
- 日期、时间函数,例如hour、quarter、next_day
- 数学函数,例如asin、atan、sqrt、tan、round等;
- 开窗函数,例如rowNumber等
- 字符串函数, concat、format_number、rexexp_extract
- 其它函数, isNaN、sha、randn、callUDF

Hive 下的 单行单词统计

- 1 select t.wd ,count(t.wd) as count from (select explode(split(line," ")) as wd from word)
- t group by t.wd;

在编写程序代码的时候如果调用函数那么需要注意的是要导入functions





utm_sour 网站客服 (http://wpa.qq.com/msgrd? v=3&uin=2431299880&site=qq&r

webmaster@csdn.net (mailto:webmaster@csdn.net)

(http://e.weibo.com/csdnsupport/r 400-660-0108

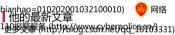
关于 招聘 广告服务

©2018 CSDN 京ICP证09002463号

(http://www.miibeian.gov.cn/)



(http://www.hd315.gov.cn/beian/view.asp?



OOZIE 4.1 中的 Java动态代理 递归 应用 (http://blog.csdn.net/qq_16103331/ar ticle/details/78687139)

Dubbo与Zookeeper、SpringMVC整合 和使用(负载均衡、容错)(http://blog. csdn.net/qq_16103331/article/details/ 78136548)

kettle 7.1 集成 HDP2.6 写入HDFS 报错 问题 (http://blog.csdn.net/qq_1610333 1/article/details/74001609)

[翻译]Kafka Streams简介: 让流处理变得 更简单 (http://blog.csdn.net/qq_16103 331/article/details/73650376)

Spark SQL 集成ElasticSearch的案例实战 (http://blog.csdn.net/qq_16103331/ar ticle/details/53562373)

文章分类

微服务 (http://blog.csdn.net... 🗘 内容举报 ETL收录 (http://blog.csdn.ne... Spark 调优 (http://blog.csdn... 🙃 返回顶部 JAVASE (http://blog.csdn.ne... 83篇 JavaEE (http://blog.csdn.net... 92篇 展开~

博主专栏

```
1 import org.apache.spark.sql.types.{DataTypes, StructType}
2 import org.apache.spark.{SparkConf, SparkContext}
3
  import org.apache.spark.sql.{Row, SQLContext, functions}
4
   object SparkSQLFunctionOps {
5
     def main(args: Array[String]): Unit = {
7
       val conf = new SparkConf().setAppName("SparkSQLFunctionOps").setMaster("local");
       val sc = new SparfkContext(conf)
8
q
       val sqlContext = new SQLContext(sc)
       val linesRDD = sc.textFile("E:/test/scala/sql-rdd-source.txt")
10
       val rowRDD = linesRDD.map(line => {
11
12
         val splits = line.split(",")
13
         Row(splits(0).trim.toInt,splits(1).trim,splits(2).trim.toInt,splits(3).trim.toInt)
14
        })
15
        val structType = StructType(Array(
          DataTypes.createStructField("id",DataTypes.IntegerType,true),
16
17
         DataTypes.createStructField("name",DataTypes.StringType,true),
         DataTypes.createStructField("age",DataTypes.IntegerType,true),
18
         DataTypes.createStructField("height",DataTypes.IntegerType,true)
19
20
        ))
21
22
        val df = sqlContext.createDataFrame(rowRDD,structType)
23
       df.registerTempTable("person")
       df.show()
24
25
          * 接下来对df中的数据进行查询
26
          * 第一个查询年龄的最大值,平均值
27
          * height的总身高
28
29
30
        sqlContext.sql("select avg(age) from person").show()
        sqlContext.sql("select max(age) from person").show()
31
        sqlContext.sql("select sum(height) from person").show()
32
33
34
     }
35 }
```

Java 版本



单片机学习



Mtp://wpa.dg://blog.csdprd? 2篇 v=3&uin=2431299880&site=qq&r webmaster@ttsgr/.hetog.csd... 13篇 (mailto:webmaster@csdn.net) 2016年11月 (http://blog.csd... (文) 微博客服

(http://e.weibo展安m/csdnsupport/r

他的疾病 广告服务





用python读与Exe高的方法9M24637表log. csdn.nettypa/www.03834aagiolendetails/ 51374497)

22220世网站备案信息

(Washing And Manager 1994) (Washing Manager tbialotgaosd01.020020004_018610083301/a (**):阿绍e 1封接無疑。345)/www.cyberpolice.cn/)

11041

Fluxion无线攻击 (搜了全网都是这一 篇) (http://blog.csdn.net/qq_1610333 1/article/details/52252646)

10486

Spark SQL 函数操作 (http://blog.csdn.n et/qq_16103331/article/details/53572 627)

3419

Java通过反射获取带参数构造方法并使用 (http://blog.csdn.net/qq_16103331/ar ticle/details/51926862)

484

 \triangle 内容举报



```
1 public class SparkSQLFunctionJava {
2 public static void main(String[] args) {
       SparkConf conf = new SparkConf( ).setAppName(SparkSQLFunctionJava.class.getSimpleNam
   e()).setMaster("local");
       JavaSparkContext sc = new JavaSparkContext(conf);
       SQLContext sqlContext = new SQLContext(sc);
7
       JavaRDD<String> linesRDD = sc.textFile("E:/test/scala/sql-rdd-source.txt");
                        8
q
       JavaRDD<Row> rowRDD = linesRDD.map(new Function<String, Row>() {
10
           @Override
11
            public Row ca⊞(String s) throws Exception {
12
               String splits[] = s.split(",");
13
               return RpmFactory.create(Integer.valueOf(splits[0].trim()),splits[1].trim
    (), Integer.valueOf(splits[2].trim()), Integer.valueOf(splits[3].trim()));
14
15
           }
                        ···
16
       });
17
       StructField structFields[] = new StructField[4];
        structFields[0] = DataTypes.createStructField("id",DataTypes.IntegerType,true);
19
        structFields[1] = DataTypes.createStructField("name",DataTypes.StringType,true);
        structFields[2] = DataTypes.createStructField("age",DataTypes.IntegerType,true);
20
21
        structFields[3] = DataTypes.createStructField("height",DataTypes.IntegerType,true);
22
       StructType structType = new StructType(structFields);
23
       DataFrame dataFrame = sqlContext.createDataFrame(rowRDD, structType);
24
       dataFrame.show():
25
       dataFrame.registerTempTable("person");
26
        * 接下来对df中的数据进行查询
27
         * 第一个查询年龄的最大值,平均值
28
        * height的总身高
29
        * */
30
31
        sqlContext.sql("select avg(age) from person").show();
        sqlContext.sql("select max(age) from person").show();
33
        sqlContext.sql("select sum(height) from person").show();
34
        sc.close();
35 }
    }
```

修改Spark运行日志级别

cp log4j.properties.template log4j.properties vim conf/log4j.properties 将INFO 修改为ERROR级别需要重启Spark集群,使其生效

HIVE 设置列结构显示

set hive.cli.print.header=true;

Spark SQL 开窗函数

- 1、Spark 1.5.x版本以后,在Spark SQL和DataFrame中引入了开窗函数,比如最经典的就是我们的row_number(),可以让我们实现分组取topn的逻辑。
- 2、做一个案例进行topn的取值(利用Spark的开窗函数),不知道是否还有印象,我们之前在最早的时候,做过topn的计算,当时是非常麻烦的。但是现在用了Spark SQL之后,非常方便。



联系我们

▲ 网站客服 (http://wpa.qq.com/msgrd? v=3&uin=2431299880&site=qq&r ■ webmaster@csdn.net (mailto:webmaster@csdn.net)

关于 招聘 广告服务



©2018 CSDN 京ICP证09002463号 (http://www.miibeian.gov.cn/)

② 经营性网站备案信息

⚠
内容举报

(記) (返回)(近部

```
1
2
      * Spark SQL 开窗函数之 row number
3
   object SparkSQLOpenWindowFunction {
5
      def main(args: Array[String]): Unit = {
        val conf = new SparkConf().setAppName("SparkSQLOpenWindowFunction").setMaster("loca
6
   1")
7
        val sc = new SpankContext(conf)
8
9
        val sqlContext = new HiveContext(sc)
       val linesRDD = sc_textFile("E:/test/scala/topn.txt")
10
        val rowRDD = linesRDD.map(line =>{
11
12
         val splits = line.split(" ")
13
         Row(splits(0).trim,splits(1).trim.toInt)
14
        })
        val structType = __taTypes.createStructType(Array(
15
          DataTypes.createStructField("class",DataTypes.StringType,true),
16
17
         DataTypes.createStructField("score",DataTypes.IntegerType,true)
18
        ))
19
        val df = sqlContext.createDataFrame(rowRDD,structType)
       df.registerTempTable("stu_score")
20
21
         * 查询操作
22
          * 先按照class 进行分组,然后对每个class分组中的数据求出TOP3
23
          * */
24
        val topNDF = sqlContext.sql("select temp.* from (select *, row_number() over(partit
25
   ion by class order by score desc) rank from stu_score ) temp where temp.rank < 4")
26
27
        topNDF.show()
28
        sc.stop()
29
     }
    }
```

Java版本

```
1 public class SparkSQLOpenWindowFunctionJava {
   public static void main(String[] args) {
        {\tt SparkConf conf = new SparkConf().setAppName(SparkSQLOpenWindowFunctionJava.class.get)} \\
4
   SimpleName()).setMaster("local");
5
        JavaSparkContext sc = new JavaSparkContext(conf);
6
       HiveContext sqlContext = new HiveContext(sc);
       JavaRDD<String> linesRDD = sc.textFile("E:/test/scala/topn.txt");
7
8
       JavaRDD<Row> rowRDD = linesRDD.map(new Function<String, Row>() {
9
10
11
            public Row call(String s) throws Exception {
                return RowFactory.create(s.split(" ")[0].trim(),Integer.valueOf(s.split(" ")
12
    [1].trim()));
13
14
            }
15
        StructField structFields[] = new StructField[2];
16
17
        structFields[0] = DataTypes.createStructField("class",DataTypes.StringType,true);
18
        structFields[1] = DataTypes.createStructField("score",DataTypes.IntegerType,true);
19
        StructType structType = new StructType(structFields);
20
       DataFrame df = sqlContext.createDataFrame(rowRDD, structType);
21
       df.registerTempTable("stu_score");
22
        * 查询操作
23
         * 先按照class 进行分组,然后对每个class分组中的数据求出TOP3
24
25
       DataFrame dataFrame = null;
26
27
       dataFrame = sqlContext.sql("select temp.* from (select *, row number() over(partitio
28 n by class order by score desc) rank from stu_score ) temp where temp.rank < 4");
29
       dataFrame.show():
30
        sc.close():
    }
    }
```



单片机学习



联系我们

▲ 网站客服
(http://wpa.qq.com/msgrd?
v=3&uin=2431299880&site=qq&r

■ webmaster@csdn.net
(mailto:webmaster@csdn.net)

关于 招聘 广告服务





©2018 CSDN 京ICP证09002463号 (http://www.miibeian.gov.cn/)

② 经营性网站备案信息

⚠
内容举报



在spark-sql 下运行,

- 先在hive 中,建一张表, create table topn (class string, score int) row format delimited fields terminated by '
- 然后把数据导入表中, load data local inpath '/opt/data/spark/topn.txt'
- 然后就可以利用开窗函数进行分组排序了, select temp.x from (select *,row_number() over(partition class order/4py score desc) rank from topn) temp where temp.rank < 4;

UDF自定义函数

≔

1、UDF: User Defined Function。用户自定义函数。 我们通常所说的UDF自定义函数,就是一对一的关系:一个输入参数和一个输出参数 创建UDF的步骤:

- 1、先创建一个自定义的函数fund
- 2、使用sqlContext.udf().register("起个名字", func_)
- 3、在我们的sql中直接使用就行了

```
1 object SparkSQLUDFOps {
   def main(args: Array[String]): Unit = {
       val conf = new SparkConf().setAppName("SparkSQLUDFOps").setMaster("local")
3
       val sc = new SparkContext(conf)
       val sqlContext = new SQLContext(sc)
7
       val linesRDD = sc.textFile("E:/test/scala/sql-rdd-source.txt")
8
       val rowRDD = linesRDD.map(line => {
9
           val splits = line.split(", ")
            Row(splits(0).trim.toInt, splits(1).trim, splits(2).trim.toInt, splits(3).trim.t
10
   oInt)
11
12
       })
13
       val structType = StructType(Array(
           DataTypes.createStructField("id", DataTypes.IntegerType, true),
14
           DataTypes.createStructField("name", DataTypes.StringType, true),
15
           DataTypes.createStructField("age", DataTypes.IntegerType, true),
16
17
           DataTypes.createStructField("height", DataTypes.IntegerType, true)
18
       val df = sqlContext.createDataFrame(rowRDD, structType)
19
       df.registerTempTable("person")
20
21
       //2、注册自定义的UDF
22
       /**
         * 这是两种注册的方式
23
24
       sqlContext.udf.register("myLen", myLen _)
25
       sqlContext.udf.register("len", (str:String, len:Int) => str.length > len)
26
27
28
       sqlContext.sql("select id, name, myLen(name) as len from person where len(name,
29
    5)").show()
30
       sc.stop()
31
   }
   //1、创建一个自定义的函数,用于求出字符串的长度
    def myLen(str: String) = str.length
```



单片机学习









联系我们

网站客服 ● ^{网站各版} (http://wpa.qq.com/msgrd? v=3&uin=2431299880&site=qq&r webmaster@csdn.net

(mailto:webmaster@csdn.net) (http://e.weibo.com/csdnsupport/r

400-660-0108





©2018 CSDN 京ICP证09002463号 (http://www.miibeian.gov.cn/)

🝘 经营性网站备案信息

(http://www.hd315.gov.cn/beian/view.asp? bianhao=010202001032100010) 110报警服务 (http://www.cyberpolice.cn/)

> \triangle 内容举报



}

发表你的评论

(http://my.csdn.net/xuhao7891)

Spark SQL内置函数的使用Spark SQL执行计划总结

个人感觉Spark SQL的优势在于UDF和DataFrame(RDD的封装,由 🦳 UnionIBM 2017年01月18日 16:46 🕮 4232 于RDD是可以是任意类型,因此可以封装一个对象在RDD里面,spa

rk2.0以后对DataSet做了优化,由于DataF...

(http://blog.csdn.net/UnionIBM/article/details/54601750)

Spark实战-Spark SQL ()

wizheyouye 2016年02月06日 18:02 □ 1993

本文是按照周志胡老师的博客实践的,再次特别感谢周老师!Spark修炼之道(进阶篇)——Spark入门到精通:第十节Spar k SQL案例实战(一) http://blog.csdn.net/l...

(http://blog.csdn.net/yizheyouye/article/details/50640620)

程序员不会英语怎么行?



老司机教你一个数学公式秒懂天下英语

Spark SQL的使用

─ hddyxl 2016年03月10日 23:35 □ 211

本文主要给出了Spark1.4之前和之后IDE编译的区别

(http://blog.csdn.net/hddyxl/article/details/50851080)

Spark SQL 初探

🚱 u014388509 2014年07月02日 19:13 🕮 7259

早在Spark Submit 2013里就有介绍到Spark SQL,不过更多的是介绍Catalyst查询优化框架。经过一年的开发,在今年Spar k Submit 2014上, Databricks宣布...

(http://blog.csdn.net/u014388509/article/details/36440821)

spark sql udf 解析json数据表的嵌套数组

🤷 baiyangfu 2014年12月15日 19:03 🕮 3368

使用spark sql 的时候,加载json数据作为表的基础数据。发现其中的json数据中有些字段对应的是一个嵌套数组:{ "sche ma1": "hello", "schema2": [{"a":1...

(http://blog.csdn.net/baiyangfu/article/details/41946447)

不再死记硬背,一个公式学懂英文



英语长难句解读,记住这个公式就够了!

一、由于具有多张宽表且字段较多,每个宽表数据大概为4000万条,根据业务逻辑拼接别名,并每张宽表的固定字段进行left join 拼接SQL。这样就能根据每个宽表的主列,根据每个宽表的不同字段关联出一...

(http://blog.csdn.net/xiaolang85/article/details/51750855)

sparksql各种数据源

🕌 HANLIPENGHANLIPENG 2016年12月15日 22:04 🕮 1758

sparksql各种数据源的测试: 大致的有json文件parquet文件,和常用的文件,jdbc等还有hbase的数据源(还没有贴出, 可能要等几天贴出来了)代码:一般过程:第一步创建:...

(http://blog.csdn.net/HANLIPENGHANLIPENG/article/details/53677697)

SparkSQL之读取数据

tangshiweibbs 2017年04月19日 13:35 435

def main(args: Array[String]): Unit = { val conf = new SparkConf().setMaster("local").setAppName("...

(http://blog.csdn.net/tangshiweibbs/article/details/70239386)











联系我们

♣ 网站客服 (http://wpa.qq.com/msgrd? v=3&uin=2431299880&site=qq&r webmaster@csdn.net

(mailto:webmaster@csdn.net)

(http://e.weibo.com/csdnsupport/r 400-660-0108

关于 招聘 广告服务

©2018 CSDN 京ICP证09002463号 (http://www.miibeian.gov.cn/)

🝘 经营性网站备案信息

(http://www.hd315.gov.cn/beian/view.asp? bianhao=010202001032100010)

110报警服务 (http://www.cyberpolice.cn/)

 \triangle 内容举报

TOP 返回顶部

SparkSQL相关语句总结

💖 zhao897426182 2017年10月26日 15:07 🕮 204

1.in 不支持子查询eg. select * from src where key in(select k ey from test); 支持查询个数 eg. select * from src wh...

(http://blog.csdn.net/zhao897426182/article/details/78354083)

spark sql根本使用方法介绍

doupeihua 2016年08月10日 13:04 Q 2351

0-...

spark sql基本使用方法介绍 spark中可以通过spark sql 直接查询hive或impala中的数据 , 一、启动方法 /data/spark-1.4.

(http://blog.csdn.net/doupeihua/article/details/52170498)







SparkSQL(下)--Spark实战应用

gongpulin 2016年06月30日 16:02 □ 1756

1、运行环境说明 1.1 硬软件环境 | 主机操作系统: Windows 64位, 双核4线程, 主频2.2G, 10G内存 | 虚拟软件: VMwar e® Workstation 9.0...

(http://blog.csdn.net/gongpulin/article/details/51791163)

联系我们

♣ 网站客服 (http://wpa.qq.com/msgrd? v=3&uin=2431299880&site=qq&r

webmaster@csdn.net (mailto:webmaster@csdn.net)

(http://e.weibo.com/csdnsupport/r 400-660-0108

关于 招聘 广告服务





©2018 CSDN 京ICP证09002463号 (http://www.miibeian.gov.cn/)

🝘 经营性网站备案信息

(http://www.hd315.gov.cn/beian/view.asp? bianhao=010202001032100010) 110报警服务 (http://www.cyberpolice.cn/)

Spark读取数据库(Mysql)的四种方式讲解

⑥ LW_GHY 2016年03月20日 21:47 ☐ 2683

一、不指定查询条件 这个方式链接MySql的函数原型是: def jdbc(url: String, table: String, properties: Properties): Data

(http://blog.csdn.net/LW_GHY/article/details/50939091)

SparkSQL之查询,过滤

tangshiweibbs 2017年04月19日 13:39

673

def main(args: Array[String]): Unit = { val conf = new SparkConf().setMaster("local").setAppName("...

(http://blog.csdn.net/tangshiweibbs/article/details/70239416)

第71课:Spark SQL窗口函数解密与实战学习笔记

第71课:Spark SQL窗口函数解密与实战学习笔记 本期内容: 1 Spark 🦳 slq1023 2016年04月13日 00:40 🕮 3308 SQL窗口函数解析 2 SparkSQL窗口函数实战 窗口函数是Spark内置函 数中最有价值的函数,因为...

(http://blog.csdn.net/slq1023/article/details/51138709)

⑥ bitcarmanlee 2016年07月23日 21:30 □ 6609

1.spark sql简介spark sql是为了处理结构化数据的一个spark 模块。不同于spark rdd的基本API, spark sql接口更多关于数

spark sql 入门详解

据结构本身与执行计划等更多信息。在spar...

(http://blog.csdn.net/bitcarmanlee/article/details/52006718)

Spark SQL UDF使用

🚱 u014388509 2014年09月19日 18:17 🕮 12900

spark sql udf编写及使用

(http://blog.csdn.net/u014388509/article/details/39401391)

12.Spark SQL: 开窗函数以及top3销售额统计案例实战

Spark 1.4.x版本以后,为Spark SQL和DataFrame引入了开窗函数,比如最经典,最常用的,row_number(),可以让我们实 现分组取topn的逻辑。 案例:统计每个种类的销售...



🌀 weixin_32265569 2017年11月02日 17:23 🕮 212

(http://blog.csdn.net/weixin_32265569/article/details/78427180)

\triangle 内容举报



SparkSQL---开窗函数(java)

(ymf827311945 2017年08月11日 17:47 및 796

package com.spark.sparksql.save;import org.apache.spark.SparkConf; import org.apache.spark.api.java....

(http://blog.csdn.net/ymf827311945/article/details/77101554)

SparkSQL中DataFrame Operations操作 (select、filter、groupBy、count)

代码+图文讲解SparkSQL DataFrame操作, select、filter、groupBy、count操作

🦲 RiverCode 2017年03月07日 14:38 🕮 4674

(http://blog.csdn.net/RiverCode/article/details/60764116)

SparkSQL之函数的操作

package com.uplooking.bigdata.sql.p3 import com.uplooking.bigdata.utils.MySparkUtil import org.apac...

(http://blog.csdn.net/tangshiweibbs/article/details/70254898)

 \odot



联系我们

♠ 网站客服 (http://wpa.qq.com/msgrd? v=3&uin=2431299880&site=qq&r webmaster@csdn.net (mailto:webmaster@csdn.net)

(http://e.weibo.com/csdnsupport/r 400-660-0108

关于 招聘 广告服务



©2018 CSDN 京ICP证09002463号 (http://www.miibeian.gov.cn/)

🕜 经营性网站备案信息

(http://www.hd315.gov.cn/beian/view.asp? 110报警服务 (http://www.cyberpolice.cn/)

> \triangle 内容举报

