# Spark Sql 部分

## Spark 2.0里变化

1. x版本里Spark Sql的入口类是SqlContext

Hive Sql的入口类是HiveContext，兼容hive特性和支撑。

2.0里统一用SparkSession，目的是编码统一，且功能更灵活。

查询结果类用DataSet，不再用DataFrame，DataFrame相当于DataSet(Row)

DateSet相当于之前Java开发读数据库的ResultSet类。

#### type **DataFrame** = [Dataset](http://spark.apache.org/docs/2.0.0-preview/api/scala/org/apache/spark/sql/Dataset.html)[[Row](http://spark.apache.org/docs/2.0.0-preview/api/scala/org/apache/spark/sql/Row.html)]

## SparkSession的数据源来自哪里？

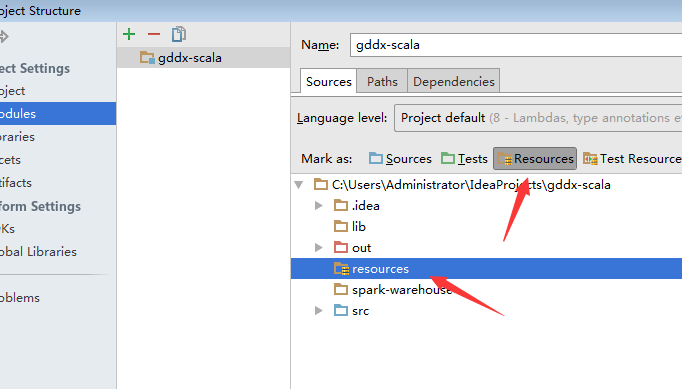
1. 程序内部，把RDD映射为临时数据表（视图），用sql进行查询。

case class可以隐式转为DataSet

*s函数，拼接字符串用的*

*//s函数多行的语法*

**val** sql =  
 **s""" select \* from log  
 | where rid<=$**top **| and 1=1  
 """**.stripMargin



1. 读外部数据，比如HDFS、Hive、Json、Parquet文件等。

## Spark sql 简单案例

dataFrame.rdd 转为普通rdd

普通rdd通过.toDF() 转为dataFrame（需要先**import** ss.implicits.\_

）

## Spark sql实现电商流量统计

|  |  |  |  |
| --- | --- | --- | --- |
| **指标名** | **全称** | **描述** | 口径 |
| UV | User views | 独立访客 | count(distinct guid)  去重型 |
| PV | Page views | 页面访问数 | count(url)   汇总型 |
| 二跳率 |  | 衡量页面吸引度或和用户契合度 | select  count(distinct case when pv>=2 then sessionid else null end)/count(distinct sessionid)  from  (select count(url) pv,session\_id from from track\_log where date='2018-01-01'  and lengnth(url)>5 group by session\_id) a group by date; |

select count(distinct guid) uv,sum(pv) pv,

count(case when pv>=2 then sessionid else null end) second\_num,

count(sessionid) visits from

(select ds date,sessionid,max(guid) guid,count(url) pv from track\_log where ds='$date' and hour='18'

and length(url)>0

group by ds,sessionid) a

group by date

郑博 20:49:29

执行性能上有很大差距不 这两种写法

答：流量统计的rdd方式和sql方式性能上哪个好？个人认为简单逻辑时，rdd方式性能好，复杂逻辑时sql性能好。

吴卓钧 20:55:09

SparkContext和SparkSession是什么关系？

答：SparkContext 是创建普通RDD的驱动类

SparkContext.parallelize 创建普通rdd

SparkContext.textFile 读取外部数据

SparkSession是创建DataFrame的驱动类

SparkSession.createDataFrame 通过rdd或者List创建DataFrame

陈少彬 21:03:23

数据在toDF的过程中进行过滤和在sparksql中用where条件进行过滤，哪种性能更优？是不是要看总数据量的大小以及过滤的数据量大小而定？

答：toDF能过滤就过滤，当然最好。

## Spark sql 读取Json、Parquet文件

Parquet 是基于列存储的一种文件格式

如果大家遇到读win下的文件失败，就把文件上传到linux上测试，程序打包在linux上执行。

黄锡恩 21:31:29

thrift 你不是说过大量数据会把服务拖死吗？

答：通过thrift服务执行Hive ql 是不建议的，即不建议通过HiveServer2访问Hive；

访问Hive推荐用Hive Cli，即hive -e -f -i

Spark读hive是通过thrift://master:9083 读Hive，容易把thrift累死。

推荐用spark-sql ，同hive Cli

spark-sql -e -f -i

注：把hive-site.xml 放到/opt/soft/spark2.0/spark-2.2.0-bin-hadoop2.6/conf 下，否则无法找到Hive。

spark-sql 是Apache Spark下自带的，cdh中去掉该命令（因为cdh推荐用该公司出品的Impala）。

小结：

1. 能用Spark sql实现的，就不写RDD方式；
2. 能用spark-sql 命令操作，就不用Scala或Java方式。

全脚本（shell+hql）是易开发和维护。

Hive ql作业可以一行不变放在spark-sql 命令中执行，运行在Spark计算引擎上。