**课后扩展题作业**

# 实操五、藏在数据中的秘密。

1. 扩展题

开发新的自定义函数，对身份证号码脱敏处理，把前后4位变为\*号输出，如

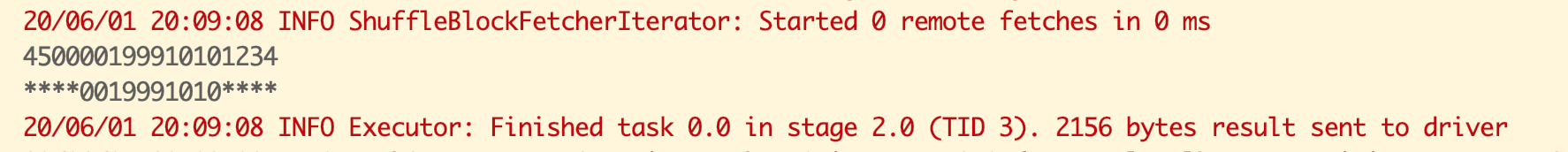
440234201203024325 变为 \*\*\*\*3420120302\*\*\*

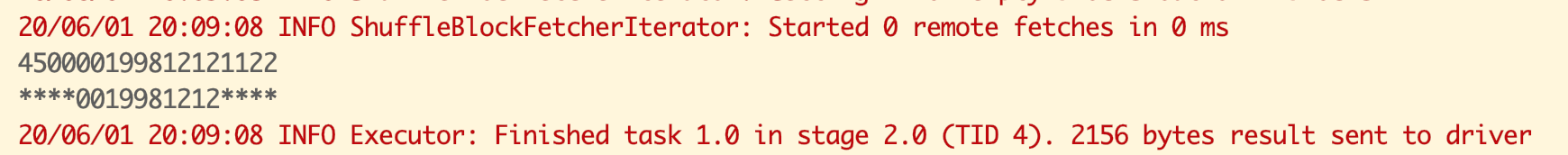
1. 关键代码

//Demo.scala文件  
 *//设置读取的文件* **val** row = sc.textFile(**"/Users/boscoli/Downloads/大数据实训\_品高/2/user\_data.txt"**).map(\_.split(**","**)).map(p => *Row*(p(0).toInt, p(1), p(2), p(3).toLong, p(4))).toDF()  
 row.registerTempTable(**"row"**)  
 spark.udf.register(**"mosaic"**, (str: String) => Utils.*makePicMosaic*(str))*//注册匿名函数* spark.udf.register(**"star"**, (str: String) => Utils.*makeStar*(str))  
**val** result = sql(**"SELECT id, name, sex, idCard, pic, star(idCard) FROM row ORDER BY id"**)  
 result.foreach(x => {  
 Utils.*makeStar*(x(3).toString)』

//Utils.java文件  
**public static** String makeStar(String inputBase64) **throws** Exception {  
 System.***out***.println(inputBase64);  
 String tmp = **"\*\*\*\*"**;  
 **for** (**int** i = 4; i < inputBase64.length()-4; i++) {  
 tmp += inputBase64.charAt(i);  
 }  
 tmp += **"\*\*\*\*"**;  
 System.***out***.println(tmp);  
 **return "\*\*\*\*"** + tmp + **"\*\*\*\*"**;  
}

1. 结果截图





# 实操七、大数据框架打擂台。

1. 扩展题

仿造坐牢人员数据，建一张拘留记录表数据，把正在拘留人员也排除在嫌疑人外。

表结构如：身份证号码、姓名、拘留开始时间、拘留结束时间，拘留原因

1. 关键代码

1）建表SQL脚本并写入拘留人员数据

drop table if exists t\_zl\_jbxx;

create table t\_zl\_jbxx(

sfzhm varchar(50),

xm varchar(50),

jlkssj timestamp,

jljssj timestamp,

jlyy varchar(50)

);

comment on table t\_zl\_jbxx is '坐牢人员基本信息表';

comment on column t\_zl\_jbxx.sfzhm is '身份证号码';

comment on column t\_zl\_jbxx.xm is '姓名';

comment on column t\_zl\_jbxx.jlkssj is '拘留开始时间';

comment on column t\_zl\_jbxx.jljssj is '拘留结束时间';

comment on column t\_zl\_jbxx.jlyy is '拘留原因';

INSERT INTO t\_zl\_jbxx(sfzhm, xm, jlkssj, jljssj,jlyy) VALUES ('522730199712062028', '江访波', '2022-04-21 10:17:06', '2023-04-21 10:17:06','猥亵儿童妇女');

INSERT INTO t\_zl\_jbxx(sfzhm, xm, jlkssj, jljssj,jlyy) VALUES ('511823197607191327', '曾美伊', '2022-03-21 10:17:06', '2023-03-21 10:17:06','故意杀人');

INSERT INTO t\_zl\_jbxx(sfzhm, xm, jlkssj, jljssj,jlyy) VALUES ('430211200805203811', '夏佳音', '2023-02-21 10:17:06', '2023-02-21 10:17:06','强奸罪');

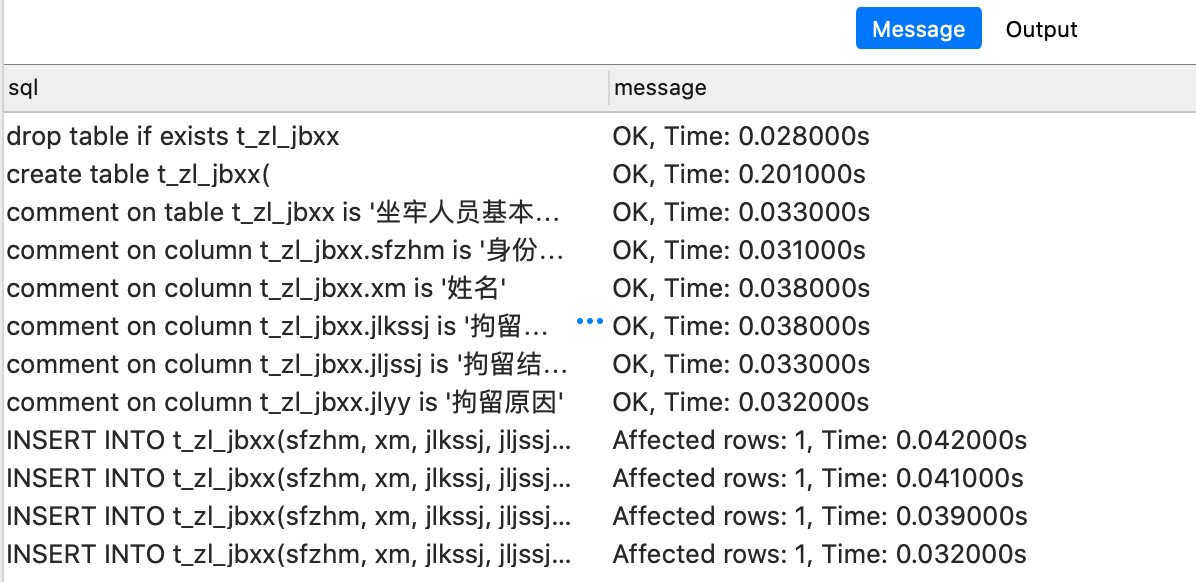
INSERT INTO t\_zl\_jbxx(sfzhm, xm, jlkssj, jljssj,jlyy) VALUES ('430211200805203832', '张三', '2023-02-21 10:17:06', '2023-02-21 10:17:06','强奸罪');

2）查询脚本：嫌疑犯与拘留人员的关系

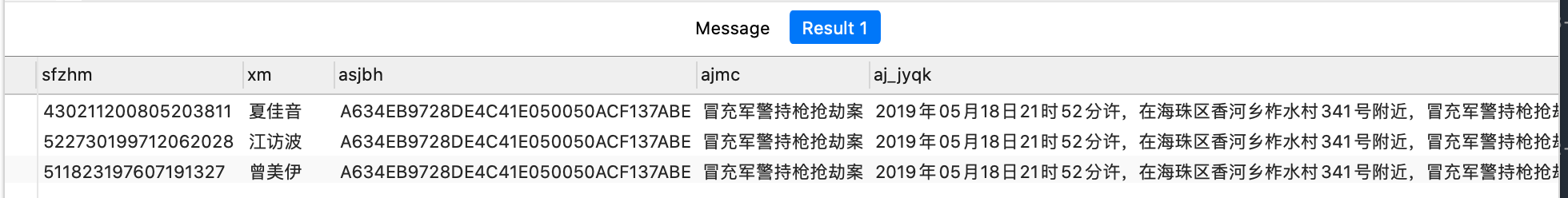
select \* from t\_rk\_jbxx\_result t where t.sfzhm in (select sfzhm from t\_zl\_jbxx where jljssj > now());

1. 结果截图

1）建表结果



2）查询结果



# 实操八、算完之后怎么办？

1. 扩展题

在Greenplum创建一个表，从spark把数据导出到刚刚创建的Greenplum表中。

1. 脚本
2. Greenplum创建数据表

create table t\_rk\_jbxx\_result\_2(

sfzhm varchar(50) ,

xm varchar(50) ,

asjbh varchar(50) ,

ajmc varchar(50) ,

aj\_jyqk varchar(500));

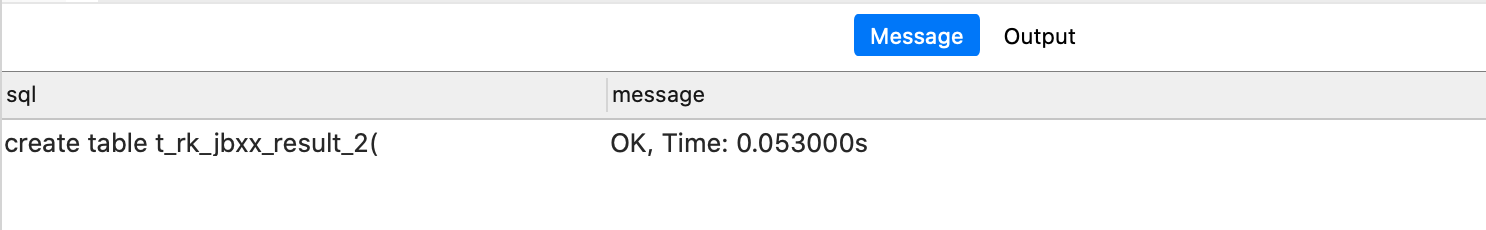
2） 将数据导入到Greenplum表中

sqoop export --connect jdbc:postgresql://bigdata28.depts.bingosoft.net:25432/user02\_db --driver org.postgresql.Driver --password pass@bingo2 --username user02 --table t\_rk\_jbxx\_result\_2 --num-mappers 1 --hcatalog-database user02\_db --hcatalog-table gp\_s3;

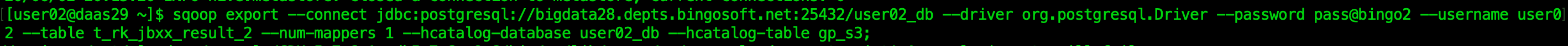
3）查询Greenplum中刚建的表

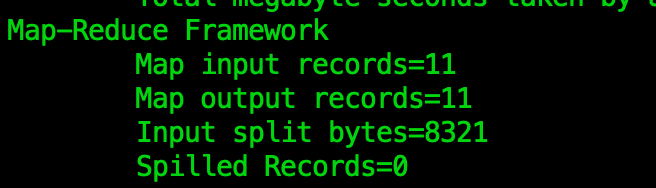
SELECT \* FROM t\_rk\_jbxx\_result\_2;

1. 结果截图
2. 建表结果



1. 脚本运行





3）查询Greenplum中的表

