捷豹数据库和
Zeppelin 数据可视化集成
测试报告

Jaguar – Zeppelin Integration
and
Data Visualization

Feb. 5, 2018

摘要

我们搜集了一些银行存款客户的数据，是银行采集的贷款人的基本信息， 包括存储人年龄，职务，婚姻状况，教育水平，银行储蓄，有无持有不动产，联系方式和其他一些相关的基本信息。这些信息存入捷豹数据库，并进行数据分析，基本动机是要了解这些银行客户的相关信息，比如：“年龄跟储蓄的关系”，” 教育程度于储蓄的关联”等。从捷豹数据库输出 SQL 的结果， 再通过集成 Zeppelin 数据可视化以后，可以更加直观的理解数据的实际意义。

测试环境

本捷豹数据库和 Zeppelin 的数据可视化集成测试是在三台 Linux 机器的分布式集群上进行的。

每一台 Linux 的配置:

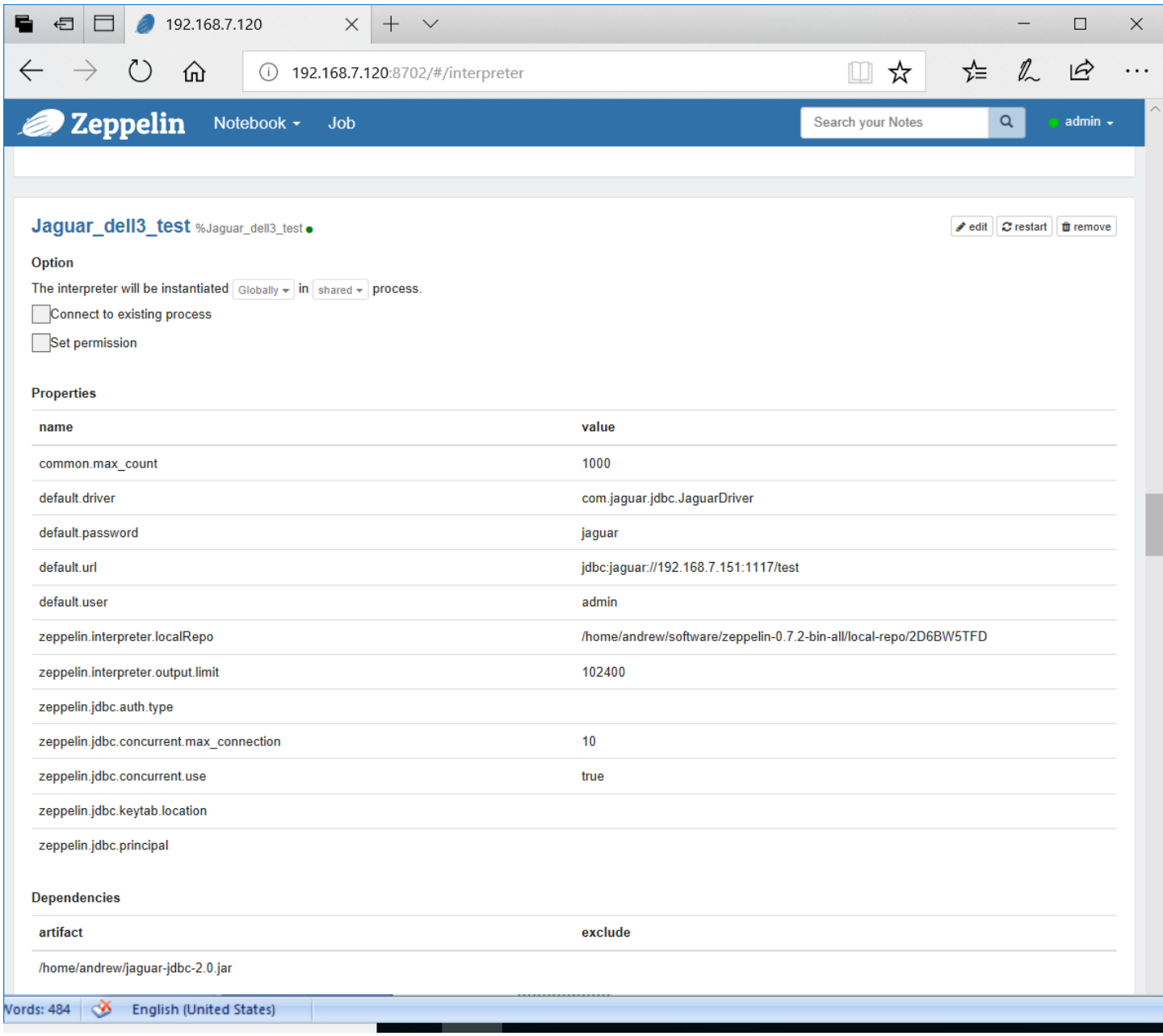
OS	CentOS 6 2.6.32-573.el6.x86_64
CPU	16 核
内存	72 GB
硬盘	5.4 TB

捷豹数据库版本： 2.8.5.2

Zeppeline **版本**: 0.7.2

Zeppelin 面向捷豹数据库的解释器配置

搭建 Zeppelin 可视化软件系统， 并对它的 Interpreter 解释器进行配置。下面的图示例配置信息。



可视化数据库数据来源

被 广泛用来作为可视化数据的原始数据来自于：

https://raw.githubusercontent.com/z-o-e/bank_data_analysis/master/bank-full.csv （此数据有许多不同的来源地）。

为便于本数据库的使用，数据已稍作清理（全本的数据是由 scala 来清理的，详情请参考：

https://zeppelin.apache.org/docs/0.5.5-incubating/tutorial/tutorial.html)

CSV 文件的部分内容列举如下：

```
head bank-full.csv
58,management,married,tertiary,no,2143,yes,no,unknown,5,may,261,1,-
1,0,unknown,no
44,technician,single,secondary,no,29,yes,no,unknown,5,may,151,1,-
1,0,unknown,no
33,entrepreneur,married,secondary,no,2,yes,yes,unknown,5,may,76,1,-
1,0,unknown,no
47,blue-collar,married,unknown,no,1506,yes,no,unknown,5,may,92,1,-
1,0,unknown,no
33,unknown,single,unknown,no,1,no,no,unknown,5,may,198,1,-1,0,unknown,no
35,management,married,tertiary,no,231,yes,no,unknown,5,may,139,1,-
1,0,unknown,no
28,management,single,tertiary,no,447,yes,yes,unknown,5,may,217,1,-
1,0,unknown,no
42,entrepreneur,divorced,tertiary,yes,2,yes,no,unknown,5,may,380,1,-
1,0,unknown,no
58,retired,married,primary,no,121,yes,no,unknown,5,may,50,1,-1,0,unknown,no
43,technician,single,secondary,no,593,yes,no,unknown,5,may,55,1,-
1,0,unknown,no
```

原本的数据“bank-full.csv” 仅有 45,211 行， 并没含有主键(Key), 而捷豹数据库必须拥有主键，所以本测试作了修改，加了 uuid 作为主键(Key), 并重复插入 663 次，让数据量达到约 3,000 万条。

数据(bank-full.csv) 是有关银行采集的贷款人的基本信息：年龄，职务，婚姻状况，教育水平，银行储蓄，有无持有不动产，联系方式和其他一些相关的基本信息。

我们的基本动机是要了解这些银行客户的相关信息，比如：“年龄跟储蓄的关系”等。从捷豹数据库输出 SQL 的结果， 再通过集成 Zeppelin 数据可视化以后，可以更加直观的理解数据的实际意义。

捷豹数据库表结构 (Table Structure)

我们创建的捷豹数据库结构如下：

```
table test.bankdata
(
  key:
    id uuid,
  value:
    age int,
    job char(50),
    marital char(10),
    education char(50),
    default_1 char(3),
    balance int,
    housing char(10),
    loan char(10),
    contact char(10),
    day int,
    month char(10),
    duration int,
    campaign int,
    pdays int,
    previous int,
    poutcome char(10),
    y char(10),
    spare_ char(87)
);
```

数据通过 捷豹客户端程序 jag 导入到数据库：

$HOME/jaguar/bin/jag -u admin -p -h localhost:8888

jag:test> load bank.csv into test.bankdata;

共有 3 千万条数据入库。

数据样本和概述

数据导入到数据库后，读取部分数据如下：

```
jaguar:test> select * from test.bankdata limit 5;
id:[000O7IT3mXOPZGKv2P9mmmo44EU3453275771546] age:[32] job:[technician]
marital:[married] education:[secondary] default_1:[no] balance:[1]
housing:[no] loan:[yes] contact:[cellular] day:[29] month:[jul] duration:[123]
campaign:[6] pdays:[-1] previous:[0] poutcome:[unknown] y:[no]
id:[000mR01eduGdnmwiKiOb3wjRoSQ3453202771459] age:[47] job:[technician]
marital:[married] education:[secondary] default_1:[no] balance:[644]
housing:[no] loan:[no] contact:[telephone] day:[28] month:[jan] duration:[54]
campaign:[3] pdays:[160] previous:[13] poutcome:[failure] y:[no]
id:[000mR01eduGdnmwiKiOb3wjRoSQ3453212771174] age:[47] job:[technician]
marital:[married] education:[secondary] default_1:[no] balance:[644]
housing:[no] loan:[no] contact:[telephone] day:[28] month:[jan] duration:[54]
campaign:[3] pdays:[160] previous:[13] poutcome:[failure] y:[no]
id:[000nFHCOFqzkZKzcFqU4lAISHQF3453228571190] age:[41] job:[management]
marital:[married] education:[tertiary] default_1:[no] balance:[5] housing:[no]
loan:[no] contact:[unknown] day:[5] month:[jun] duration:[178] campaign:[1]
pdays:[-1] previous:[0] poutcome:[unknown] y:[no]
id:[000nFHCOFqzkZKzcFqU4lAISHQF3453258571403] age:[41] job:[management]
marital:[married] education:[tertiary] default_1:[no] balance:[5] housing:[no]
loan:[no] contact:[unknown] day:[5] month:[jun] duration:[178] campaign:[1]
pdays:[-1] previous:[0] poutcome:[unknown] y:[no]
Done in 239 milliseconds (5 rows)
```

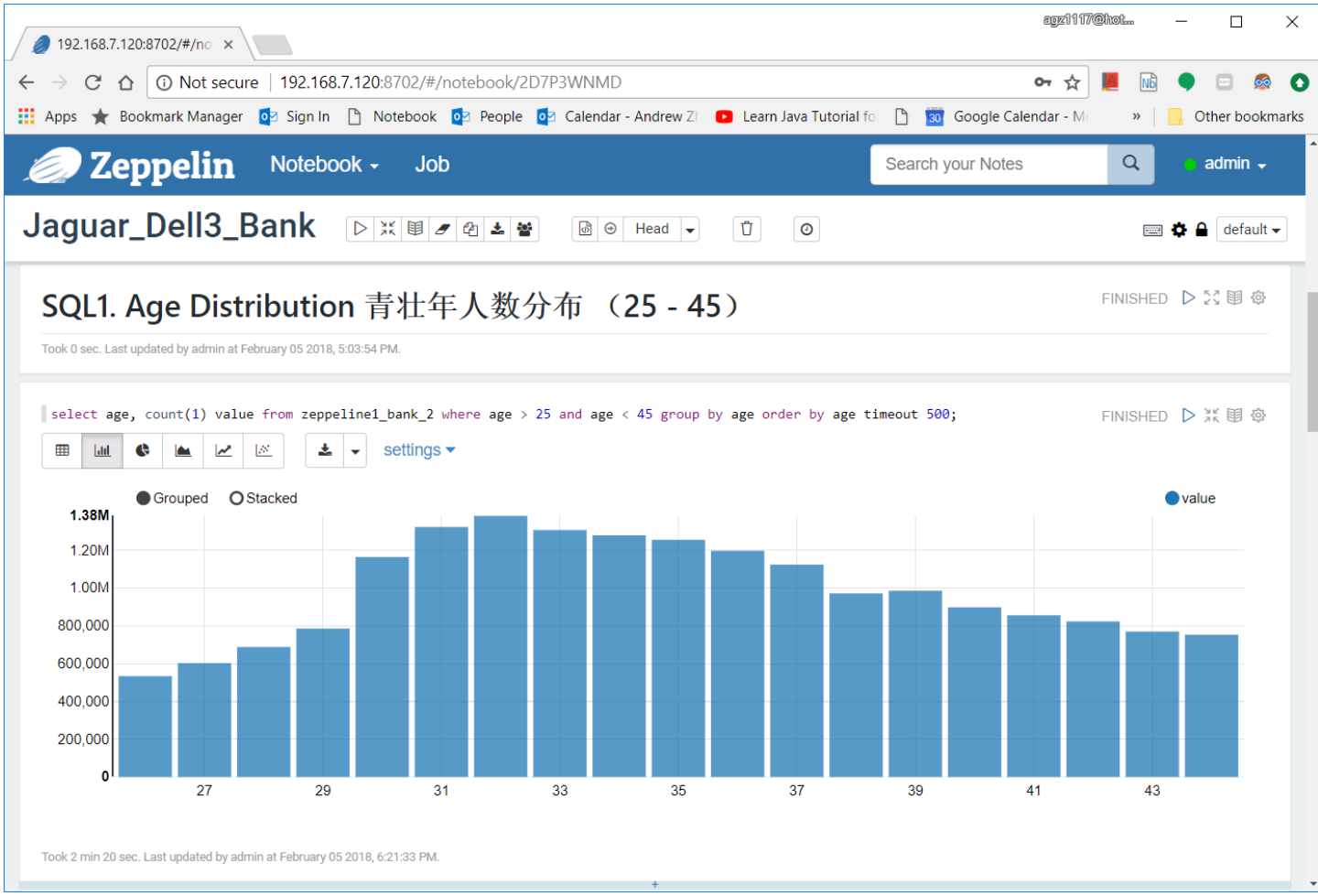
可视化集成测试结果

SQL1. Age Distribution 青壮年人数分布 （25 - 45）

在 Zeppelin 界面输入下面查询语句：

select age, count(1) value from bankdata where age > 25 and age < 45 group by age order by age timeout 500;

点击可视化按钮即可观察根据年龄的分布情况。



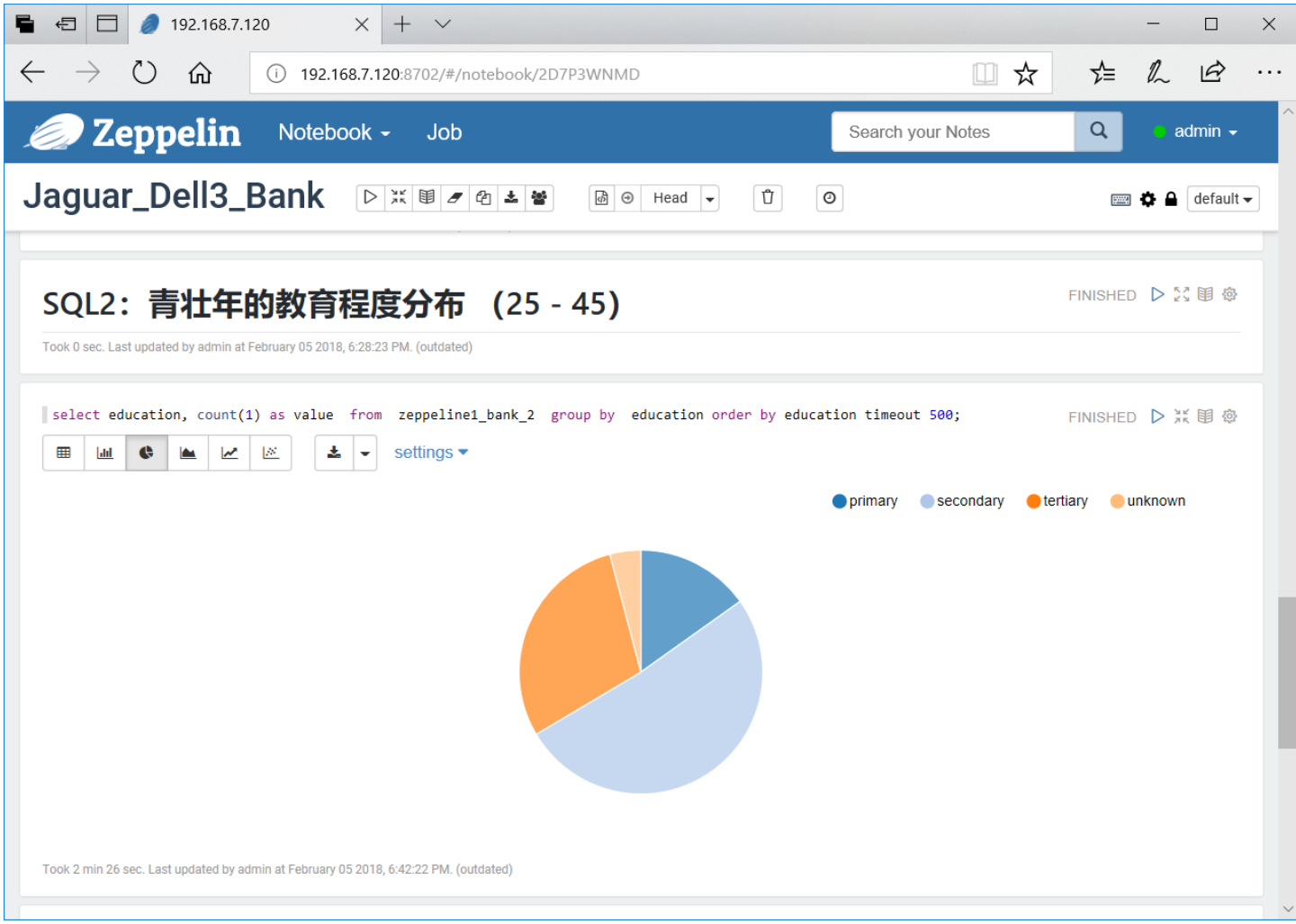
花费时间：2 分 20 秒

SQL2：青壮年的教育程度分布 （25 - 45）

在 Zeppelin 界面输入下面查询语句：

select education, count(1) as value from bankdata group by education order by education timeout 500;

点击可视化按钮即可观察根据教育程度的分布情况。



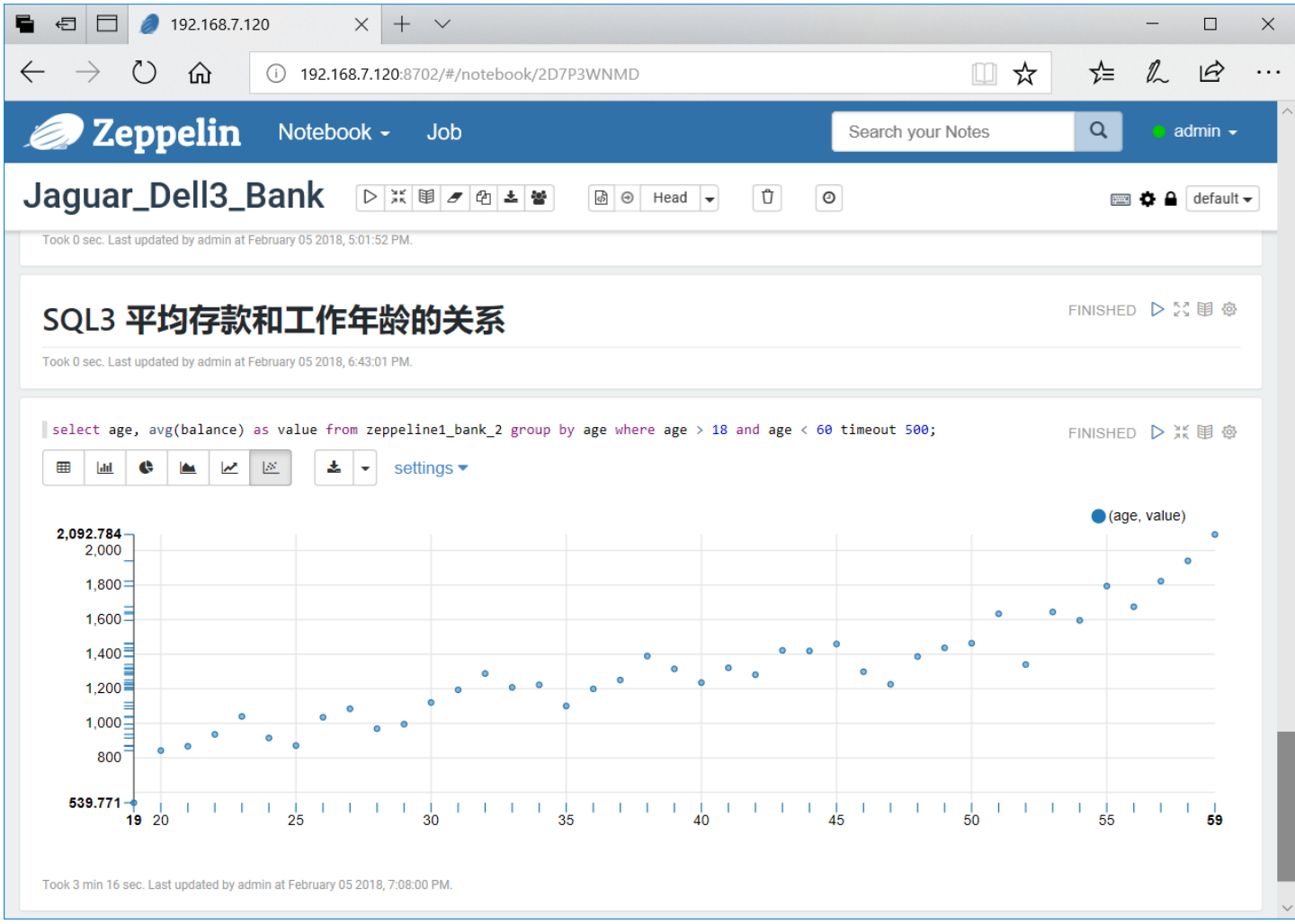
花费时间：2 分 26 秒

SQL3 平均存款和年龄的关系

在 Zeppelin 界面输入下面查询语句：

select age, avg(balance) as value from bankdata group by age where age > 18 and age < 60 timeout 500;

点击可视化按钮即可查看平均存款数量和年龄的分布情况。



花费时间：3 分 16 秒

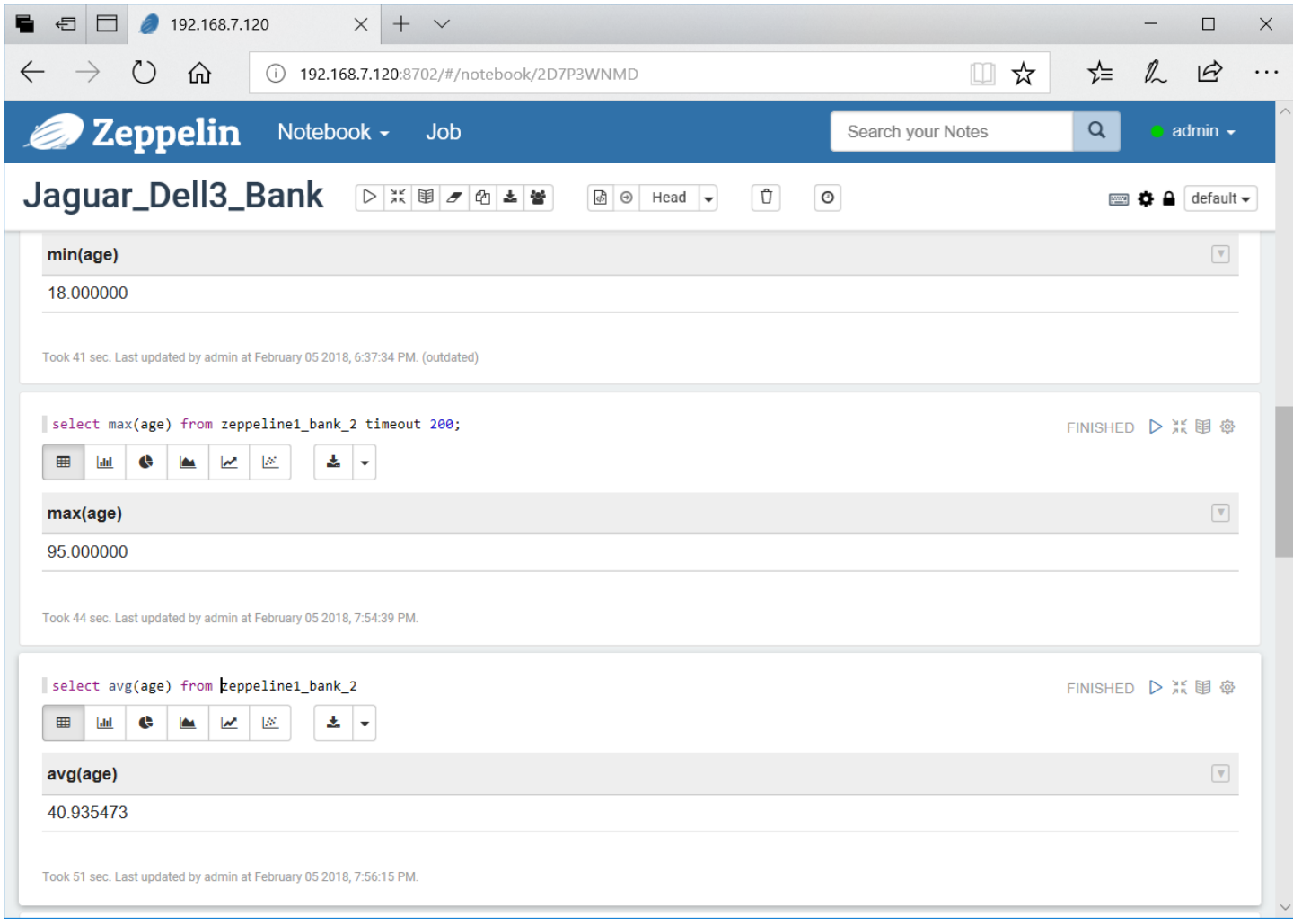
其他函数: 最小，最大，平均和总和

查看最小年龄：

select min(age) from bankdata timeout 200;

查看最大年龄：

select max(age) from bankdata timeout 200;

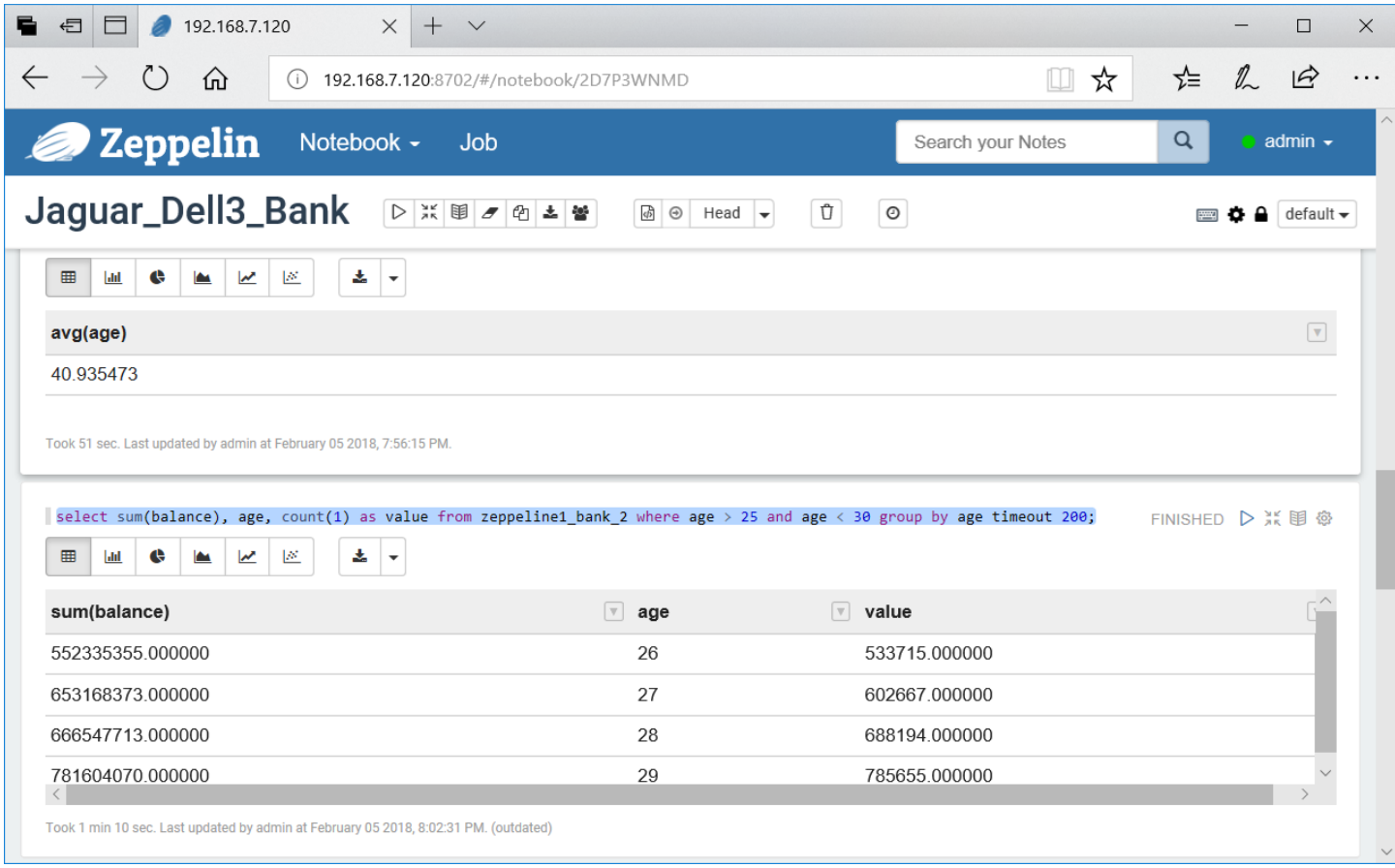


查看平均年龄：

select avg(age) from bankdata

查看根据年龄的分组存款统计：

select sum(balance), age, count(1) as value from bankdata where age > 25 and age < 30 group by age timeout 200;



所需时间大概在 40 到 70 秒之间。

结束语

通过银行实际业务数据证明分布式捷豹数据库能快速导入大量数据，并且利用 SQL 语句对数据进行各种各样的统计分析。结合 Zeppelin 可视化工具，将数据查询结果方便地展现出来，直观的了解数据的隐藏价值。捷豹数据库在 www.datajaguar.com 有更多介绍材料。