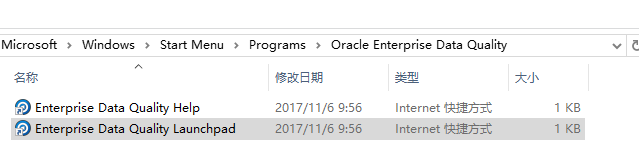
**一：启动EDQ**

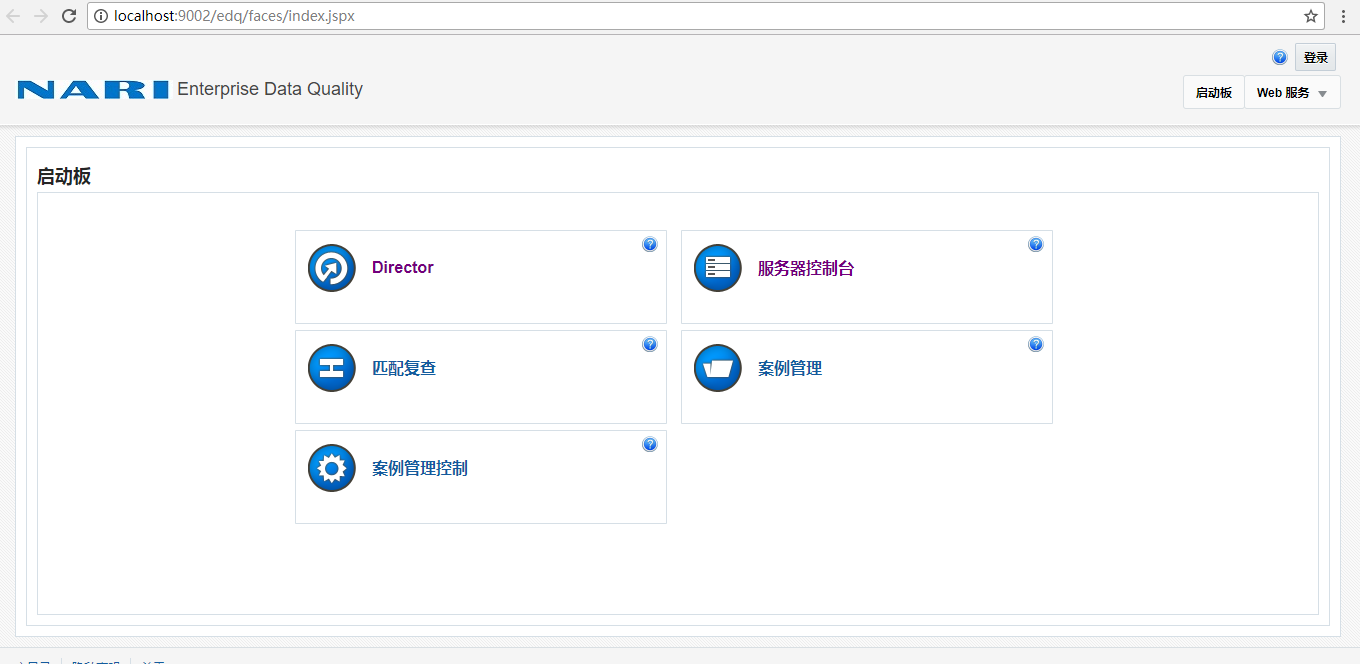
登录到EDQ然后启动director

**请确保您已安装最新的 Java运行环境（JRE7/JRE8）**

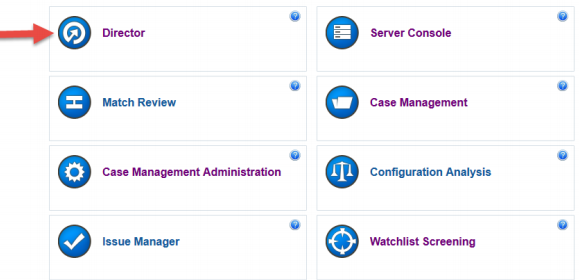
1. 双击开始菜单EDQ目录下的链接Enterprise Data Quality Launchpad



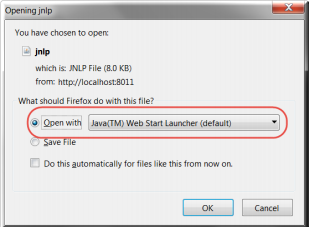
浏览器会打开如下网址:http://localhost:9002/edq/faces/index.jspx



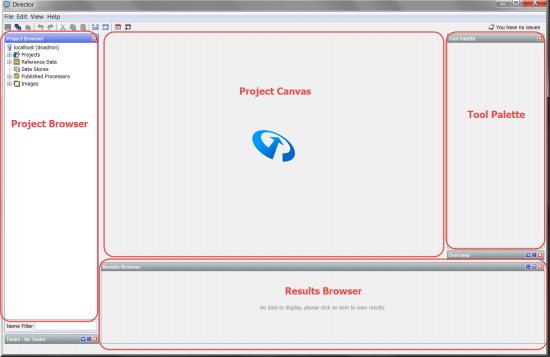
2.单击director图标以启动控制器用户界面



 运行jnlp文件：选择java web start application(javaws.exe)



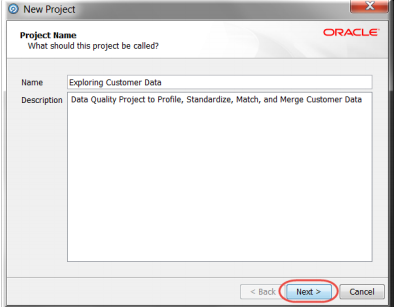
控制器用户界面

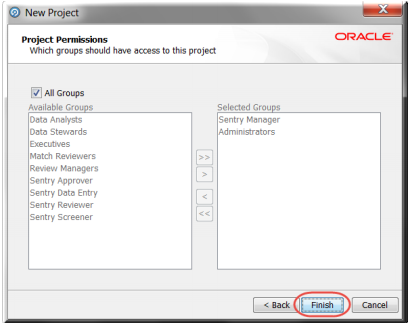
3.以下为用户界面的四个窗口

创建新项目

1. 在**项目浏览器**中, 右键单击**项目**, 然后选择**新建项目，**启动向导

2. 填入**名称**，添加**说明**。单击**下一步 >**继续



1. 确保 "**所有组**" 复选框处于选中状态。这将确保任何用户都可以查看和使用该项目。单击**完成**以创建新项目

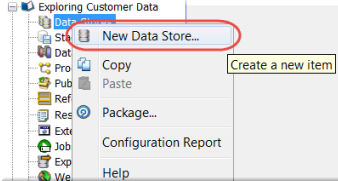
项目现在出现在**项目**列表中!

添加数据存储区

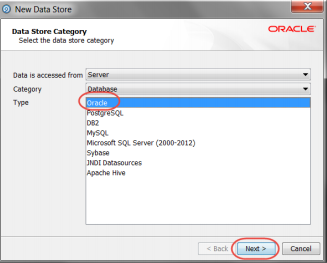
数据存储区存储数据的连接，通常建议通过服务器连接到数据存储区上。

EDQ 支持对多种类型的数据库的连接。添加一个连接到 Oracle 数据库的数据存储区:

1. 展开新创建的项目, 右键单击**数据存储**, 然后选择**新建数据存储**



1. 在两个下拉框中选择Sever和Database 然后选择**Oracle**。单击**下一步 >**继续



3. 输入以下连接详细信息:

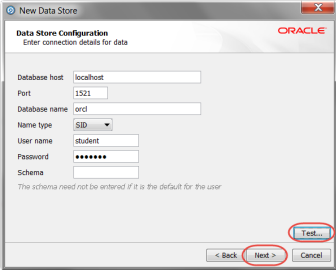
· **数据库主机**: 本地主机

· **端口**: 1521

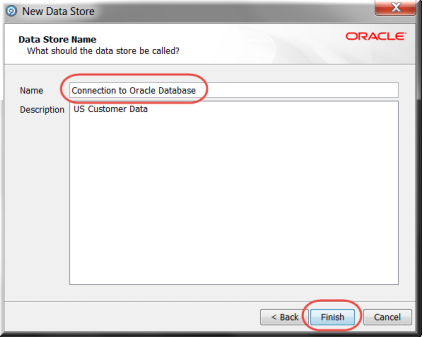
· **数据库名称**(SID):orcl

· **用户名**: student

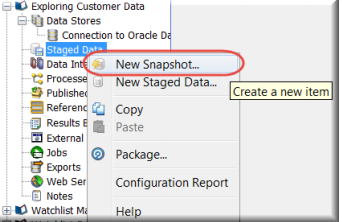
· **密码**: student

单击**测试**以检查控制器是否可以访问服务器数据库, 然后单击**下一步 >**继续

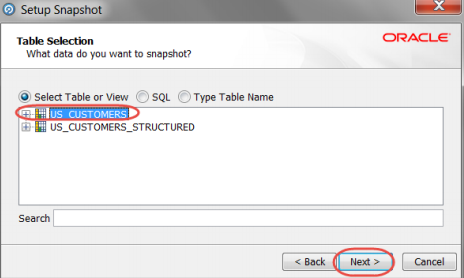
1. 给数据存储命名，单击**完成**



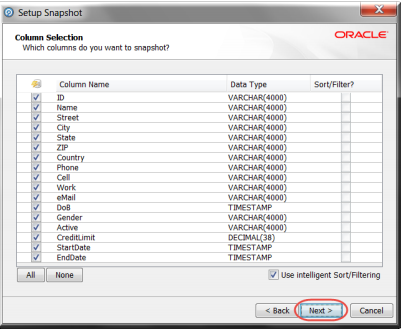
1. 到**项目浏览器，**右键单击**暂存数据**在您的项目下, 然后选择**新建快照...**



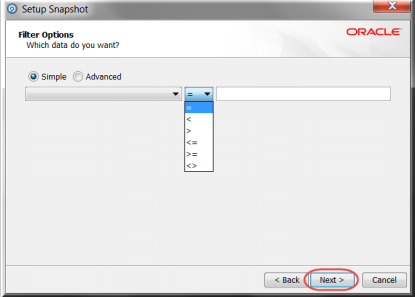
6. 选择快照的数据源，保留选中的**选择表或视图**单选按钮。单击**下一步 >**继续



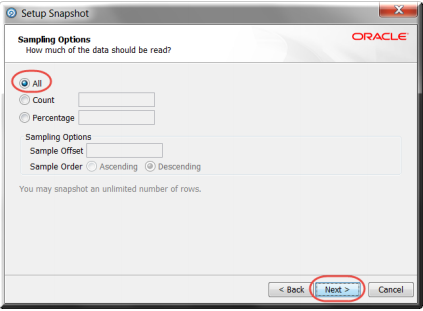
7. **列选择**, 确保选择了所有列以设置此数据快照, 然后单击**下一步 >**继续



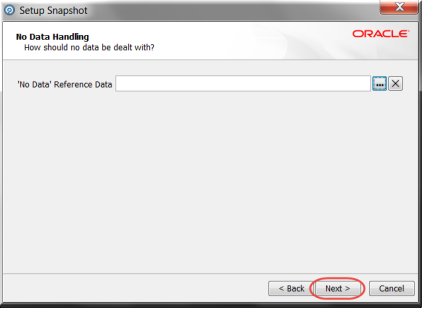
1. 保留默认值, 然后单击**下一步 >**继续



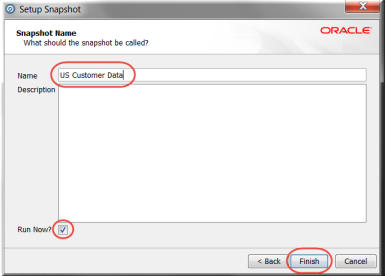
1. 为快照指定要读取的数据的特定**数目**或**百分比**。在此示例中,选择**全部**, 然后单击**下一步 >**继续



10. 保留**引用数据**的默认空值文本。



11. 确保**立即运行？**复选框被选中, 然后单击**完成**

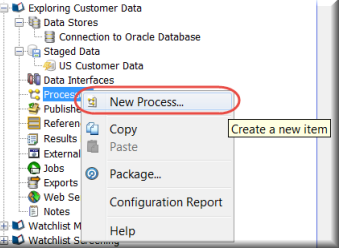


在短暂的延迟之后,**结果浏览器**将会被来自 Oracle 数据库快照中的数据填充。

**二 ：分析数据**

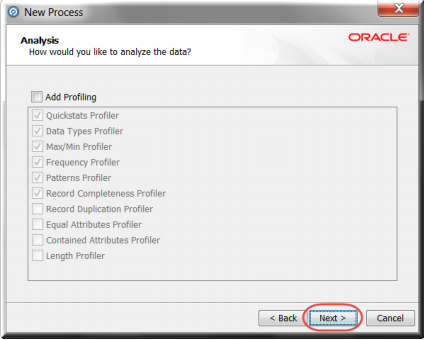
创建和运行分析进程

1. 项目浏览器并右键单击**进程**, 然后单击**新建进程**



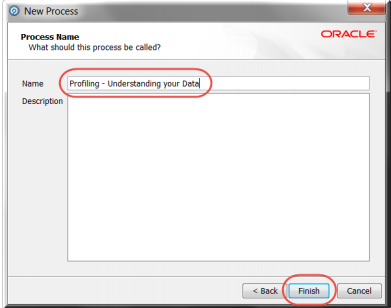
1. 选择数据快照US Customer Data, 然后单击**下一步 >**继续

3. 请注意, 在创建此新进程时, 可以选择 "**添加分析**"。保留复选框未选中, 然后单击**下一步 >**继续



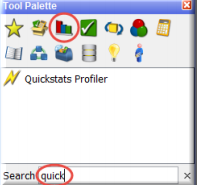
在创建新进程时选择 "**添加性能分析**" 时, 将为选定的所有分析器对数据进行描述。在创建大型或宽数据集的进程时, 建议在创建进程后单独添加分析器, 因为执行分析器可能需要很长时间。

4. 为该进程命名: **Profiling-Understanding your Data** –然后单击**完成**

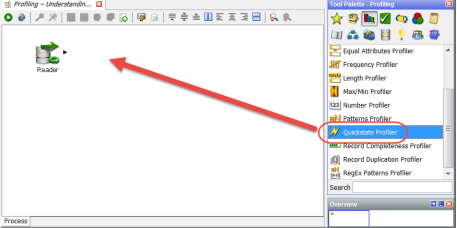


新创建的进程将显示一个项目画布。**读取器**分析器会自动添加到项目画布中。

5. 在**工具箱**中查找**分析**图标。接下来, 在分析器系列中查找并选择**快速统计分析器**

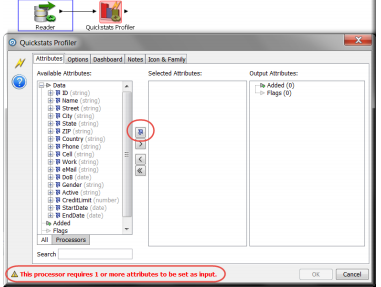


6. 拖放**快速统计分析器**到项目画布上



任何分析器上的**绿色圆形箭头**都意味着它尚未执行。

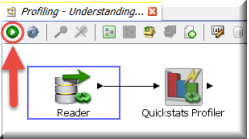
1. 悬停在**读取器**分析器的输出三角形上。将出现一个信息工具提示, 其中带有分析器的名称和简要说明。单击**读取器**分析器的输出三角形拖动到**快速统计分析器**的输入三角形。成功连接和释放鼠标后, 将出现**快速统计分析器**配置对话框:



8. 注意红色消息, 通知您**此分析器至少需要将一个属性设置为输入**。单击**全选**图标,如上面的屏幕截图所示，单击**确定**以保存

新分析器还显示 "尚未运行" 绿色图标。一旦进程运行, 绿色 "尚未运行" 图标将消失。

9. 进程现在有一个**读取**器和一个**快速统计分析器**。单击工具箱中的**运行**图标以运行进程

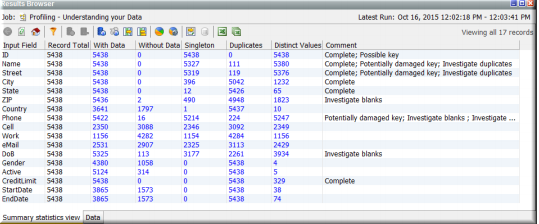


当进程运行时, 可以在控制器左下角的**任务箱**中观察进度。当进程完成后, "尚未运行" 图标将从画布中消失, 以显示分析器具有与之关联的数据。



10. 单击**读取器**分析器以查看存储在暂存数据快照中的原始输入数据。这将显示在**结果浏览器**

11. 下一步单击**快速统计分析器**查看分析器的输出



" **快速统计分析器** "为许多记录或事务提供基本质量度量, 突出显示 :

· 候选键列

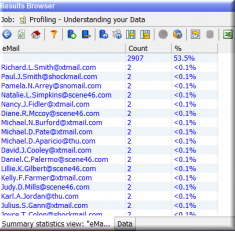
· 完整性和缺少的数据

· 重复

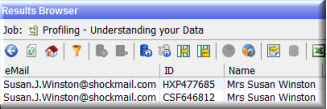
· 值的唯一性和多样性

这些结果可以观察和调查, 以快速发现数据异常。例如, 对于**性别**属性, 有**4 不同的值**, 您还可以向下点击任何蓝色文本以查看里面的数据。

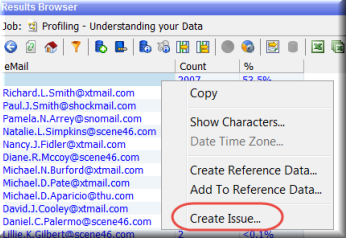
12. 单击**结果浏览器**中**重复**列下的列出的**电子邮件**的编号**3113** 。单击包含无内容/空白的**电子邮件**地址的**Count**超链接



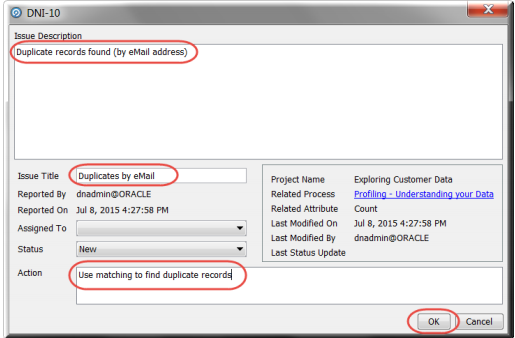
13. 返回到上一个非空值的视图. 我们观察到系统中有许多重复的**电子邮件**值 ( **Count=2** ), 可能需要从重复的记录点进行进一步的调查



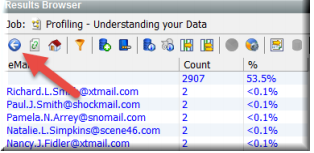
14.要为重复 ( **count** > 1) 电子邮件记录添加问题, 请右键单击**计数**值在**结果浏览器**中为**2**的超链接字段, 然后选择**创建问题...**



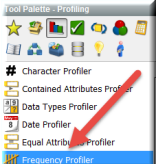
15. 可以将问题分配给自己 ( **dnadmin@ORACLE** ) 或其他用户。还可以键入后续**操作**: **使用匹配查找重复记录**。该问题还包括一个指向创建问题的进程和结果视图的链接。在**问题描述**下, 键入 **重复记录 (电子邮件地址)**. 单击确定 并创建问题



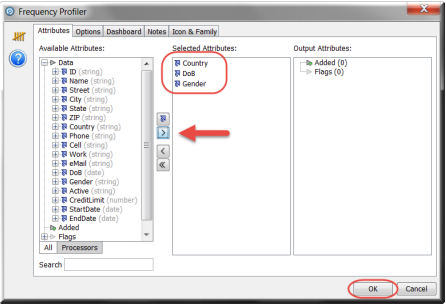
16. 单击返回图标**，**返回到**快速统计分析器**的结果



17. 返回**工具箱-分析**并查找**频率分析器**。将分析器拖放到项目画布上, 并将**快速统计分析器**的输出三角形链接到**频率分析器**的输入.



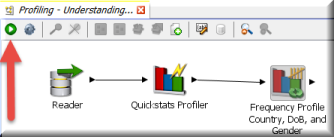
18. **频率分析器**"对话框出现。多选**国家**、**出生日期**和**性别**到**选定属性**，然后单击**确定**



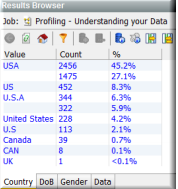
19. 可以通过双击画布中分析器的名称来重命名分析器。双击**频率分析器**的现有标签, 以重命名分析器

\*右键单击分析器图标并选择**重命名**也允许重命名分析器

20. 单击**运行**图标启动进程。等待执行完成



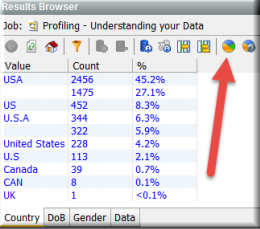
21. 单击**频率**分析器在**结果浏览器**中查看结果。请注意左下角4个不同的选项卡:**国家、出生日期、性别**和**数据**。

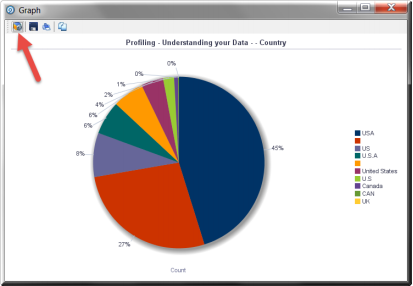


对于每个**选定属性**都有单独的选项卡

美国有多少种不同的表述形式:USA、US、U.S.A、United States和U.S。

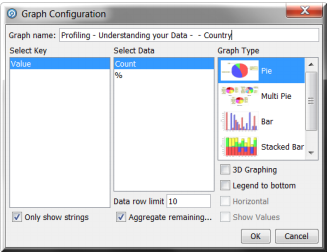
22. 单击**结果浏览器**中的**图形结果**图形按钮以查看图表

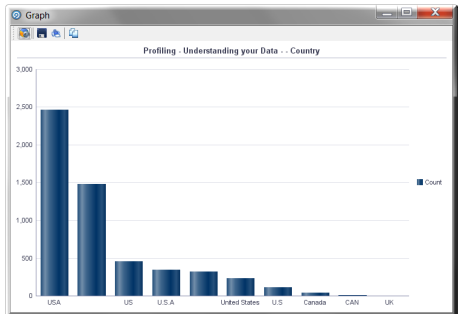




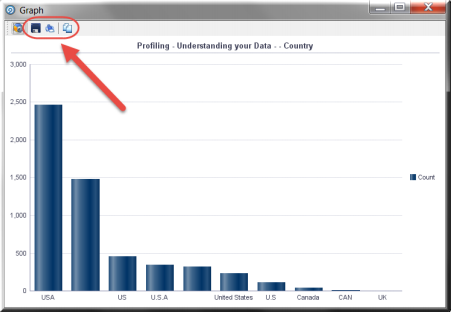
23. 工具箱中的第一个按钮如上面的屏幕截图所示, 将允许您更改图表的标题, 将图表配置为不同的可视化类型, 或修改显示的数据类型。单击此按钮以打开**图形配置**对话框。

24. 选择**条形图图形类型**, 然后单击**确定**





25. 图表中的下一个按钮允许您将其保存到图像文件、打印或复制到剪贴板, 以便与他人共享

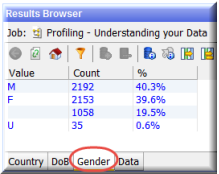


26. 关闭**图形**对话框窗口并单击**结果浏览器**底部的出生日期选项卡, 以查看**频率分析器**的结果



可以通过单击各种列标题来对列进行排序。

27. 单击**结果浏览器**底部的**性别**选项卡, 查看**频率分析器**的结果



例如, 上述19.5% 的性别的值是空白的。

28. 返回**工具箱-分析**并找到**记录完整性分析器**和**记录重复分析器**。在 **搜索** 文本中键入 **记录** 以查找分析器并将这些分析器拖放到项目画布

29. 将**快速统计分析器**的输出三角形连接到**记录完整性分析器**的输入三角形

30. **记录完整性分析器**配置对话框随即出现。单击 "全选" 图标使所有数据列都参与, 然后单击**确定**



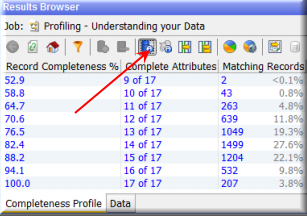
31. 单击**记录完整性分析器**的输出三角形到**记录重复分析器**的输入三角形

32. **记录重复分析器**配置对话框随即出现。将**名称**属性、**Zip**属性移动到**选定属性**。然后单击**确定**

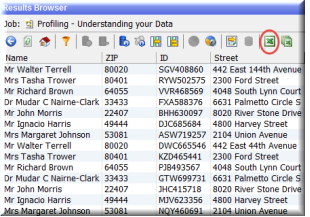
33. 单击工具箱中的**运行**图标以运行进程

**记录完整性分析器**将分析所有选定属性的记录以显示完整性。**记录重复分析器**将在选定的名称和 Zip 属性中分析重复项的记录。

34. 单击**记录完整性分析器**分析器查看**结果浏览器**中的结果。



35. 单击**记录重复分析器**查看结果。向下点击表示重复记录的**14** 。将结果导出到一个 Excel 文件, 以便将其发送到组织中的个人进行进一步调查。单击**结果浏览器**工具箱中的**导出到 Excel**图标以保存该文件。



如果结果浏览器中有多个选项卡, 则可以单击 "导出到 excel" 旁边的按钮以将**所有选项卡导出到 excel**.

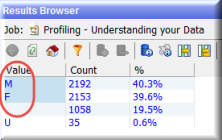
**三: 审计数据**

审计分析器根据**检查**将每个输入记录分类为**有效**或**无效**。在分析时, 审计分析器隐式使用您应用于给定数据属性的业务规则。对于您可以应用的每种类型的业务规则, 都有一个审计分析器。

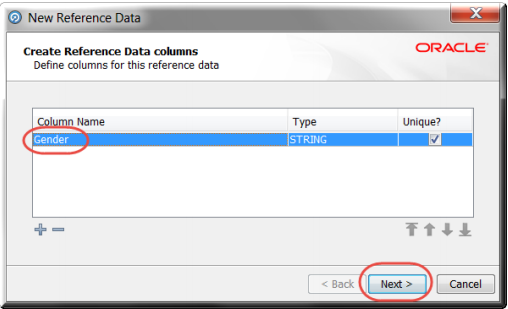
创建引用数据

1. 首先, 返回到**Profiling-Understanding your Data**进程，单击**频率分析器，**然后单击**结果浏览器**左下角的**性别**选项卡

2. 按住 CTRL 键并单击**M**和**F**值



3. 右键单击并选择**创建引用数据**。将出现**新建引用数据**对话框。将属性名称重命名为 **性别**, 单击**下一步**继续

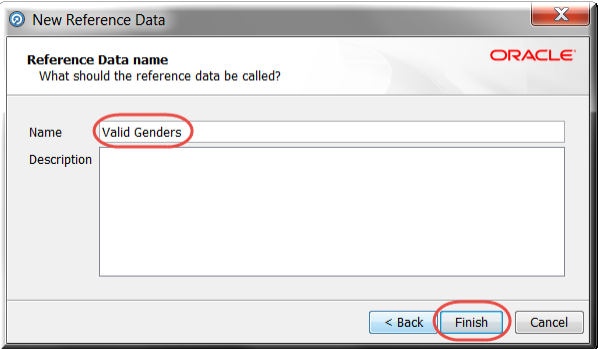


4. 添加**性别**到**查找列** 然后单击**下一步**。单击下两个屏幕上的**下一步 >**以保持默认设置继续





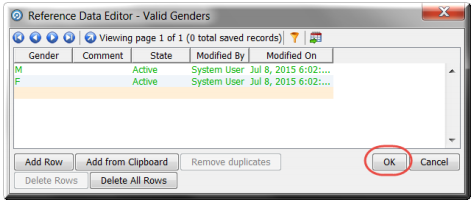
5. 最后, 输入引用数据的名称: **有效性别** , 然后单击**完成**



6. **引用数据编辑器**显示

在这里, 您可以修改引用数据以添加行或删除行。EDQ 带有许多不同类型的引用数据, 可以显著加快创建数据检查过程所需的时间。

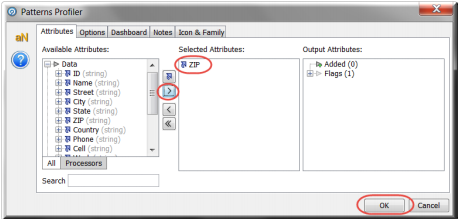
7. 在这里, 您可以看到 EDQ 从性能分析结果中获取了 M 和 F 值, 以创建自己的自定义引用数据。单击**确定**返回到项目画布



接下来, 我们将需要为有效的邮政编码类型创建引用数据 。在这种情况下, 我们需要5位数字字符串或一个5位数字, 后跟一个连字符和一个4位数字。为了方便地创建此引用数据, 我们需要使用模式分析器来分析 ZIP 的代码属性。

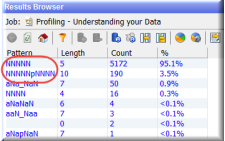
8. 在**工具箱中**查找**模式分析器**。将此拖放到**项目画布**中, 并将输出三角形从**记录重复分析器**拖放到**模式分析器**

9. **模式分析器**配置对话框随即出现。从**可用属性**中选择**ZIP** , 然后按**>**按钮将其添加到**选定属性**, 单击**确定**继续



10. 单击**进程选项卡**上方左上角的**工具箱**中的**运行**按钮.

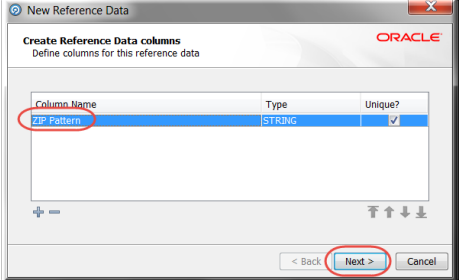
进程完成后, 单击**模式分析器**并在**结果浏览器**中查看结果



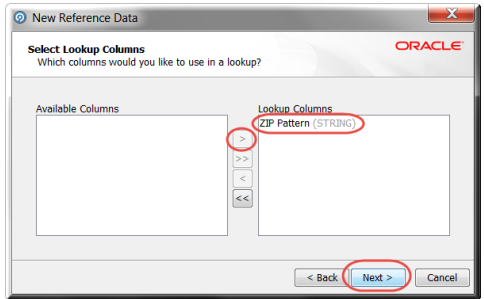
**N**表示一个数字, **p**表示标点符号, **a**表示 alpha 字符, **\_**表示空格。

11. 现在我们有了创建引用数据的模式, 请重复创建有效的性别引用数据时所采取的步骤。因为我们需要5位数字或5位数字, 后跟4位数字, 所以在**NNNNN**和**NNNNNpNNNN**上单击

12. 右键单击**NNNNN**并选择**创建引用数据**, 显示**新建引用数据**对话框。将**列名**重命名为 **Zip 模式** 通过双击**模式**下的列名, 单击**下一步 >**继续



13. 添加**ZIP 模式**到**查找列**, 然后在接下来的两个屏幕上单击**下一步 >** 以继续

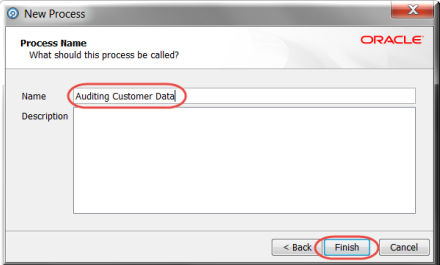


14. 最后, 为这个新的引用数据命名 **有效的 ZIP 模式** , 单击**完成**。将显示**引用数据编辑器**。如果任何值是意外选择的, 请删除这些行。否则, 请单击**确定**继续。

我们现在将开始创建一个新的流程来审计我们的US Customers Data。在我们的审计过程中, 我们在过去几步中创建的引用数据将被一些现成的审计分析器所利用。

创建审计过程

15. 单击**进程**, 然后单击**新建进程...**

16. 选择**US Customers Data** 然后单击**下一步**。单击下一个屏幕上的**下一步**。命名此新进程 **审计 客户数据**, 然后单击**完成**继续

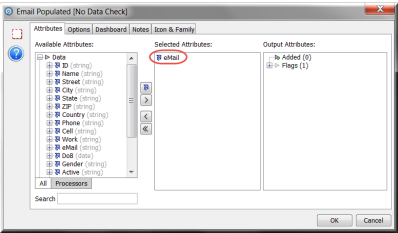
17. 与我们创建的第一个进程一样,**读取器分析器**自动添加到

项目画布。在**工具箱中**查找审计图标

18. 首先, 在进程画布上拖放**无数据检查**分析器。右键单击**无数据检查**分析器, 然后选择**重命名**以将其命名为 **电子邮件填充。**将**读取器的**输出三角形拖放到**电子邮件填充**审计分析器

19. 将出现**电子邮件填充**的配置对话框.从**可用属性**中选择**电子邮件**

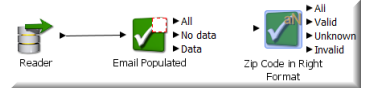
属性并单击按钮将其添加到**选定属性**, 单击**确定**继续



20. 单击**运行**, 然后选择**电子邮件填充**的审计分析器以查看结果

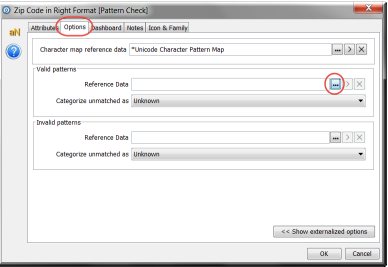
请注意**结果浏览器**中的**无数据**和**有数据**值。如果需要, 我们可以通过选择**数据**、**无数据**或**全部**, 继续使用来自分析器的一个或多个端点输出数据流三角形来开发此过程。

21. 下一步, 查找**模式检查分析器**。将其拖放到画布中, 将其重命名为 **Zip 编码的正确格式**

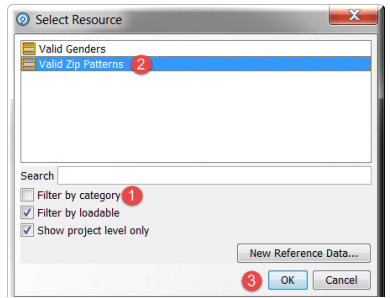


22. 连接**电子邮件填充**的All输出三角形, 到 **Zip 代码以正确格式**分析器。模式检查的配置对话框分析器出现。从**可用属性**中选择**ZIP**作为待验证的字段。

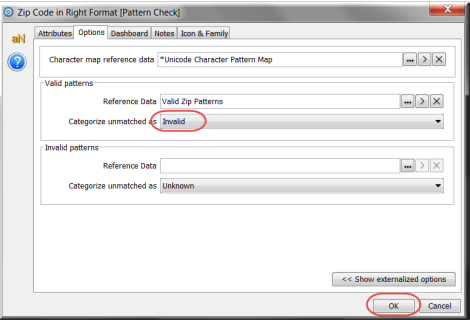
23. 单击对话框顶部的**选项**选项卡, 然后单击窗口中间**有效模式**部分中的按钮



24. 取消选中**按类别筛选**。选择我们为不同类型的有效邮政编码创建的引用数据。单击**有效的 Zip 模式**, 然后单击**确定**



25. 在 "**有效模式**" 下的部分中, 单击要更改的下拉框**分类为**到**无效**, 然后单击**确定**继续

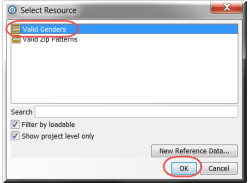


26. **运行**。单击以**正确格式显示邮政编码**以查看结果

请注意, 有**5362 有效记录**和**76 无效记录**。即, 有76条记录不符合 NNNNN 或 NNNNNpNNNN 的 "适合使用" 规则。

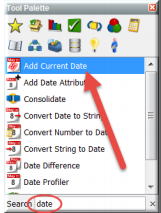
27. 返回**工具箱**并查找**列表检查**分析器。将其拖放到项目画布上, 并将**Zip 代码以正确格式**的**所有**三角形链接到**列表检查**分析器

28. 选择 "**性别**" 属性并添加到**字段以进行验证**。然后单击对话框顶部的**选项**选项卡, 为性别添加引用数据.单击**有效值**部分中**引用数据**文本字段右侧的图标, 以浏览引用数据并选择**有效性别**引用数据, 然后单击**确定**继续



29. 单击**确定**关闭**列表复选**对话框。双击**列表检查**分析器将其重命名为 **检查有效的性别** 最后 , 单击**运行**按钮以启动进程

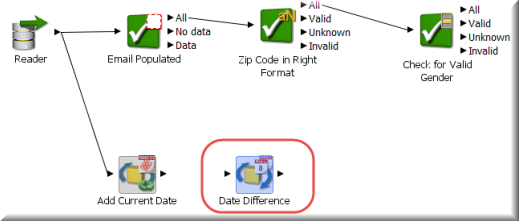
30. 返回**工具箱**, 并搜索**添加当前日期**分析器。将其拖放到项目画布上 (您可以通过在工具箱搜索文本字段中输入 **日期** 来找到分析器)



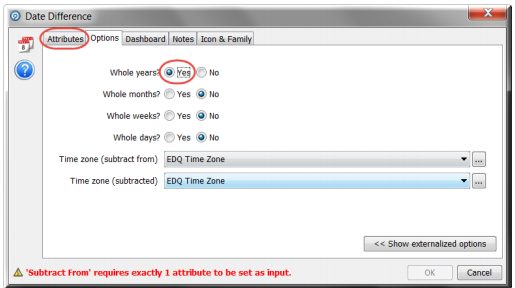
31. 将输出三角形从**读取器**连接到**添加当前日期**分析器的输入三角形。此分析器不需要其他配置, 请单击**确定**

32. 返回**工具箱**查找**日期差**分析器, 并将其拖放到项目画布上。将**添加当前日期**的输出三角形连接到**日期差**分析器的输入三角形

**日期差**和**添加当前日期**分析器将用于计算每个客户的年龄。



33. 单击日期差对话框顶部的**选项**选项卡。选择**整个年份**的**是**



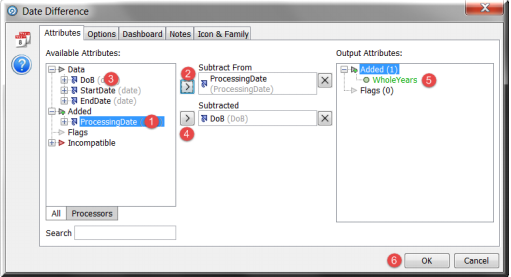
如果要计算月、周或日的年龄 (或日期) 差异, 可以通过选择这些选项来完成。此外, 由于出生日期和使用 **添加当前**时间分析器添加的日期恰好位于同一时区, 因此我们将留下最后两个下拉列表。

34. 返回**属性**选项卡。使用按钮将**ProcessingDate**添加到**被减数**选项

35. 使用按钮将出生日期添加到**减数**选项中, 然后单击 "时差" 对话框顶部的**选项**选项卡

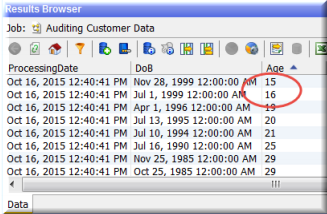
现在已将输出属性添加到对话框的右侧。

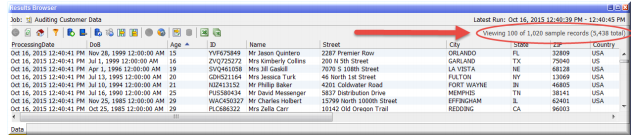
36. 在**日期差**对话框右侧的**输出属性**中,双击**WholeYears** 并将其重命名为 **年龄** , 单击**确定**继续



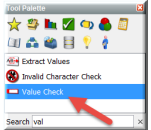
37. 运行该进程之后, 通过单击**年龄**列, 对结果浏览器中的记录进行排序

18岁以下的顾客。此时只有**2**个。

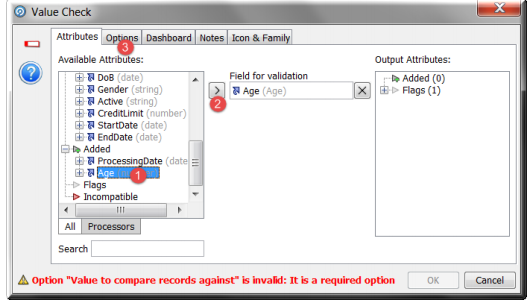




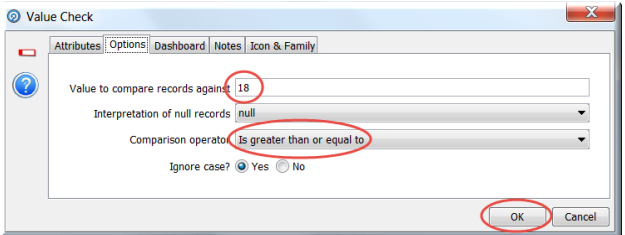
38. 在**工具箱**中查找**值检查**分析器。将**值检查**分析器拖放到项目画布上，将用于查找大于或等于18的值。



39. 将**日期差**分析器的输出三角形拖动到**值检查**分析器的输入三角形。选择**年龄**属性作为**验证字段**, 然后单击顶部的**选项**选项卡继续

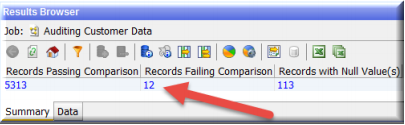


40. 输入 **18 作为与记录进行比较的值**。将比较运算符更改为**大于或等于**, 然后单击**确定**继续



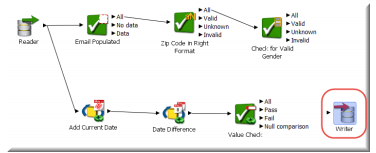
41. 运行进程

请注意, **12**个客户的年龄低于 18, 而**5313**符合标准。

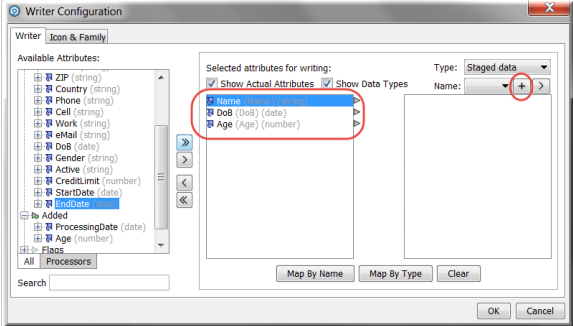


我们还可以将一个或多个**编写**器分析器拖到项目画布上, 以便将结果写回数据存储区。

42. 返回**工具箱**以搜索**编写**器, 并将其拖放到项目画布



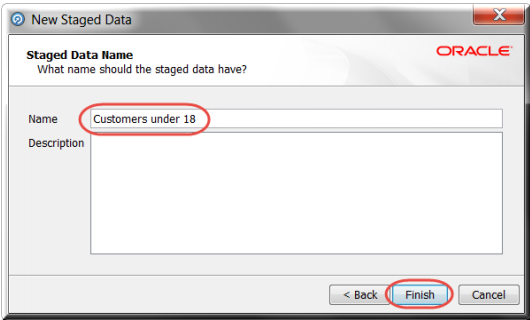
43. 将**值检查**分析器的**失败**输出三角形连接到**编写**器的输入三角形

44. 从**可用属性**中选择**名称**、**出生日期、**和**年龄**属性, 并将其移动到**选定属性**。然后单击**类型: 暂存数据**下拉列表下面的**+** , 创建一个新的**暂存数据**对象 

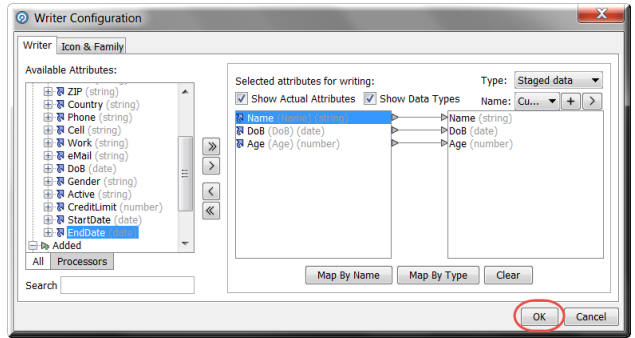
45. **新的暂存数据**第一个对话框将提示您选择要包括的列。在这种情况下, 我们将保留默认值, 单击**下一步 >**继续



46. 为此阶段的数据集命名: **18 以下的客户** , 然后单击**完成**

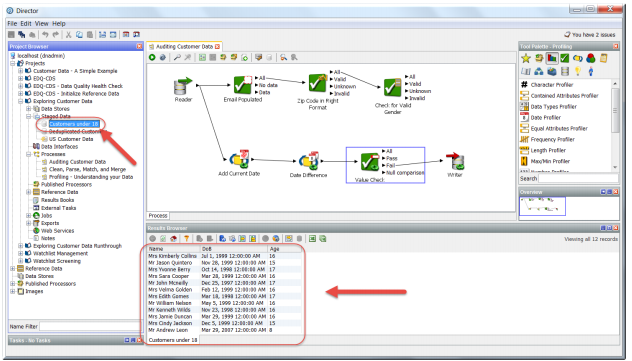


47. **名称** 、**出生日期**和**年龄**是映射到刚刚创建的**暂存数据**集的属性。单击**确定**完成配置编写器

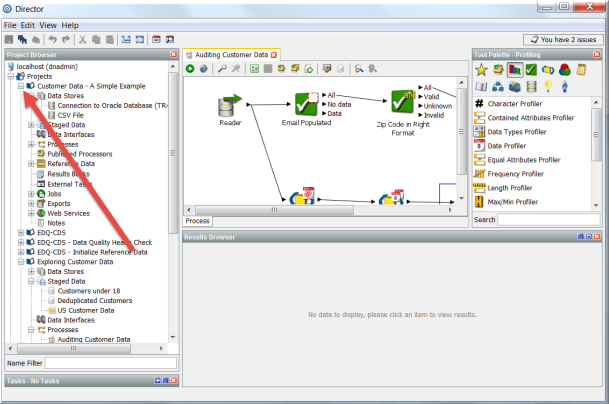


48. 运行进程

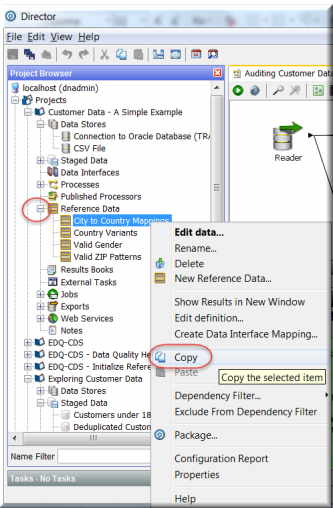
您将注意的焦点移回到**项目浏览器**和新的分暂存数据, **18 以下的客户**被选中。**结果浏览器**也已更改, 显示**6 客户**, 该用户的值对于大于或等于18的检查失败。



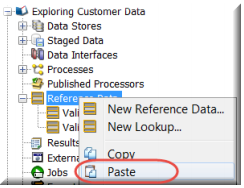
49. 到项目浏览器中查找标题为Customer Data -A Simple Example



50. 展开**引用数据**项并右键单击以复制**city to Country from City Mappings**



51. 返回**项目-浏览客户数据**并查找**引用数据**项目, 右键单击并**粘贴**剪贴板中的内容



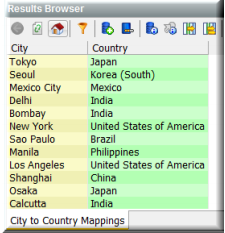
52. 在**项目浏览器中**查找标题为**Customer Data -A Simple Example**

53. 展开**引用数据**项 , 请右键单击并**复制Country Variants**引用数据

54. 返回**项目-浏览客户数据**并右键单击**引用数据**和**粘贴**剪贴板的内容

55. 在**引用数据**的项目中,**浏览客户数据**中, 单击**City to Country Mappings**

请注意, 在**结果浏览器**中, 这是两列引用数据集。**城市**(黄色) 列表示查找列。**Country**(绿色) 列表示返回列,在这种情况下, 只要查找列值包含数据, 就可以使用**City to Country Mappings**引用数据集返回该城市所在的国家。



56. 单击引用数据中的**Country Variants**

请注意, 这也是一个两个列引用数据集, 其中国家变体 (US、USA、U.S**等)**作为查找列和**标准化值 (United States)**作为返回列。这一引用数据将使我们能够将单个国家的多个表示标准化为一个国家的单一标准化代表.

**四 创建和自动化数据标准化**

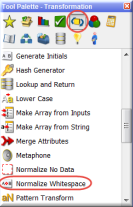
创建标准化流程

1. 在您的项目下创建新进程, 右键单击**进程**并单击**新建进程...**

2. 选择**US Customers Data** 然后单击**下一步** 。单击**下一步**而不添加任何分析

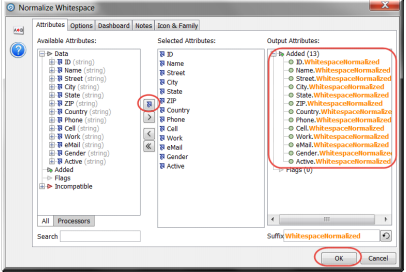
3. 命名进程 **清除、分析、匹配和合并** 然后单击**完成**

4. 在**工具箱中查找** **规范化空白**分析器

·

5. 将**读取器**连接到**规范化空白**分析器

6. **规范化空白 "对话框**出现。单击**全选**图标, 然后单击**确定**。运行以启动进程。

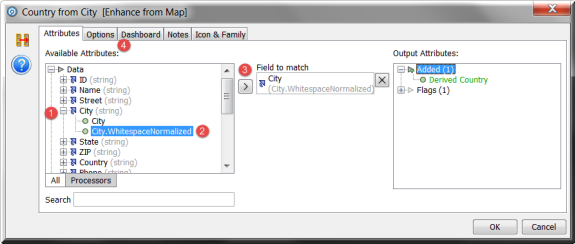


接下来, 我们将开始对国家属性进行标准化。

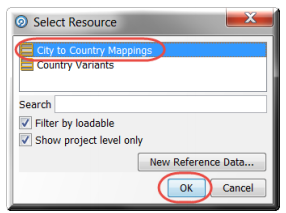
7. 查找**从映射增强分析器** 。将分析器拖放到进程画布上。右键单击**从映射增强分析器**, 然后选择 "重命名" 以将其重命名为 **Country from City**

8. 将**规范化空白**分析器的输出三角形连接到**Country from City**的输入三角形

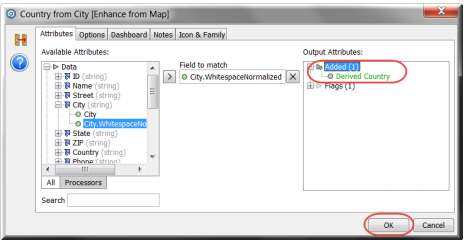
9. 在**Country from City "对话框**中, 从**可用属性**中展开**城市**字段，双击**城市**将**城市**添加到**字段以匹配**。接下来, 单击对话框顶部的**选项**选项卡



10. 单击**值映射**区域内的浏览按钮, 以连接先前复制的引用数据。选择**City to Country Mappings**, 单击**确定**继续



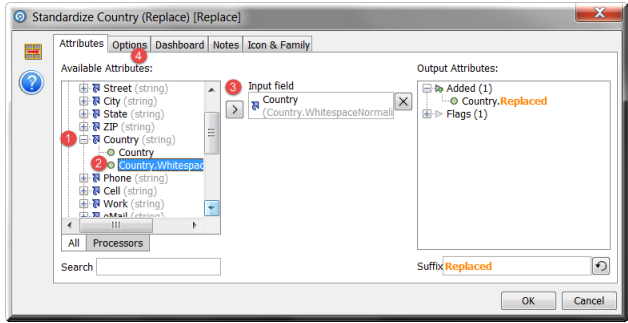
11. 单击**属性**在对话框的左上角, 重命名**输出属性: EnhancedResult**



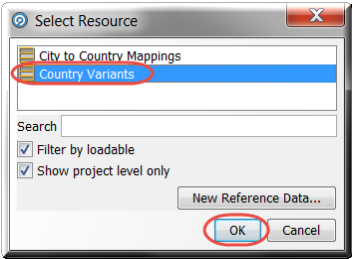
12. **运行，启动Country from City**进程

13. 返回**工具箱**搜索**替换**分析器。将**替换**分析器拖放到项目画布上。右键单击**替换**分析器并选择**重命名**以重命名它 **Standardize Country**

14. **Standardize Country**配置对话框出现. **将Country from City**中的All输出三角形与**Standardize Country的输入三角形连接**。双击**Country**字段, 并将**country.WhitespaceNormalized**添加到输入字段。然后单击对话框顶部的**选项**选项卡以设置替换



15. 单击**替换**项旁边的浏览按钮添加**Country Variants**引用数据, 单击**确定**



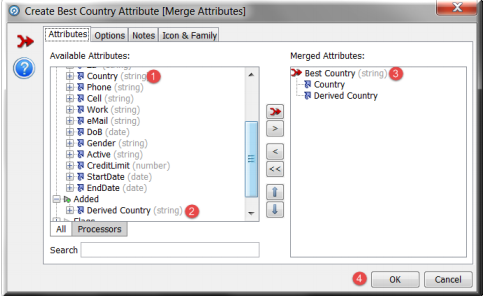
16. **运行**以启动进程

17. 返回**工具箱**并搜索**合并**。将**合并属性**分析器拖放到项目画布上并将其重命名为 **Creat** **Best Country Attribute**

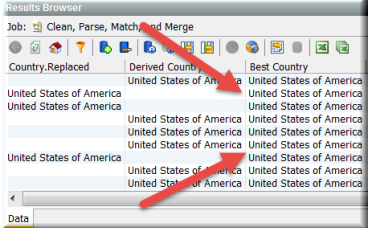
我们将使用此方法将国家和派生的国家列合并到一个可以在过程中进一步使用的单一属性

18. 将**Standardize Country**的**All**输出三角形连接到**Creat** **Best Country Attribute**分析器。在**可用属性**中双击**Country** 和**Derived Country**

添加到**合并属性**。最后, 将合并的属性重命名为 **Best Country** 然后单击**确定**继续

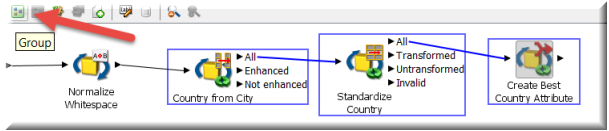


19. 单击项目画布左上角的**运行**按钮, 然后单击**Creat** **Best Country Attribute**分析器, 在**结果浏览器**中查看结果



请注意, **Country.Replaced**和**Derived Country**合并为单个列**Best Country**

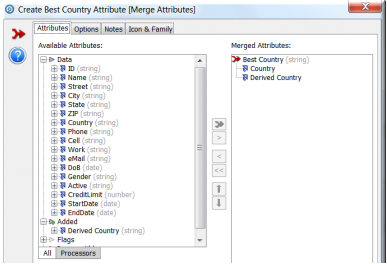
20. 将**Country from City**,**Standardize Country**和**Create Best Country**括起来，然后单击工具箱中的**组按钮**



21. 之后, 可以在 "绿色" 框中单击，重命名分组。在这种情况下, 请将其重命名为 **Standardize/Add Country**

22. 返回**工具箱**并输入 **频率分析器** 在**搜索**文本中。将**频率分析器**拖放到项目画布上, 并将输出三角形从**Create Best Country**连接到**频率分析器**, 并显示 "配置" 对话框

23. 双击**Create Best Country**将其添加到**选定属性**列中, 然后单击**确定**.



24. 通过右键单击鼠标重新命名**频率分析器**, 为**Frequency Profile Countries** 单击**运行**按钮以启动进程

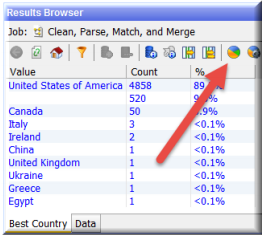


我们已经标准化了五个不同的代表美国的字段 (USA, US, U.S.A, United States和U.S) 到一个规范化的表示法:United States。

25. 右键单击**520**和**创建问题...**

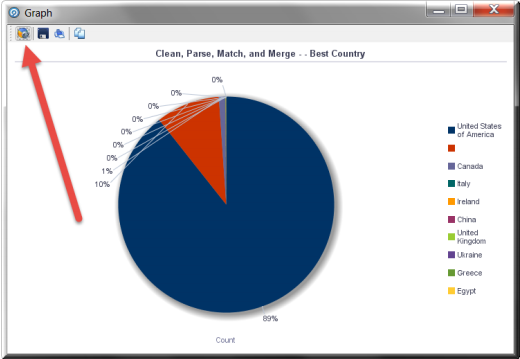
26. 在**问题说明**的文本区域中键入问题的说明, 以便在下一次实验练习期间返回到该描述.

27. 单击**图形结果结果浏览器**的工具箱中的图标



请注意, 此图表与我们在实验1b 中创建的表单的外观截然不同: 将美国的各种各样的表示法现在全部被规范化了到United States

28. 单击图表工具箱中的**图形配置按钮**

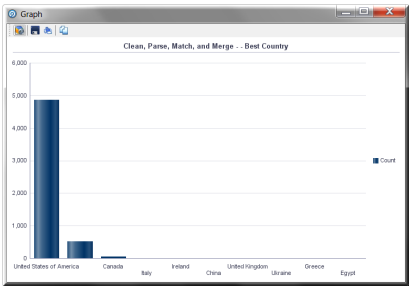


29. 更改**图形类型**到条形, 单击**确定**继续

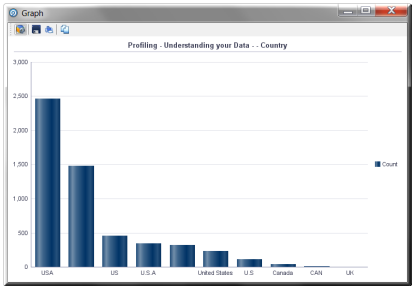


与以前创建的图表相比,可以说国家属性有更高的适用性。

**标准化后**



**标准化前**

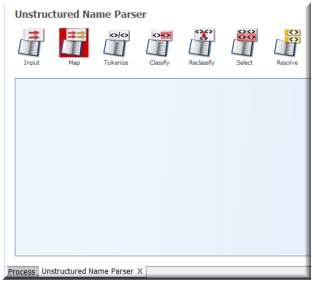


下一步, 我们将开始检查名称字段。

30. 在**工具箱**中搜索 **解析**。将**解析**拖放到项目画布上, 并将**规范化空白**端点连接到**解析**

31. **解析 "对话框**出现, 从**可用属性**中选择**名称**, 并将其添加到**选定属性**中, 单击**确定**继续

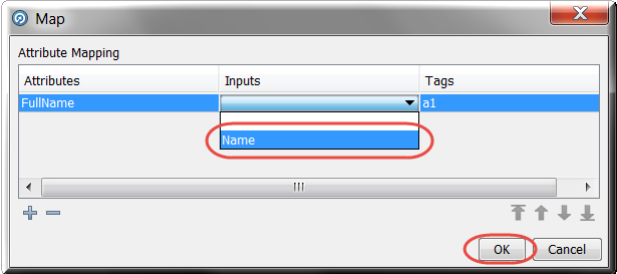
解析分析器由几个配置屏幕组成。此外, 您会注意到项目画布的左下角有多个选项卡 (在**结果浏览器**上方)。您可以在项目画布和**非结构化名称分析器**配置之间根据需要来回切换。



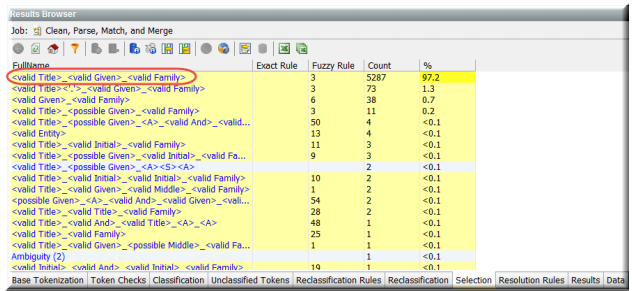
在项目画布顶部, 您将注意到几个 sub-processors**: 输入、映射、**

**标记、分类、重新归类、 选择和解析**。需要配置任何**以红色为底纹**的 sub-processor。

32. 双击**映射**将其配置。因为我们为这个分析器选择了一个属性, 所以它只有一个输入– **FullName**。在**输入**下拉列表中选择**Name**属性, 然后单击**确定**继续

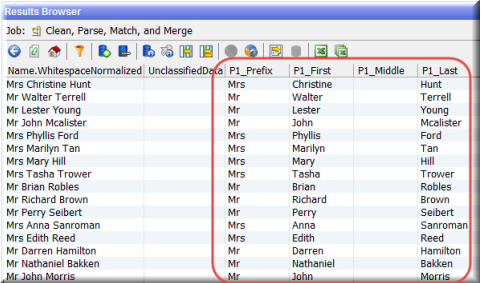


33. 运行该进程。几分钟后, 分析器结果将显示在结果浏览器的**选择**选项卡内



请注意, 超过97% 的记录有以下模式 < 有效的标题 > \_ < 有效给定 > \_ < 有效的家庭 >。

34. 点击模式**< 有效标题 > \_ < 有效给定 > \_ < 有效家族 >**查看此分析器如何拆分**名称**属性



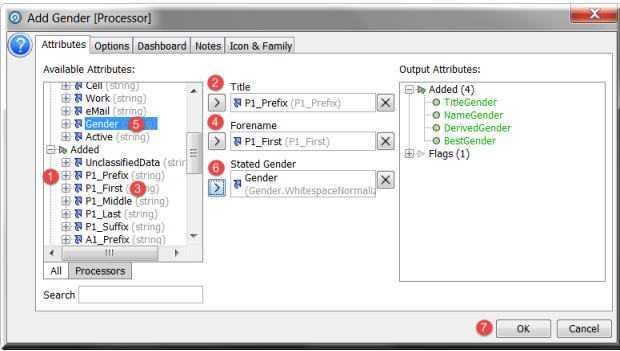
现在, 我们可以在这个过程中进一步使用这些附加属性 (P1\_Prefix、P1\_First 等)。例如, 我们可以根据 P1\_Prefix 属性得出性别, 以进一步提高我们在实验1中注意到的性别数据的质量。

35. 单击项目画布左下角的**非结构化名称分析器**上的**X** , 以返回到**进程画布**

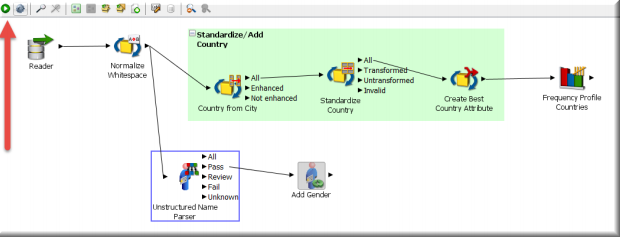


36. 返回**工具箱**搜索 **添加性别** 。将**添加性别**分析器拖放到项目画布上, 并将**解析分析器**的通过输出三角形连接到**添加性别**分析器的输入三角形

37. 选择**P1\_Prefix**并按**>**图标将其移动到**标题.**选择**P1\_First**并移动到**名**。选择**性别**并将其移至**Stated Gender**, 然后单击**确定**继续

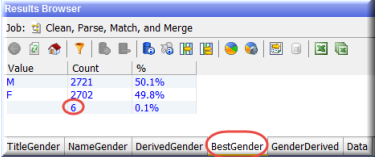


38. 单击工具箱中的**运行**图标以启动进程

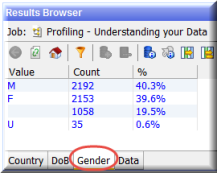


39. 在**结果浏览器**中, 单击**BestGender**选项卡。这将显示从标题和名称派生的记录的成功和不成功的计数和百分比。单击**计数**列下面的**6**

**在增加性别 (丰富/标准化性别) 后:**

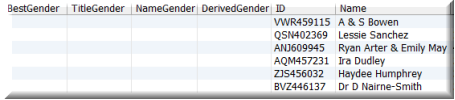


**在添加性别之前 (从实验1b 的截图)..。使用的数据 un-fit**



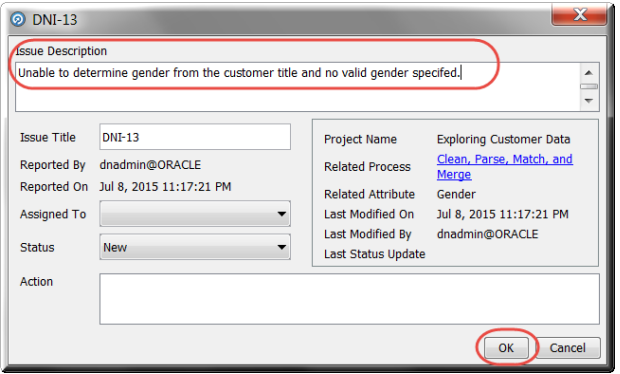
我们将缺少的**性别**的百分比从**19.5%**减少到小于**0.1%.**

40. 通过观察下面的**Name**列, 可以看到为什么未导出**性别**



41. 向右滚动直到看到**性别**列

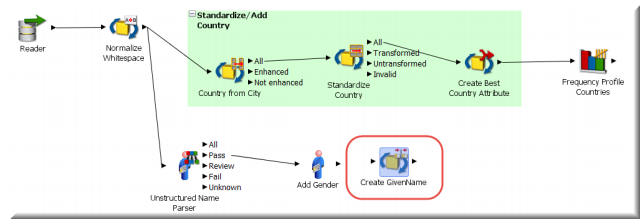
42. CTRL 键单击**性别**列下面的 4 **U**值。右键单击并单击**创建问题...**键入 **无法从客户名称中确定性别, 并且没有**供有效的性别指定



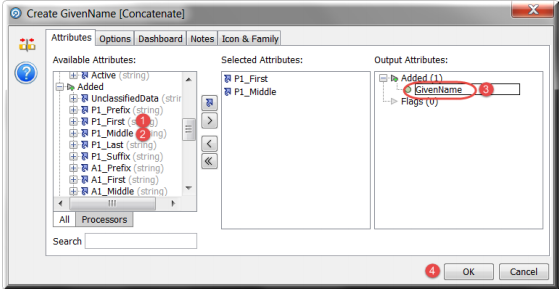
**四：匹配、合并和重复数据消除, 为关键数据实体创建单一的真实来源**

1. 返回**工具箱**并通过键入**连接**来搜索连接分析器

2. 将**连接**分析器拖到项目画布中。双击**连接**将分析器重命名为 **Create GivenName**



3. 将**Add Gender的输出三角形连接**到**Create GivenName**。选择属性**P1\_First**和**P1\_Middle**。请确保按此顺序选择这些属性。在**输出属性**中双击**Concat**并将其重命名为 **GivenName**, 单击**确定**继续



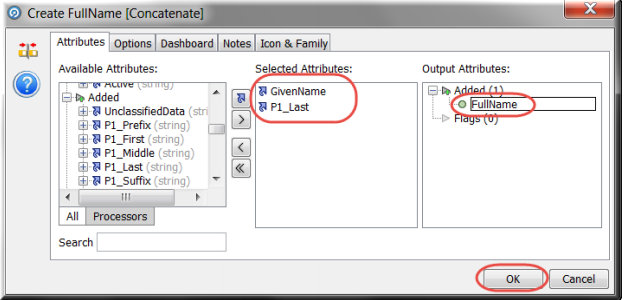
4. 右键单击**Create GivenName**分析器, 然后单击**复制**

5. 右键单击项目画布, 然后单击**粘贴**添加**Create GivenName**分析器的另一个副本。双击复制的分析器的名称, 将其重命名为 **Create FullName**。接下来, 将分析器**Create GivenName**连接到**Create FullName**

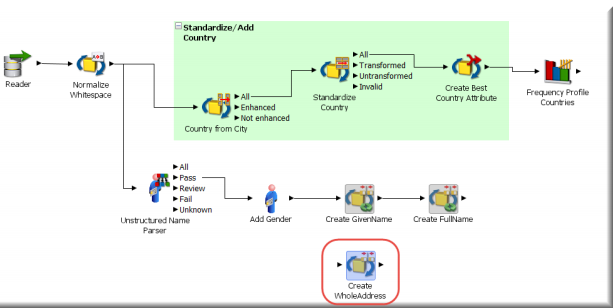


6. 将"**Create GivenName** " 的输出三角形连接到**Create FullName**的输入三角形。首先, 通过双击, 删除现有的**选定属性GivenName** 和**P1\_Last**

7. 按此顺序选择**GivenName**和**P1\_Last** 。 通过双击**现有输出属性** , 将对话框右侧的输出属性重命名为 **FullName**



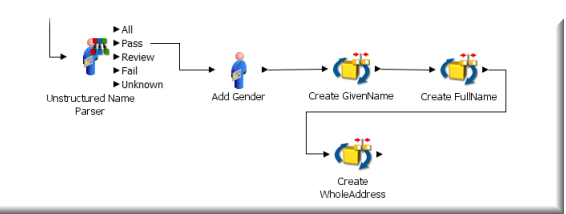
8. 返回**工具箱**并将另一个**连接**分析器添加到项目画布，双击分析器名称, 将其重命名为**Create WholeAddress**



您可以双击连接不同分析器的管线 (管道) 将其更改为弯头连接。

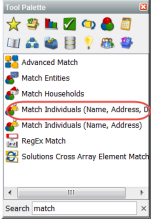
9. 将**Create FullName**的端点连接到**Create WholeAddress**分析器的输入。选择**Stree**t**、City、State**和**ZIP**属性作为**选定属性**。通过双击现有值来重命名**输出属性** 为**WholeAddress**, 单击**确定**继续

10. 双击连接**创建 FullName**和**创建 WholeAddress**分析器的管道将其更改为弯头连接

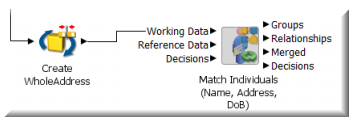


接下来, 我们将查找并添加匹配分析器以匹配个人。

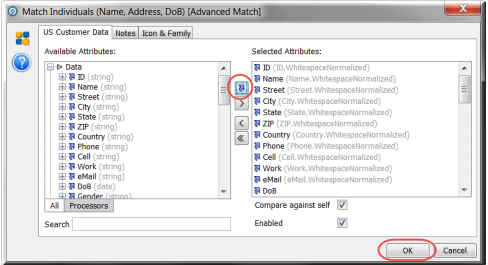
11. 返回**工具箱**并按键入方式搜索 **匹配** 。您可能需要展开**工具箱**以查找**匹配个人 (姓名、地址、**出生日期)。



12. 将**匹配个人 (姓名、地址、出生日期**) 拖到项目画布上, 并将**Create WholeAddress**的输出三角形连接到**匹配个人 (姓名、地址、出生日期**)分析器的输入端口



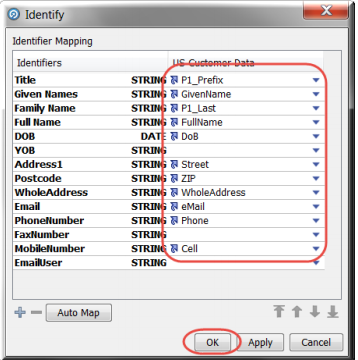
13. 使用**选择所有**按钮选择所有属性, 单击**确定**继续



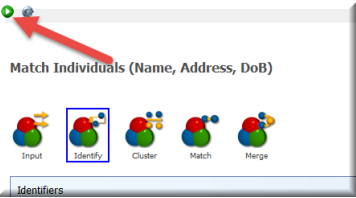
14. 双击Identify



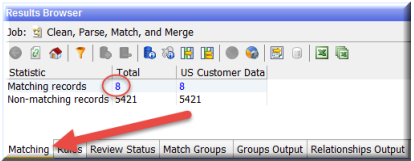
15. 将输入属性与下一页上的图形中描述的匹配标识符连接起来



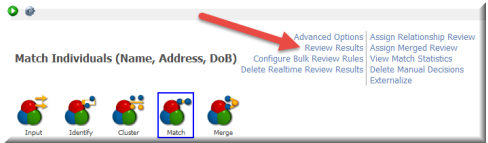
16. 单击工具箱中的**运行**图标以运行进程



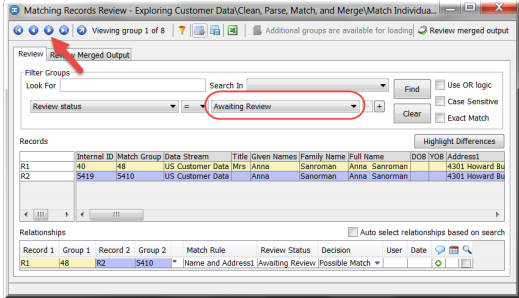
17. 当进程完成时, 单击**匹配**sub-processor 以在**结果浏览器**中查看结果。选择结果浏览器底部边缘的**匹配**选项卡, 您将看到匹配分析器已找到自动匹配的**8条记录**



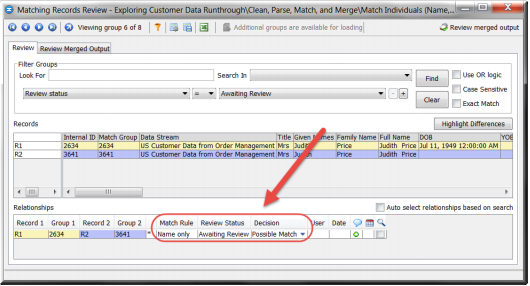
18. 单击**审阅结果**链接**项目画布**的右上角的以查看更详细的匹配结果



19. 使用左上的按钮在分组中切换

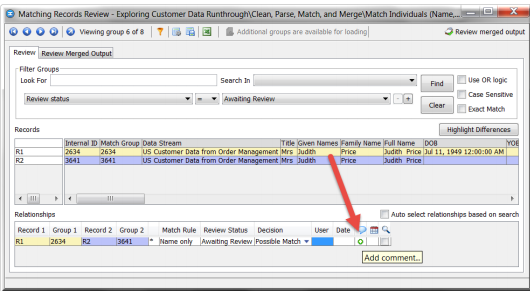


注意在**匹配记录审阅**窗口的底部, 有其他列，显示**匹配规则**、**审阅状态**和**决定**.

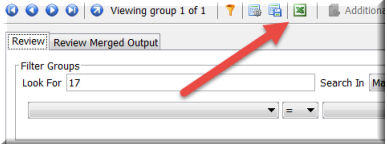


20. 单击**决定**下拉按可钮查看可用选项

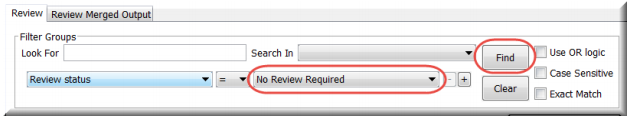
21. 此外, 如果需要进一步的详细信息，或者用户希望添加注释, 则可以单击如图的绿色图标



22. 与以前的结果浏览器屏幕一样, "**导出到 Excel** " 按钮可用于将匹配的记录审阅结果发送到另一个用户进行验证



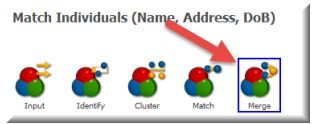
23. 若要查看自动匹配的内容, 请更改筛选器以显示具有**审阅状态的组 = 不需要审阅**的组



24. 在**匹配记录审阅**对话框左上角使用**左**和**右**箭头浏览组以查看匹配结果。完成后, 关闭**匹配记录审阅**窗口

到目前为止, 我们已经从这个客户数据集中清理、分析和匹配了一些信息。此时, 我们将基于默认合并规则使用合并 sub-processor 从匹配进程中取出重复数据消除的记录

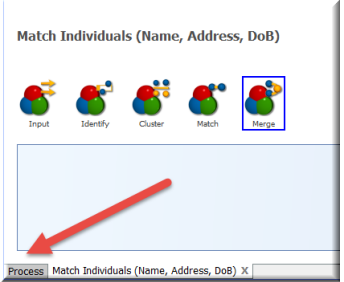
25. 双击**合并**sub-processor 开始并打开**合并**对话框



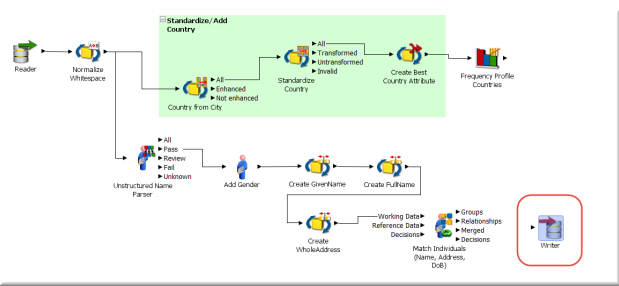
26. 在**合并**对话框的 "**输出不相关的记录**" 的旁边勾选方框, 单击**确定**继续



27. 单击画布中间左侧的**进程选项卡**以返回到项目画布

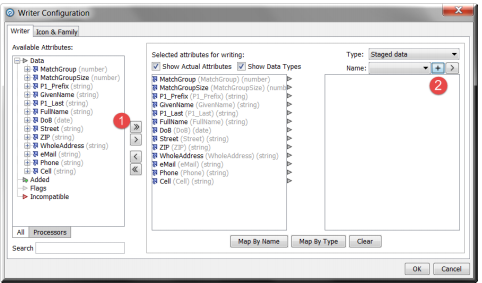


28. 在**工具箱**中, 搜索 **编写** 器。此分析器使 EDQ 进程能够将数据写入不同类型的数据存储区, 例如, 暂存数据。将**编写**器拖放到项目画布上

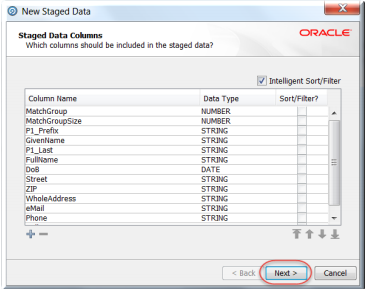


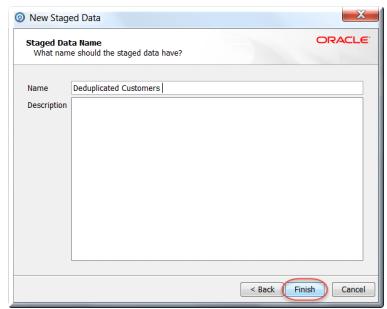
1. 将**匹配个人分析器**的**合并**输出三角形连接到**编写**器

30. 单击按钮以选择所有**可用属性**。然后按**编写器配置**对话框右侧的按钮添加新的 **暂存数据**

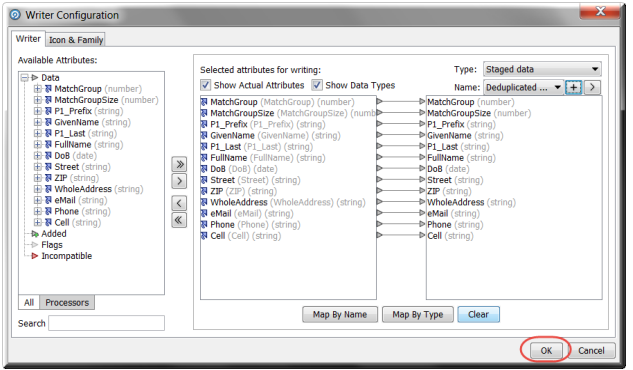


31. 单击**下一步**接受创建暂存数据集的默认配置。命名为Deduplicated Custimers, 单击**完成**





32. 单击**确定**完成设置**编写器配置**



1. 单击工具箱中的**运行**图标以运行该进程。完成此过程后, 右键单击**编写**器并单击**在新窗口中显示结果**

