# RapidMiner (https://rapidminer.com/)

# 

# หน้าจอ Design ประกอบด้วยส่วนหลัก ๆ คือ

# Repositories – เป็นที่เก็บข้อมูล

# Operator – เป็นที่รวมฟังก์ชันต่าง ๆ เมื่อเรียกใช้งานแล้ว อาจต้องมีการกำหนดค่าพารามิเตอร์ที่ส่วนของ Parameters ด้วย

# Process – เป็นส่วนที่บอกถึงกระบวนการในการทำงานเป็นขั้นตอน โดยหลักของการสร้าง process จะใช้วิธี drag and drop คือ ลาก dataset มาวาง ลาก operator ที่ต้องการใช้งานมาวาง แล้วเชื่อมต่อแต่ละกล่องเข้าด้วยกัน และสุดท้ายเชื่อมต่อไปยังจุดทางออก res เมื่อสั่งรัน จะเห็นผลลัพธ์ในหน้า Results

# Parameters – สำหรับกำหนดพารามิเตอร์ของแต่ละ operation

# Help – คำอธิบาย

# ถ้าลองเปิดไฟล์ iris จาก Repository จะปรากฏข้อมูลใน tab Results ดังนี้

# 

# ในส่วนของ Results คลิกเลือกเพื่อดูส่วนต่าง ๆ ได้ดังนี้

# Data – ข้อมูลใน dataset

# Statistics – metadata และบอกค่าสถิติของข้อมูล

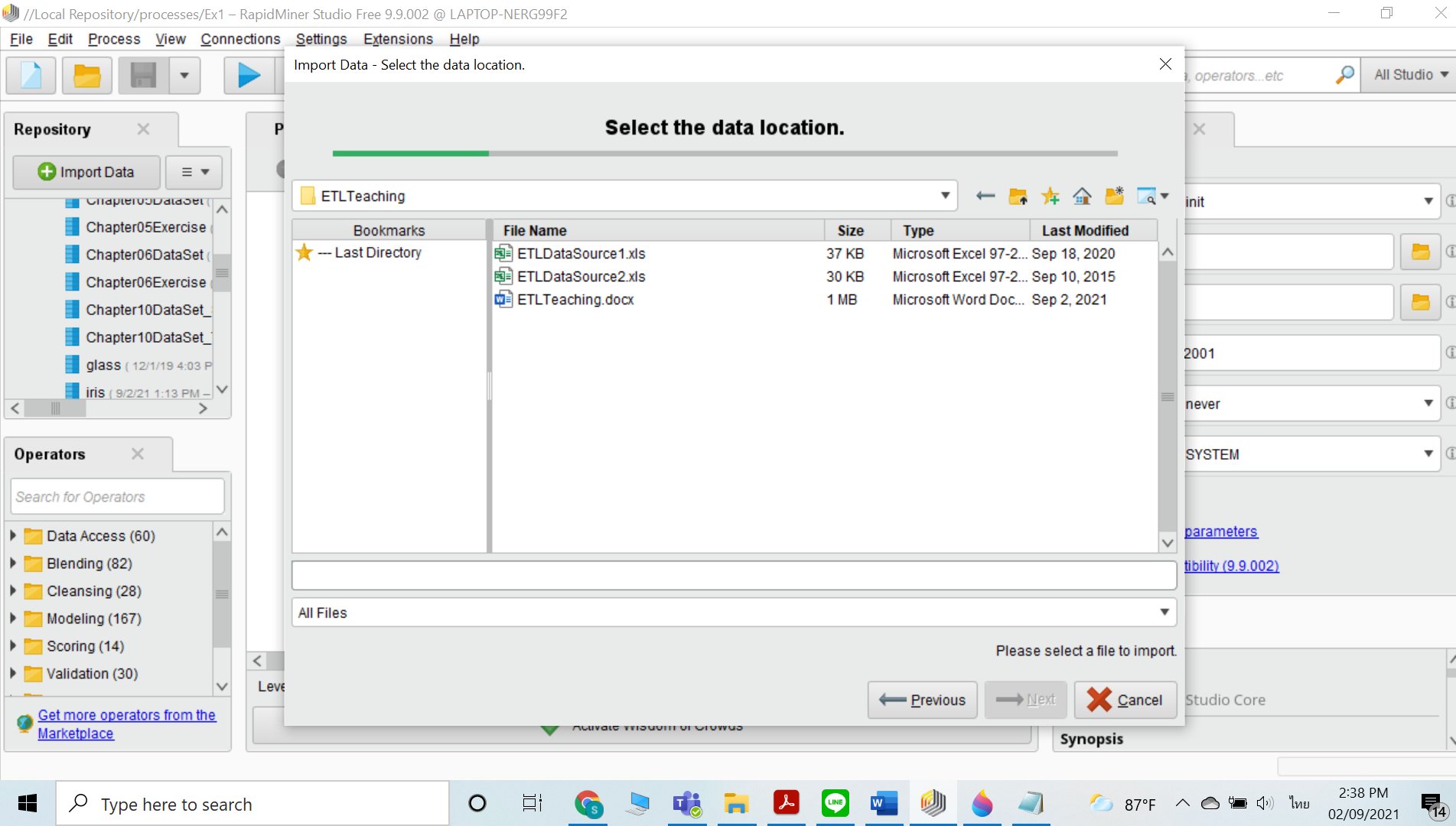
# Visualizations – สร้าง chart เพื่ออธิบายข้อมูล

# Annotation – การสร้างหมายเหตุประกอบหรือคำอธิบายเพิ่มเติม

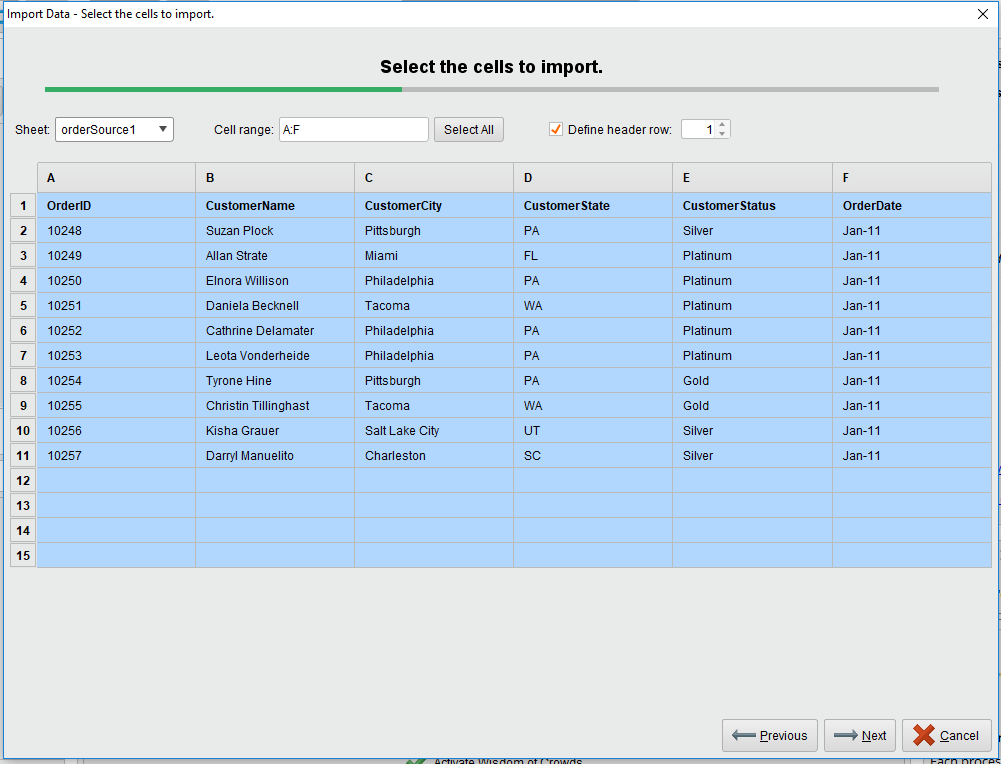
**ETL**

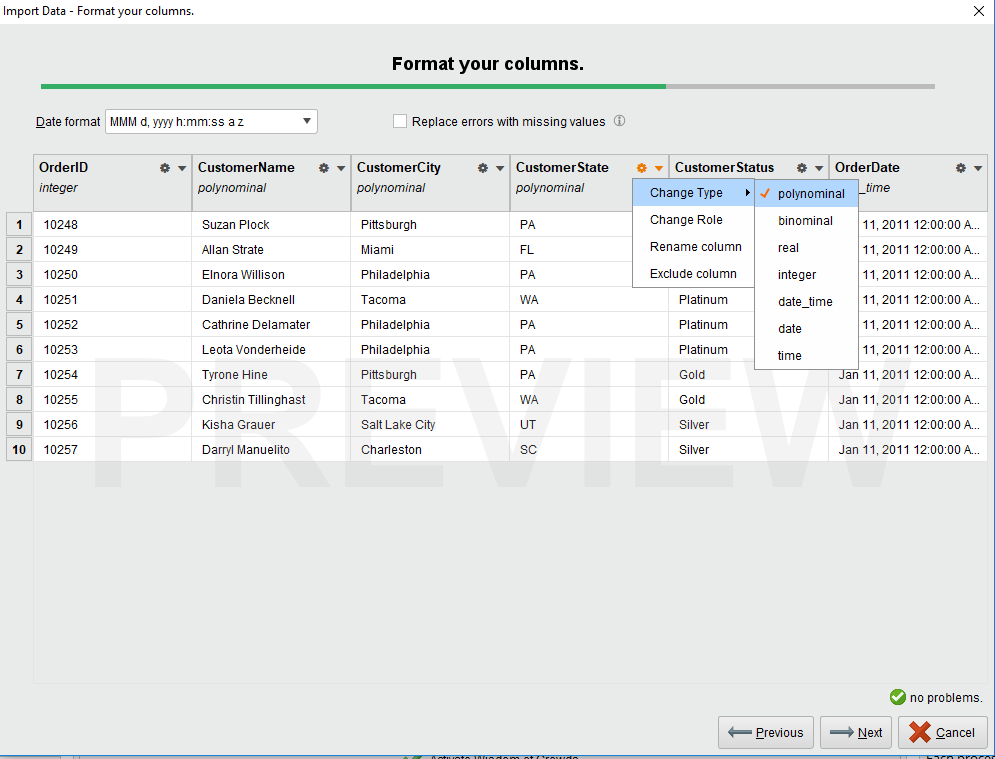
ข้อมูลตั้งต้นเป็น excel 2 ไฟล์ ETLDataSource1.xlsx ภายในไฟล์ประกอบด้วย 3 sheet คือ order, product และ stateLookUp และ ETLDataSource2.xlsx ภายในไฟล์ประกอบด้วย 2 sheet คือ order และ product มี orderID เป็นฟิลด์ที่เชื่อมข้อมูลระหว่าง 2 sheet

1. สร้าง repository เพื่อเก็บ data และ process ของงานที่จะทำ
2. Import file (เลือก Add Data) มาเก็บไว้ใน data ถ้าทำเสร็จเราจะได้ data รวม 5 อัน (ได้จาก 5 sheet ของ excel 2 file)

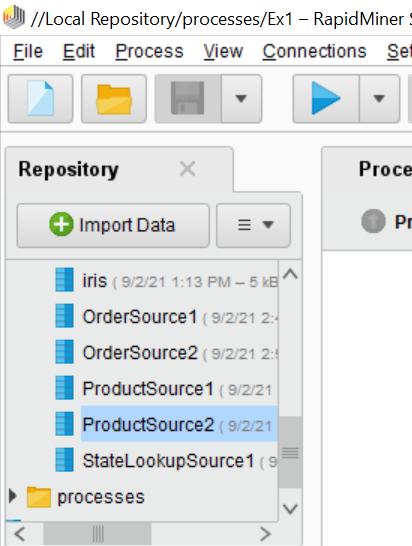


เปิด ETLDataSource1.xlsx เลือกชีทแรกก่อน





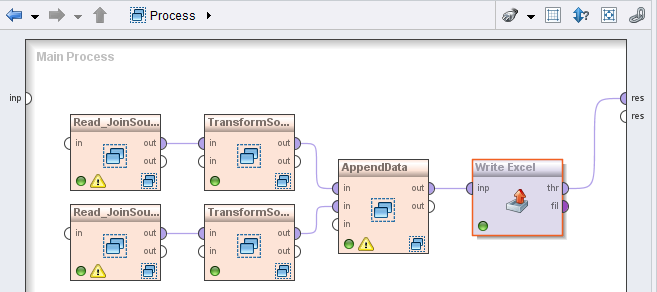
1. การจัดการกับแต่ละฟิลด์
   * ตอน import ไฟล์เข้ามาใน Rapidminer โปรแกรมจะเดาชนิดให้จากข้อมูลที่มีอยู่ เลือกเปลี่ยนได้ตามต้องการ
   * ถ้ายังไม่ได้คลิกเปลี่ยนชนิดตอนนี้ สามารถใช้ operator เพื่อแปลงชนิดข้อมูลได้ในภายหลัง
   * ในแต่ละฟิลด์ นอกจาก Change Type ยังมี Change Role, Rename column, Exclude column
   * ให้ Change Role ฟิลด์ orderID จากทุกชีทให้มีหน้าที่เป็น ID
2. เอาข้อมูลเข้าจนได้ 5 อันใน repository ตามรูป



1. **อันนี้อย่าเพิ่งรีบทำตาม** อันนี้ลองสร้างเป็นแบบ subprocess ให้ดู ซึ่งจะทำให้เห็นภาพรวมก่อนว่ากำลังจะทำอะไร (ที่ทำแบบฝึกหัดกันนี้จะไม่ให้สร้างเป็น subprocess แบบที่ครูทำ **เดี๋ยวข้างล่างจะมีสอนแบบให้เริ่มทำทีละก้อน ก็คือเดี๋ยวเริ่มทำตามที่ ข้อ 6**)

ครูลองสร้าง process ชื่อ ETLprocess ให้เป็นโครงเรื่องทั้งหมดที่จะทำ โดยให้แตกเป็น process ย่อย (subprocess) และภายใน subprocess เราจะเข้าไปบรรจุ process ที่เป็นขั้นตอนละเอียดด้านในนั้นอีกครั้ง

เมื่อร่างโครงของ process แล้ว อย่าเพิ่งเปิดให้สามารถ execute process ได้ (ยังไม่ enable process) ให้ค่อย ๆ execute และตรวจผลลัพธ์ดูทีละกล่อง ตามรูป ซึ่งเป็นรูปของ RM เวอร์ชันเก่า

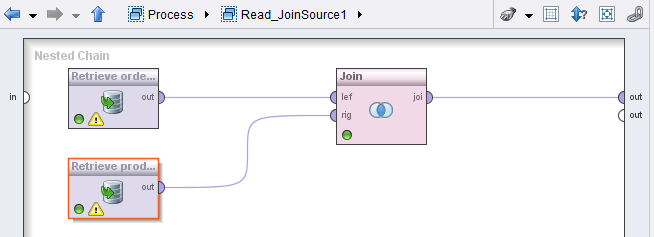


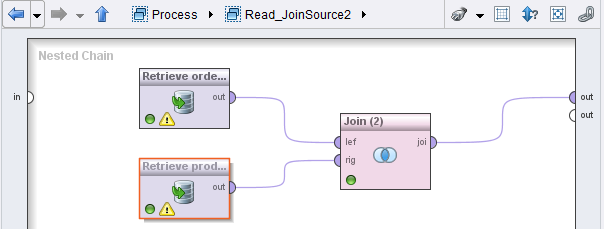
Operator ส่วนใหญ่ที่ใช้จะอยู่ในหมวด data transformation

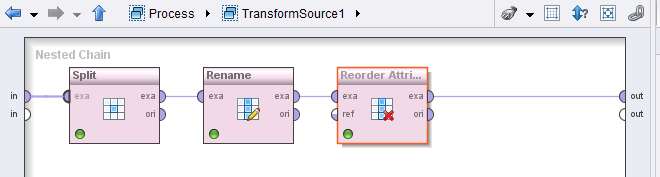
เมื่อเรียกใช้ operator แล้ว จะต้องไปกำหนดค่า parameter ให้เหมาะสมแล้วแต่กรณีด้วย (ด้านล่างของกรอบที่ให้กำหนด parameter เป็น Help บอกรายละเอียดของ operator นั้น ๆ)

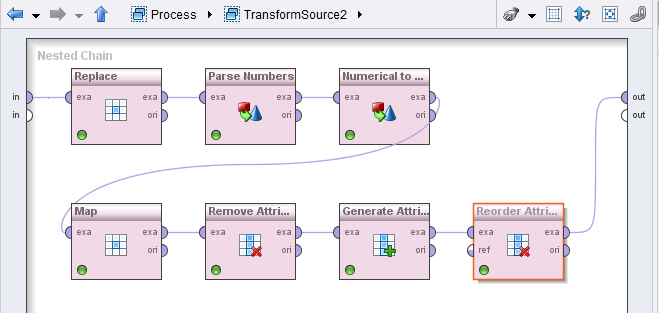
การสร้าง subprocess (อยู่ใน Utility -> Subprocess)

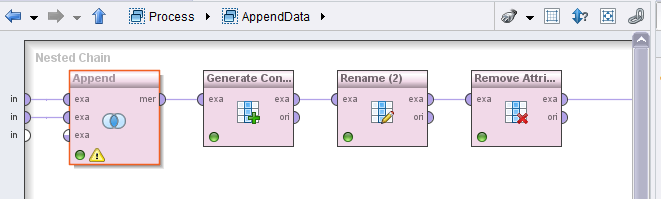
**จอภาพต่อไปนี้แสดงถึง subprocess ในแต่ละ process ใหญ่**





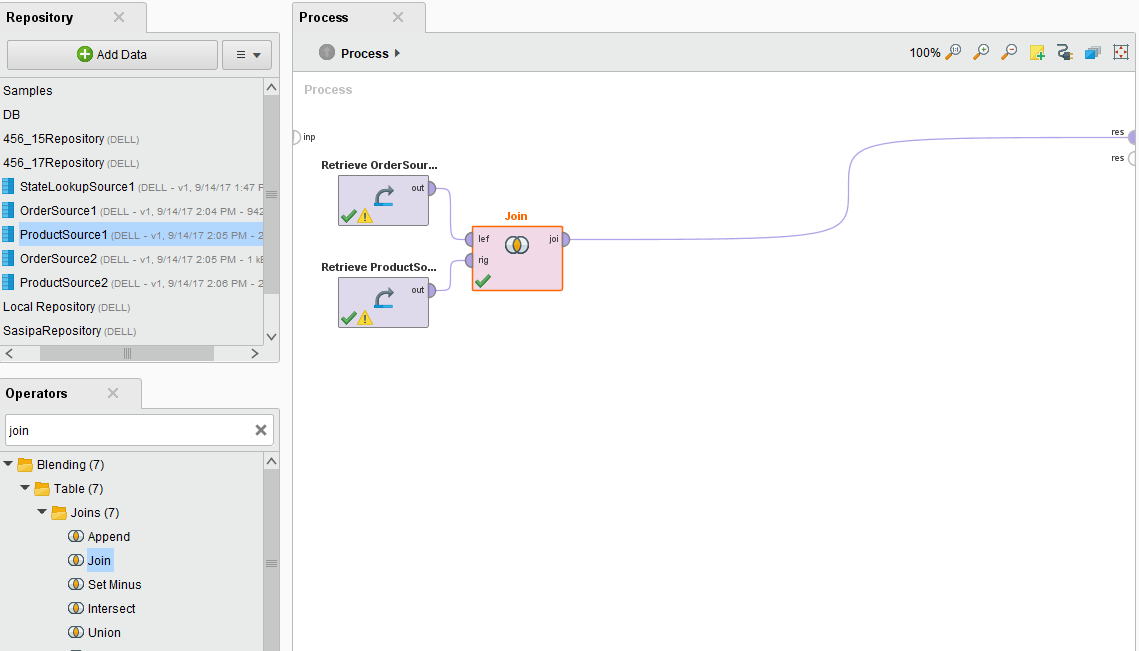
ถ้าใน excel file ETLDataSource1 sheet orderSource1 มีการใช้ฟังก์ชัน vlookup() ช่วยแปลงข้อมูลโดยแทนค่าชื่อย่อของรัฐด้วยชื่อเต็มของรัฐให้เรียบร้อยก่อน แล้วจึงนำไฟล์เข้ามาใน RapidMiner (คือทำความสะอาดข้อมูลตั้งแต่ในไฟล์ excel เลย) ตรง TransformSource1 จะได้รูปดังนี้ แต่ถ้ายังไม่ได้แปลงมาตั้งแต่ใน excel เดี๋ยวดูรายละเอียดการทำที่ด้านล่างได้



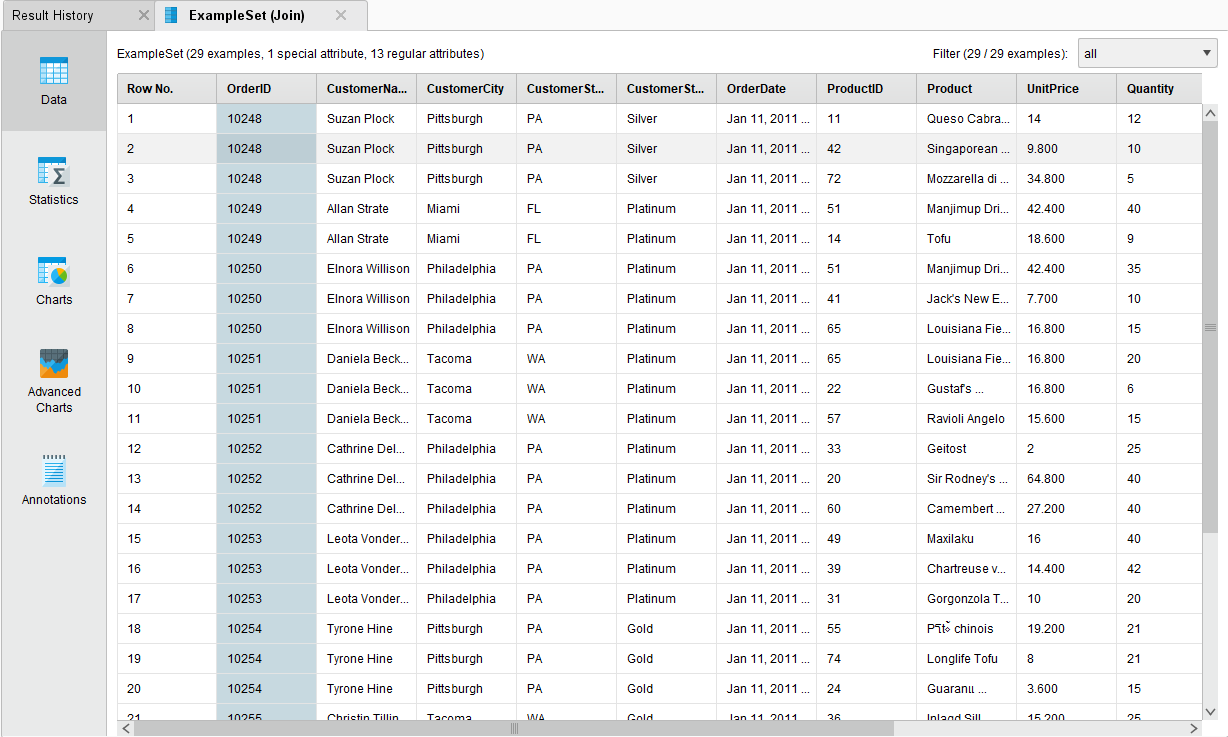


จากตัวอย่างจอภาพด้านบน ถ้าทำตามจะเสียเวลามากกว่า เพราะต้องคอย disable หรือ enable process ลองทำเองจริง ๆ อยากให้ทำไปทีละขั้นตอน ทำหนึ่งก้อน test ผลลัพธ์ดูก่อน ถ้า ok ค่อยไปสร้างก้อนถัดไป จะทำได้เร็วกว่า ตามนี้

1. retrieve file เข้ามาและเรียกใช้ join operator (ใน set operation) เพื่อเชื่อมข้อมูล 2 ชีท เข้าด้วยกัน กำหนดการเชื่อมเป็น แบบ inner join และใช้ตัวเชื่อมคือ orderID



เปลี่ยนจาก Tab Design ไปเลือก Tab Results จะได้ดังจอภาพ



ทำต่อสำหรับ Data Source 2

1. transform data ในส่วนของ data source 1 ประกอบด้วย (ให้ค่อย ๆ ทำทีละ operator แล้วดูผลลัพธ์ ถ้าถูกต้องค่อยต่อด้วย operator ถัดไป)

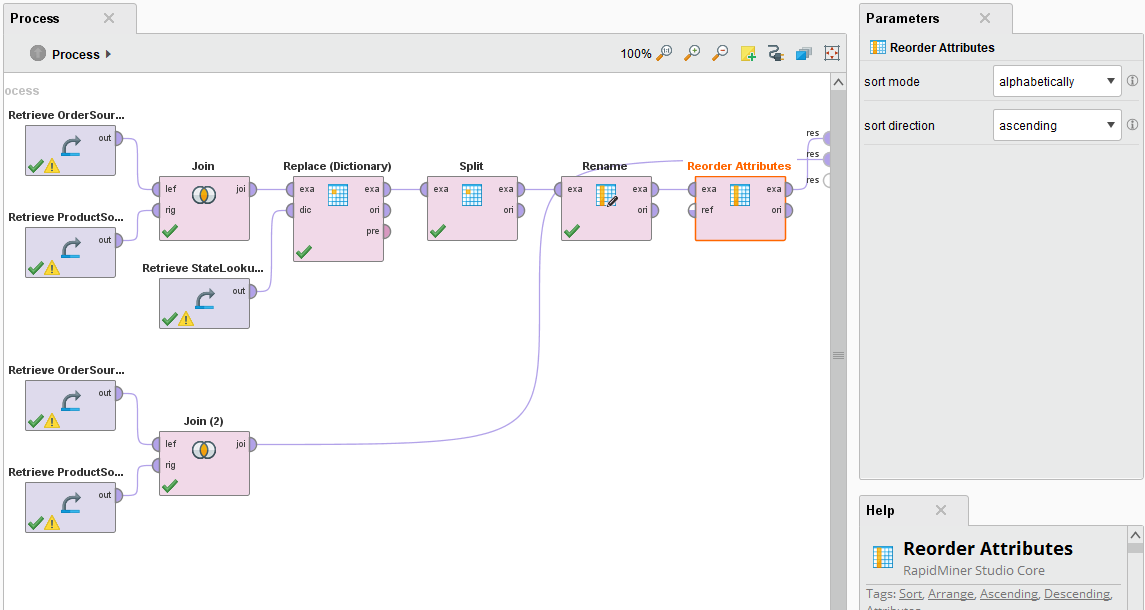
Retrieve StateLookup data เข้ามา แล้วสั่ง Replace(Dictionary) แทนค่าชื่อย่อของรัฐด้วยชื่อเต็มของรัฐ

Split เพื่อแยกฟิลด์ CustomerName ออกมา ตอนใส่ parameter ใส่ split pattern เป็น เคาะ space 1 ที จะได้เป็น CustomerName1, CustomerName2

rename ชื่อฟิลด์ที่มันแยกออกมาให้เป็น CustomerFirstName และ CustomerLastName

reorder attribute เตรียมไว้ก่อน เพราะต่อไปเราจะเอา ไฟล์ 2 ไฟล์มา merge กัน ซึ่งมันจะเช็คว่า ข้อมูลต้อง compatible คือ ฟิลด์ตรงกัน ทั้งตำแหน่ง ชื่อฟิลด์ และชนิดข้อมูล (set parameter ตามรูป)

* + Replace(Dictionary) (อยู่ใน Blending -> Values) – แทนค่าชื่อย่อของรัฐด้วยชื่อเต็มของรัฐ
  + Split (อยู่ใน Value Modification -> Nominal Value Modification) – แยกฟิลด์
  + Rename (อยู่ใน Name and Role Modification) – เปลี่ยนชื่อฟิลด์
* Reorder Attributes (อยู่ใน Attribute Set Reduction and Transformation -> Selection)



1. transform data ในส่วนของ data source 2 ประกอบด้วย

Replace เพื่อทำให้ OrderID เป็นตัวเลข (set parameter เลือก single attribute เลือก OrderID และต้องบอกว่า include special attribute เพราะฟิลด์นี้เป็น ID อยู่ เปลี่ยนจาก A เป็น null ตรง null นี่ไม่ต้องใส่อะไรเลย) แต่เมื่อ replace เสร็จแล้วข้อมูลยังเป็น nominal ก็ให้สั่ง parse numbers ให้แปลงเป็นตัวเลข

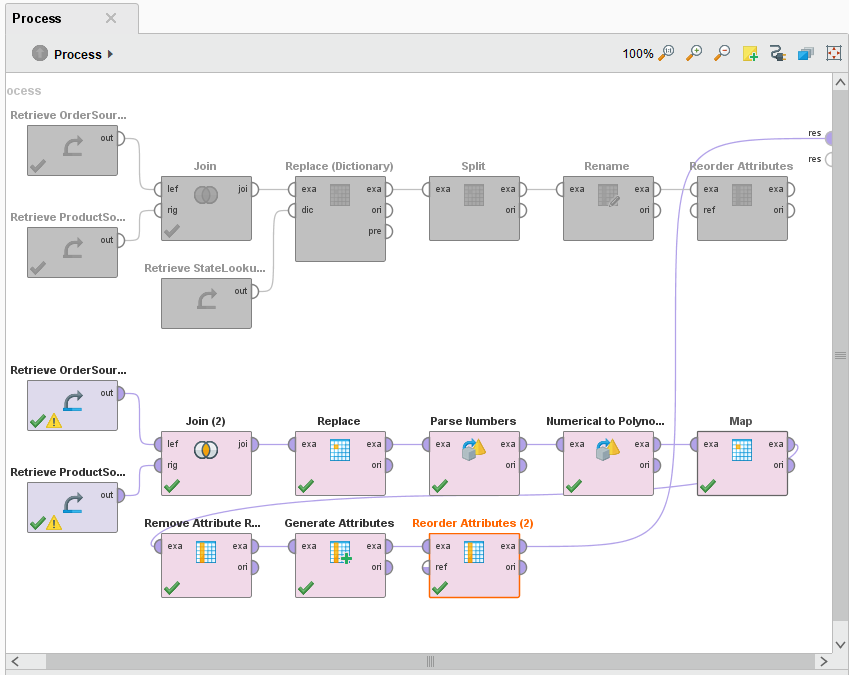
สำหรับ ฟิลด์ customerStatus ให้ map ค่า จาก 1 2 3 เป็น Silver Gold Platinum ซึ่งทำให้เราต้องเปลี่ยนชนิดข้อมูลโดยใช้ Numerical to polynominal และเรียก map operator

Total discount ไม่มีค่าข้อมูลเลย missing values หมด ดังนั้นขอเลือกลบฟิลด์นี้ทิ้งไปก่อนด้วย remove attribute range (เลือกลบฟิลด์ที่ 14) แล้วค่อยเลือกสร้างขึ้นมาใหม่ด้วย generate attribute (เอา FullPrice – ExtendedPrice หรือ FullPrice \* Discount ก็ได้) สุดท้าย reorder attribute

* + Replace (อยู่ใน Value Modification -> Nominal Value Modification) – แทนที่
  + parseNumber (อยู่ใน Type Conversion -> Discretization) มันจะแปลงเป็น integer ให้
  + Numerical to Polynominal (อยู่ใน Type Conversion) – เปลี่ยนชนิดข้อมูลจากตัวเลขเป็น nominal แล้วจึงจะเรียกใช้ map ได้

ฟิลด์ Customer Status ที่มีค่าเป็น 1, 2, 3 แบบ numeric เปลี่ยนให้เป็น nominal ก่อน

* + Map (อยู่ใน Value Modification -> Nominal Value Modification) – map ค่าหนึ่งไปเป็นอีกค่าหนึ่ง
  + Remove Attribute Range (อยู่ใน Attribute Set Reduction and Transformation -> Selection) – ลบฟิลด์ในช่วงที่กำหนด (ฟิลด์แรกนับเป็น 1)
  + Generate Attribute (อยู่ใน Attribute Set Reduction and Transformation -> Generation) – สร้างฟิลด์ใหม่ โดยคำนวณจากฟิลด์ที่มีอยู่
  + Reorder Attributes (อยู่ใน Attribute Set Reduction and Transformation -> Selection)



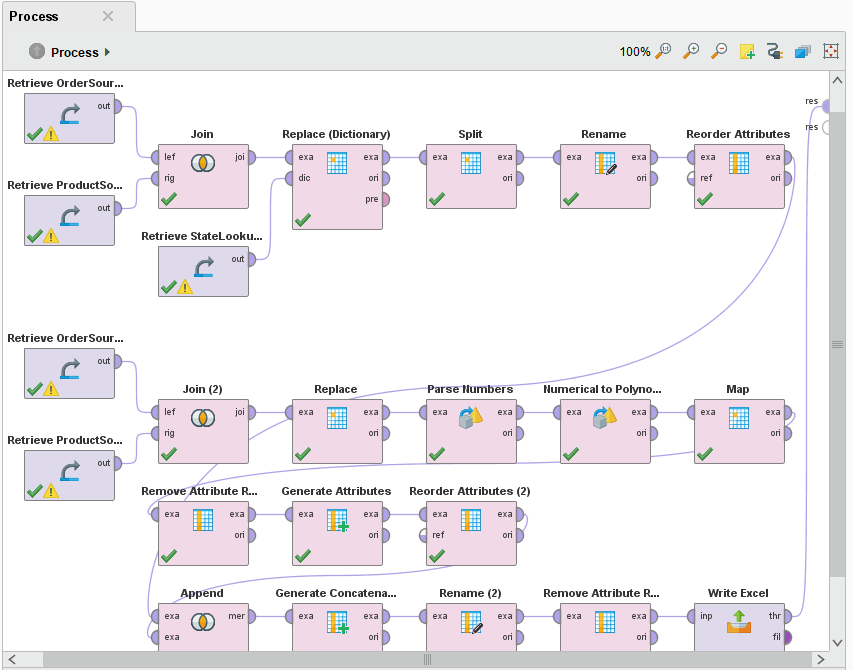
1. เอาผลลัพธ์จากทั้งสอง source มาต่อรวมกันเป็นตารางเดียว

สั่ง Append ให้เอา 2 ตารางมาต่อกัน

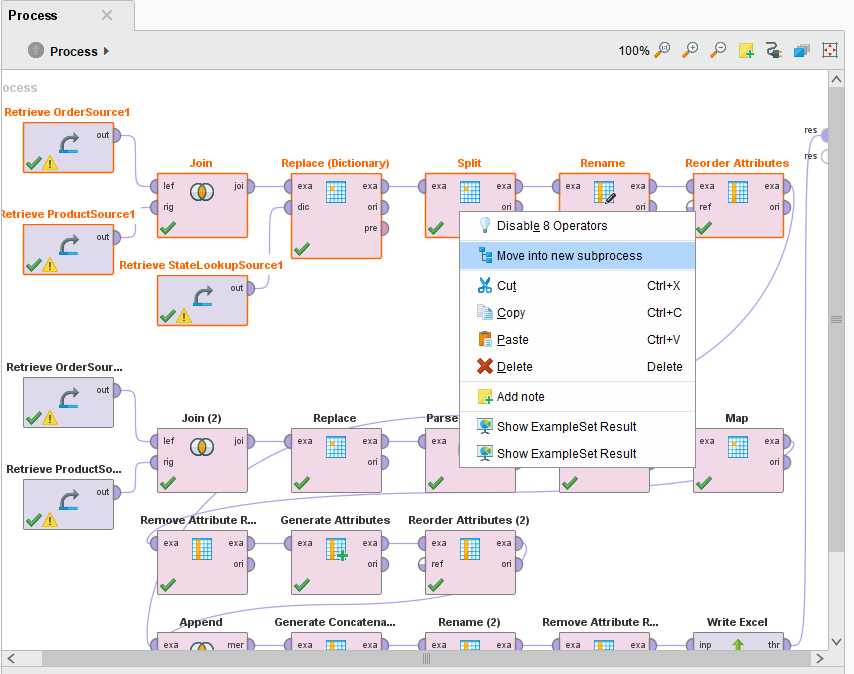
Generate concatenation ทำเพื่อเชื่อมฟิลด์ ชื่อ สกุล ให้เป็น ฟิลด์รวม แล้วตั้งชื่อฟิลด์ใหม่ (rename) เป็น customerName และลบ customerFirstName และ customerLastName ทิ้งไป

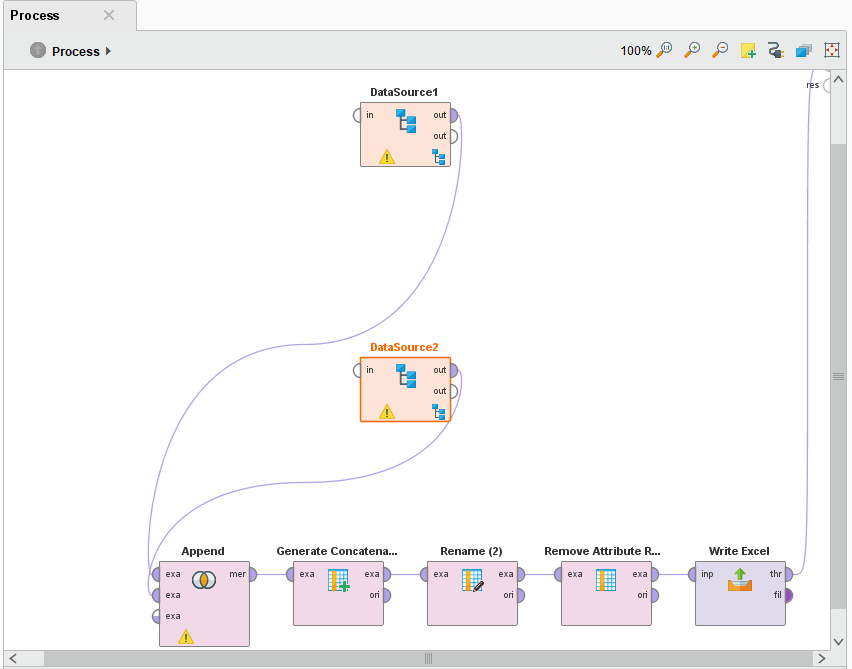
* + Generate Concatenation (อยู่ใน Attribute Set Reduction and Transformation -> Generation) – รวมฟิลด์

1. สุดท้าย สั่ง write excel เพื่อเขียนไฟล์ผลลัพธ์เก็บลงดิสก์
   * Write Excel (อยู่ใน Data Access -> Files -> Write) connect from thr port to res port



ลองสร้าง subprocess ทีหลังก็ได้





Save process ไว้ใน local repository

จะได้ file 2 file นามสกุล .rmp และ .properties

อยู่ใน folder

C:\Users\ชื่อuserของเรา\.RapidMiner\repositories\Local Repository

(ถ้า save ใน subfolder ชื่อ process ก็จะเข้าไปอีกชั้นนึง)