Data Analytics Foundations

Online Workshop 2
Introduction to Data Analytics

Questions?

KNIME

- Download the iris.csv file from the workshop page in Canvas

- Import it into KNIME using CSV Reader or File Reader

- Filter the data to exclude all rows where Sepal.Length > 6
  1. Using Row Filter
  2. Using Rule-based Row Filter

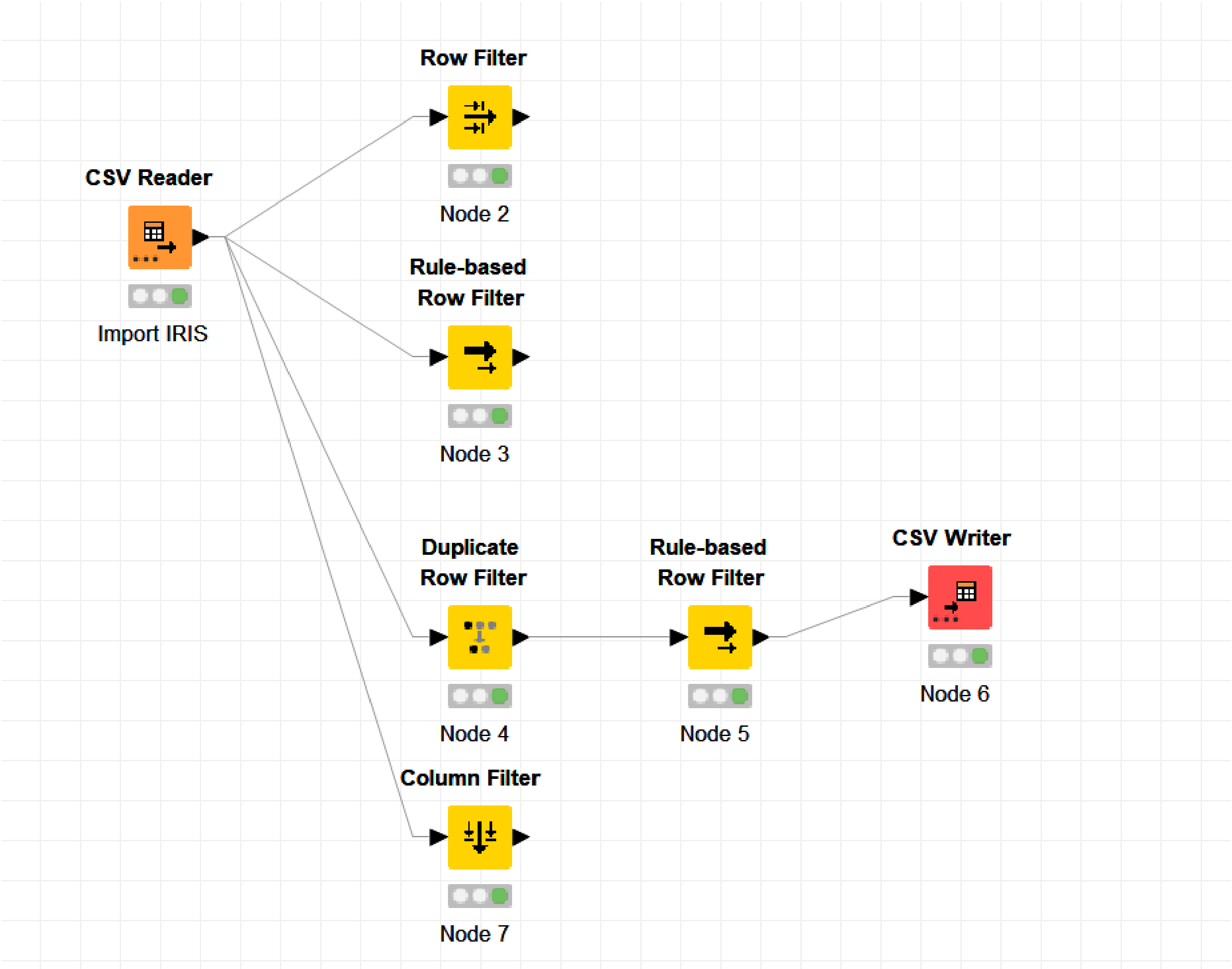
- Export the filtered dataset using CSV Writer and verify that both methods give the same result

KNIME

- Explore LIKE, IN, AND, OR, NOT options in Rule-based Row Filter
  - Include all the rows contains “se” in Species
  - Include all the rows where Sepal.length >=4 or Sepal.Width>=3
  - Exclude the rows where Species equal to “versicolor”

- Remove the duplicate rows using “Duplicate Row Filter”
- Filter the columns using KNIME node “Column Filter”

Workflow


Row Filter
Node 2
CSV Reader
Import IRIS
Rule-based
Row Filter
Node 3
Duplicate
Row Filter
Node 4
Rule-based
Row Filter
Node 5
CSV Writer
Node 6
Column Filter
Node 7

KNIME

Try the following rules:

```
//exclude all the rows in sepal length greater than 6
$Sepal.Length$  >= 6 => TRUE
$Petal.Length$>1 AND $Petal.Width$<1 => TRUE
($Petal.Length$>2 OR $Petal.Width$<1) AND $Species$="setosa" => TRUE
$Species$ IN ("setosa","versicolor") => TRUE
$Species$="setosa" OR $Species$="versicolor" => TRUE
$Species$ LIKE "*to*"=>TRUE
$Species$ LIKE "ve*"=>TRUE
$Species$ LIKE "*sa"=>TRUE

$Sepal.Width$>=4 => FALSE
TRUE=>TRUE

//Include all the rows contains “se” in Species
$Species$ LIKE "*se*" => TRUE
//Include all the rows where Sepal.length >=4 or Sepal.Width>=3
$Sepal.Length$ >= 4 OR $Sepal.Width$ >= 3 =>TRUE
//Exclude the rows where Species equal to “versicolor”
$Species$ = "versicolor" => TRUE
```

Using data (Q&A session)

Thinking of your own area of work within your company…

- How could this data be used to better understand the business processes?
- How could this data be used to improve the business processes?
- Are these supervised or unsupervised problems?