KNIME Cluster Extension

Unified Cluster Execution Framework for KNIME Workflows

Nico Hoffmann (TU Dresden, ZIH)

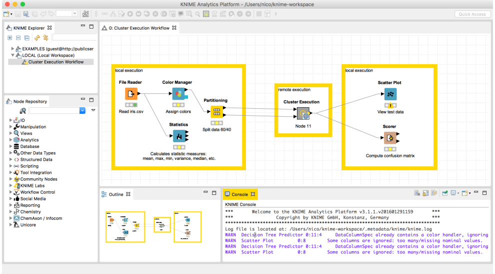
June 23, 2016

Table of Contents

1. KNIME Analytics Platform
2. Case Study
3. KNIME Cluster Extension
4. UNICORE Integration
5. Summary

Table of Contents

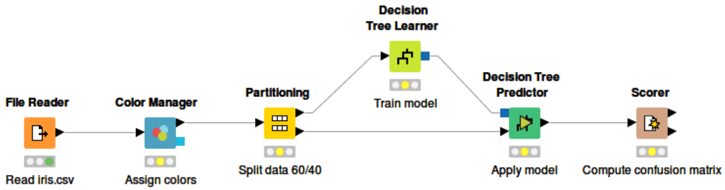
KNIME Analytics Platform



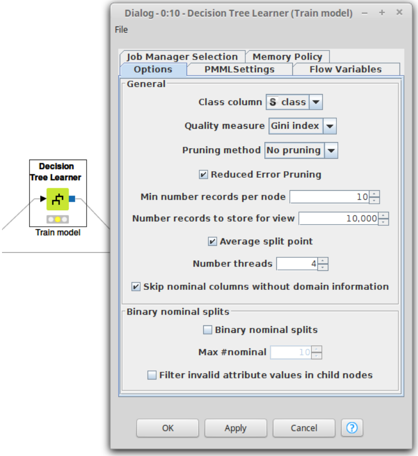
```
KNIME Console
***       Welcome to the KNIME Analytics Platform v3.1.1.v201601291159       ***
***                Copyright by KNIME GmbH, Konstanz, Germany                ***
******************************************************************************
Log file is located at: /Users/nico/knime-workspace/.metadata/knime/knime.log
WARN  Decision Tree Predictor 0:11:4     DataColumnSpec already contains a color handler, ignoring
WARN  Scatter Plot          0:8       Some columns are ignored: too many/missing nominal values.
WARN  Decision Tree Predictor 0:11:4     DataColumnSpec already contains a color handler, ignoring
WARN  Scatter Plot          0:8       Some columns are ignored: too many/missing nominal values.
```

- open source software for advanced analytics
- integration of many other tools and data sources
- easy to use graphical workbench
- extensible through plug-ins

Exemplary KNIME workflow



Node: single algorithm



General

Class column: S class

Quality measure: Gini index

Pruning method: No pruning

☑ Reduced Error Pruning

Min number records per node: 10

Number records to store for view: 10,000

☑ Average split point

Number threads: 4

☑ Skip nominal columns without domain information

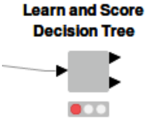
Binary nominal splits

☐ Binary nominal splits

Max #nominal: 10

☐ Filter invalid attribute values in child nodes

Metanode: mechanism to wrap subworkflow



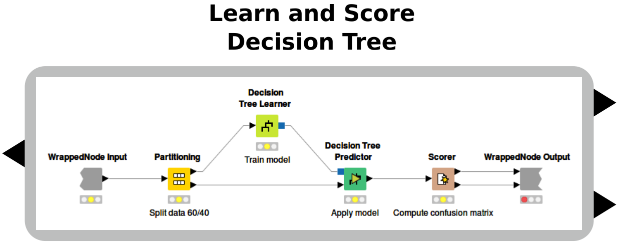
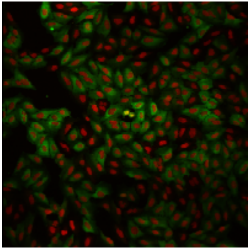
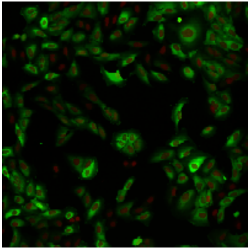
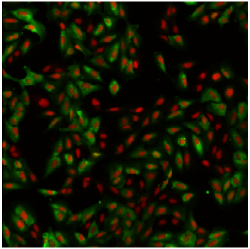


Table of Contents

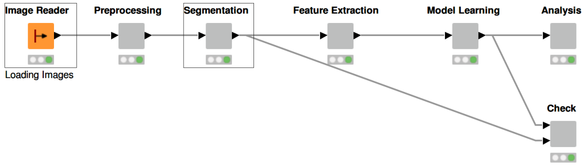
1. KNIME Analytics Platform
2. Case Study
3. KNIME Cluster Extension
4. UNICORE Integration
5. Summary

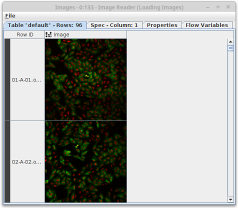
Life Sciences

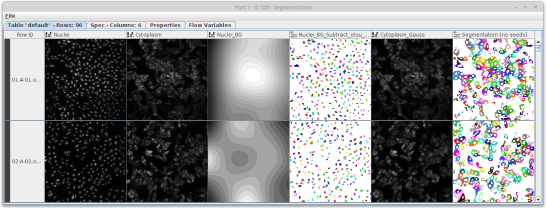
- biologist has many cell images
- goal: infer knowledge about cell behaviour
- no scripting knowledge
- basic knowledge of image analysis
- access to a computer cluster

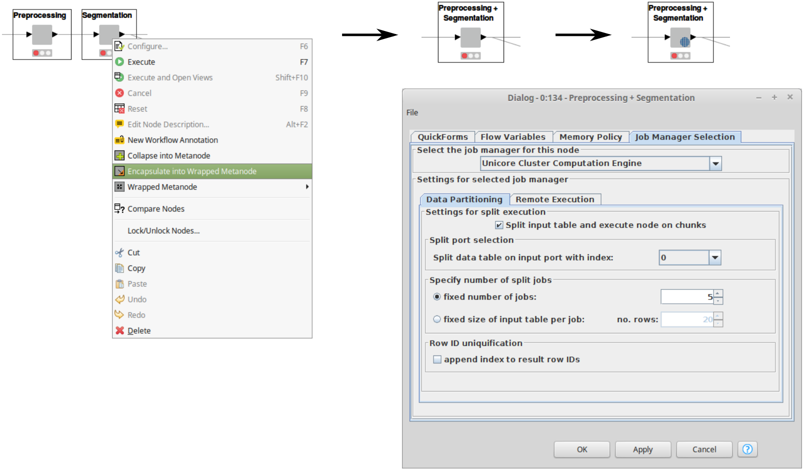
Data-parallel workflows are commonly observed







Aim



Dialog - 0:134 - Preprocessing + Segmentation

Select the job manager for this node: Unicore Cluster Computation Engine

Settings for selected job manager

Settings for split execution

☑ Split input table and execute node on chunks

Split port selection

Split data table on input port with index: 0

Specify number of split jobs

● fixed number of jobs: 5

○ fixed size of input table per job: no. rows: 20

Row ID uniquification

□ append index to result row IDs

Table of Contents

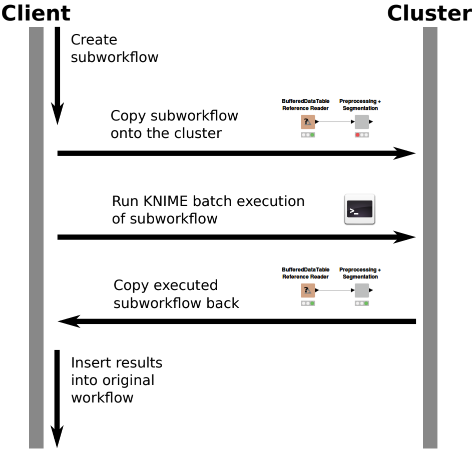
1. KNIME Analytics Platform
2. Case Study
3. KNIME Cluster Extension
4. UNICORE Integration
5. Summary

Overview

General steps on cluster

1. provide data
2. set environment variables
3. execute knime on subworkflow
4. signal knime termination
5. fetch results

Detailed View on Distributed Processing Workflow


BufferedDataTable Reference Reader
Preprocessing + Segmentation
BufferedDataTable Reference Reader
Preprocessing + Segmentation

Client

Cluster

Create subworkflow

Copy subworkflow onto the cluster

Run KNIME batch execution of subworkflow

Copy executed subworkflow back

Insert results into original workflow

Creation of Subworkflows

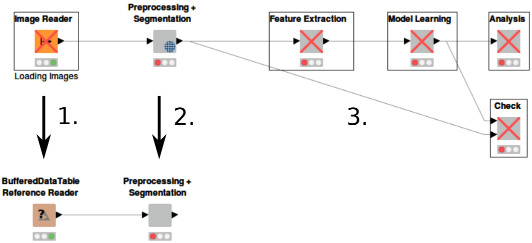
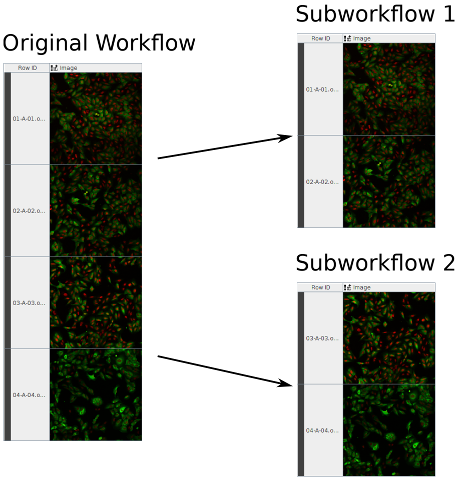


Image Reader
Loading Images
Preprocessing + Segmentation
Feature Extraction
Model Learning
Analysis
Check
1.
2.
3.
BufferedDataTable Reference Reader
Preprocessing + Segmentation

1. replace previous nodes with readers containing the input table/object
2. copy node and replace executor with local
3. remove all nodes after the node that should be executed on the cluster

Splitting the Input Data



Original Workflow

Row ID	Image
01-A-01.o...	
02-A-02.o...	
03-A-03.o...	
04-A-04.o...	

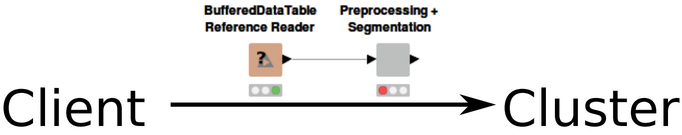
Subworkflow 1

Row ID	Image
01-A-01.o...	
02-A-02.o...	

Subworkflow 2

Row ID	Image
03-A-03.o...	
04-A-04.o...	

Pushing the Workflow onto the Cluster


BufferedDataTable Reference Reader
Preprocessing + Segmentation
Client
Cluster

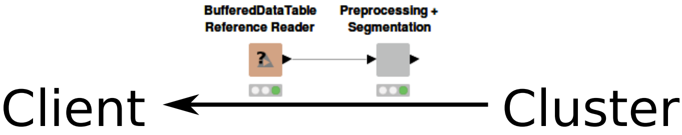
Shared Filesystem

Upload

Cluster Execution

- create job description
- allocate resources for job
- execute KNIME subworkflow on each allocated node

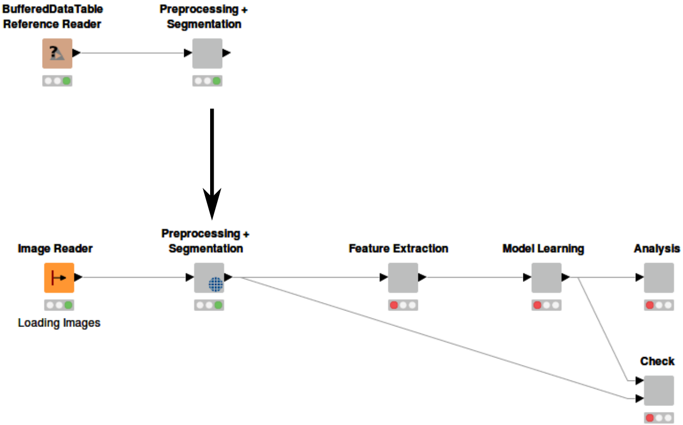
Pulling the Executed Workflow back to the Client


BufferedDataTable Reference Reader
Preprocessing + Segmentation

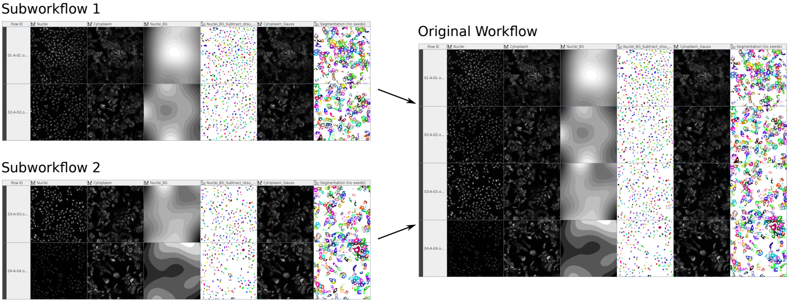
Client ← Cluster

- Shared Filesystem
- Download

Inserting the Results back into the Original Workflow


BufferedDataTable Reference Reader
Preprocessing + Segmentation
Image Reader
Loading Images
Preprocessing + Segmentation
Feature Extraction
Model Learning
Analysis
Check

Concatenating the Result Data



Subworkflow 1

Row ID	Nuclei	Cytoplasm	Nuclei_BG	Nuclei_BG_Subtract_otsu_...	Cytoplasm_Gauss	Segmentation (no seeds)
01-A-01.o...						
02-A-02.o...						

Subworkflow 2

Row ID	Nuclei	Cytoplasm	Nuclei_BG	Nuclei_BG_Subtract_otsu_...	Cytoplasm_Gauss	Segmentation (no seeds)
03-A-03.o...						
04-A-04.o...						

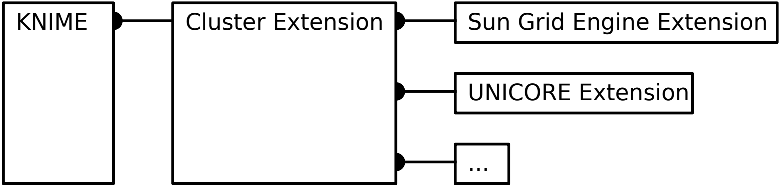
Original Workflow

Row ID	Nuclei	Cytoplasm	Nuclei_BG	Nuclei_BG_Subtract_otsu_...	Cytoplasm_Gauss	Segmentation (no seeds)
01-A-01.o...						
02-A-02.o...						
03-A-03.o...						
04-A-04.o...						

Table of Contents

1. KNIME Analytics Platform
2. Case Study
3. KNIME Cluster Extension
4. UNICORE Integration
5. Summary

The Framework and Specific Integrations



Cluster extension:

- subworkflow creation
- script for execution
- join results

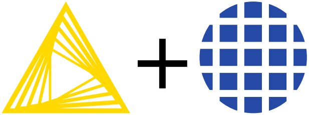
Specific extension:

- cluster specific settings
- communication with cluster

UNICORE

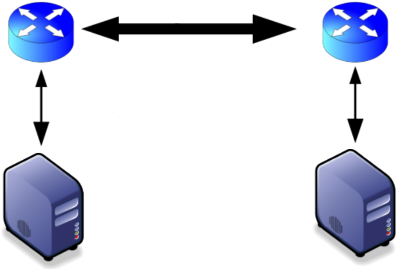
- open source
- offers connection via its own client or via RESTful API
- middleware in between client and batch system
- connects to the most popular systems like SLURM, Torque, LSF, . . .

Communication by Unicore's RESTful API



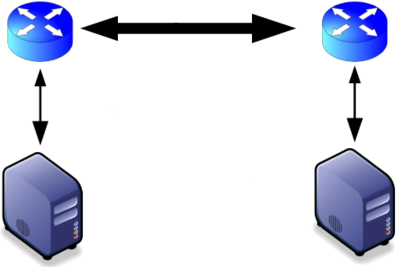
- cooperation with Patrick Winter, Universität Konstanz
- based on UNICORE's REST interface
- job description created automatically (user can specify required resources)
- uses HTTP GET/PUT for download/upload
- no shared filesystem required
- efficient handling of data that is already present on the cluster

UNICORE Data Staging



- transferring data to the worker that processes it
- multiple transfers in parallel
- optional encryption and/or compression
- using transfer protocols like UFTP, FTP, BFT, . . .

UNICORE Data Staging

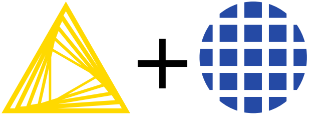


- prototype locally on subset, execute remotely on whole data
- import data from Lustre, FTP, HTTP, cloud storages, ...
- only data required for remote workflow execution is imported
- future: enforce data privacy laws

Table of Contents

1. KNIME Analytics Platform
2. Case Study
3. KNIME Cluster Extension
4. UNICORE Integration
5. Summary

KNIME Analytics Platform + Cluster Extension



KNIME Analytics Platform:

- software for advanced analytics
- integration of many other tools and data sources
- open source
- extensible through plug-ins

Cluster Extension:

- works with all of KNIME's integrations
- open source
- cluster support extensible through plug-ins

Thank you for your attention!

More Information. . .

KNIME:
https://www.knime.org/

KNIME Learning Hub:
https://www.knime.org/learning-hub/

KNIME Beginner's Luck:
https://www.knime.org/knimepress/beginners-luck
Promotion code for a free copy:
ScaDS2016