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

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

- File Reader: Read iris.csv
- Color Manager: Assign colors
- Partitioning: Split data 60/40
- Decision Tree Learner: Train model
- Decision Tree Predictor: Apply model
- Scorer: Compute confusion matrix
Node: single algorithm
Metanode: mechanism to wrap subworkflow
| 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

KNIME Cluster Extension

UNICORE Integration

Summary

Nico Hoffmann, Patrick Winter

KNIME Analytics Platform
Case Study
KNIME Cluster Extension
UNICORE Integration
Summary
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

Client

Create subworkflow

Copy subworkflow onto the cluster

Run KNIME batch execution of subworkflow

Copy executed subworkflow back

Insert results into original workflow

Cluster
Creation of Subworkflows

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

Subworkflow 1

Subworkflow 2
Cluster Execution

- create job description
- allocate resources for job
- execute KNIME subworkflow on each allocated node
Pulling the Executed Workflow back to the Client

Client ← Shared Filesystem ← Cluster

Download

Nico Hoffmann, Patrick Winter

KNIME Cluster Extension
Inserting the Results back into the Original Workflow
Concatenating the Result Data

Subworkflow 1

Subworkflow 2

Original Workflow

Nico Hoffmann, Patrick Winter
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!
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