Running KNIME Workflows in a Compute Cluster Environment

Compute clusters often run idle because of a lack of applications that can be run in a cluster environment and the enormous effort required to operate, maintain, and support applications on the grid. KNIME Cluster Execution tackles this problem by providing a thin connection layer between KNIME and the cluster. This allows every node running in KNIME and every application integrated in KNIME to be executed on the cluster. Submission of data to the cluster and collection of results is very simple. Long-running analysis workflows can be executed on the compute cluster, freeing up local resources for other productive work. Expensive licenses can be shared by running the referring nodes on the compute cluster, reducing costs.

Cluster Execution Concept

Workflows are created in the familiar manner using KNIME Analytics Platform. Now the user can make selections for each individual node to specify whether it should be executed locally on the PC or submitted to the cluster. A preference setting in KNIME specifies which cluster engine is used. Node-specific settings can also be entered to define how the cluster is used.

The data that will be processed can be split into multiple subsets (see picture above), which are subsequently sent to the master server in the cluster. The master coordinates execution of the task.

In split workflows, parallel nodes can be executed on separate resources in the cluster (see picture overleaf). Once the remote execution of all tasks is finished KNIME collects the results from the remotely-executed nodes and completes the workflow execution locally.

Applications

Workflows that process a large number of data sets or that require CPU-intensive calculations are ideal candidates for using KNIME Cluster Execution. The extension is already successfully used for 3-D modelling calculations and virtual docking experiments in pharmaceutical research.

Another area of use is the analysis and mining of huge data volumes found in CRM and telecom databases, or financial systems.

KNIME Cluster Execution is used in these industries by power-users to provide calculation results in a reasonable time frame, which would be too time consuming with single system execution.

This enables the KNIME user to provide mission critical results to business processes and management.

KNIME Integration

KNIME Cluster Execution is an extension of KNIME Analytics Platform. It can be easily integrated with KNIME TeamSpace for advanced use of shared metanodes that can be submitted to the compute cluster as well.

The KNIME Server can also be combined with KNIME Cluster Execution to start calculation intensive tasks on the server where the workflow containing cluster enabled nodes can call out to the compute cluster. On workflow completion, the results can be downloaded from the server or full reports can be saved in the desired report document format including PDF, Word, Excel or PowerPoint.

In combination with KNIME WebPortal, the power of the cluster can be used by casual users in a controlled and simple web browser interface.
Advantages

An important benefit is the gain in performance of calculation intensive workflows.

KNIME Cluster Execution enables you to disconnect from running jobs to continue work on other urgent tasks and later reconnect to those jobs to check for status changes and retrieve the results.

Third-party nodes can also be routed to dedicated servers, making it possible to distribute software that does not usually offer cluster support. Operation of the cluster is simple and transparent to the user.

Interaction

KNIME Cluster Execution is a KNIME commercial productivity extension.

In combination with the KNIME Server the value of the overall solution can be increased by adding remote workflow storage, user access classes, and web access functionality.

Contact KNIME to obtain more details about KNIME Cluster Execution or the other products in the KNIME.com AG product family.

Software Prerequisites

Submit Clients: Linux (32 and 64 bit)
Cluster Engine: Grid Engine 6.2
Cluster Slaves: Linux (32 and 64 bit)

KNIME Collaborative Extensions

...activate the full potential of KNIME for teams. Features include user rights & authentication, remote & scheduled execution, shared workflow repository, data space, and metanodes plus access to workflows, reports and web services.

KNIME Analytics Platform

...is an open source platform for integrated data access, data mining, statistics, visualization and reporting.

KNIME Productivity Extensions

...increase the speed at which KNIME workflows can be created, reused and maintained for both individuals and partners who use KNIME for their clients.

KNIME Performance Extensions

...enable distributed storage and scalable execution of KNIME workflows, enabling scale for large data sets and complex computation requirements.

KNIME Community & Partner Extensions

...are nodes created by and given back to the KNIME community. These range from broadly useful application nodes simplifying the work of anyone using KNIME to domain-specific solutions to pressing problems for specialists.

About KNIME

KNIME is the leading open platform for data-driven innovation helping organizations to stay ahead of change. Innovative organizations use our open-source, enterprise-grade analytics platform to discover the potential hidden in their data, mine for fresh insights or predict new futures.

Quick to deploy, easy to scale, and intuitive, KNIME is used in over 60 countries on data of every kind: from numbers to images, molecules to humans, signals to complex networks, from kilo- to petabytes or simple reports to complex analyses.

KNIME is developed and supported by KNIME.com AG. Learn more at www.knime.com