KNIME Newsletter

April 2011 - Vol. 1, No. 2

Welcome to the second KNIME Newsletter!

Tripled attendance: over 130 attendees at the 4th KNIME User Group Meeting.

The first quarter of 2011 has been busy. The latest Rexter Analytics Survey indicated that KNIME has grown strongly, also in the area of data analytics setups, with reports of strong user satisfaction across all domains. However, our 4th KNIME User Group Meeting was undoubtedly the highlight of the past three months! Not only did we almost triple attendance (133!) in comparison to last year but we also saw various presentations about the productive use of KNIME and our server tools in diverse application areas. In addition, ten KNIME partners showed what they do with KNIME in our exhibition area. The demo of upcoming KNIME improvements raised a lot of interest—look out for information in future KNIME Newsletters!

Modern Research needs Freedom for all Types of Users

Andy Palmer
Global Head of Software and Database Engineering
Novartis

Freedom is what modern life science research needs most. The freedom to work easily with heterogeneous data sources. The freedom to switch quickly between tools of different origin. The freedom to integrate existing legacy tools into a flexible processing environment. And the freedom for different types of users to work the way they want. With the advent of KNIME, this desire is now being fulfilled.

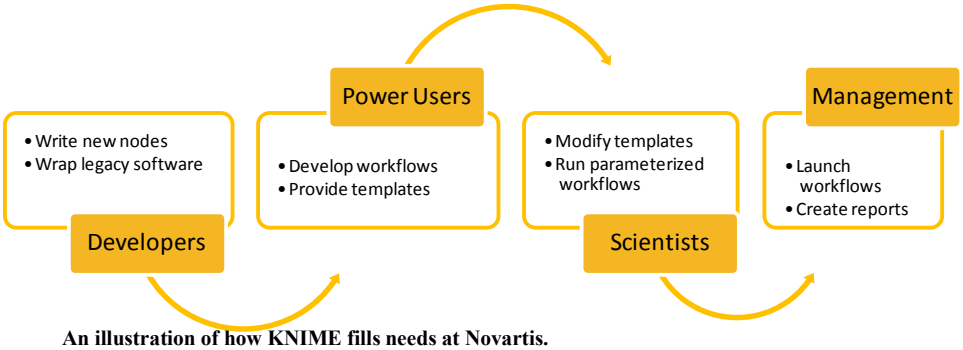
"In modern life science research, KNIME gives users the freedom to work as THEY choose to work."

At Novartis a whole variety of users are benefiting from KNIME and the KNIME Server infrastructure. Developers are creating workflows and packaging their own analytic and modeling modules in the KNIME platform for automation and sharing. Power users can focus on using preconfigured workflows and modify them to fit their current work at hand. The occasional user accesses KNIME through an easy to use web interface, which only exposes the relevant parameters and resources required for the underlying workflow. And finally, research management sees tremendous value in having a preexisting set of template workflows that standardize preferred types of data processing, analysis, and modeling across the teams – regardless of user type.

The support provided by Novartis for the continued development of KNIME is a great indicator of the true power of KNIME as an open-source program. But Novartis not only sponsors KNIME development. Under the guidance of Greg Landrum, one of the founders of the RDKit project and now Global Head of Chemical Information Systems at Novartis, the integration of RDKit into KNIME has produced an impressive set of nodes from which other pharmas and researchers can benefit. Other community contributions to the KNIME platform are also available to extend the breadth of the Novartis KNIME installation. Another strength of KNIME is the integration of in-house legacy tools. Nodes that call shell scripts, web services but also in-house developed wrappers around existing tools make the KNIME workflows more powerful by leveraging Novartis' existing IP.

"We are very impressed with our early work with KNIME. Being able to satisfy the needs of our different types of users from within one modern platform is very important and is the ultimate measure of success." says Andy Palmer, Global Head of Software and Database Engineering of Novartis.

While power users are extremely important, they are in the minority. KNIME enables them to create and share powerful workflows that encapsulate the entire process from data loading and integration all the way to analysis, modeling and visualization. And all this without being forced to migrate every piece of their existing work. User can incorporate it into their own KNIME nodes, which allow the use of their existing programs and tools. The resulting workflows are uploaded to the KNIME Server as templates or starting points for other users. And finally, for less experienced users who really do not care about the underlying technology but want to get a job done, KNIME workflows can be called from the server web front-end directly. The QuickForm setup which was demonstrated at the recent KNIME User Group Meeting will make this even more powerful .



An illustration of how KNIME fills needs at Novartis.

KNIME news

KNIME Press Launched

Under KNIME Press we are publishing a series of e-books on the use of KNIME.

The first book, a beginners guide, is already available for online purchase under:
www.knime.org/knimepress/

Upcoming User Training

If you want to learn more about how to use KNIME and KNIME Reporting, you can now enroll for another one of our very popular monthly training courses.
May 9 - 11, 2011
Technopark
Zurich, Switzerland
Visit:
www.knime.org/training
to register and for more information.

KNIME to present at the ChemAxon UGM

Meet us in Budapest on **May 16 - 18, 2011.**

Get an update on the KNIME.com product suite.
Talk to the people from KNIME.
Learn how to use the JChem Extensions for KNIME.

Details can be found at:
www.knime.org/ChemAxon2011UGM

KNIME comes to San Francisco!

After a great event in Boston last year, we will be meeting you and our Life Science Partners this year in the SF bay area on **July 28, 2011.**

Details are available at:
www.knime.org/LifeScienceDay2011

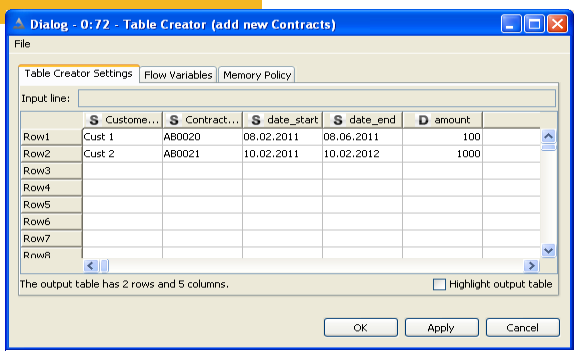
What (else) is new in KNIME 2.3?

The *Table Creator* node

KNIME is a powerful tool for data analysis and reporting. Its data manipulation category helps to build data warehousing solutions to consolidate and keep data for analysis up to date at all times. Data consolidation deals with many different files and/or databases. There are files that are automatically exported at regular times; files that are generated on demand; and unfortunately files that can only be updated by the hand of some employee.

It is particularly ironic if you have to manually re-import data that has been collected by KNIME generated survey forms. In this case, the user generates the form with KNIME Reporting, closes KNIME, imports the collected data with Excel, reopens KNIME and runs the data analysis.

In KNIME 2.3 the Table Creator node also acts as a substitute for Excel.



Configuration window of the Table Creator node

From your workflow, simply enter the new data by hand and append it to the original data archive by means of the File Writer.

The configuration window for the Table Creator node shows a table whose cells can be overridden. You can also change the table header's names and types by right-clicking the column headers.

Tips and Tricks

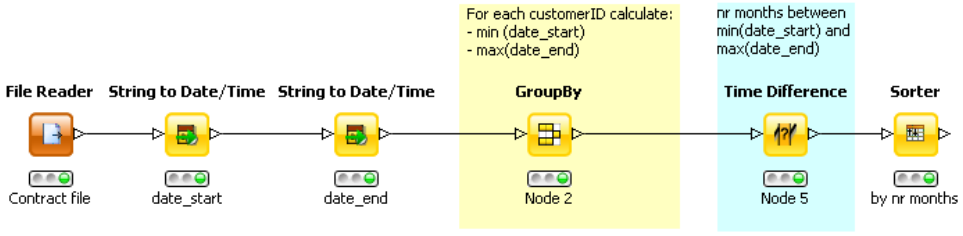
Thomas Gabriel
KNIME Team

Calculate how many months a customer has been around

In order to measure customer loyalty, it is mandatory to know how long a customer relationship has been in existence. For example, it would be interesting to know for each customer how many months have passed between the subscription of the first contract and the expiration of the last one. First of all we need to get the full list of contracts from a file. This list would be more or less organized as follows: customer ID, contract ID, date start of contract, date end of contract, amount of money.

KNIME represents times and dates by using a particular data type: the *DateTime* type. In case the input values date_start and date_end can only be read as *String* type, let's convert them immediately from *String* to *DateTime* type with two *String To Date/Time* nodes.

A *GroupBy* node can then be used to group all contracts by customer ID; even better it can be used to calculate the minimum of the date_start and the maximum of the date_end in the group of contracts for each customer ID. The output data table will be organized as follows: customer ID, minimum of date_start, maximum of date_end. Finally, a *Time Difference* node for each row calculates the time difference, in terms of number of months, between the minimum of date_start and the maximum of date_end. As a result, each row displays the customer ID together with the number of months between the earliest date of a valid contract and the latest expiring date.



The "Time Series" category

For those who have not been able to find the "Time Difference" node, it is located in the "Time Series" category. This category contains a number of nodes to transform and manipulate *DateTime* values.

We have already encountered the *Time Difference* node that calculates the time elapsed between the values in two *DateTime* columns. In the same category we can also find a few nodes to convert from String to *DateTime* type and vice versa. Two nodes (*DateField-Extractor* and *TimeField-Extractor*) isolate a particular field, like month or minutes, in a *DateTime* value. Finally, the *Extract Time Window* node filters out the rows that do not fall into a pre-defined time window.

GroupBy node

The *GroupBy* node is a very powerful node; not only because it groups the data rows that have the same values in the grouping columns, but especially because it aggregates each group of data rows into only one row by means of an aggregation method. The power of the *GroupBy* node lies mainly in the high number of aggregation methods available.

A set of basic aggregation methods is available for String columns. This set is expanded with additional aggregation methods specific for Integer/Double cells and DateTime cells.

Basic aggregation methods for String columns

(unique, value) Count	-> number of (unique, not missing) values in each group
(unique) Concatenate	-> String with concatenated (unique) values of each group
List / Set	-> like Concatenate, but it outputs a ListCell (SetCell)
Mode	-> most frequently occurring value in each group
First/Last	-> first (last) occurence in each group

Additional aggregation methods for Integer/Double columns

Mean/Stand. Dev./Median/Variance	-> stat values of each group
Range	-> range covered by the column values in each group
Sum	-> sum of the column values for each group

Additional aggregation methods for DateTime columns

Date Range (day, ms)	-> range in terms of days or ms covered in each group