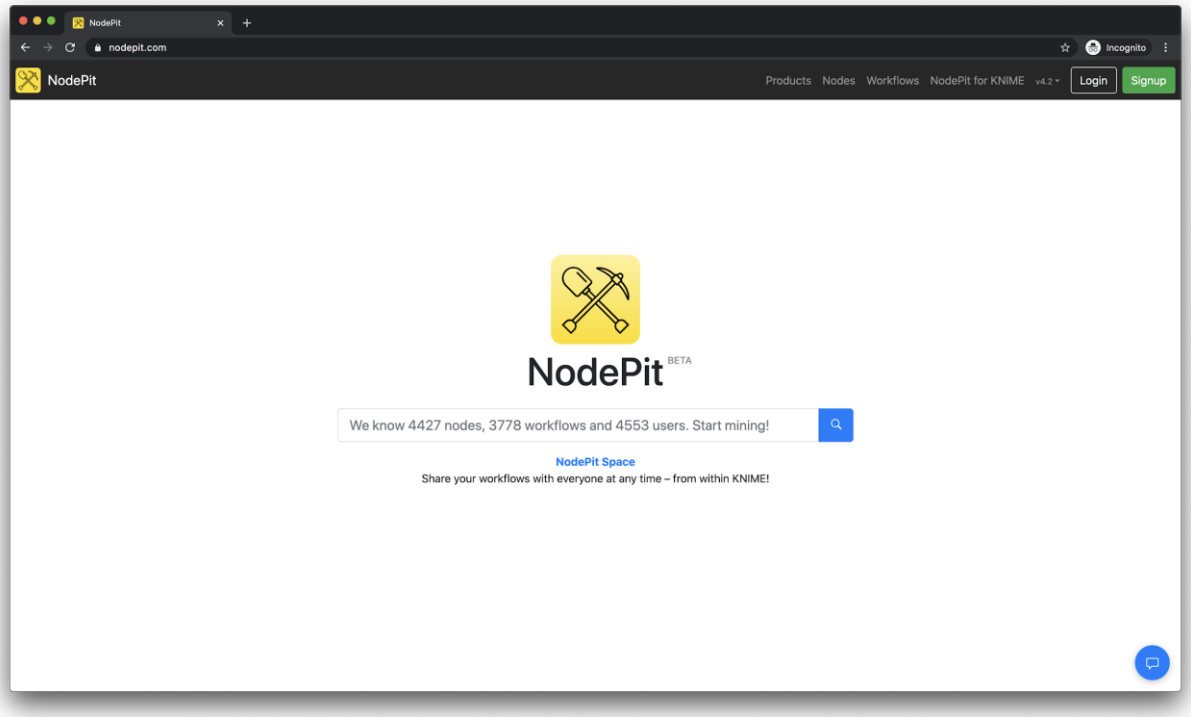
Easy management of cloud resources for searching and exploring ETL functionality

- by Philipp Katz

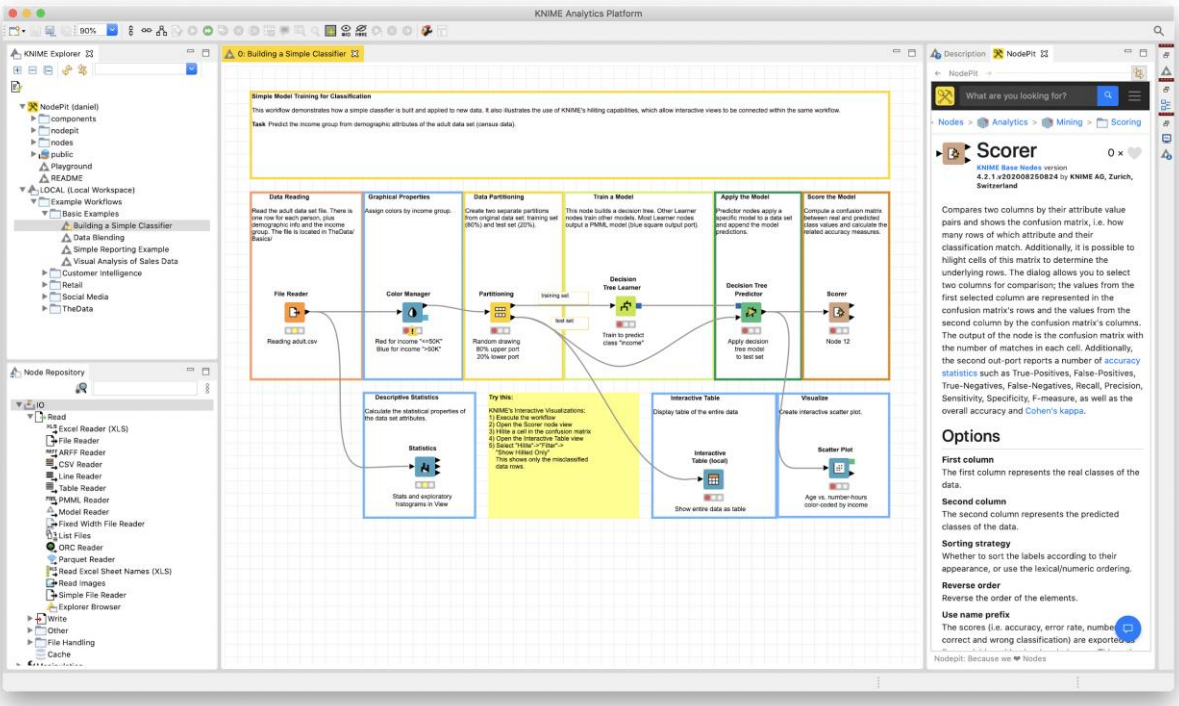
NodePit

NodePit loves ETL! It's one of the largest platforms for searching and exploring ETL (extract, transform, load) functionality, and examples for commercial and free data science portals. With over 4,500 registered users, NodePit caters for one of the largest communities of ETL enthusiasts and provides them first-class access to valuable resources and guidance.



How did we start

NodePit started as a community-driven, free and independent search engine, solely living within the KNIME universe initially. KNIME is a free and partially open-source data analytics platform that allows users to easily create data pipelines (called workflows) by dragging and dropping small modules (called nodes) on a worksheet and connecting them to solve a larger task.



We – a bunch of people from Dresden working with data in different commercial and scientific contexts – had been using KNIME, beside other tools, for our data wrangling, extraction and machine learning tasks for many years. So, we know the tool's strengths and weaknesses from many years of first-hand experience, and thus consider ourselves as independent users, developers, contributors, and consultants, but not evangelists. For almost ten years we have built several free and commercially successful extensions for KNIME, ranging from custom-tailored to publicly available ones, such as the Selenium and Palladian Nodes for Web data mining and automatization, the QR Code / Barcode Nodes, and the Elasticsearch Nodes.

Back in 2017, when we started "NodePit", the sheer amount of available nodes (>= 4,000) made it – especially for starters – very hard to figure out what functionalities existed and how to use them to solve their specific problems. It's hard to believe, but at the time, the nodes' documentation was not visible on the Web. Instead, it was deeply buried into the arcanes of the Eclipse-based KNIME desktop application, and you needed to install each extension before you could even view its documentation to top it off!

Fast-forward a few years, KNIME AG has now built a strikingly similar, separate platform for discovering nodes and workflows called "the Hub". We strongly believe that having a choice is a great advantage for

the community. Meanwhile, the steadily growing NodePit usage rate has strengthened our sentiment in offering a platform independently from the KNIME AG which advaned users have been longing for.

Beyond that, we are willing to keep pushing incremental and disruptive innovations into the KNIME universe. Our efforts have so far focused on discovering and sharing nodes and workflows, but this is just the beginning. We have learnt from our own experience and countless feedback from our users that, with the options currently available, executing and integrating these workflows in big or small enterprise environments is still considered a big technical painpoint; it causes infrastructure friction and financial hardship. Thus, this gaping hole feeds huge opportunities for the new NodePit product which we are currently building.

Let's get technical

Now, let's take a step back, put on the nerd glasses and see why Cloud&Heat is a perfect fit for our current and future infrastructure needs:

We fully rely on Cloud&Heat for hosting our services. The size and growth of the community and its spread over the whole world requires a hosting provider that is able to guarantee high availability and easy management of cloud resources. Cloud&Heat is a natural match and strengthens the bonds between Dresden-based companies and projects.

We've always been following the "keep it simple" credo when building NodePit. This also reflects on our stack – we avoid adding unnecessary complexity or following any "cargo cults". In line with this credo, the following description might seem unspectacular, but it has proven rock-solid, hassle-free to maintain for us, and easy to understand for new team members.

Our stack consists of a small amount of Docker services. Also, we mostly use NodeJS for our server-side code written in TypeScript combined with some shell scripting magic. The "Web-facing" part uses the Express web framework and all the pages are rendered on the server from Pug templates. We deliberately avoid client-side JavaScript, as KNIME is often used in "enterprise" contexts with archaic browser versions.

NodePit automatically crawls new information from many sources – our crawler runs in a separate Docker container which upon automated execution creates a KNIME environment, installs extensions, downloads workflows and then runs the indexation tasks which are eventually sent to the Web container's REST API.

For persistence, we now exclusively use MongoDB. When we started the project, we decided for Elasticsearch as an index (which could basically be ditched and re-created any time by the crawlers), then later added MongoDB for persistent, user-contributed data. Realizing that our requirements for a full-text search are rather modest, we removed Elasticsearch from our stack, and we now rely on MongoDB text indices.

Our services sit behind the "jwilder/nginx-proxy" which itself runs as a Docker container. It allows the effortless configuration of subdomains for different services and automatically generates the nginx configuration, while custom settings, e.g. for rate-limiting, can be easily added any time. Paired with the "letsencrypt-nginx-proxy-companion", we can automatically generate Let's Encrypt SSL certificates without much ado.

For deploying new code to production, we merge into a Git release branch (hosting our code at GitLab) and we build a deployable Docker image. We run the "watchtower" service on our servers, which continuously checks for new image versions, pulls them, and restarts the corresponding service. Dead simple.

The service zoo is completed by the obligatory backup tools and a headless Chrome instance, which we use for rasterizing SVG images.

The simplicity of this approach comes with great additional benefits for our customers and users. As the stack is rather simple and the maintenance efforts are quite low, we are also able to provide the NodePit experience as a self-hosted package to all kinds of enterprises from various business sectors such as pharmaceuticals, automobile or IT, and more. This allows companies to host and maintain their own installation of "NodePit Enterprise" in a private and highly secure environment behind their own firewall.

The road ahead and how Cloud&Heat can support

With our new product strategy, we're transforming NodePit from a purely informative and collaborative tool into a fully-fledged execution and integration platform. As a result, our requirements will naturally increase and we will need a more flexible approach to managing and provisioning additional computing capacity. Cloud&Heat's Managed Kubernetes Service is a perfect fit to allow a smooth scaling of resources when we launch NodePit's exciting new features.