

The KBase Platform

Shane Canon*¹ (scanon@lbl.gov), Adam P. Arkin¹, Chris Henry², Bob Cottingham³ and the KBase Team at the following institutions

¹Lawrence Berkeley National Laboratory, Berkeley, CA; ²Argonne National Laboratory, Argonne, IL; ³Oak Ridge National Laboratory, Oak Ridge, TN; ⁴Brookhaven National Laboratory, Upton, NY; ⁵ Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

<http://kbase.us>

Project Goals: The DOE Systems Biology Knowledgebase (KBase) is a free, open-source software and data platform that enables researchers to collaboratively generate, test, compare, and share hypotheses about biological functions; analyze their own data along with public and collaborator data; and combine experimental evidence and conclusions to model plant and microbial physiology and community dynamics. KBase's analytical capabilities currently include (meta)genome assembly, annotation, comparative genomics, transcriptomics, and metabolic modeling. Its web-based user interface supports building, sharing, and publishing reproducible, annotated analysis workflows with integrated data. Additionally, KBase has a software development kit that enables the community to add functionality to the system.

In order to deliver on KBase's ambitious goals, the project has developed a robust, open-source, highly-extensible platform that leverages various cutting edge technologies. This service-oriented platform consists of several interconnected subsystems that handle authentication, data storage and access, job execution, user interaction, and SDK services. These services provide the key functionality to enable reproducibility and sharing, and provide the foundation for supporting knowledge propagation. In this poster we will describe the overall architecture and some of its innovative features.

The KBase platform consists of a collection of interacting distributed services backed by multiple databases, storage components, and computational resources. We divide the system into core services that provide the low-level infrastructure and software development kit (SDK) modules that capture the specific science functionality and data models. The core services include components such as low-level data services, SDK support services, execution engine, user interface services, and (in progress) knowledge services. Figure 1 shows the schematic layout of the architecture.

Two key innovations of KBase are the Narrative Interface and the SDK. The Narrative Interface builds on top of the popular Jupyter Notebook platform and provides the key user interface for KBase. KBase's enhancements allow users to easily interact with data, perform drag-and-drop analysis on their data, and share and publish results. Another key innovation is the KBase SDK which enables external developers to easily add functionality to KBase by adding new applications or services that are integrated into the platform. The KBase SDK heavily leverages

the Docker Container technology to allow these extensions to be compartmentalized and reliably executed.

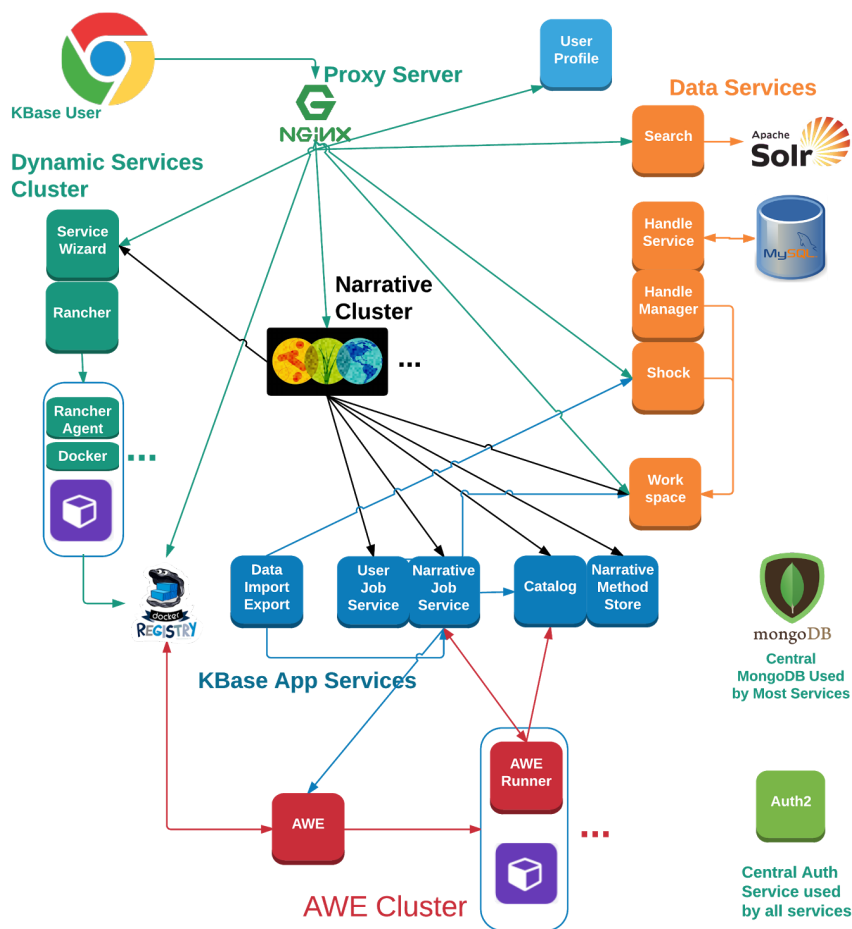


Figure 1: Schematic overview of the KBase platform architecture.

KBase leverages HPC resources provided by the DOE ASCR Facilities like NERSC and the Leadership Computing Facilities at Oak Ridge and Argonne. These resources open the door to running large-scale analysis and provide additional capacity to complement the limited dedicated compute resources in KBase. To enable this, the KBase execution engine and SDK have been extended to offer basic support for HPC. These enhancements enable SDK developers to create optimized applications that can run at large scales by KBase users to tackle large scale assembly, comparative analysis and other grand challenge problems.

KBase is funded by the Genomic Science program within the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under award numbers DE-AC02-05CH11231, DE-AC02-06CH11357, DE-AC05-00OR22725, and DE-AC02-98CH10886.