

Docs » Foundation » The Main Dataiku DSS Concepts

# The Main Dataiku DSS Concepts

#### Data Stack

Dataiku DSS is an on-premises/on-cloud software product (*not* SaaS) that operates as part of your data stack's existing infrastructure.

A data stack typically includes development, production, and deployment environments; in order to work across these environments, a separate instance of Dataiku DSS is installed in each environment.

#### Instances

A Dataiku DSS instance is an installation of the product to serve the needs of a particular environment:

- The Design node instance, in the development environment, is used to create the pipelines that turn data into outputs [Dashboards/Reports, Data (to build reports), Models]
- The Automation node instance, in the production environment, puts pipelines from the Design node into production to turn your enterprise data into the final outputs
- The API node instance, in the deployment environment, makes model outputs from the Automation node available for use in real-time scoring

Pipelines in the Design and Automation nodes are organized into projects, which can be accessed from the main page after logging in to the Dataiku DSS instance.



## Projects

A Dataiku DSS project is a container for all your work on a particular activity. The project home acts as the command center from which you can see the overall status of a project, view recent activity, and collaborate through comments, tags, and a project to-do list.

🖊 Haiku T-Shirts 🐤 🛈	ê •••				III 😗 🛹 (	🔥 વ
•			Summary Ac	tivity Changes	Metrics	TIONS -
Haiku T-Shir Create a short description (appear Create a short description (appear	ts s on homepage)		● WATCH ▼ 1 ★ STAR	0 WEDNESS	TIMELINE AV, 9 MAY Heremy edited model High revenue prediction	11:01
Automation No active scenarios yet. Activate one from the list 5/8	12/8	19/8 26/8	2/9	WEDNESI A	ou tagged project ✔ Halku T-Shirts	08:23
Flow Flow Flow Flow Flow Flow Flow Flow Flow Flow	Lab	Dashboards W	liki Tasks Constructe Construction Constructico Construction Construction Construction Construction Construc		Y, 21 DECEMBER You edited metadata on project ▲ Haiku T-Shirts	13:00
◆ 1 MODEL	1 ANALYSIS				ou edited recipe compute_Pivot_by_Country ou created recipe	12:47

#### Flow, Datasets, and Recipes

A Dataiku DSS **dataset** is a tabular view into your data that allows you to access, visualize and write data in the same way, regardless of the underlying storage system. You can connect to a variety of storage systems (file system, SQL database, Hadoop, etc), and file formats (CSV, JSON, Hadoop file formats, etc).

Creating your first DSS dataset and learning how to cleanse it is the subject of the Basics Tutorial.

A Dataiku DSS **recipe** is a set of actions to perform on one or more input datasets, resulting in one or more output datasets. Each time you prepare, join, group... your datasets, this will be through a recipe. A recipe can be visual or code.

- A visual recipe allows a quick and interactive transformation of the input dataset through a number of prepackaged operations available in a visual interface.
- A code recipe allows a user with coding skills to go beyond visual recipe functionality to take control of the transformation using any supported language (SQL, Python, R, etc).

Dataiku allows "coders" and "clickers" to seamlessly collaborate on the same project through code and visual recipes.

The lineage of a dataset (or a model) is thus defined by the inputs and outputs of its ancestor recipes. The Flow is a visual representation of your work as a set of dependencies between datasets and the recipes used to produce them.



The knowledge of these dependencies helps the Dataiku DSS engine minimize the number of data  $^3$  of  $^5$  processes to be launched when (re)building a dataset.  $^{2/14/20, 9:43 \text{ AM}}$ 

The MiaDatai Visual Amalysis iku Ac...



The Visual Analysis lab allows you to experiment with your data in a code-free environment where you can:

- Perform interactive analyses with built-in charts and data preparation (cleaning, filtering, enriching). These steps can be deployed to the Flow as Prepare recipes.
- Use machine learning algorithms (unsupervised and supervised training) to generate insights and build predictive models. These models can be deployed to the Flow.

## Lab - Code



The code lab allows you to experiment with your data in Jupyter notebooks (for Python / R) or SQL notebooks when working with SQL DB's, Hive, or Impala. You can perform interactive analysis in these notebooks and then deploy them to the Flow as code recipes.

You can also use the code lab to create interactive web apps with R Shiny, Python Bokeh, or Javascript. Templates and code samples are provided to help you get started. Web apps can be shared on dashboards.

The code lab also provides the ability to create some advanced R Markdown reports that mix text, code, and complex visualizations using Python and R. These reports can be shared on dashboards or distributed in various printable formats.

## Jobs, Scenarios, and Monitoring



Jobs are created when you build a dataset. Dataiku DSS provides a full job log to let you monitor what works and what does not, along with the ability to debug potential errors.

The Manodeds aix always to date apredictive scoring. Reports/anatemarity that complexity have be their results are shown in Monitoring.

#### Wikis



Wikis are a widely-used tool for collaboration, allowing team members to create and edit articles that document a project, a group of projects, or more. Every Dataiku DSS project is wiki-enabled.

### **Dashboard & Insights**



The Dashboard is a communication tool to organize, share, and deliver the Insights of your data project. Insights can include any Dataiku DSS object, such as charts, datasets, web apps, and reports.

