Collaborative Data Science Development Platform to Turn Raw Data Into Predictions. Quickly.

Build and Deploy. Collaboratively.

Dataiku DSS is the collaborative data science platform that enables teams to explore, prototype, build, and deliver their own data products more efficiently.

Dataiku DSS’ all-in-one analytics & data science platform includes:

- **Coding &/or Visual Interface**: use notebooks (R, Python, Hive, Pig, Spark, etc.) or a customizable drag and drop visual interface at any step of the predictive dataflow prototyping process – from wrangling, to analysis, to modeling.

- **Data Agnostic Integration**: with 30+ data connectors and the ability to extend with custom plugins, connect to your existing infrastructure, automatically detect data format and schema, and push computation to your existing SQL, Hadoop, or Spark infrastructure.

- **Prepare, Blend, Visualize**: thanks to a visual profile of the data at every step of the analysis, interactively explore, prepare, enrich, blend, and clean data using 80+ built-in functions, from simple filters to custom Python.

- **Machine Learning**: leverage ML technologies (Scikit-Learn, MLlib, XGboost, etc.) in a visual UI, build & optimize models in Python or R, integrate any external ML library through code APIs (H2O, Dato, Skytree, etc.), and get instant visual & statistical feedback on model performance.

- **Robust Production Deployment**: bundle your whole workflow, optionally including data & models, as a single deployable and reproducible package for real-time predictions with our REST API.

- **Monitoring & Version Control**: make sure deployments run smoothly with dashboard monitoring & data validation policies (model metrics, drift, data consistency, etc.). If something’s off, easily roll back to a previous version.

“Before Dataiku DSS, we’d externalize the whole research phase that leads to concrete solutions. Our work consisted mainly of descriptive analytics on past data. With Dataiku, we’ve internalized the design & deployment of our data solutions. We can now predict the future actions of our customers with a high degree of accuracy and act accordingly.”

Damien Garzilli, Chief Data Officer, Showroomprive

www.dataiku.com
End-to-End: From Raw to Predicted

With Dataiku DSS, transforming your raw data into real-time predictions requires only one interface from beginning to end.

One Stop Shop to Explore, Wrangle, Prepare

**Step 1 – Connect**
Don’t let different formats or varying storage locations stop you from accessing the data you need on demand. With over 25 connectors and the possibility to create your own, access any and all data – structured or unstructured, big or small, internal or external – at any time.

**Step 2 – Integrate & Explore:**
In the blink of an eye, Dataiku’s Quick Columns View enables anyone to instantaneously see the quality of data at hand - from duplicates and invalids, to completeness and accuracy, to distribution and outliers. Plus, with one click, access a full statistical summary of any feature and/or column.
Step 3 – Prepare & Enrich:
Preparing data for advanced analytics and predictive modeling should not take up 80% of anyone’s time. In Dataiku DSS, data prep is quick. A spreadsheet type interface paired with automatically suggested contextual transformations makes performing mass actions on data a painless process for anyone involved.

Plus, filter and search your data as you would in Excel and use over 90 built-in visual processors for code-free data wrangling. You can also create business-specific tailored transformation types or use formulas and Python scripts for custom processing.

Step 4 – Share Insights & Visualizations:
Engaging business stakeholders or other members of the team is essential when it comes to proving value or sharing insights.

- **Dashboards:** Dataiku DSS Dashboards make it easy to create interactive charts and visualizations from datasets. With 25 built-in chart formats – from histograms, to boxplots, to maps – drag and drop your data to automatically compute charts on your existing Big Data infrastructure (SQL or Impala) for optimal performance.

- **Web-apps:** create pixel-perfect, HTML, Javascript (d3.js, Leaflet, plot.ly, …), and Python based web applications that you can keep secure with API keys management.
Machine Learning: Unleashed, Accessible, and Scalable

Dataiku DSS enables anyone to apply analytics algorithms to data in order to uncover patterns and predict future trends and activities. Whether the project at hand necessitates clustering or predicting, users can leverage existing libraries in the visual interface, use custom code directly in R, Python, and SQL, or even integrate any external ML library through code APIs. Thanks to specialized pre-processing, users can leverage features of all kinds (numerical, categorical, textual) to improve model performance. Plus, instant visual feedback makes testing, assessing, and comparing several models’ performances a quick and painless process.

Supervised Machine Learning: Prediction
Dataiku DSS supports different types of modeling to predict three types of target variables:
- **Regression** for a numerical target
- **Two-way Classification** for a target than can be one of two categories
- **Multi-class Classification** for a target that can be one of many categories

With 10+ algorithms to pick from in the visual interface – from Ridge Regression, to Random Forest, to Stochastic Gradient Descent – or by coding your custom model, understand what drives your business and predict what’s coming next.

Unsupervised Machine Learning: Clustering
Unsupervised machine learning in Dataiku DSS easily lets anyone understand the structure of the data at hand. Through a simple point and click interface, let the algorithms (K-means, Ward Hierarchical or Spectral Clustering, DBSCAN, …) detect groups within your data based on specified dimensions.

With sampling (when data at hand is too large to run on RAM), dimensionality reduction (to reduce the number of variables into “principal components”), and automatic outlier detection (to avoid skewed clusters), quickly discover hidden patterns in data of any size or format.
Machine Learning with Spark: MLib
Are your datasets too large to fit in memory? No problem. Dataiku DSS comes complete with MLib, Spark’s machine learning library. Using either the visual interface or code-based analysis, easily fine-tune your model, deploy it in the Dataiku Workflow as a re-trainable model, and apply it in a scoring recipe to perform predictions on unlabeled datasets. Mining at scale has never been so effective.

From Input to Output in One Flow

Dataiku DSS Workflow
The Dataiku DSS Workflow is where users can navigate datasets and recipes that make up the process from raw to predicted. The Workflow also enables users to define dependencies between datasets, build custom automation scenarios with advanced triggering policies and notifications, or use partitioning to manage incremental data. With the Workflow, teams can assemble code and visual components all together as an end-to-end pipeline.

Bye Bye Data Plumbing, Hello Automation
But that’s not all. Your days of data plumbing are over. Dataiku DSS will automatically manage storage formats, schemas, and data movement for you. Thanks to execution scenarios, rebuilding data and retraining models based on conditions (data availability, schedule, etc.) can be automatic. Plus, users can dig into each individual scenario to inspect failure causes and execution timing.
**All-in-One: From Design to Production**

Dataiku DSS is the world’s first integrated visual and code-based environment for the design and production of data science applications, batch or real-time.

**Development & Production, United at Last**

**Design & Iterate**
Quickly create your predictive models and the associated workflows by combining visual components and programming languages in a common environment. For clickers: acquire, prepare, filter, join, copy your data in the drag and drop visual interface. For coders: use your favorite (big data) programming languages to add arbitrary custom logic.

**Deploy & Run in Hadoop, Spark, or in Database**
Deploy your predictive applications and data products in production using advanced automation of workflows and expose your machine learning models via APIs. Plus, easily set up model as well as workflow monitoring and version control.

**Real-time Scoring**
Deploy your model to production to get real-time predictions with Dataiku DSS’ REST API; handle large quantities of real-time predictions with queuing, parallelism, and load balancing. Plus, avoid model drift with a feedback loop and choose to run multiple versions of the same model for automated A/B testing.
Who Said Coders & Clickers Can’t Get Along?

Click Like There’s No Tomorrow
Whether to prepare your data, do feature engineering, generate charts and other data visualizations, or to build models, Dataiku DSS’s visual interface is great for the less technical members of the team – or for those who are just beginning to learn the machine learning process.

Code Like a 1st Class Citizen
Does a GUI feel like an extra barrier between you and the data? If so, Dataiku DSS lets you integrate your favorite – and most effective – programming languages (Python, R, Shell, SQL) into a project. You can even code over Hadoop using MapReduce (Hive, Pig), Impala, or Spark (PySpark, SparkR, SparkSQL). Finally, develop and automate your scripts interactively using Jupyter or SQL Notebooks.

For Clickers & Coders

Quickly create predictive models and the associated workflows by combining visual components and programming languages in a common environment.

For Clickers
Acquire, prepare, filter, join, copy data with visual components…

For Coders
Use your favorite (big data) programming languages to add arbitrary custom logic…

www.dataiku.com
For the Whole Data Team: From Beginner to Expert

By focusing on the needs of the people behind data-driven initiatives – from the data analysts and scientists who build the prototypes to the IT teams that implement and deploy them – Dataiku DSS was designed for data teams from the ground up. Collaboration features make it easy to share knowledge amongst team members and to onboard new users quickly. After all, Dataiku DSS is designed for:

- **Data Analysts** who are more efficient in an interactive visual interface where they can point, click, and build or use languages like SQL to wrangle data, model, easily re-run workflows, visualize results, and get up-to-date insights on demand;

- **Data Scientists and Developers** who need to draft data preparation & modeling in seconds, that wish to leverage their favorite ML libraries (Scikit-Learn, R, MLlib, H2O, and so on), and that rely on automating their work in a completely customizable interface;

- **DataOps & Data Engineers** that coordinate development and operations by handling workflow automations, creating predictive web services, monitoring data & model health on a daily basis, and who don’t want to worry about multi-technology platforms.

- **Team Leads & Managers** who need to ensure team productivity in order to deliver quality data products on time. Managing growing teams made of data scientists, analysts, and engineers with favorite tools and different levels of expertise is not an easy task. With Dataiku DSS, managers can easily track a project’s evolution all the while ensuring optimal productivity by enabling individual team members to work together using the tools and technologies they know best.

Complete with integrated documentation, knowledge sharing features, change management, version control, and team activity monitoring, Dataiku DSS enables teams to collaborate effectively throughout the design and production phases of data science projects.

“Dataiku DSS is a tool for all members of the data team – from data engineer to data analyst. With the help of Dataiku DSS, we have created a structure that allows us to empower teams of analysts around the world. One of our primary goals was that iterations no longer be dependent on technicians. With Dataiku DSS, we’ve succeeded.”

Gaelle Periat, BI Manager - Blablacar

www.dataiku.com
About Dataiku

Dataiku strives to be the preferred software solution in helping organizations succeed in the world's rapidly evolving data-driven business ecosystem. Guided by the belief that true innovation comes from the effective combination of diversity of cultures, of mindsets, and of technologies, Dataiku's purpose is to enable all enterprises to imagine and deliver the data innovations of tomorrow.

• Because 90% of the innovations to come in the next 10 years will be made possible by the effective use of data.
• Because faced with the complexity and volume of today’s ever evolving data ecosystem, a one-man approach is an unrealistic bet for success.
• Because the successful reconciliation of business and tech profiles will undoubtedly unleash tomorrow's greatest data-driven advances.

With this in mind, Dataiku develops the unique advanced analytics software solution that enables companies to build and deliver their own data products more efficiently. Thanks to a collaborative and team-based user interface for data scientists and beginner analysts, to a unified framework for both development and deployment of data projects, and to immediate access to all the features and tools required to design data products from scratch, users can easily apply machine learning and data science techniques to all types, sizes, and formats of raw data to build and deploy predictive data flows.

To learn more, please visit our website: [www.dataiku.com](http://www.dataiku.com)