DATASHEET

data
iku

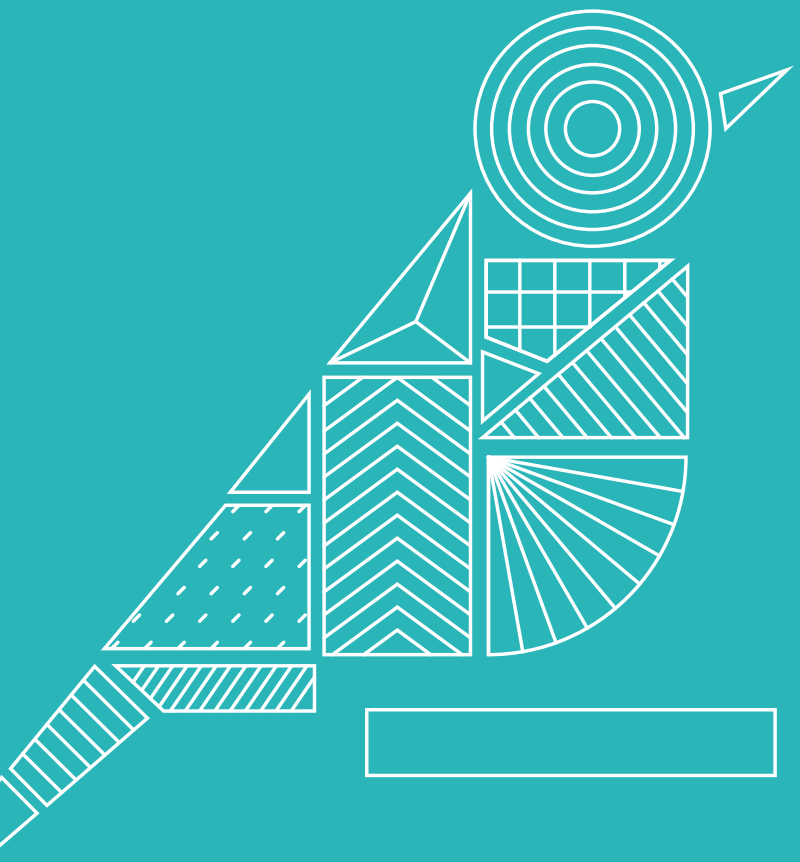
data iku

DATASHEET

Your Path to Enterprise AI

To succeed in the world's rapidly evolving ecosystem, companies (no matter what their industry or size) must use data to continuously develop more innovative operations, processes, and products. This means embracing the shift to Enterprise AI, using the power of machine learning to enhance - not replace - humans.

Dataiku is the centralized data platform that moves businesses along their data journey from analytics at scale to Enterprise AI, powering self-service analytics while also ensuring the operationalization of machine learning models in production.



Key Features

- Seamless connectivity to any data, no matter where it's stored or in what format.
- Faster data cleaning, wrangling, mining, and visualization.
- The latest in machine learning technology (including AutoML and deep learning) all in one place and ready to be operationalized with automation environments, scenarios, and advanced monitoring
- Every step in the data-to-insights process can be done in code or with a visual interface.
- Enterprise-level security with fine-grained access rights.

data iku

DATASHEET

Feature Overview

A Centralized, Controlled Environment

- Connect to existing data storage systems and leverage plugins and connectors for access to all data from one, central location.
- Maintain enterprise-level security with data governance features like documentation, projects, task organization, change management, rollback, monitoring, etc.
- Reuse and automate work (via data transformations, code snippets, centralized best practices, automated assistants, and more), to go from raw data to insights faster.

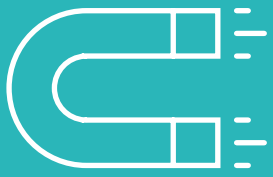
A Common Ground for Experts and Explorers

- Built from the ground up for data scientists, analysts, IT/data engineers, and data team managers.
- From code-free data wrangling using built-in visual processors to the visual interface for machine learning, non-technical users aren't left in the dust when it comes to contributing to machine learning projects.
- But with interactive Python, R, and SQL notebooks, coders get what they want too.

A Shortcut to ML/AI Operationalization

- Centrally manage models and update them from one location, integrating with an API without having to modify or inject anything into existing applications.
- Prevent model drift with the ability to easily monitor performance and easily make necessary adjustments.

data iku



DATASHEET

Connectivity

Dataiku allows you to seamlessly connect to your data no matter where it's stored or in what format. That means easy access for everyone - whether technical or not - to the data they need.

SQL Databases

- ☑ MySQL
- ☑ PostgreSQL
- ☑ Vertica
- ☑ Amazon Redshift
- ☑ Pivotal Greenplum
- ☑ Teradata
- ☑ IBM Netezza
- ☑ SAP HANA
- ☑ Oracle
- ☑ Microsoft SQL Server (incl. SQL DW)
- ☑ Google BigQuery
- ☑ IBM DB2
- ☑ Exasol
- ☑ MemSQL
- ☑ Snowflake
- ☑ Custom connectivity through JDBC

NoSQL Databases

- ☑ MongoDB
- ☑ Cassandra
- ☑ ElasticSearch

Hadoop & Spark Supported Distributions

- ☑ Cloudera
- ☑ Hortonworks
- ☑ MapR
- ☑ Amazon EMR

Hadoop File Formats

- ☑ CSV
- ☑ Parquet
- ☑ ORC
- ☑ SequenceFile
- ☑ RCFile

Remote Data Sources

- ☑ FTP
- ☑ SCP
- ☑ SFTP
- ☑ HTTP

Cloud Object Storage

- ☑ Amazon S3
- ☑ Google Cloud Storage
- ☑ Azure Blob Storage
- ☑ Azure Data Lake Store

Custom Data Sources - extended connectivity through Dataiku Plugins

- ☑ Connect to REST APIs,
- ☑ Create custom file formats
- ☑ Connect to databases

data iku

DATASHEET

Exploratory Analytics

Sometimes you need to do a deep dive on your data, but other times, it's important to understand it at a glance. From exploring available datasets to dashboarding, Dataiku makes this type of analysis easy.

Data Analysis

- Automatically detect dataset schema and data types
- Assign semantic meanings to your datasets columns
- Build univariate statistics automatically & derive data quality checks
- Dataset audit
  - ☑ Automatically produce data quality and statistical analysis of entire Dataiku datasets
  - ☑ Support of several backends for audit (in-memory, Spark, SQL)
- Leverage predefined Python-based Jupyter Notebooks
  - ☑ Univariate analysis and statistical tests on a single population.
  - ☑ Statistics and tests on multiple populations
  - ☑ Correlations analysis
  - ☑ Principal Components Analysis
  - ☑ High dimensional data visualization (t-SNE)
  - ☑ Topic modeling
  - ☑ Time-Series analytics
  - ☑ Time series forecasting

Data Cataloging

- Search for data, comments, features, or models in a centralized catalog.
- Explore data from all your existing connections

Data Visualization

- Create standard charts (histogram, bar charts,etc). and scale charts' computation by leveraging underlying systems (in-database aggregations)
- Create custom charts using using:
  - ☑ Custom Python-based or R-based Charts
  - ☑ Custom Web Application (HTML/JS/CSS/Flask)
  - ☑ Shiny Web Application ®
  - ☑ Bokeh Web Application (Python)

Dashboarding

- User-managed reports and dashboards
  - ☑ RMarkdown reports
  - ☑ Jupyter Notebooks reports
  - ☑ Custom Insights (GGplot, Plotly, Matplotlib)
  - ☑ Custom interactive, web-based visualisations

data iku

DATASHEET

Data Preparation

Traditionally, data preparation takes up to 80 percent of the time of a data project. But Dataiku's data prep features makes that process faster and easier, which means more time for more impactful (and creative) work.

Visual Data Transformation

- Design your data transformation jobs using a point-and-click interface
  - ☑ Group
  - ☑ Filter
  - ☑ Sort
  - ☑ Stack
  - ☑ Join
  - ☑ Window
  - ☑ Sync
  - ☑ Distinct
  - ☑ Top-N
  - ☑ Pivot
  - ☑ Split
- Scale your transformations by running them directly in distributed computations systems (SQL, Hive, Spark, Impala)
- See and tune the underlying code generated for the task

Dataset Sampling

- First records, random selection, stratified sampling, etc.

Interactive Data Preparation

- Processors (80 built-in from simple text processing to custom Python- or formula-based transformations
- Automatically turn data preparation scripts into Spark or MapReduce jobs

data iku

DATASHEET

Machine Learning

Dataiku offers the latest machine learning technologies all in one place so that data scientists can focus on what they do best: building and optimizing the right model for the use case at hand.

Automated Machine Learning (AutoML)

Automated ML strategies

- ☑ Quick prototypes
- ☑ Interpretable models
- ☑ High performance

Features handling for machine learning

- ☑ Support for numerical, categorical, text and vector features
- ☑ Automatic preprocessing of categorical features (Dummy encoding, impact coding, hashing, custom preprocessing, etc.)
- ☑ Automatic preprocessing of numerical features (Standard scaling, quantile-based binning, custom preprocessing, etc.)
- ☑ Automatic preprocessing of text features (TF/IDF, Hashing trick, Truncated SVD, Custom preprocessing)
- ☑ Various missing values imputation strategies
  - + Features generation
    - ◊ *Feature-per-feature derived variables (square, square root…)*
    - ◊ *Linear and polynomial combinations*
  - + Features selection
    - ◊ *Filter and embedded methods*

Choose between several ML backends to train your models

- ☑ Scikit-learn
- ☑ XGBoost
- ☑ MLLib
- ☑ H20

Algorithms

- ☑ Python-based
  - + Ordinary Least Squares
  - + Ridge Regression
  - + Lasso Regression
  - + Logistic regression
  - + Random Forests
  - + Gradient Boosted Trees
  - + XGBoost
  - + Decision Tree
  - + Support Vector Machine
  - + Stochastic Gradient Descent
  - + K Nearest Neighbors
  - + Extra Random Trees
  - + Artificial Neural Network
  - + Lasso Path
  - + Custom Models offering scikit-learn compatible API's (ex: LightGBM)
- ☑ Spark MLLib-based
  - + Logistic Regression
  - + Linear Regression
  - + Decision Trees
  - + Random Forest
  - + Gradient Boosted Trees
  - + Naive Bayes
  - + Custom models
- ☑ H20-based
  - + Deep Learning
  - + GBM
  - + GLM
  - + Random Forest
  - + Naive Bayes

data iku

DATASHEET

Machine Learning

Dataiku offers the latest machine learning technologies all in one place so that data scientists can focus on what they do best: building and optimizing the right model for the use case at hand.

Automated Machine Learning (AutoML)

Hyperparameters optimisation

- ☑ Freely set and search hyperparameters
- ☑ Cross validation strategies
  - + Support for several Train/test splitting policies (incl. custom)
  - + K-Fold cross testing
  - + Optimize model tuning on several metrics (Explained Variance Score, MAPE, MAE, MSE, Accuracy, F1 Score, Cost matrix, AUC, etc.)
- ☑ Interrupt and resume grid search
- ☑ Visualize grid search results

Analyzing model training results

- ☑ Get insights from your model
  - + Scored data
  - + Features importance
  - + Model parameters
  - + Partial dependencies plots
  - + Regression coefficients
- ☑ Publish training results to Dataiku Dashboards
- ☑ Audit model performances
  - + Confusion matrix
  - + Decision chart
  - + Lift chart
  - + ROC curve
  - + Probabilities distribution chart
  - + Detailed Metrics (Accuracy, F1 Score, ROC-AUC Score,MAE, RMSE, etc.)

Automatically create ensemble from several models

- ☑ Linear stacking (for regression models) or logistic stacking (for classification problems)
- ☑ Prediction averaging or median (for regression problems)
- ☑ Majority voting (for classification problems)

Scoring Engines

- ☑ Local, in-memory
- ☑ Optimized scoring
- ☑ Spark
- ☑ SQL (in-database scoring)

data iku

DATASHEET

Machine Learning

Dataiku offers the latest machine learning technologies all in one place so that data scientists can focus on what they do best: building and optimizing the right model for the use case at hand.

Model Deployment

- Model versioning
- Batch scoring
- Real-time scoring
  - ☑ Expose your models through REST API's for real-time scoring by other applications
- Expose arbitrary functions and models through REST API's
  - ☑ Write custom R, Python or SQL based functions or models
  - ☑ Automatically turn them into API endpoints for operationalisation
- Easily manage all your model deployments
  - ☑ One-click deployment of model
- Docker & Kubernetes
  - ☑ Deploy models into Docker containers for operationalisation
  - ☑ Automatically push images to Kubernetes clusters for high scalability
- Model monitoring mechanism
  - ☑ Control model performances over time
  - ☑ Automatically retrain models in case of performance drift
  - ☑ Customize your retraining strategies
- Logging
  - ☑ Log and audit all queries sent to your models

Deep Learning

- Support for Keras with Tensorflow backend
- User-defined model architecture
- Personalize training settings
- Support for multiple inputs for your models
- Support for CPU and GPU
- Support pre-trained models
- Extract features from images
- Tensorboard integration

data iku

DATASHEET

Machine Learning

Dataiku offers the latest machine learning technologies all in one place so that data scientists can focus on what they do best: building and optimizing the right model for the use case at hand.

Unsupervised Learning

- Automated features engineering (similar to Supervised learning)
- Optional dimensionality reduction
- Outliers detection
- Algorithms
  - ☑ K-means
  - ☑ Gaussian Mixture
  - ☑ Agglomerative Clustering
  - ☑ Spectral Clustering
  - ☑ DBSCAN
  - ☑ Interactive Clustering (Two-step clustering)
  - ☑ Isolation Forest (Anomaly Detection)
  - ☑ Custom Models

Scale Your Model Trainings

- Train models over Kubernetes

data iku

DATASHEET

Automation Features

When it comes to streamlining and automating workflows, Dataiku allows data teams to put the right processes in place to ensure models are properly monitored and easily managed in production.

Data Flow

- ☑ Keep track of the dependencies between your datasets
- ☑ Manage the complete data lineage
- ☑ Check consistency of data, schemas or data types

Partitioning

- ☑ Leverage HDFS or SQL partitioning mechanisms to optimize computation time

Metrics & Checks

- ☑ Create Metrics assessing data consistency and quality
- ☑ Adapt the behavior of your data pipelines and jobs based on Checks against these Metrics
- ☑ Leverage Metrics and Checks to measure potential ML models drift over time

Monitoring

- ☑ Track the status of your production scenarios
- ☑ Visualize the success and errors of your Dataiku jobs

Scenarios

- ☑ Trigger the execution of your data flows and applications on a scheduled or event-driven basis
- ☑ Create complete custom execution scenarios by assembling set of actions to do (steps)
- ☑ Leverage built-in steps or define your own steps through a Python API
- ☑ Publish the results of the scenarios to various channels through Reporters (Send emails with custom templates; attach datasets, logs, files, or reports to your Reporters; send notifications to Slack or Hipchat)

Automation Environments

- ☑ Use dedicated Dataiku Automation nodes for production pipelines
- ☑ Connect and deploy on production systems (data lakes, databases)
- ☑ Activate, use or revert multiple Dataiku project bundles

data iku

DATASHEET

Code

Work in the tools and with the languages you already know - everything can be done with code and fully customized. And for tasks where it's easier to use a visual interface, Dataiku provides the freedom to switch seamlessly between the two.

Support of multiple languages for coding "Recipes"

- ☑ Python
- ☑ R
- ☑ SQL
- ☑ Shell
- ☑ Hive
- ☑ Impala
- ☑ Pig
- ☑ Spark Scala
- ☑ Spark SQL
- ☑ PySpark
- ☑ SparkR
- ☑ Sparklyr

Create and use custom code environments

- ☑ Support for multiple versions of Python (2.7, 3.4, 3.5, 3.6)
- ☑ Support for Conda
- ☑ Install R and Python libraries directly from Dataiku's interface
- ☑ Open environment to install any R or Python libraries
- ☑ Manage packages dependencies and create reproducible environments

Scale code execution

- ☑ Submit R and Python jobs to Kubernetes clusters transparently

Interactive Notebooks for data scientists

- ☑ Full integration of Jupyter notebooks with Python, R or PySpark kernels
- ☑ Use pre-templated Notebooks to speed up your work
- ☑ Interactively query databases or data lakes through SQL Notebooks (support for Hive)

Python & R Libraries

- ☑ Create your own R or Python libraries or helpers
- ☑ Share them within all the Dataiku instance
- ☑ Easily use your pre-existing code assets

Create reusable custom components

- ☑ Package and distribute arbitrarily complex code-based functions to less-technical users
- ☑ Extend native Dataiku capabilities through code-based Plugins (Custom connectors, custom data preparation processor, etc.)
- ☑ Create Python-based custom steps for your Dataiku scenarios

APIs

- ☑ Manage the Dataiku platform through CLI or Python SDK
- ☑ Train and deploy ML models programmatically
- ☑ Expose custom Python & R functions through REST API's

Collaboration

Dataiku was designed from the ground up with collaboration in mind. From knowledge sharing to change management to monitoring, data teams - including scientists, engineers, analysts, and more - can work faster and smarter together.

- ## Shared platform [for data scientist, data engineer, analyst, etc.]
- ## Version control
  - ☑ Git-based version control recording all changes made in Dataiku
- ## Knowledge management and sharing
  - ☑ Create Wikis to document your projects
  - ☑ Engage with other users of the platform through Discussions
  - ☑ Tag, comment and favorite any Dataiku objects
- ## Team Activity Monitoring
- ## Shared code-based components
  - ☑ Distribute reusable code snippets for all users
  - ☑ Package arbitrary complex function, operation or business logic to be used by less-technical users

data iku

DATASHEET

Governance & Security

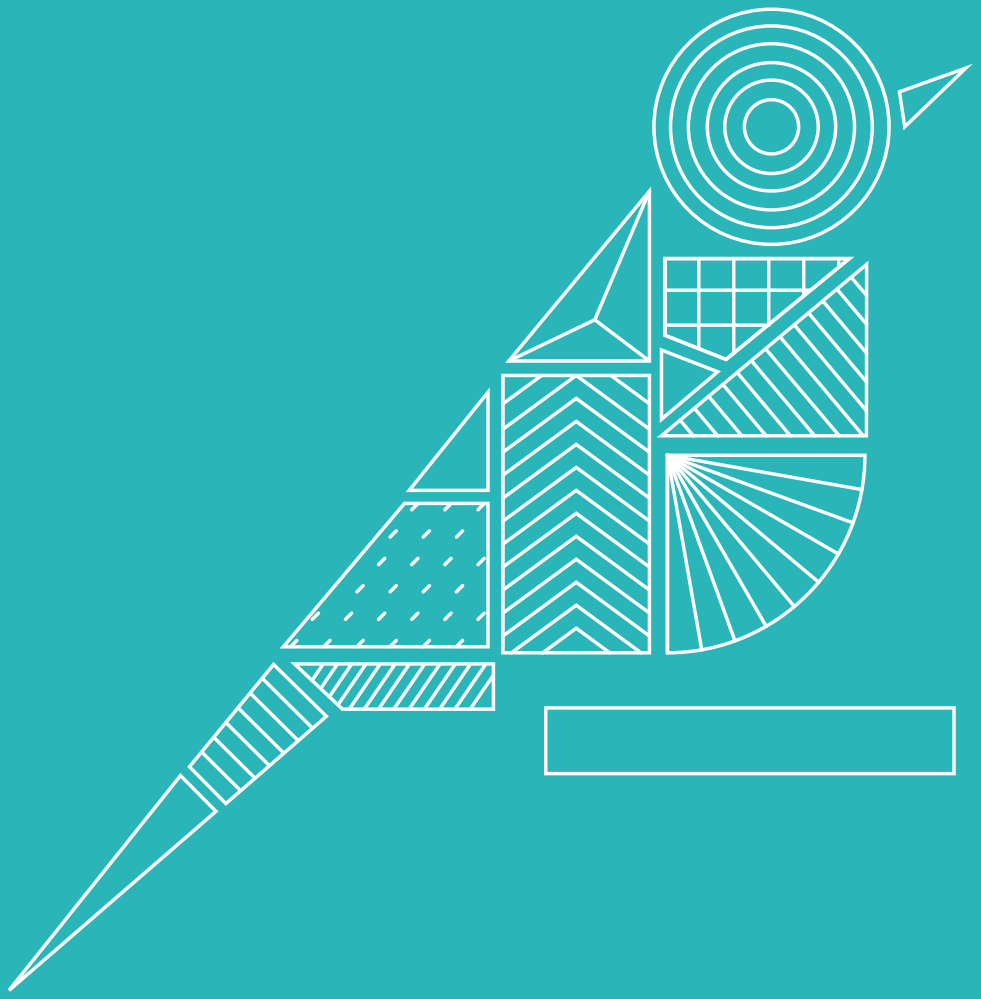
Dataiku makes data governance easy, bringing enterprise-level security with fine-grained access rights and advanced monitoring for admins or project managers.

- ## User profiles
- ## Role-based access (fine-grained or custom)
- ## Authentication management
  - ☑ Use SSO systems
  - ☑ Connect to your corporate database (LDAP, Active Directory…) to manage users and groups
- ## Enterprise-grade security
  - ☑ Track and monitor all actions in Dataiku using an audit trail
  - ☑ Authenticate against Hadoop clusters and databases through Kerberos
  - ☑ Supports users impersonation for full traceability and compliance
- ## Resources management
  - ☑ Dynamically start and stop Hadoop clusters from Dataiku
  - ☑ Control server resources allocation directly from the user interface
- ## Platform management
  - ☑ Integrate with your corporate workload management tools using Dataiku CLI and APIs

Architecture

Dataiku was built for the modern enterprise, and its architecture ensures that businesses can stay open (i.e., not tied down to a certain technology) and that they can scale their data efforts.

- No client installation for Dataiku users
- Dataiku nodes (use dedicated Dataiku environments or nodes to design, run, and deploy your ML applications)
- Integrations
  - ☑ Leverage distributed systems to scale computations through Dataiku
  - ☑ Automatically turn Dataiku jobs into SQL, Spark, MapReduce, Hive, or Impala jobs for in-cluster or in-database processing to avoid unnecessary data movements or copies
- Modern architecture (Docker, Kubernetes, GPU for deep learning)
- Traceability and debugging through full system logs
- Open platform
  - ☑ Native support of Jupyter notebooks
  - ☑ Install and manage any of your favorite Python or R packages and libraries
  - ☑ Freely reuse your existing corporate code assets
  - ☑ Extend the Dataiku platform with custom components (Plugins - see full list)



data
iku