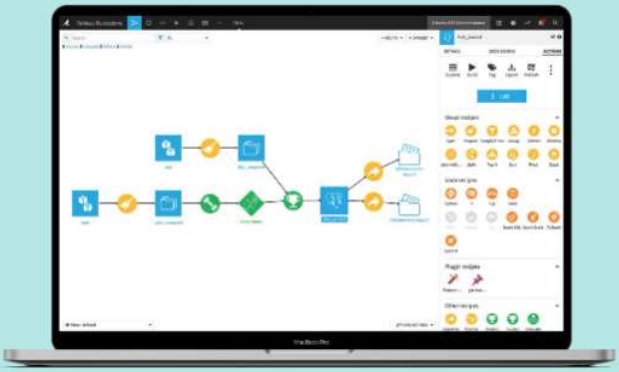
Data Sheet

2023



Summary

Everyday AI, Extraordinary People

Dataiku Data Sheet

Dataiku is the platform for Everyday AI, systemizing the use of data for exceptional business results. Organizations that use Dataiku elevate their people (whether technical and working in code or on the business side and low- or no-code) to extraordinary, arming them with the ability to make better day-to-day decisions with data.

More than 500 companies worldwide use Dataiku to systemize their use of data, analytics, and AI, driving diverse use cases from fraud detection to customer churn prevention, predictive maintenance to supply chain optimization, and everything in between.

Data Management & Preparation

Connectivity

With dozens out of the box connectors, Dataiku allows you to access data from public or private cloud as well as on prem sources without the need to create copies of the data.

SUPPORTED CONNECTIONS

→ SQL Databases

- ☑ Snowflake
- ☑ Amazon Redshift
- ☑ Azure Synapse
- ☑ Google BigQuery
- ☑ PostgreSQL
- ☑ MySQL
- ☑ Microsoft SQL Server
- ☑ Oracle
- ☑ Teradata
- ☑ Pivotal Greenplum
- ☑ Vertica
- ☑ Amazon Athena
- ☑ SAP HANA
- ☑ IBM Netezza
- ☑ Exasol
- ☑ kdb+
- ☑ Other SQL databases (through JDBC)

→ Remote data sources

- ☑ FTP
- ☑ HTTP
- ☑ SCP/SFTP

→ Other databases

- ☑ ElasticSearch
- ☑ Cassandra
- ☑ MongoDB

→ File formats

- ☑ CSV
- ☑ Excel
- ☑ Parquet
- ☑ JSON
- ☑ GeoJSON
- ☑ Shapefile
- ☑ Delta Lake
- ☑ ORC
- ☑ Avro
- ☑ XML

→ Business application data connectors

- ☑ Salesforce
- ☑ PowerBI
- ☑ Tableau
- ☑ One Drive
- ☑ Box.com
- ☑ Sharepoint
- ☑ Google Drive
- ☑ Google Sheets
- ☑ Splunk
- ☑ Airtable
- ☑ OSISoft PI
- ☑ Dropbox
- ☑ Confluence
- ☑ Stripe
- ☑ Intercom
- ☑ Jira
- ☑ SAP OData
- ☑ Zendesk
- ☑ HubSpot
- ☑ Qlik
- ☑ Microstrategy

→ Cloud object storage

- ☑ Amazon S3
- ☑ Azure Blob Storage / ADLS
- ☑ Google Cloud Storage

→ Hadoop ecosystem

- ☑ HDFS
- ☑ Hive
- ☑ Impala

→ Streaming data sources

- ☑ Kafka
- ☑ AWS SQS

Data Preparation & Transformation

Traditionally, data preparation takes up to 80% of the time of a data project. But Dataiku makes that process 10x faster and easier, which means more time for more impactful (and creative) work for everyone from analysts to data scientists.
Diverse teams can use both visual and code-based transformations to clean, wrangle, and enrich your data in modularized, reusable steps and pipelines with full traceability.

VISUAL DATA TRANSFORMATION

- ✓ Simple to advanced analytic functions and processing logic in a point and click, guided visual interface
- ✓ Group
- ✓ Join (inner, left, right, outer)
- ✓ Fuzzy join (text, numbers, dates, etc.)
- ✓ Geo join (contains, intersects, within distance, etc.)
- ✓ Sample/Filter
- ✓ Split
- ✓ Stack
- ✓ Window processing
- ✓ Distinct
- ✓ Top-N
- ✓ Pivot
- ✓ Sort
- ✓ Sync

SPECIALIZED DATA PREPARATION

- ✓ Geospatial enrichment (create geo points,reverse geocoding, geo join)
- ✓ Time series preparation (resampling, windowing, extrema extraction, interval extraction)
- ✓ Image preparation and augmentation (preprocessing, color, affine, crop transformations, annotation with bounding boxes)

→ **Text preparation**

- ☑ Language detection
- ☑ Tokenization
- ☑ Normalization
- ☑ Filtering
- ☑ Lemmatization
- ☑ Spell checking
- ☑ Translation
- ☑ Fuzzy matching

INTERACTIVE DATA PREPARATION

- ✓ Smart transformation suggestions based on inferred data types

→ **More than 100 built-in processors**

- ☑ Column and row operations
- ☑ Reshaping
- ☑ Data filtering and cleaning
- ☑ Text processing
- ☑ Custom formula language
- ☑ Dates handling
- ☑ Semantic extraction
- ☑ Geoprocessing
- ☑ Specialized web logs processing
- ☑ Support for objects and arrays
- ☑ Custom Python transformations

Exploratory Data Analysis (EDA)

Before you build a data product, analytics dashboard, or predictive model, understanding the source data is critical to success. Dataiku makes EDA easy with immediate visual exploration as well as automated guided suggestions for types of insights to compute depending on detected data types.

DATA PROFILING

- ✓ Instant data quality analysis: semantic meaning detection, color cues for missing and invalid values, and context-based recommendations for data cleansing
- ✓ Automatic descriptive column statistics (histogram/bar chart of value distribution, summary statistics, outlier detection, etc.)

INTERACTIVE VISUAL STATISTICS

→ Univariate analysis)

- ☑ Side by side variable comparison
- ☑ Histograms
- ☑ Box plots
- ☑ Quantile tables
- ☑ Frequency tables
- ☑ Cumulative distribution function

→ Bivariate analysis

- ☑ Variables relationships
- ☑ Histogram
- ☑ Box plot
- ☑ Scatter plot
- ☑ Frequency table

→ Fit curves and distributions

- ☑ Probability density, Q-Q plots, goodness of fit metrics and estimated parameters
- ☑ Fit distributions supported - beta, exponential, Laplace, log-normal, normal, normal mixture, Pareto, triangular, Weibull
- ☑ 2D Fit distribution: kernel density estimate (KDE) or joint normal (Gaussian) distribution
- ☑ Fit curve - polynomial or isotonic

→ Statistical tests

- ☑ Hypothesis testing
- ☑ One sample t-test, sign test, Shapiro-Wilk test
- ☑ Pairwise t-test, median Mood, K-S
- ☑ N-sample One-way ANOVA, median mood, pairwise student t-test, pairwise median mood
- ☑ Categorical chi-square independence test

→ Multivariate analysis

- ☑ Numerical variables distribution
- ☑ Principal component analysis (PCA)
- ☑ Correlation matrix
- ☑ Scatter plot (3D)
- ☑ Parallel coordinates plot

→ Time series analysis

- ☑ Stationarity and unit root tests: Augmented Dickey-Fuller (ADF), Zivot-Andrews, Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests
- ☑ Mann-Kendall trend test
- ☑ (Partial) Auto-correlation function plot and Durbin-Watson statistic

Data Science & Visual Machine Learning

Dataiku offers the latest machine learning (ML) technologies and techniques all in one place so that data scientists of all kinds, from citizen data scientists to advanced data scientists, can focus on what they do best: building and optimizing the right model for the use case at hand.

AUTOML

- ✓ Develop highly optimized models with visual intervention
- ✓ Teams can modify automatic settings to customize the models
- ✓ Templates to prioritize model performance, interpretability, or training speed

→ **Automated model documentation**

- ☑ Customizable templates with training and model metadata

ALGORITHMS

→ **Visual ML Engines:**

- ☑ In-memory Python (scikit-learn / LightGBM / XGBoost)
- ☑ MLLib (Spark) engine

SUPERVISED LEARNING

- ✓ Ordinary least squares
- ✓ Ridge regression
- ✓ Lasso regression
- ✓ Logistic regression
- ✓ Linear Regression
- ✓ Random forest
- ✓ Gradient boosted trees
- ✓ XGBoost
- ✓ LightGBM
- ✓ Decision trees
- ✓ Naive bayes
- ✓ Support vector machines
- ✓ Stochastic gradient descent
- ✓ K-Nearest neighbors
- ✓ Extra random trees
- ✓ Artificial neural network
- ✓ Lasso path
- ✓ Extensible with custom Python and Scala models

→ **Ensembling**

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

UNSUPERVISED LEARNING (CLUSTERING)

- ✓ K-means
- ✓ Gaussian mixture
- ✓ Mini-batch K-means
- ✓ Agglomerative clustering
- ✓ Spectral clustering
- ✓ DBSCAN
- ✓ Interactive clustering (two-step clustering)
- ✓ Isolation Forest (Anomaly Detection)
- ✓ Custom Models

MODEL DESIGN

→ **Dataset sampling**

- ☑ Default 80/20 random training and testing set split
- ☑ Customizable sampling strategy
- ☑ Time-based ordering
- ☑ Subsampling for memory optimization
- ☑ Random, column values subset, class rebalancing sampling
- ☑ K-fold cross-test
- ☑ Explicit extracts

→ Automatic features handling depending on data type

- ☑ Category handling (Dummy, target, ordinal, frequency encoding, feature hashing)
- ☑ Target encoding (Impact coding, GLMM encoding)
- ☑ Missing values
- ☑ Numerical handling (rescaling, binarize, quantize)
- ☑ Datetime cyclical encoding
- ☑ Text handling (count and tf/idf vectorization, hashing trick, truncated SVD, sentence embedding)
- ☑ Vector unfolding
- ☑ Extensible with custom Python preprocessings

→ Feature generation

- ☑ Per-feature derivatives
- ☑ Linear combinations
- ☑ Polynomial combinations

→ Feature reduction

- ☑ Filtering methods (correlation with target, tree-based feature importance)
- ☑ Lasso
- ☑ PCA

→ Clustering settings

- ☑ Per-feature derivatives
- ☑ Dimensionality reduction

→ Advanced model optimization

- ☑ Multiple cross-validation strategies
  - + Train/test splitting policies
  - + K-Fold cross-validation and cross-testing
  - + Time-aware folds
- ☑ Hyperparameter search strategies
  - + Grid
  - + Random
  - + Bayesian
- ☑ Interruptible and resumable, time-boundable search
- ☑ Distributions and multi-threading powered by Elastic AI on Kubernetes
- ☑ Real time result and progress visualization

PREDICTION RESULTS AND EVALUATION

✓ Multiple performance evaluation metrics

→ For regression

- ☑ Explained Variance Score, MAPE, MAE, MSE, RMSE, RMSLE, R2 Score, custom code

→ For classification

- ☑ F1 Score, Accuracy, Precision, Recall, Cost matrix, AUC, Log Loss, Cumulative lift, custom code

✓ Adjustable and optimized thresholds for F1 score, accuracy, cost matrix

→ Advanced result visualization

- ☑ Confusion matrix
- ☑ Decision chart
- ☑ Lift charts
- ☑ Calibration curve
- ☑ ROC curve
- ☑ Error bins
- ☑ Scatter plots
- ☑ Metrics and assertions

→ ML Diagnostics

- ☑ Alerts for unbalanced data, outliers, training speed, underfitting, overfitting, or data leakage
- ☑ Model assertions to validate and detect counter-intuitive predictions

→ Explainability, interpretability and bias detection

- ☑ Decision tree analysis
- ☑ Individual prediction explanations (Shapely values, ICE)
- ☑ Feature importance
- ☑ Regression coefficients
- ☑ Partial dependence
- ☑ Subpopulation analysis
- ☑ Model fairness report
- ☑ Model error analysis

→ Interactive what-if simulations and analysis

- ☑ Counterfactuals and actionable recourse analysis for classification tasks
- ☑ Outcome optimization analysis for regression tasks

TIME SERIES FORECASTING

- ✓ Trivial identity
- ✓ Seasonal naive
- ✓ AutoARIMA
- ✓ Seasonal trend
- ✓ Non-parametric time series
- ✓ Simple feed forward
- ✓ DeepAR
- ✓ Transformer
- ✓ Multi-horizon quantile - Convolutional Neural Network (MQ-CNN)

NO-CODE COMPUTER VISION TASKS

- ✓ No-code tasks for object detection and image classification
- ✓ Built-in pre-trained models for transfer learning
- ✓ Image labeling with collaborative, managed labeling system for annotators
- ✓ Active learning
- ✓ Image augmentation (color, affine, crop)
- ✓ Train and deploy with CPU or GPU
- ✓ What-if live scoring for images with activation heatmap overlays for explainability

NATURAL LANGUAGE PROCESSING (NLP)

→ Text extraction:

- ☑ Optical Character Recognition (OCR)
- ☑ Speech to Text

→ Text analysis

- ☑ Sentiment analysis
- ☑ Named entity recognition
- ☑ Summarization
- ☑ Classification
- ☑ Key phrase extraction
- ☑ Categorization
- ☑ Ontology tagging
- ☑ Zero-shot classification
- ☑ Natural language generation (NLG) with GPT-3
- ☑ Similarity search with TF-IDF or embeddings
- ☑ Text labeling
- ☑ Transformers (BERT, GPT, etc.)
- ☑ Language detection
- ☑ Named entity recognition
- ☑ Topic extraction

→ Text visualization

- ☑ Entity visualizer
- ☑ Word clouds

- ✓ Machine translation (AWS, Azure, DeepL, Dataiku, Google)
- ✓ Visual interface for third-party NLP APIs (Crowlingo, Google Cloud, Azure)

LOW-CODE DEEP LEARNING

- ✓ Same feature management capabilities for visual ML
- ✓ Feature processing, epoch tracking and charts, early stopping capabilities, multiple input tensors
- ✓ Support for image and text features
- ✓ Keras with Tensorflow backend, Tensorboard integration
- ✓ Support for code environments, CPU and GPU, including multiple GPUs and cloud-enabled dynamic GPUs clusters
- ✓ Support pre-trained models, transfer learning

Coding & Extensibility

The Dataiku platform supports coders developing processes and models using notebooks or IDEs with Dataiku APIs and SDKs. Coders can also extend Dataiku's native functionality with plugins and custom code.

CODE-BASED DATA EXPLORATION, TRANSFORMATION, AND ML

→ Coding recipes in the flow:

- ☑ Python
- ☑ R
- ☑ Shell
- ☑ SQL
- ☑ Hive & Impala
- ☑ Spark (Scala, SQL, Python, R)

→ Code Notebooks

- ☑ Full Jupyter Notebook support for Python, R, and Scala (including Spark)
- ☑ SQL / Hive / Impala / SparkSQL notebooks
  - + Interactive query environment
- ☑ Templated notebooks for statistical analysis, dimensionality reduction, time series and topics modeling using NMF and LDA.
- ☑ Containerized execution (local Docker, Kubernetes cluster)

CODING ENVIRONMENTS

→ Managed custom package directory

- ☑ Open environment to install any Python or R package
- ☑ Packages dependency management
- ☑ Python 2.7, 3.6, 3.7, 3.8, 3.9 and 3.10
- ☑ Support for conda, pip, virtualenv, R

→ Full Git integration

- ☑ Project version control and traceability
- ☑ Manual or automatic commits, reverting, branching, remote use.
- ☑ Import any third-party library, Jupyter notebooks, or existing code assets from Git
- ☑ Share repositories per project or globally

→ Reuse with plugins

- ☑ Package and ship complex code-based functions as GUI-wrapped components
- ☑ Extend native Dataiku capabilities with custom connectors, processors, recipes, interactive visualizations and applications, and scenario steps

→ Powerful APIs

- ☑ Advanced data manipulation, model training and deployment, and MLOps through the Dataiku public API (Python API client, HTTP REST API) and R API

→ Code studioss

- ☑ Editing and debugging code in built-in code studios synchronized with a Dataiku project
  - + Visual Studio Code
  - + RStudio
  - + JupyterLab
- ☑ Containerized backends running on elastic Kubernetes infrastructure

→ IDE Integrations

- ☑ Support for IDE extensions for Visual Studio Code, PyCharm, RStudio

DataOps - Pipeline and Production Automation

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

DATA-AWARE DATA FLOW

- ✓ Dataset dependency checks
- ✓ Data lineage
- ✓ Consistency checks (data, schema, data types)
- ✓ Flow zones for logical grouping
- ✓ Dynamic dataset rebuild

AUTOMATED METRICS & CHECKS

→ Gather and historize measurements on your datay
- ☑ Size
- ☑ Records counts
- ☑ Per-column statistics (min, max, sum, number of distinct values, most frequent values, etc.)
- ☑ Custom SQL queries
- ☑ Custom Python code

→ Checks
- ☑ Numeric range
- ☑ Value in set
- ☑ Custom Python chec

- ✓ Different execution engines: Hive, Impala, SQL DB, streaming data engine

AUTOMATION SCENARIOS

→ Trigger execution of data flows and applications based on
- ☑ Schedule
- ☑ Dataset changes
- ☑ SQL query
- ☑ Model performance
- ☑ Input data drift
- ☑ Custom triggers (Python)

→ Assemble actions (steps)
- ☑ Build datasets
- ☑ Synchronize Hive table
- ☑ Check consistency
- ☑ Compute metrics and run checks
- ☑ Train or deploy models
- ☑ Define, set, and update project and global variables
- ☑ Run custom Python or SQL code
- ☑ Create notebook export
- ☑ Create dashboard export
- ☑ Package API service
- ☑ Implement loops and conditions
- ☑ Run another scenario

→ Publish results
- ☑ Send emails with custom templates
- ☑ Send notifications to Slack or Microsoft Teams
- ☑ Attach datasets, dashboards, reports, etc.

PARTITIONING

- ✓ File-based (using file system hierarchy) or column based (for SQL DBs, Mongo and Cassandra) partitioning

FLEXIBILITY FOR EXECUTION ENGINE

→ Dataiku server (streamed data, no need for in-memory)

→ Hadoop clusters
- ☑ Cloudera's CDP
- ☑ Amazon Elastic MapReduce (EMR)

- ✓ Spark clusters
- ✓ Run in SQL database
- ✓ Additional supported transformations for push-down in Snowflake
- ✓ Containerized execution (Kubernetes)

MLOps

Implement complete machine learning operations (MLOps) lifecycles with confidence, from initial deployment to ongoing management at scale.

PREPARE

→ Model description

- ☑ Business initiative with model expectations and responsibilities (Govern)

→ Data Discovery

- ☑ Catalog & Feature Store to quickly find the most appropriate data
- ☑ Connector richness
- ☑ Interactive statistics

→ Data preparation

- ☑ Visual & Code recipes
- ☑ Flow lineage

BUILD

→ Model development

- ☑ Auto ML
- ☑ Custom model (Python & MLflow)
- ☑ Experiment tracking - code & visual

→ Model readiness

- ☑ Model testing - Diagnostics, assertions & Model Stress tests
- ☑ Model & Flow document generator
- ☑ Model & project versioning
- ☑ Native integration with Git and Git repositories

→ Model Export

- ☑ Java class/ JAR
- ☑ PMML
- ☑ Snowflake UDF
- ☑ SQL
- ☑ ONNX
- ☑ Python
- ☑ MLflow

DEPLOY

→ Model Sign-off

- ☑ Model registry (Govern)
- ☑ Project blueprints (Govern)
- ☑ Model review & approval (Govern)

→ Model Deployment

- ☑ Bundles & API packages for consistent
- ☑ Controlled self-service deployment with Dataiku Deployer
- ☑ Automated deployment with scenarios or API-based
- ☑ Support for multiple deployment environments - DEV, UAT, PROD…

→ Model consumption

- ☑ Easily expose and turn into realtime REST APIs:
  - + Models trained in Visual ML
  - + Custom models trained in code
  - + Custom Python or R functions
  - + Lookups in datasets
  - + SQL queries
  - + Query enrichments (lookup mapping, retrieved columns, error handling)
- ☑ Deploy and run in dedicated production environments:
  - + Data flows
  - + Interactive web applications to consume data and models

MONITOR

→ Model Monitoring

- ☑ Historize training and performance metadata for all runs through evaluate recipes and Model Evaluation Store
- ☑ Drift monitoring and decay tracking in Model Evaluation Store
  - + Input data drift
  - + Prediction drift
  - + Performance drift
- ☑ Monitor external models using their prediction logs (Model exported from Dataiku or even completely foreign models)

→ Continuous Improvement

- ☑ Design monitoring dashboards & webapp
- ☑ Compute and display actionable metrics & checks
- ☑ Build automated actions through scenarios or API-based
- ☑ CI/CD-ready through integrations with Jenkins, Azure DevOps or JFrog Artifactory

Visualization and Applications

Create insights across the various stages of the project from charts, to statistical insights, model information, and results that can be centralized in dashboards and shared with team members and stakeholders.

CHARTS AND GRAPHS

- ✓ Execute on Dataiku optimized CPU engine or in-database
- ✓ Preview on sample of the data for real-time results.
- ✓ Charts: bar, histogram, line, curve, pie, donut, scatter, boxplot
- ✓ Maps: scatter, binned, administrative, density, geometry
- ✓ Tables, heat map
- ✓ KPI chart with conditional formatting
- ✓ Treemap
- ✓ Custom Python-based or R-based charts (GGplot, Plotly, Matplotlib)
- ✓ Customizable color and meaning-associated palettes

DASHBOARDS

- ✓ Charts, metrics, notebook exports
- ✓ Customizable visibility and access control
- ✓ Export to PDF or image
- ✓ Integrations with PowerBI, Tableau, Qlik, Microstrategy

CODE REPORTS

- ✓ RMarkdown reports
- ✓ Jupyter Notebooks reports

NO-CODE APPLICATIONS

- ✓ Visual application framework
- ✓ Application-as-recipe builder

WEB APPLICATIONS

- ✓ Dash
- ✓ Shiny
- ✓ Bokeh
- ✓ Streamlit
- ✓ Custom Flask + HTML, CSS, Javascript applications

AI Governance

Achieve enterprise-grade governance and AI portfolio oversight to safely scale and deliver AI

GOVERNABLE ITEMS

- ✓ Dataiku projects
- ✓ Models
- ✓ Model versions
- ✓ Project bundles

GOVERNANCE WORKFLOW

- ✓ Standardized governance workflows
- ✓ Adaptive governance plans
- ✓ Risk and value assessment

SIGN-OFFS AND APPROVALS

- ✓ Multiple reviewers at any stage
- ✓ Final sign-off before production deployment
- ✓ Configurable email notifications

CENTRALIZED REGISTRIES

- ✓ Model registry
- ✓ Project bundle registry
- ✓ Metrics by project or model

Collaboration

Dataiku was designed from the ground up with collaboration in mind so business, data, and production teams can work faster and smarter together. Coders and non-coders can work side by side on the same project, each contributing based on their business or technical knowledge.

PROJECT MANAGEMENT AND DOCUMENTATION

- ✓ Project wiki, description, task management
- ✓ Tag, annotate and view status of data assets
- ✓ Team activity monitoring

CENTRALIZED KNOWLEDGE

- ✓ Global catalog for data assets and custom transformations
- ✓ Feature store
- ✓ Code samples
- ✓ Project libraries

SHARED ENVIRONMENT FOR ALL ROLES

→ **Shared visual environment for multiple collaborators:**

☑ data analysts, data scientists, data engineers, business stakeholders, IT operators, etc.

- ✓ Ability to work visually, via code, or anywhere in between
- ✓ Discussion boards and comment function for users
- ✓ Git-based collaboration with full traceability and branching control
- ✓ Package arbitrary complex functions, operations, or business logic via plugins
- ✓ Gather insights from multiple projects in a single workspace
- ✓ Manage user access and permissions for different roles
- ✓ Share applications, dashboards, webapps, datasets, and wiki articles
- ✓ Object discussions
- ✓ Discussion inbox
- ✓ Email notifications

Data Protection & Security

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

USER ACCESS

- ✓ User profiles with sets of permissions
- ✓ Granular permissions per project for user or user groups
- → **Authentication management**
  - ☑ SSO systems (support for OIDC and SAML v2)
  - ☑ LDAP, Active Directory
  - ☑ Native integrations to cloud identity and access management services and OAuth
- ✓ User secrets with encrypted credentials

DATA MANAGEMENT

- ✓ Create, configure, and limit access to data connections
- ✓ Passwords encryption for 3rd party systems
- ✓ Global service credential or personal user credentials data access
- ✓ Granular permissions at object level
- ✓ Objects can be made shareable across projects or instances

SECURITY

- → **Enterprise-grade security**
  - ☑ Audit trail
  - ☑ Kerberos authentication
  - ☑ Users impersonation.
- ✓ GDPR functionalities for policy frameworks

Architecture

Dataiku is built for the modern enterprise with a powerful, flexible and open architecture so IT teams can focus on scaling AI initiatives, not on infrastructure.

DATAIKU ARCHITECTURE

✓ Browser-based (Google Chrome, Mozilla Firefox, Microsoft Edge) user interface

→ Installing and setting up options
- ☑ Dataiku Cloud: Hosted/SaaS
- ☑ Deploy on all major cloud providers
- ☑ Custom Dataiku: own Linux server deployed either on-premises or on any cloud

✓ Highly available deployments, load balancing

✓ Scalable, self-contained, multi-process architecture for high availability, resilience and disaster recovery

✓ All instances can be centrally administered, managed and updated

✓ Complete monitoring infrastructure

→ Modular platform deployment through Dataiku server nodes:
- ☑ Design
- ☑ Deploy
- ☑ Automation (batch deployment)
- ☑ API (realtime endpoints)
- ☑ Govern

SCALING AND PUSH DOWN COMPUTATION

→ Elastic AI computation strategies: Processes within Dataiku can run on one or several hosts powered by Docker or Kubernetes
- ☑ Data transformations
- ☑ Python and R recipes or notebooks
- ☑ Spark-powered code and visual recipes, notebooks
- ☑ Machine learning models training, scoring, and evaluation
- ☑ Code environments with containerized execution (CPU and GPU support)
- ☑ Code studios IDEs
- ☑ API nodes can run multiple containers orchestrated by Kubernetes

→ Other external leverageable execution engines:
- ☑ Dataiku server (streamed data, no need for in-memory)
- ☑ Hadoop clusters
  - + Cloudera's CDP
  - + Amazon Elastic MapReduce (EMR)
- ☑ Spark clusters
- ☑ Run In-SQL database
- ☑ Additional supported transformations for push-down in Snowflake

SCORING INFRASTRUCTURE

→ Real-time scoring APIs
- ☑ Complete API administration and management
- ☑ Zero-downtime model updates and upgrades
- ☑ Static infrastructure or deploying on Kubernetes
- ☑ Horizontally scalable and highly available

→ Batch Scoring Engines
- ☑ Local (Dataiku server)
- ☑ Spark
- ☑ SQL (Regular and Snowflake)

DATAIKU CLOUD STACKS

→ **Deploy a complete modern data, AI, and analytics stack on**

- ☑ Amazon Web Services (AWS)
- ☑ Microsoft Azure
- ☑ Google Cloud Platform (GCP)

✓ Central fleet manager to deploy, upgrade, backup, restore, and configure one or several Dataiku instances

✓ Templated, automated deployment and configuration of Dataiku instances and nodes

✓ Can be deployed in virtual private cloud (VPC) with secure connectivity (no public IP)

✓ Full administration environment for onboarding, maintenance and upgrades

✓ Integrated to cloud security and disaster recovery