データシート

2021

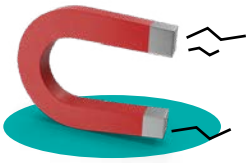
Everyday AI Extraordinary People

- 日々の AI で、皆が一歩先へ -

Everyday AI, Extraordinary People - 日々のAIで、皆が一歩先へ -

Dataikuは、Everyday AIのためのプラットフォームであり、比類ない業績を実現できるようにデータ活用を仕組み化します。Dataikuを使用する企業は、技術分野でコーディングに取り組む従業員であろうと、ビジネス分野でコーディングへの取り組みが少ないか、あるいは全くない従業員であろうと、データを利用してより優れた日々の意思決定を行えるような並外れたレベルへと彼らを引き上げることができます。

全世界450以上の企業がDataikuを利用することでデータとAIの活用を仕組み化し、不正検知をはじめとして顧客の離反防止、予知保全、サプライチェーン最適化、そしてその間の全てのものまで、さまざまなユースケースを推進しています。



接続性

Dataiku では、データの格納場所や形式に関わらず、あらゆるデータにシームレスに接続することが可能です。これは - 技術的であるかどうかは関係なく - 全員が必要なデータに簡単にアクセスできることを意味します。

SQL データベース

- ☑ MySQL
- ☑ PostgreSQL
- ☑ Vertica
- ☑ Amazon Redshift
- ☑ Pivotal Greenplum
- ☑ Teradata
- ☑ IBM Netezza
- ☑ SAP HANA
- ☑ Oracle
- ☑ Microsoft SQL サーバー（SQL DW を含む）
- ☑ Google BigQuery
- ☑ IBM Db2
- ☑ Exasol
- ☑ MemSQL
- ☑ Snowflake
- ☑ JDBC を介したカスタム接続

NoSQL データベース

- ☑ MongoDB
- ☑ Cassandra
- ☑ ElasticSearch

Hadoop および Spark 対応のディストリビューション

- ☑ Cloudera
- ☑ Hortonworks
- ☑ Google DataProc
- ☑ MapR
- ☑ Amazon EMR
- ☑ DataBricks

Hadoop ファイル形式

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

ストリーミングデータソース

- ☑ Kafka
- ☑ AWS SQS
- ☑ Spark

リモートデータソース

- ☑ FTP
- ☑ SCP
- ☑ SFTP
- ☑ HTTP

クラウドオブジェクトストレージ

- ☑ Amazon S3
- ☑ Google Cloud ストレージ
- ☑ Azure Blob ストレージ
- ☑ Azure Data Lake Store Gen1 & Gen2

カスタムデータソース - Dataiku プラグインを介した拡張接続性

- ☑ REST APIへの接続
- ☑ カスタムファイル形式の作成
- ☑ データベースへの接続

以下の間で最適化された同期：

- ☑ Snowflake と WASB
- ☑ S3 と Amazon Redshift
- ☑ Snowflake と S3

Spark Driver での Snowflake に対するネイティブサポート

探索的分析

時にはデータを深く掘り下げて調べる必要がありますが、その他の時は一目で理解することが重要です。分析に使えるデータセットを探し出すことからダッシュボードを作ることまで、Dataiku によってこうした分析が簡単になります。

データ分析

- 自動的にデータセットスキーマおよびデータタイプを検出
- セマンティックな意味をデータセット列に割り当て
- 一変量解析を自動的に構築し、データの品質をチェック
- データセット監査
  - ☑ Dataiku の全データセットに対するデータ品質と統計的な分析を自動的に作成
  - ☑ 監査用に複数のバックエンドに対応（インメモリ、Spark、SQL）

高度な分析

- インタラクティブなビジュアル統計
  - ☑ 単一または複数の母集団での一変量解析および統計テスト
  - ☑ 複数母集団での統計およびテスト
  - ☑ 相関分析
  - ☑ 主成分分析
- 事前定義済の Python ベースの Jupyter ノートブックを活用
  - ☑ すべての分析をビジュアル統計でサポート
  - ☑ 高次元データ可視化 （t-SNE）
  - ☑ トピックモデル
- 時系列
  - ☑ 再サンプリング、窓掛け、局地抽出、感覚抽出のための、ビジュアルレシピを用いた時系列データ準備
  - ☑ 時系列可視化
  - ☑ 時系列予測

データのカタログ化

- 一元管理されたカタログで、データ、コメント、機能、またはモデルを検索
- 既存のすべての接続からデータを探索

データの可視化

- 標準的なチャート（ヒストグラムや棒グラフなど）を作成でき、基盤とするシステムを活用しチャート処理の対象を拡大可能（インデーターベース集計）
- 以下を使用してカスタムチャートを作成
  - ☑ カスタム Python ベースまたは R ベースのチャート
  - ☑ カスタムウェブアプリケーション（HTML/JS/CSS/Flask）
  - ☑ Shiny Web Applications®
  - ☑ Bokeh Web Applications (Python)

ダッシュボード化

- ユーザー管理レポートとダッシュボード
  - ☑ RMarkdown レポート
  - ☑ Jupyter Notebooks レポート
  - ☑ カスタマイズされたインサイト（GGplot、Plotly、Matplotlib）
  - ☑ カスタマイズ可能なインタラクティブでWebベースのビジュアルダッシュボード

データ準備

一般的に、データ準備にはデータプロジェクトの 80 %の時間がかかっています。Dataiku のデータ準備機能によって、そのプロセスはより速くより簡単になります。これにより、より重要（かつクリエイティブ）な作業により多くの時間を割けるようになります。

ビジュアルツールによるデータ変換

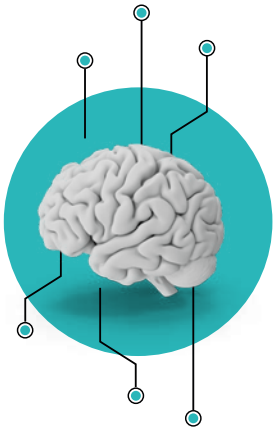
- データ変換ジョブを、ポイントアンドクリックインターフェースを使用して設計
  - ☑ グループ
  - ☑ フィルター
  - ☑ 並べ替え
  - ☑ 積み重ね
  - ☑ 結合
  - ☑ ウィンドウ
  - ☑ 同期
  - ☑ 重複削除
  - ☑ 上位-N
  - ☑ ピボット
  - ☑ 分割
- 分散コンピューティングシステム（SQL、Hive、Spark、Impala）で直接実行することによって、大量データの変換に対応
- タスクのために生成されたコードを確認して調整

データセットのサンプリング

- 先頭レコード、ランダム選択、層化抽出など

インタラクティブなデータ準備

- プロセッサーにより（90内蔵、シンプルなテキスト処理から）カスタマイズされたPythonベースあるいは数式ベースのデータ変換に対応
- データ準備に使うスクリプトをインデータベース（SQL）やインクラスタ（Spark）で処理することで大量データに対応



機械学習

Dataiku は最先端の機械学習テクノロジーをまとめて提供しているため、データサイエンティストはモデルの開発や最適化など自身の得意な作業に集中できます。

自動機械学習 (AutoML)

・自動 ML 戦略

- ☑ 迅速なプロトタイプ作成
- ☑ 説明可能なモデル
- ☑ 高性能

・機械学習のための機能

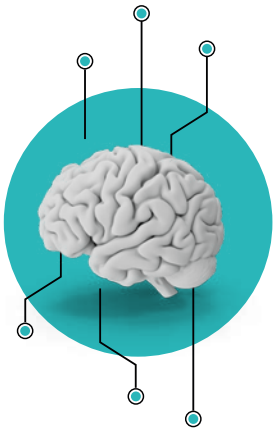
- ☑ 数値変数、カテゴリ変数、テキスト変数、ベクトル変数に対応
- ☑ カテゴリ変数の前処理自動化（ダミーエンコーディング、インパクトコーディング、ハッシング、その他のカスタマイズ）
- ☑ 数値変数の前処理自動化（標準化、分位点によるビン分割、その他のカスタマイズ）
- ☑ テキスト変数の前処理自動化（TF/IDF、ハッシュトリック、切り捨て SVD、カスタム前処理）
- ☑ さまざまの欠損データの補完術
  - + 特徴量生成
    - ◊ *特徴量ごとの派生変数（平方、平方根...）*
    - ◊ *線形および多項式の組み合わせ*
  - + 特徴量選択
    - ◊ *フィルターおよび内蔵メソッド*

・複数の ML バックエンドから選択して、モデルをトレーニング

- ☑ TensorFlow
- ☑ Keras
- ☑ Scikit-learn
- ☑ XGBoost
- ☑ MLLib
- ☑ H20

・アルゴリズム

- ☑ Python ベース
  - + 最小二乗
  - + リッジ回帰
  - + ラッソ回帰
  - + ロジスティック回帰
  - + ランダムフォレスト
  - + 勾配ブースティングツリー
  - + XGBoost
  - + 決定木
  - + サポートベクターマシン
  - + 確率的勾配降下法
  - + k近傍法
  - + エクストラランダムツリー
  - + 人工ニューラルネットワーク
  - + Lasso パス
  - + LightGBM など scikit-learn の仕様に似た形のカスタムモデル
- ☑ Spark MLLib ベース
  - + ロジスティック回帰
  - + 線形回帰
  - + 決定木
  - + ランダムフォレスト
  - + 勾配ブースティングツリー
  - + 単純ベイズ
  - + カスタムモデル
- ☑ H20 ベース
  - + 深層学習
  - + GBM
  - + GLM
  - + ランダムフォレスト
  - + 単純ベイズ



機械学習

Dataiku は最先端の機械学習テクノロジーをまとめて提供しているため、データサイエンティストはモデルの開発や最適化など自身の得意な作業に集中できます。

自動機械学習 (AutoML)

・ハイパーパラメーターの最適化

- ☑ ハイパーパラメーターを自由に設定して検索
- ☑ グリッド、ランダム、ベイジアンのハイパーパラメーターの最適化と検索をサポート
- ☑ 交差検証戦略
  - + 学習用・テスト用データの様々な分割方法に対応
  - + K-分割の交差テスト
  - + 様々な評価指標（Explained Variance Score、MAPE、MAE、MSE、正解率、F1スコア、コストマトリックス、AUCなど）を使ったモデル最適化
- ☑ グリッドサーチの中断と再開
- ☑ グリッドサーチ結果の視覚化
- ☑ 確率予測の自動再キャリブレーション
- ☑ Kubernetes上での分散ハイパーパラメーター検索

・モデルトレーニング結果の分析

- ☑ モデルから洞察を得る
  - + データのスコアリング
  - + 特徴量の重要度
  - + モデルパラメーター
  - + 部分従属プロット
  - + 回帰係数
  - + 部分母集団のバイアスおよびパフォーマンス分析
  - + 個々の予測の説明
  - + モデル公平性のレポート
  - + インタラクティブスコアリング（What-if分析）
  - + ML診断
  - + モデルアサーション
- ☑ トレーニング結果を Dataiku ダッシュボードで公開
- ☑ 監査モデルのパフォーマンス
  - + 混同行列
  - + デシジョンチャート
  - + リフトチャート
  - + ROC 曲線
  - + 確率分布チャート
  - + その他詳細な評価指標（正解率、F1スコア、ROC-AUC、MAE、RMSEなど）

・アンサンブルモデルの自動生成

- ☑ 線形スタッキング（回帰モデル用）またはロジスティックスタッキング（分類問題用）
- ☑ 予測の平均値あるいは中央値（回帰問題用）
- ☑ 多数決（分類問題用）

・スコアリング機能

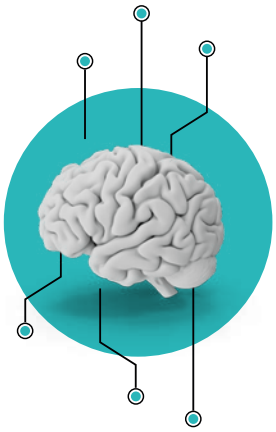
- ☑ リアルタイムサーバーレススコアリング API
- ☑ Spark による分散バッチ
- ☑ SQL (インデータベーススコアリング)
- ☑ Dataiku 内蔵エンジン

・モデルのエクスポート

- ☑ 学習したモデルを Java クラスのセットとしてエクスポートすることで、あらゆるJVMアプリケーションで非常に効率的なスコアリングが可能
- ☑ 学習したモデルを PMML ファイルとしてエクスポートし、PMMLに対応したスコアラーでスコアリング可能

・モデルレポートの自動作成

- ☑ モデルの学習結果を自動的に文書へ落とし込み。テンプレートも使用可



機械学習

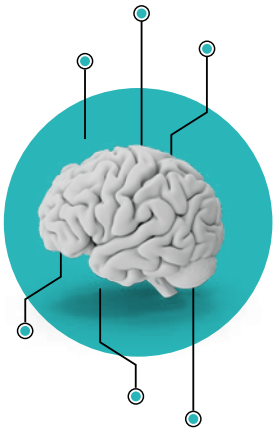
Dataiku は最先端の機械学習テクノロジーをまとめて提供しているため、データサイエンティストはモデルの開発や最適化など自身の得意な作業に集中できます。

デプロイメント

- モデルのバージョニング
- バッチスコアリング
- リアルタイムスコアリング
  - ☑ 他のアプリケーションによるリアルタイムスコアリング用に、モデルを REST API 経由で公開
- どんな関数やモデルでもREST APIで公開
  - ☑ Python、R、SQLで書いた関数やモデルを開発
  - ☑ それらをAPIエンドポイントに自動変換
- 簡単なモデルのデプロイメント
  - ☑ ワンクリックだけでデプロイOK
- Docker & Kubernetes
  - ☑ Dockerコンテナにモデルをデプロイ
  - ☑ イメージを Kubernetes クラスタに自動で push、スケーラブルなデプロイメントを実現
  - ☑ 特別な設定なしに Kubernetes 上でSparkを利用可
- モデルのモニタリング
  - ☑ モデルのパフォーマンスを常時監視
  - ☑ データドリフトの検知
  - ☑ 性能が落ちたら自動的に再学習
  - ☑ 再学習の戦略をカスタマイズ
- ロギング
  - ☑ モデルに送信されたすべてのクエリをログ記録して監査

深層学習

- Tensorflow バックエンドで Keras をサポート
- ユーザー定義のモデルアーキテクチャ
- トレーニング設定をパーソナライズ
- 使用するモデルへの複数の入力に対応
- CPU および GPU をサポート
- 事前トレーニング済モデルをサポート
- 特徴量を画像から抽出
- Tensorboard 統合



機械学習

Dataiku は最先端の機械学習テクノロジーをまとめて提供しているため、データサイエンティストはモデルの開発や最適化など自身の得意な作業に集中できます。

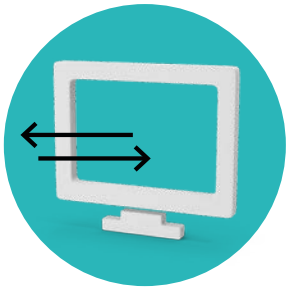
教師なし学習

- 特徴量エンジニアリングの自動化（教師あり学習も同様）
- 次元の削減
- 外れ値検出
- アルゴリズム

☑ K-means
☑ ガウス混合
☑ 凝集型クラスタリング
☑ スペクトルクラスタリング
☑ DBSCAN
☑ インタラクティブクラスタリング（二段階クラスタリング）
☑ 分離フォレスト（異常検知）
☑ カスタムモデル

スケーラブルなモデル学習

- モデルを Kubernetes でトレーニング



出力機能

インサイトを発見した後、それを組織の意思決定者に効果的に伝えて行動を引き出すことが重要です。Dataiku はすべての人がインテリジェンスに基づいた意思決定をできるよう、AI のパワーを提供します。

チャート

- ☑ 棒グラフ、線グラフ、円グラフ、ドーナツグラフ、散布図、箱ひげ図、2D分布、リフトチャート
- ☑ マップ: 分布、ビン分割、行政区域
- ☑ テーブル

Dataiku アプリ

- ☑ プロジェクトの上にユーザーフレンドリーなインターフェースを作り、ユーザーがコードなしで数クリックするだけでカスタマイズして、パラメーター化できるようにします
- ☑ アプリを Dataiku でレシピまたは API のように共有

ダッシュボード

- ☑ チャート、テーブル、エクスポートしたノートブック、Web アプリなどでインタラクティブに洞察を得ることが可能

Dataiku Webアプリ

- ☑ コードを使って、ユーザー向けの API として活用可能な高度にカスタマイズされたアプリケーションを作成

自動化機能

ワークフローのストリーミングと自動化についてデータチームが適切なプロセスを導入することで、本番環境でモデルを適切に監視し、容易に管理できるようにします。

データフロー

- ☑ データセット間の依存状態の追跡を維持する
- ☑ データ系列全体の管理
- ☑ データ、スキーマまたはデータタイプの一貫性を確認
- ☑ フローをゾーンに分けて管理

パーティション分割

- ☑ HDFS または SQL のパーティション分割メカニズムを使用して、コンピュテーションの時間を最適化

評価指標と確認

- ☑ データの一貫性と品質を評価する測定基準を作成
- ☑ 上記測定基準に対する確認に基づいて、データパイプラインとジョブの動作を適合させる
- ☑ 測定基準と確認を使用して、ML モデルの潜在的な経時ドリフトを測定

モニタリング

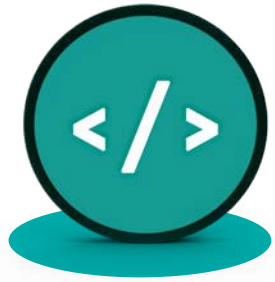
- ☑ 本番シナリオのステータスを追跡
- ☑ Dataiku ジョブの成功とエラーを視覚化

シナリオ

- ☑ データフローとアプリケーションの実行を、スケジュールベースまたはイベントベースでトリガー
- ☑ （ステップを）実行するための一連のアクションを組み立てて、カスタム実行シナリオ全体を作成
- ☑ Python API を通して、内蔵ステップを活用するか、独自のステップを定義
- ☑ シナリオの結果を複数のチャンネルにレポーター経由で公開（カスタムテンプレートを使い、データセット、ログ、ファイルまたはレポートを添付してレポーターに送信し、通知を Slack または Hipchat に送信）

自動化環境

- ☑ 本番用パイプライン専用の Dataiku 自動化ノードを使用
- ☑ 本番システム（データレイク、データベース）に接続して展開
- ☑ 複数の Dataiku プロジェクトのバンドルをアクティブ化、使用、または元に戻す



コード

様々な作業にコードを使用でき、カスタマイズ可能です。馴染みのあるツールと言語で作業することができます。ビジュアルインターフェースを使用する方が簡単な作業では、コードとビジュアルインターフェースをシームレスに切り替えられます。

「レシピ」をコーディングするための多言語をサポート

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

カスタムコード環境の作成と使用

- ☑ 複数バージョンの Python（2.7、3.4、3.5、3.6）のサポート
- ☑ Conda に対応
- ☑ R および Python ライブラリーを Dataiku のインターフェースから直接インストール
- ☑ 任意の R または Python ライブラリーをインストールするためのオープンな環境
- ☑ パッケージの依存性を管理し、再現可能な環境を作成

スケールコードの実行

- ☑ オンプレミスの、あるいはクラウドサービス（EKS, AKS, GKE）経由で Kubernetes クラスタに Python または R のジョブを送信し、コードを拡張

・データサイエンティスト用のインタラクティブノートブック

- ☑ Python、R または PySpark カーネルでの Jupyter ノートブックの完全統合
- ☑ テンプレート作成済ノートブックを使用して作業をスピードアップ
- ☑ SQL ノートブック（Hive に対応）を使ってデータベースまたはデータレイクをインタラクティブにクエリ
- ☑ Jupyter ノートブックを Kubernetes で実行

・Python & R ライブラリー

- ☑ 独自の R または Python ライブラリーまたはヘルパーの作成
- ☑ 上記をすべての Dataiku インスタンス内で共有
- ☑ 既存のコードアセットを簡単に使用
- ☑ 開発ワークフローを合理化する Git 統合のメリット

・再利用可能なカスタムコンポーネントの作成

- ☑ 技術力の低いユーザーのために、Dataiku プラグインで複雑なコードベースの関数をビジュアルインターフェースにパッケージ化
- ☑ Dataiku 生来の機能をコードベースのプラグインで拡張（カスタムコネクター、カスタムデータ準備プロセッサー、インタラクティブ分析および視覚化用のカスタムウェブアプリケーションなど）
- ☑ 使用する Dataiku シナリオ用に、Python ベースのカスタムステップを作成

・API

- ☑ Dataiku プラットフォームを CLI または Python SDK 経由で管理
- ☑ ML モデルをプログラムでトレーニングして展開
- ☑ カスタム Python & R 関数を REST API 経由で公開

・お気に入りの IDE を使用してコードを開発およびテスト

- ☑ R コード用 RStudio
- ☑ Sublime Text
- ☑ VS Code
- ☑ PyCharm

コラボレーション

Dataiku はコラボレーションを念頭に置いて、ゼロから作り込まれています。知識の共有から変更管理、さらにモニタリングまで、データチーム（データサイエンティスト、エンジニア、アナリストなど）は、より迅速によりスマートに一緒に作業できます。

共有プラットフォーム（データサイエンティスト、データエンジニア、アナリストなど）

バージョン管理

- ☑ Dataiku で加えられたすべての変更を記録する、Git ベースのバージョン管理

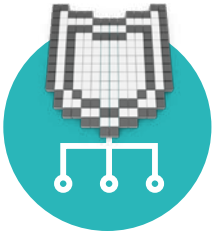
知識の管理と共有

- ☑ Wiki を作成してプロジェクトを文書化
- ☑ ディスカッション機能によりプラットフォームの他のユーザーとコミュニケーション
- ☑ Dataiku オブジェクトへのタグやコメント付け、お気に入り登録

チーム活動のモニタリング

- グローバル検索による、すべてのプロジェクトアセット、プラグイン、Wiki、参照文書などの素早い検索
- カスタマイズされたコードベースの機能を、ビジュアルインターフェースを使って技術力の低いユーザーと共有
- コードベースのコンポーネントの共有

- ☑ 全ユーザーに対して、再使用可能なコードスニペットを配布
- ☑ 技術力が低いユーザーが使用できるように任意の複雑な関数、操作、またはビジネスロジックをパッケージ化
- ☑ Github などのリモート Git レポジトリに統合



ガバナンスとセキュリティー

Dataiku により、データガバナンスが簡単になり、細やかなアクセス権設定、および管理者またはプロジェクトマネージャー用の高度なモニタリングによって、エンタープライズレベルのセキュリティーが実現されます。

ユーザープロファイル

・ロールベースのアクセス権（細やかな設定またはカスタム設定）

・認証管理

- ☑ SSO システムの使用
- ☑ コーポレートデータベース（LDAP、アクティブディレクトリなど）に接続して、ユーザーおよびグループを管理

・エンタープライズグレードのセキュリティー

- ☑ 監査証跡を使って、Dataiku での全アクションを追跡しモニター
- ☑ Hadoop クラスタとデータベースを Kerberos で認証
- ☑ 全体に対するトレーサビリティーとコンプライアンスで、ユーザーのなりすましに対応

・リソース管理

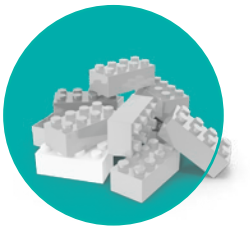
- ☑ Hadoop クラスタを、Dataiku からダイナミックに開始および停止
- ☑ サーバーリソースの割り当てを、ユーザーインターフェースから直接コントロール

・プラットフォーム管理

- ☑ Dataiku CLI および API を使って、利用中のコーポレートワークロード管理ツールを統合

データ保護および外部規制順守のためのカスタムポリシーフレームワーク

- ☑ GDPR 規則とプロセスを直接実装
- ☑ フレームワーク機能
  - ◊ *機密情報が含まれているデータソースを文書化し、グッドプラクティスを実行*
  - ◊ *機密情報が含まれているプロジェクトやデータソースへのアクセスを制限*
  - ◊ *Dataiku インスタンス内の機密情報を監査*



アーキテクチャー

Dataiku は現代の企業向けに構築されており、そのアーキテクチャーにより企業は（特定の技術に捕われることなく）オープンな環境を維持し、データに関する取り組みを拡張できます。

- Dataiku ユーザーはクライアントをインストールする必要はありません
- Dataiku ノード（専用の Dataiku 環境またはノードを使用して、ML アプリケーションを設計、実行、展開）
- 統合
  - ☑ 分散システムを使用して Dataiku 全体にわたるコンピュテーションを拡張可能。
  - ☑ 自動的に Dataiku ジョブを SQL、Spark、MapReduce、Hive、または Impala のジョブに変換し、インクラスタやインデータベースでの処理での不要なデータの移動やコピーを回避
- 最新のアーキテクチャ（深層学習用の Docker、Kubernetes、GPU）
- システムログ全体のトレーサビリティーとデバッグ
- オープンプラットフォーム
  - ☑ Jupyter ノートブックのネイティブサポート
  - ☑ お好みの Python または R パッケージおよびライブラリーをインストールして管理
  - ☑ 既存の企業コードベースを自由に再使用
  - ☑ Dataiku プラットフォームをカスタマイズされたコンポーネントで拡張