*Group 13*

*DATA SCIENCE 3: BIG DATA MANAGEMENT | UNIVERSITY OF WATERLOO*

**Compare and Contrast Pentaho and Dataiku**

# Team Members

1. Garcia, Ronald, r3garcia
2. Bujokas, Michael, mbujokas
3. Beckett, Paul Kenneth, pkbecket
4. Batra, Gaurav, g6batra

# Table of Contents

[**Team Members**](#_ta53xqo92599)2

[**Table of Contents**](#_3agvtf9u376i)2

[**Objective**](#_70c6xmcbqrii)3

[**Exploration**](#_rpl6msvt66xs)3

[Pentaho](#_pz3udw12jo) 3

[Dataiku](#_i95z3j6317hd) 4

[**Analysis**](#_55cfb25k9ki6)5

[Pentaho](#_4wmjvse784f) 5

[Dataiku](#_8wfxwnp7ei5d) 7

[**Conclusions**](#_30j0zll)15

[Pentaho](#_c2abnvczkgt) 16

[Dataiku](#_hfahw13fr1p7) 16

[**References**](#_1fob9te)16

# Objective

To compare and contrast 2 Big Data tools, Pentaho and Dataiku, and identify the purpose of each one.

We have used sample data sets to explore the different functions available to data scientists across these tools. Although the data sets used are different across the tools assessed, the purpose of exploration of these tools and analysis stays the same.

# Exploration

Data science platforms have been around since 2017, but recently advanced analytics is driving the enterprise, and the concept is becoming a must have. A data science platform is the structure in which the entire lifecycle of a data science project takes place.

## Pentaho

For Pentaho the approach we used was to download a data set from Kaggle (*World Covid-19 Vaccination Progress*), pull the csv file into the Data Integration tool, run cleaning and transformations, show results, and dump data into a database.

Key Capabilities

Pentaho is a business intelligence system written in java. It offers:

* Analytics
* Visual data integration
* OLAP
* Reports
* Dashboards
* Data mining
* Extract, Transform and Load (ETL)

|  |  |  |
| --- | --- | --- |
| lt is offered as:   * Free community edition * Purchased enterprise edition | Available for:   * Windows * Linux * Mac OSX | Other features:   * Community supported * Open source plugins available * Uses the Apache Java Application Server |

The objective is to demonstrate how to build a data integration transformation and a job using the features and tools provided by Pentaho Data Integration (PDI). We used a combination of steps to cleanse, format, standardize, and categorize the sample data. The seven basic steps are:

1. Extract and load data
2. Filter for missing codes
3. Resolve missing data
4. Clean the data
5. Format data
6. Run the transformation
7. Orchestrate with jobs

Data Preparation

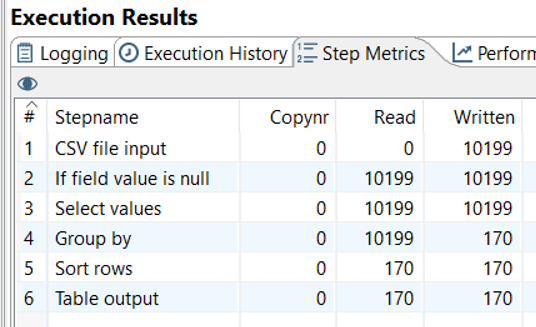
Kaggle data source: <https://www.kaggle.com/gpreda/covid-world-vaccination-progress>

Challenges with the data are as following:

* Data was missing because it’s recent and new data
* There is a duplicate row
* Data set is small for Big Data analysis

Data characteristics:

* Data is recent data up to early April
* Size: 2.06 MB; 15 Columns; 10,199 rows
* Data was poor quality because there were many null value in the dependent variables; every row had a null value (see figure below):



## Dataiku

Our process for exploring Dataiku was based on the current Dataiku Science Studio (DSS) capabilities.

Dataiku is the platform democratizing access to data and enabling enterprises to build their own path to AI in a human-centric way.

Key Capabilities

**Data Preparation:** The Dataiku visual flow allows coders and non-coders alike to easily build data pipelines with datasets, create recipes to join and transform datasets, and the ability to build predictive models. The visual flow also has code and reusable plugin elements for customization and advanced functions.

**Visualization:** Dataiku saves time with quick visual analysis of columns, including the distribution of values, top values, outliers, invalids, and overall statistics. For categorical data, the visual analysis includes the distribution by value, including the count and % of values for each value.

**Machine Learning:** To aid in the feature engineering process, Dataiku AutoML automatically fills missing values and converts non-numeric data into numerical values using well-established encoding techniques. Users can also create new features using formulas, code, or built-in visual recipes to provide additional signals to improve model accuracy. Once created, Dataiku stores feature engineering steps in recipes for reuse in scoring and model retraining.

**DataOps:** Dataiku projects are the central place for all work and collaboration for users. Each Dataiku project has a visual flow, including the pipeline of datasets and recipes associated with the project. Users can view the project and associated assets (like dashboards), check the project’s overall status, and view recent activity.

**MLOps:** The Dataiku unified deployer manages project files’ movement between Dataiku design nodes and production nodes for batch and real-time scoring. Project bundles package everything a project needs from the design environment to run on the production environment. With Dataiku, data scientists can see all the deployed bundles, and data engineers of IT operations can quickly know when a new bundle requires testing and roll-out.

**Analytic Apps:** Dataiku makes it easy to create project dashboards and share them with business users. Scheduling updates for dashboards or triggering updates is easy and ensures the latest information is available. With dashboards as part of a Dataiku project, business users and project stakeholders can easily see the outputs of AI projects and track KPIs and value.

**Collaboration:** Real advanced analytics projects require a series of steps that transform data from one state to the next, resulting in new datasets, features, metrics, charts, dashboards, predictive models, and applications. The Dataiku visual flow is the canvas where teams collaborate on data projects. With the visual flow, everyone on the team can use common objects and visual language to describe the step-by-step approach and document the entire data process for future users.

**Governance:** Dataiku permissions control who on the team can access, read, and change a project. Permissions also include creating projects, executing code, executing applications, reading only content, and more. With Dataiku, users can belong to more than one group and have different permissions across projects, or organizations can have global permissions.

**Explainability:** Dataiku provides critical capabilities for explainable AI, including reports on feature importance, partial dependence plots, subpopulation analysis, and individual prediction explanations. Together, these techniques can help explain how a model makes decisions and enable data scientists and key stakeholders to understand the factors influencing model predictions.

**Architecture:** Dataiku can run on-premise or in the cloud—with supported instances on Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure—integrating with storage and various computational layers for each cloud.

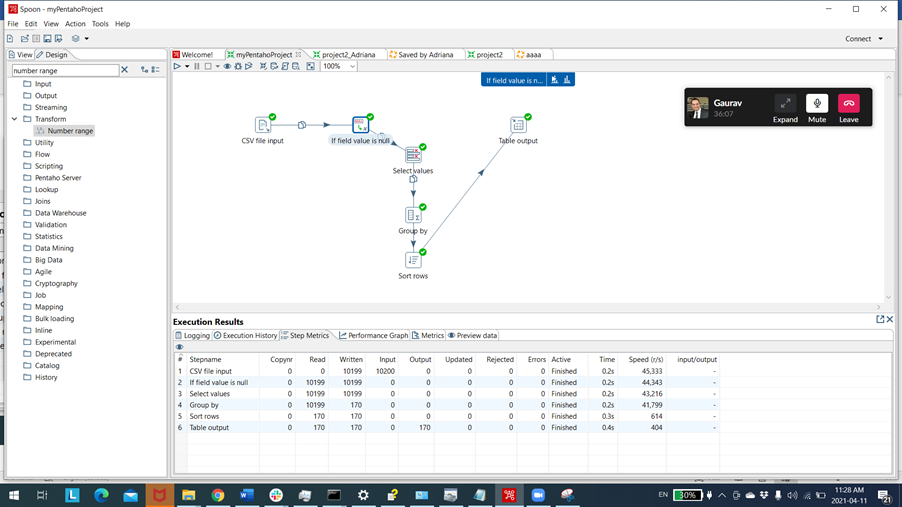
# Analysis

## Pentaho

The total vaccinations per country were highest in the US, China, and India for the top 3. Only 1 country had 0 vaccinations (Laos). It would have been interesting to see total vaccinations per capita but this metric was not available in the data set.

Experiment 1

**Load, Clean, Sort, Group, ETL (upload to Oracle DB)**



**Analysis:** With transformation functions in Pentaho Spoon, data can be sorted, grouped, and cleaned easily. The left hand navigation makes it possible to search for functions or expand categories such as Database or Big Data, which make it easy for the user to find what they are looking for.

Grouping/aggregate functions like *sum*, *average* and multiple others were found in the node dropdown and the options were rich in different types.

There are Database bulk loaders which make data migration simplified and integrated into one process/job.

Pentaho is able to clean data, select columns and group by and aggregate on daily vaccines across distinct countries.

**Clean Data:** We have removed nulls for integers by replacing by 0 and removed the *<null>* values from string by replacing with empty string. Pentaho has UI features per field and per type of field to perform the latter transformation.

**Select Columns:** The tool removes a subset of fields we were not interested in seeing via *Remove Fields* from the data tab. In this case we removed all fields except country and daily vaccinations to see the countries with the highest total vaccinations.

**Grouping:** The data was grouped into 2 columns by country and aggregate sum on daily vaccinations.

**Upload to Oracle Database:** The 170 rows of data was plugged into an existing Oracle database table in milliseconds. The data could be migrated or queried in the future.

Experiment 2

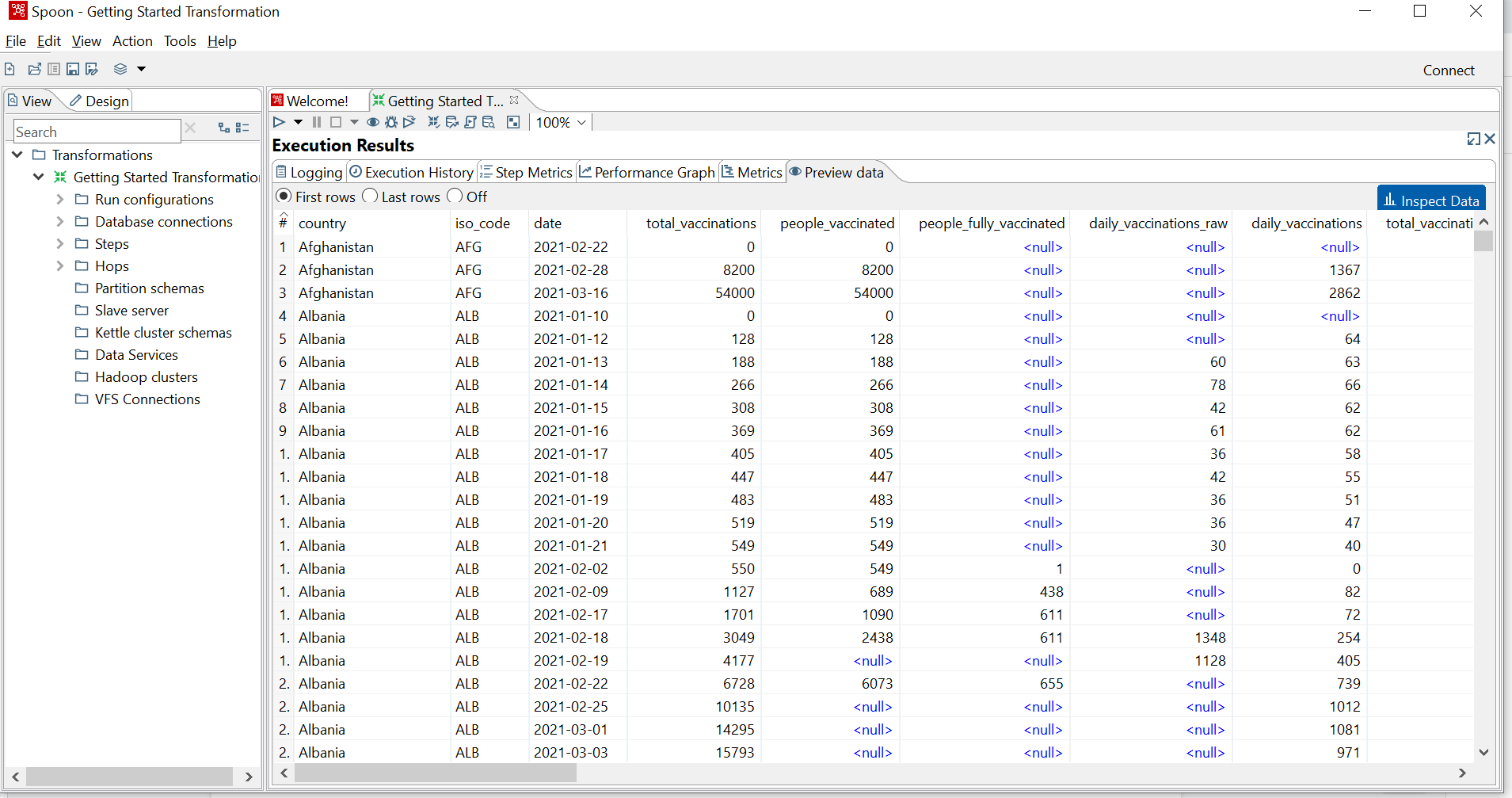
**Extract and load data, add a transform component to the data pipeline (filter missing and resolving data, apply formatting, set value properties, ranged data), and schedule the job to run the transformation.**

**Extract and Load Data:** Use Spoon PDI window to create a new transformation, retrieve data from a csv flat file and use the *Text File Input* step to: connect to a repository, view the file schema, and retrieve the data contents. Create connection type as *Hypersonic* and load data into DB table output.

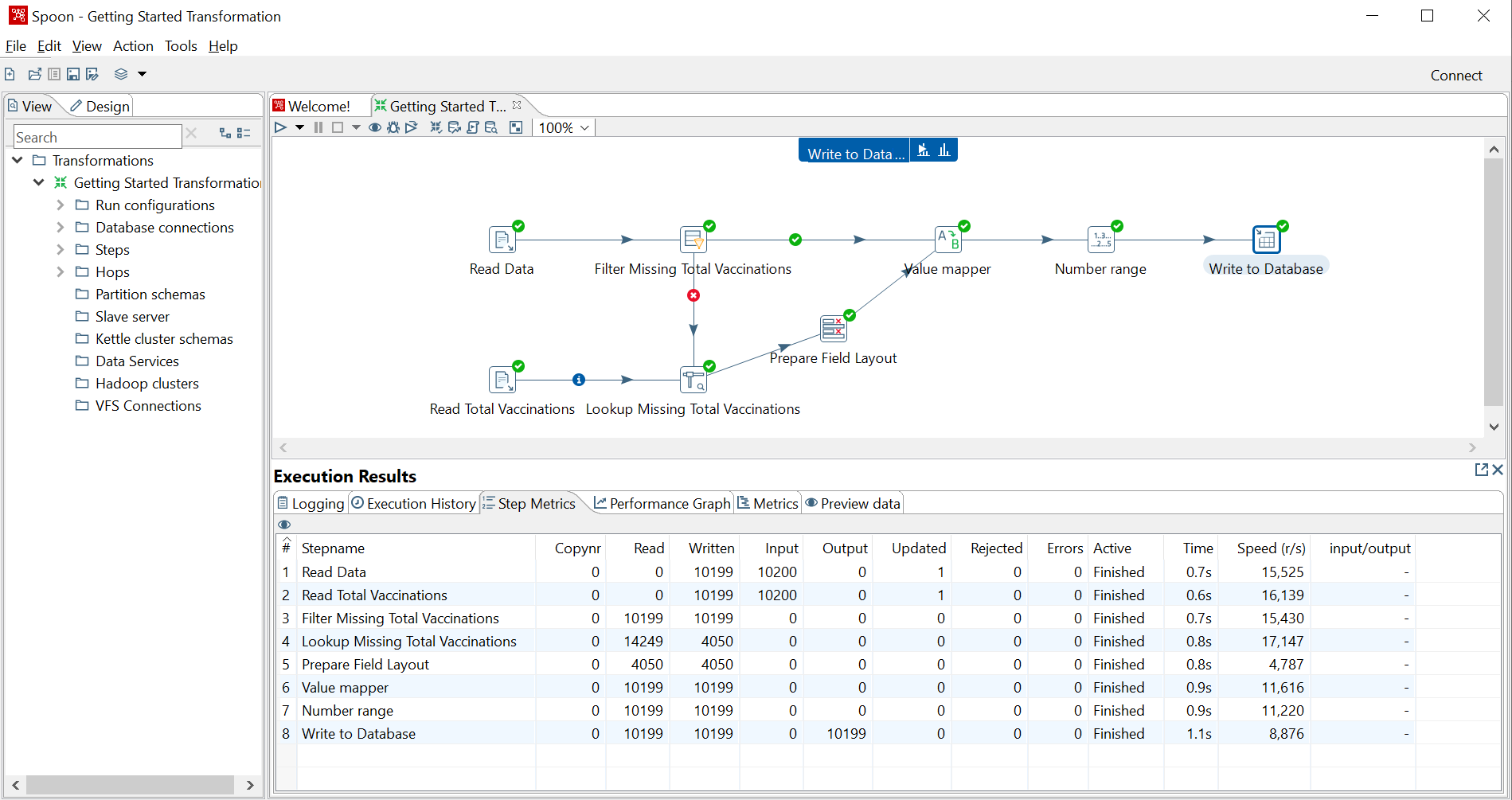
**Add Transformation:** The source file contains several records that are missing *total\_vaccinations*. By filtering the records with missing data into stream lookup, the data can be resolved. Apply formatting, set the value properties, range data to the transformation and further cleanse and categorize the data into buckets before loading it into a relational database.

**Orchestrate with Jobs:** Create a job that will verify the file has arrived and then run the transformation to load the records into the database.

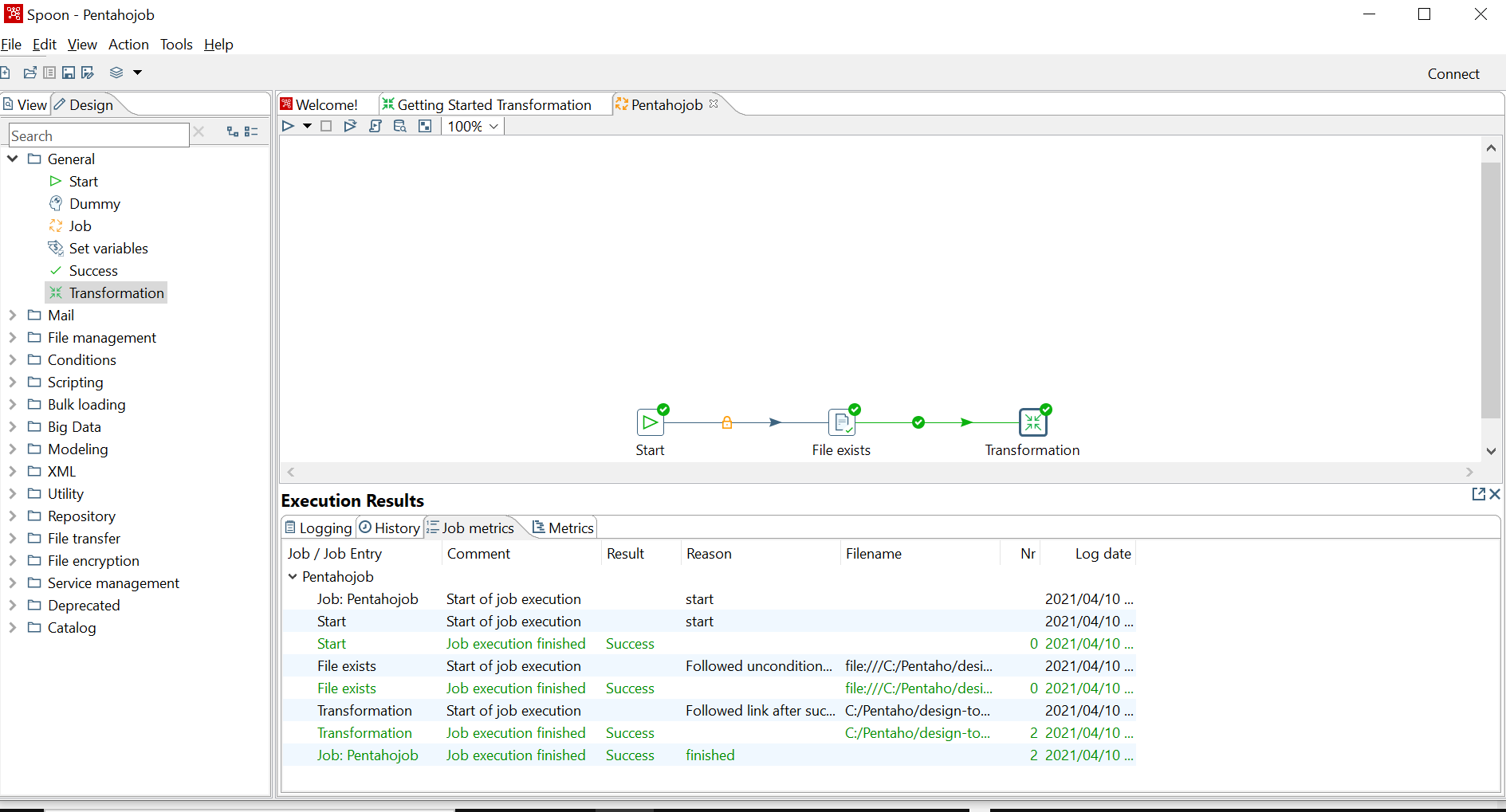
Data Before Clean of Null



Transformed Data



Job Pipeline



## Dataiku

Dataiku provides a solution (DSS) that enables an incremental approach to allow advanced analytics users and data scientists to use the same tool and develop on their own learning pathways, while delivering insights to the right audience.

Dataiku Model pipelines, once built, can be easily automated which helps to expedite the expected deliverables.

One of the biggest benefits of the Dataiku Science Studio (DSS) is a user with no coding experience can work in this tool together with a user who wants a more flexible experience through integrating their own code. Dataiku allows users to connect to make sense of the data to work towards enterprise-wide standardization and eliminate technology debt.

Dataiku DSS offers data integration to more than 30 connectors and has customized plugin extensions that link users to their current system. DSS can also detect schema and data formats and promotes computation to Spark, Hadoop, and SQL systems. A clear profile of each organization’s information at every area of analysis helps users to search, prepare, collate, and clean data effectively. DSS also takes advantage of current Machine Learning tools like XGboost and Scikit-Learn.

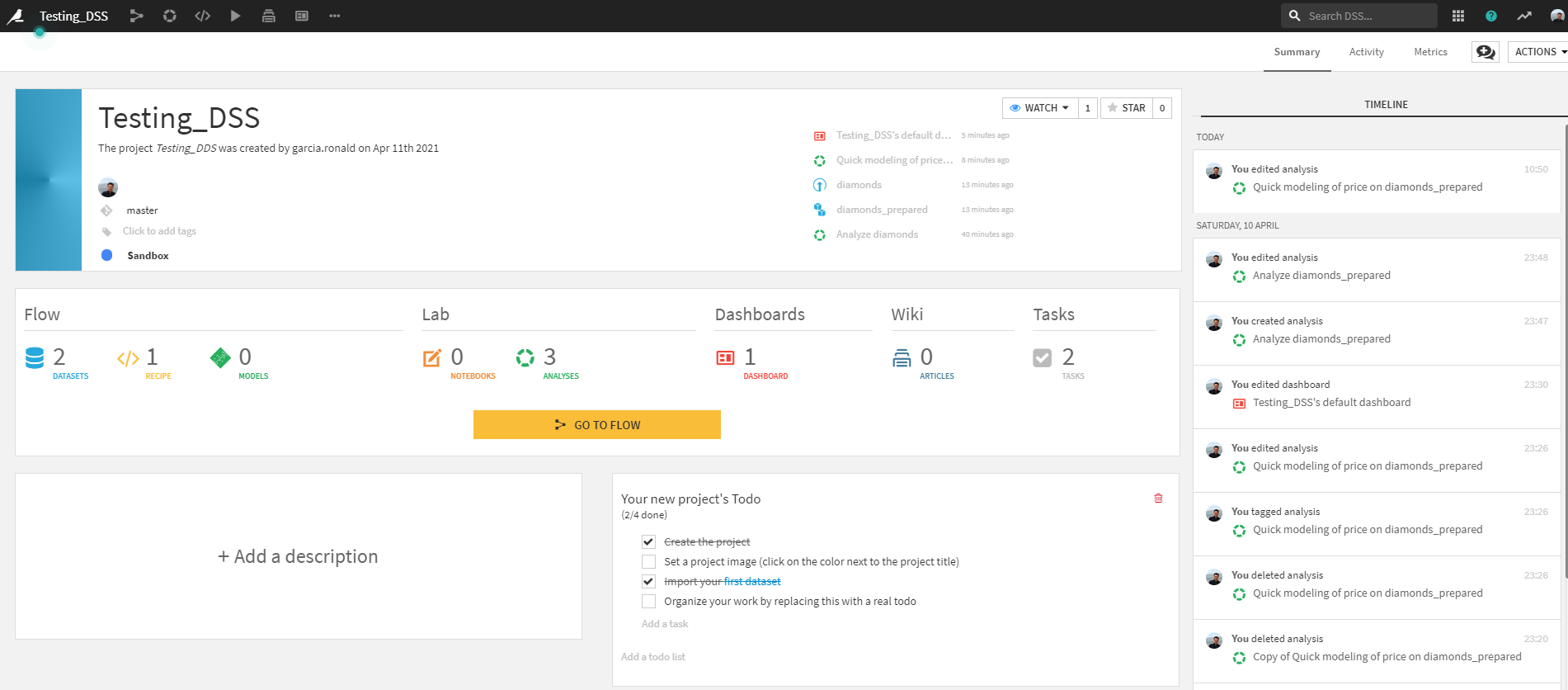
DSS has a dashboard functionality that makes it simple for users to generate visualizations and interactive charts from their datasets. The dashboard consists of more than 20 pre-installed chart formats while allowing drag-and-drop data to compute changes automatically on today’s Big Data infrastructure, which significantly improves performance.

Dataiku and Diamonds

**Academy and Community:** Before proceeding with the using Dataiku Science Studio (DSS) for our Assignment 4 – Diamonds dataset. We want to highlight different resources regarding product documentation:

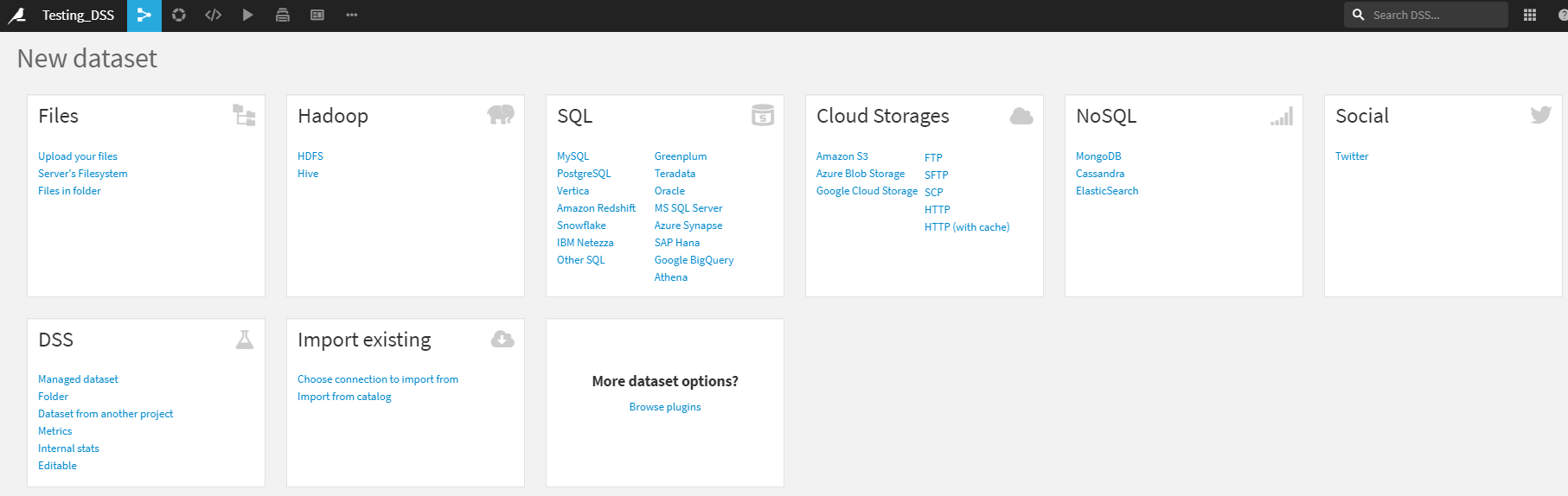
* The Knowledge Base has a variety of topics that can help you to learn more about Dataiku DSS or find solutions to problems without having to ask for help.
* Dataiku Academy provides guided learning paths for you to follow, upskill, and gain certification on Dataiku DSS.
* Dataiku Community is a place where you can join the discussion, get support, share best practices, and engage with other Dataiku users.

**Initial Project - Testing\_DSS:** The first step is to create a project. A Dataiku DSS project is a container for all your work on a particular activity. See screenshot below with the created project - Testing\_DSS.



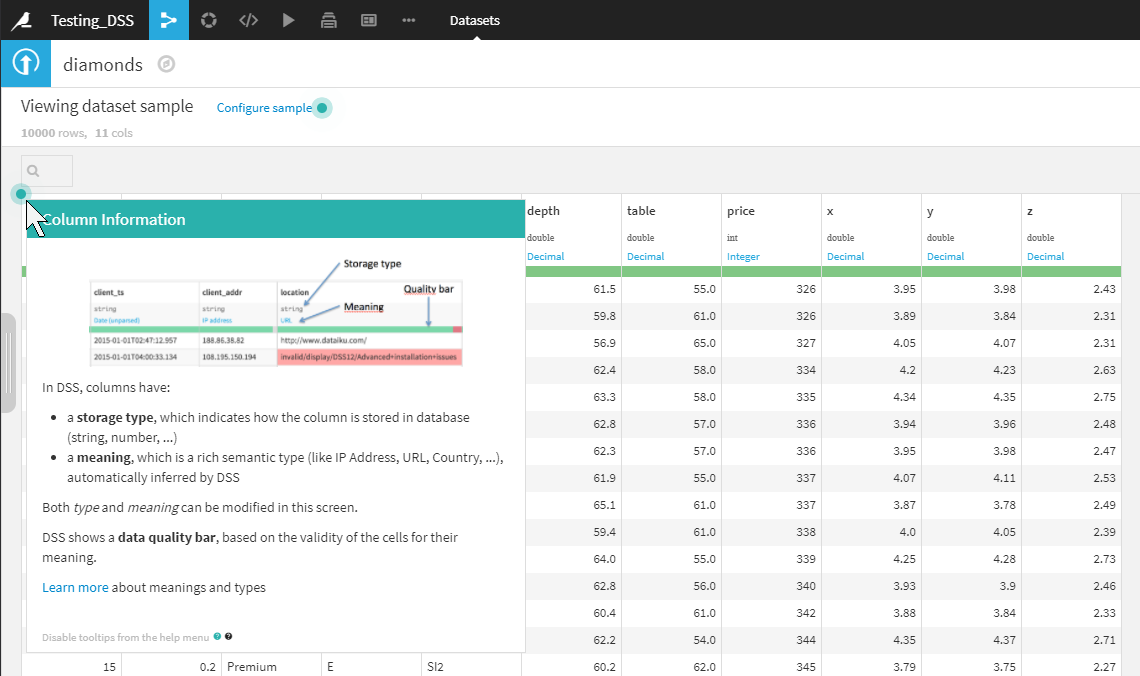
**Importing the Diamonds Dataset:** Many different options to import a new dataset. One of the easiest ways to create datasets in DSS is to upload your files to the DSS server.

After upload, DSS automatically parses and detects format, settings, and schema for your new dataset.

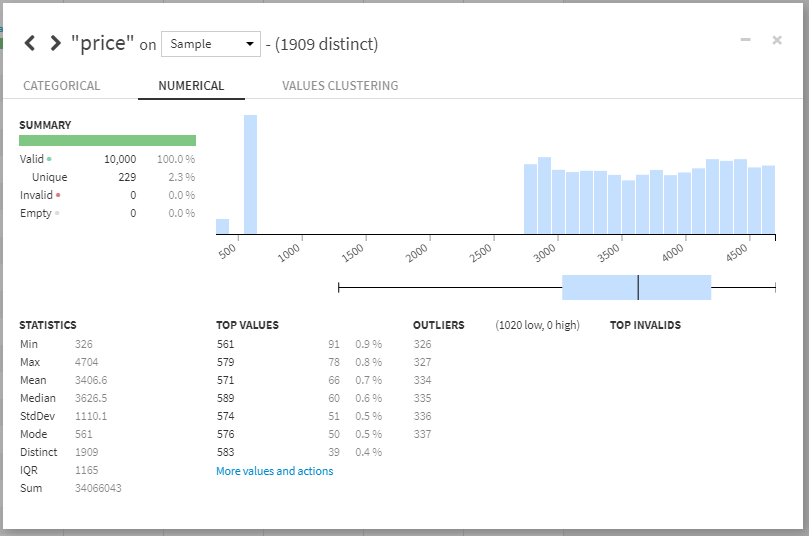


**Data Exploration:** As part of the DSS features, there is an interesting concept, *Quality Bar*. By default, the explore view shows a data quality bar showing which rows are valid for their meaning. When you are in data exploration, meanings are mostly useful for informational purposes.

If you choose to filter a column (click on column header then on *Filter*), the filter shows a quality bar with checkboxes that allow you to focus on rows that are valid, invalid or empty for their meanings.



Another useful feature, *Analyze per column*. In the *Explore* tab of your dataset, you can do some exploratory analysis by clicking on a column and choosing *Analyze* from the dropdown menu. Image below shows *Analyze* results for the price column.

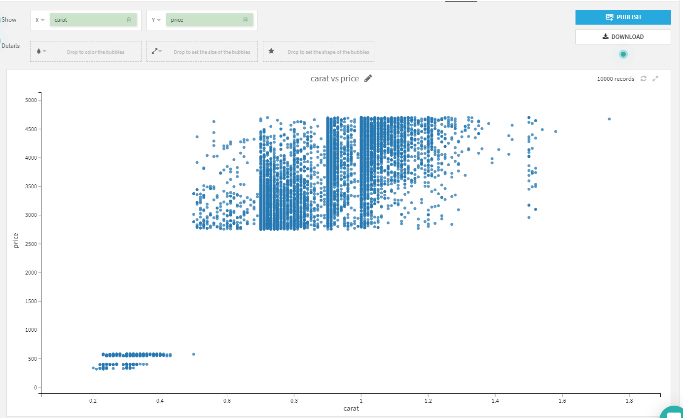
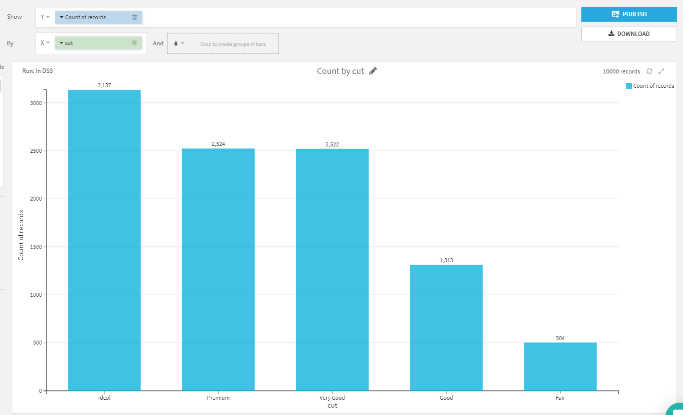


**Creating Charts:** We created two charts as part of the data exploration.

1. Count by depth and price (histogram)
2. Carat and Price (scatter plot)

DSS delivers an advanced data visualization engine through the Charts tab of a dataset for visual analysis. The chart-building interface is essentially the same in both locations, with the following important caveats.

* Charts in a visual analysis can work in real-time on the output of a data preparation script. Instead of rebuilding a dataset, simply add a step to the script and view the result immediately.
* Charts in a dataset can be published as insights for inclusion in dashboards, while charts in a visual analysis cannot. However, when a visual analysis is deployed as a *Prepare* recipe, its charts can be transferred during deployment to the output dataset.
* Charts in a dataset can make use of the in-database execution engine, while charts in a visual analysis are always run in the DSS engine.



**Options for Statistical Analysis:** An interactive statistics worksheet in Dataiku DSS provides a dedicated interface for performing exploratory data analysis (EDA) on datasets. Using this feature, you can:

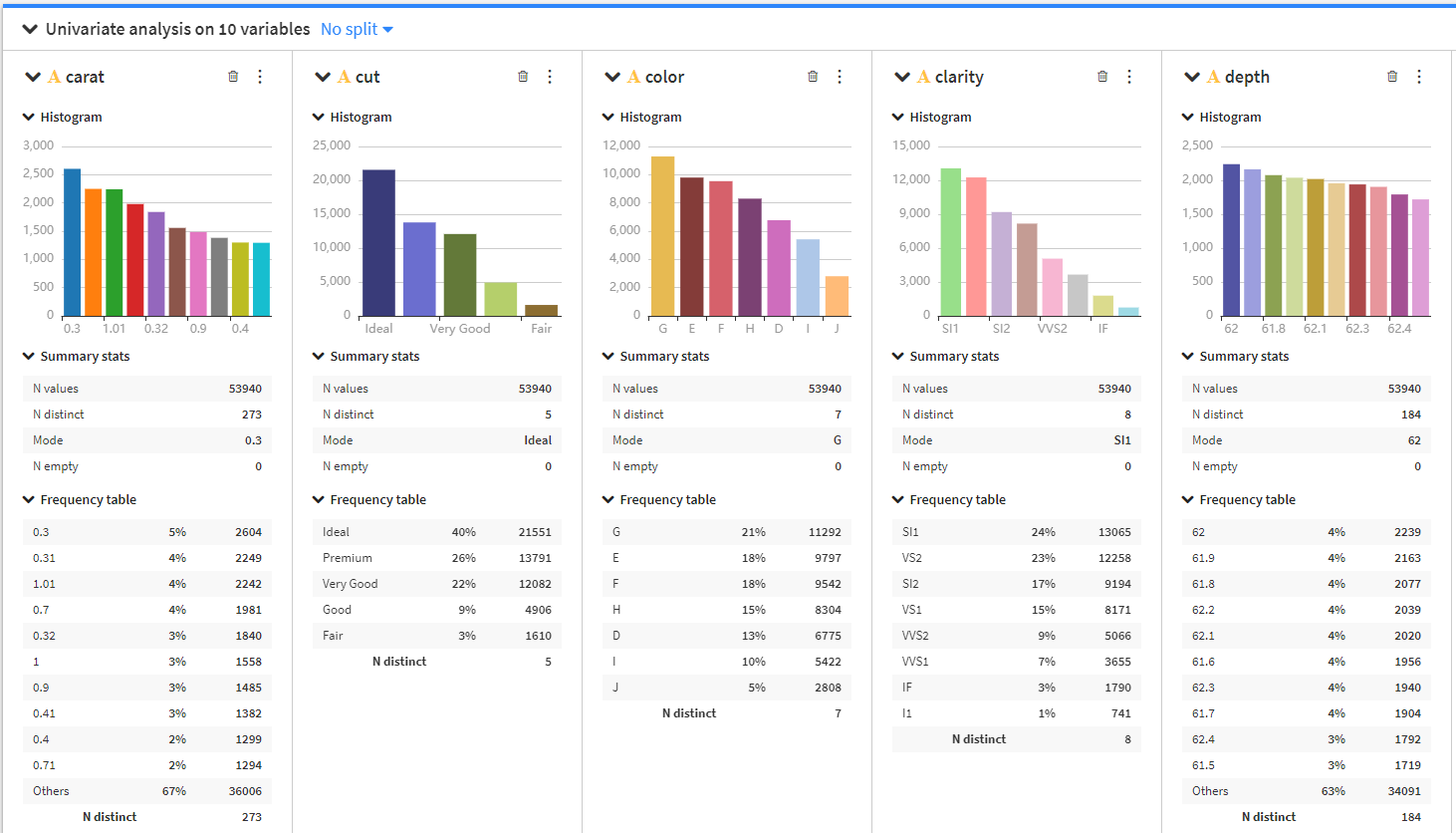
* Summarize or describe data samples, e.g. using univariate analysis, bivariate analysis, distribution and curve fitting, and correlation matrices. This falls under the area of descriptive statistics.
* Draw conclusions from a sample dataset about an underlying population, e.g. using hypothesis testing. This falls under the area of inferential statistics.
* Visualize the structure of the dataset in a reduced number of dimensions, using principal component analysis. This falls under the area of dimensionality reduction.

We created two cards for exploratory data analysis in the diamonds dataset.

1. Univariate analysis
2. Correlation matrix

Univariate analysis

Univariate analysis is useful for exploring a dataset one variable at a time. This kind of analysis does not consider relationships between two or more variables in your dataset. Rather, the goal here is to describe and summarize the dataset using a single variable.



Correlation Matrix – Pearson

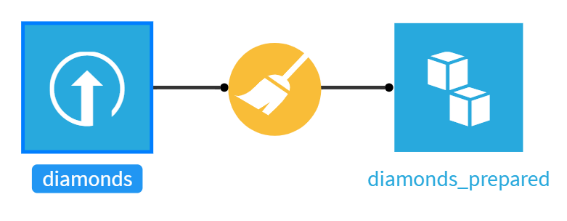
The Correlation matrix option allows you to examine the degree to which pairwise relationships may exist for variables in the dataset.



**Creating a *Flow*:** In DSS, the datasets and the recipes together make up the flow.

Using the flow, DSS knows the lineage of every dataset in the flow. DSS, therefore, is able to dynamically rebuild datasets whenever one of their parent datasets or recipes has been modified.

Flow was unnecessary in this analysis; to explore the functionality we removed the id column from the original dataset and created a new dataset (diamonds\_prepared).

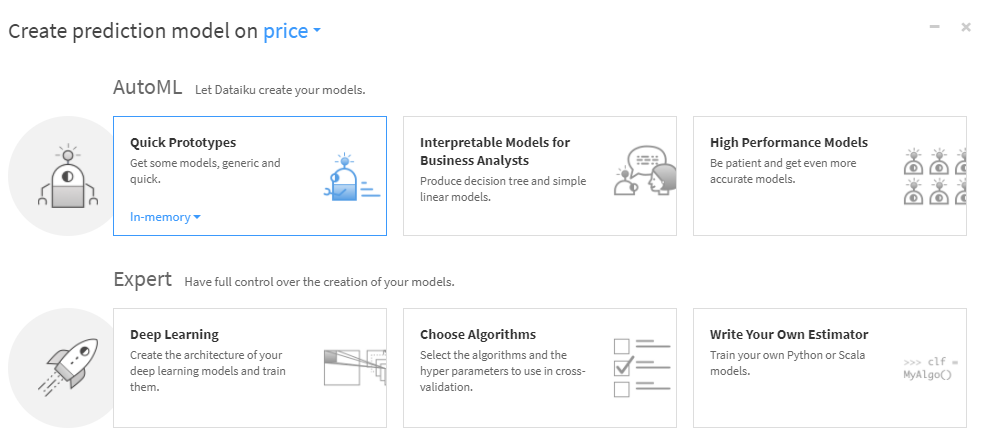


**Letting Dataiku Create Models – AutoML:** DSS contains a powerful automated machine learning engine that allows the user to get highly optimized models with minimal intervention.

At your choice with DSS you can select between:

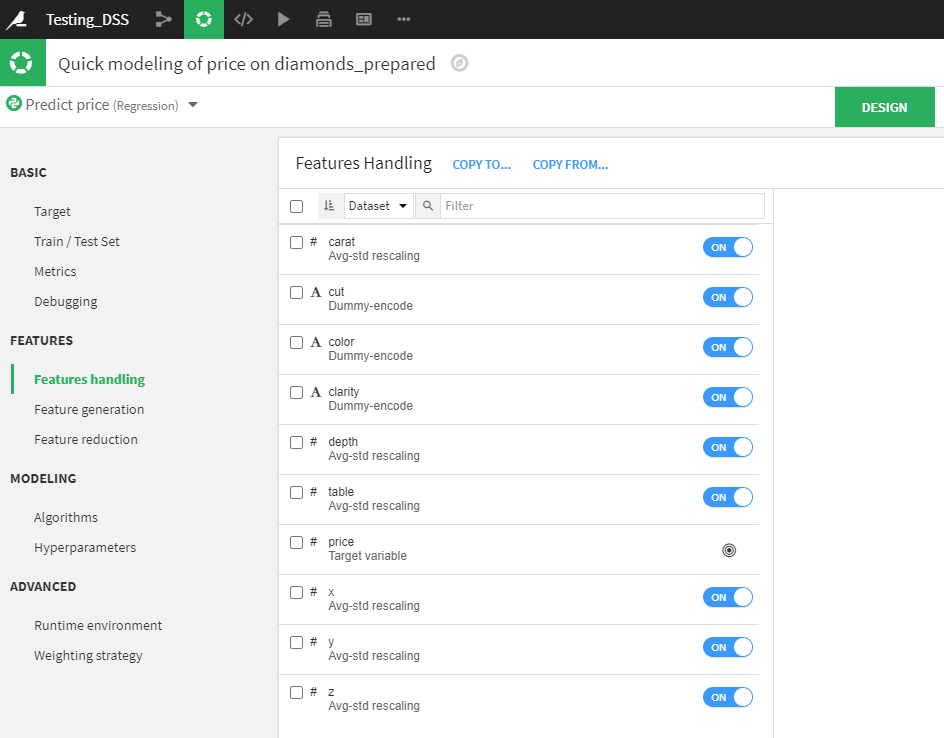
* Having the full control over all training settings, algorithm settings and optimization processes, including writing your own custom models and using advanced deep learning models
* Using DSS powerful automatic machine learning engine to effortlessly get models

The Automated Machine Learning engine of DSS will analyze your dataset, and depending on your preferences, select the best features handling, algorithms, and hyper parameters.



**Feature Handling:** Categorical features need to be encoded as numbers so the machine learning algorithm can understand them.

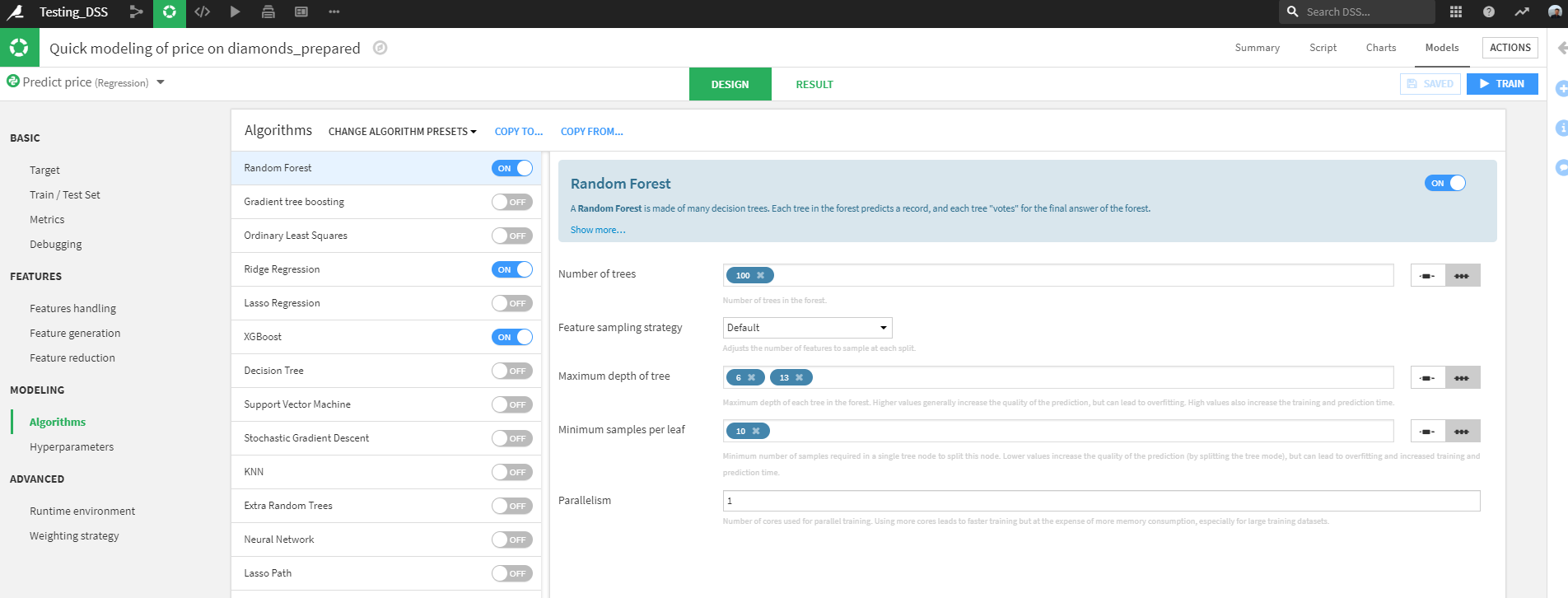
Categorical encoding for cut, color, and clarity variables.

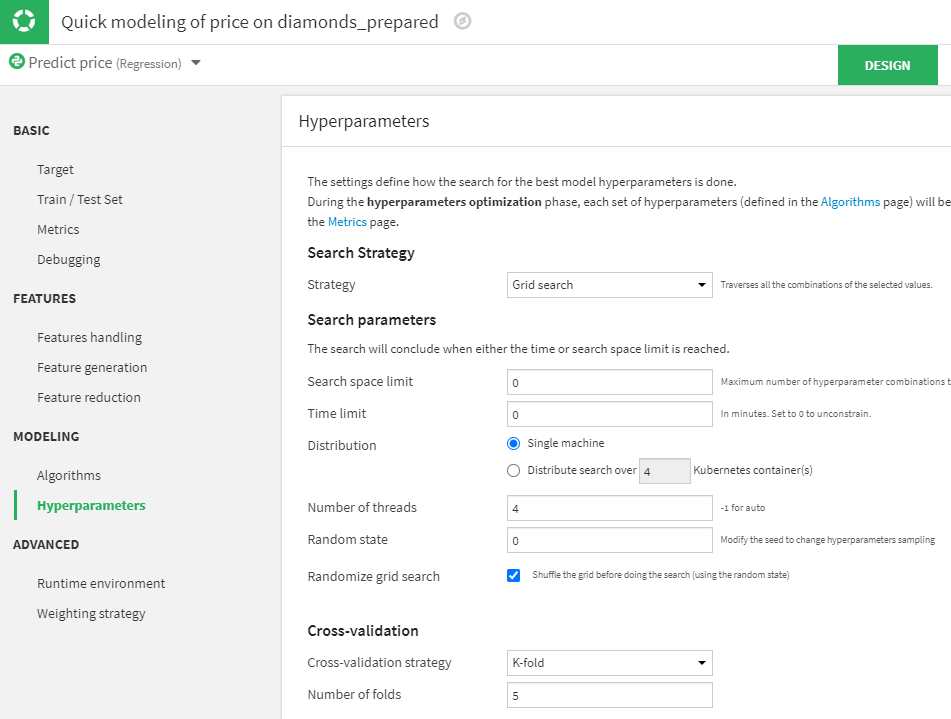


**Modeling – Algorithms and Hyperparameters:** DSS’s visual machine learning comes with support for 4 different Machine Learning engines:

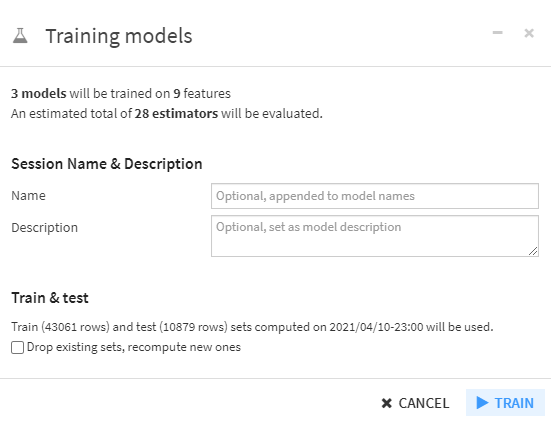
* In-memory Python (Scikit-learn/XGBoost)
* MLLib (Spark) engine
* H2O (Sparkling Water) engine

Each time you create a new machine learning model in DSS, you can select the corresponding training engine. The models will be trained with this engine. Once trained, models can be applied to new records to make predictions. This is called scoring and is handled by various scoring engines.

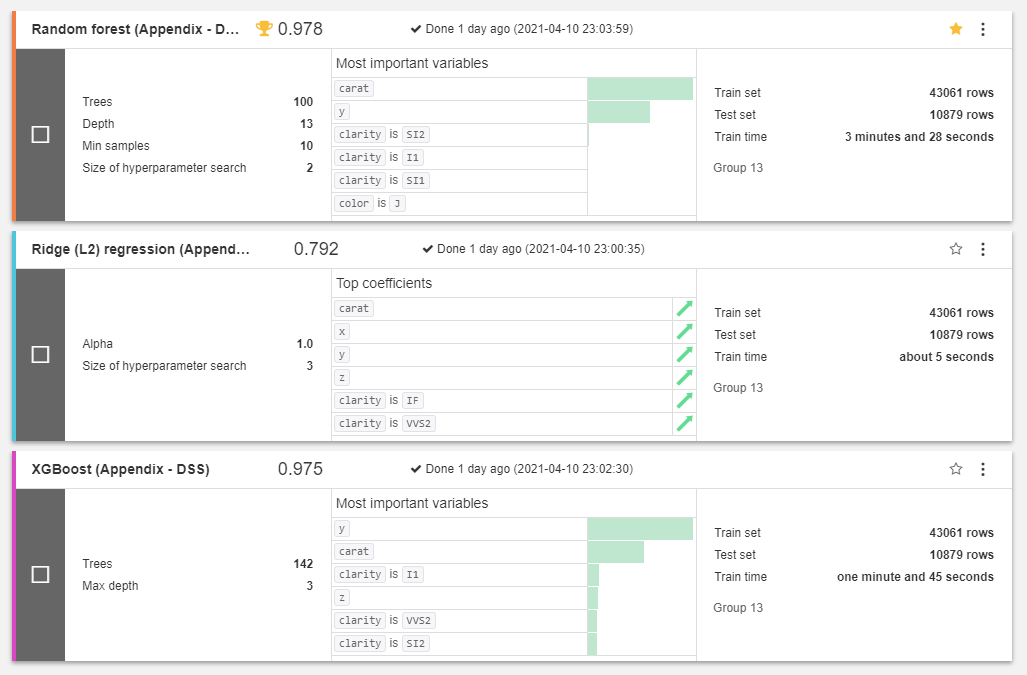
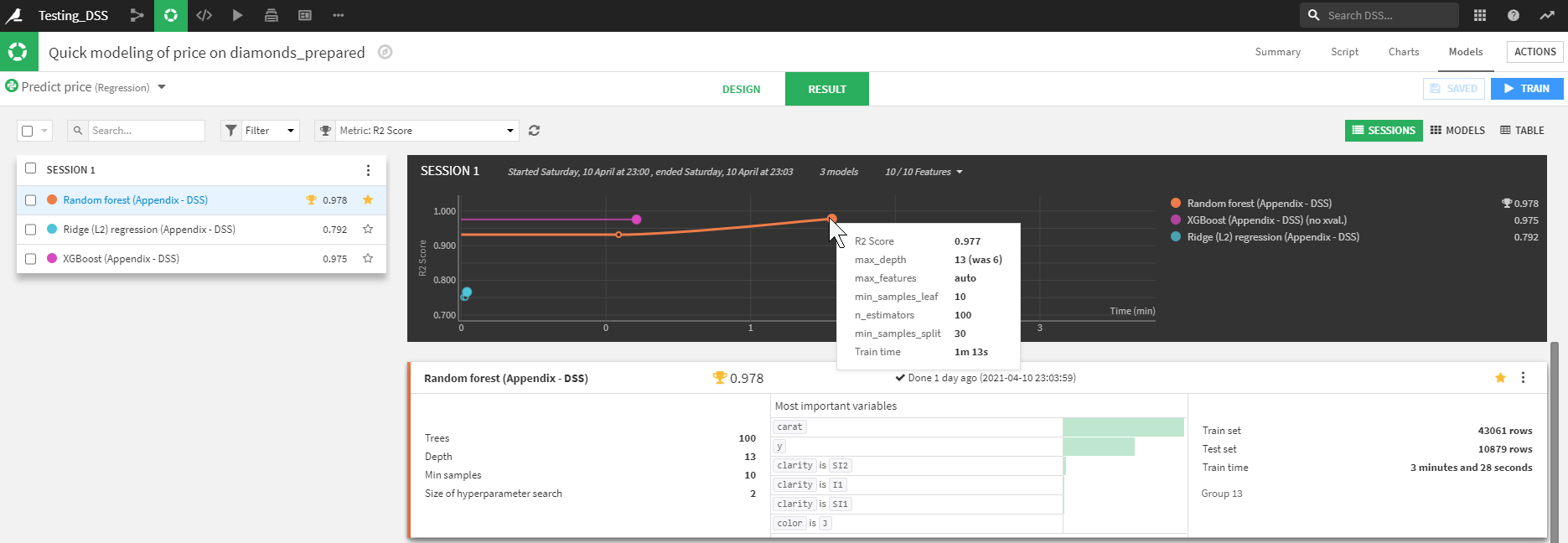




Training models for our diamonds\_prepared dataset



Model Results



From the results, the best R2 score was 0.978 for the Random Forest algorithm.

**Dashboard:** We created a dashboard with some of our data analysis artifacts.

Dashboards allow you to share elements of your data project, either with other analysts working on the project, or with users who do not have full access to the project.

Dashboard



# Conclusions

We recommend testing two or more types of architecture to ensure complete compatibility across any organization as part of the analysis. Businesses have distinctive business-related requirements, it is only wise they steer clear of purchasing a one-size-fits-all, ideal software product.

When testing different types of architecture/platforms, organizations should ensure they partner with both technology and the business to develop Big Data use cases and real-business cases so that they have a strong vision for the platform.

One aspect that was beyond the scope of this report is the support service quality of the selected platforms.

## Pentaho

The initial aim of this project was to compare and contrast the various data integration tools. While working on Pentaho Data Integration (aka Kettle), we learned many different concepts such as:

* Spoon – GUI Tool to develop all jobs & transformations
* DI Server (Server Application) – Data integration server executes jobs and transformations using PDI engine. Here, we can store the transformations and jobs stored at one common place.
* Design Tool (standalone) – It is for designing jobs and transformations

It is relatively easy to build complex operations, using the “blocks” Kettle makes available. A Kettle job contains the high level and orchestrating logic of the ETL application, the dependencies and shared resources, using specific entries. Each entry is connected using a hop, that specifies the order and the condition (can be “unconditional”, “follow when false” and “follow when true” logic). A job can contain other jobs and/or transformations, that are data flow pipelines organized in steps.

The major drawback using a tool like this is, logic will be scattered across jobs and transformations and could be difficult, at some point, to maintain the “big picture”. At the same time, it’s an enterprise tool allowing advanced features like parallel execution, task execution engine, detailed logs, and the possibility to modify the business logic without being a developer.

## Dataiku

Our overall approach regarding Dataiku can be summarized as follows:

* Start with the business question and then build a solution to help the business answer the question. This approach for the first few projects will ensure that the business can see the value.
* Access to the data is key. Dataiku is a great solution for transforming, manipulating, and processing data but you need to have good access and understanding of the data you have to enable the greatest insights. Empower users to get the most from the data sources they need and understand most, and really experiment using Dataiku to get the insights.

Dataiku has broad applications for machine learning with AutoML which is tailored for a user with little experience, while also allowing for advanced customization and integration with Python and Scala code. It also supports collaboration across an organization.

# References

1. Dataiku DSS - <https://doc.dataiku.com/dss/latest/>
2. 2021 Magic Quadrant for Data Science and ML Platforms - <https://pages.dataiku.com/gartner-2021>
3. Win your colleagues over with Dataiku - <https://towardsdatascience.com/win-your-colleagues-over-with-applications-in-dataiku-8-0-19ddd9d9754f>
4. Compare Data Science and Machine Learning Platforms - <https://www.g2.com/categories/data-science-and-machine-learning-platforms>
5. Pentaho - Kaggle data source: <https://www.kaggle.com/gpreda/covid-world-vaccination-progress>