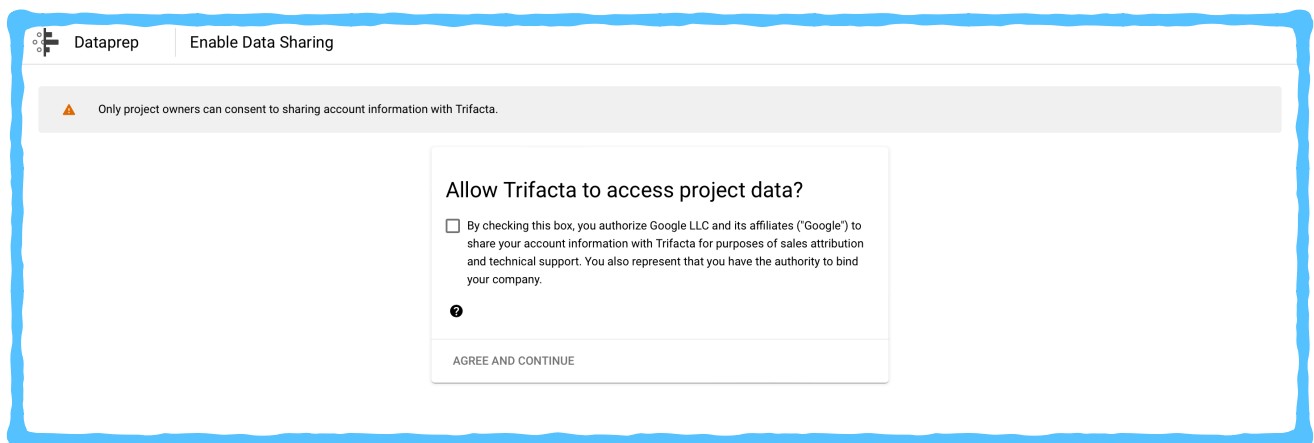


Cloud Dataprep by Trifacta

- An intelligent cloud data service to visually explore, clean, and prepare data for analysis and machine learning.
- Build by Trifacta – Third Party tool, not cloud native one
- Play with this tool without any code, with just click
- Dataprep is serverless and works at any scale
- No infrastructure to deploy or manage
- Automatically detect schema and anomalies
- Concern – Need to share data with Trifacta



What is Dataprep by Trifacta?

- Dataprep by Trifacta enables you to explore, combine, and transform diverse datasets for downstream analysis.
- Data preparation (or data wrangling) has been a constant challenge for decades, and that challenge has only amplified as data volumes have exploded.

Why use Dataprep by Trifacta?

- Estimates vary, but something like 60% of an analyst's time is consumed with preparing data for use, leaving two days per week to actually analyze it. That's expensive and inefficient.
- Some organizations have pushed these cleansing efforts onto IT, which may take weeks to come up with a custom, scripted solution that requires inevitable back-and-forth between coder and analyst. .
- The Dataprep by Trifacta solution delivers the tools to wrangle data to the people who understand the meaning of the data.
- With Dataprep by Trifacta, analysts have the means to apply their expertise to the preparation of the data, in a way that is faster and more productive.
- Dataprep by Trifacta helps to do the following:
 - Cut time to prepare actionable data
 - Avoid IT bottlenecks and reliance on data scientists in data prep
 - Deliver tools to prepare data to the people who understand the data
 - Eliminate manual prep work
 - Surface data quality issues in a way that's easy to fix them
 - Featuring a leading-edge interface, powerful machine intelligence, and advanced distributed processing, Dataprep by Trifacta renders the time-consuming, complex, and error-prone process of preparing datasets of any volume into a point-and-click exercise.

Basic Steps

- Users interact with the platform through the user interface to complete the following tasks:
 - Create connections to sources of data.
 - Through those connections, create references to specific datasets (files and tables).

- Load samples of data into the user interface.
- Build sequences of steps to transform the sampled data through innovative user interface controls.
- Execute a job to perform that sequence of steps across the entire dataset, yielding an output dataset delivered to the preferred destination in the proper format.
- Export the results from the platform.

Dataprep ELT pipeline architecture



Lab on Cloud Dataprep

Pricing of Cloud Dataprep

- Dataprep comes into two edition.
 - Starter edition
 - Professional edition

- Pricing is split across two variables;
 - **Design** – is priced on a per-project basis for an unlimited number of users.
 - **Execution** – consists of the Dataflow usage for running jobs in Dataprep.

Advantages of Cloud Dataprep

- Data preparation is nothing but a process of exploring, cleaning, and modifying raw and messy data so that it can be used for reporting, analysis, and evaluation of machine learning algorithms.
- It is done to prepare and structured or unstructured data ready for implementation.
- With Cloud Dataprep there is no need to worry about the size of the data file.

Limitations Of Cloud Dataprep

- Sort transform is not supported.
- It doesn't support user-defined functions.
- It doesn't support custom dictionaries and data types.
- User access to administrative functions is not supported.
- There may also be some limitations to file formats.

Resources

<https://docs.trifacta.com/display/DP/Product+Overview>