Data Wrangling

Jennifer Morales (C)
Contingent Worker

Definition

Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. The goal of data wrangling is to assure quality and useful data. Data analysts typically spend much of their time in the process of data wrangling compared to the actual analysis of the data.

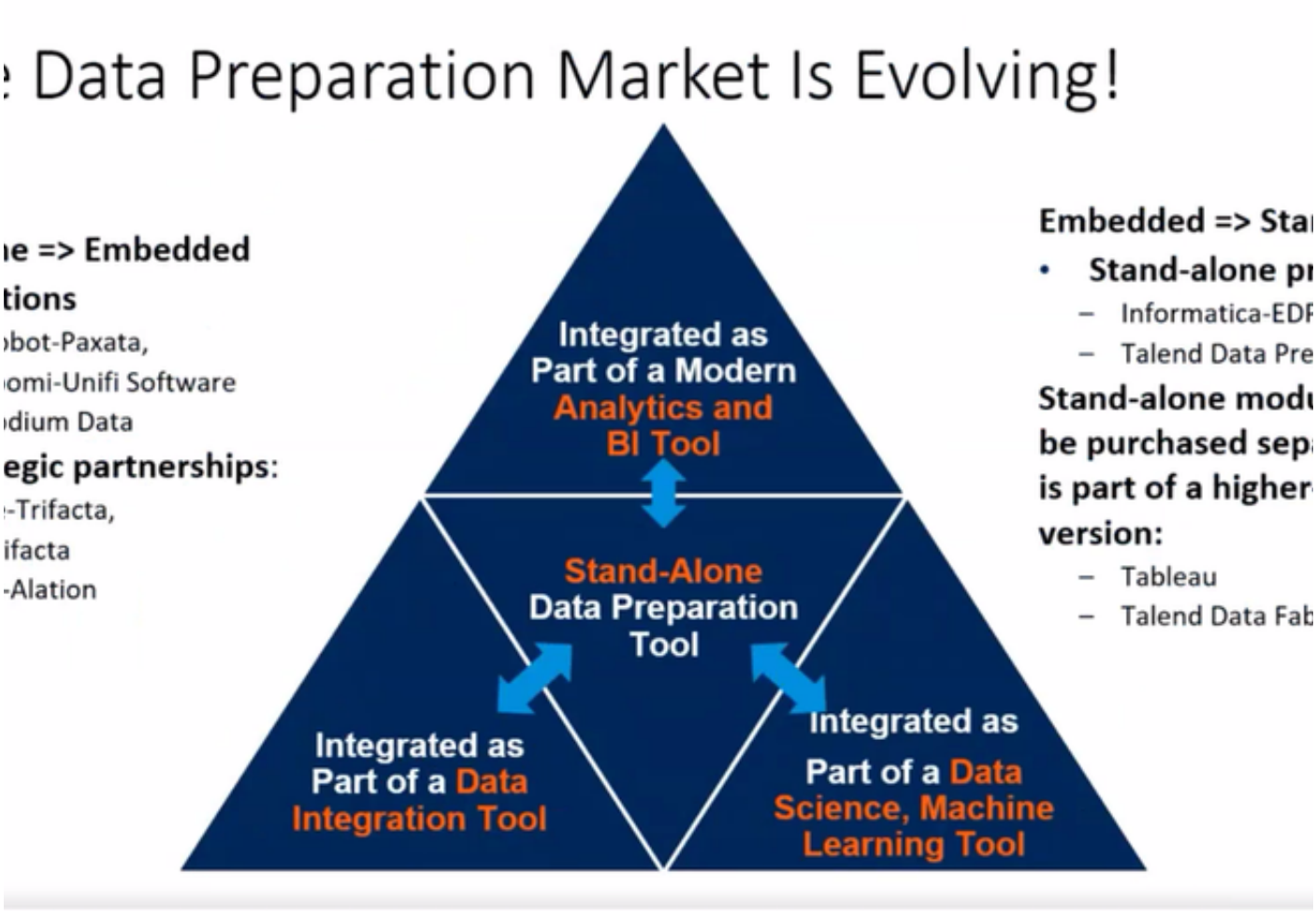
Background

In a landscape with increasing complexity and data volumes, tax departments leverage data wrangling tools to address data transformation and improve data quality. Data wrangling produces data accuracy to feed downstream processes such as compliance and analytics. Tax departments are challenged with producing data more frequently for tax authorities to meet new regulatory requirements. Traditional manual processes need to be improved on. Tools such as Alteryx have transformed and modernized tax departments.

Current State of Data Wrangling

Gartner analyst call on 2/18/2022 discussed and confirmed the current state of Data Wrangling.

Data Wrangling has been transforming from a stand-alone category to and embedded capability. This can be seen by the acquisitions/consolidation of the independent data wrangling tools as well as the advancements in the areas of Data Integration, Data science /ML, and BI / analytics. The chart below illustrates what is happening:



The major cloud providers (AWS, GCP, Azure) are also actively providing low-code tools for data wrangling, both for independent and integrated workflows. The use of the cloud provider's tools is recommended for the best ease of use, most capabilities, single pane of management, and cost efficiencies. The one independent data wrangling tool that has surfaced as a leader is Alteryx, and with its acquisition of Trifecta it is expected to embrace a cloud offering as part of its suite to effectively compete.

One of the major trends that is influencing the inclusion of data wrangling capabilities in these other areas is the advance of citizen data engineers and their access to end-to-end capabilities. This is enabling people familiar with the data to have access to easy to use tools to perform actions on the data.

More and more providers are also using dbt to capture and process the transformations of data. Although not a standard, the increased use is notable.

Also discussed was the use of data wrangling tools to offer in a multi-tenant SaaS solution. The current field of commercial tools is not setup to support this model – the business models and capabilities are directed toward a corporation using their tool, not exposing them as part of another system. Companies that have had a requirement to expose this type of tool have embraced open source (eg. Open refine) as a way to get access to the capability without having to create all of it. This model does introduce more complexity into the solution and maintenance of the solution.

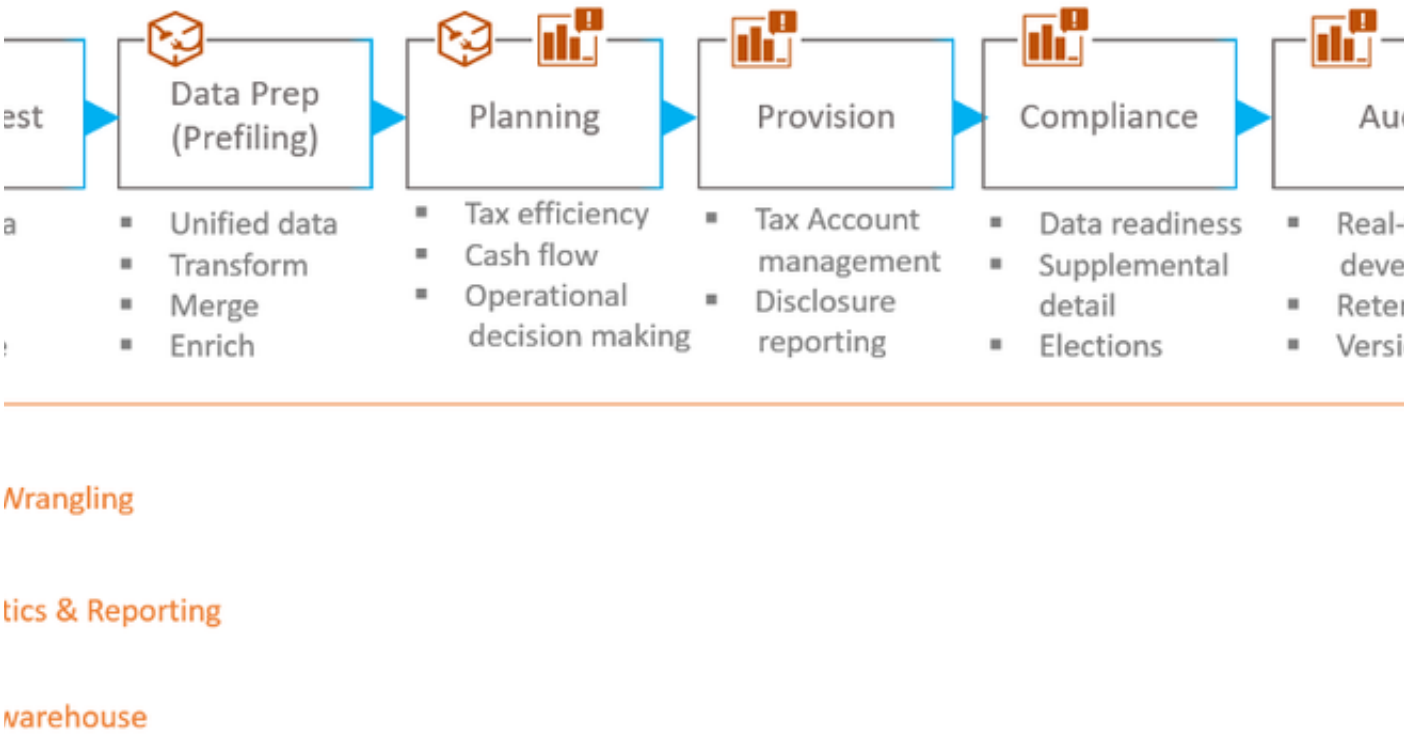
In light of the discussion and above notes, recommending the following technical direction:

Vertex should be focusing on the storage, organization, reporting, analytics, and insights of the data. This may include data enrichment, mapping, and correlations, but would exclude data preparation/data wrangling capabilities. Customers would be expected to use their own tools to access, transform, wrangle, prep, and publish their data into Vertex solutions as well as use their own tools to access the data outputs from Vertex. Vertex should adopt data usability patterns and accelerators to enable customers to publish and receive data more easily. This would include approaches like: data APIs, Python libraries to interact with Vertex data, integrations into the low-code data tools like Alteryx, etc

Tax Lifecycle Management

Tax lifecycle management is the process of bringing in raw source data from multiple systems, validating/cleansing data for accuracy in order to support various needs of the tax department: planning, provision, compliance, audit, etc. Several major processes are involved in the overall tax lifecycle management. The diagram flow below illustrates these steps and helps delineate data wrangling from other data pipelines. Each step in the process is associated with

the applicable data management pipeline (data warehouse, data wrangling, analytics & reporting). Generally, data wrangling functions apply during tax transaction data ingestion, data preparation, and data planning. Analytics & reporting capabilities are offered as separate solutions. Tax departments typically need to integrate several systems in order to fulfill all capabilities required for full tax lifecycle management.



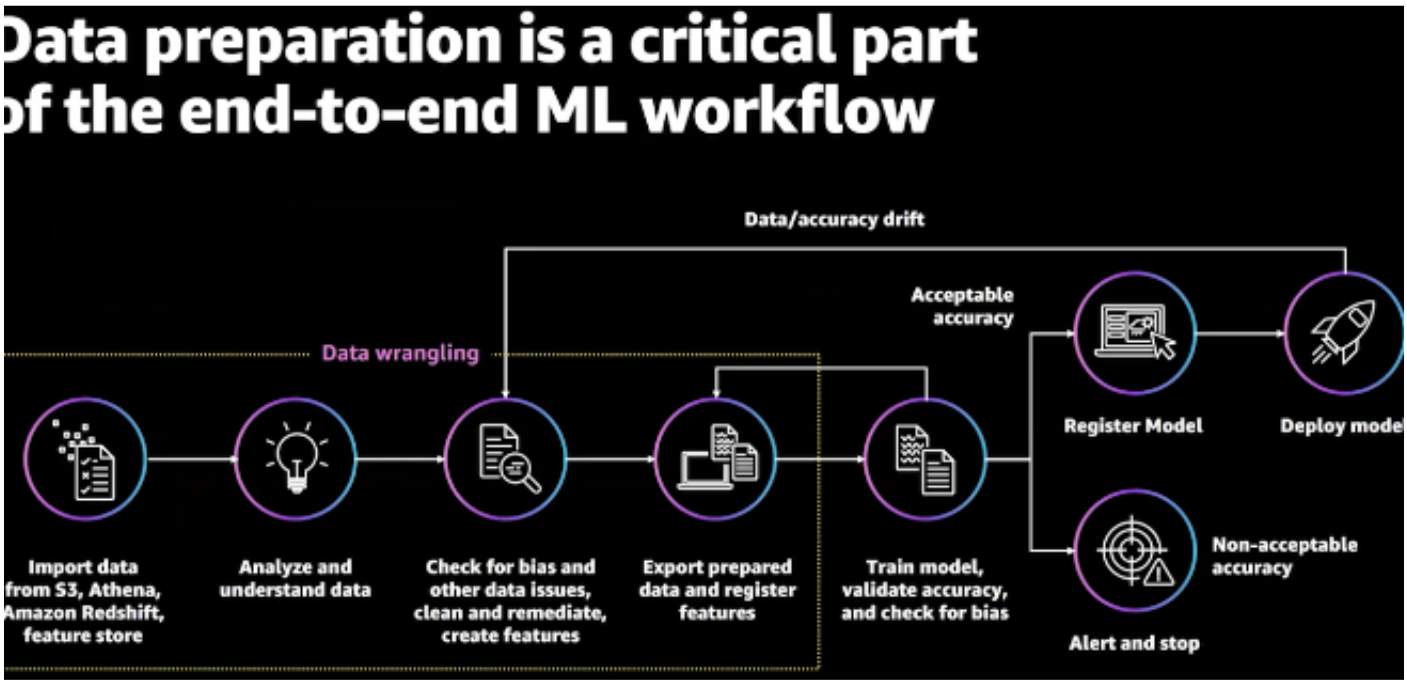
Next Generation Data Wrangling (Data Science and Machine Learning)

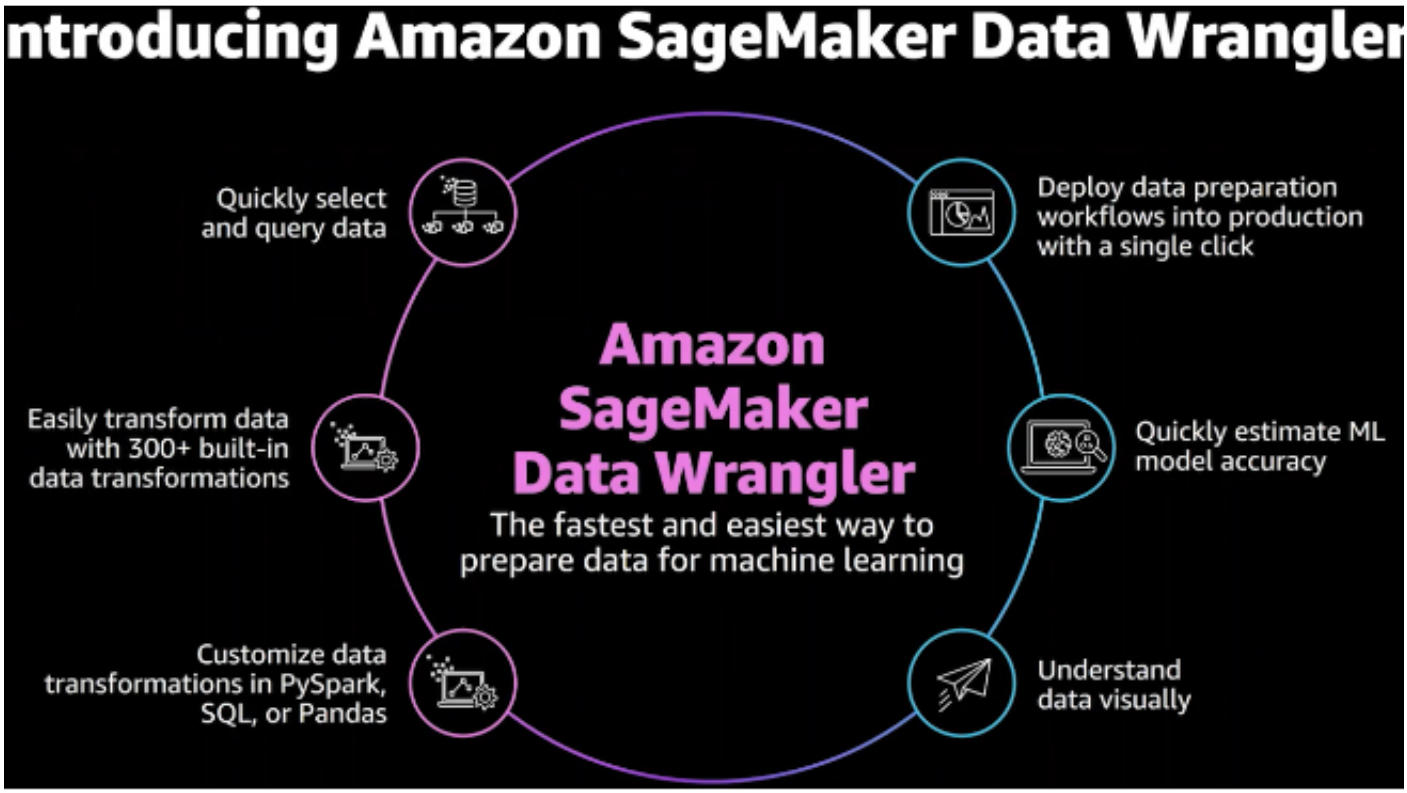
Next generation data wrangling solutions combine with machine learning components to reduce the time required to clean and provide accurate data sets. Machine learning models help identify commonly used data sets and label them for enriching/merging with other datasets.

Amazon SageMaker Data Wrangler

https://aws.amazon.com/sagemaker/data-wrangler/

Amazon SageMaker Data Wrangler reduces the time it takes to aggregate and prepare data for machine learning (ML) from weeks to minutes. With SageMaker Data Wrangler, you can simplify the process of data preparation and feature engineering, and complete each step of the data preparation workflow, including data selection, cleansing, exploration, and visualization from a single visual interface. Using SageMaker Data Wrangler's data selection tool, you can choose the data you want from various data sources and import it with a single click. SageMaker Data Wrangler contains over 300 built-in data transformations so you can quickly normalize, transform, and combine features without having to write any code. With SageMaker Data Wrangler's visualization templates, you can quickly preview and inspect that these transformations are completed as you intended by viewing them in Amazon SageMaker Studio, the first fully integrated development environment (IDE) for ML. Once your data is prepared, you can build fully automated ML workflows with Amazon SageMaker Pipelines and save them for reuse in the Amazon SageMaker Feature Store."

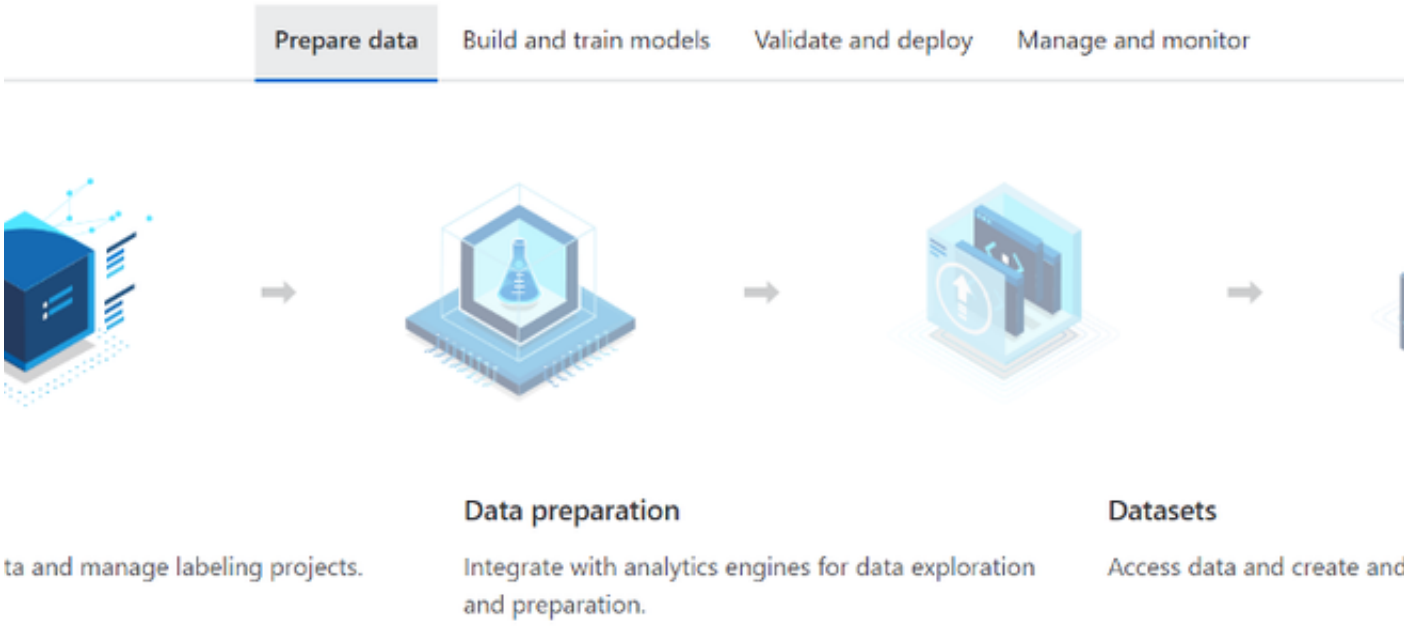




Vendor Examples

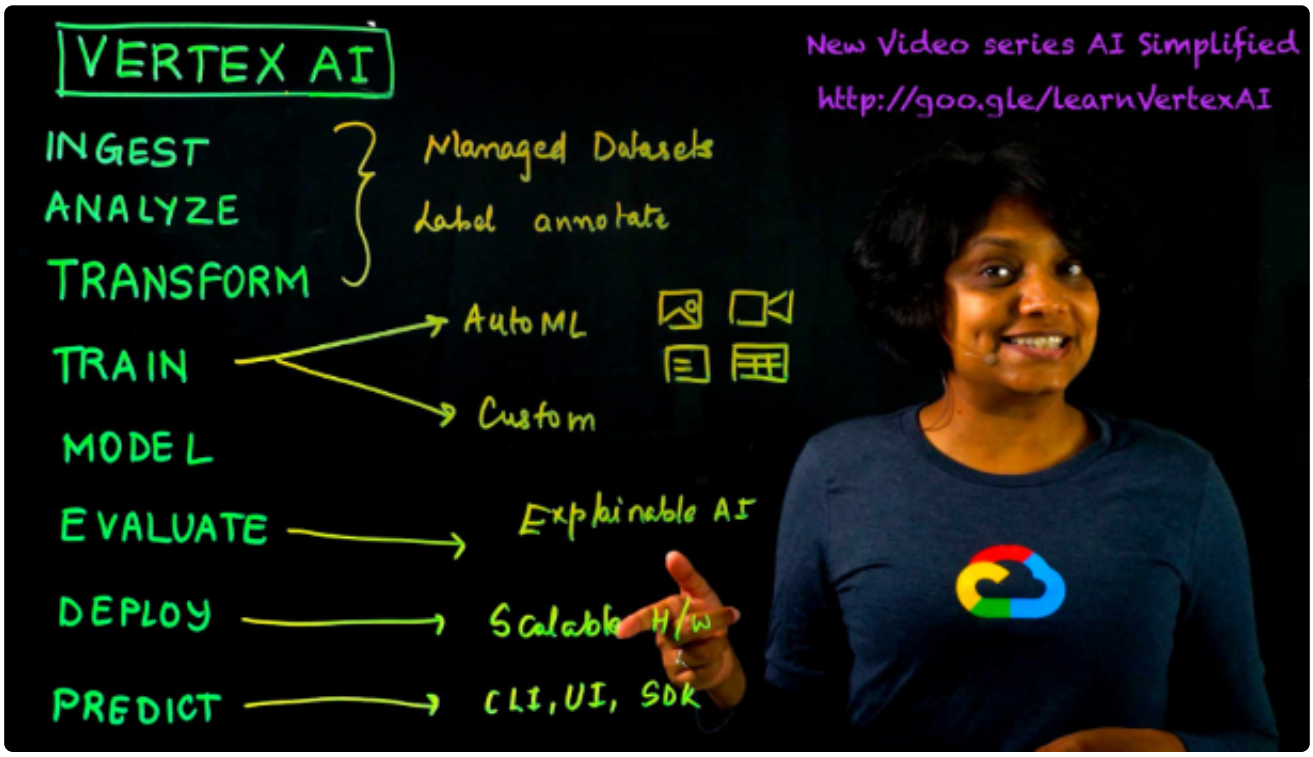
Azure ML

[Azure Machine Learning - ML as a Service | Microsoft Azure]



Google Cloud Vertex AI

Vertex AI | Google Cloud

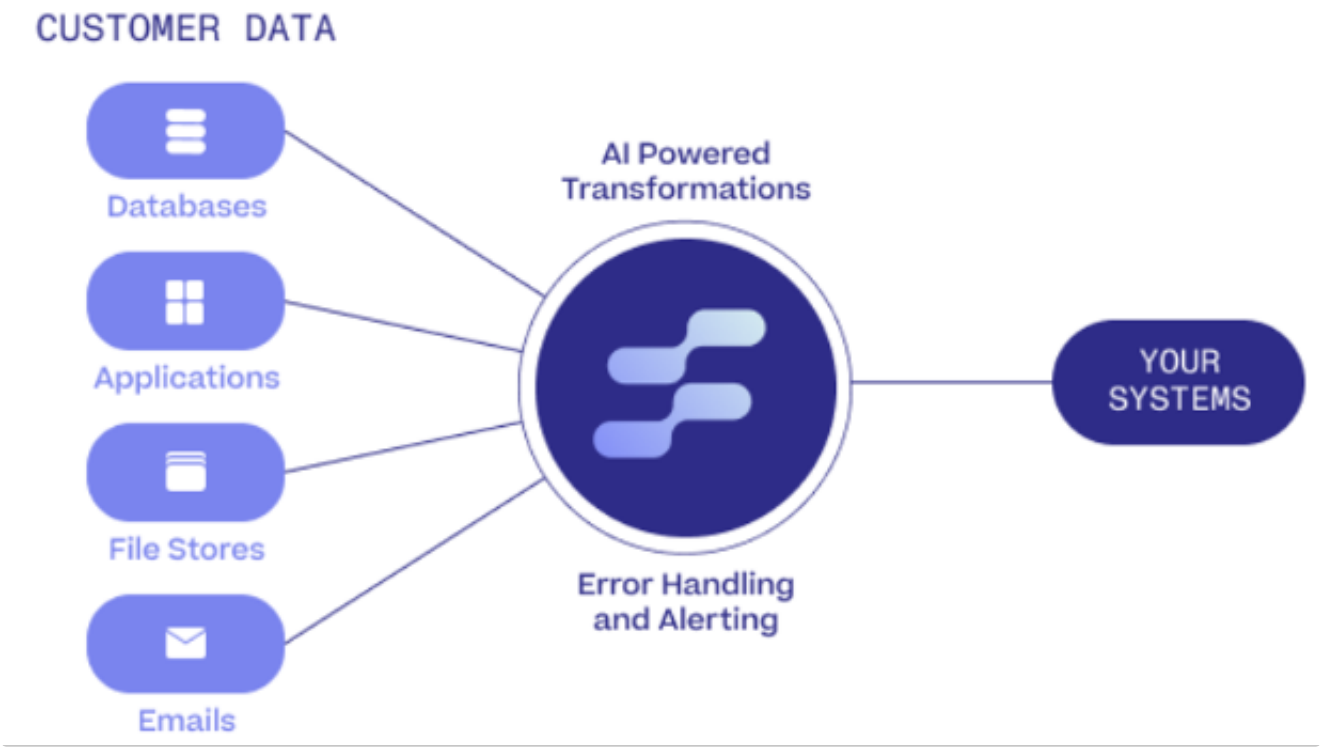


Vertex AI overview | Google Cloud Blog

Osmos

Osmos provides capabilities to clean and automate data imports without writing code.

Osmos Pipelines



Product Market Research pre 2020

-
-
-
-
-
-

-
-
-
-
-
-
-
-
-
-

-
-
-
-
-
-
- Data wrangling
- https://aws.amazon.com/sagemaker/data-wrangler/

Ideation & Emerging Technology Edit **Contact Us**