

EMBEDICA.AI Anvil

To succeed in the world’s rapidly evolving ecosystem, companies (no matter what their industry or size) must use data to continuously develop more innovative operations,processes, and products.

This means embracing the shift to Enterprise AI, using the power of machine learning to enhance - not replace – humans.Working with data is a hard task and can be very distractive.The preparation of data only takes up to 70% of a data science project's lifespan.

Our product is an "anvil" for data science projects. Instead of steel, the raw material here is data. Anvil is there to bring that hard work to a comfort zone.

We do not aim to simplify algorithms, we aim to provide a comfy tool for running complex data science projects. Targeted users are data scientists, analysts and data engineers. Providing a simplistic user interface also enables participation of less advanced data workers. This is pretty useful when thinking of SME.

Anvil is a cloud platform for data scientists who need powerful compute, fast configuration, secure collaboration and easy deployment.

Anvil is the centralized data platform that moves businesses along their data journey from analytic at scale to Enterprise AI,powering self-service analytic while also ensuring the operationalization of machine learning models in production.

Features like unlimited file-size upload, automated data exploration and scalable computing power mean less time accessing and cleaning data and more time building models.

Anvil allows you to share data and code securely with team members, publish version-controlled reports of your analysis .

#### Key Features

* Seamless connectivity to any data, no matter where it’s stored or in what format.
* Faster data cleaning, wrangling, mining,and visualization.
* Unlimited file-size upload, automated data exploration and scalable computing power.
* The latest in machine learning technology(including AutoML and deep learning) all inone place and ready to be operationalized with automation environments, scenarios, and advanced monitoring.
* Every step in the data-to-insights process can be done with a visual interface.

#### Feature Overview

◆ DATA REPOSITORIES

Repositories are logical divisions for datasets. Meaning that Anvil keeps and lets access the datasets within these divisions. The user can define data repositories like for finance, products, and sales, and can relate countless uploaded datasets to these divisions for easy access at later stages. Data Repositories are bounded to teams.

◆TEAMS

Users of Embedica Anvil can be assigned to Teams. Teams are participated within projects, doing all data-related work.

◆ PROJECTS

Once a project is created, teams can be assigned to a project. Teams come along with their accessible data sets (repos).

When developing "projects" we strictly hold on implementing the CRISP-DM approach. Meaning that within a project's life cycle (and even after) users can switch back and forth between different processes of data science project.

##### A Centralized,Controlled Environment

* Connect to existing data storage systems and leverage plugins and connectors for access to all data from one, central location.
* Maintain enterprise-level security with data governance features like documentation, projects, task organization, change management, rollback, monitoring, etc.
* Reuse and automate work (via data transformations, code snippets, centralized best practices, automated assistants, and more), to go from raw data to insights faster.

##### A Common Ground for Experts and Explorers

* Built from the ground up for Data Scientists, Data Engineers, Data Analysts, Business Leaders or Manager and Team Managers.
* From code-free data wrangling using built-in visual processors to the visual interface for machine learning, non-technical users aren’t left in the dust when it comes to contributing to the machine learning projects.

##### A Shortcut to ML/AI Ope-rationalization

■ Centrally manage models and update them from one location,integrating with an API without having to modify or inject anything into existing applications.

##### AI TRUST - Explainability

##### Explainability is creating diverse explanations, including training highly optimized directly interpretable models, creating contrastive explanations of black box models, using information flow in a high-performing complex model to train simpler, interpretable classifiers, learning disentangled representations, and visualizing information flows in neural networks.

In many applications, trust in an AI system will come from its ability to ‘explain itself’.

Anvil explainability is implemented in two ways:

The first one is an assertive feature during the model development process, giving the developer insight on the model.

The second implementation is foreseen for post-production that means after a model is deployed if there are any concerns on the model's behavior, explainability module provides more detailed inspection.

## Connectivity

Anvil allows you to seamlessly connect to your data no matter where it’s stored or in what format. That means easy access for everyone - whether technical or not - to the data they need.

we can say the system will be capable of connecting to Big Data environment for fetching structured and structured data from clustered environments.

■  **SQL Databases**

◆ MySQL

◆ PostgreSQL

◆ Vertica

◆ Amazon Redshift

◆ Pivotal Greenplum

◆ Teradata

◆ IBM Netezza

◆ SAP HANA

◆ Oracle

◆ Microsoft SQL Server (incl. SQL DW)

◆ Google BigQuery

◆ IBM DB2

◆ Exasol

◆ MemSQL

◆ Snowflake

◆ Custom connectivity through JDBC

■ NoSQL Databases

◆ MongoDB

◆ Cassandra

◆ ElasticSearch

■ Hadoop & Spark Supported Distributions

◆ Cloudera

◆ Hortonworks

◆ Google DataProc

◆ MapR

◆ Amazon EMR

◆ DataBricks

■ Hadoop File Formats

* csv
* pARQUET
* orc
* sEQUENCEfILE
* rcfILE

■ Cloud Object Storage

◆ Amazon S3

◆ Google Cloud Storage

◆ Microsoft Azure

* Cloud Object Storage
* Amazon S3
* Google Cloud Storage
* Azure Blob Storage
* Custom Data Sources - extended connectivity through Anvil Plugins
* Connect to REST APIs,
* Create custom file formats
* Connect to own databases

Optimized sync between:

* Amazon S3

ANVİL

* Train the best model in the least amount of time to save human hours.
* Reduce the need for expertise in machine learning by reducing the manual code-writing time.
* Improve the performance of machine learning models.
* Increase reproducible and establish a baseline for scientific research or applications.
* Scales training data set to clusters (Hadoop, [Spark](https://docs.h2o.ai/sparkling-water/2.4/latest-stable/doc/tutorials/sw_automl.html" \t "/home/linus/Documents\\x/_blank), Kubernetes)

So far so good, but what is an Anvil Project?

An Anvil project is where all data science related tasks are held in one place: From data processing to running various models and evaluating them, along comparing model performances and selecting best performing ones for deployment. Anvil provides powerful explainability feature during model development as well.

We provide:

\* easy data preprocessing functionalities through simplistic UI

\* automated data preprocessing.

\* a well organized, enterprise-level tool to leverage benefits of various (and counting) AI/ML algorithms

\* time for the data scientist to focus more on the algorithms

\* ease of reporting

\* bringing EXPLAINABILITY to model, a very useful feature especially in understanding complex models' prediction behaviours

\* importing custom (pick) models for bringing explainability to model behaviour.

Summary of main features Anvil provides:

\* Run all data science processes within projects, and access them any time.

\* Create automated data processing

\* Powerful explainability functionalities where model developers need to interpret advanced models

\* Loading custom models as pkl files

\* Deploying trained models for decision makers use

\* Reporting functionality for model results

INDUSTRIES: Despite to the agnostic nature of ML models, Anvil is specialized with use cases from cyber security and finance. Anomaly detection, real-time analysis

DATA SECURITY: Anvil is provided as a Docker image. Residing on customers' on-prem or virtual private could, the data is stored on client's own machines. Connection to the outside world is made when importing data from Cloud providers (Amazon S3, Microsoft Azure, etc.) or for Embedica Anvil licence validation.

BIG DATA CONNECTIVITY: Dealing with big data, more and more companies migrate to big data envronments like Hadoop. With Big Data Connectivity current Anvil's resources will be extended to a Big Data scale.

##### Exploratory Analytics

##### Sometimes you need to do a deep dive on your data, but other times,its important to understand it at a glance. From exploring available datasets to dashboarding, Anvil makes this type of analysis easy.

###### Data Analysis

* Automatically detect dataset schema and data types
* Assign semantic meanings to your datasets columns
* Build univariate statistics automatically &derive data quality checks
* Dataset audit
* Automatically produce data quality and statisticalanalysis of entire Anvil datasets
* Support of several back-ends for audit (in-memory,Spark, SQL)

###### Data Cataloging

* Search for data, comments, features, or models in a centralized catalog.
* Explore data from all your existing connections

###### Advanced analysis

Interactive visual statistics

* Univariate analysis and statistical tests on single or multiple populations.
* Statistics and tests on multiple populations
* Correlations analysis
* Principal Components Analysis

###### Data Visualization

* Create standard charts (histogram, bar charts,etc) and scale charts’ computation by leveraging underlying systems (in-database aggregations)
* Create custom charts using:
* Custom Python-based or R-based Charts

###### Dashboarding

* User-managed reports and dashboards
* Shows streaming history and activity
* Custom Insights Plotly, Matplotlib)
* Custom interactive, web-based visualisations

#### Data Preparation

Traditionally, data preparation takes up to 80 percent of the time of a data project. But Anvil’s data prep features make that process faster and easier, which means more time for more impactful (and creative) work.

we can say the system will be capable of connecting to Big Data environment for fetching structured and structured data from clustered environments.

###### Visual Data Transformation.

Design your data transformation jobs using a point-and-click interface

* Group
* Filter
* Sorts
* Stack
* Join
* Window
* Sync
* Distinct
* Top-N
* Pivot
* Split

■ Scale your transformations by running them directly in distributed computations systems (SQL, Spark)

Data repository:

Repositories are logical divisions for datasets. Meaning that Anvil keeps and lets access the datasets within these divisions. The user can define data repositories like for finance, products, and sales, and can relate countless uploaded datasets to these divisions for easy access at later stages. Data Repositories are bounded to teams.

Data Pool: Pools have changed datasets which any changing on the dataset that supports change data type, rename removes, or add new a column that means you can reshape according to your requests additional a still continues stored the original dataset on the data repository.

Table Join: According to join types the user can merge two different datasets to one new that store of permanently.

###### Interactive Data Preparation:

* Scale data data preparation scripts using in-database (SQL) or in-cluster (Spark) processing
* Automatically turn data preparation scripts into Spark or Map Reduce jobs

###### Dataset Sampling:

First records, random selection, stratified sampling,specifical sampling, etc.

#### Machine Learning

Anvil offers the latest machine learning technologies all in one place so that data scientists can focus on what they do best: building and optimizing the right model for the use case at hand.

Automated Machine Learning(AutoML)

AutoML or Automatic Machine Learning is the process of automating algorithm selection, feature generation, hyperparameter tuning, iterative modeling, and model assessment.

AutoMl technology makes it easy to train and evaluate machine learning models for everyone and everysize. Automating repetitive tasks allows people to focus on the data and the business problems they are trying to solve.

AutoML Technology made Easy for EVERYONE & EVERYSIZE

###### Automated ML strategies

* Quick prototypes
* Interpretable models
* High performance

###### Features handling for machine learning

* Support for numerical, categorical, text and vector features
* Automatic preprocessing of categorical features (Dummy encoding, impact coding, hashing, custom preprocessing, etc.)
* Automatic preprocessing of numerical features (Standard scaling, quantile-based binning, custom preprocessing, etc.)
* Automatic preprocessing of text features (TF/IDF, Hashing trick, Truncated SVD, Custom preprocessing)
* Various missing values imputation strategies
* Features generation
* Feature-per-feature derived variables (square, square root...)
* Linear and polynomial combinations
* Features selection
* Filter and embedded methods
* Choose between several ML backends to train your models
* TensorFlow
* Keras
* Scikit-learn
* XGBoost
* MLLib
* H20
* Algorithms
* Python Based

+ Logistic regression

+ Random Forests

+ XGBoost

+ Decision Tree

+ K-Means

* Spark MLLib-based

+ Logistic Regression

+ Linear Regression

+ Decision Trees

+ Random Forest

* H20-based

+ Random Forest

* Linear Regression
* Logistic Regression
* Naïve Bayes
* KNN
* SVM
* Decision Trees
* Random Forest

## How can AutoML Help You?

If you’re part of the majority of data scientists who work with tabular or “relational” data (tables with numeric and/or categorical columns),

then Embedica Anvil is great tool to use.

#### Machine Learning

Anvil offers the latest machine learning technologies all in one place so that data scientists can focus on what they do best: building and optimizing the right model for the use case at hand.

Model Deployment

* Model version
* Batch scoring
* Real-time scoring
* Easily manage all your model deployments
* One-click deployment of models
* Export Model
* Users can download the trained model for them to use wherever they want.
* The model is saved as a pickle file that the process of converting a Python object into a byte stream to store it in a file/database, maintain program state across sessions, or transport data over the network.
* Deploy as API
* Users can deploy their model as API and can run predictions and other stuff related to that model from external sources.
* Users can use the API on their own apps or programs for their customer or themselves
* Expose arbitrary functions and models through API’s
* Write custom R, Python or SQL based functions or models
* Automatically turn them into API endpoints for operationalisation
* Docker & Kubernetes
* Deploy models into Docker containers for operationalisation
* Automatically push images to Kubernetes clusters for high scalability
* Model monitoring mechanism
* Control model performances over time
* Automatically retrain models in case of performance drift
* Customize your retraining strategies
* Logging
* Log and audit all queries sent to your models

Deep Learning

* Support for Keras with Tensorflow backend
* User-defined model architecture
* Personalize training settings
* Support for multiple inputs for your models
* Support for CPU and GPU
* Support pre-trained models
* Extract features from images
* Tensorboard integration

#### Machine Learning

Anvil offers the latest machine learning technologies all in one place so that data scientists can focus on what they do best: building and optimizing the right model for the use case at hand.

###### Unsupervised Learning

* Automated features engineering (similar to Supervised learning)
* Optional dimensionality reduction
* Outliers detection
* Algorithms

Model Training

#### Automation Features

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

###### Data Flow

* Keep track of the dependencies between your datasets
* Manage the complete data lineage
* Check consistency of data, schema or data types

Partitioning

* Leverage HDFS or SQL partitioning mechanisms to optimize computation time

Metrics & Checks

* Create Metrics assessing data consistency and quality
* Adapt the behavior of your data pipelines and jobs based on Checks against these Metrics
* Leverage Metrics and Checks to measure potential ML models drift over time

Monitoring

* Track the status of your production scenarios
* Visualize the success and errors of your Anvil Jobs

Automation Environments

* Trigger the execution of your data flows and applications on a scheduled or event-driven basis
* Create complete custom execution scenarios by assembling a set of actions to do (steps)
* Leverage built-in steps or define your own steps through a Python API
* Publish the results of the scenarios to various channels through Reporters (Send emails with custom templates; attach datasets, logs, files,or reports to your Reporters; send notifications

Automation Environments

* Use dedicated Anvil Automation nodes for production pipelines
* Connect and deploy on production systems (data lakes, databases)
* Activate, use or revert multiple Anvil project bundles

### Code

Work in the tools and with the languages you already know everything can be done with code and fully customized. And for tasks where it’s easier to use a visual interface, Anvil provides the freedom to switch seamlessly between the two.

### Collaboration

Anvil was designed from the ground up with collaboration in mind. From knowledge sharing to change management to monitoring, data teams including scientists, engineers, analysts, and more can work faster and smarter together.

Shared platform for Data Scientists, Data Engineers, Data Analysts, Business Leaders and Manager, etc.

Team Activity Monitoring

* Global search to quickly find all project assets, plugins, wiki, reference docs, etc.
* Shared code-based components
* Distribute reusable code snippets for all users
* Package arbitrary complex function, operation or business logic to be used by less-technical users

#### Governance & Security

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

###### User profiles

###### Role-based access(fine-grained or custom)

###### Authentication management

* Use SSO systems
* Connect to your corporate database (LDAP, Active Directory...) to manage users and groups

###### Enterprise-grade security

* Track and monitor all actions in Anvil using an audit trail
* Authenticate against Hadoop clusters and databases through Kerberos
* Supports users impersonation for full traceability and compliance

###### Custom policy framework for data protection and external regulations compliance

* Framework capabilities

◊ Document data sources with sensitive information, and enforce good practices

◊ Restrict access to projects and data sources with sensitive information

◊ Audit the sensitive information in an Anvil instance

#### Architecture

Anvil was built for the modern enterprise, and its architecture ensures that

businesses can stay open (i.e., not tied down to a certain technology) and that

they can scale their data efforts.

* No client installation that easy to use web-based Cloud Deployment for Anvil users
* Anvil nodes that use dedicated Anvil environments or nodes to design, run, and deploy your ML applications.
* Integrations
* Leverage distributed systems to scale computations through Anvil
* Automatically turn Anvil jobs into SQL, Spark, MapReduce, Hive, or Impala jobs for in-cluster or in-database processing to avoid unnecessary data movements or copies
* Modern architecture (Docker, Kubernetes, )
* Full AutoML Deployment with Docker Containers for suitable for all environments
* Trace-ability and debugging through full system logs

|  |
| --- |
| AutoML Technology made Easy for EVERYONE & EVERYSIZE-added |
| Developed for companies having no prior Data science and AI teams and technologies |
| For Data Analysts, Data Engineers, Data Scientists, and Business Leaders and Managers -added |
| Sales, Marketing, Risk, Operations, HR and Finance, IT, Procurement and Management Teams |
| Apply and Deploy AutoML Models for in Health care, Financial Services, Banking, Retail,  Services, Logistics, e-commerce, Media & Advertising, Marketing Data |
| Power of Big Data, Cloud Platforms, and AI Deep Learning Technologies and Algorithms Combined |
| Natural Language Processing  (NLP) Module - for Advanced Text & Language Analytics, Documents, Email, and Free Text |
| Built-in EM EASI AI TRUST module -  Train and Explain models to boost AI Trust and adaptability in AI Projects |
| Easy to navigate Dashboard for all your Projects and Teams - Run multiple Data Science,  NLP, and Computer Vision Projects |
| 15 Data Connectors - most commonly used for Databases Flat files and Cloud Platforms |
| Data Pre-Processing Module - for Uploading, Cleaning, and viewing data fields |
| Data Training and Testing Module - enables continuous Training and Testing, modifying and changing parameters on the fly |
| Full AutoML Deployment with Docker Containers for suitable for all environments added |
| Python & R Workbench -  Pre-installed Deep Learn Frameworks & Libraries for Advanced Users and Data Scientists |
| Open-Source Tools & API integration and customization available |
| Beautiful and easy Visualization and Charts for Business Leaders and Managers |
| High Accuracy, Gradient Boosting with High Interpretability of models |
| Enterprise Level Security for Teams and Departments |
| Easy to use web based Cloud deployment-added |
| Maintenance Customization Monitoring and Alerting Services |
| Save time Improve Productivity Transform into Digital AI company Gain Advantage |