kamu

Bringing true collaboration to modern data science

$3 trillion / year	> 1,800	80%	100%
of value is locked in open data	ML / AI startups hungry for data	of time still being spent on data preparation	of this work is **bound to be repeated**
[McKinsey]	[AngelList]	[Forbes]	[kamu]

Problem

Data is the most powerful decision making tool at our disposal, but we've just begun to learn how to use it.

Most data is of poor quality and needs to go through many stages of transformations to unlock value.

Modern data science encourages routine copying of data, producing new data that:

› Disjoint from its source - it's hard to tell where data came from and how it was altered

› Not trustworthy - there is no practical way to verify that no malicious or accidental alterations to the data were made

The data science is stuck in a loop where all forward progress is constantly lost because no mechanism exists for making incremental improvements to data.

Solution

Kamu brings technology that for the first time can enable true collaboration around data on the same scale that is seen today in software.

We aim to diminish following limitations:

› Break the vicious cycle of low investments and returns for data publishers - make it easy to publish good data

› Stop the spread of non-trustworthy derivative data sets - make good data easy to find

› Provide a standard way to make transparent and verifiable improvements to data - enable collaboration around data on the global scale

› Empower enterprise, startups, governments, and communities - make quality data readily available for everyone

Data is our modern age history book - it's time we treat it as such

Technology

We build our products around these core traits:

✓ **Complete history**
Every bit of data is immutable and stored forever

✓ **Automatic lineage**
Trace any data to its origin down to the cell level

✓ **Transparency**
See every transformation the data went through

✓ **Repeatability**
Guaranteed consistent results of your data science projects

✓ **Verifiability**
Anyone can verify the transformation history, tampering is made impossible

✓ **Time travel**
Rewind time and examine past states of data with bitemporal data model

✓ **Living data**
Get near real-time updates as new data arrives with streaming data transformations engine

There is no collaboration without trust, and blind trust is not an option

Products

Our suite of products targets the foundation layer of any modern data science project:

› **DataFabric** - an open standard for decentralized metadata exchange for transparent and verifiable data that captures the evolution of its schema, and tracks all applied transformations throughout the dataset lifetime

› **Kamu CLI** - an open-source tool that implements the above standard, providing all the amazing data management capabilities to anyone, ranging from professional to hobbyist data scientists
https://github.com/kamu-data/kamu-cli

› **Kamu Hub** - a platform for open exchange and collaboration on data, built to close the feedback loop between data publishers and consumers and to create self-organizing communities in the world of data

https://kamu.dev   info@kamu.dev