



Project Name: GENESIS

Team Name: GENESIS

Email: xxxxxxxxxxxx@u.nus.edu

Phone: +65 xxxxxxxx

Demonstration URL:

<https://marymountlabs-genesis-test-streamlit-app-pps9r8.streamlitapp.com/>

/ You make all the difference /

Huawei Developer Competition

Spark Infinity



— 2 0 2 2 —

China, Asia Pacific, Latin America, the Middle East, Africa, and Europe
June to November, 2022





Overview



Project Overview

Project Name	<i>GENESIS 合创</i>
Team Name	<i>GENESIS 合创</i>
Contacts	<i>Zacchaeus Chok</i>
Technical Field	<i>Synthetic Data Generation, Machine Learning Analytics</i>
Technologies	<i>OBS, GaussDB, ModelArts, Generative Adversarial Networks, Deep Learning</i>
Keywords	<i>Cloud-native, Synthetic, Analytics</i>
Applicable Fields	<i>Healthcare, Insurance, Government, Finance</i>
Description (in 500 words)	<p><i>GENESIS is a cloud-native synthetic data start-up, providing end-to-end data analytics capabilities starting in healthcare. Most real-world healthcare datasets are incomplete, poorly annotated and lack diversity. The lack of representative, large-scale datasets in real-world settings results in a critical bottleneck prohibiting the adoption of AI applications. In healthcare industries, data also needs to be treated sensitively, mandating the de-identification of datasets to support the adoption of data analytics. Resultantly, data gaps mean that data-driven analyses and interventions cannot be adopted, especially in local settings. GENESIS supplements the shortfall in real-world data with synthetic data. Clients upload can upload single table datasets or even complex, multi-table, relational datasets onto the cloud server. Using a proprietary generative adversarial algorithm, GENESIS ensures that the generated data preserves key ground truth characteristics by preserving the relationship of biological variables and retaining correlational structure. Further, the preservation of patient privacy is achieved through automatic de-identification and anonymisation of datasets. GENESIS leverages Huawei's ModelArts platform to deploy machine learning analytics on synthetic datasets. A low-code interface enables users to rapidly identify trends and patterns from the enriched datasets, and the results and visualisations can be exported to other systems. Finally, GENESIS assures data utility through evaluation benchmarks based on biological relationship preservation, univariate distance, multivariate distance, and privacy preservation. The fidelity benchmark ensures that synthetic data is faithful to the original datasets and the fairness benchmark ensures that the synthetic data fairly represents all sensitive attributes.</i></p>

Contents



1 Team Introduction

2 Project Introduction

3 Functions

4 Technical Architecture

5 Innovations

6 Business Value

7 Achievements

8 Project Planning



Team Introduction

Near-term plans to recruit technical founding members



Zacchaeus Chok 卓永信
Founder

Education

- NUS Computer Science + Business
- Stephen Riady Young Entrepreneur Scholar

Awards

- SMU Social Engine | Champion
- National Youth Council | Young Changemaker

Product Experience



*Resume diagnosis
using NLP*



*Digital marketing using
NLP*



*Data management
interface for clinics*

Our Advisory Panel



Dr Lim Chien Chuan
Director
i-Care Primary Care Network

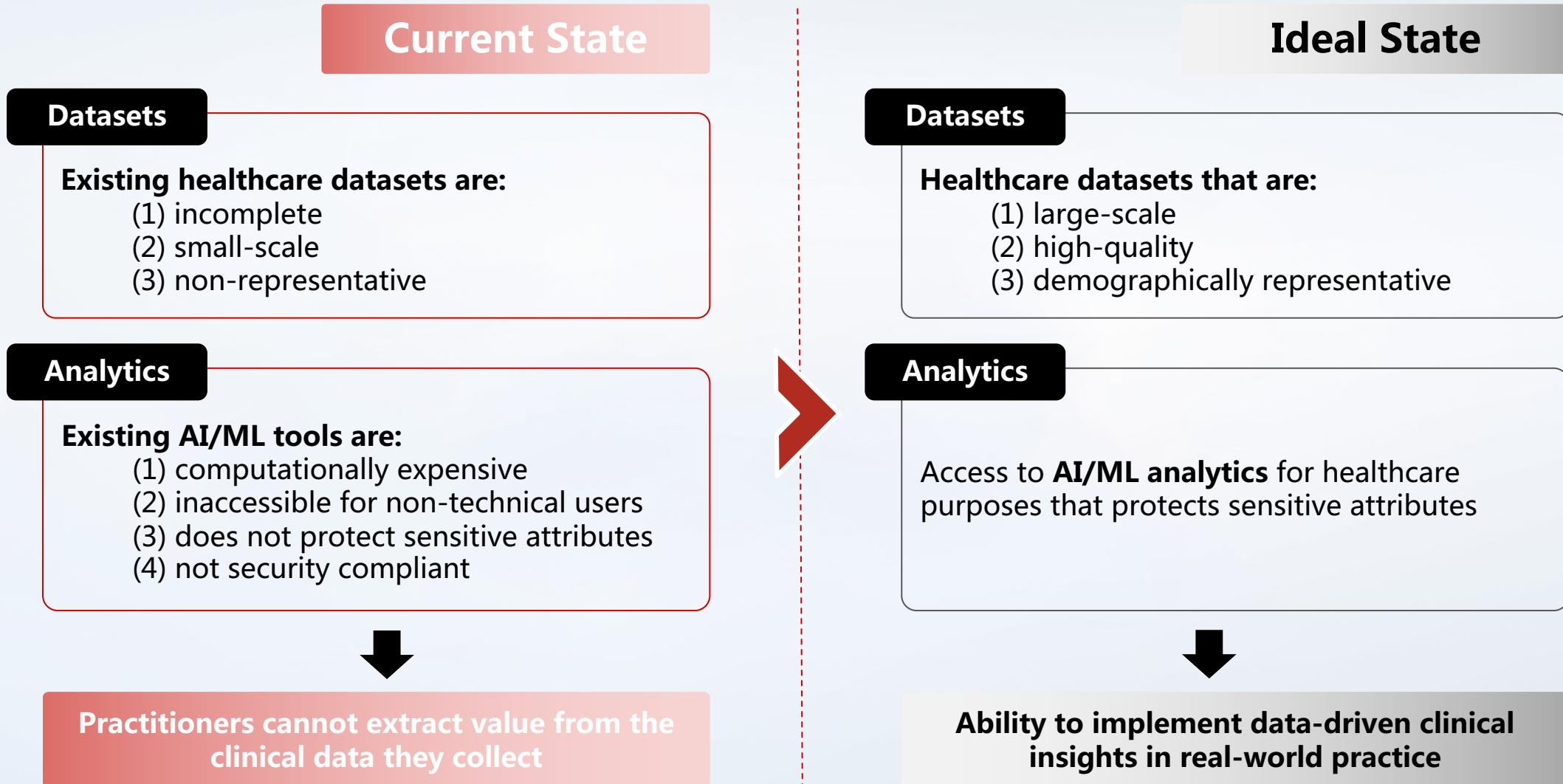


Karen Tsang
Head (Operations)
i-Care Primary Care Network



Project Introduction

Healthcare practitioners **lack access to data-driven clinical insights**





Project Introduction

Synthetic data generation **overcomes existing flaws in clinical datasets**

DIAGNOSIS

01

Incomplete data

Poorly annotated, fragmented datasets across legacy systems

02

Small-scale data

Datasets are usually too small (1000s of records) for meaningful machine learning analysis

03

Non-representative data

Does not represent underlying population demographics patterns

REFERENCES

Wang, Z., Myles, P. & Tucker, A. (2021). 'Generating and evaluation cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy' , *Computational Intelligence*, 37(2), pp. 819 - 851.

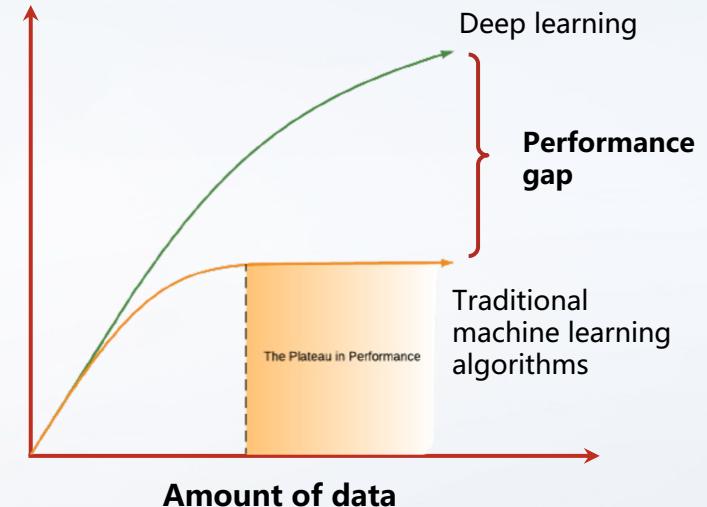
Tucker, A., Wang, Z., Rotalinti, Y. & Myles, P. (2020). 'Generating high-fidelity synthetic patient data for assessing machine learning healthcare software' , *npj Digital Medicine*, 3(147).

Chen, R.J., Lu, M.Y., Chen, T.Y., Williamson, F.K. & Mahmood, F. (2021). 'Synthetic data in machine learning for medicine and healthcare' , *Nature Biomedical Engineering*, 5, pp. 493 - 497.

Das, H.P., Tran, R., Yue, X., Tison, G., Sangiovanni-Vincentelli, A. & Spanos C.J. (2022). 'Conditional Synthetic Data Generation for Robust Machine Learning Applications with Limited Pandemic Data' , *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11).

SOLUTION

Performance



Algorithmically generated datasets that statistically mimic real-world datasets

- Creates complete dataset by **filling in missing attributes** based on global population attributes
- Creates **larger population-size datasets** that are amenable to machine learning analysis
- Ensures **strong fidelity with underlying demographic attributes** that can be statistically proven

Clinics currently hold small amounts of data that cannot be used to train powerful, deep learning models.

Rather than purchasing more data, or spending more resources to collect larger datasets, **synthetic data generation makes SMEs' existing datasets compatible with the big data, deep neural nets paradigm**.



Project Introduction

Cloud-native analytics enables practitioners to derive insights quickly and fairly

DIAGNOSIS

01

Computationally Expensive

Not feasible for individual clinics to set up GPU/TPUs for deep learning

02

Not Security Compliant

Local computers typically cannot store and transmit data securely

03

Not Accessible

Most healthcare practitioners are not familiar with data science tools

04

Not Fair

Most tools do not account for ML fairness in their analysis, which is critical for healthcare analytics

SOLUTION

Cloud-native AI/ML analytics

- Clustering on cloud allows for **greater efficiency and scale**, with minimal set-up costs for individual users
- Leverage full-stack cloud security system offered by Huawei

Low-code, modern UI/UX platform for non-technical users

ML fairness evaluation metrics

- Built into analytics suite available for users to automatically confirm fairness of ML analysis

REFERENCES

Friedler, S.A., et al. (2019). 'A comparative study of fairness-enhancing interventions in machine learning' , *Proceedings of the Conference on Fairness, Accountability and Transparency*, pp. 329 - 338.

Chouldechova, A. & Roth, A. (2020). 'A snapshot on the frontiers of fairness in machine learning' , *Communications of the ACM*, 63(5), pp. 82 - 89.

Rajkomar, A. et al. (2018). 'Ensuring fairness in machine learning to advance health equity' , *Annals of Internal Medicine*.



Functions

GENESIS is the most cost-effective way for clinics to transform their datasets for the **new analytics paradigm**

The following packages are currently under testing

Upload

De-identification and anonymisation

Dynamic data masking to break link between individual and attributes

Transform

Data Schema Conversion

Encodes data into JSON elements for efficient retrieval and analysis

Synthetic Data Generation

Proprietary generative adversarial algorithm to generate synthetic data at scale

Faithfulness Evaluation

State-of-the-art fidelity benchmarks to ensure synthetic data is faithful to the original dataset

Analyse

- i) Suite of **low-code machine learning and data science tools** to rapidly identify trends and patterns in synthetic datasets
- ii) Results and **visualisations** can be exported out to local networks

Most synthetic data engines focus only on synthetic data generation and are not tightly integrated into existing data workflows

betterdata

MOSTLY.AI

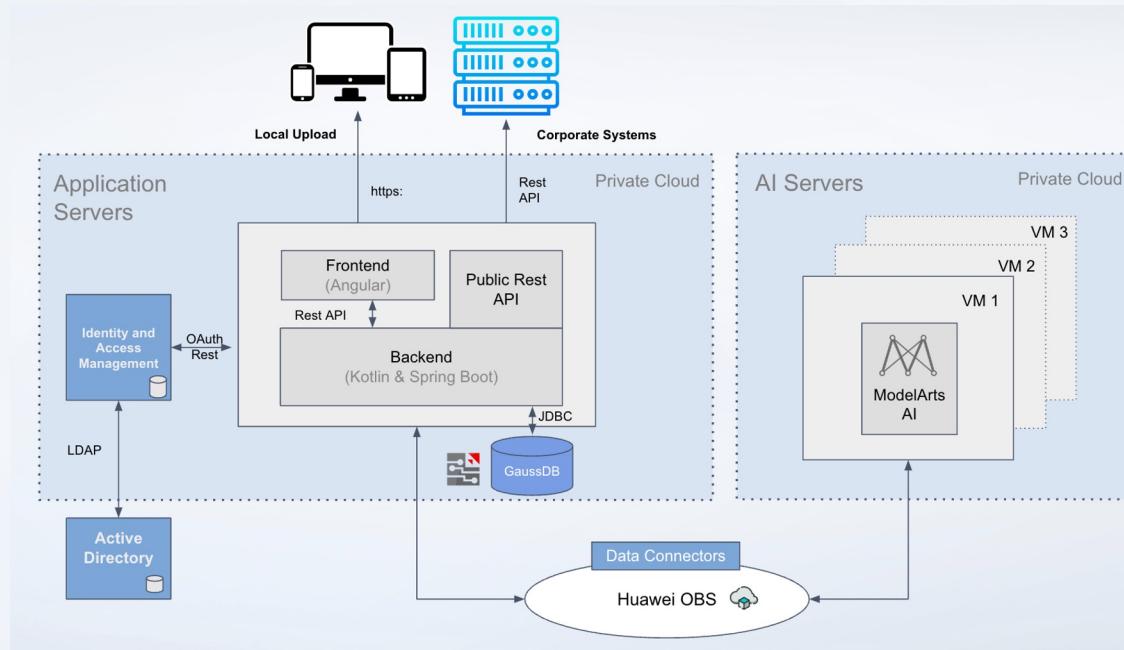
SDV
The Synthetic Data Vault



Technical Architecture

Overview of key Huawei cloud services used in GENESIS

SIMPLIFIED OVERVIEW



SERVICE DESCRIPTION

Synthetic data generation, evaluation benchmarks and data analytics are performed on the cloud. The process is as follows:

1. Dataset uploads can be performed locally or through corporate servers.
2. Once uploaded via APIs on the frontend, they are transmitted securely via REST API. Information is encrypted in storage and transmission.
3. Data is stored securely on **GaussDB** to interface with AI servers via **OBS**. AI servers host **RapidGAN** and evaluation tool sets deployed using **ModelArts AI**.
4. Aggregated analytical insights are stored securely on **OBS** and can be exported to local and corporate servers with the appropriate permission controls.

KEY TECHNOLOGIES

01 **OBS**

02 **GaussDB**

03 **ModelArts**

DESCRIPTION

- Elastic, scalable storage system handling high volume of uploaded data
- Real-time file processing that triggers file operations
- Real-time deployment of RapidGAN models on highly-concurrent data

ADVANTAGES

- Secure authentication and fine-grained permission control to **guard sensitive patient data**
- Automatically **scale out resources** to run more function instances as number of requests increase
- Edge services capable of completing inference locally flexibly; enables **efficient import/export**



Cloud Services

Deployment of Huawei Cloud Services

For more information, please see Appendix

The screenshot shows the Huawei Cloud Console interface. The top navigation bar includes 'HUAWEI CLOUD', 'Console', 'Singapore', 'Search', 'More', 'Intl-English', and a user icon. The main content area is titled 'My Resources [Global]' and displays resource counts for various services: Buckets (2), Disks (1), ECSs (1), VPC Endpoints (1), Regions for ModelArts (1), Models (1), Bandwidths (1), Security Groups (3), VPCs (1), EIPs (1), Nodes (3), Instances (1). Below this is a 'Recently Visited Resources' section with tabs for ECSs, Instances, Models, and Buckets. To the right are three cards: 'Cloud Eye [Singapore]' (Alarm: 0, Insufficient Data: 0, Alarms from Last 7 Days: 0), 'ECS Resource Monitoring' (CPU Usage: 0.62%, Disk Read Rate: 0 KB/s, Inbound Network Speed: 0 Kbit/s), and 'Security Overview' (with a shield icon and a message about enabling Security Analytics). A sidebar on the right contains a 'Hello!' message, a 'Billing Center' link, and a 'Not authenticated' status. It also features a 'Tutorial' section with links to purchasing an ECS, creating an EVS disk, setting up an IPv4 network, best practices for websites, and uploading files through the OBS console.



ECS

Cloud servers for low-cost, low-latency AI inference



FunctionGraph

Scalable services compute for API backend



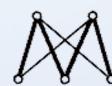
GaussDB for NoSQL

Reliable, responsive database backend



API Gateway

Frontend-backend communication proxy



ModelArts

GPU-accelerated training for TensorFlow



OBS

Stores objects such as bulk files



Cloud Services

Model deployment on ModelArts

Huawei Cloud

Cloud service provider with data storage and AI platform features

ECS

Secure, scalable, on-demand compute resources

OBS

Cloud storage service optimised for storing massive amounts of data

ModelArts

One-stop development platform for AI developers from data to output

METHODOLOGY

1

Upload code to OBS

2

Create **synthetic data generation algorithm** on Huawei Cloud ModelArts

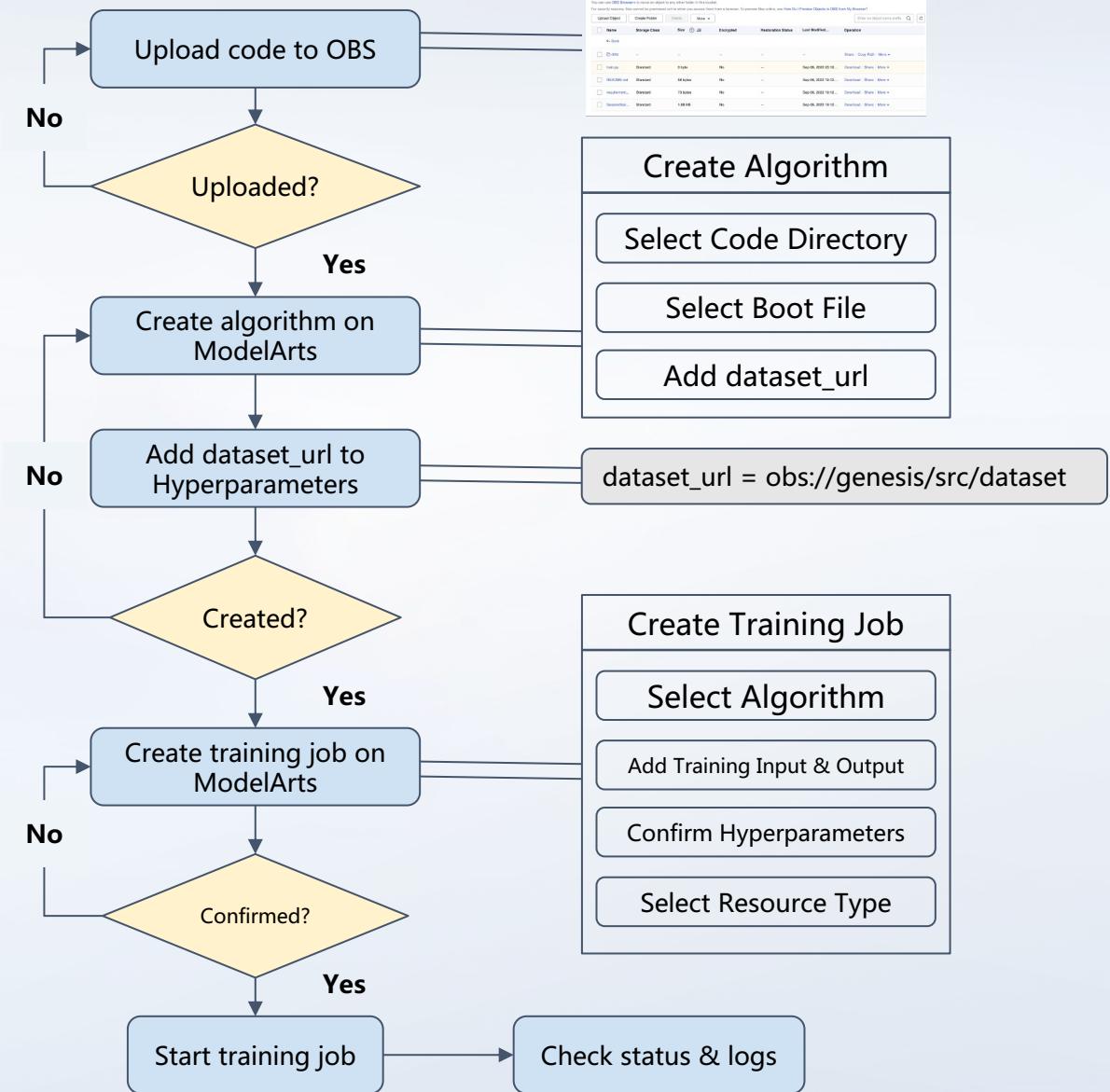
3

Create **Training Jobs** on Huawei Cloud ModelArts

4

Start training and **post-training**

PROCESS FLOW





Innovations

End-to-end synthetic data generation and analytics platform **powered by cloud**

Paradigm shift in healthcare data architecture

Product Uniqueness

Cloud-native synthetic data generation



Business Model

Platform-as-a-service

- Proprietary GAN algorithm that is optimised for **complex, multi-table, relational datasets**
- End-to-end process for users to perform **ML analytics** directly on the dataset, thus reducing user complexity and eliminating issue of big data transfer
- **Evaluation benchmarks** available for users to validate **fidelity** and **fairness** of synthetic datasets
- Entirely **cloud-native** to improve **data security compliance**

- **PDPA/GDPR** compliance under Huawei Cloud services
- Not selling data, but selling **analytics services**
- **Subscription model** rather than once-off payments for datasets

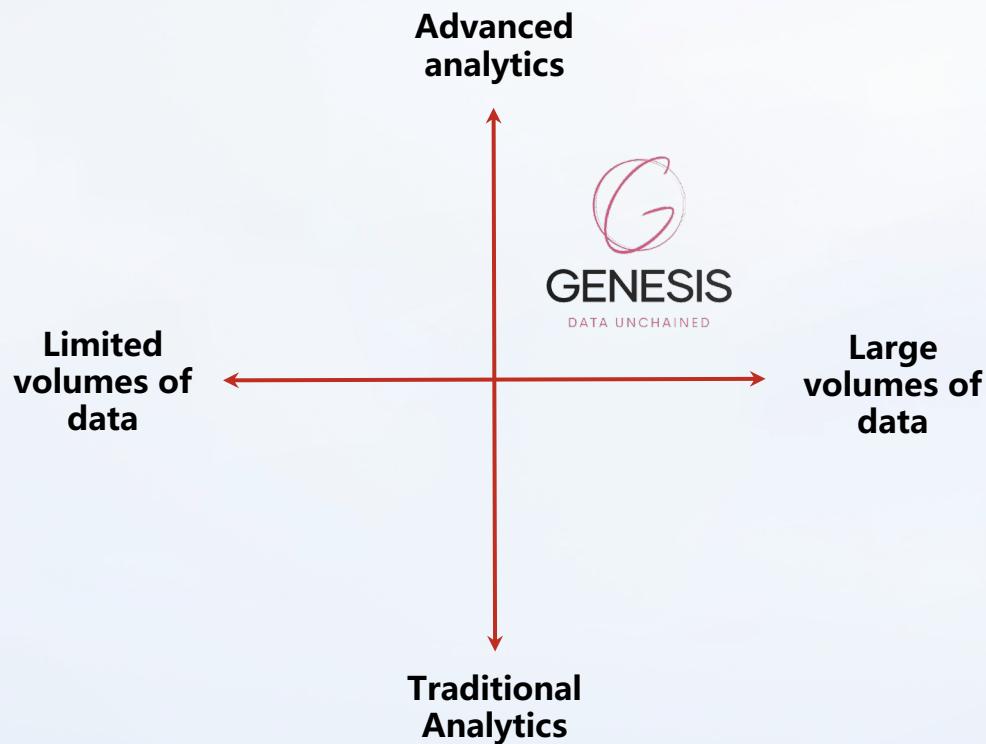


Business value

Primary care clinics are willing to spend up to **\$5,000 per month*** on analytics

Healthcare is rapidly moving to the front quadrant

- Primary Care Networks want accessible tools to perform cutting-edge analysis on clinical data
- But patient data privacy concerns limit meaningful data analysis



PRIMARY CARE NETWORKS (PCN)

PCNs are networks of GPs supported by nurses and care coordinators in providing coordinated care for patients with chronic conditions

Patient-facing

Ancillary services for chronic disease management
Nurse counselling

Backend

Chronic disease registry
Monitoring clinical outcomes
Care coordination

CASE STUDY



- Wants **data-driven insights** on how to improve health outcomes for chronic disease patients e.g. predictive modelling
- Lacks **data volume** and **technical expertise** to derive actionable insights



Business value

GENESIS is solving a difficult problem in the new data paradigm

Pattern of incomplete and unrepresentative data prevalent across industries

- AI training dataset market to reach **> USD 8.6bn** by 2030; with CAGR over **22%**
- GENESIS' underlying technology can be rapidly customised for many sectors

The market for synthetic data is rapidly growing...

GP Clinical Records

Insurance

Other industries

- Direct sales to GP clinics, on subscription model
- Partnership with Primary Care Networks to distribute GENESIS to affiliate GP clinics, on larger contracts
- Pursue research partnerships with IHLs, GPs to validate feasibility of synthetic data analytics

Reasons for insurance market:

- Sensitive attributes need to be protected
- Actuarial scenarios require large amounts of data for testing and validation
- Challenges in data protection are similar to healthcare

To be further evaluated:

- Government services
- Finance and trading
- Political campaigning

 datagen

2022

\$50mn (Series B)

MOSTLY.AI

2022

\$25mn (Series B)

 Synthesis.ai

2022

\$17mn (Series A)

*Based on our customer validation and PCNs' estimated expenditure on data science products, we assess a willingness to pay of up to \$5,000/mth for QuickForm



Achievements

GENESIS prototype is in **development** and **consultation** with potential clients

Achievements

Technical

Key Targets

- Benchmark RapidGAN against state-of-the-art synthetic data generation models (**4Q 2022**)
- Develop RapidGAN models and evaluation metrics (faithfulness, fairness) for complex multivariate datasets, relational schemas, time-series, images (**2Q 2023**)
- Publish benchmarking papers at reputable AI/ML conferences (**4Q 2023**)

Product

- Build MVP for synthetic data generation and evaluation metrics on Python notebooks (**4Q 2022**)
- Build beta version of modern UI/UX for synthetic data generation platform with user testing (**2Q 2023**)

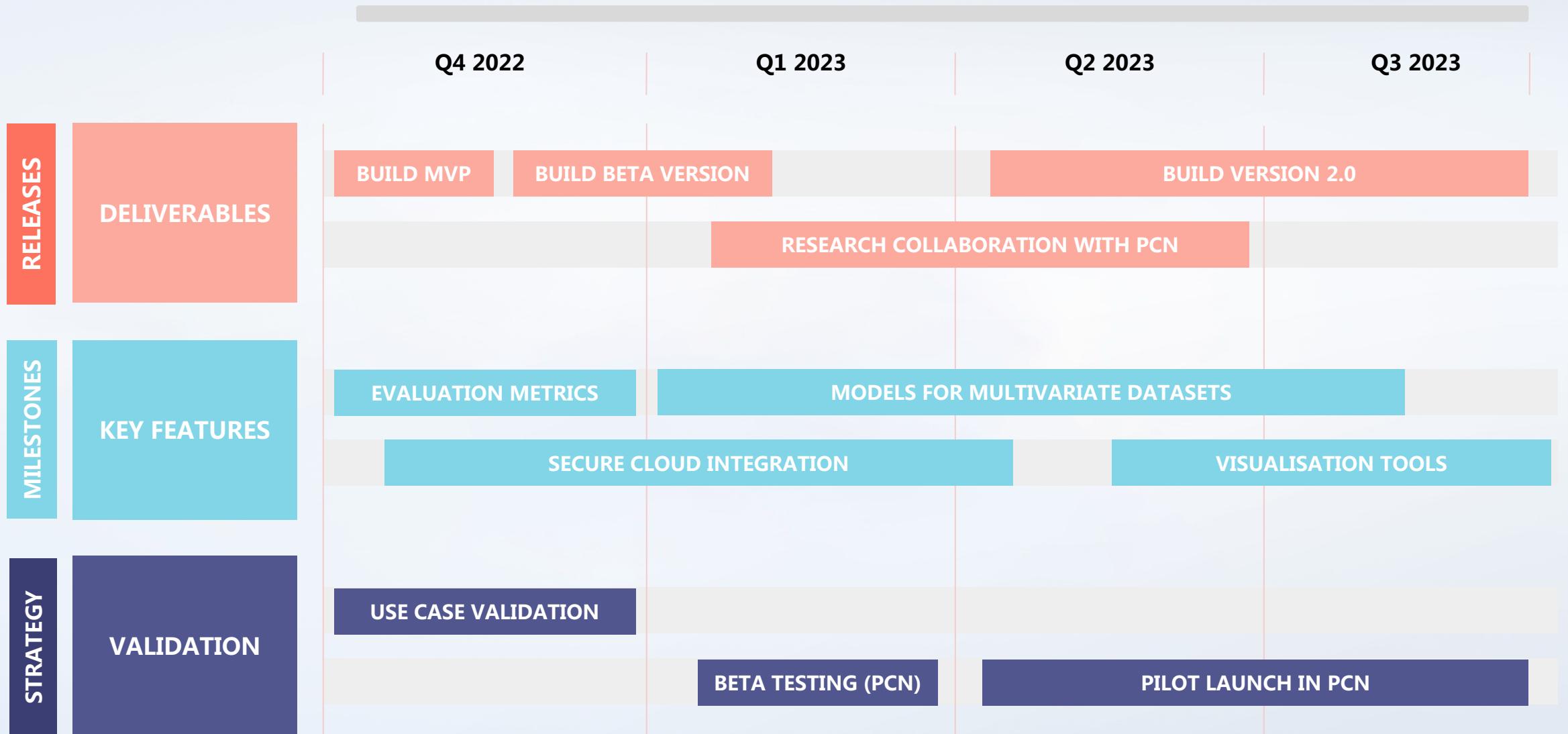
Commercial

- Research collaboration with PCN in Singapore (**2Q 2023**)
- Pilot testing with 3 - 5 healthcare clinics in Singapore (**4Q 2023**)
- Establish MoU with primary care clinic networks for further research (**4Q 2023**)



Project Planning

Indicative timeline for proposed **feature rollout** and **commercial traction**





Appendix: Cloud Services

Use of Huawei Cloud Services



ECS

Cloud servers for low-cost, low-latency AI inference



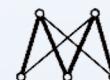
API Gateway

Frontend-backend communication proxy



FunctionGraph

Scalable services compute for API backend



ModelArts

GPU-accelerated training for TensorFlow



GaussDB for NoSQL

Reliable, responsive database backend



OBS

Stores objects such as bulk files



Appendix: Cloud Services

Virtual Private Cloud (VPC)

HUAWEI CLOUD | [Console](#) | [Singapore](#) | [Search](#) | [More](#) | [Intl-English](#) | [hid_d17q_mm9bctzome](#) | [8](#)

vpc-apac

Summary Topology Tags

VPC Information

Name	vpc-apac	ID	7889ccdd-dff0-46a8-bbe7-86898d50e47e
Status	Available	CIDR Block	192.168.0.0/16 Edit CIDR Block
Description	--		

VPC Resources

ECSs	Add	1
BMSs	Add	0
Load Balancers	Add	0
Network Interfaces	Add	4

Networking Components

Subnets	1
Route Tables	1

Related Services

NAT Gateway [Learn more](#)

A public NAT gateway allows cloud servers in a VPC to share EIPs for Internet access, which keeps costs down while also improving security by hiding your servers behind EIPs.

VPC Peering Connections [Learn more](#)

A VPC peering connection enables you to route traffic between two VPCs by using private IP addresses. ECSs in either VPC can communicate with each other just as if they were in the same VPC. You can create a VPC peering connection between your own VPCs, or between your VPC and a VPC of another account within the same region.

⋮

⋮

⋮



Appendix: Cloud Services

Elastic Cloud Server (ECS)

HUAWEI CLOUD | Console | Singapore | Search | Billing Center | Resources | Service Tickets | Enterprise | Support | English | hid_dl7q_mm9bctzome | 8

< | ecs-apac | Remote Login | Start | Stop | Restart | More | C

Summary | Disks | NICs | Security Groups | EIPs | Monitoring | Tags

ECS Information

ID	0ad238ff-0009-4256-92c7-5257907ca731
Name	ecs-apac
Region	Singapore
AZ	AZ1
Specifications	General computing s3.medium.2 1 vCPUs 2 GiB
Image	CentOS 7.6 64bit Public image
VPC	vpc-apac

Billing Information

Billing Mode	Pay-per-use
Obtained	Aug 28, 2022 00:15:41 GMT+08:00
Launched	Aug 28, 2022 00:15:57 GMT+08:00

Management Information

ECS Group	-- Create ECS Group
Agency	-- Create Agency
License Type	Use license from the system

Running | Monitoring | Monitoring

Disks
System Disk
[ecs-apac](#) High I/O | 40 GiB

NICs
Primary NIC
[subnet-web](#) 192.168.0.66 | 119.8.181.197

Security Groups
[default](#)

EIPs
[119.8.181.197](#) | 5 Mbit/s

Cloud Backup and Recovery
The ECS has not been backed up.
After an ECS is backed up, you can use the backup data for server or disk restoration, ensuring service security. [Back Up Now](#)



Appendix: Cloud Services

Elastic IP (EIP)

HUAWEI CLOUD | [Console](#) | [Singapore](#) | [Search](#) | [More](#) | [Intl-English](#) | hid_dl7q_mm9bctzome | [8](#)

Summary Bandwidth Tags Bind Unbind C

EIP 119.8.181.197 EIP Type Dynamic BGP
Assigned Aug 28, 2022 00:16:06 GMT+08:00 Name --
ID 8acfa591-99fa-4a3e-8cf5-0a7c258cb5ba [View Metric](#) Status Bound

Associated Instance

Instance Name	ecs-apac	VPC	vpc-apac
Instance ID	0ad238ff-0009-4256-92c7-5257907ca731	Subnet	subnet-web
Instance Type	Server	Status	Running
AZ	AZ1	Bound NICs	192.168.0.66

Help Call Support



Appendix: Cloud Services

Elastic Cloud Server (ECS)

HUAWEI CLOUD | Console | Singapore | Search | Billing Center | Resources | Service Tickets | Enterprise | Support | English | hid_dl7q_mm9bctzome | 8

< | ecs-apac | Summary | Servers | Backups | Snapshots | Tags | Expand Capacity | C

Basic Information

ID	77894113-d5c1-4d70-80fd-fe105c30b912
Name	ecs-apac
Region	Singapore
AZ	AZ1
Disk Type	High I/O
Capacity (GB)	40
Max. IOPS	IOPS limit: 2,120, IOPS burst limit: 5,000
Function	System disk
Image	CentOS 7.6 64bit
Created	Aug 28, 2022 00:15:43 GMT+08:00

In-use

Servers	1
Backups	0
Snapshots	0

Servers

- ecs-apac View Metric Running

Backups

You have not created any backup yet.
CBR allows you to create backups for EVS disks on the console without stopping servers.

Snapshots

You have not created any snapshot yet.
Snapshots can be created to quickly save disk data at specified time points.

Attach Disk | Create Backup | Create Snapshot

Configuration Information

Disk Sharing	Disabled
Device Type	VBD
Encrypted	No
KMS Key Name	--
KMS Key ID	--
Source Backup ID	--

...



Appendix: Cloud Services

Object Storage Service (OBS)

HUAWEI CLOUD | [Console](#)

Search [More](#) [Intl-English](#) hid_dl7q_mm9bctzome | [8](#)

Versioning [Disabled](#) Storage Class [Standard](#) Task Center

Overview Objects Metrics NEW Permissions Basic Configurations Domain Name Mgmt Cross-Region Replication Back to Source Data Processing Inventories

Objects Deleted Objects Fragments

Objects are basic units of data storage. In OBS, files and folders are treated as objects. Any file type can be uploaded and managed in a bucket. [Learn more](#)
You can use [OBS Browser+](#) to move an object to any other folder in this bucket.
For security reasons, files cannot be previewed online when you access them from a browser. To preview files online, see [How Do I Preview Objects in OBS from My Browser?](#)

Upload Object Create Folder Delete More Enter an object name prefix. [C](#)

<input type="checkbox"/>	Name	Storage Class	Size	Encrypted	Restoration Status	Last Modified	Operation
<input type="checkbox"/>	out	--	--	--	--	--	Share Copy Path More ▾
<input type="checkbox"/>	src	--	--	--	--	--	Share Copy Path More ▾
<input type="checkbox"/>	synthetic_data	Standard	227.34 KB	No	--	Sep 06, 2022 22:01:14 GMT+0...	Download Share More ▾

Help (?) Call (?)