HUAWEI

GENESIS

DATA UNCHAINED

/ You make all the difference /

Huawei Developer Competition

Spark Infinity

— 2022 —

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

HUAWEI CLOUD
EVERYTHING AS A SERVICE

Project Name: GENESIS

Team Name: GENESIS

Email: xxxxxxxxxxx@u.nus.edu

Phone: +65 xxxxxxxx

Demonstration URL:

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

Overview

Project Overview

Project Name	*GENESIS 合创*
Team Name	*GENESIS 合创*
Contacts	*Zacchaeus Chok*
Technical Field	*Synthetic Data Generation, Machine Learning Analytics*
Technologies	*OBS, GaussDB, ModelArts, Generative Adversarial Networks, Deep Learning*
Keywords	*Cloud-native, Synthetic, Analytics*
Applicable Fields	*Healthcare, Insurance, Government, Finance*
Description (in 500 words)	*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.*

Contents

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

NYC NATIONAL YOUTH COUNCIL SINGAPORE

Resume diagnosis using NLP

500

Digital marketing using NLP

SMU SINGAPORE MANAGEMENT UNIVERSITY

Data management interface for clinics

Our Advisory Panel

Dr Lim Chien Chuan
Director
i-Care Primary Care Network

icare Primary Care Network

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

Project Introduction

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

Current State

Datasets

Existing healthcare datasets are:

(1) incomplete
(2) small-scale
(3) non-representative

Analytics

Existing AI/ML tools are:

(1) computationally expensive
(2) inaccessible for non-technical users
(3) does not protect sensitive attributes
(4) not security compliant

Practitioners cannot extract value from the clinical data they collect

Ideal State

Datasets

Healthcare datasets that are:

(1) large-scale
(2) high-quality
(3) demographically representative

Analytics

Access to **AI/ML analytics** for healthcare purposes that protects sensitive attributes

Ability to implement data-driven clinical insights in real-world practice

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

DIAGNOSIS

01 **Incomplete data**
Poorly annotated, fragmented datasets across legacy systems

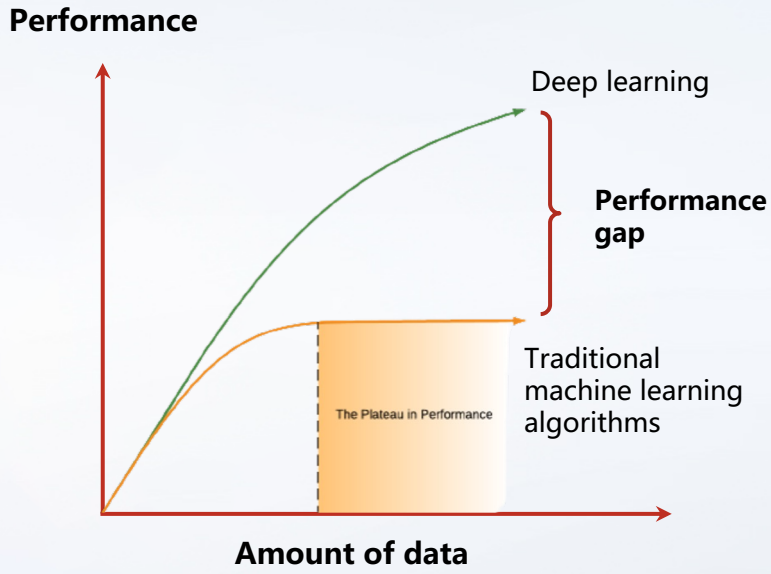
02 **Small-scale data**
Datasets are usually too small (1000s of records) for meaningful machine learning analysis

03 **Non-representative data**
Does not represent underlying population demographics patterns

SOLUTION

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.**

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).

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

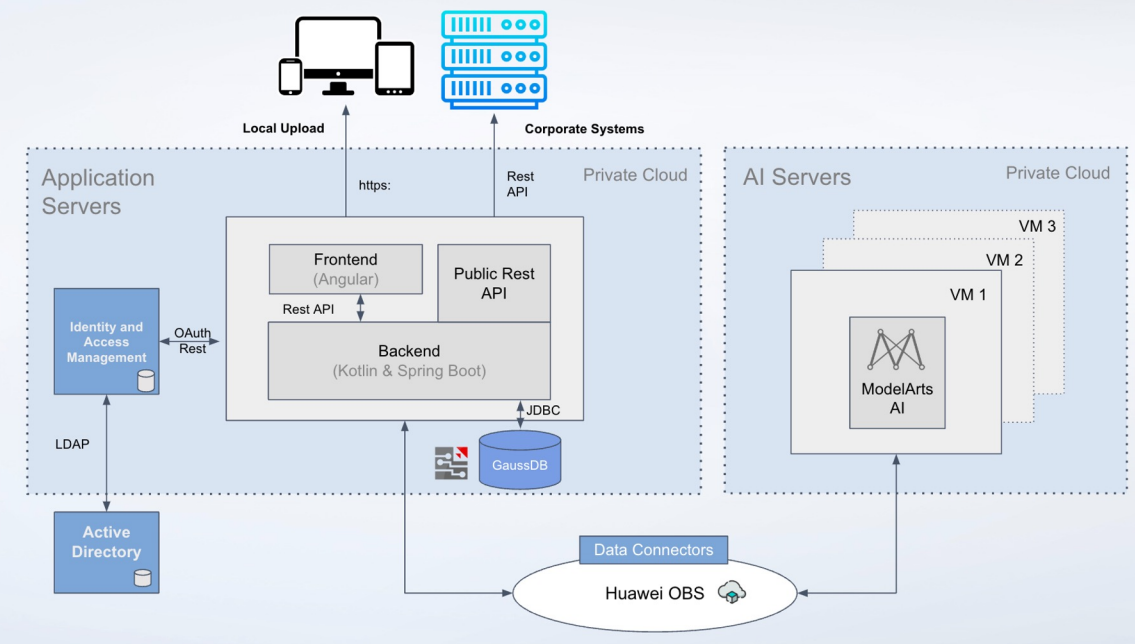
ii) Results and **visualisations** can be exported out to local networks

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

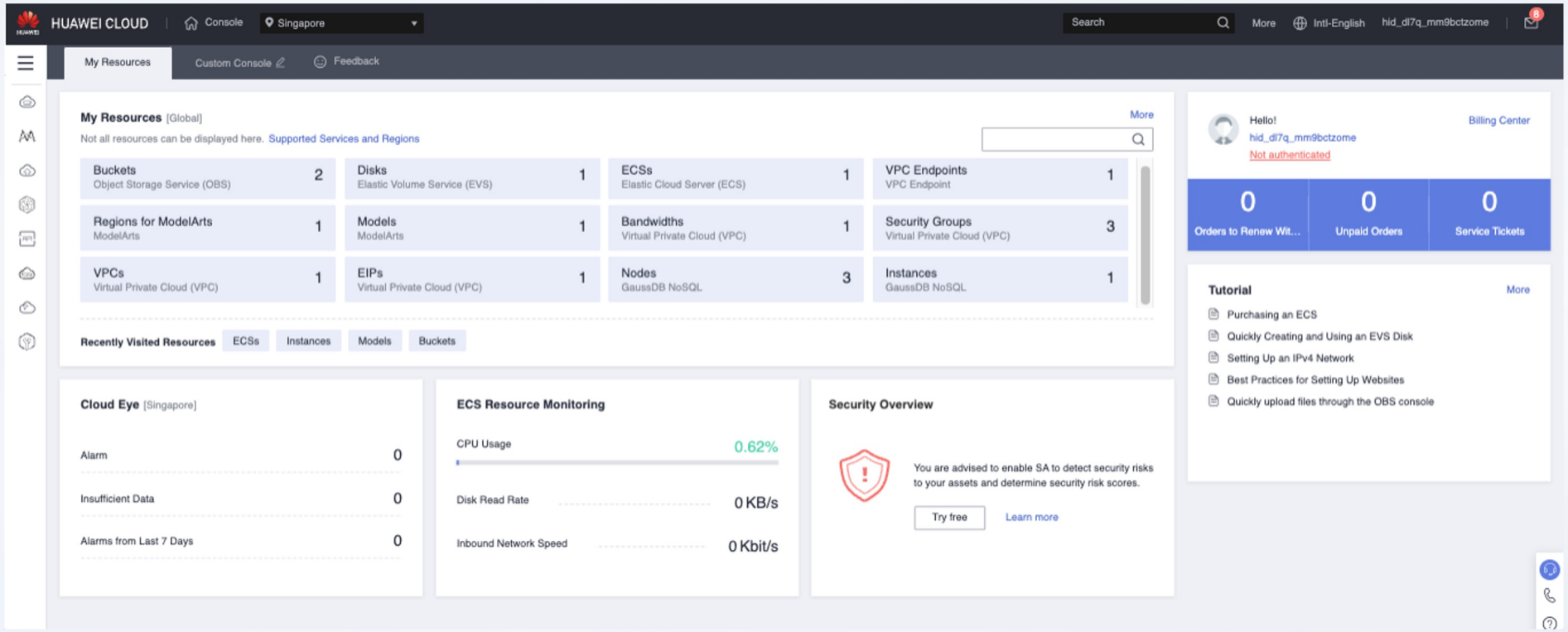
- Elastic, scalable storage system handling high volume of uploaded data
- Real-time file processing that triggers file operations
- Real-time deployment of RapidGAN models on highly-concurrent data

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



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

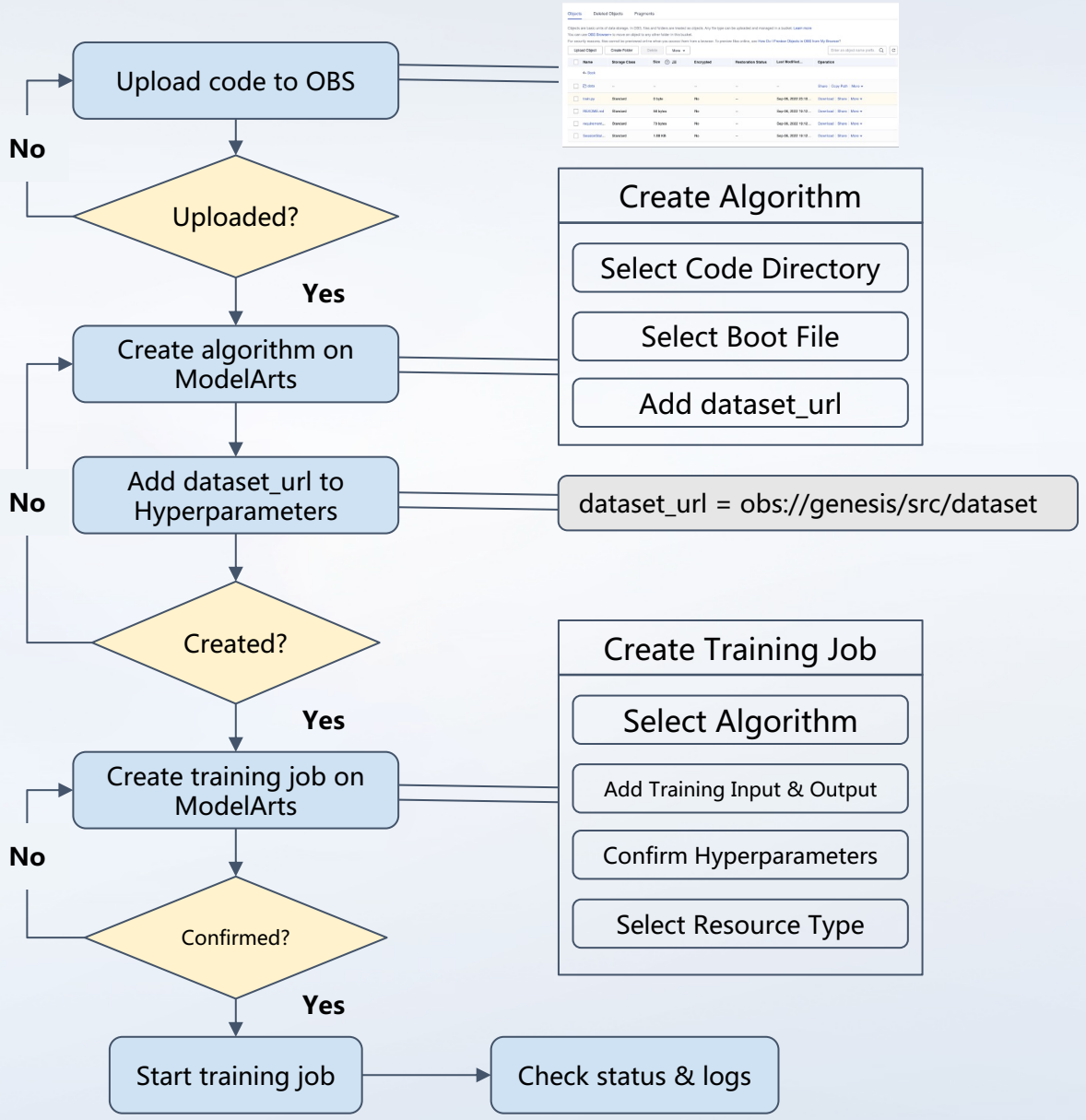
Cloud Services
Model deployment on ModelArts

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

- 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**

Business Model
Platform-as-a-service

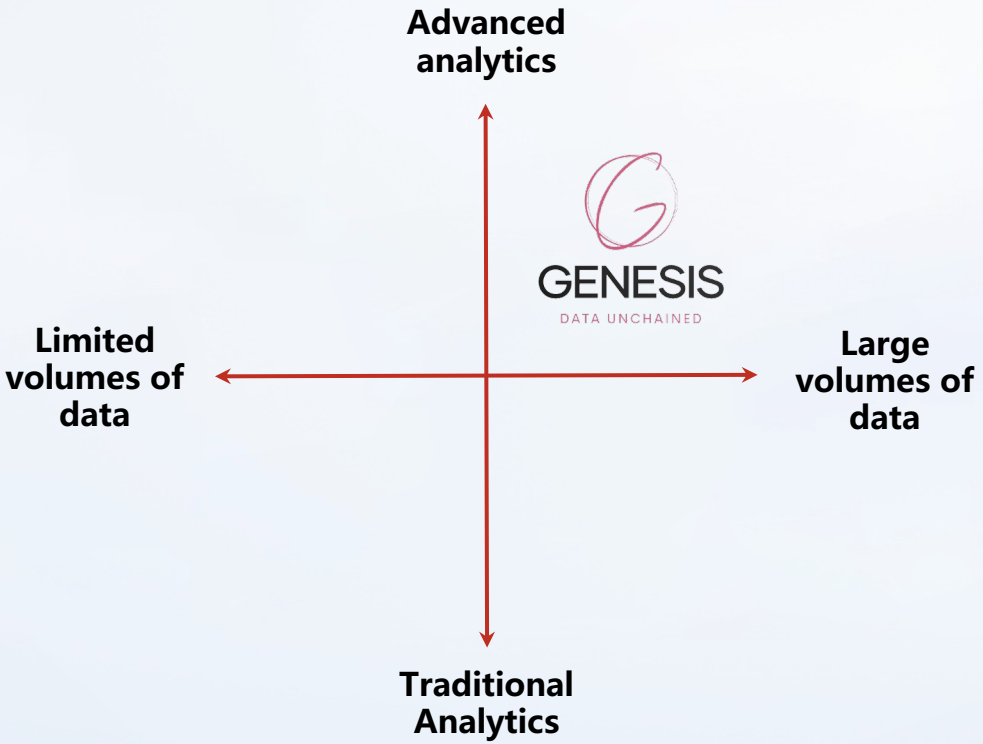
- **PDPA/GDPR** compliance under Huawei Cloud services
- Not selling data, but selling **analytics services**
- **Subscription model** rather than once-off payments for datasets

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	Backend
Ancillary services for chronic disease management	Chronic disease registry
Nurse counselling	Monitoring clinical outcomes
	Care coordination

CASE STUDY

Jcare
Primary Care Network

- 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*

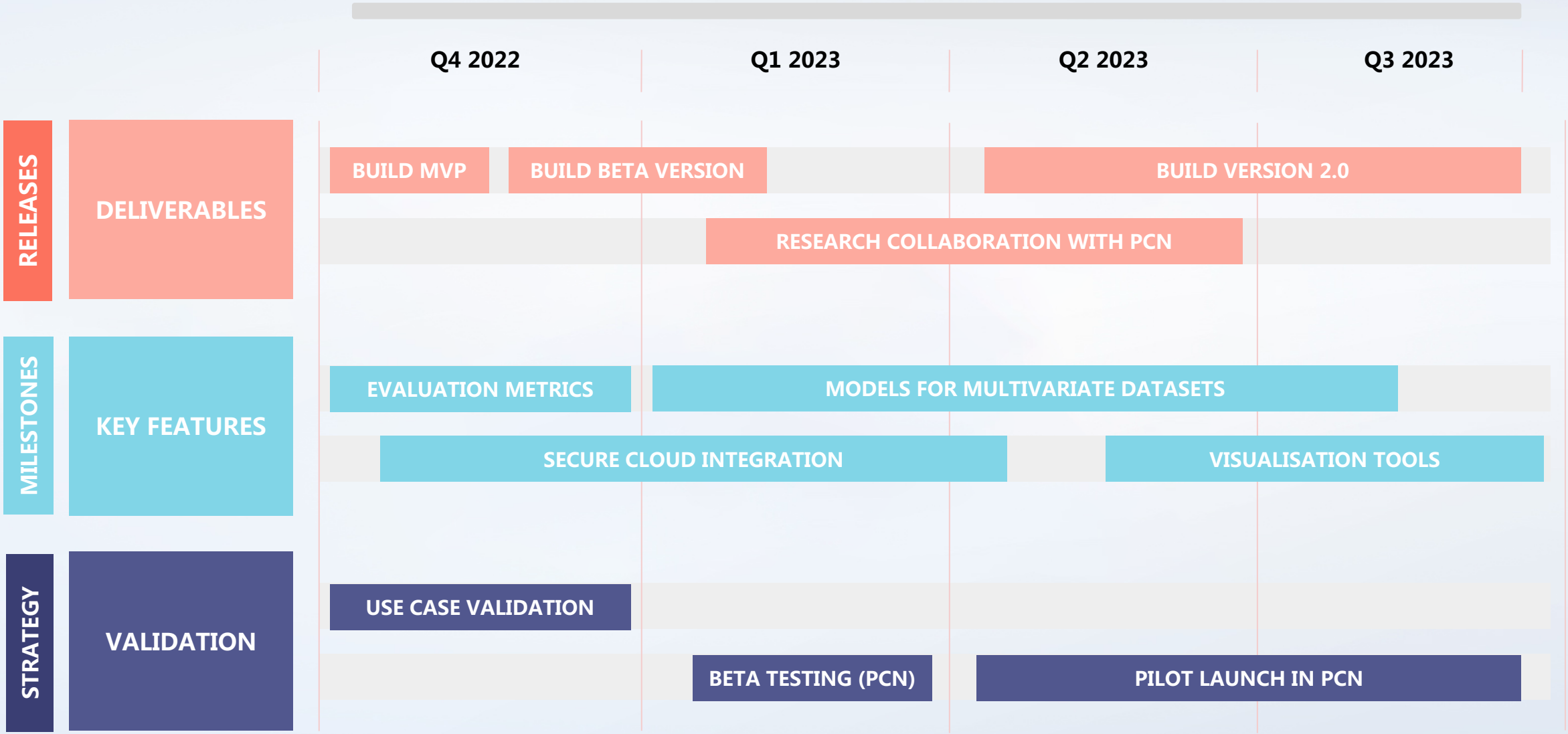
Achievements

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

Achievements	Key Targets
Technical	• 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

FunctionGraph
Scalable services compute for API backend

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

HUAWEI CLOUD
Grow with Intelligence

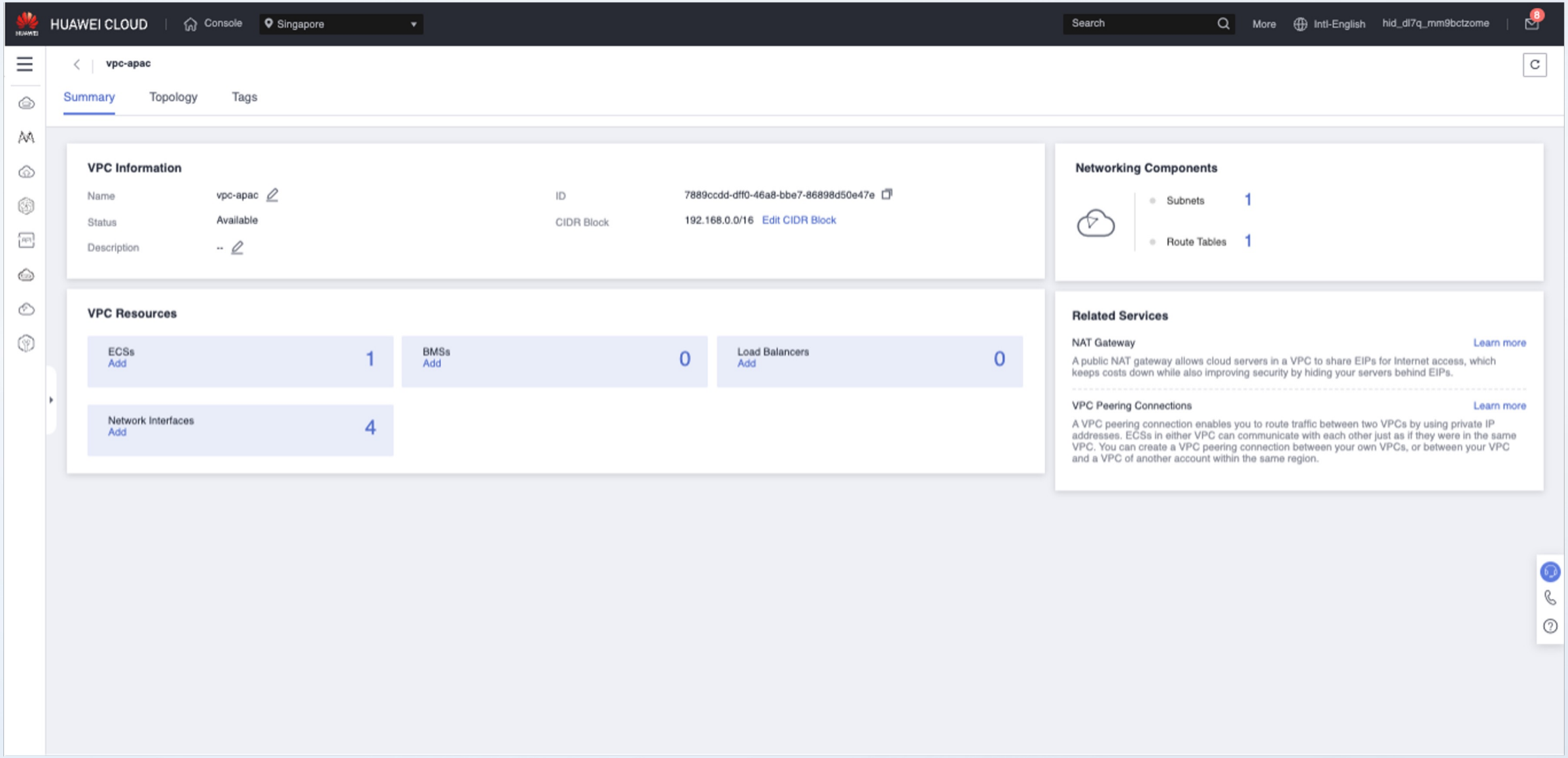
GaussDB for NoSQL
Reliable, responsive database backend

API Gateway
Frontend-backend communication proxy

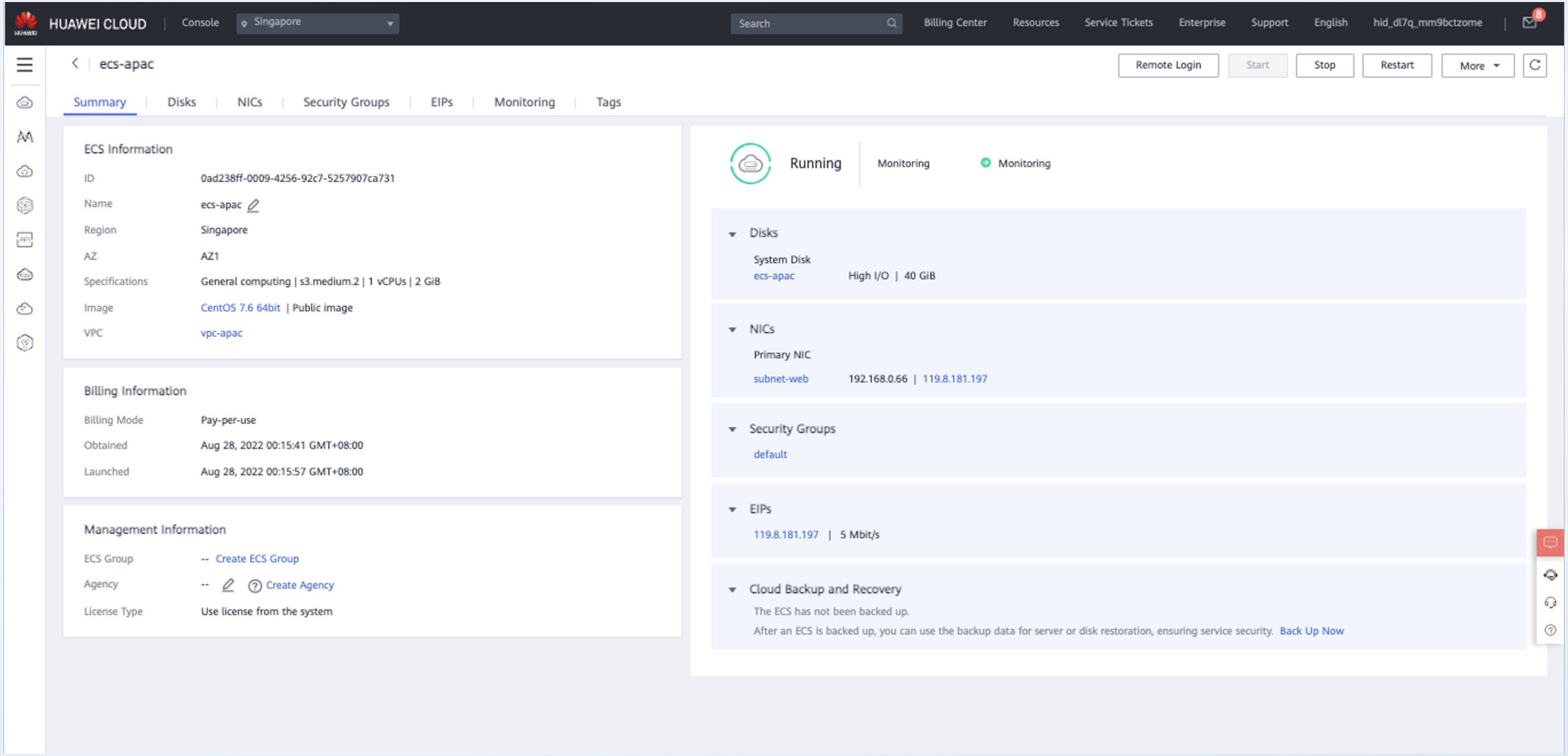
OBS
Stores objects such as bulk files

ModelArts
GPU-accelerated training for TensorFlow

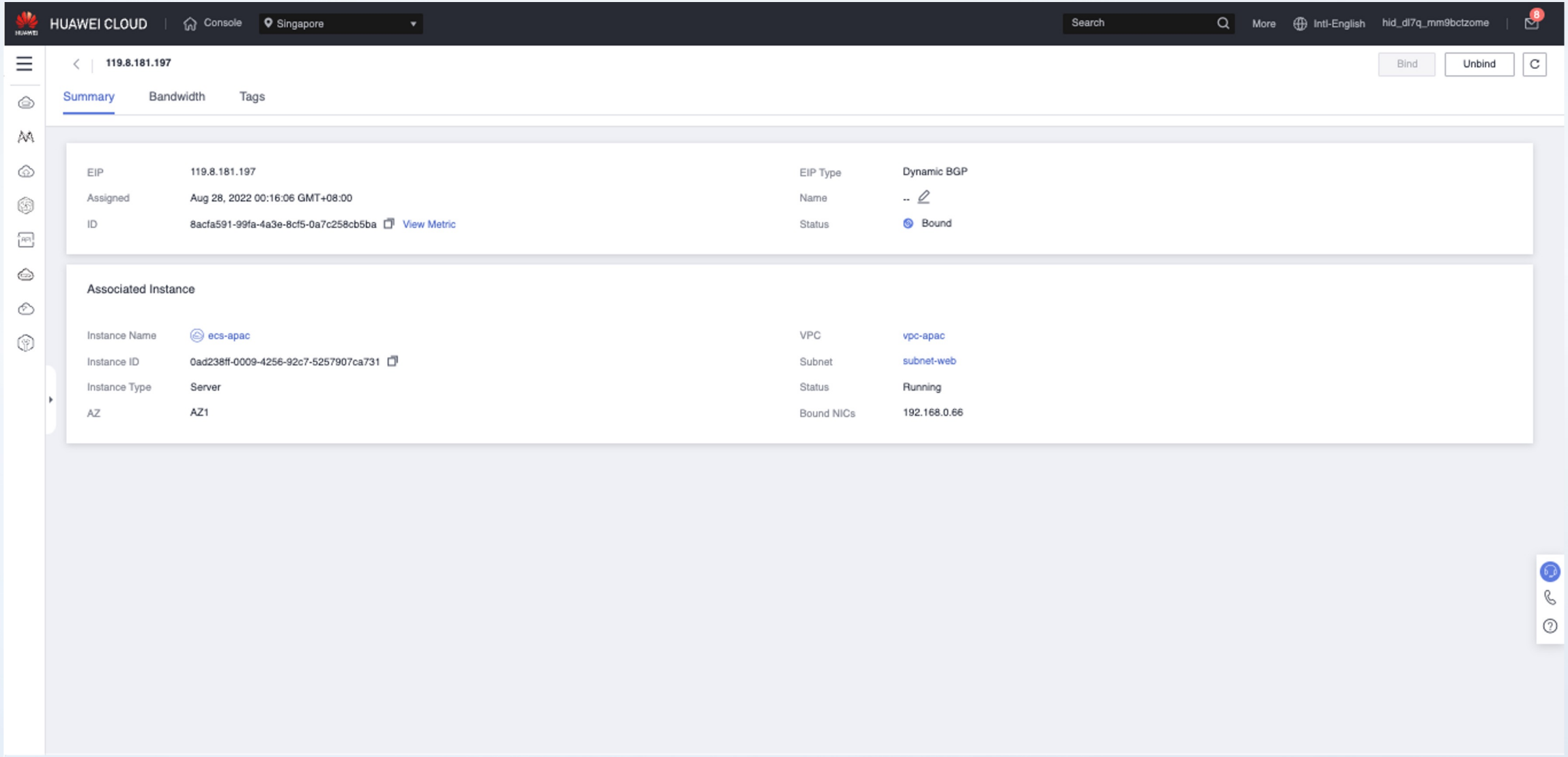
Appendix: Cloud Services
Virtual Private Cloud (VPC)



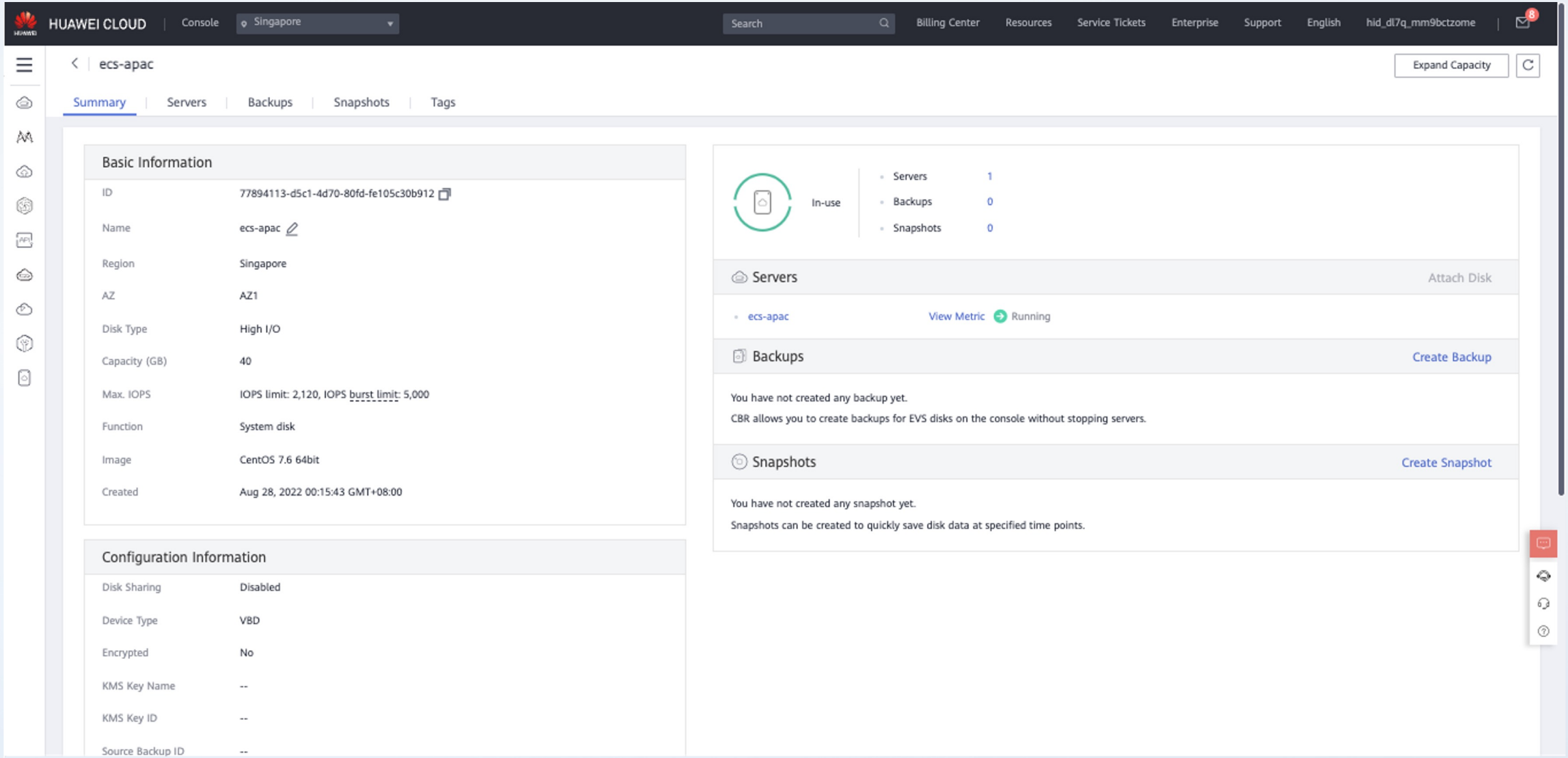
Appendix: Cloud Services
Elastic Cloud Server (ECS)



Appendix: Cloud Services
Elastic IP (EIP)



Appendix: Cloud Services
Elastic Cloud Server (ECS)



Appendix: Cloud Services
Object Storage Service (OBS)