**Health Chatbot Deployment Proposal**

**By Terry Tu**

**Server Requirement:**

#### **1. On-Premises Server for Model Deployment**

**Hardware Specifications**:

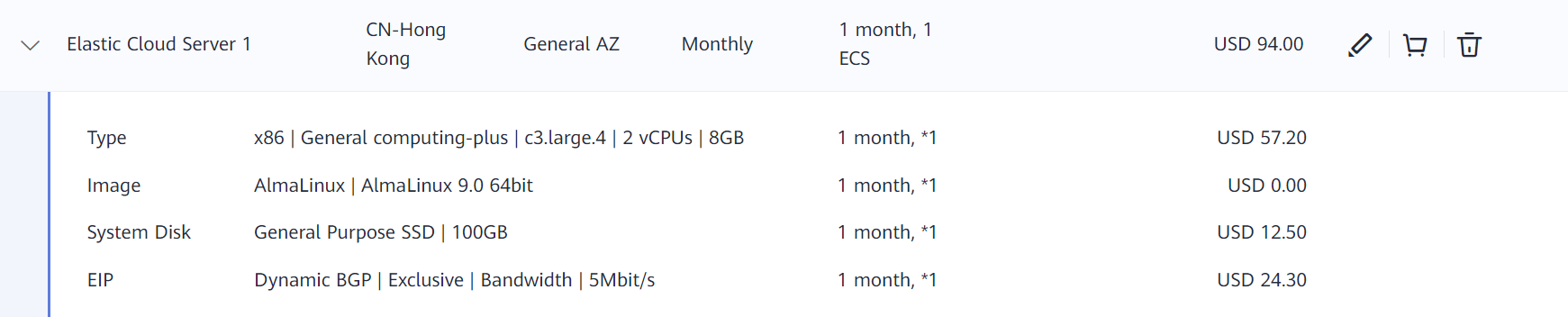
* **GPU**: 2 x NVIDIA RTX 3090
* **CPU**: High-performance multi-core CPU
* **Memory**: 64GB to 128GB RAM
* **Storage**: 1TB NVMe SSD for fast access

#### **2. Cloud Database for User Data**

**Cloud Provider Options**:

* **Huawei Gauss DB (Recommended)**

100 GB SSD, 8 GB RAM, 2vCPUs, 5 Mbit/s network



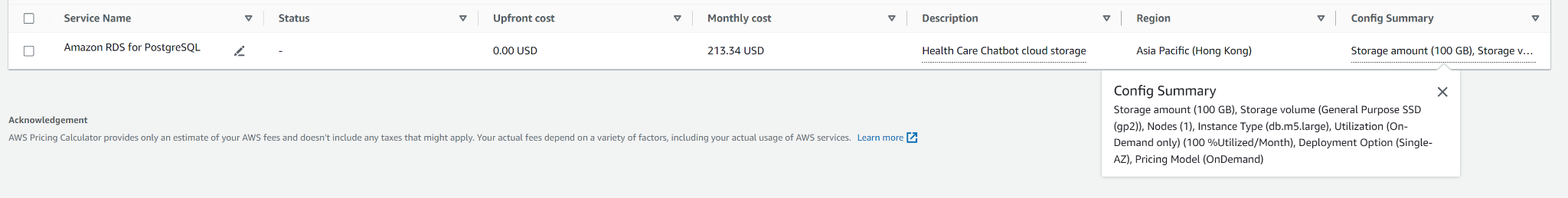
Cost: 94 USD/Month

Price Calculator: <https://www.huaweicloud.com/intl/en-us/pricing/#/ecs>

* **Amazon RDS (PostgreSQL)**

Storage amount (100 GB), Storage volume (General Purpose SSD (gp2)), Nodes (1), Instance Type (db.m5.large), Utilization (On-Demand only) (100 %Utilized/Month), Deployment Option (Single-AZ), Pricing Model (OnDemand)

Cost: 213.34 USD/Month



Price Calculator: <https://calculator.aws/#/estimate>

* **Google Cloud SQL (PostgreSQL)**

2 vCPUs, 8 GB RAM, 100 GB SSD

Cost: 117.01 USD/Month (Have discount when reserved more than 1 year)

Price Calculator: <https://cloud.google.com/products/calculator?dl=CiQzYTRlMDM4OS02MzI2LTRlNzMtOWVjNy01MTAzZWU0ZDI5MjgQBxokN0FBNTI0NUQtMkNBNi00RjY5LTk1QjctODIyODhBRDgwQzYz>

Notes:

1. We can use less RAM if the price is too high (i.e. 4GB RAM)
2. We can use less Storage for the user data (i.e. 50 GB SSD)
3. 2 vCPUs is the minimum we can get
4. According to online reviewers, AWS has the best performance, but Huawei’s Gauss DB is the cheapest for the same hardware specs.