Business Capability Model

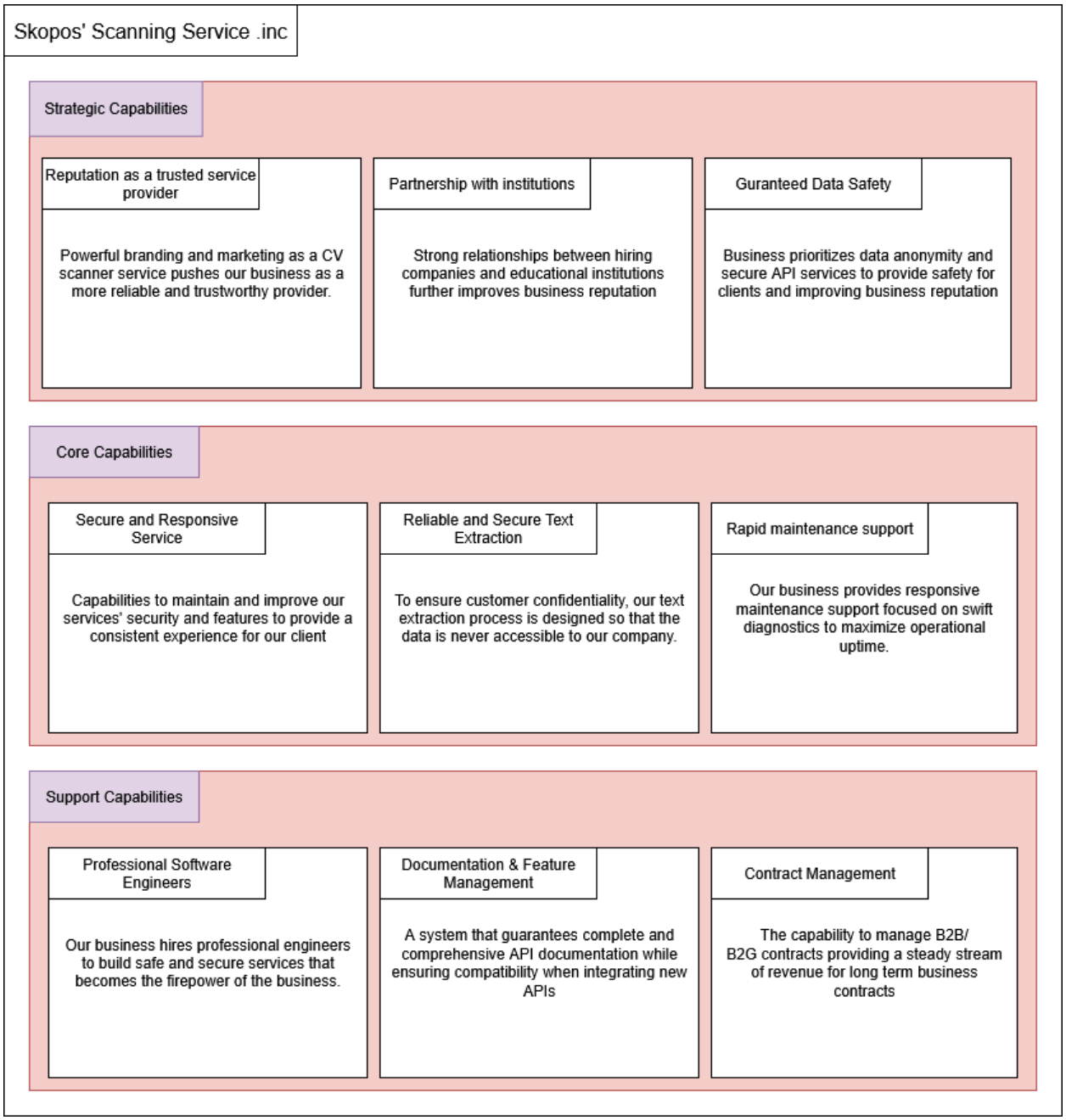

Image 1. Business Capability Model

Our business model provides a service API for clients, which utilizes Oauth2 authentication for safety and restriction use of our service. The service is a freemium, it divides clients into free tier users and premium tier users, where the premium tier provides more advanced features. The free tier only extends to text extraction from a PDF file, but the premium tier involves OCR technology scanning text

from images in the PDF. Accessing the premium tier requires a client to pay a subscription fee to access the premium service. Aside from text extraction, the service supports keyword search using a fast algorithm “Aho-Corasik” and supports multiple keyword search at once, this feature is free for all clients. The API service is run in a cloud infrastructure of Azure, provided by Microsoft, with a containerized environment with Docker to ensure consistency between development and production while also opening more ports for different applications in a cloud computing environment.

For the service itself, we utilize the programming language python using a backend framework, FastAPI, with documentation provided by SWAGGER API Documentation (access the [host]/docs endpoint). The framework is paired with database & data modelling libraries, i.e. sqlalchemy as the ORM and Pydantic to model the data for client request. The service is connected to a Postgresql database to store client information, like client id, client email, and its URI. For authentication and security, the service utilizes python-jose’s cryptography library (JWT) for access token, and Bcrypt for generating client secrets when registering. The endpoint is protected with FastAPI’s built-in OAuth2PasswordBearer class so unauthorized access isn’t allowed. For text extraction and keyword search, we utilize libraries such as PyMuPDF (Fitz), Pytesseract, and Pyahocorasick.