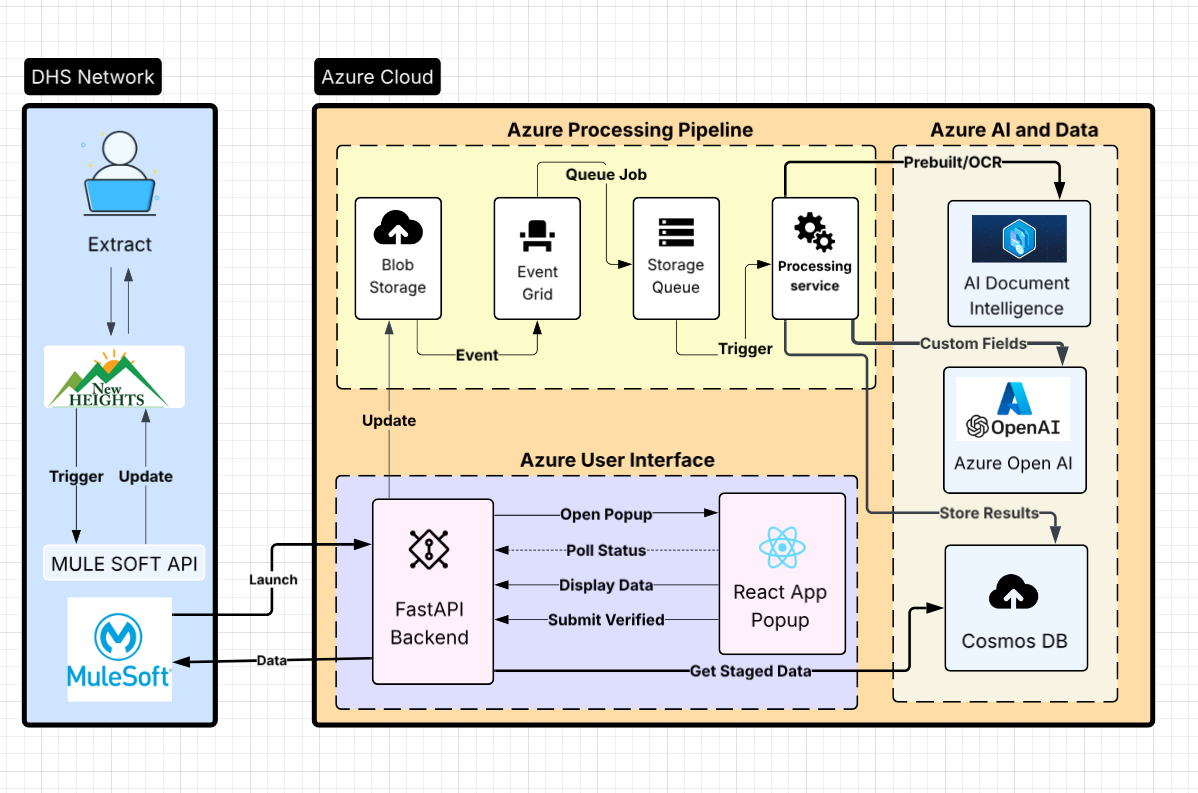
****

**Architectural Overview**

The proposed architecture utilizes an event-driven, asynchronous model to process pay stub documents efficiently and at scale. This design decouples the user-facing legacy application from the intensive background processing, ensuring high availability and responsiveness for the end-user. The workflow is initiated from the on-premise environment and leverages a suite of managed Azure services for processing, AI-driven data extraction, and user interaction, before returning the final data to the source system.

**Component Breakdown and Data Flow**

**1. DHS Network (On-Premise Environment)**

* **Caseworker & New Heights Application:** This is the point of initiation. The end-user operates exclusively within the existing New Heights application.
* **MuleSoft API:** This established integration gateway acts as the secure bridge between the internal DHS network and the Azure cloud platform. When a user triggers an extraction, the New Heights application makes an API call to MuleSoft.

**2. Azure Cloud (Processing Engine)**

MuleSoft forwards the request and document payload to the dedicated Azure environment where the intelligent processing occurs.

**A. Azure User Interface (Service Type: Container & Web Hosting)**

* **FastAPI Backend (Azure Container Apps):** This Python application, running as a managed **container**, serves as the primary API endpoint and orchestration layer for the entire workflow. It receives the initial request, launches the user interface, and coordinates the data flow between all services.
* **React App Popup (Azure Static Web Apps):** A modern user interface, hosted on a specialized **web hosting** service, which launches in a popup window. It provides the caseworker with real-time status updates and a final screen for Human-in-the-Loop (HITL) verification.

**B. Azure Processing Pipeline (Service Type: Storage & Messaging)**

* **Blob Storage (Object Storage):** Provides secure, ephemeral storage for the pay stub document during the analysis phase.
* **Event Grid & Storage Queue (Messaging Services):** These services form a resilient asynchronous pipeline. Upon file upload to Blob Storage, **Event Grid** publishes an event that places a job message into the **Storage Queue**. This queuing mechanism reliably manages high volumes of concurrent requests.
* **Processing Service (Azure Container Apps):** A dedicated background service, running as a **container**, responsible for executing the AI analysis. It polls the Storage Queue for new jobs to process.

**C. Azure AI and Data (Service Type: AI & Database)**

* **AI Document Intelligence (AI Service):** A specialized AI model used to perform OCR and extract standard, structured fields from the pay stub document.
* **Azure OpenAI (AI Service):** large language model (LLM) leveraged for more complex extraction tasks, using reasoning to identify and structure custom fields from the OCR text.
* **Cosmos DB (NoSQL Database):** A flexible database that serves as the staging and job-tracking repository. It stores the status of each job and the final, consolidated JSON data extracted by the AI services.

**Monitoring and Logging**

**Azure Monitor** is a foundational service which is used for providing comprehensive observability across the entire platform.

* **Function:** Azure Monitor automatically collects logs, performance metrics, and error data from all services within the Azure Cloud environment, including the FastAPI Backend, the Processing Service, Cosmos DB, and the Storage Queue.
* **Implementation:** Both the FastAPI and Processing Service applications are instrumented to send structured logs (e.g., job start/end times, AI model performance, error details) to Azure Monitor's Application Insights feature.
* **Purpose:** This centralized logging enables efficient troubleshooting and operational oversight. It allows support personnel to trace a specific job ID through the entire workflow, analyze performance bottlenecks, and diagnose errors without direct access to the application servers. This capability is critical for maintaining the health and reliability of the system in a production environment.
* **Azure Policy:** It is used to set security hard stops for azure environment like AWS guardrails.

**Detailed Cost Breakdown**

This estimate is based on the production workload of **30,000 documents per month** and **100+ daily active caseworkers**. All prices are based on the "Pay-As-You-Go" model in a standard US region (e.g., East US) and are subject to change based on Microsoft's pricing.

**1. Generative AI Services**

* **Service:** **Azure AI Document Intelligence**
* **Estimated Cost:** **$435 / month**
* **Calculation:** The total cost is a sum of the models used for each of the 30,000 documents.
  + Read (OCR): $1.50 per 1,000 pages = $45
  + Custom Classification: $3.00 per 1,000 pages = $90
  + Prebuilt: Paystub: $10.00 per 1,000 pages = $300
  + **Total:** $45 + $90 + $300 = **$435**
* **Pricing Link:** [Azure AI Document Intelligence Pricing](https://azure.microsoft.com/en-us/pricing/details/ai-document-intelligence/)
* **Service:** **Azure OpenAI Service**
* **Estimated Cost:** **~$90 / month**
* **Calculation:** Based on our earlier analysis of ~432 tokens per document. Using the GPT-4 model at approximately $0.03/1k input tokens and $0.06/1k output tokens.
  + 30,000 documents \* 432 tokens/doc ≈ 13 million tokens
  + The cost for this volume is highly dependent on the input/output ratio but is estimated to be around $90. This is a conservative estimate.
* **Pricing Link:** [Azure OpenAI Service Pricing](https://www.google.com/search?q=https://azure.microsoft.com/en-us/pricing/details/openai-service/)

**2. Infrastructure**

* **Service:** **Azure Container Apps**
* **Estimated Cost:** **$40 - $75 / month**
* **Calculation:** This covers both the FastAPI Backend and the Processing Service. The cost is based on sustained vCPU and Memory usage. We estimate two containers, each averaging 0.5 vCPU and 1 GiB of memory, with significant usage during business hours and scaling down at night.
* **Pricing Link:** [Azure Container Apps Pricing](https://azure.microsoft.com/en-us/pricing/details/container-apps/)
* **Service:** **Azure Static Web Apps**
* **Estimated Cost:** **$9 / month**
* **Calculation:** For a production application with multiple daily users, the "Standard" plan is required. This plan has a fixed monthly cost.
* **Pricing Link:** [Azure Static Web Apps Pricing](https://azure.microsoft.com/en-us/pricing/details/app-service/static/)

**3. Storage**

* **Service:** **Azure Cosmos DB**
* **Estimated Cost:** **$60 - $120 / month**
* **Calculation:** To support 100+ concurrent users, we move from the serverless model to **Provisioned Throughput**. This estimate is for ~1000 Request Units per second (RU/s) plus the cost for storing several gigabytes of JSON data.
* **Pricing Link:** [Azure Cosmos DB Pricing](https://azure.microsoft.com/en-us/pricing/details/cosmos-db/) (See "Provisioned Throughput" section).
* **Service:** **Azure Blob Storage**
* **Estimated Cost:** **$5 - $10 / month**
* **Calculation:** Since files are deleted daily, the storage volume cost is negligible. This cost is almost entirely driven by the high number of transactions: 30,000 uploads + 30,000 reads + 30,000 deletes per month.
* **Pricing Link:** [Azure Blob Storage Pricing](https://azure.microsoft.com/en-us/pricing/details/storage/blobs/)

**4. Monitoring & Security**

* **Service:** **Azure Monitor**
* **Estimated Cost:** **$15 - $25 / month**
* **Calculation:** While there is a large free tier, the high volume of logs from two container apps and other services will likely exceed it. This estimate covers the cost of ingesting and retaining several gigabytes of log data per month.
* **Pricing Link:** [Azure Monitor Pricing](https://azure.microsoft.com/en-us/pricing/details/monitor/)
* **Service:** **Azure Key Vault**
* **Estimated Cost:** **~$5 / month**
* **Calculation:** The cost is based on the number of transactions (requests for secrets). With two applications frequently accessing secrets, we estimate a few million transactions per month, resulting in a small fee.
* **Pricing Link:** [Azure Key Vault Pricing](https://azure.microsoft.com/en-us/pricing/details/key-vault/)