Analysis of Alternatives (AoA)

**S&ID -- Automated Security‑Form Pre‑Population & Validation**

**By: Mitch Marsden**

Purpose & Scope

**The Personnel Security (PERSEC) team needs a FedRAMP‑compliant system that will**

1. Pre‑populate incoming security forms with authoritative data from the on‑prem SIMS database.
2. Validate every remaining field (e‑mail, phone, ZIP, empty signature boxes, etc.) using deterministic rules plus LLM reasoning.
3. Generate a concise correction report that cites the exact section needing fixes and can be viewed, copied, downloaded, or attached to a ticket.
4. Meet “no raw form retained > 24 h” policy and keep monthly OPEX (recurring cloud spend) as low as practical.

# Key Assumptions

|  |  |
| --- | --- |
| **Assumption** | **Rationale** |
| Volume : 1000 SF‑85P forms per month, 95 pages each (≈ 95000 pages) | Estimate |
| Processing window 09:00‑17:00 ET, Mon–Fri | Enables scheduled VM power‑on/off |
| SIMS access via Private Endpoint (read‑only) | No egress fees; stays in Azure Gov boundary |
| Power Automate GCC‑High licenses already procured | No extra license cost line |

# Baseline Unit Prices (Azure Gov)

|  |  |  |
| --- | --- | --- |
| **Service** | **Gov Price** | **Source** |
| Azure Document Intelligence – Pre‑built Document | $0.010 / page | <https://learn.microsoft.com/azure/ai-services/document-intelligence/overview#pricing> |
| Azure OpenAI GPT‑4o | $0.005 / 1 000 tokens | <https://learn.microsoft.com/azure/ai-services/openai/how-to/gov-pricing#token-pricing> |
| Logic Apps (consumption) | $0.000025 / action | <https://azure.microsoft.com/pricing/details/logic-apps/#pricing> |
| Azure Functions (consumption) | $0.000016 / GB‑s | <https://learn.microsoft.com/azure/azure-functions/functions-consumption-costs-government> |
| Azure AI Search Basic S1 | $0.14 / search‑unit‑h (≈ $108 / mo) | <https://learn.microsoft.com/azure/search/search-sku-tier> |
| VM Standard B1s | $0.010 / h | <https://learn.microsoft.com/azure/virtual-machines/pricing-burstable#government-pricing> |

Illustrative Monthly Cost — Alternative A (Azure-Native Baseline)

|  |  |  |  |
| --- | --- | --- | --- |
| **Line‑item** | **Quantity** | **Unit price** | **Cost** |
| Logic Apps actions | 1000 x 10 | $0.000025 | $0.25 |
| Azure Functions (PyPDF parse) | 1000 x 0.2s x 0.5 GB | $0.000016/ GB‑s | $0.002 |
| GPT‑4o tokens | 1000 forms × 1000 | $0.005 / 1000 | $5 |
| Blob + ADX storage | 2 GB | $0.0184 / GB | $0.04 |
| ADX ingest & queries | ≈ 100 rows | est. | $1.00 |
| Total |  |  | ≈ $6.3 / month |

***All other alternatives reuse these unit prices; swap in their own hosting & index costs.***

# Architectural Alternatives

|  |  |  |  |
| --- | --- | --- | --- |
| **Alt** | **Workflow (high level)** | **Monthly OPEX** | **Highlights / Caveats** |
| A – Azure-Native Stack | Logic Apps + Azure Functions (PyPDF) → Regex → GPT‑4o → Blob + ADX | $6.3 | Serverless, lowest cost, no VM |
| B – Alt A + Azure AI Search | Alt A + Azure AI Search (1 Search Unit) | $114 | Adds RBAC, full‑text & semantic search, RAG‑ready |
| C – Power Automate | Power Automate (pre-paid licences) → Azure Functions (PyPDF) → GPT‑4o → Blob storage | $6.3 | 100 % no‑code; flow run‑time limits |
| D – n8n | VM $1.76 + disk $2.88 → PyPDF → GPT‑4o → Blob storage | $9.6 | Open‑source flexibility; VM patching |
| E – Document Intelligence fallback | Any scanned/flattened pages → DI OCR ($.010 / page) | Usage-based - $950 / mo *only if every page is scanned* | Triggers only when PDF lacks AcroForm |

***Design assumes pdf extraction can be done with python PyPDF due to majority of forms being AcroForm (editable).***

# Acceptance‑Criteria Trace Matrix

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Requirement** | **A** | **B** | **C** | **D** | **Notes** |
| Detect corrections | ✓ | ✓ | ✓ | ✓ | Regex + GPT‑4o |
| Summarized report, cited sections | ✓ | ✓ | ✓ | ✓ | JSON/Markdown |
| User uploads form | ✓ | ✓ | ✓ | ✓ |  |
| View report | ✓ | ✓ | ✓ | ✓ |  |
| Download report | ✓ | ✓ | ✓ | ✓ |  |
| Copy text | ✓ | ✓ | ✓ | ✓ |  |
| Batch upload (nice-to-have) | ✓ | ✓ | Limited | ✓ | Flow timeout |
| No raw forms > 24 h | ✓ | ✓ | ✓ | ✓ |  |
| SIMS pre‑populate | ✓ | ✓ | ✓ SQL connector | ✓ |  |
| Validate phone/email/ZIP ect. | ✓ | ✓ | ✓ | ✓ |  |

# Cost & Accuracy Summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Alt** | **Monthly OPEX** | **OCR/Extraction Acc.** | **Validation Acc.** | **Requirements Met** |
| A | $6.3 | ≈ 100 %\* | 95 % | Yes |
| B | $114 | ≈ 100 %\* | 95 % | Yes + semantic |
| C | $6.3 | ≈ 100 %\* | 95 % | All but batch upload |
| D | $9.6 | ≈ 100 %\* | 95 % | Yes |
| E | Usage-based | 98% | 95% | Yes |

\*PyPDF reads only AcroForm fields; accuracy drops to 0% on scanned PDFs → covered by Alt E.

Accuracy Baselines & Sources

|  |  |  |
| --- | --- | --- |
| **Metric** | **Value** | **Source** |
| Document Intelligence OCR | 98 % field‑level F‑score | <https://learn.microsoft.com/azure/ai-services/document-intelligence/concept-accuracy> |
| Tesseract OCR | 88 – 92 % F‑score on 300 dpi scans | <https://arxiv.org/abs/2102.11811> |
| GPT‑4o recall on fuzzy errors | 97 % | <https://openai.com/research/gpt-4o> |
| Combined validation F‑score (DI + GPT‑4o) | ≈ 95 % | Derived Estimate: 98 % × 97 % |

# n8n Hosting Options

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Option** | **Ideal use** | **Cost (Gov)** | **Pros** | **Cons** |
| VM Std B1s | Daytime load ≤ 10 k forms/day | $1.76 | Cheapest, simple, Private Endpoint friendly | Patch OS; remember shutdown |
| Azure Container Apps | Autoscale / future micro‑services | $21 base | Managed OS; scale‑to‑zero | Higher base cost; cold start |
| Azure Container Instances | Occasional bursts | ≈ $1 | Zero idle cost | Cold‑start ~45 s |
| AKS | Org already uses Kubernetes | $60 | GitOps, Helm, ingress control | Highest ops overhead |

# Blob + ADX vs AI Search

|  |  |  |
| --- | --- | --- |
| **Capability** | **Blob Storage + ADX** | **Azure AI Search Basic S1** |
| Encryption / CMK | ✔ (SSE + CMK) | ✔ |
| Row‑level RBAC | ADX role assignments | Index‑level RBAC (coarse‑grained) |
| Keyword search | Exact Kusto queries | Full‑text analyzers |
| Semantic / vector search | ✘ | ✔ Neural & Vector (~92 % recall @ top‑5) |
| Built‑in purge API | Retention policy (soft‑delete) | ✔ Purge API deletes individual docs |
| Cost / TB hot data | ≈ $122 / TB‑mo | ≈ $432 / TB‑mo |
| Ready for RAG chatbot | Need extra vector store | Native vector fields plug directly into Azure OpenAI |

# Recommendation

1. Adopt Alternative A — Azure native Logic Apps + Functions + PyPDF extraction.
   * Serverless, FedRAMP, no VM ops, ≈ $6 / month for 1 000 fillable PDFs.
2. Enable Alternative E (Document Intelligence) as an automatic fallback for any scanned or flattened PDFs at $0.010 / page.
3. Add AI Search (Alt B) only if semantic search or RAG chatbot becomes a requirement.
4. Keep n8n (Alt D) as a sandbox option; production stays on the Azure native stack.
5. Power Automate (Alt C) remains viable for M365 only teams but watch flow run time limits.