Secure OCR Financial Document Extractor – Project Plan & Cost-Benefit Analysis

# 1. Project Summary

This project aims to build a secure, self-hosted OCR-based financial document extractor that allows financial professionals to upload documents such as Social Security statements, 401(k) reports, and brokerage statements. The system will extract key financial data and output it in a format compatible with Nitrogen or future internal tools.

# 2. Objectives & Scope

- Accept PDF/JPG/PNG uploads with secure handling  
- Let users select specific pages to process  
- Extract structured data using Tesseract OCR and regex parsing  
- Output results in Excel and JSON formats  
- Ensure strong compliance (encryption, access control, audit trail)

# 3. Development Phases & Costs

|  |  |  |  |
| --- | --- | --- | --- |
| Phase | Description | Hours | Cost ($80/hr) |
| Discovery & Planning | Finalize doc types, fields, and compliance scope | 4 | 320 |
| OCR Engine Setup | Dockerized app with Tesseract, PDF/image pipeline | 6 | 480 |
| Upload UI + Page Preview | Upload form + pdf.js viewer with page selector | 8 | 640 |
| Extraction Logic | Regex/NLP parsing for 2–3 doc types | 10 | 800 |
| Export Formatter | Excel and/or JSON file generation | 4 | 320 |
| Security Layer | Encryption, access control, audit logs | 10 | 800 |
| AWS Deployment | EC2 setup, SSL, S3, IAM config | 6 | 480 |
| QA + Testing | Bug fixes, edge cases, document testing | 6 | 480 |
| Documentation & Handoff | User/admin docs, GitHub repo | 4 | 320 |

Total Estimated Hours: 58

Total Development Cost: $4,640

# 4. Estimated Monthly AWS Hosting Cost

- EC2 t3.small: ~$17/month (on-demand) or ~$9/month (1-year reserved)  
- S3 (2GB + 5,000 requests): ~$0.05/month  
- Bandwidth (5 GB out): ~$0.45/month  
- Optional Route 53 domain: $0.50/month  
- Total Estimated: ~$18–20/month (or ~$10/month with reserved instance)

# 5. Security & Compliance Measures

- AES-256 file encryption at rest  
- Auto-deletion of temporary files post-processing  
- Virus scanning (ClamAV or equivalent)  
- Role-based access control (RBAC)  
- HTTPS via AWS ACM or Let's Encrypt  
- Upload validation and logging

# 6. Cost-Benefit: Textract vs. Self-Hosted Tesseract

| Feature | Amazon Textract | Self-Hosted Tesseract |  
|----------------------------|--------------------------------|------------------------------|  
| Cost per Page | $0.10 (Forms + Tables) | $0 (free) |  
| Monthly at 2,000 Pages | ~$200 | ~$10 (server only) |  
| Accuracy | Very high | High (with clean documents) |  
| Setup Time | Low | Moderate (regex needed) |  
| Maintenance | None | Light (parser updates) |  
| Data Control | Cloud-based (limited) | Full control |  
| Security Compliance | AWS Managed | You control configuration |  
| Breakeven Point | ~46,000 pages/year | After that, Tesseract wins |

Recommendation: For your use case—structured documents (SSA, 401(k)), focus on value fields, need for low cost and full control—the self-hosted Tesseract solution offers the best return. It provides strong accuracy at minimal cost and can be enhanced incrementally as your needs grow.