SMEVENTURES AI ENGINEER TECHNICAL ASSESSMENT

Take-Home Test

Company Contact Information Extraction Service

Document Version: 1.0
Assessment Duration: Take-Home (Submit within 96 hours)
Platform: n8n Workflow Automation

Executive Summary

This technical assessment evaluates your ability to build a production-ready automation system using n8n, Google Vertex AI (Gemini), PostgreSQL, and HubSpot CRM. You will create a service that extracts business contact information (phone numbers, email addresses, and physical addresses) from enriched company data using AI, validates the results, and updates CRM records.

This assessment simulates real-world engineering work. You are expected to deliver a fully functional system with proper error handling, documentation, and a demonstration of it working.

Attribute	Details
Duration	96 hours from receipt
Deliverables	Working n8n workflow (.json), Metrics dashboard (Vercel), Documentation (README), Screen recording demo
Tech Stack	n8n, HubSpot (free sandbox), Google Vertex AI (Gemini), PostgreSQL
Evaluation	Functionality, Error Handling, Documentation, Code Quality

1. System Requirements

1.1 Purpose

Build an automated service that extracts and validates business contact information (phone, email, physical address) from enriched company data using AI, then safely updates HubSpot CRM records.

1.2 Required Architecture

Component	Technology
Orchestration	n8n (self-hosted or cloud)
AI Provider	Google Vertex AI (Gemini 2.0 Flash or 2.5 Pro)
CRM	HubSpot (free sandbox account)
Database	PostgreSQL (any provider: Supabase, Neon, Railway, local)

1.3 Data Flow

Your system must implement this exact data flow:

```
Schedule Trigger → HubSpot Search → Deduplication Check → AI Extraction → Data Validation → PostgreSQL Storage → Batch Update Preparation → Safety Validation → HubSpot Update → Audit Logging
```

2. HubSpot Configuration

You must create a HubSpot free sandbox account and configure the following custom properties on the Company object. This simulates a real production environment.

2.1 Required Custom Properties

Create these custom properties in HubSpot Settings > Properties > Company:

Property Name	Type	Description
enrichment_trigger	Date	Timestamp when enrichment was triggered
scraped_website_content	Multi-line text	Scraped website text content
scraped_contact_page	Multi-line text	Scraped contact page content
scraped_about_page	Multi-line text	Scraped about page content
extracted_phone	Single-line text	AI-extracted phone number
extracted_email	Single-line text	AI-extracted email address
extracted_address	Multi-line text	AI-extracted street address
extraction_model	Single-line text	Model version used (e.g., gemini-2.0-flash)
extraction_timestamp	Date	When extraction was performed
extraction_confidence	Number	Confidence score 0-1
fields_found_count	Number	Count of non-empty fields (0-3)
processing_status	Single-line text	Status: pending, complete, error, partial
error_message	Multi-line text	Error details if processing failed
manual_override	Single checkbox	If true, skip automatic processing
clearbit_enriched	Single checkbox	If true, already enriched by another service
clearbit_confidence	Number	Confidence from external enrichment

2.2 Trigger Conditions

Your system must process companies ONLY when ALL of these conditions are met:

1. enrichment_trigger timestamp exists and is within 6 months
2. ALL clearbit fields are empty: clearbit_enriched = false, clearbit_confidence is null
3. ALL extracted fields are empty: extracted_phone, extracted_email, extracted_address
4. No manual override (manual_override ≠ true)
5. Has enrichment data (at least one of: scraped_website_content, scraped_contact_page, scraped_about_page)
6. Not processed within last 30 days (deduplication window)
7. No errors within last 24 hours (error cooldown)
8. Less than 5 total error attempts

3. Test Companies & Expected Results

You must create these 5 test companies in your HubSpot sandbox and populate them with the provided mock enrichment data. Your system will be evaluated on whether it produces results matching the expected output format.

3.1 Required Output Format

Your AI extraction must produce output in EXACTLY this format:

```
Phone: <phone_number_with_country_code_no_formatting>Address:
<full_street_address_google_format>Email: <generic_company_email>
```

If a field cannot be found, return "none" (lowercase). You must design your own prompt to achieve this output consistently.

3.2 Test Company 1: Patagonia

Field	Value
Company Name	Patagonia
Domain	patagonia.com
Country	United States
City	Ventura
State	California

Mock Enrichment Data (scraped_contact_page):

Contact Us - Patagonia. Customer Service: For questions about orders, returns, or products, call us at 1-800-638-6464 or email customerservice@patagonia.com. Hours: Monday-Friday 6am-6pm PT, Saturday-Sunday 6am-6pm PT. Corporate Headquarters: Patagonia, Inc., 259 W Santa Clara St, Ventura, CA 93001. For press inquiries: press@patagonia.com. For wholesale inquiries: wholesale@patagonia.com

Expected Output:

```
Phone: +18006386464Address: 259 W Santa Clara St, Ventura, CA 93001, USAEmail: customerservice@patagonia.com
```

3.3 Test Company 2: Atlassian

Field	Value
Company Name	Atlassian
Domain	atlassian.com
Country	Australia
City	Sydney
State	New South Wales

Mock Enrichment Data (scraped_about_page):

About Atlassian - We believe teamwork is possible anywhere. Atlassian builds software that helps teams collaborate better. Founded in 2002 in Sydney, Australia. Global Headquarters: Level 6, 341 George Street, Sydney NSW 2000, Australia. US Office: 350 Bush Street, Floor 13, San Francisco, CA 94104. For general inquiries contact info@atlassian.com. Investor Relations: investors@atlassian.com

Expected Output:

```
Phone: noneAddress: Level 6, 341 George Street, Sydney NSW 2000, AustraliaEmail: info@atlassian.com
```

3.4 Test Company 3: Basecamp

Field	Value
Company Name	Basecamp
Domain	basecamp.com
Country	United States
City	Chicago
State	Illinois

Mock Enrichment Data (scraped_website_content):

Basecamp - Project management software that helps teams stay organized. We have been a fully remote company since 1999. Headquarters address: 30 N Racine Ave #200, Chicago, IL 60607. Phone: (312) 555-0123. For support questions, visit our help center at help.basecamp.com or email support@basecamp.com. Press contact: press@basecamp.com

Expected Output:

```
Phone: +13125550123Address: 30 N Racine Ave #200, Chicago, IL 60607, USAEmail: support@basecamp.com
```

3.5 Test Company 4: TechStartup XYZ (Edge Case - Minimal Data)

Field	Value
Company Name	TechStartup XYZ
Domain	techstartupxyz.io
Country	United States
City	Austin
State	Texas

Mock Enrichment Data (scraped_website_content):

TechStartup XYZ - We are building the future of AI. Join our waitlist to be the first to know when we launch. Follow us on Twitter @techstartupxyz. Backed by Y Combinator.

Expected Output:

```
Phone: noneAddress: noneEmail: none
```

Note: This edge case tests your system's ability to gracefully handle companies with no extractable contact information.

3.6 Test Company 5: Global Imports Ltd (Edge Case - Invalid Data)

Field	Value
Company Name	Global Imports Ltd
Domain	globalimportsltd.co.uk
Country	United Kingdom
City	London

Mock Enrichment Data (scraped_contact_page):

Contact Global Imports Ltd. Mailing Address: PO Box 12345, London EC1A 1BB. For test orders use test@example.com. Phone: 0000000000. Our warehouse is located at 45 Industrial Way, Manchester M1 2AB, UK. General enquiries: enquiries@globalimportsltd.co.uk. Phone support: +44 20 7946 0958

Expected Output:

```
Phone: +442079460958Address: 45 Industrial Way, Manchester M1 2AB, UKEmail: enquiries@globalimportsltd.co.uk
```

Note: This edge case tests your validation logic. Your system must reject: PO Box addresses (extract warehouse address instead), test/example emails, and invalid phone numbers (all zeros).

4. Data Validation Rules

Your system must implement these validation rules. Invalid data should be rejected and replaced with "none".

4.1 Phone Number Validation

- **Length:** 10-20 digits (including country code)
- **Format:** Must start with + or digit, no spaces/parentheses/hyphens in output
- **Reject Patterns:**
  All zeros: ^0{5,}
  All ones: ^1{5,}
  Sequential: ^123456
  Repeated: ^555555
- **Valid Country Codes:** +1 (US/CA), +44 (UK), +61 (AU), +64 (NZ), +49 (DE), +33 (FR), +81 (JP), +86 (CN), +91 (IN), +65 (SG)

4.2 Email Validation

- **Valid Format:** ^[a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]+\.[a-zA-Z]{2,}$
- **Length:** 5-254 characters
- **Must contain:** Exactly one @ symbol
- **Reject Patterns:**
  test@, example@, admin@localhost
  @test., @example., @localhost
  noreply@, no-reply@, donotreply@

4.3 Address Validation

- **Length:** 10-500 characters
- **Reject:** PO Box patterns (PO Box, P.O. Box, Post Office Box, GPO Box)
- **Reject Test Data:** ^test, ^example, ^none$, ^n/a$
- **Should contain:** Street number and street type indicator (St, Ave, Rd, Way, etc.)

5. PostgreSQL Database Schema

You must create these tables in your PostgreSQL database. These tables support deduplication, audit logging, and error tracking.

5.1 Main Results Table

```
CREATE TABLE contact_extraction_results (    id SERIAL PRIMARY KEY,    crm_record_id
VARCHAR(50) NOT NULL,    company_name VARCHAR(255),    domain VARCHAR(255),        --
Extracted data    extracted_phone VARCHAR(50),    extracted_address TEXT,
extracted_email VARCHAR(255),        -- Metadata    extraction_timestamp TIMESTAMP
WITH TIME ZONE NOT NULL,    processing_run_id UUID,    model_version VARCHAR(50),
extraction_time_seconds DECIMAL(10,2),        -- Status    status VARCHAR(20),
error_message TEXT,    crm_updated_at TIMESTAMP WITH TIME ZONE,        -- Context
data_sources_used JSONB,    raw_ai_response TEXT,    confidence_score DECIMAL(3,2),
fields_found INTEGER,        created_at TIMESTAMP WITH TIME ZONE DEFAULT
CURRENT_TIMESTAMP);CREATE INDEX idx_crm_record_id ON
contact_extraction_results(crm_record_id);CREATE INDEX idx_processing_run ON
contact_extraction_results(processing_run_id);CREATE INDEX idx_extraction_date ON
contact_extraction_results(extraction_timestamp);
```

5.2 Deduplication Tracking Table

```
CREATE TABLE processing_status (    crm_record_id VARCHAR(50) PRIMARY KEY,
last_trigger_timestamp TIMESTAMP WITH TIME ZONE,    last_processed_timestamp TIMESTAMP
WITH TIME ZONE,    last_error_timestamp TIMESTAMP WITH TIME ZONE,    processing_count
INTEGER DEFAULT 0,    error_count INTEGER DEFAULT 0,    last_status VARCHAR(20),
last_run_id UUID,    updated_at TIMESTAMP WITH TIME ZONE DEFAULT CURRENT_TIMESTAMP);
```

5.3 Error Logging Table

```
CREATE TABLE processing_errors (    id SERIAL PRIMARY KEY,    crm_record_id
VARCHAR(50) NOT NULL,    error_timestamp TIMESTAMP WITH TIME ZONE NOT NULL,
error_type VARCHAR(50),    error_message TEXT,    error_details JSONB,
processing_run_id UUID,    retry_count INTEGER DEFAULT 0);CREATE INDEX
idx_error_crm_id ON processing_errors(crm_record_id);CREATE INDEX idx_error_time ON
processing_errors(error_timestamp);
```

6. Metrics Dashboard

Build a custom metrics dashboard and deploy it to Vercel. The dashboard must connect to your PostgreSQL database and display the required metrics in real-time.

6.1 Technical Requirements

- **Framework:** Your choice (Next.js, React, Vue, etc.)
- **Hosting:** Vercel (free tier is fine)
- **Database Connection:** Must query PostgreSQL directly (use connection pooling)
- **Styling:** Your choice - clean and professional

6.2 Required Metrics

Your dashboard must display the following metrics:

Volume Metrics

- Total companies processed (all time, last 7 days, last 30 days)
- Success rate (completed vs. total)
- Companies with data found vs. no data
- Processing throughput (companies per hour/day)

Quality Metrics

- Field extraction rates: % with phone, % with email, % with address
- Average confidence score
- Distribution of fields_found (0, 1, 2, 3)

Error Metrics

- Error count by type
- Average retry count
- Recent errors list (last 10)

6.3 Dashboard Deliverables

- **Live URL:** Working Vercel deployment link
- **Source Code:** Include in your GitHub repository
- **Auto-refresh:** Dashboard should update automatically (every 30-60 seconds)

7. API Rate Limits & Configuration

7.1 HubSpot API Limits

Limit Type	Value
Rate Limit	100 requests per 10 seconds
Batch Update	Max 100 companies per request
Search API	Max 100 results per request (use pagination)
Daily Limit (Free)	250,000 API calls

7.2 Vertex AI Configuration

Configure your Vertex AI integration with the following parameters:

- **Model:** gemini-2.0-flash OR gemini-2.5-pro (recommended)
- **Temperature:** 0.5

All other configuration parameters (location, max_output_tokens, top_p, top_k, safety settings) are up to you. Document your choices in your README.

7.3 Retry Logic Requirements

- **Max Retries:** 3 attempts
- **Backoff:** Exponential (2^attempt seconds)
- **Timeout:** 30 seconds per request

8. Deliverables

You must submit all of the following:

8.1 Working n8n Workflow

- Export your complete workflow as a .json file
- Workflow must be importable and runnable (with credential updates)
- Include all nodes: triggers, API calls, data transformations, error handling

8.2 Documentation (README.md)

Your README must include:

9. **Setup Instructions:** Step-by-step guide to configure HubSpot, Vertex AI, and PostgreSQL
10. **Architecture Overview:** Diagram or description of your workflow structure
11. **AI Prompt:** The exact prompt you designed for Vertex AI extraction
12. **Error Handling Strategy:** How your workflow handles failures at each stage
13. **Design Decisions:** Explain key technical choices and trade-offs
14. **Known Limitations:** What would you improve with more time?

8.3 Screen Recording Demo

- **Duration:** 3-5 minutes
- **Format:** MP4, MOV, or Loom link
- **Must demonstrate:**
    1. Workflow execution from trigger to completion
    2. Processing at least 2 test companies
    3. HubSpot records updated with extracted data
    4. PostgreSQL records showing audit trail
    5. Error handling in action (trigger an error deliberately)

9. Submission Instructions

9.1 What to Submit

15. **n8n Workflow Export:** contact-extraction-workflow.json
16. **Documentation:** README.md
17. **Database Schema:** schema.sql (your CREATE TABLE statements)
18. **Dashboard:** Live Vercel URL + source code in repository
19. **Demo Recording:** Link to Loom/YouTube or attached MP4

9.2 How to Submit

- Create a GitHub repository (can be private, invite us as collaborators)
- OR submit as a ZIP file via email
- Include all deliverables in a single submission

9.3 Timeline

- **Deadline:** 96 hours from receiving this document
- **Questions:** You may ask clarifying questions via email (response within 24 hours)
- **Extensions:** Request at least 24 hours before deadline if needed

Important Notes

- **Use your own accounts:** Set up HubSpot sandbox, Google Cloud, and PostgreSQL under your own accounts. Do not use company credentials.
- **Cost awareness:** Vertex AI has free tier limits. Design your workflow to be cost-efficient.
- **No AI assistance disclosure required:** You may use AI tools to help build this. We care about the final result and your ability to explain it.
- **Focus on production quality:** This should be code you would deploy to production, not a quick prototype.

Good luck! We look forward to reviewing your submission.

— End of Document —