Lead Ingestion & Outreach Platform — Architecture Specification

1. Overview

This system is a modular, agent-based lead ingestion, validation, routing, export, and retry pipeline designed for high-volume business lead acquisition and downstream sales/outreach workflows.

Key design goals: - Deterministic, testable pipeline contracts - Idempotent exports with per-sheet deduplication - Fail-fast validation and atomic multi-sheet export - Retry-safe ingestion for transient failures - Horizontal extensibility for enrichment, scoring, and outreach

The platform is organized into discrete processing phases implemented as composable agents.

---

2. High-Level Architecture

The platform operates in two execution modes:

Normal Mode (Primary Ingestion)

MapsSearchAgent → BusinessNormalizeAgent → WebsitePresenceValidator → LeadRouterAgent → LeadFormatterAgent → GoogleSheetsExportAgent

Retry Mode (Failure Recovery)

RetryInputLoaderAgent → WebsitePresenceValidator → LeadRouterAgent → LeadFormatterAgent → GoogleSheetsExportAgent

Both modes share the same downstream contract after ingestion.

---

3. Core Architectural Principles

3.1 Pipeline Contract Enforcement

Each agent produces and consumes explicitly named keys:

Stage	Output Contract Key
MapsSearchAgent	raw_places
BusinessNormalizeAgent	normalized_businesses

Stage	Output Contract Key
WebsitePresenceValidator	validated_businesses
LeadRouterAgent	routed_leads
LeadFormatterAgent	formatted_leads
GoogleSheetsExportAgent	export_stats

Contracts are protected by fail-fast runtime validation and unit tests.

3.2 Deduplication Ownership Model

Deduplication is centralized:

Owner: BusinessNormalizeAgent

Rules: - dedup_key is generated once - downstream agents must pass-through unchanged - exporters consume pre-computed dedup_key - no recomputation allowed

This guarantees deterministic idempotency across pipeline runs.

3.3 Routing Architecture

Leads are classified post-validation into routing buckets:

Route	Description
TARGET	No website, high outreach priority
EXCLUDED	Has website, deprioritized
RETRY	Website validation error

LeadRouterAgent outputs: - routed_leads (flat list) - routing_stats

Flattened ordering is deterministic: TARGET → EXCLUDED → RETRY

3.4 Export Fan-Out Architecture

Exports are atomic, multi-sheet, batch-safe operations.

Three-phase execution:

Phase 1 — Preflight (No Writes)

- Validate spreadsheet
- Create or validate worksheets
- Load per-sheet dedup sets

Phase 2 — Write (Sequential)

- Deterministic sheet order
- Batch append (MAX_BATCH_SIZE = 200)
- Per-sheet idempotency
- Abort remaining sheets on any failure

Phase 3 — Backup Commit

- Write JSON snapshot
- Write CSV snapshot
- Only executed if ALL writes succeed

3.5 Retry Pipeline Design

Retry pipeline supports controlled reprocessing of failed leads.

RetryInputLoaderAgent:

Inputs: - spreadsheet_id - retry_sheet_name (default: WEBSITE_CHECK_ERRORS)

Behavior: - Parses retry_attempt field - Enforces PIPELINE_MAX_RETRIES (env configurable) - Skips maxed leads - Increments retry counter - Preserves ordering

Outputs: - validated_businesses - retry_stats

---

4. Agent Responsibilities

MapsSearchAgent

- Queries external map provider
- Returns raw business listings

BusinessNormalizeAgent

- Canonicalizes business fields
- Computes dedup_key
- Produces normalized_businesses

WebsitePresenceValidator

- Validates website existence
- Applies blacklist rules
- Handles DNS, HTTP, timeout errors
- Produces validated_businesses

LeadRouterAgent

- Applies routing rules
- Assigns lead_route
- Assigns target_sheet
- Preserves dedup_key
- Produces routed_leads

LeadFormatterAgent

- Transforms to export schema
- Pass-through routing and dedup fields
- Fail-fast on contract violations
- Produces formatted_leads

GoogleSheetsExportAgent

- Fan-out export by target_sheet
- Enforces per-sheet dedup
- Performs atomic export
- Generates export_stats

RetryInputLoaderAgent

- Reads failed leads from Sheets
- Applies retry rules
- Produces validated_businesses

---

5. Testing Strategy

Test layers:

Contract Tests

- Field preservation
- Ordering guarantees
- Fail-fast enforcement

Export Fan-Out Tests

- Batch chunking
- Atomic abort behavior
- Per-sheet dedup isolation
- Export stats structure

Retry Logic Tests

- MAX_RETRIES enforcement
- Parsing edge cases
- Ordering preservation

Integration Tests

- Full pipeline dry-run
- Mocked external services

Current coverage: 122 tests passing

---

6. Configuration Model

Environment Variables:

Variable	Purpose	Default
PIPELINE_MODE	normal / retry	normal
PIPELINE_MAX_RETRIES	Retry cap	3
MOCK_SHEETS	Disable live export	false

---

7. Extensibility Points

Planned integrations:

- Landing Page Generator Agent
- Lead Scoring Agent
- CRM Sync Agent
- Email Outreach Agent

- Campaign Analytics Agent

Each new module plugs into the pipeline via contract-based chaining.

---

8. Non-Functional Guarantees

- Deterministic output ordering
- Idempotent exports
- Zero partial writes
- Fail-fast validation
- Test-first contract enforcement
