“# SSIS Blueprint #1 v1.4 — Implementation Checklist

Status: DERIVED SUMMARY (non-authoritative)

Canonical Source: SSIS Blueprint #1 v1.4 (uploaded)

Last updated: 2026-01-07

Purpose: Implementation checklist + constraint recall

**Source of Truth**: Blueprint #1 v1.4 | **Implementation Detail**: Research Pack v1.0

---

## HARD CONSTRAINTS (Non-Negotiable)

| ID | Constraint | Blueprint Ref |
|----|------------|---------------|
| C1 | **Offline-first, CPU-only** — no cloud dependencies; SA power resilience | §0 |
| C2 | **Idempotency at every stage** — safe to rerun after crash/power loss | §0 |
| C3 | **Atomic publish** — write temp → flush (best-effort fsync) → rename to final | §4 |
| C4 | **Stage locks** — key `(asset_id, stage, feature_spec_alias|null)`; TTL ~10min; reclaim stale | §7 |
| C5 | **Versioned contracts** — every artifact: `schema_id` + `version` + `asset_id` + `computed_at` | §6 |
| C6 | **FeatureSpec alias immutability** — alias→spec frozen at first registration; collision = `FEATURE_SPEC_ALIAS_COLLISION` | §5 |
| C7 | **HDF5 single-writer rule** — only `worker_features` writes `.h5`; all others write JSON | §4 |
| C8 | **Never overwrite artifacts** — new config → new FeatureSpec → new artifact; keep prior intact | §1, §12 |

---

## LOCKED TECHNICAL DECISIONS

| Component | Decision | Detail |
|-----------|----------|--------|
| Orchestrator | Huey + SQLite | Durable queue state offline; 1 worker default |
| Audio format | 22050 Hz mono PCM WAV | Canonical derivative |
| Embeddings | YAMNet via ONNX Runtime CPU | 3.7M params, 1024-D, MIT license |
| Segmentation | inaSpeechSegmenter | speech/music/noise; silence derived from gaps |
| Feature storage | HDF5 + gzip | ~300KB/min; random access; `.h5.tmp` atomic |
| Preview | Heuristic v1 | Boundaries + energy variance + embedding variance; fallback=intro |

| Concurrency | 1 worker (baseline) | SQLite queue best with single consumer |

---

## STAGE REQUIREMENTS

### Stage A — Ingest (FastAPI) `services/ingest_api/`

| Requirement | Spec |
|-------------|------|
| Input | Local path OR file upload |
| Copy to | `data/audio/{asset_id}/original.<ext>` |
| Hash | SHA256 of file content |
| Metadata | Duration, channels, sample_rate (best-effort) |
| Idempotency | Unique constraint on `(owner_entity_id, content_hash)` |
| Output | `AudioAsset` row, `PipelineJob(stage="ingest")`, enqueue orchestrator |
| **Metrics** | file_size, hash_time_ms, format_guess |
| **Errors** | `INGEST_FAILED`, `FILE_NOT_FOUND`, `HASH_FAILED` |

### Stage B — Decode (worker_decode) `services/worker_decode/`

| Requirement | Spec |
|-------------|------|
| Input | `AudioAsset.source_uri` |
| Output path | `data/audio/{asset_id}/normalized.wav` |
| Format | 22050 Hz, mono, 16-bit PCM WAV |
| Chunking | Process in 30-60s chunks |
| Checkpoint | Every ~60s of processed audio |
| Publish | Atomic: `.wav.tmp` → rename |
| **Metrics** | output_duration_sec, chunk_count, resample_time_ms |
| **Errors** | `CODEC_UNSUPPORTED`, `FILE_CORRUPT`, `FILE_TOO_SHORT` (<1.7s) |

### Stage C — Features (worker_features) `services/worker_features/`

| Requirement | Spec |
|-------------|------|
| Input | `normalized.wav` |
| FeatureSpec ID | `mel64_h10ms_w25ms_sr22050__yamnet1024_h0.5s_onnx` (v1.4 default) |

| Alias | First 12 chars of `sha256(feature_spec_id)` |
| **Log-mel** | 64 mel bands, 25ms window (n_fft=1024), 10ms hop (hop_length=220), sr=22050 |
| **Embeddings** | YAMNet 1024-D, 0.5s hop, ONNX Runtime CPU |
| Output path | `data/features/{asset_id}.{alias}.h5` |
| HDF5 datasets | `/mel` (float32), `/embeddings` (float32), attrs: sample_rate, mel_hop, embedding_hop_sec, version |
| Publish | Atomic: `.h5.tmp` → rename (never write directly to final) |
| Validation | No NaN/Inf; shape checks |
| Upsert rule | If alias exists + spec matches → no-op; if alias exists + spec differs → hard error |
| **Metrics** | inference_time_ms, mel_shape, embedding_shape, nan_inf_count, spec_alias, spec_id |
| **Errors** | `FEATURE_NAN`, `MODEL_OOM`, `FEATURE_EXTRACTION_FAILED`, `FEATURE_SPEC_ALIAS_COLLISION` |

### Stage D — Segments (worker_segments) `services/worker_segments/`

| Requirement | Spec |
|-------------|------|
| Input | `normalized.wav` |
| Model | inaSpeechSegmenter (`vad_engine='smn'`, `detect_gender=False`) |
| Labels | `speech`, `music`, `noise` |
| Silence | Derived from gaps (>0.5s low-energy) + optional Silero VAD |
| Post-process | Min duration filter (speech≥0.8s, music≥3.4s, silence≥0.5s); merge adjacent same-class |
| Confidence | Heuristic, not calibrated; include `confidence_type: "heuristic_v1"` |
| Output path | `data/segments/{asset_id}.segments.v1.json` |
| Publish | Atomic |
| **Metrics** | segment_count, class_distribution, flip_rate |
| **Errors** | `SEGMENTATION_FAILED` |

### Stage E — Preview (worker_preview) `services/worker_preview/`

| Requirement | Spec |
|-------------|------|
| Inputs | `normalized.wav`, segments JSON, embeddings from HDF5 |
| **FeatureSpec selection** | 1) `SSIS_ACTIVE_FEATURE_SPEC_ALIAS` env if set + exists → 2) pipeline default → 3) fail `FEATUREPACK_MISSING` |

| | |
|---|---|
| Boundaries | Pause boundaries (>200ms low-energy), segment boundaries in speech-heavy regions |
| Candidates | 60s windows at boundaries |
| Scoring | `0.6 * energy_variance + 0.4 * embedding_variance` (tune empirically) |
| Selection | Best above threshold; else **fallback = intro** (first non-silent window) |
| Output path | `data/preview/{asset_id}.preview.v1.json` |
| Fields | `mode` (smart/intro/fallback), `start_sec`, `end_sec`, `duration_sec`, `confidence`, `fallback_used`, `reason` |
| **Metrics** | candidate_count, best_score, fallback_used, spec_alias_used |
| **Errors** | `PREVIEW_LOW_CONF` (not fatal → triggers fallback), `FEATUREPACK_MISSING` |

---

## ORCHESTRATOR (Huey + SQLite)

| Responsibility | Spec |
|----------------|------|
| Stage planning | Based on artifact existence + `artifact_index` |
| Lock acquisition | Before stage dispatch; log in `pipeline_jobs.metrics_json` |
| Stale lock reclaim | If lock age > TTL (~10min) |
| Retry policy | **3 attempts**, delays: **60s, 300s, 900s** |
| Dead-letter | After 3 failures; log with error taxonomy |
| Idempotency | If final artifact exists → skip; temp artifacts → ignore/clean |

---

## DATABASE TABLES

1. `audio_assets` — canonical identity + source metadata

2. `pipeline_jobs` — per-stage logs, metrics, error codes, attempts

3. `stage_locks` — `(asset_id, stage, feature_spec_alias|null)`, acquired_at, worker_id

4. `feature_specs` — alias (PK), feature_spec_id, created_at, notes

5. `artifact_index` — tracks which artifacts exist per asset

---

## OBSERVABILITY (Must Ship in v1.4)

| Level | Metrics |
|-------|---------|

| | |
|---|---|
| **ingest** | file_size, hash_time, format_guess |
| **decode** | output_duration, chunk_count, resample_time |
| **features** | inference_time, mel_shape, embedding_shape, nan_inf_count, spec_alias/id |
| **segments** | segment_count, class_distribution, flip_rate |
| **preview** | candidate_count, best_score, fallback_used, spec_alias_used |
| **pipeline** | end_to_end_latency (p50/p95/p99), success_rate (>95% target), backlog_depth, error_breakdown |

---

## ERROR TAXONOMY (Locked)

| Code | Stage | Action |
|------|-------|--------|
| `CODEC_UNSUPPORTED` | decode | Skip, log for review |
| `FILE_CORRUPT` | decode | Skip, alert if frequent |
| `FILE_TOO_SHORT` | decode | Skip (<1.7s inaSpeechSegmenter limit) |
| `FEATURE_NAN` | features | Retry with different params |
| `MODEL_OOM` | features | Reduce chunk size, retry |
| `SEGMENTATION_FAILED` | segments | Use heuristic fallback |
| `PREVIEW_LOW_CONF` | preview | Use intro fallback (not fatal) |
| `FEATUREPACK_MISSING` | preview | Re-run features stage |
| `FEATURE_SPEC_ALIAS_COLLISION` | features | Hard error, log job failure |

---

## RESILIENCE TESTS (Required)

- [ ] **Contract tests**: Validate API + artifact JSON against `/specs/*.schema.json`

- [ ] **E2E smoke**: ingest → orchestrator → decode → features → segments → preview

- [ ] **Kill mid-write tests**:

  - Kill during decode (WAV write)

  - Kill during HDF5 `.tmp` write

- [ ] **Restart assertions**:

  - No corrupt final artifacts

  - Temp files ignored/cleaned

  - Stale locks reclaimed
  - Pipeline completes successfully
  - Idempotency prevents duplication

---

## CANONICAL PATHS

```
data/
  audio/{asset_id}/
    original.<ext>
    normalized.wav
  features/{asset_id}.{feature_spec_alias}.h5
  segments/{asset_id}.segments.v1.json
  preview/{asset_id}.preview.v1.json
logs/
```

---

## BUILD ORDER (§14)

1. Contracts + DB primitives (schemas, models, atomic_io, hashing, paths)
2. Ingest API (local + upload, idempotency, enqueue)
3. Orchestrator (locks, retries, dead-letter)
4. Decode worker (canonical WAV, chunking, checkpoint, atomic)
5. Features worker (mel + ONNX embeddings, HDF5 atomic, FeatureSpec immutability)
6. Segments worker (inaSpeechSegmenter, silence derivation)
7. Preview worker (FeatureSpec selection, scoring, fallback)
8. Resilience harness (kill/restart tests)
9. MVP acceptance (offline CPU, deterministic, safe restart, telemetry)