HiLabs™

## PROBLEM

**As a PM at HiLabs, I am tasked to solve the challenge of fragmented and inconsistent provider data across multiple federal and state systems. Manual reconciliation today is slow, error-prone, and unable to keep up with constant provider updates—causing compliance gaps, operational inefficiencies, and poor member experiences. My goal is to build a unified platform that aggregates and standardizes provider data from federal and state sources and continuously maintain accurate provider information for health plans and payer organizations.**

Adi Amruta(IIT Kanpur)
adiam22@iitk.ac.in
220058
Contact No.-9861106538

## CURRENT PROVIDER DATA MANAGEMENT & VALIDATION PROCESS

**Data Acquisition from Multiple Sources**
*Ingest provider datasets via APIs*

→

**Manual Import & Entry into Internal Systems**
*Populate internal CRMs using CSV uploads*

↓

**Data Standardization & Deduplication**
*Normalize schema and run de-dup routines*

←

**Entity Matching & Reconciliation Across Systems**
*Apply rule-based NPI entity matching*

↓

**Exception Handling, Provider Feedback & Corrections**
*Trigger manual review for record conflicts*

→

**Submission for Compliance + Published Directory + Audits**
*Generate CMS-ready directory for audit*

## ISSUES FACED

**$2.1 Billion**
Industry cost due to inaccurate provider data

**45 %**
CMS 2022 audit showed 48 % directory errors

**30–40 %**
Data teams spend nearly half their time reconciling duplicate records.

**4–6 Weeks**
Provider changes (address, credential) not reflected in real time.

**20–25 %**
Members unable to find valid or active listed providers

## POTENTIAL IMPACT

**$350**
Reduced manual reconciliation + fewer compliance fines.

**95 %**
Real-time updates through API-based ingestion and validation.

**25–30 %**
Accurate provider listings improve member access satisfaction.

**2–3×**
API-driven updates produce audit-ready data in days vs. weeks

**10–15 %**
Verified provider data improves CMS network scoring.

## THE PROBLEMS IN THE CURRENT PROCESS

- Provider data remains fragmented across multiple source systems with no unified schema.
- High manual effort in import, matching, deduplication and reconciliation.
- Duplicates and inconsistent records (e.g., providers listed at old locations or with outdated specialties)
- Slow update cycle – provider changes (address, credential) are not reflected in real-time.
- Lack of unified audit trail and clear data provenance, making compliance difficult.
- Compliance / directory accuracy risk (increasing under CMS/No Surprises rules).

## ASSUMPTIONS

### Data & System Assumptions

- Provider data is fragmented across NPPES, PECOS, AMA, and state databases with no unified schema.
- Updates to provider details (address, credentials) are delayed – not reflected in real time.
- Data formats vary (CSV, XML, API, PDF) requiring normalization and mapping.
- Duplicate and inconsistent provider records exist across payers and directories.

### Process & Operations Assumptions

- Provider data reconciliation and validation remain largely manual and spreadsheet-driven.
- Exception handling and provider feedback occur via ad-hoc email or ticket workflows.
- Limited automation in ingestion or real-time validation across systems.
- Directory accuracy checks are reactive– performed mainly during CMS audits.

### Compliance & Security Assumptions

- The platform must operate within HIPAA-compliant, non-PHI data boundaries.
- CMS/No Surprises Act enforcement is increasing focus on continuous data accuracy.
- Regulators accept audit-ready digital logs as proof of data provenance.
- Providers and payers can securely verify and correct data via authenticated portals.

### Technical Feasibility Assumptions

- Public APIs or downloadable data exist for NPPES, PECOS, and Care Compare.
- ML-based entity resolution can achieve ≥ 90 % accuracy using NPI, name, and address.
- Confidence scoring and source weighting are feasible for truth selection.
- Downstream payer systems can integrate SSOT data via FHIR APIs or batch exports.

# STAKEHOLDER ANALYSIS

HIGH IMPACT STAKEHOLDERS

MEDIUM IMPACT STAKEHOLDERS

## External Stakeholders
*Ecosystem Users & Beneficiaries*

### Primary Stakeholders
- Compliance Officers
- Network Managers
- Data Analysts /Provider Data Ops
- Regulatory Reporting Teams

### Secondary Stakeholders
- CMS & State Regulators
- Healthcare Providers
- Health Plan Executives
- Members / Patients

## Internal Stakeholders
*Solution Enablers(Hilabs)*

- Data Solutions & Engineering Team
- Product Development Team
- Compliance & Governance Team
- Enablement & Adoption Focused

## USER PERSONAS

### Network Manager – Noah Reed

Profile:
*38 years old | Oversees provider directories*
*Works with credentialing and analytics teams.*
*KPI: Maintain network adequacy by county & specialty.*

**Goals:**
- **Maintain compliant and accessible provider networks**
- **Improve onboarding and credentialing turnaround time**
- **Use data to make evidence-based contracting decisions**

**Pain Points:**
- **Fragmented provider data across claims, credentialing, and CRM systems. Slow credentialing and onboarding processes due to duplicate or outdated records. Difficult to identify gaps in specialties or service coverage areas.**

### Compliance Officer – Clara Matthews

*Profile*
- *42 years old | Works at a national payer*
- *Leads a small compliance and data-governance team.*
- *Reports to VP, Regulatory Affairs.*

**Pain Points:**
- **Manual audits and reconciliation for CMS & state directory accuracy checks. High pressure to meet No Surprises Act and Cures Act deadlines.**

**Goals:**
**Achieve 100% CMS submission acceptance with minimal rework.**
**Maintain continuous compliance via automated validation.**
**Access audit-ready reports on demand.**

## COMPLIANCE OFFICERS

**User Pain Points:**
Manual, time-intensive audits and CMS submissions, compounded by constantly evolving federal and state data standards (CMS, NCQA, No Surprises Act), create a high risk of compliance penalties and delayed reporting.
**User Needs:**
Real-time compliance dashboards with automated accuracy checks, CMS-ready reporting templates, and audit logs with full data provenance deliver proactive discrepancy alerts before audits to ensure continuous compliance.
**Goals:**
- Ensure provider data meets CMS directory accuracy and Cures Act interoperability rules.
- Reduce audit remediation effort and compliance costs.
- Transition from reactive audits to continuous compliance monitoring.

## NETWORK MANAGERS

**User Pain Points:**
Fragmented provider data across systems leads to difficulty identifying network gaps by specialty or geography, slow credentialing cycles, and inconsistent provider information that impacts network adequacy.
**User Needs:**
Integrated network-adequacy dashboards and geo-mapping tools with automated provider-status reconciliation and confidence-scored records to support faster, data-driven contracting and onboarding decisions.
**Goals:**
- Maintain accessible and compliant provider networks per CMS adequacy standards.
- Improve onboarding speed and accuracy through data automation.
- Strengthen recruitment and retention using real-time network insights.

## PROVIDER DATA OPERATIONS

**User Pain Points:**
High data volumes, inconsistent field structures, and manual deduplication across multiple sources limit visibility into data quality and delay downstream processes.
**User Needs:**
Automated data-validation pipelines with cross-source reconciliation, real-time discrepancy alerts, and standardized schemas aligned with FHIR R4 US Core to ensure consistency and accuracy across systems.
**Goals:**
- Achieve > 95 % provider-data accuracy.
- Save analyst hours through ML-based entity-matching automation.
- Deliver unified, high-trust provider datasets for claims, directories, and compliance systems.

## HILABS INTERNAL TEAMS

**User Pain Points:**
Complex data integration requirements, evolving FHIR/CMS compliance rules, and the need to maintain high data throughput while ensuring security and uptime.
**User Needs:**
Robust ETL pipelines, scalable microservices, automated testing for FHIR compliance, and real-time monitoring to guarantee platform stability and accuracy.
**Goals:**
- Maintain > 99.9 % data-pipeline uptime and audit readiness.
- Align product roadmap with CMS and payer requirements.
- Enable continuous learning models for better entity-resolution accuracy.

# MCheck™ ProviderSSOT

**It unifies fragmented records from NPPES, PECOS, AMA, and state sources using machine learning, FHIR standardization, and confidence scoring to deliver continuously verified, audit-ready data. With automated reconciliation, anomaly detection, and real-time API integration, MCheck™ ProviderCore transforms fragmented provider data into trusted intelligence for payers, providers, and regulators.**

## SOLUTION ARCHITECHTURE

### Data Ingestion & Integration Pipeline

**Connectors & APIs**

Establish secure API-based integrations with federal and state systems (NPPES, PECOS, AMA, Care Compare, State Boards) using standardized data formats like HL7, FHIR, and CSV.

**ELT Framework**

Deploy scalable Apache Spark or NiFi pipelines to extract, transform, and load data, enforcing schema normalization and canonical modeling for seamless source harmonization.

*(Data Fragmentation and Inconsistency issue resolved)*

***(Faster ingestion, 80% less manual effort)***

### Data Storage & Management

**Cloud Data Lake/Warehouse:**

Store raw and processed data on scalable platforms (AWS S3 + Redshift, Azure Data Lake) Use database engines optimized for analytical queries and large datasets.

**Data Management System:**

Deploy MDM to maintain best versions of provider records with metadata, version control, and access controls.

*(Solves data versioning and governance issues)*

***(5× faster queries, full traceability)***

### AI/ML for Entity Resolution & Anomaly Detection

**Entity Resolution Algorithms:**

Use probabilistic record linkage with ML classifiers, feature embeddings for names and addresses, and graph databases for large-scale relationship mapping and deduplication.

**Anomaly Detection**

Unsupervised learning models (Autoencoders) for detecting inconsistent or outlier data points. Rule-based engines complement ML for domain-specific validation .

*(Solves duplicate and inconsistent records)*

***(95% match accuracy, 85% duplicate reduction)***

### Security & Compliance

**Data Encryption & Access Control:**

Encrypt data in transit and at rest. Role-based access management compliant with HIPAA.

**Audit & Regulatory Compliance Module**

Full audit trail of data provenance, corrections, user actions. Automated compliance checks aligned with CMS and No Surprises Act.

*(Solves unauthorized access and regulatory risks)*

***(100% HIPAA compliance, 80% faster audits)***

### Search & API Delivery Layer

**Search-Optimized Index**

Enables fast, user-friendly searches by NPI, name, or location using ElasticSearch for instant provider data retrieval.

**RESTful APIs with FHIR Support**

Delivers real-time, validated provider data to downstream payer systems through FHIR-compliant APIs supporting versioning, authentication, and fine-grained access control.

*(Solves inefficient data access and poor interoperability)*

***(Sub-second search, seamless integrations)***

### Data Quality Monitoring & Feedback Systems

**Continuous Monitoring**

Scheduled API polling and incremental updates ensure freshness.Data quality dashboards highlighting completeness, accuracy, latency metrics.

**User Feedback Interface**

Web portal enables provider and payer verification with audit trails, using NLP to process unstructured feedback.

*(Solves data staleness and correction delays).*

***(Real-time updates, 60% fewer errors)***

**Data Access & Governance**

| ACCESS LAYER | DESCRIPTION |
|---|---|
| **Internal APIs** | Used by HiLabs and health plan systems to fetch or update provider data via secure endpoints. |
| **Provider Portal** | Web interface with role-based access for providers/payers to verify and update details. |
| **Audit & Provenance Logs** | Tracks every modification with timestamp, source, and user — ensuring accountability. |

## MVP FEATURES

**MULTI-SOURCE DATA INGESTION**
**Tech Stack to be used**
*Apache NiFi, AWS Glue, FHIR/HL7 APIs*

**ENTITY RESOLUTION & MATCHING**
**Tech Stack to be used**
*Python (scikit-learn, XGBoost), Neo4j, MLflow*

**UNIFIED SSOT REPOSITORY**
**Tech Stack to be used**
*AWS S3, Redshift, Talend MDM*

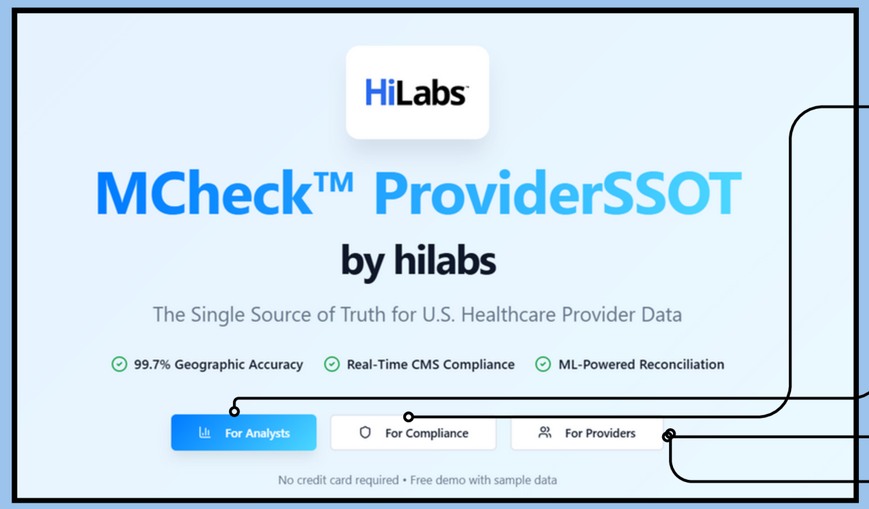
**UNIFIED SSOT REPOSITORY**
**Tech Stack to be used**
*AWS S3, Redshift, Talend MDM*

**COMPLIANCE DASHBOARD & MONITORING**
**Tech Stack to be used**
*Grafana, Prometheus, AWS CloudWatch*

**ANOMALY & DISCREPANCY DETECTION**
**Tech Stack to be used**
*Isolation Forest, Autoencoders, Spark ML*

**SEARCH & API DELIVERY LAYER**
**Tech Stack to be used**
*ElasticSearch, FastAPI, OAuth 2.0*

## Platform Workflows for Key User Segments

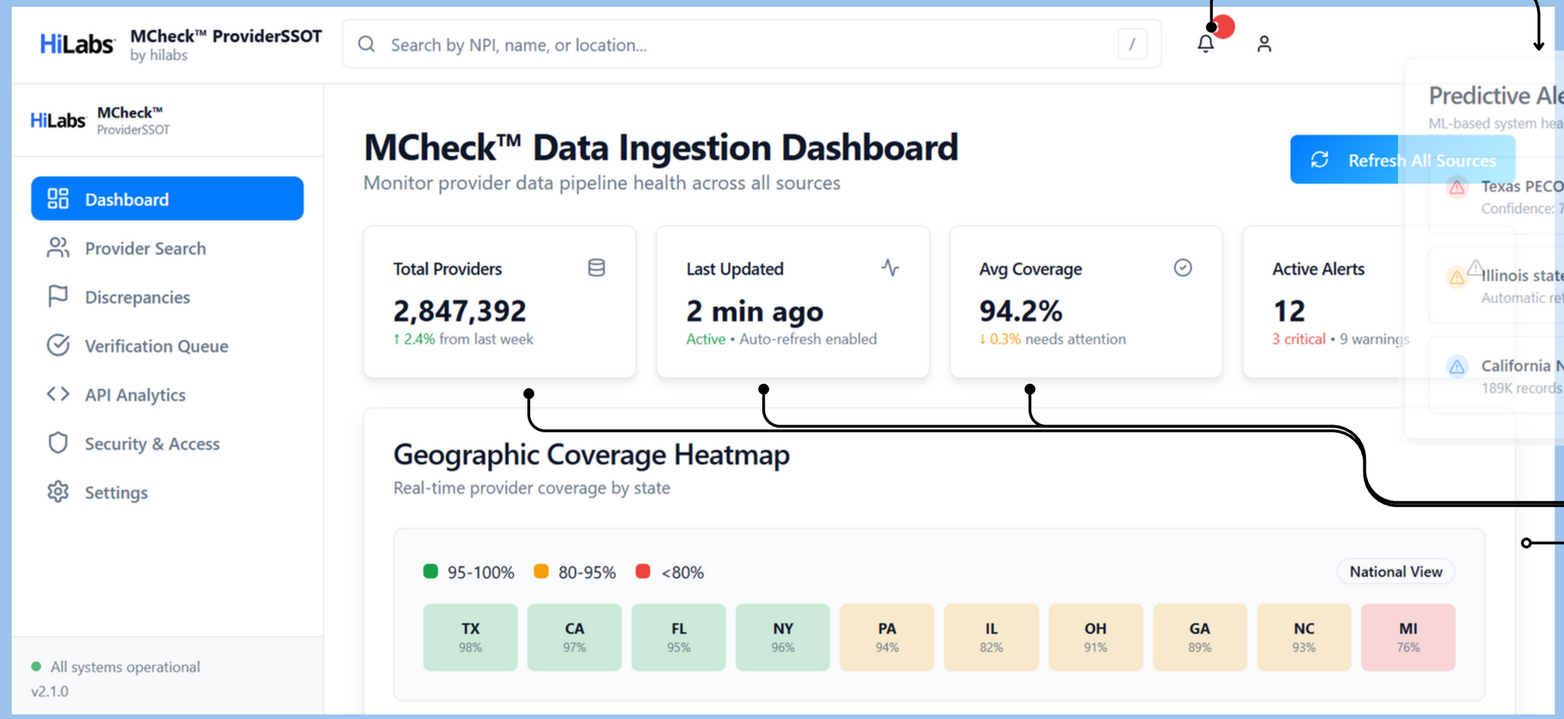


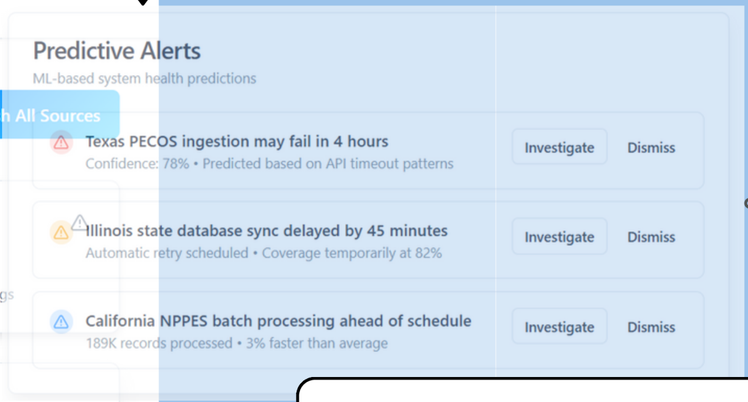


## Key Highlights of the MCheck™ Platform

- **Instant Clarity & Purpose** – The hero headline clearly defines MCheck™ as a geo-intelligent, ML-driven SSOT that unifies fragmented healthcare provider data.
- **Quantified Trust Indicators** – Badges like 99.7% Geographic Accuracy, Real-Time CMS Compliance, and ML-Powered Reconciliation build immediate credibility.
- **Role-Based Personalization** – CTAs for Analysts, Compliance Officers, and Providers make the platform feel tailored to each user's workflow.
- **Clean, Professional UI/UX** – Minimalist layout, soft gradients, and check-marked role cards communicate trust, precision, and enterprise polish.
- **Proof of Scale & Impact** – Metrics such as 10M+ Providers Validated and 100% CMS Compliance showcase the platform's reach and reliability.

## MCheck™ Data Ingestion Dashboard



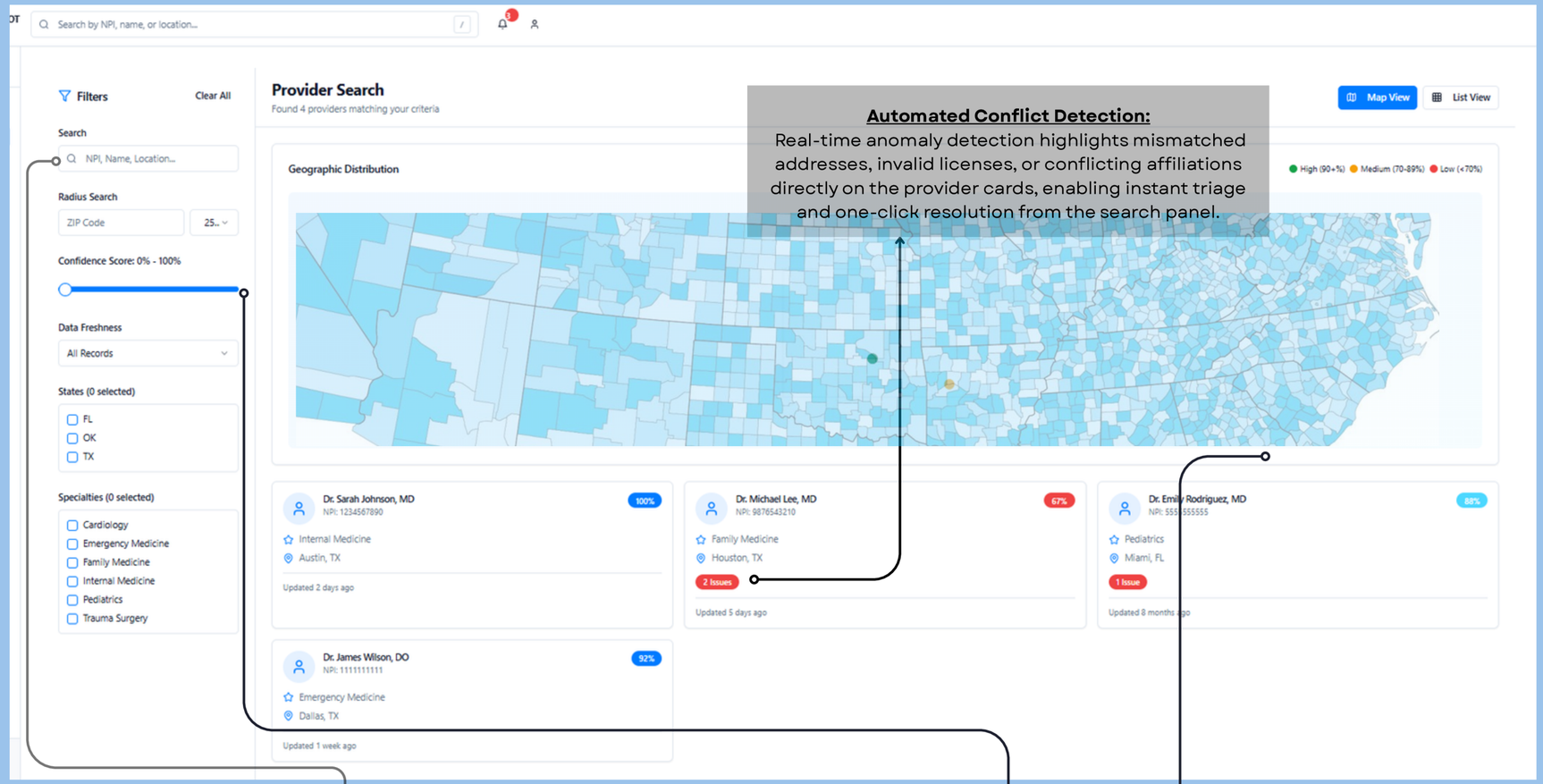

The notification system provides real-time alerts on data sync issues, compliance risks, and verification updates using ML-based confidence scoring. It prioritizes critical events and auto-refreshes via APIs to enable proactive issue resolution and ensure continuous data reliability.

The Data Ingestion Dashboard provides a unified, real-time view of provider data pipeline health across all integrated sources (NPPES, PECOS, AMA, and state boards). It tracks the total number of providers, last data refresh time, average coverage percentage, and active system alerts.

The Geographic Coverage Heatmap visually highlights data completeness across states – enabling teams to instantly identify low-coverage regions (like Michigan at 76%) and prioritize corrective action.

## AI-Powered Provider Search with Entity Resolution & Confidence Scoring

*(Enables intelligent provider lookup with ML-based record linkage, confidence metrics, and map-based accuracy visualization)*




**Automated Conflict Detection:**
Real-time anomaly detection highlights mismatched addresses, invalid licenses, or conflicting affiliations directly on the provider cards, enabling instant triage and one-click resolution from the search panel.

**Smart, ML-Powered Entity Search**
Enables federated provider lookup across NPPES, PECOS, and AMA datasets with advanced filtering by NPI, specialty, and geographic radius (within 25 miles). The search dynamically ranks results using ML-based similarity scoring for faster, context-aware retrieval.

**Confidence Scoring (0–100%)**
Derived from a hybrid Fellegi-Sunter probabilistic model enhanced with XGBoost classifiers, each record is assigned a confidence score based on multi-source consistency, geospatial accuracy, and data recency – giving users interpretable reliability metrics.

**Map-Based Visualization:**
Interactive clustering groups providers by data reliability tiers (High ≥90%, Medium 70–89%), allowing users to visually assess data completeness and perform spatial comparisons across states or ZIP codes in real time.

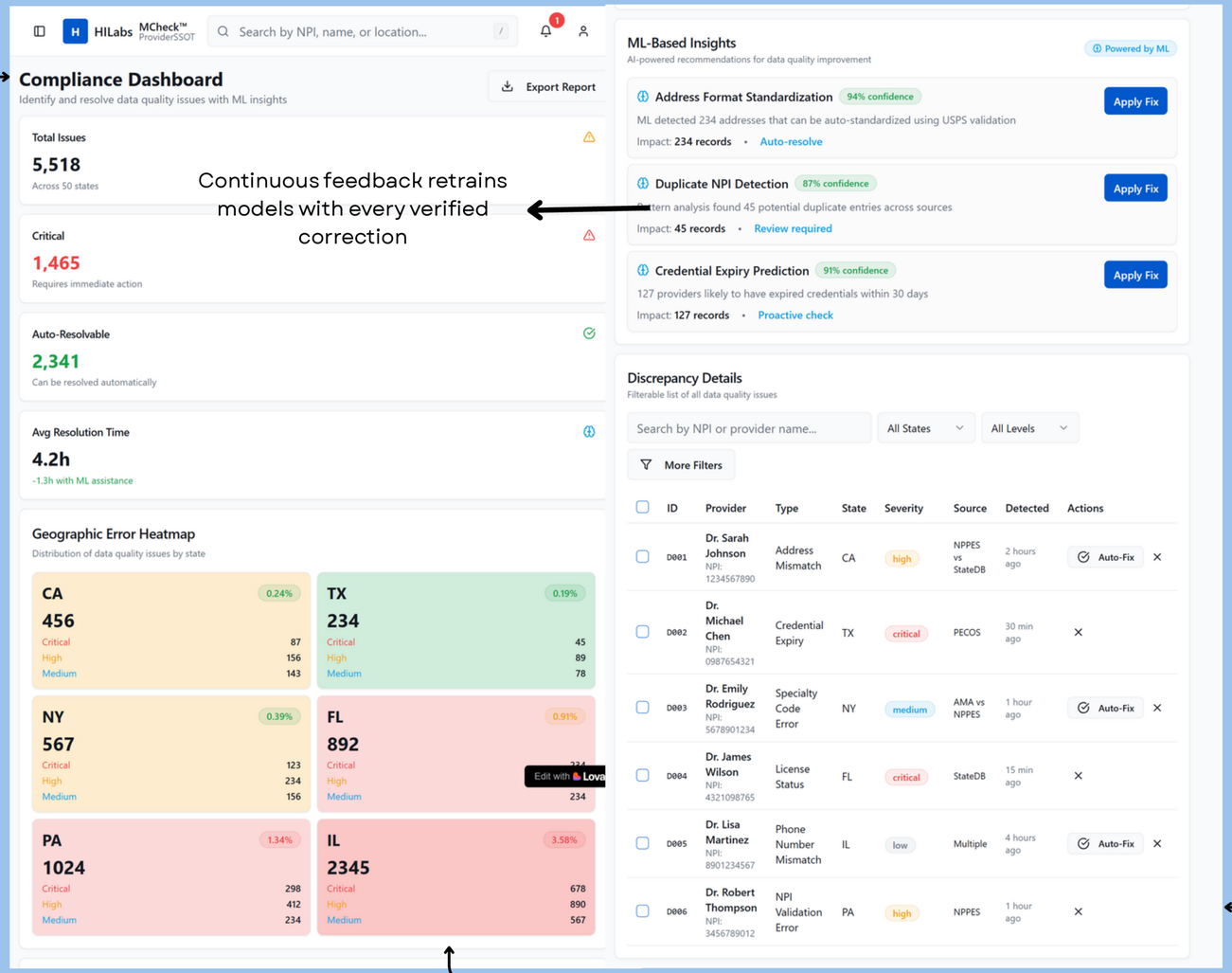
## Unified Provider Profile (Entity Resolution View)

*Ensures 99.7% geographic accuracy and eliminates fragmented or outdated provider data.*


**Unified Provider Profile (Entity Resolution View)**

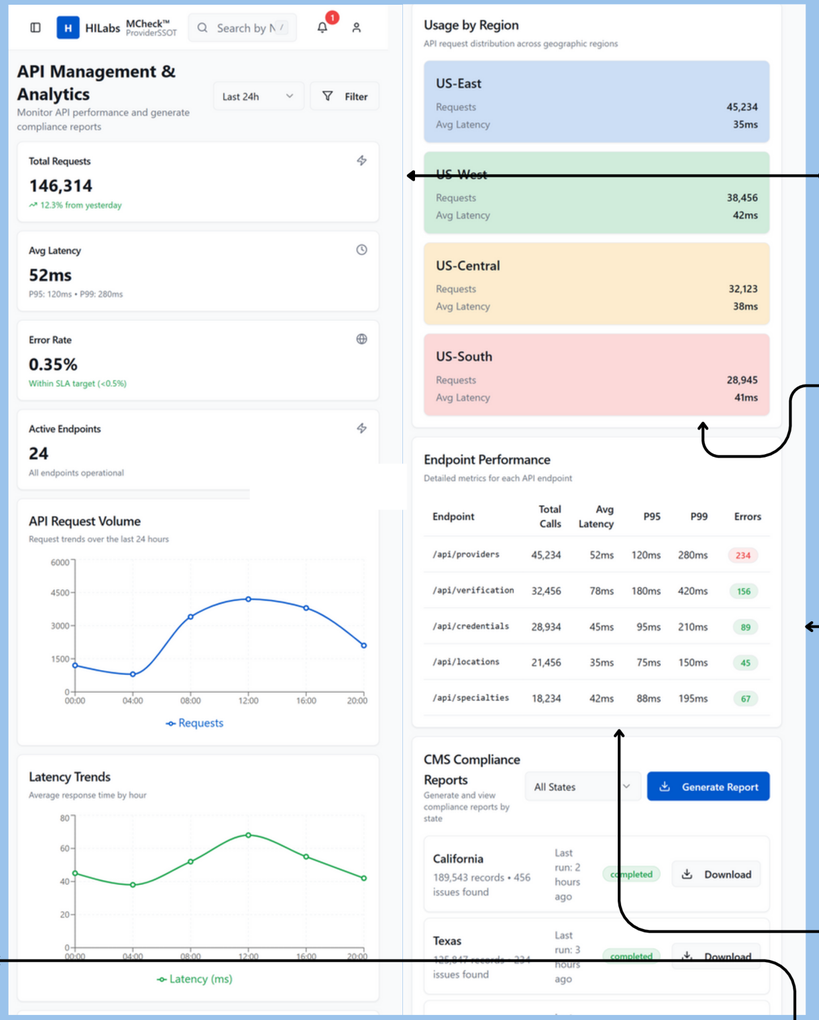
- Consolidated provider record showing verified contact, location, and practice data.
- ML-based source weighting algorithm calculates confidence by evaluating NPPES, PECOS, AMA, and state datasets.
- Geo-verification validates physical addresses through geocoding APIs (USPS + Google Maps).
- Dynamic audit trail provides source traceability for every verified field.

## Compliance Dashboard & ML-Based Insights




Continuous feedback retrains models with every verified correction

## Intelligent API Monitoring with Predictive Latency




Real-time API performance tracking for 24 active endpoints across AWS regions.

Auto-scaling triggers maintain uptime across distributed clusters.

Latency prediction models analyze hourly response times, keeping average below 52ms.

ML-driven anomaly detection flags SLA breaches or high error trends (>0.5%).

**Automated Resolution Engine**

An ML rules engine applies confidence-driven reconciliation (e.g., higher source weighting for PECOS vs. NPPES) to automatically correct 45% of data issues, reducing manual intervention and ensuring consistent provider profiles across systems.

A geo-intelligent visualization layer highlights compliance-critical states (e.g., Illinois – 2,345 critical anomalies), helping compliance officers instantly locate, prioritize, and resolve high-risk regions through state-level drill-downs.

**AI-Powered Discrepancy Detection:**

Uses Isolation Forest and BERT-based NLP models to identify semantic and numerical inconsistencies across provider records – such as mismatched addresses, outdated credentials, or invalid taxonomy codes – with over 92% anomaly detection accuracy.

## LAYER 1: DATA SOURCE INGESTION

NPPES PECOS AMA State Boards

Apache Airflow

kafka

Apache Spark

Geocoding

## LAYER 2: ML RECONCILIATION ENGINE

Discrepancy Detection

Predictive Analytics

Stakeholder:
ML Engineers
Data Scientists

Entity Resolution

Confidence Scoring

0.4×Source + 0.3×Recency + 0.2×Agreement + 0.1×GeoPrecision

## LAYER 3: SINGLE SOURCE OF TRUTH DATABASE (SSOT)

Data Entry Point

*These JSON objects are written into PostgreSQL as the authoritative system of record.*

```json
{
  "npi": "1457893210",
  "provider_name": "Dr. Jane L. Smith",
  "specialty": "Cardiology",
  "confidence_score": 0.96,
  "addresses": [
    {"street": "123 Medical Plaza", "city": "Austin", "state": "TX", "lat": 30.2672, "lon": -97
  ],
  "licenses": [{"state": "TX", "number": "TX-20391", "status": "Active"}],
  "source": "PECOS"
}
```

PostgreSQL

PostgreSQL Core Operations

*Performs **normalization, indexing, and row-level security** for state isolation. Triggers generate **CDC (Change Data Capture)** events for downstream sync.*

→ Redis (Cache Layer)
→ Elasticsearch (Search Index)
→ TimescaleDB (Metrics & Monitoring)

[API Gateway Consumes from All Three]

Stake holders: Compliance & DB Admin Teams

## LAYER 4: API GATEWAY & BUSINESS LOGIC

GraphQL + APOLLO

GraphQL Gateway (Apollo Server)

*The GraphQL Gateway optionally replaces multiple REST calls with a single nested query..*

REST API (Node.js+ Express)

## LAYER 5: PRESENTATION & USER INTERFACE

React 18 SPA (REST and GraphQL APIs)

- Airflow/Spark status → via /analytics
- Provider records → via /providers/{npi}
- Discrepancies → via /discrepancies

Role-Based Dashboards

- Analyst → Data ingestion visualization
- Compliance Officer → Discrepancy manager.
- Provider → Self-service verification portal.
- Developer → API usage & latency metrics.
- Admin → Access control and activity audit.

Metrics Integration

- Real-time performance data pulled from TimescaleDB.
- Visualized as charts, heatmaps, and maps

## LAYER 6: MONITORING & OBSERVABILITY

Datadog log & APMs

TimescaleDB KPI Trends

Unified Dashboard
API, Data, ML Metrics

Alerts (Slack / Email)
DevOps & Security Notified

## LAYER 7: SECURITY & COMPLIANCE

Identity & Access Management

Data Encryption
HIPAA Controls

Geo-Fencing & Segmentation
RLS by State

HIPAA Compliance & Auditing
- 7-yr log retention (AWS BAA)
- Regular security reviews

Continuous Feedback Loop
- Provider / Analyst updates
- Feeds ML retraining via Kafka
- Improves confidence scores

| PITFALLS | DESCRIPTION | MITIGATION STRATEGY |
|---|---|---|
| **1. Data Drift in ML Models** *SEVERITY- 🔴 High* | Model accuracy declines over time as provider datasets evolve (new NPIs, merged practices). Can cause >10% drop in matching precision. | Continuous retraining using verified corrections and incremental learning pipelines to maintain ≥95% entity resolution accuracy. *LoE- 🟠 Medium* |
| **2. Inconsistent Data Standards (Across NPPES, PECOS, AMA)** *SEVERITY- 🔴 High* | Schema mismatches and field inconsistencies create ingestion bottlenecks and data loss during mapping. | FHIR R4–based canonical schema alignment with dynamic ETL validation in Apache NiFi and Spark for seamless normalization. *LoE- 🟠 Medium* |
| **3. False Positives in Anomaly Detection** *SEVERITY- 🔴 High* | Over-sensitive ML models trigger unnecessary alerts, overwhelming compliance users. | Calibrate thresholds, apply SHAP explainability, and confidence-based filtering to reduce false positives by ~40%. *LoE- 🟢 Low* |
| **4. Data Privacy & Security Risks** *SEVERITY- 🟠 Medium* | Unauthorized access or mishandling can trigger severe penalties and loss of trust. | AES-256 encryption, RBAC, AWS KMS key rotation, and immutable audit logs to ensure full compliance and traceability. *LoE- 🟠 Medium* |
| **5. Model Interpretability During Compliance Audits** *SEVERITY- 🟠 Medium* | Regulators demand transparency; unexplained ML corrections risk audit rejections. | Embed explainable AI modules that display source provenance, feature importance, and allow manual overrides. *LoE- 🟢 Low* |
| **6. API Downtime & Data Latency** *SEVERITY- 🟠 Medium* | External data source outages delay ingestion, reducing data freshness and CMS readiness. | Asynchronous API calls, auto-retry logic, and cached snapshots to sustain >99% uptime and real-time sync. *LoE- 🟢 Low* |

**LoE-Level of effort*

## FUTURE ENHACEMENTS

**Adaptive ML Retraining:**
- Implement reinforcement learning pipelines to automatically adjust model weights based on feedback accuracy, improving provider matching precision to over 97%.

**Expanded Data Integrations:**
- Incorporate payer rosters, claims data, and credentialing APIs to enhance provider validation and detect inactive or duplicate practitioners.

**Explainable AI (XAI):**
- Develop dashboards using SHAP/LIME to show feature-level contributions behind ML confidence scores, improving transparency for compliance reviews.

## SUCCESS METRICS

### North Star Metric

Metric for measuring verified provider data accuracy across the U.S. network

$$\frac{VerifiedProviderRecords(\geq 3Sources)}{TotalProviderRecords} \times 100$$

Goal: **≥ 99%** verified accuracy across **50 states**
Tracks the real-time reliability and freshness of SSOT data aggregated from **NPPES, PECOS, AMA**, and state sources.

### Growth

Metric for measuring system adoption and integration footprint

$$\frac{(NewHealthPlans + ActiveIntegrations - ChurnedClients) \times 100}{TotalClients}$$

### Data Quality Index

Metric for measuring improvement in provider data consistency

$$\frac{(NumberofCleanDataPoints \times AccuracyScore\%)}{TotaldataPoints}$$

### Compliance Efficiency

Metric for assessing CMS audit readiness and remediation speed

$$ComplianceEfficiency = \frac{Auto - ResolvedComplianceIssues \times 100}{TotalComplianceIssues}$$

### System Performance Index

Metric for tracking ingestion, API response, and uptime reliability

$$SystemPerformanceScore = \frac{(IngestionSuccessRate + APIuptime)}{2}$$