Deconstruction, Requirement Analysis & Clarifying Questions

1. Requirement Categorization (30 mins)

The system requirements are classified into four domains to ensure design decisions directly map to business, regulatory, and technical needs.

Functional Requirements (System Capabilities)

- **Multi-Modal Ingestion:** Support for 7 modalities — Clinical Notes (Text, RTF, PDF), Audio Streams, Medical Imaging (DICOM), Structured Records (HL7, FHIR, CDA, X12), Genomic Data (VCF), Insurance/Billing Records, and Clinical Trial Data.
- **Advanced Processing:** OCR for scanned docs/PDFs, speaker diarization for audio, parsing of nested healthcare formats.
- **PII/PHI Detection:** 4-tier entity recognition with recall targets (>99.5% for Tier 1, >90% for Tier 4).
- **Format-Preserving Redaction:** Tokenization/redaction that maintains original structure (e.g., (###) ###-#### → (XXX) XXX-XXXX).
- **Stable Pseudonymization:** Consistent pseudonyms for patients, providers, and family links across all modalities/timeframes.
- **System Interfaces:** REST, GraphQL, gRPC, WebSocket APIs, plus real-time dashboard and mobile apps.
- **EHR Integration:** Native integration with Epic, Cerner, Allscripts.
- **Failure Analysis:** Automated near_miss_analysis reports in JSON for performance diagnostics.

Non-Functional Requirements (The "-ilities")

- **Latency:** <100ms (p95), <200ms (p99). Model inference <10ms/document.
- **Throughput:** 100k documents/sec, 1k audio streams, 50k API req/sec.
- **Reliability & Availability:** 99.99% uptime, 11 nines durability, <30s recovery.
- **Scalability:** 1k+ compute nodes, petabyte data, 10k+ tenants.

Privacy & Compliance Requirements (Constraints)

- Compliance with HIPAA, HITECH, GDPR, FDA 21 CFR Part 11, and global/state laws.
- Formal differential privacy guarantees (ε, δ) with budget tracking.
- Cryptographically verifiable, immutable audit logs.
- Robust against adversarial ML attacks (model inversion, membership inference).
- Role-Based Access Control (RBAC) + “minimum necessary standard.”

ML-Specific Requirements (Intelligence Layer)

- **Custom Model Architecture:** Multi-modal transformer for text, audio, structured data.
- **Privacy-Preserving ML:** Federated Learning, Homomorphic Encryption, Secure Multi-Party Computation.
- **Continuous Improvement:** Active Learning pipeline with human-in-the-loop.
- **Robustness & Explainability:** Adversarial training + interpretable redaction decisions.
- **MLOps:** Concept drift detection, multi-tenant model management, A/B testing.

---

2. Ambiguities & Stated Assumptions (20 mins)

- **Privacy Budgets (ε, δ):** Not specified.
  Assumption: Configurable privacy profiles (strict: ε=1 for research, lenient: ε=8 for analytics).
- **Entity Resolution / MPI:** Unclear if existing MPI is provided.
  Assumption: System builds its own entity graph but supports MPI integration if available.
- **Use-Case Context:** HIPAA's "minimum necessary" depends on purpose (billing vs. research).
  Assumption: Every API request must include a `use_case` parameter. A policy engine will map use cases to tailored redaction/retention rules.

---

3. Core Problem Statement (10 mins)

The mission is to design an enterprise-grade, cloud-native platform capable of:

- **Unifying and processing multi-modal healthcare data in real time.**
- **Detecting, redacting, and pseudonymizing sensitive entities with high precision and sub-100ms latency.**
- **Providing provable differential privacy guarantees and tamper-proof audit trails for compliance with HIPAA/GDPR.**
- **Scaling horizontally to handle petabyte-scale data and extreme throughput.**
- **Employing privacy-preserving ML techniques so the system continuously improves without compromising patient confidentiality.**

In short: The challenge is to architect a secure, privacy-first, high-performance healthcare data processing platform that balances regulatory compliance, ML intelligence, and extreme scale.