National Biomedical Research Network (NBRN) Database Request for Proposal

The NBRN coordinates cutting-edge infectious disease research across 300+ institutions globally. We require a comprehensive database to track:

1. Institutions & Governance

- **Core Attributes**: Institution ID, legal name, year founded, tax status (nonprofit/for-profit), accreditation status (ISO, WHO, etc.), disaster recovery tier (1-4).
- **Subtypes**:
  - *Academic Institutions*: QS World Ranking, number of PhD programs, IRB approval capacity.
  - *Corporate Labs*: Parent company, stock ticker (if public), quarterly R&D budget.
  - *Government Facilities*: Security clearance level (confidential/secret/top secret), agency affiliation (CDC, NIH, etc.).
- **Locations**: Each institution has multiple campuses (address, square footage, BSL lab levels 1-4).

2. Research Personnel

- **Researchers**: NIHHIS ID, ORCID, visa status (for international scholars), security clearance, conflict-of-interest disclosures.
- **Employment History**: Track concurrent appointments (start/end dates, FTE %, department).
- **Qualifications**: Degrees (institution, year, field), certifications (e.g., BSL-3 training), languages spoken.

3. Projects & Compliance

- **Projects**: Protocol ID, pre-registration DOI (e.g., ClinicalTrials.gov), DSMB oversight flag, embargo expiration date.
- **Funding**: Grant numbers (NIH, Wellcome Trust, etc.), quarterly disbursements, indirect cost rate.
- **Regulatory**: IRB approval dates, FDA phase (for trials), export control restrictions (ITAR/EAR).

4. Clinical Trials

- **Participants**: Screening logs, withdrawal reasons, adverse event reports (CTCAE severity grade).
- **Biospecimens**: Repository ID, aliquot counts, freezer location (GPS coordinates), chain-of-custody logs.

5. Intellectual Property

- **Patents**: Filing dates, jurisdictions, licensing revenue.
- **Publications**: Embargo periods, altmetrics scores, preprint server links.
- **Data Sharing**: DUAs, de-identification methods (k-anonymity vs. differential privacy).

6. Reporting Requirements

- Real-time dashboards for:
  - Funding utilization by institution tier
  - Protocol deviations by principal investigator
  - Biospecimen chain-of-custody audits
  - Dual-use research of concern (DURC) flagging

Constraints

- HIPAA/GDPR compliance for participant data.
- Support for 10M+ adverse event records/year.
- Integration with external systems (REDCap, PubMed API).

Deliverables:

1. EERD with all entities/relationships (no missing attributes).
2. SQL schema with constraints (PKs, FKs, CHECKs for validations).
3. Stored procedures for:
   a. Annual conflict-of-interest reconciliation
   b. Export-controlled project flagging

4. Python script to anonymize participant data (FERPA/HIPAA compliant).

Evaluation Criteria:

- Handling of multi-jurisdictional legal requirements.
- Scalability for genomic datasets (50TB+/year).
- Audit trail design for FDA 21 CFR Part 11 compliance.

This problem forces you to:

- Model hierarchical institution types with divergent attributes.
- Handle temporal data (employment history, project phases).
- Design for regulatory complexity (security, privacy, compliance).
- Support advanced analytics (biospecimen logistics, DURC monitoring).

We want to track data from time to time to use SQL to gather the data for analysis.

1. Institutional Analytics

Problem:
"List all government facilities with BSL-4 labs that have conducted Phase III trials in the last 5 years, including their total funding received and average trial severity grade. Exclude institutions under export control restrictions."
Skills Tested:

- Multi-table joins (Institutions → Trials → Funding)
- Date filtering
- Exclusion logic

2. Researcher Workload

Problem:
"Identify researchers holding concurrent appointments at multiple institutions who are PIs on more than 3 active projects. Include their qualification level and total FTE percentage across all roles."
Skills Tested:

- Self-joins or subqueries for concurrent positions
- Aggregation with HAVING
- Percentage calculations

3. Compliance Monitoring

Problem:
"Generate a report of all clinical trials missing IRB renewal dates within the next 30 days, flagged by institution type and PI contact info. Include trials with past-due renewals in red."
Skills Tested:

- Date arithmetic (CURRENT_DATE + INTERVAL '30 days')
- Conditional formatting (use CASE WHEN)
- Hierarchical joins (Institution → Department → Researcher → Trial)

4. Biospecimen Chain-of-Custody

Problem:
"Find all biospecimen aliquots stored in freezers at locations with temperature violations (≥ -70°C) in the last week, tracing back to the originating trial and PI."
Skills Tested:

- Time-series filtering
- Multi-hop joins (Freezer → Biospecimen → Trial → Researcher)
- Threshold validation

5. Publication Impact

Problem:
"Calculate the 5-year h-index for each academic institution, considering only publications linked to NIH-funded projects. Rank institutions by h-index but exclude those with retraction rates >5%."
Skills Tested:

- Window functions (for citation counting)
- Advanced metrics (h-index calculation)
- Anti-joins (exclude retracted papers)

6. Adverse Event Analysis

Problem:
"Compare adverse event rates (events/participant) between corporate and academic trials for Phase II/III COVID-19 studies, adjusted for trial duration. Highlight trials with rates exceeding 2σ above the group mean."
Skills Tested:

- Statistical aggregates (STDDEV, AVG)
- Cohort comparison
- Z-score calculation

7. Funding Efficiency

Problem:
"Identify the top 10% most 'efficient' researchers (publications per $1M funding) in malaria research, but only if they've authored at least 3 papers with impact factor ≥10 in the last 3 years."
Skills Tested:

- Percentile calculation (NTILE)
- Multi-condition filtering
- Cost-benefit ratios

8. Conflict of Interest

Problem:
"Detect researchers who reviewed NIH grant proposals while being co-investigators on projects competing for the same funding pool within ±6 months of the review date."
Skills Tested:

- Temporal overlap detection (BETWEEN)
- Many-to-many relationship traversal
- Ethical constraint checks

9. Data Anonymization

Problem:
"Create a HIPAA-compliant view of participant data that masks direct identifiers (name, email) but preserves age brackets (18-30, 31-45, etc.) and trial outcomes for analysis."
Skills Tested:

- Data masking (SUBSTRING, HASH)
- Bucketization (WIDTH_BUCKET)
- View creation with access control

10. Geospatial Logistics

Problem:
"Optimize biospecimen transfer routes by finding all freezer pairs within 50 miles where one has >80% capacity and the other has >90% utilization. Include driving time estimates via HERE API."
Skills Tested:

- Geospatial joins (PostGIS or equivalent)
- External API integration
- Resource allocation logic

Bonus: Machine Learning Prep

Problem:
"Export a dataset for predicting trial delays: include project start/end dates, PI publication history, institution funding trends, and quarterly adverse event counts—featurized as time-series arrays."
Skills Tested:

- JSON/array aggregation
- Time-series windowing
- ML-ready data structuring

National Biomedical Research Network (NBRN) - Advanced Analytics RFP

1. Core Deliverables

A. Snowflake Data Warehouse

- Ingest structured/unstructured data from:
    - Clinical trial EDC systems (REDCap, Medidata)
    - IoT freezer monitors (temperature logs)
    - Researcher ORCID profiles (API JSON)
- Implement **CDC (Change Data Capture)** for IRB protocol amendments.
- Design **aggregation strategies** for:
    - Daily biospecimen inventory snapshots
    - Real-time adverse event monitoring

B. Real-Time Anomaly Detection

- Deploy a **Kafka pipeline** detecting:
    - Abnormal temperature fluctuations in specimen freezers (±2σ from 24h rolling avg)
    - Suspicious login attempts to trial databases (geo/IP anomalies)
- Stream results to Snowflake via **Snowpipe** with severity tagging.

C. Multi-Model ML Comparison

- **Predictive Task**: Forecast trial delays (classification ±14 days from protocol) using:
    - **XGBoost** (feature importance for compliance audits)
    - **Neural Network** (LSTM for temporal patterns in amendment history)
- Compare metrics:
    - Precision/recall for FDA audit risk cases

    - Inference latency (Snowpark vs. external ML serving)

D. Cost-Optimized Architecture

- Provide **monthly cost projections** for:
    - Warehouse sizing (X-Small vs. Medium for Spark clusters)
    - Storage tiers (transient vs. permanent trial data)
    - Query acceleration (CACHE vs. materialized views)
- **Constraint**: Budget cannot exceed $12k/month at 50TB scale.

2. Visualization Requirements

A. Dynamic Dashboards

- **Tableau/PowerBI**:
    - Participant dropout rates by institution type (academic vs. corporate)
    - Freezer capacity heatmap (Plotly GeoJSON + Snowflake coordinates)
- **Streamlit App**:
    - Real-time anomaly alerts with Kafka event playback
    - ML model drift monitoring (Evidently reports)

B. Automated Reports

- **Weekly PDFs** (generated via Python):
    - Top 5 PIs by publication impact factor (matplotlib tables)
    - Cost variance analysis (Snowflake credit usage vs. budget)

3. Compliance & Validation

- **Audit Trail**:
    - Snowflake TIME TRAVEL for all protocol changes (7-day retention)
    - GDPR right-to-be-forgotten workflow (anonymize participant IDs)
- **Model Cards**:
    - SHAP values for delay prediction model
    - FDA 21 CFR Part 11 compliance checklist