In the banking industry, especially when dealing with **corporate clients**, understanding the **full picture of financial risk** is critical. Large companies often have complex structures with multiple subsidiaries, shared directors, loan obligations, pending legal cases, or connections to high-risk entities. These risk factors are not always apparent from a single document.

U.S. banks are under increasing pressure to maintain **transparency and accountability** in their corporate lending and onboarding processes. Regulatory frameworks such as the:

- **Bank Secrecy Act (BSA)**
- **Anti-Money Laundering (AML) laws**
- **Customer Due Diligence (CDD) Rule**
- **USA PATRIOT Act**
- **OFAC sanctions compliance**

...require banks to not only verify a corporate client's identity but also **understand the network of entities and individuals behind them**—including **ultimate beneficial owners**, **shared executives**, **cross-border links**, and **hidden liabilities**.

However, this information is buried across disparate, unstructured sources:

- SEC filings (e.g., 10-K, 8-K)
- Loan agreements and credit memos
- KYC documents and internal onboarding forms
- Legal documents and litigation reports
- Public news, regulatory alerts, or watchlists

Manually connecting the dots across these documents to uncover **non-obvious but high-risk relationships** is time-consuming, inconsistent, and prone to oversight.

⚠️ Compliance & Risk Gaps This Solution Addresses

Today's Questions That Are Hard to Answer Quickly:

1. **"Is this company ultimately owned or linked to a sanctioned individual or entity?"**

    - Cross-document extraction needed to identify UBOs and check them against OFAC or FinCEN lists.

2. **"What is the total credit exposure across all entities controlled by or connected to a parent group?"**
    - Relationships hidden across separate credit files, subsidiaries, and historical data.
3. **"Are there any indirect links between two clients via common directors, law firms, or shell companies?"**
    - Needs entity and relationship mapping across onboarding docs, filings, and external sources.
4. **"Is there any pattern of circular lending, hidden guarantees, or layered transactions that resemble money laundering?"**
    - Requires complex multi-hop traversal of entities and financial relationships.
5. **"Has this corporate client previously been involved in litigation or regulatory scrutiny that we missed in onboarding?"**
    - Buried in unstructured memos, media reports, and case documents.

🧠 Why This is a Data + AI Problem

While banks have the documents, they often lack the tools to extract and **connect entity and relationship intelligence** across them. Traditional rule-based systems struggle with:

- **Cross-document linking** of the same entities (e.g., “XYZ Inc.” vs “XYZ Holdings”)
- **Context-aware relationship detection** (e.g., “guaranteed a loan”, “controlled through a trust”)
- **Temporal analysis**, e.g., tracing changes over time in ownership or board structure

This is where **Large Language Models (LLMs)** and a **graph database like Neo4j** provide a breakthrough. LLMs can extract and normalize complex entities and relationships across unstructured text, while Neo4j stores them in a queryable graph that allows analysts and systems to reason over the entire relationship network.

✅ Desired Solution Capabilities

A successful solution should allow analysts to answer queries such as:

Query Type	Example
Ownership/Control Risk	"List all entities directly or indirectly owned by John Doe (a PEP or sanctioned person)."
Credit Exposure	"What is the cumulative loan exposure across all subsidiaries of XYZ Corp?"
Conflict of Interest	"Are any of our loan recipients connected by shared executives or board members?"
Litigation Risk	"Show me all clients involved in lawsuits over the past 5 years, regardless of which entity name appears."
AML Pattern Detection	"Trace all money flows and guarantees linked to ABC LLC and identify circular or suspicious chains."

🏁 Outcome

The final system will provide:

- A **risk knowledge graph** in Neo4j constructed from extracted relationships
- A user interface or API to **query, visualize, and monitor** risk connections
- An auditable pipeline for compliance and audit teams to validate the extraction and reasoning

---

Recommended Datasets for This Use Case

1. SEC EDGAR Filings (10-K, 8-K, 10-Q)

- **What it contains**: Annual and quarterly reports from U.S. public companies, including details on subsidiaries, executives, litigation, risk factors, and financial performance.
- **Useful for**:
    - Entity extraction (subsidiaries, officers, legal entities)
    - Relationship extraction (ownership, lawsuits, debts)

- **Access**:
    - https://www.sec.gov/edgar/searchedgar/companysearch.html
    - Bulk download via EDGAR FTP

---

2. OpenCorporates

- **What it contains**: Corporate registration data (names, jurisdictions, officers, status, parent/child relationships).
- **Useful for**:
    - Cross-document entity resolution
    - Building corporate ownership structures
- **Access**:
    - Free tier API (limited)
    - https://opencorporates.com/info/about

---

3. Global Sanctions Lists (OFAC, EU, UN)

- **What it contains**: Lists of sanctioned individuals and organizations, often with aliases and metadata.
- **Useful for**:
    - Flagging high-risk individuals/entities
    - Link analysis with sanctioned parties
- **Access**:
    - OFAC SDN List: https://sanctionssearch.ofac.treas.gov/
    - Download in CSV/XML formats

---

4. Financial News Articles & Legal Datasets (CourtListener / RECAP)

- **What it contains**: Case law, litigation filings, and news coverage of corporate legal disputes.

- **Useful for**:
    - Identifying legal exposure across multiple entities
    - Entity/event extraction
- **Access**:
    - https://www.courtlistener.com/
    - Use APIs or web scraping (public domain)

---

5. Synthetic or Generated Datasets

For a hackathon, you can also create or augment data to simulate:

- Interlinked loan agreements
- Board member overlaps
- Shell company structures

Use LLMs (like GPT-4) to generate realistic text for internal memos, emails, or onboarding documents that resemble what a bank might actually have internally.

🔧 **How to Use These Datasets in the Hackathon**

Step	Description
1. **Document Collection**	Pull 10-Ks, sanctions list entries, and company profiles. Add sample internal docs (synthetic).
2. **LLM-based Extraction**	Use GPT-4 or LangChain agents to extract entities (companies, people, roles, relationships) from each doc.
3. **Entity Resolution**	Match "XYZ Corp", "XYZ Inc.", and "XYZ Holdings" as the same entity using embeddings or rules.
4. **Graph Construction**	Load nodes and edges into Neo4j — each doc becomes a source for relationships with metadata.
5. **Query Demo**	Showcase Cypher queries: “Find all companies connected to John Doe in 2 hops”, “Show litigation links”, etc.

🚀 **Bonus: Ready-to-Use Starter Projects**

- Neo4j + SEC Filings Graph Example (public repo)
- OpenSanctions – unified API for global watchlists
- FinCEN Files (ICIJ) – historical SAR (suspicious activity reports), can be used as a model

---

Let me know if you want help building a small starter dataset or notebook that parses filings and pushes into Neo4j with relationships.