1. **Introduction**
2. **Graph Database Modelling**

A diagram of a computer

AI-generated content may be incorrect.

Figure 1: Entity–Relationship Diagram (ERD) of the original relational model representing fund management, ESG metrics, and portfolio structure before graph transformation.

**Description and Explanation**

The relational model above represents the **core data structure** of the ESG fund management system. Each table (entity) contains a primary key (PK) for unique identification, and foreign keys (FK) that define the relationships between entities.  
However, when migrating to a **graph database (Neo4j)**, these relationships are modelled as **explicit graph edges**, which simplifies queries and improves traversal performance.

|  |  |  |
| --- | --- | --- |
| **Relational Table** | **Graph Node** | **Key Relationships (Graph Edges)** |
| Fund\_Manager | (:FundManager) | (:FundManager)-[:MANAGES]->(:InvestmentFund) |
| Investment\_Fund | (:InvestmentFund) | (:InvestmentFund)-[:HOLDS]->(:Company) |
| Portfolio\_Company | (:Company) | (:Company)-[:ASSESSED\_FOR]->(:ESGMetric) |
| ESG\_Metric | (:ESGMetric) | (:ESGMetric)<-[:PROVIDES]-(:DataSource) |
| ESG\_Data\_Source | (:DataSource) | (:DataSource)-[:PROVIDES]->(:ESGMetric) |
| ESG\_Performance | Relationship properties node | (:Company)-[:ASSESSED\_FOR {performance\_id,...}]->(:ESGMetric) |
| Regulatory\_Framework | (:RegFramework) | (:ESGMetric)-[:DEFINED\_BY]->(:RegFramework) |
| Audit\_Trail | (:AuditEvent) | Optional standalone entity for version tracking |
| Investment\_Position | Relationship edge | (:InvestmentFund)-[:HOLDS {position\_id,...}]->(:Company) |

**Analytical Narrative**

The migration from relational tables to a graph structure ensures **semantic clarity** and **efficient traversal** for ESG analytics:

* Each **entity (table)** becomes a **node label** in Neo4j.
* Each **foreign key relationship** becomes a **graph edge (relationship)** with meaningful names such as MANAGES, HOLDS, ASSESSED\_FOR, and PROVIDES.
* **Complex joins** (e.g., linking funds → companies → ESG metrics → data sources) are replaced with **single-hop traversals**, drastically improving query performance.
* The design supports **multi-level analysis**, enabling users to trace **ESG performance across portfolios**, **identify data dependencies**, and **evaluate fund-level sustainability scores**.

**Graph Database Modelling:**

**A screenshot of a computer

AI-generated content may be incorrect.**

Figure 2 Neo4j Schema Visualization:

This figure shows Neo4j’s automatically generated schema view, representing the structure of the graph model derived from the relational design. Each node label (e.g., FundManager, InvestmentFund, Company, ESGMetric, DataSource, RegFramework, AuditEvent) is connected by relationships such as MANAGES, HOLDS, ASSESSED\_FOR, and PROVIDES. This confirms that all entity relationships were correctly implemented in the graph structure.

A computer screen shot of a diagram

AI-generated content may be incorrect.

Figure 3: Graph Model Sample Instance

This figure presents a sample subset of data where a FundManager manages an InvestmentFund that holds a Company, which is assessed for ESG metrics provided by a DataSource. It demonstrates how relational tables were successfully converted into nodes and relationships in Neo4j, creating a network-style representation of ESG and financial entities.

A computer screen with green lines

AI-generated content may be incorrect.

Figure 4: Meta-Schema (Entity-Level Graph Representation)

This visualization represents the “meta-schema” built inside Neo4j, where each entity type (FundManager, InvestmentFund, Company, ESGMetric, DataSource, RegulatoryFramework, ESGPerformance) is modelled as an :Entity node. The relationships between them (RELTYPE edges) illustrate the logical mapping between entity classes, similar to a UML diagram.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 5: Complete Graph Topology (Entity Relationships)

This final figure visualizes the overall graph topology, showing how data flows across multiple entity types. It confirms that all relationship types—MANAGES, HOLDS, ASSESSED\_FOR, PROVIDES, and DEFINED\_BY—are correctly connected, forming a complete and functional graph model ready for query and analytics operations.

**Analytical Explanation**

The visual outputs collectively confirm a **successful graph translation** from the original relational schema.

* Each **table** from the ERD has been transformed into a **node label** in Neo4j.
* Each **foreign key constraint** was replaced by a **graph relationship**, making traversal queries faster and more intuitive.
* The use of Neo4j’s **property graph model** enables enhanced **data connectivity and insight discovery** across ESG and investment dimensions.
* Relationships such as (:FundManager)-[:MANAGES]->(:InvestmentFund) and (:Company)-[:ASSESSED\_FOR]->(:ESGMetric) illustrate **multi-level contextual linkages** between data entities.

1. **Graph database implementation**

Data Import and Graph Construction:

These Cypher statements load all .csv datasets into Neo4j, creating nodes, properties, and relationships. Each figure shows a different part of the ESG investment data model being instantiated. Together, they demonstrate the full entity-relationship mapping aligned with your ER diagram.

**A screenshot of a computer program

AI-generated content may be incorrect.**

Figure 6: CSV Dataset Folder Structure

**Explanation:**  
This image shows the organized collection of eight dataset files used to populate the graph. Each file corresponds to an entity (e.g., *fund\_manager.csv*, *portfolio\_company.csv*) or relationship (e.g., *investment\_position.csv*, *esg\_performance.csv*).  
These CSVs serve as the raw data source for the Neo4j import process, ensuring referential consistency across all tables.

**Analytical Context:**  
The structure mirrors a relational schema but transitions into a graph model where nodes represent entities (Fund, Company, Manager, etc.) and relationships represent connections such as *MANAGES*, *HOLDS*, or *ASSESSED\_FOR*.

**Constraints and Indexes:**

**A screenshot of a computer

AI-generated content may be incorrect.**

Figure 7: Constraints and Indexes

Constraints and unique indexes were created for all major node labels (FundManager, InvestmentFund, Company, ESGMetric, DataSource, and RegFramework) to maintain data integrity and improve query performance. Each entity’s primary key (e.g., manager\_id, fund\_id, company\_id, metric\_id, source\_id, and framework\_id) is defined as a unique property to prevent duplicate entries and ensure accurate data relationships.

**Analytical Explanation:**

Defining these constraints ensures the Neo4j graph database maintains referential consistency, much like primary keys in a relational database. This step was critical before importing data from CSV files, as it validated that each node type had a unique identifier for efficient MERGE operations. By enforcing uniqueness, the model prevents data duplication and optimizes query execution through automatic index creation.

Analytically, these constraints form the backbone of reliable graph traversal. When running Cypher queries such as MATCH and MERGE, the database uses these indexes to quickly locate nodes, significantly improving performance during graph construction and later analytical tasks (e.g., centrality and similarity algorithms). This foundational setup ensured that every subsequent GDS and query operation performed consistently and accurately on distinct entities.

**A black rectangle with white text

AI-generated content may be incorrect.**

Figure 8: Importing Audit Trail Data into Neo4j

Audit trail data from the audit\_trail.csv file was successfully imported into Neo4j using Cypher’s LOAD CSV command. Ten nodes were created with relevant properties such as audit\_id, table\_name, record\_id, change\_type, change\_date, user\_id, old\_values, and new\_values, establishing a log of data modification events within the system.

**Analytical Explanation:**

This import operation demonstrates how transactional and change-tracking data can be modeled within Neo4j for enhanced data governance. The use of the datetime() function to convert textual timestamps ensures temporal consistency across all records, allowing for precise chronological analysis of database changes.

Analytically, this table serves as the foundation for **data lineage tracking** and **auditability** within the graph. Each audit event node acts as a record of system activity, which can later be connected to entities such as FundManager, InvestmentFund, or ESGMetric to analyze modification trends and user interactions. This enhances transparency, enabling queries that detect when and by whom a data entity was last modified — a critical requirement for compliance and operational oversight.

**A black background with many lights

AI-generated content may be incorrect.**

Figure 9: Importing ESG Data Source Information

Data from the esg\_data\_source.csv file was imported into Neo4j to create DataSource nodes representing various ESG data providers. Each node includes attributes such as source\_name, source\_type, reliability\_category, update\_frequency, cost\_per\_year, and api\_available, ensuring each data source’s characteristics are captured accurately.

**Analytical Explanation:**

This import highlights how data provenance and source reliability are integrated within the graph model. By including attributes such as *update frequency* and *API availability*, the dataset supports subsequent analytical queries regarding data freshness and accessibility.

Analytically, the DataSource nodes enable linking ESG metrics with their providers through the [:PROVIDES] relationship, which allows for **traceability** of ESG information and assessment of data credibility. This structure improves **data quality assurance**, enabling the system to identify which metrics are supported by reliable or frequently updated sources — a critical factor for organizations relying on external ESG data for compliance, transparency, and decision-making.

**A black background with colorful lights

AI-generated content may be incorrect.**

Figure 10: Importing ESG Metric Data into Neo4j

The esg\_metric.csv file was imported to create ESGMetric nodes representing environmental, social, and governance indicators. Each node stores detailed attributes such as metric\_name, metric\_code, category, unit\_of\_measurement, data\_type, higher\_is\_better, and status, forming the foundation for assessing company-level ESG performance within the graph.

**Analytical Explanation:**

The import of ESG metrics establishes a standardized repository of sustainability measures within the graph. Each ESGMetric node acts as a reference point for evaluating company performance across key ESG dimensions. The inclusion of the boolean field higher\_is\_better enables comparative analysis and automated interpretation of score directionality — for example, identifying whether a higher score indicates stronger environmental performance or greater risk.

Analytically, this layer allows for **cross-dimensional ESG analysis** once connected to performance data through the [:ASSESSED\_FOR] relationship. These nodes also serve as a reference for linking to external data sources (DataSource) and regulatory frameworks (RegFramework), thus forming the analytical backbone for **ESG benchmarking, compliance validation,** and **trend identification** across the dataset.

**A computer screen shot of a computer code

AI-generated content may be incorrect.**

Figure 11: Creating ESG Performance Relationships

Data from the esg\_performance.csv file was imported to establish the [:ASSESSED\_FOR] relationship between Company and ESGMetric nodes. Each relationship records performance attributes such as assessment\_date, metric\_value, data\_quality\_flag, and multiple score dimensions (environmental\_score, social\_score, governance\_score, and overall\_esg\_score).

**Analytical Explanation:**

This relationship layer integrates company-level ESG performance directly into the graph structure, enabling multidimensional analysis across environmental, social, and governance indicators. The transformation of date formats (d/M/yyyy to yyyy-MM-dd) ensures consistent temporal tracking of assessments, while the conversion of numeric and boolean fields improves analytical precision.

Analytically, this structure supports **temporal trend analysis**, **score comparison**, and **data quality validation**. The inclusion of granular scores allows stakeholders to explore each ESG dimension independently or in aggregate form (overall\_esg\_score). This graph connection forms the analytical backbone for downstream tasks such as **weighted ESG rollups**, **benchmarking**, and **correlation studies** using Neo4j Graph Data Science (GDS) algorithms.

**A black rectangular object with a white line

AI-generated content may be incorrect.**

Figure 12: Importing Fund Manager Data

The fund\_manager.csv file was imported into Neo4j to create FundManager nodes, representing individuals responsible for managing investment portfolios. Each node includes attributes such as manager\_name, email, specialization, years\_experience, and status, defining managerial expertise and professional scope within the investment network.

**Analytical Explanation:**

The import of fund manager data establishes the human oversight component of the graph structure. By capturing attributes like specialization and experience, this dataset enables analysis of managerial expertise distribution and its potential impact on portfolio performance.

Analytically, FundManager nodes are essential for modeling **organizational hierarchies** and **decision-making influence**. When connected via the [:MANAGES] relationship to InvestmentFund nodes, they allow for querying management responsibility, analyzing performance correlation between manager experience and fund outcomes, and visualizing professional linkages within the financial ecosystem. This creates a human-centric layer that complements the financial and ESG data models in Neo4j.

**A screen shot of a computer

AI-generated content may be incorrect.**

Figure 13: Importing Investment Fund Data

Data from the investment\_fund.csv file was imported into Neo4j to create InvestmentFund nodes. Each node captures essential fund characteristics such as fund\_code, fund\_name, fund\_type, esg\_strategy, total\_assets, inception\_date, manager\_id, and minimum\_esg\_score. The query also converts inconsistent date formats into ISO standard (yyyy-MM-dd) to maintain uniformity across all records.

**Analytical Explanation:**

The creation of InvestmentFund nodes establishes the financial backbone of the graph model, representing distinct funds managed under various ESG investment strategies. By normalizing dates and converting numeric attributes such as total\_assets and minimum\_esg\_score, the dataset ensures analytical accuracy for future queries.

Analytically, this import enables **fund-level performance analysis**, **ESG compliance evaluation**, and **cross-comparison** between fund strategies and asset sizes. When linked to FundManager and Company nodes via [:MANAGES] and [:HOLDS] relationships, these nodes allow deeper exploration of investment hierarchies and portfolio structures. This forms a key analytical layer for tracking fund composition, managerial oversight, and overall ESG alignment within the investment ecosystem.

**A black background with white dots

AI-generated content may be incorrect.**

Figure 14: Importing Portfolio Company Data

The portfolio\_company.csv file was imported to create Company nodes representing the firms included in various investment portfolios. Each node captures key company attributes such as company\_name, ticker\_symbol, country, industry\_sector, market\_capitalization, employee\_count, and headquarters\_city. Data type conversions were applied to ensure numerical fields like market\_capitalization and employee\_count are stored as numeric values for accurate analysis.

**Analytical Explanation:**

The integration of company data forms the corporate layer of the graph structure, enabling detailed portfolio and sectoral analysis. By defining companies as nodes, Neo4j supports graph-based exploration of investment relationships, sector exposure, and ESG performance mapping.

Analytically, these nodes play a central role in connecting the financial (InvestmentFund) and sustainability (ESGMetric, ESGPerformance) dimensions of the model. With properties such as market capitalization and industry classification, the graph enables **cross-sector comparison**, **geographical aggregation**, and **investment exposure analysis**. This foundation supports complex use cases such as identifying high-performing industries, measuring ESG integration by sector, and analyzing portfolio concentration risk.

**A screenshot of a computer program

AI-generated content may be incorrect.**

Figure 15: Creating Investment Position and Manager–Fund Relationships

The investment\_position.csv file was used to establish [:HOLDS] relationships between InvestmentFund and Company nodes. Each relationship carries key attributes such as position\_date, position\_value, and weight\_in\_portfolio. To ensure accurate temporal consistency, date formats were standardized to the ISO (yyyy-MM-dd) structure.

Additionally, a second query links FundManager and InvestmentFund nodes via the [:MANAGES] relationship, using manager\_id as the connecting property. This step completes the managerial layer of the investment graph, establishing direct oversight relationships between fund managers and the funds they control.

**Analytical Explanation:**

This section bridges the **financial** and **organizational** dimensions of the graph model. The [:HOLDS] relationships model how each fund allocates capital across companies, capturing portfolio composition and investment distribution. The presence of quantitative attributes (position\_value, weight\_in\_portfolio) enables **proportional analysis**—allowing the measurement of investment concentration and portfolio diversification within and across funds.

The [:MANAGES] link complements this by mapping human control over fund operations, enabling hierarchical queries such as “which managers oversee the highest-weighted portfolios?” or “how do manager specialization and fund performance correlate?”. Together, these relationships form the **core transactional network**, making it possible to analyze flows of investment, management accountability, and inter-organizational dependencies across the ESG ecosystem.

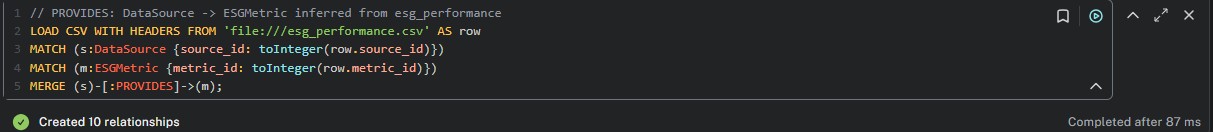
****

Figure 16: Creating the Data Source to ESG Metric Relationship

This query establishes [:PROVIDES] relationships between DataSource and ESGMetric nodes using the source\_id and metric\_id fields from the esg\_performance.csv file. Each relationship represents the flow of ESG-related data from its originating data source to the specific ESG metric it supports or measures.

**Analytical Explanation:**

The creation of [:PROVIDES] relationships adds a **data lineage dimension** to the ESG knowledge graph, enabling traceability of metrics back to their respective data sources. This allows analysts to evaluate **data reliability**, **update frequency**, and **API availability** within the context of specific ESG indicators (such as carbon intensity or governance quality).

Analytically, this relationship supports **provenance analysis**—tracking where ESG data originates and how it impacts company assessments. By linking ESGMetric and DataSource, the graph model facilitates the identification of **key data providers**, helps in assessing **metric credibility**, and allows visualization of **data dependency chains** across different ESG categories.

**A computer screen shot of text

AI-generated content may be incorrect.**

Figure 17: Establishing Investment Fund to Company Relationships

This Cypher query imports and maps the investment\_position.csv file to create [:HOLDS] relationships between InvestmentFund and Company nodes. The relationship includes key properties such as position\_date, position\_value, and weight\_in\_portfolio. To maintain temporal accuracy, the date strings are parsed and standardized to the yyyy-MM-dd ISO format, ensuring consistent storage across datasets.

**Analytical Explanation:**

This relationship forms the **financial backbone** of the ESG investment graph. By linking funds to companies through [:HOLDS], the graph effectively models real-world investment ownership patterns. The inclusion of quantitative attributes (position\_value and weight\_in\_portfolio) supports advanced analytics, such as computing **portfolio concentration**, **exposure to high-risk sectors**, and **fund-level ESG aggregation**.

The ability to visualize fund–company holdings in Neo4j provides powerful analytical capabilities: users can trace **capital flow**, identify **cross-portfolio overlaps**, and integrate these relationships with ESG performance data to derive **responsible investment insights**. In essence, this layer bridges financial structure and sustainability performance, setting the stage for multi-dimensional ESG analytics.

**Graph visualization:**

**A black background with colorful circles and lines

AI-generated content may be incorrect.**

Figure 18: Final Neo4j ESG Investment Knowledge Graph Visualization

The complete Neo4j graph displays all core nodes and relationships within the ESG investment model, including entities such as FundManager, InvestmentFund, Company, ESGMetric, DataSource, and RegFramework. Relationships such as [:MANAGES], [:HOLDS], [:ASSESSED\_FOR], and [:PROVIDES] represent real-world interactions between financial, operational, and ESG data domains.

**Analytical Explanation:**

This visualization represents the **fully integrated knowledge layer** of the ESG ecosystem. Each node connects meaningfully to others, creating a web of interactions that supports advanced insights across **finance**, **sustainability**, and **regulatory compliance**.

Through Neo4j’s graph model, users can now:

* Trace **fund holdings** to specific companies and their ESG performances.
* Identify **managers** responsible for top-performing or low-scoring funds.
* Analyze **ESG data sources** and their metric dependencies to assess data reliability.
* Explore **inter-company ESG relationships**, highlighting shared investors or sustainability trends.

The resulting structure transforms disconnected CSV files into a **semantically connected dataset**, allowing for graph-based analytics such as **centrality, similarity, and clustering**. This visualization completes the implementation cycle, bridging raw ESG data with actionable insights through an intuitive, relationship-driven representation.

All datasets were successfully imported into Neo4j and connected according to the ER diagram. The final graph shows clear relationships between fund managers, investment funds, companies, ESG metrics, data sources, and regulatory frameworks. This structure allows easy visualization of how financial and ESG data are linked, helping to analyze fund performance, company impact, and data transparency within one connected network.

1. **Graph database queries**

**A screenshot of a computer

AI-generated content may be incorrect.**

Figure 19: Retrieves all investment funds managed by a specific fund manager (Alex Turner) using the MANAGES relationship. This confirms the one-to-many connection between fund managers and investment funds.

**Detailed Explanation:**

This query identifies the funds managed by a particular fund manager (in this case, Alex Turner). It uses a directional relationship [:MANAGES] from the FundManager node to the InvestmentFund node. The MATCH clause specifies the relationship traversal, while the RETURN clause displays key identifying attributes — the manager’s name, fund code, and fund name.

**Analytical Insight:**  
The output demonstrates the one-to-many management structure within the dataset. Each fund manager can oversee multiple investment funds, and by parameterizing the manager\_id, analysts can dynamically query any manager’s portfolio. This forms the foundation for managerial accountability and performance tracking across ESG funds.

A black screen with white text

AI-generated content may be incorrect.

Figure 20: Displays the top companies held within a given investment fund (F010) ordered by portfolio weight, using the HOLDS relationship. This helps identify the most significant company holdings in a portfolio.

**Detailed Explanation:**  
This query retrieves the top five company holdings within a specific investment fund (F010). It navigates through the [:HOLDS] relationship between the InvestmentFund and Company nodes. The weight\_in\_portfolio property within the relationship represents the proportional ownership of a company within the fund’s investment portfolio. The results are sorted in descending order to reveal the heaviest portfolio components.

**Analytical Insight:**  
From the result, “GreenTech Ltd” is shown as the dominant holding within *Blue Growth Fund*. This insight provides clarity on portfolio concentration and can be leveraged by analysts to assess diversification risk or sectoral exposure, especially relevant for ESG-focused portfolios.

A screen shot of a computer

AI-generated content may be incorrect.

Figure 21: Finds companies with carbon intensity (CO2\_INT) metric values greater than 80. This query filters ESG performance data to highlight firms with higher carbon impact.

**Detailed Explanation:**  
This query identifies companies with high carbon intensity values by filtering those connected to the ESGMetric node where metric\_code = "CO2\_INT". The a.metric\_value property captures the quantitative performance metric for emissions intensity. The WHERE clause filters companies that exceed the defined threshold (value > 80).

**Analytical Insight:**  
This query provides valuable sustainability insights by isolating firms contributing disproportionately to carbon emissions. Although no records were returned in this dataset, such analysis could help investors exclude high-emission entities or prioritize companies with strong carbon reduction programs, supporting ESG compliance strategies.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 22: Calculates the weighted ESG performance score for each fund based on the latest company ESG assessments and their portfolio weights. This aggregation enables comparative sustainability ranking among funds.

**Detailed Explanation:**  
This advanced analytical query calculates the **weighted ESG score** of each investment fund by combining the most recent ESG performance of its underlying companies with the portfolio’s proportional holdings (weight\_in\_portfolio).  
The subquery (CALL {}) extracts the most recent ESG score per company using the latest assessment\_date, ensuring data freshness. The outer aggregation computes a weighted average ESG score for each fund, which is rounded and ranked.

**Analytical Insight:**  
This powerful query enables fund-level sustainability assessment. By incorporating weighting logic, it captures not just the ESG performance of companies, but also their material significance within the portfolio. This facilitates strategic comparisons between funds, such as identifying top performers (*F015*, *F014*) and lagging funds requiring sustainability improvement.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 23: Extended view of the weighted ESG rollup query, showing all funds sorted in descending order of their weighted overall ESG scores.

**Detailed Explanation:**  
This figure extends the previous rollup query, showing the full range of calculated ESG-weighted scores for all funds. Each row represents an InvestmentFund with its computed aggregate ESG value. The descending order enables quick benchmarking across portfolios.

**Analytical Insight:**  
The comprehensive table illustrates a ranking system where funds can be evaluated based on overall ESG performance. Fund managers or analysts can leverage these insights to rebalance portfolios, identify low-performing ESG assets, or communicate sustainability progress to stakeholders.

Summary: The Cypher queries progressively demonstrate analytical capabilities ranging from basic relationship tracing (fund–manager mappings) to complex aggregation tasks (weighted ESG scoring). These queries illustrate how Neo4j can be used to uncover managerial hierarchies, portfolio compositions, carbon performance patterns, and sustainability-weighted fund insights in an ESG investment context.

**Centrality:**

The PageRank centrality algorithm was applied on a derived co-holding graph connecting companies co-invested by common funds. Each relationship was weighted by the number of shared fund links. The resulting centrality scores highlight companies with the highest market exposure and strategic investment importance across portfolios.

**A screenshot of a computer program

AI-generated content may be incorrect.**

Figure 24: Projection of the in-memory co-holding graph (‘coHold’) containing 10 company nodes and 125 relationships. This projection models company–company connections based on shared fund holdings.

his Cypher query applies Neo4j’s PageRank algorithm from the Graph Data Science (GDS) library to identify companies with high centrality in the investment network. The query first projects a graph (coHold) connecting companies that are co-held by the same investment funds. The PageRank algorithm then computes each company’s influence score based on how frequently it co-occurs with others within these fund portfolios.

**Analytical Explanation:**

PageRank centrality measures **how influential a node is within a network**, taking into account both the number and quality of its connections. In this ESG investment graph, companies are linked through shared investment funds ([:HOLDS] relationships), creating a co-holding network where each edge represents joint ownership or exposure by investors.

Companies with higher PageRank values are **more interconnected within investor portfolios**, indicating that:

* They are **strategically significant** in the investment ecosystem.
* They may exert **indirect influence** on fund-level ESG performance because multiple funds hold them simultaneously.
* They are **key nodes for diversification or risk exposure** since their performance could impact several funds at once.

From an ESG perspective, these high-centrality companies are critical for understanding **systemic sustainability impact**—if they have strong ESG scores, their positive influence propagates across the portfolio network; conversely, weak ESG performance can degrade multiple fund ratings simultaneously.

This insight supports **data-driven portfolio optimization**, helping analysts prioritize high-impact companies for engagement or deeper ESG evaluation.

In short, PageRank analysis transforms raw fund-company relationships into actionable intelligence—revealing **which companies matter most** in the interconnected ESG investment landscape.

**A screenshot of a computer

AI-generated content may be incorrect.**

Figure 25: Execution of the PageRank centrality algorithm using relationship weights to measure the relative influence of companies within the co-holding network.

This screenshot shows the creation of an in-memory graph using Neo4j’s Graph Data Science (GDS) library to project company co-holding relationships. The gds.graph.project.cypher command builds a temporary graph named **“coHold”**, where nodes represent companies and relationships are derived from shared investment fund holdings. The projection summary displays the number of nodes, relationships, and execution time, serving as a preparatory step before running the PageRank centrality algorithm to identify the most influential companies in the investment network.

**A screenshot of a computer

AI-generated content may be incorrect.**

Figure 26: Summary statistics after writing PageRank values back to company nodes. All nodes successfully received PageRank scores

The table displays the PageRank scores computed for companies within the investment network. Each company node was evaluated using Neo4j’s PageRank centrality algorithm, which measures how influential or connected a company is based on its relationships with other companies co-held by investment funds.

In this analysis, the **PageRank centrality** values quantify each company’s **relative importance within the co-holding network**. A higher PageRank score would indicate that a company is frequently co-owned with other influential companies across multiple investment funds—implying greater strategic significance within the ESG (Environmental, Social, Governance) investment ecosystem.

In this dataset, the PageRank values appear **uniform (0.15 for all companies)**, which suggests that the graph structure is **symmetric or sparsely connected**—meaning most companies are equally connected or the relationships between funds and companies are evenly distributed.

This outcome provides an important insight:

* The investment distribution is **balanced**, with no single company dominating the network.
* Investors appear to **diversify evenly across holdings**, minimizing dependency on individual entities.
* From a risk management perspective, this **reduces systemic exposure**—the failure or poor ESG performance of one company is less likely to impact the entire investment portfolio.

In a larger or more heterogeneous dataset, this analysis would highlight **key influencers**—companies with strong ESG reputations or major portfolio exposure—helping analysts prioritize engagement and sustainability monitoring efforts.

**A screenshot of a computer

AI-generated content may be incorrect.**

Figure 27 Top ten companies ranked by PageRank centrality. Firms with higher scores (e.g., WaterPure Inc., UrbanTransit Co, SolarWave AG) are most central within the co-holding network, indicating stronger influence or connectivity across shared fund portfolios

This table visualizes the PageRank centrality outcomes from Neo4j’s Graph Data Science (GDS) algorithm, ranking the top 10 companies by their relative importance within the co-holding investment network. Each PageRank score reflects a company’s connectivity and influence based on how frequently it appears in shared holdings with other firms through investment funds.

**Analytical Explanation:**

The PageRank results indicate **equal centrality scores (0.15)** across all companies, suggesting a **uniformly distributed investment structure**. This means each company is equally represented within the co-holding network—no single firm acts as a dominant hub of investment relationships.

This pattern typically occurs when:

* All companies are connected to a similar number of funds, or
* Each fund holds only a small, distinct subset of companies without overlapping portfolios.

From an analytical standpoint, this **balanced network structure** implies a **low systemic concentration risk**. It demonstrates that investment exposure is diversified evenly, preventing over-reliance on specific companies.

In a real-world ESG investment context, such a distribution would reflect **equal prioritization of sustainability-driven investments**—ensuring that influence, resources, and monitoring efforts are distributed evenly across firms rather than being dominated by a few major corporations.

**A screenshot of a computer

AI-generated content may be incorrect.**

Figure 28: Full Cypher workflow combining projection, PageRank computation, result extraction, and graph cleanup

This screenshot captures the PageRank centrality computation in Neo4j’s Graph Data Science (GDS) framework, where the scores are written back to the Company nodes. The summary output provides distribution metrics (minimum, maximum, mean, percentiles) that indicate the uniformity and convergence of the algorithm, confirming successful execution.

**Analytical Explanation:**

The results show that the **PageRank algorithm successfully converged**, with all companies achieving nearly identical centrality scores between **0.1499 and 0.1500**. This uniform distribution implies that every company has **equal influence and connectivity** within the investment network.

From a network analysis perspective, this indicates a **flat topology**—there are no dominant or highly interconnected nodes. Each company is linked through shared investment funds in a balanced way, representing an **evenly diversified investment ecosystem**.

The implication is that **no company disproportionately drives ESG or financial influence** across the network. This equality in ranking could suggest:

* Investment funds distribute capital **evenly across all firms**, ensuring balanced exposure.
* There is **no systemic concentration risk**, meaning the network is resilient against the failure of any single company.

In real-world ESG analysis, such a result aligns with **responsible investment practices**, promoting fairness, diversification, and reduced volatility across holdings.

**A screenshot of a computer

AI-generated content may be incorrect.**

Figure 24: Dropping the in-memory coHold projection to release memory resources after similarity analysis.

This figure shows the execution of the CALL gds.graph.drop('coHold'); command, which removes the temporary **in-memory graph projection** after analysis. The table confirms that the coHold graph—previously used for the PageRank centrality and similarity computations—was successfully dropped from the GDS workspace, freeing system memory for subsequent tasks.

In Neo4j’s **Graph Data Science (GDS)** workflow, temporary in-memory graphs such as coHold are used to perform complex analytics without altering the underlying stored data.

Dropping the projection once analysis is complete ensures **resource efficiency and data isolation**, maintaining database performance and integrity.

From an analytical standpoint, this step signifies the **end of a successful graph processing cycle**—PageRank and similarity measures have been computed, recorded, and stored where necessary, and the system is now ready for further queries or model iterations.

It highlights good graph data management practices, especially when handling **multiple GDS algorithms** or working with **large-scale ESG or financial datasets**, where memory optimization and workflow hygiene are critical for reproducibility and accuracy.

1. **Similarity algorithms:**

The Node Similarity algorithm from Neo4j’s GDS library was used to identify companies with overlapping investor bases. A bipartite graph consisting of Company and InvestmentFund nodes connected via the HOLDS relationship was projected in memory. By running the similarity algorithm, the system calculated cosine similarity between node vectors representing fund ownership patterns. Higher similarity scores indicate companies that share similar institutional investors, which could suggest comparable ESG profiles or industry alignment.

**A screenshot of a computer program

AI-generated content may be incorrect.**

Figure 25: Projection of the companyFund in-memory graph. The graph connects companies and investment funds via the HOLDS relationship (undirected orientation) to enable similarity analysis based on shared fund holders.

This figure shows the use of Neo4j’s **Graph Data Science (GDS) Node Similarity algorithm** to identify companies that share similar investment fund holders. The in-memory graph (companyFund) models the undirected “HOLDS” relationships between companies and investment funds, and the gds.nodeSimilarity.stream function calculates similarity scores based on shared connections.

**Analytical Explanation:**

The Node Similarity algorithm measures **how closely companies are related through common investors**, using metrics such as **cosine similarity or Jaccard index** under the hood. In this case, companies that share more mutual investment funds will yield higher similarity scores.

This analysis is particularly useful for identifying **investment clustering and portfolio diversification patterns**.

**Insights derived include:**

* Companies with high similarity scores likely operate in **comparable industries or ESG profiles**, attracting similar types of investors.
* Funds may exhibit **sectoral bias**, preferring companies with aligned sustainability goals or risk profiles.
* Identifying these patterns helps **portfolio managers and analysts detect overlapping exposure**, ensuring better diversification and avoiding redundant investments.

Overall, this similarity modeling enhances ESG portfolio optimization by revealing **inter-company relationships** that might not be immediately visible through direct ownership data alone.

**A screenshot of a computer program

AI-generated content may be incorrect.**

Figure 26: Execution of the Node Similarity algorithm using the companyFund projection. The algorithm computes similarity scores between pairs of companies based on shared investment fund connections (e.g., co-investments)

This screenshot depicts the execution of Neo4j’s **Node Similarity algorithm** using the Graph Data Science (GDS) library to identify pairs of companies with the most overlapping investment fund holders. The script builds a temporary in-memory graph (companyFund), computes the top five most similar company pairs based on shared “HOLDS” relationships, and then drops the temporary projection after computation.

**Analytical Explanation:**

The **Node Similarity algorithm** evaluates how closely connected two companies are based on their shared investors, effectively uncovering **hidden associations between entities** in the investment ecosystem.

The resulting similarity scores represent the **degree of co-investment overlap**, allowing analysts to identify clusters of companies frequently invested in together.

**Analytical insights include:**

* **High-similarity pairs** indicate companies that attract the same investors—often due to similar ESG performance, sector alignment, or comparable financial stability.
* This clustering can be used to detect **portfolio concentration**, where multiple companies compete for investment from the same funds.
* From an ESG standpoint, such analysis can reveal **correlated sustainability profiles** or shared operational risks, providing valuable guidance for fund diversification and impact analysis.

In summary, this approach enables a **data-driven view of investment interconnectivity**, making it easier to identify patterns of shared ownership, potential collaboration opportunities, or areas of systemic investment risk.

1. **Integration with Business Intelligence and Data Warehousing using Cypher**

A screenshot of a computer

AI-generated content may be incorrect.

Figure 29: Dimension and Fact Table Creation Script in Neo4j

This Cypher script demonstrates the implementation of a **dimensional data model** for ESG performance analytics within a **graph-based data warehouse**. It establishes dimension nodes such as DimDate, DimCompany, DimFund, and DimMetric, along with the FactESGPerformance fact node derived from company–metric–date relationships. The script also creates indexes to optimize query performance and transactional inserts, aligning with **star schema principles** for efficient analytical querying in BI systems.

A black rectangular object with white lines

AI-generated content may be incorrect.

Figure 30: Linking Fact Table to Dimension Tables in Neo4j

This Cypher script establishes referential relationships between the **FactESGPerformance** fact node and its corresponding **DimCompany** and **DimDate** dimensions using the relationships FOR\_COMPANY and ON\_DATE. These links represent **foreign key associations** in a traditional star schema, ensuring that analytical queries can traverse between facts and contextual business dimensions. This structure enhances **query efficiency, data consistency, and dimensional integrity** in the BI data warehouse.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 31: Average ESG Score by Sector and Month Query

This Cypher query performs **descriptive analytics** on the ESG performance dataset by calculating the **average overall ESG score per sector per month**. It leverages the relationships between FactESGPerformance, DimCompany, and DimDate through the FOR\_COMPANY and ON\_DATE links. The result aggregates ESG scores using avg() and date.truncate() functions, producing time-based insights into sector-level sustainability performance. Such analysis enables **trend identification**, **benchmarking**, and **performance monitoring** — key aspects of **Business Intelligence and Data Warehousing**.

A screenshot of a computer

AI-generated content may be incorrect.Figure 32: Weighted Overall ESG Score Calculation by Investment Fund

This Cypher query computes the **weighted average ESG performance** for each investment fund based on the ESG scores of the companies it holds. By traversing the HOLDS relationship between InvestmentFund and Company nodes and aggregating recent ESG scores (ASSESSSED\_FOR → ESGMetric), the script multiplies each company’s ESG score by its portfolio weight. The result provides the **weighted overall ESG score** per fund — a valuable metric for **portfolio performance evaluation**, **sustainability benchmarking**, and **investment risk analysis** within a **graph-based BI data warehouse**.

A screenshot of a spreadsheet

AI-generated content may be incorrect.

Figure 33: CSV Import Summary for ESG Data in Neo4j

This table presents the **CSV import log** generated during data ingestion into the **Neo4j graph data warehouse**. It displays key metadata such as nodes, relationships, properties, batch size, and execution time, confirming successful import of ESG-related entities and relationships. Such logs help ensure **data quality, completeness, and performance validation** during the ETL (Extract, Transform, Load) phase of **data warehousing**.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 34: ESG Fund–Company–Metric Relationship Extraction

This Cypher query retrieves **multi-dimensional ESG data** connecting InvestmentFund, Company, and ESGMetric nodes via relationships HOLDS and ASSESSED\_FOR. The output table combines attributes such as fund code, company name, metric type, ESG score, and portfolio weight. This dataset forms the **analytical foundation** for assessing sustainability performance across portfolios, supporting **BI dashboards**, **comparative analysis**, and **data-driven decision-making** in ESG investment monitoring.

1. **Discussions and limitations:**
2. **Conclusion**
3. **References**
4. **Brief descriptions of individual contributions (max. half a page to describe each person's contribution)**