

ESG Data Ingestion Agent - Objective

Objective

Objective

The goal of the ESG Data Ingestion Agent is to create a modular, automated pipeline that can:

1. Ingest and Process Diverse Document Types

- Accept input files in PDF, Excel, and Image formats
- Automatically detect the file type and apply the appropriate parsing method

2. Extract Relevant ESG Data

- Identify and extract key Environmental, Social, and Governance (ESG) metrics such as:
 - Scope 1, 2, and 3 emissions
 - Energy consumption
 - Water usage
 - Other sustainability-related indicators

3. Map Extracted Data to ESG Metrics

- Use keyword-based matching to tag raw information with standardized ESG metric labels

4. Validate and Deduplicate Data

- Ensure completeness of critical ESG fields
- Remove duplicate entries to maintain data integrity

ESG Data Ingestion Agent - Objective

5. Format and Store Structured Data

- Convert processed data to JSON
- Store in MongoDB collections for persistence and downstream consumption
- Log processing steps and any encountered issues for traceability

6. Enable Easy Data Retrieval

- Provide query interfaces to search stored ESG data via:
 - Direct MongoDB queries
 - Keyword-based CLI search

This objective ensures that ESG data-whether hidden in scanned reports, spreadsheets, or raw documents-can be automatically extracted, cleaned, structured, and made queryable for analytics, reporting, or compliance.