**Understanding the project requirements**

* **Contextualizing AI in finance**: This task immerses you in the real-world application of AI in finance. By extracting and analyzing data from 10-K documents, you'll understand how AI can transform raw financial data into insightful analytics.
* **Identifying key financial indicators**: The ability to discern which data points are critical for financial assessment is fundamental. This task will enhance your acumen in recognizing significant financial metrics crucial for AI analysis.

**Determining important factors for AI integration**

* **Data quality assessment**: The success of AI heavily depends on the quality of data it is fed. Through this task, you'll learn to identify and extract high-quality, relevant financial data, setting a strong foundation for accurate AI modeling.
* **Understanding data structure**: AI models require data in specific formats. This task will help you comprehend the structuring of financial data, which is a pivotal step in preparing it for AI integration.

**Project context:**

* Our goal is to extract meaningful insights from 10-K financial reports.
* These insights will feed into the AI chatbot, enabling it to provide in-depth financial performance analysis.

**Your role and responsibilities**

**Data extraction**:

* Research and review 10-K documents.
* Focus on key financial figures and ratios.

**Basic analysis**:

* Identify significant financial trends and indicators.
* Assess the financial health and performance of the companies.

**Data preparation**:

* Format and clean the data for AI model integration.

**Deliverable**:

* A comprehensive data analysis report, which should include:
  + your findings
  + a summary providing insights into the financial health of the analyzed companies.

10-K reports are comprehensive annual reports filed with the SEC by publicly traded companies. They provide a detailed account of a company's financial performance, including audited financial statements, management's discussion and analysis (MD&A), and disclosures about market risk, controls, and legal proceedings.

**Key sections to focus on for financial data extraction:**

* **Income statement**: This section provides information about the company's revenue, costs, and expenses over a specific period of time.
  + **Key data points**: Total revenue, cost of goods sold (COGS), operating expenses, and net income.
  + **Extraction technique**: Look for the income statement summary, typically in the early pages of the reports. Pay attention to year-over-year changes.
* **Balance sheet**: This section outlines the company’s assets, liabilities, and the shareholders’ equity at a specific point in time.
  + **Key data points**: Current assets, long-term assets, current liabilities, long-term liabilities, and total shareholders’ equity.
  + **Extraction technique**: Focus on the balance sheet summary. Compare assets against liabilities to understand the company’s financial health and note any large changes in assets or liabilities.
* **Cash flow statement**: This shows how changes to the balance sheet and income impact cash and cash equivalents.
  + **Key data points**: Cash from operating activities, investing activities, and financing activities.
  + **Extraction technique**: Analyze the cash flow statement to understand how the company generates and spends its cash. This can provide insights into a company's liquidity.

**Effective techniques for data extraction:**

* **Manual extraction**: Start by manually reviewing the documents to understand their layout and where key information is typically found.
* **Highlight and annotate**: Use digital tools to highlight and annotate key figures and notes for easy reference.
* **Excel and spreadsheet tools**: For quantitative data, using Excel or similar spreadsheet tools can be effective. You can input key figures into a spreadsheet for analysis and comparison.
* **Automated extraction tools**: For more advanced users, tools such as Python (in particular, libraries such as Beautiful Soup or Pandas) can automate the extraction of data from these documents, especially if they are available in digital formats.

**Data preparation steps:**

* **Data cleaning**: Involves correcting or removing incorrect, corrupted, or duplicate data.
  + Techniques include filling in missing values, smoothing noisy data, and resolving inconsistencies.
* **Data transformation**: This step is about normalizing and standardizing data to ensure it's in a usable format for AI models.
  + Includes converting all financial figures to a consistent format (e.g., all figures in thousands or millions) and adjusting for inflation or currency changes where necessary.

**Preprocessing for AI models:**

* **Feature engineering**: The process of using domain knowledge to create features that make machine learning algorithms work. In financial data, this might involve creating ratios or deriving financial health indicators from raw data.
* **Data encoding and formatting**: Many AI models require data in a specific format. This may involve encoding categorical data (like fiscal quarters) into numerical values or restructuring data sets for time-series analysis.
* **Dealing with time-series data**: Financial data often involves time-series analysis. Special care should be taken to handle trends and seasonality and potentially integrate lag features that capture past values.

Step 1: Data extraction

**Navigate to the SEC's EDGAR database**:

* [**Microsoft**](https://www.sec.gov/edgar/browse/?CIK=789019&owner=exclude)
* [**Tesla**](https://www.sec.gov/edgar/browse/?CIK=1318605&owner=exclude)
* [**Apple**](https://www.sec.gov/edgar/browse/?CIK=320193&owner=exclude)

**Manual extraction**:

* For each company, find the 10-K filings for the last three fiscal years.
* Extract the following financial figures: Total Revenue, Net Income, Total Assets, Total Liabilities, and Cash Flow from Operating Activities.

**Organize Your Data**:

* Compile the extracted data into an Excel spreadsheet for easy reference during your Python analysis.