Standard DataFeed User Guide

May 2025

Table of Contents

1 Introduction

This guide provides an overview of the research databases available to professors and students at ESCP Business School, along with instructions for requesting access and retrieving data. It includes a wide range of data types and offers key details such as coverage, time periods, and access platforms.

The databases are organized into two main sections:

- **Section 1: Subscribed Databases**
- **Section 2: Free and Open Access Databases**

2 Section I: Subscribed Databases

- **Bloomberg**
- **Eurofidai**
- **FactSet**
- **LSEG**
- **RavenPack**
- **WRDS**
    - **BoardEx**
    - **Compustat**
    - **CRSP**
    - **CRSP/Compustat Merged (CCM)**
    - **I/B/E/S**
    - **ISS ESG**

2.1 Bloomberg

2.1.1 Content

- Real-time and historical stock market data (prices, volumes, indices)
- Bond, foreign exchange (FX), commodities, and derivatives data
- Economic indicators (GDP, debt levels, employment figures, etc.)
- Corporate finance data for listed companies (financial statements, ratios, ownership structures)
- Up-to-date financial and economic news from global sources

2.1.2 Common Uses

- Return and performance analysis
- Beta and risk analysis
- Company valuation (Discounted Cash Flow (DCF) models, comparables)
- Analysis of index constituents and sector compositions
- Access to real-time news, economic reports, and financial forecasts

2.1.3 Access

- Available in the Trading Rooms in Paris, London and Turin — login credentials can be created immediately onsite

2.1.4 Introduction to Bloomberg Terminal

The Bloomberg Terminal is a powerful software platform used by finance professionals around the world. It provides real-time data, news, analytics, and trading tools across global financial markets. With Bloomberg, users can monitor market movements, analyze financial securities, build investment strategies, and execute trades, all in one place.

The Terminal is widely used in investment banking, asset management, trading floors, and research departments because it offers unparalleled access to data and decision-making tools. It covers stocks, bonds, currencies, commodities, derivatives, economic indicators, and much more.

Why use Bloomberg?

- To access real-time market data and breaking financial news.
- To perform deep analysis using advanced charting, financial models, and valuation tools.
- To connect with other finance professionals through Bloomberg's messaging system.
- To place and manage trades efficiently and securely.

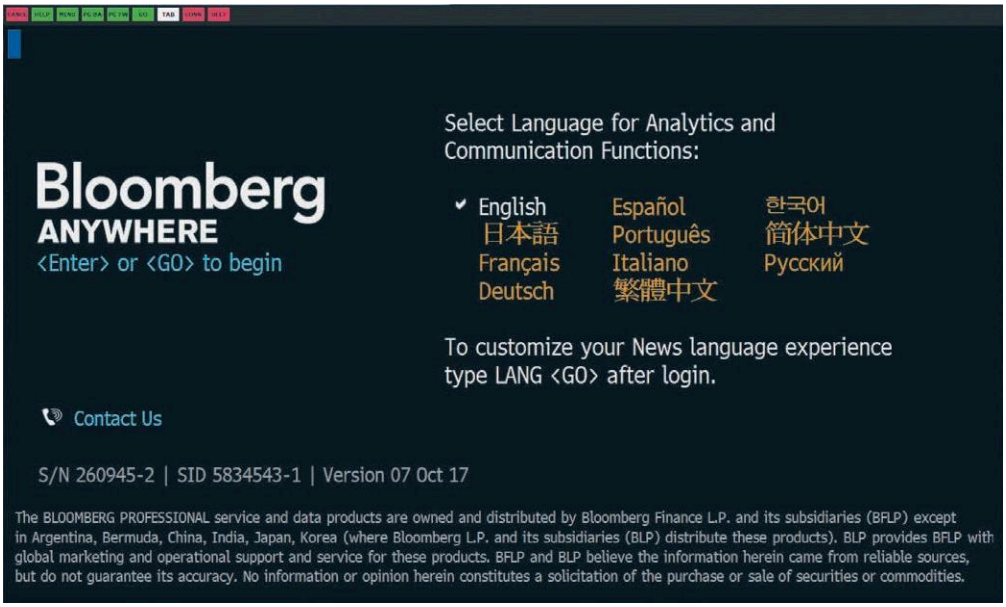
When to use Bloomberg?

Students/teachers can use Bloomberg when they need reliable market information for assignments, simulations, or research papers. It is especially valuable for trading simulations, portfolio management exercises, financial analysis, and preparing for careers in finance.

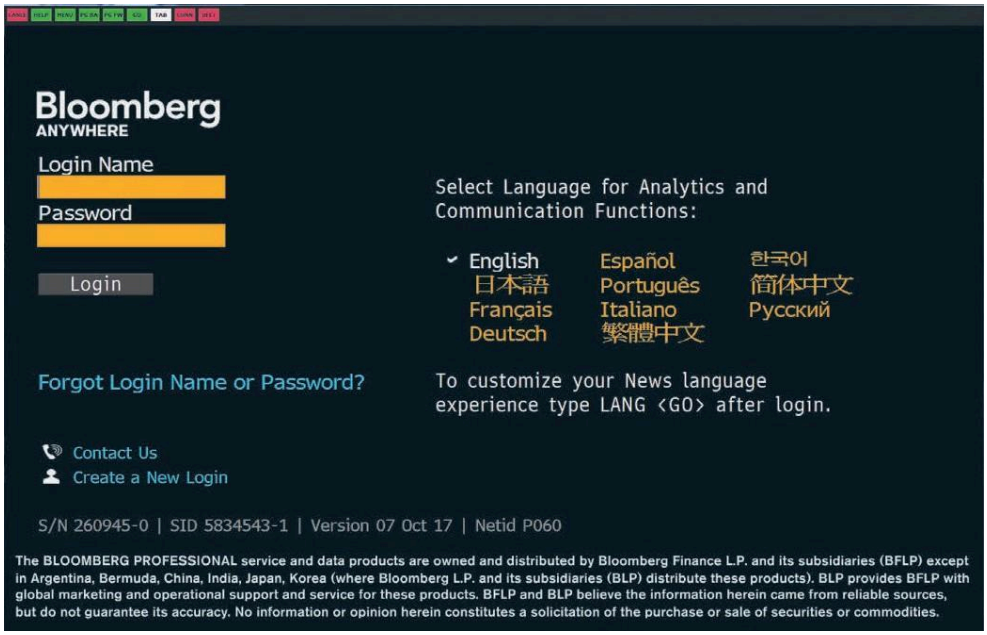
2.1.5 The Bloomberg Terminal

Accessing the application

The Bloomberg Terminal delivers news, data, and analytics to your desktop. You can access the application by clicking on this application in the Bloomberg Terminal:



Once you open the Bloomberg Terminal application, you must log in with a login name and



password.

2.1.6 The Bloomberg keyboard

The red stop keys, green action keys and yellow market sector keys help you access information quickly and easily.

Helpful keys

The Escape or Cancel key allows you to exit the current function and cancels the current activity on the screen.

Click on the Help button once to access a help page; click on it twice to access the Help Desk.

Enables keyword search of the entire Bloomberg database.

The yellow market sector keys enable you to:

- Load securities. **Example —** IBM US <EQUITY> <GO>
- Access market sector menus. **Example —** <CORP> <GO>

The Menu key opens a menu of related functions.

The <GO> or Enter key executes the command typed in the command line.

2.1.7 The Bloomberg panels

When you first log in to Bloomberg, up to four Bloomberg panels appear. The panels are independent workspaces that enable you to multi-task within the Bloomberg system. You can move from one panel to another using the blue <PANEL> key on the keyboard or by clicking on the

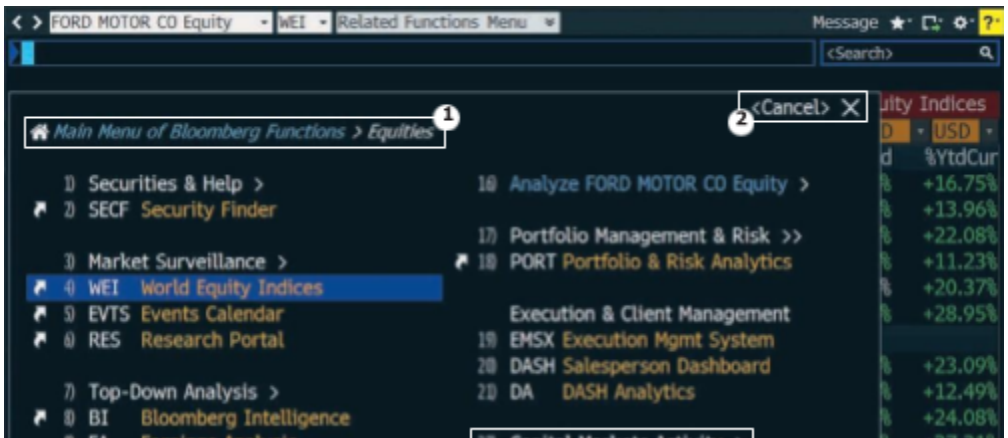
The four Bloomberg panels enable you to work with multiple functions simultaneously. As shown above, each panel is divided into three main sections:

1. Toolbar — The left side of the toolbar includes the menu tab and a drop-down list of recently loaded securities, with the current loaded security visible. The right side features icons to help you perform key tasks, including exporting data, viewing favorite places and securities, accessing Help, and adjusting your defaults and display.

2. Command line — Here you enter commands for functions and securities. You can also perform a keyword search for securities and functions from the command line. This autocomplete feature makes the Bloomberg Terminal entirely discoverable from the command line.

3. Function area — Here, you will see the actual function content displayed.

2.1.8 Navigation: General



1. Menu Breadcrumbs
2. Cancel
3. Category/Heading navigation

Menu layout

Bloomberg menus are intuitively organized to speed your search efforts.

Menu breadcrumbs

These show your path in the overall menu hierarchy and enable you to navigate backward and forward.

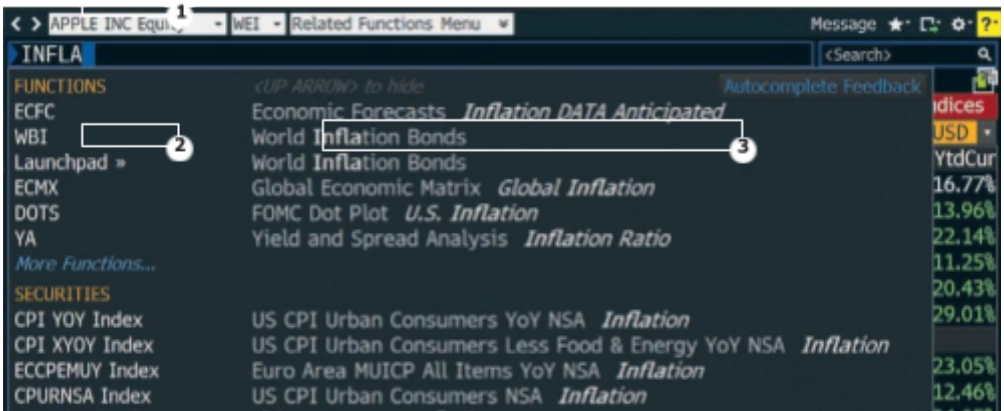
<CANCEL> X

Click <Cancel> X located in the upper right-hand corner to close the menu.

Categories & functions

Menus organize functions under categories. For category headings followed by ">", click the category to see the next menu in the hierarchy.

2.1.9 Navigation: Working with securities



1. Keyword in the Command line.
2. Function mnemonic.
3. Function title.

About securities

Securities are financial instruments, like stocks and bonds, that you can analyze with Bloomberg functions. Once you have loaded a security on a panel, it appears in the loaded security field on the panel's toolbar. You can run a series of functions to analyze the loaded security.

Loading securities

There are three main ways to load security, depending on whether you know the security's ticker symbol or identifier. If you know the ticker symbol or identification number for the security you want to load:

- Enter the ticker symbol/identifier in the command line.
- Press the yellow market sector key corresponding to the security type (Corp, Muni, Equity, etc.).
- Press the <GO> key.

Example — Using Ford Motor Company, enter F <EQUITY> <GO>.

Or, using the CUSIP for Wal-Mart Stores Inc., enter 931142DD2 <CORP> <GO>.

2.1.10 Navigation: Analyzing a company

1. The DES function analyzes the loaded security.
2. DES is the mnemonic for the Description function.

Step 1

Load the company you want to analyze. For Example — Enter F US <EQUITY> <GO>.

The company appears in the panel's toolbar as the loaded security.

Step 2

Run the analysis function on the loaded security in one of the following ways:

•Click on a category or function from the menu window — or type the mnemonic in the command field — to explore the full range of analysis options.

•Enter the mnemonic for the specific function you want to run, then press <GO>.

The analysis function runs on the loaded security.

Example — Enter DES <GO>.

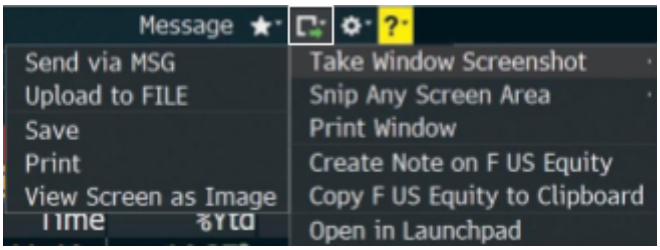
2.1.11 Exporting data

Drag & drop

Some functions display a drag and drop icon in the top right corner of the screen. You can click on and drag this icon to move securities from the current screen into another application, such as a Microsoft® Excel spreadsheet or a Bloomberg Wizard as part of the Bloomberg Excel Add-in. Once in a wizard or a spreadsheet, the tickers appear with Bloomberg market identifiers.

Printing & other export options

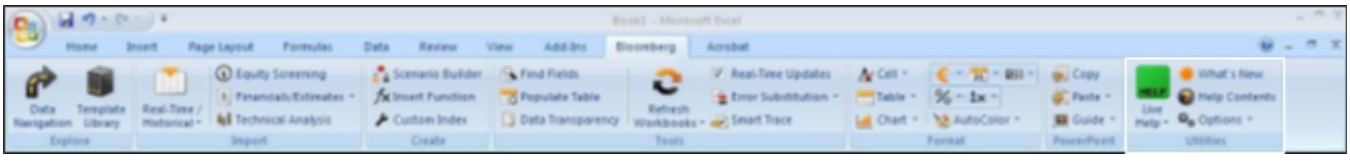
There are various ways you can save or export screenshots from Bloomberg. To display a list of export options, click on the Export icon at the top right of a panel's toolbar.



The Bloomberg Excel Add-in

The Bloomberg Excel Add-in is a powerful tool that delivers Bloomberg data into a Microsoft Excel spreadsheet for custom analysis and calculations. All data must remain on a licensed Bloomberg workstation.

On a computer where the Bloomberg software is active, you can access the Bloomberg Excel Add-in from the Excel taskbar.



Bloomberg Data Wizard

The easiest way to download data to Excel via the Bloomberg Excel Add-in is to use one of the Bloomberg Data Wizard tools. The wizards provide a guided process to draw data from Bloomberg into a spreadsheet.

Step 1

On the Bloomberg tab of the Excel spreadsheet, in the Import group, choose from the following options:

To download most forms of data, click the Real-Time/ Historical icon.

The Bloomberg Data Wizard window appears with four data type options.

•To download values from financial statements and/or earnings data, click the Financials/Estimates icon.

The Bloomberg Fundamental Analysis Wizard window appears with two data type options.

Step 2

Move your mouse over each of the data type icons to display a blurb describing the type of data that is available with each Wizard option.

Step 3

Click on the icon representing the type of data you want to download. Step 1 of the Wizard appears.

Step 4

Follow the instructions that appear in the Wizard window to build your data set.

Once you have completed all of the steps in the Wizard, the spreadsheet updates with the data you requested.

2.1.12 Useful Functions

Charting & Graphs

GP: Chart securities and technical studies on the Bloomberg Terminal.

HS: Visualize and compare the performance of two securities.

GC: Chart yield curves and see interest rate movements over time.

GF: Visual analysis of a company's fundamentals.

ECWB: Standardize and manipulate economic data.

Commodities

BGAS: See pricing for the North American natural gas spot market.

BOIL: View global oil & refined products spot pricing.

CPF: Analyze expert forecasts for future commodity prices.

FDM: Access comprehensive fundamental commodities data.

GLCO: Monitor price movement and performance of global commodities.

NGAS: Analyze natural gas statistics for North America.

WETR: Analyze current, historical and forecasted weather trends.

Company Analysis

DES: Access financial data and company background info.

BI: Industry analyses from Bloomberg Intelligence.

BIP: Primers on companies, industries, or topics.

BICO: Company primer to quickly educate yourself.

QUIC: QuickTakes to contextualize important news events.

ANR: Analyst recommendations and price targets for equities.

Currencies

FXFC: Analyze currency price forecasts by contributors.

FXCA: Currency conversion calculator.

FXC: Matrix of currency rates.

WCRS: Compare current and historical currency rates visually.

Derivatives

OMON: Real-time pricing and data for options.

OVME: Price and back-test equity derivative strategies.

OVML: Structure and price multi-leg FX option strategies.

Economics

FED: Monitor US Federal Reserve activities.

CENB: Portals to international central bank resources.

ECO: Economic calendars for industry and bank releases.

ECST: Monitor economic data from multiple sources.

ECFC: Economic forecasts and regional comparisons.

Equities

EQS: Screen companies to validate trade ideas.

FA: Company financial statements and fundamental data.

IPO: Monitor equity offerings and financing strategies.

CAST: Debtor organization and capital structure.

CACS: Calendar of corporate actions and events.

WEI: Global equities market surveillance.

RV: Compare a stock to its peers on key metrics.

EQRV: Evaluate if a stock is fairly valued vs peers.

CM: Track company stock price events.

Excel Tools

XLTP: Excel templates library for data analysis.

DAPI: Retrieve data into Excel with the Bloomberg API.

FLDS: Find Excel fields for formula use.

Fixed Income

SRCH: Search fixed income securities for trends and ideas.

DDIS: Maturity distribution of an issuer's debt.

YAS: Price fixed-income securities and calculate yields.

BTMM: Assess a country's interest rate environment.

ALLQ: Real-time corporate bond prices from dealers.

DEBT: See which holders are exposed to a country's debt.

WB: Monitor sovereign bond yields and performance.

CRPR: Assess the creditworthiness of an issuer/security.

CRVF: Search for relevant fixed income curves.

SRSK: Estimate 1-year default risk and 5-year CDS spread.

Mergers & Acquisitions

BUYP: View M&A buyer profiles for asset acquirers.

MA: Track M&A and arbitrage spread data.

MARB: Monitor real-time merger arbitrage spreads.

MRGC: Evaluate hypothetical merger scenarios.

PE: Access PE fundraising, screening, and market tools

News

TOP: Day's top global news stories.

TWTR: Search Twitter news on Bloomberg.

CN: Top news on a specific company.

FIRS: Summarized stories of market-moving news.

SALT: Set news email alerts for tracked companies.

BRIE: Read Bloomberg newsletters by topic.

Portfolio Management

PRTU: Create and manage portfolios for analysis.

PORT: Analyze portfolio performance and exposure.

EQBT: Back-test fundamental investment strategies.

UNCL: Custom classifications for portfolios/benchmarks.

BBU: Upload portfolios and custom data to PORT.

CDE: Create custom data fields across Bloomberg.

Training & Navigation

BMC: Get Bloomberg certified and learn more about the financial markets.

BHL: Visit the Bloomberg Help and Learning Center.

BPS: Locate topic-specific cheatsheets and videos.

BU: Search for, enroll in and launch webinars and training resources.

2.2 Eurofidai

2.2.1 Content

EUROFIDAI (CNRS/ESSEC) Daily Database provides:

- Verified and controlled data on stocks, mutual funds, indices, exchange rates and corporate events
- Historical data on financial markets all over Europe, Asia, Oceania, the Middle East, Africa and South America
- Stock market data going back to 1977 for France, 1980 for Europe, 1986 for Asia, the Middle East and Oceania, 1988 for Africa and 1998 for South America
- Data on mutual fund traded over-the-counter or in organized markets worldwide since 1975
- Spot exchange rates since 1975

2.2.2 Access

Send an email to Chantal Gueudar-Delahaye (gueudar-delahaye@escp.eu)

2.2.3 User guide

- Benchmark indices: https://drive.google.com/file/d/1J79uY0suCvbypmpCG0rqkSOlOooaCDK6/view?usp=sharing
- Corporate events: https://docs.google.com/spreadsheets/d/167STTUz7x2nBCkjVTHHbcXiHpTEZk5KD/edit?usp=sharing&ouid=115337863328258596010&rtpof=true&sd=true
- Global/Market Indices: https://drive.google.com/file/d/1XfVfpGlLI56naptUe9z-QydcQ-Os2bds/view?usp=sharing
- Mutual Funds: https://drive.google.com/file/d/1E0Cw2oVnw0JH4xjlKPTjZEtgGmtzft1q/view?usp=sharing
- Spot exchange rates: https://drive.google.com/file/d/1WPb77qMhEWLenJGJH5L45bFkUQ6gROsM/view?usp=sharing
- Stocks: https://drive.google.com/file/d/109oCLohgOCudXaRRI2liboRZiXr2QGAR/view?usp=sharing

2.3 FactSet

FactSet data for ESCP users consists of FactSet's SharkRepellent database and the FactSet Ownership database.

2.3.1 FactSet Ownership

2.3.1.1 Content

The FactSet Ownership database collects global equity ownership data for institutions, funds, and non-institutional "insider/stake holders". FactSet also collects fund holders of fixed-income securities. Ownership data is published as-reported when possible. This includes fund holdings, 13F holdings, pricing and shares outstanding data, as well as some insider transactions data. In other cases, data is provided adjusted for corporate actions. Adjustments are made in a monthly update where symbol changes are processed, and historical pricing and shares outstanding data are adjusted for all related corporate actions that have taken place during the period specified.

2.3.1.2 Coverage

The exact time period varies by dataset. Typically, equity ownership history is available from 1999 onward, while data for fixed income security holders extends through September 2013. The FactSet Ownership database includes both American and global coverage.

2.3.1.3 Access

Access is provided through the Nuvolos[1] platform.

Factset ownership Access via ESCP WRDS Account:

- Own_ent_13f_combined_inst: This table contains those institutions that FactSet LionShares has designated aggregate management entities for specific 13F filers.
- Own_ent_coverage: This table shows entity-level coverage information related to ownership.
- Own_ent_inst_identifiers: This table contains mappings from Central Registration Depository (CRD) identifiers maintained by FINRA to FactSet's FactSet_Entity_ID for institutions.
- Own_inst_13f_detail: This table contains historical institutional holdings of equity securities obtained from 13F filings. Quarterly holdings for active and terminated 13F filers are included.
- Own_sec_coverage: This table contains security level data for securities covered by the ownership database. Active and terminated securities are included.
- Own_sec_entity: This table contains FactSet Permanent Identifier (FSYM_ID) to FactSet Entity Identifier (FactSet_Entity_ID) mappings for those securities included in the FactSet Symbology universe. These mappings are provided at the security level.
- Own_sec_map: This table maps securities to the FSYM_ID the Ownership data.

2.3.1.4 Identifiers / linking to other products

- FactSet uses unique identifiers to facilitate consistent referencing across its datasets. The FACTSET_ENTITY_ID is a unique, FactSet-generated identifier representing a

[1] For the moment, the ESCP FactSet access is limited for ESCP professors through Nuvolos by requesting access via contacting daniel.sali@nuvolos.intercom-mail.com

company or issuer entity. The FSYM_ID is a unique identifier assigned to individual securities.

- Mappings between FSYM_IDs and FACTSET_ENTITY_IDs can be found in the SHRK_SEC_ENTITY table.
- Use the own_sec_entity table to map FactSet Permanent Identifier (FSYM_ID) to FactSet Entity Identifier (FactSet_Entity_ID) mappings for those securities included in the FactSet Symbology universe.
- To convert these identifiers into ISIN and CUSIP identifiers, use the sym_isin and sym_cusip mapping tables.

2.3.1.5 Navigation path

Nuvolos/ESCP Business School/ FactSet Ownership/Master/Current State

2.3.1.6 Retrieving the Data

For instructions on retrieving the data, see the section "How to use Nuvolos to retrieve the data".

2.3.2 FactSet SharkRepellent

2.3.2.1 Content

SharkRepellent provides comprehensive data on shareholder activism, takeover defenses, and corporate governance for thousands of U.S.-incorporated companies.

FactSet collects activism-related information from a wide range of public sources, including:

- Company filings (e.g., proxy statements),
- Dissident investor filings (e.g., Schedule 13D, alternative proxy filings),
- Press releases,
- Financial news outlets,
- Company websites, and
- Financial trade publications

Poison pill activity is tracked from EDGAR and SEDAR filings, as well as press releases. Corporate governance data is similarly sourced from publicly available documents, including SEC and SEDAR filings, news sources, and company communications.

2.3.2.2 Coverage

FactSet's SharkRepellent database offers extensive historical and current coverage on shareholder activism, poison pill activity, and corporate governance measures, primarily for U.S.-incorporated public companies.

- Poison Pill Coverage:
    - Comprehensive tracking of all poison pill activity since 2001, including original adoptions, amendments, and replacements.
    - Additional data is available before 2001 for select companies.
    - Coverage includes all in-force U.S. poison pills, Canadian poison pills (in force or pending ratification), and non-U.S. filers (e.g., companies incorporated in Bermuda or the Cayman Islands) that adopt a U.S.-style poison pill.
- Activism Coverage:
    - All corporate activism targeting U.S.-incorporated companies since 2006, with select campaign data from 2005.
    - Full coverage of all proxy fights since 2001.
    - Historical campaigns by key activist investors (e.g., the SharkWatch50) even prior to 2006.
- Corporate Governance & Takeover Defenses:
    - Data sourced from articles of incorporation, bylaws, and other publicly available materials. Includes approximately 7,000 U.S.-incorporated public companies, as well as companies traded in the U.S. but incorporated in Caribbean jurisdictions. FactSet also tracks significant activism events disclosed through SEC filings and news sources, making the dataset both deep and up to date.

2.3.2.3 Access

Access is provided through the Nuvolos[2] Platform.

Factset SharkRepellent Access via ESCP WRDS Account:

The whole Factset SharkRepellent Package consists of data on shareholder activism, takeover defenses, and corporate governance

2.3.2.4 Identifiers / linking to other products

- FactSet uses unique identifiers to facilitate consistent referencing across its datasets. The FACTSET_ENTITY_ID is a unique, FactSet-generated identifier representing a company or issuer entity. The FSYM_ID is a unique identifier assigned to individual securities.
- Mappings between FSYM_IDs and FACTSET_ENTITY_IDs can be found in the SHRK_SEC_ENTITY table.
- Use the own_sec_entity table to map FactSet Permanent Identifier (FSYM_ID) to FactSet Entity Identifier (FactSet_Entity_ID) mappings for those securities included in the FactSet Symbology universe.
- To convert these identifiers into ISIN and CUSIP identifiers, use the sym_isin and sym_cusip mapping tables.

2.3.2.5 Navigation path

Nuvolos/ESCP Business School/ FactSet Ownership/Master/Current State

2.3.2.6 Retrieving the Data

For instructions on retrieving the data, see the section "How to use Nuvolos to retrieve the data".

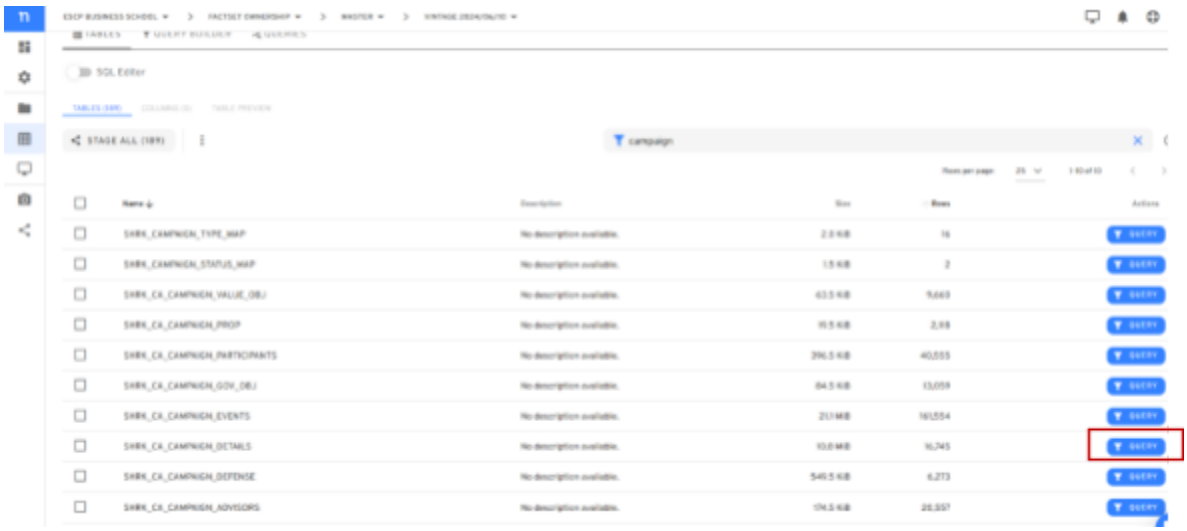
[2] For the moment, the ESCP FactSet access is limited for ESCP professors through Nuvolos by requesting access via contacting daniel.sali@nuvolos.intercom-mail.com

2.3.2.7 How to use Nuvolos to retrieve the data

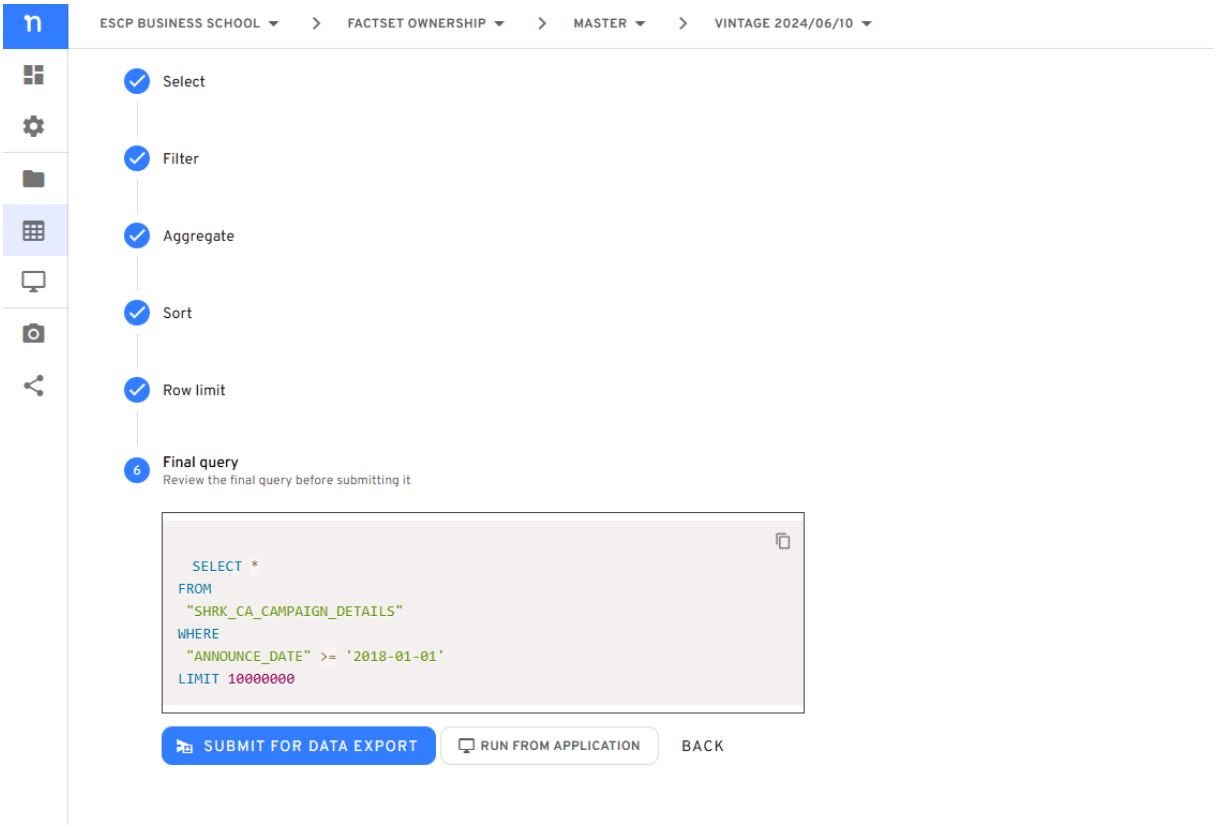
There are three alternative ways to access FactSet data through Nuvolos. The most straightforward method is via the platform's **query interface**, as outlined below:

1) Direct Query Access via Nuvolos Interface

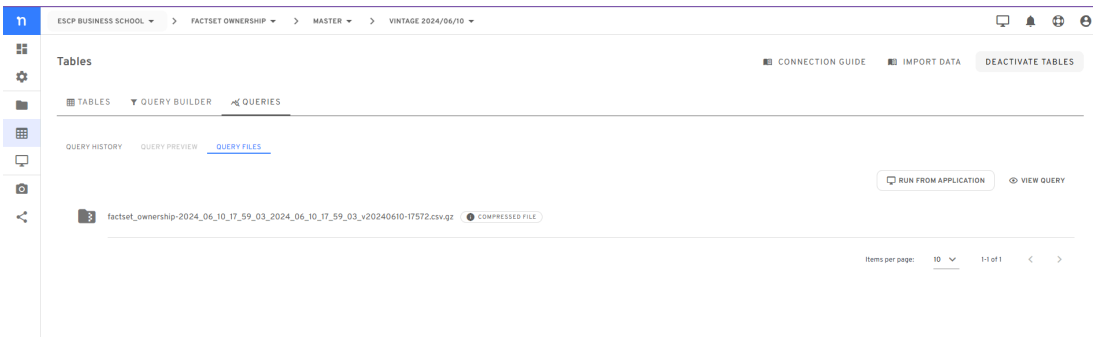
1. Log in to your Nuvolos account: https://app.nuvolos.cloud/org/21/dashboard
2. Navigate to the “FactSet Ownership” dataset
3. Click on the “Tables” tab located in the left-hand menu
4. Find your table of interest and click on the “Query” button
5. You can preview the table structure and explore columns by clicking directly on the table name



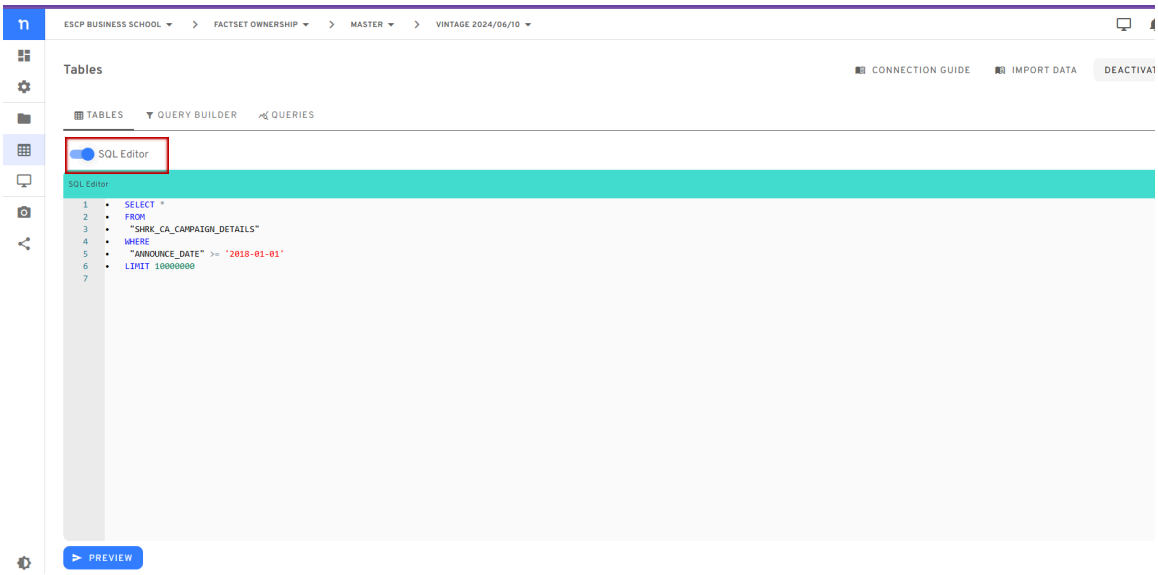
6. Extract data, first select the variables (columns) you need, then apply any necessary filters to narrow down your results, for example, by date range or company name. You can also use aggregation functions such as sum or average, sort the output by one or more columns, and specify the number of records to retrieve. Once your query is configured, submit it to extract the data.



7. Download the resulting CSV file provided after the query completes to save the extracted data



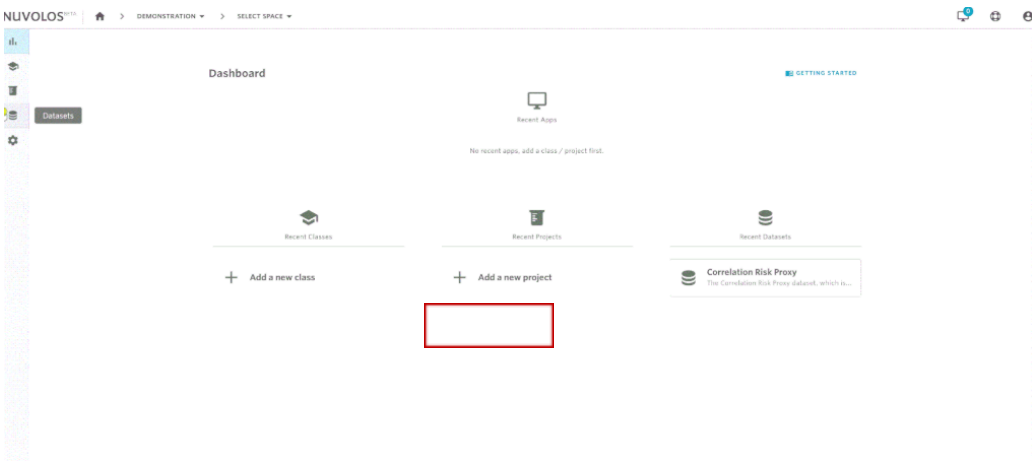
- Alternatively, you can submit your query directly in the query bar located above the table names, without going through the guided steps. For example:



- For more details, please refer to the official Nuvolos documentation: https://docs.nuvolos.cloud/features/database-integration/build-queries.

2) Accessing and Working with FactSet Data through Nuvolos Applications

1. Set up a research project: You may create research projects directly from the dashboard.



2. Provide a project name, set its privacy level, and optionally include an Overview.
3. Choose an application.

4. Add material to the project: You can upload files and source code via Nuvolos' file management features. Nuvolos also supports other storage and synchronization options.
5. Create and Launch an Application:

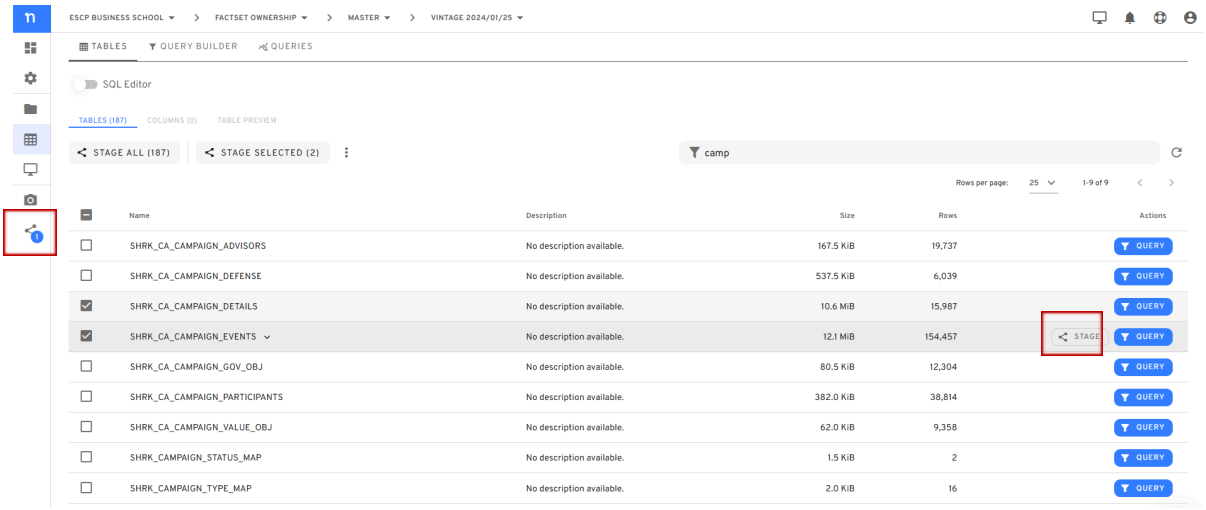
   Navigate to the Master Instance Overview via the dashboard or breadcrumb context. Then:

   i. Click the **Applications** tile or menu

   ii. Select **ADD NEW APPLICATION**

   iii. Choose your desired environment (e.g., JupyterLab or RStudio)
6. Distribute Required Data to Your Project:

   Use the **Distribution** tab to select and distribute the data to your workspace.
7. Once distributed, you will be able to work with the data inside your chosen application.



- For more details on the data integration workflow: https://docs.nuvolos.cloud/features/database-integration
- The example below assumes that the user is uploading tables to Nuvolos using Python Nuvolos Applications:

```
from nuvolos import get_connection, to_sql
import pandas as pd
```

```
df = pd.read_csv(...) #read your data
con = get_connection()
to_sql(df=df, name="table_name", con=con, if_exists='replace', index=False)
```

- For code examples in Python and other supported applications: https://docs.nuvolos.cloud/features/database-integration/upload-data

3) Uploading Data from Your Local Computer

If you wish to upload data to Nuvolos from your local device after distributing it to your instance, please follow the official documentation.[3] For steps on how to:

- Obtain your **connection tokens**
- Identify the correct **database name** and **schema name** to use

```
from nuvolos import get_connection, to_sql
import pandas as pd

df = pd.read_csv(...) #read your data
con = get_connection(dbname = "dbname", schemaname="schemaname")
to_sql(df=df, name="table_name", con=con, if_exists='replace', index=False)
```

[3] Instructions on how to obtain connection tokens

2.4 LSEG[4]

2.4.1 Content

Datastream: stock prices, financial indicators, financial markets (incl. Indexes), macroeconomic data over long time series (> 60 years)

- LSEG (Refinitiv Eikon): corporate finance data, stock prices, and other detailed information

2.4.2 Common usage

- Datastream :
    - Chartering
    - Long time series data (esp. macroeconomics)
    - Macroeconomic watch and forecasts
- LSEG (Refinitiv Eikon):
    - Return analysis
    - Beta analysis
    - Firm performance analysis and forecasts
    - Transaction data (comparables for M&A, IPO)
    - Firm valuation (DCF, comparables)
    - Ownership analysis
    - Debt information (syndicate, costs)
    - ESG and financial ratings
    - News watch
    - Alternative to Bloomberg with a weaker market side and a stronger corporate side

[4] LSEG (London Stock Exchange Group), previously known as Refinitiv

2.4.3 Access

- Trading Room (Montparnasse) (login created onsite immediately)
- Online access via the library website (login created online within a few days)
- 2 on-site licenses and 100 remote licenses at the same time (two weeks' booking)

Step 1: Fill Google form

- Click on the Link
- Fill out the Google form
- Sign up with the **school email**

Step 2: Wait for Administrative Approval

- The approval should happen within 5 Business days
- You will receive a message from Refinitiv once your license has been unlocked
- From this point, your license is available for **7 days Past those 7 days, you will need to start a new request**

2.5 RavenPack (Restricted to selected faculty only)

2.5.1 Content

Using over 40,000 sources, RavenPack provides real-time news analytics, including sentiment analysis and event data focused on business and financial applications. Data includes news and social media content, allowing for comprehensive analysis of financial markets.

Equity Entity Coverage by Region

- Europe: 40.46%
- North America: 30.95%
- Asia: 21.85%
- Latin America: 2.43%
- Oceania: 2.13%
- Africa: 1.11%
- Middle East: 1.06%

2.5.2 User guide link

https://drive.google.com/file/d/1Kb9Oqy1FeyE-4tVNkE5Lhi3E2DC0KTe6/view?usp=sharing

2.6 WRDS

Content

- Stock Exchange data for listed American firms
- Corporate finance data for listed American firms

Common usage

- Advanced Financial analysis
- Beta calculations
- Fama French Factor Analysis

Access

- Trading Room (Montparnasse and Champerret) (login created onsite immediately)
- Online access via the library website
- Unlimited licenses

2.6.1 BoardEx

2.6.1.1 Content

BoardEx provides extensive data on executive compensation, employment history, and professional relationships for over 20,000 companies. It includes valuable information on approximately 1.7 million corporate executives and board members across 2.2 million organizations worldwide. Biographical data includes details such as age, gender, positions held, educational qualifications, compensation, and stock holdings, as well as insights into the executive's professional network.

2.6.1.2 Coverage

Access through the ESCP WRDS account includes BoardEx North America, with datasets offering varying historical coverage, typically dating back to the early 2000s.

2.6.1.3 Access

Access is provided through the WRDS platform.

BoardEx Access via ESCP WRDS Account: Boardex North America[5]

- Announcements
- Committee Details
- Company Profile
- Compensation Analysis
- Individual Profile
- Networks / Associations
- Organization Summary

[5] Director profiles are divided into regions based on the director's employment records. For the Boardex-North America: HOCountryName = United States or Canada **OR** Company is in one of the following indices: S&P 500, NASDAQ 100, S&P MID CAP 400, S&P SMALL CAP 600, DOW JONES INDUSTRIAL AVG, S&P/TSX 60

2.6.1.4 Identifiers / Linking to Other Products

The primary identifiers are:

- BoardID and CompanyID (for company-level data).
- DirectorID (for individual executive/board member data).
- BoardID and CompanyID are interchangeable; they represent the same entity and can be used seamlessly across datasets.
- Many WRDS-provided tables supply both identifiers to facilitate easier merging and data integration with other products or databases.

2.6.1.5 Navigation path

WRDS/Home/Get Data/BoardEx/BoardEx - North America

2.6.1.6 Retrieving the Data

For instructions on downloading data, see the section "How to use WRDS to download the data".

Detailed BoardEx user guide link:

https://metalib.ie.edu/ayuda/Varios/BoardExWRDSDataDictionary.pdf

2.6.2 Compustat

2.6.2.1 Content

Compustat Fundamentals provides standardized financial statements and market data for over 80,000 active and inactive publicly traded companies across North America and globally. For more than 50 years, financial professionals have relied on this dataset. Compustat includes information such as Global Industry Classification Standards (GICS), pricing data, earnings data, insider and institutional holdings, and other metrics used by investors, analysts, economists, academics, and researchers.

2.6.2.2 Coverage

Compustat provides data for companies across North America and globally, with datasets offering varying historical coverage typically dating back to 1950.

2.6.2.3 Access

Access is provided through the WRDS platform.

Compustat Access via ESCP WRDS Account:

- Compustat (North America, Global, Bank, Historical Segments, Execucomp)
- Other Compustat (Marginal Tax Rates)
- Tools (Complete Financial Statements (XLS), CUSIP Converter, Data and Ratios (XLS), Financial Statements (HTML))
- Legacy Compustat (North America Ratings, Short Interest - 1973 to 2024)

2.6.2.4 Identifiers / Linking to Other Products

The key identifier used in Compustat is GVKEY, a unique and permanent company code. It ensures consistent identification of firms across time, even if names or other identifiers change. GVKEY is also the primary key used in the CRSP/Compustat Merged Database, facilitating seamless linking between Compustat and CRSP datasets.

2.6.2.5 Navigation path

WRDS/Home/Get Data Compustat - Capital IQ

2.6.2.6 Retrieving the Data

For instructions on downloading data, see the section "How to use WRDS to download the data".

Detailed Compustat user guide link:

https://wrds-www.wharton.upenn.edu/documents/1583/Compustat_Data_Guide.pdf

2.6.3 CRSP

2.6.3.1 Content

The Center for Research in Security Prices, LLC (CRSP) maintains the most comprehensive collection of security price, return, and volume data for the NYSE, AMEX and NASDAQ stock markets. Additional CRSP files provide stock indices, beta-based and cap-based portfolios, treasury bond and risk-free rates, mutual funds, and real estate data.

CRSP U.S. Stock database contains the following information:

- Price and quote data (e.g. Open, close, bid/low, ask/high, trade-only).
- Holding period returns with and without dividends.
- Excess returns and other derived data items.
- Market capitalization.
- Shares outstanding.
- Trading volume.
- Security delisting information.
- Corporate actions.
- Identifiers, descriptors, and supplemental data items.

2.6.3.2 Coverage

CRSP U.S. Stock data coverage for the various exchanges includes:

- NYSE: All data series begin on December 31, 1925.
- NYSE MKT: All data series begin on July 2, 1962.
- NASDAQ: All data series begin on December 14, 1972.
- Arca: All data series begin March 8, 2006.

The market indices cover:

- Equal and value-weighted returns for CRSP NYSE, NYSE MKT, NASDAQ, and Arca with and without dividends.

- Composite Indices for S&P 500 and NASDAQ.

At the share type (SHRCD) level, CRSP U.S. Stock includes:

- Common Stocks
- Certificates
- ADRs
- Shares of Beneficial Interest
- Units (e.g. Depository Units, Units of Beneficial Interest, Units of Limited Partnership Interest, Depository Receipts)
- ETFs
- Closed-End Mutual Funds
- Foreign companies traded on NYSE, NYSE MKT, NASDAQ, and NYSE Arca
- Americus Trust Components (Primes and Scores)
- HOLDRs Trusts
- REITs (Real Estate Investment Trusts)

CRSP U.S. Stock database excludes:

- Rights and warrants
- Preferred shares
- Units representing common stocks bundled with rights or warrants
- Over-the-Counter Bulletin Board Issues
- When Issued Trading

2.6.3.3 Access

Access is provided through the WRDS platform.

Detailed CRSP User guide link:

https://drive.google.com/file/d/1WmjEPhITeCPy1HsVpifFWsgj9_gHL42R/view?usp=sharing

Linking Matrix between datasets on the WRDS platform:

https://wrds-www.wharton.upenn.edu/pages/wrds-research/database-linking-matrix/

2.6.4 CRSP/Compustat Merged (CCM)

2.6.4.1 Content

The CRSP/Compustat Merged Database () is a product developed by CRSP that integrates company-level fundamental data from Compustat with security-level market data from CRSP. This unified structure enables researchers to seamlessly match financial statement variables with stock return and price data using consistent identifiers, such as CRSP's PERMNO/PERMCO and Compustat's GVKEY. The CCM database is essential for conducting longitudinal financial research and is widely used in academic studies across finance, accounting, and investment disciplines.

2.6.4.2 Coverage

The CCM database primarily covers companies based in North America, though some global firms may be included depending on the source datasets. With datasets offering varying historical coverage, typically dating back to 1950.

2.6.4.3 Access

Access is provided through the WRDS platform.

CCM Access via ESCP WRDS Account:

- Bank Annual
- Bank Quarterly
- Compustat CRSP Link
- Compustat Fundamentals Annual with CRSP Stock Monthly
- CRSP Stock Monthly with Compustat Fundamentals Annual
- Fundamentals Annual
- Fundamentals Quarterly
- Security Daily
- Security Monthly
- Segments

2.6.4.4 Identifiers / Linking to Other Products

The key linking variable is GVKEY, which serves as a unique and permanent identifier for each company in the Compustat database. In the merged CCM data, GVKEY is combined with CRSP's PERMNO (permanent security identifier) and PERMCO (company-level identifier). These fields allow users to match firms across datasets even when ticker symbols or names change over time. Other important identifiers include CUSIP (used for security-level linking), TICKER, and CIK (SEC Central Index Key), which can aid in cross-referencing with external datasets such as SEC filings or Bloomberg.

2.6.4.5 Navigation path

WRDS/Home/Get Data/CRSP/Annual Update/ CRSP/Compustat Merged

2.6.4.6 Retrieving the Data

For instructions on downloading data, see the section "How to use WRDS to download the data".

Detailed CCM user guide:

https://drive.google.com/file/d/1A3s6q2pBQFifsjaoBNmjNy7fWsEkMLrt/view?usp=sharing

2.6.5 I/B/E/S

2.6.5.1 Content

I/B/E/S (IBES), the Institutional Brokers' Estimate System, provides summary and individual analyst forecasts of company earnings, cash flows, and other important financial items, as well as buy-sell-hold recommendations. The data includes

- Summary (consensus)
- Individual analyst-level forecasts on
    - company earnings
    - cash flows
    - dividends
    - EBITDA
    - other important financial items.
- Price targets
- Buy-sell-hold recommendations.

2.6.5.2 Access

Access is provided through the WRDS platform.

Detailed I/B/E/S User guide link:
https://drive.google.com/file/d/1LZ4GlSLeycdvqb_kIL8hEdHPpovC209Q/view?usp=sharing

2.6.6 ISS ESG

2.6.6.1 Content

ISS ESG offers a comprehensive and evolving dataset used by institutional investors and academics worldwide to inform decision-making. The database includes detailed information on corporate directors, executive compensation, corporate governance practices, company vote results, climate and emissions data, and a variety of other Environmental, Social, and Governance (ESG) metrics.

2.6.6.2 Coverage

Geographic coverage includes North America, with dataset-specific availability varying: voting datasets are available from as early as 2003, while shareholder proposals data begins in 2006.

2.6.6.3 Access

Access is provided through the WRDS platform.

ISS ESG Access via ESCP WRDS Account:

- Other (SHAREHOLDER PROPOSAL): Data for the S&P 1500 including proposals that were submitted but were omitted or withdrawn.
- Voting Analytics (Company Vote Results US, Mutual Fund Vote Records, Shareholder Proposals): U.S. mutual fund voting records for all institutions filing the SEC form N-PX - Global institutions covered by ISS outside SEC U.S. disclosures.

2.6.6.4 Identifiers / linking to other products

The primary company identifier in ISS ESG data is iss_companyid. For mutual funds in the voting data, the key identifier is fundid. Datasets such as Company Vote Results US and Shareholder Proposals also include CUSIP codes for linking to other financial datasets.

2.6.6.5 Navigation path

WRDS/Home/Get Data/ISS ESG

2.6.6.6 Retrieving the Data

For instructions on downloading data, see the section “How to use WRDS to download the data”.

2.6.6.7 Detailed ISS ESG user guid link

https://wrds-www.wharton.upenn.edu/pages/support/manuals-and-overviews/iss-formerly-riskmetrics/iss-data-field-definitions/

2.6.7 How to Create a WRDS Account

- Visit WRDS Registration and select "**Register for a WRDS Account**" (located just below the login fields).
- Your registration request will be automatically routed for approval. Once approved, your account will be active for **one year**.

2.6.8 How to Use WRDS to Retrieve the Data

There are several ways to retrieve data using WRDS. Below is a step-by-step example using **BoardEx** data as an example.

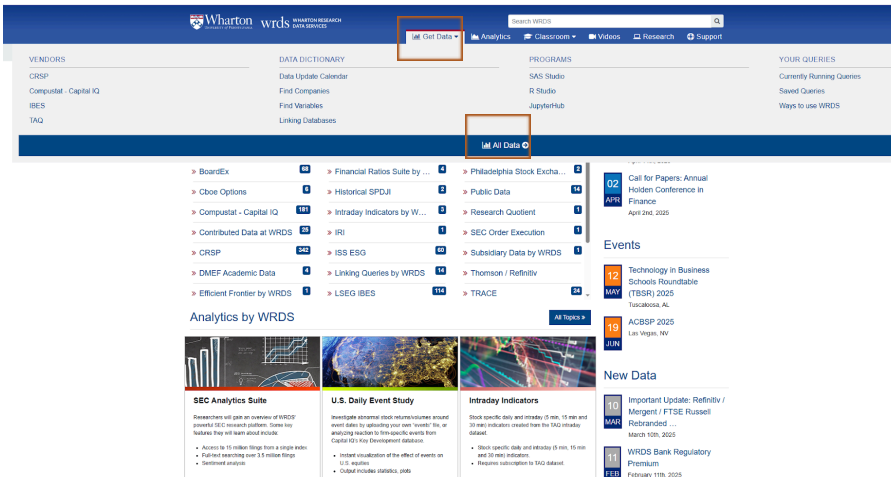
1. Accessing WRDS Data via the Web Interface

The most straightforward way to access WRDS data is by navigating **directly through the WRDS platform** using the steps outlined below. to illustrate the process:

1. **Log in to your WRDS account**:
   Go to https://wrds-www.wharton.upenn.edu/
2. **Navigate to the data section**:
   Click on **“Get Data”** in the top navigation bar.
3. **Access available datasets**:
   Click on **“All Data”** to view the full list of WRDS products and data vendors.

4. **Select your dataset**:

   From the list, choose **BoardEx** (or any other dataset depending on your research needs).



5. **Example Path – Accessing Organization Summary (BoardEx):**

   From the homepage, follow the path:

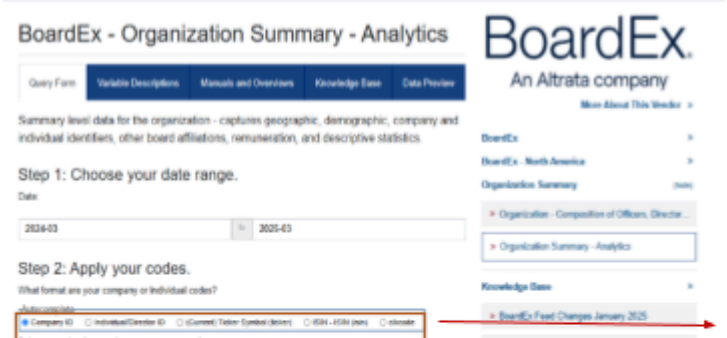
   Home > Get Data > BoardEx > BoardEx - North America > Organization Summary >Organization Summary – Analytics

6. **Select Your Date Range:**

   Use the calendar fields to define the desired time period for your data query.

7. **Search or Upload Company Identifiers:**

   - To retrieve information for specific companies, enter the **company name** or **identifier** (e.g., Ticker, CUSIP).
   - If uploading a list of companies by code, first select the appropriate **identifier type** (e.g., Ticker, CUSIP).
   - To upload a list, click **"Browse"** and select a plain text file (.txt) with **one code per line**.
   - To extract data for **all available companies**, choose the option **"Search the entire database."**



8. **Select the variables you need:**

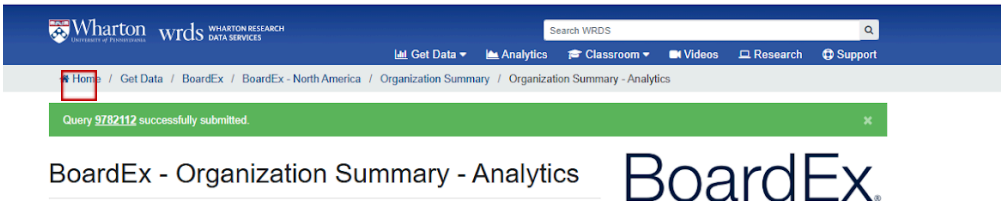
   You can view a description of each variable by clicking the "?" icon next to its name.

9. **Choose your desired output format:**

   Select the format in which you'd like to receive the data (e.g., CSV, Excel, or text).

10. **Submit your query:**

    Once your query is submitted, a message will appear in the top bar. Click on the query number to download your data.



2. Connection Using Your PC

WRDS data is also accessible remotely from your personal workstation or laptop through several popular programming environments. These methods allow for seamless integration into research workflows and enable advanced data manipulation directly within your preferred tools.

You can connect to WRDS and retrieve datasets using the following options:

- **SAS** via PC-SAS/Connect
- **R** via R Console or RStudio
- **MATLAB**
- **Stata**
- **Python** (via local language installation)

- **ODBC/JDBC-compatible applications**

WRDS provides **ODBC** and **JDBC** interfaces, which means any software supporting these protocols can retrieve data from the WRDS system. These methods connect either to the **PostgreSQL database cluster** or to the **SAS datasets on disk**. While each method requires a brief setup, they enable efficient and direct access to WRDS data from within your programming environment.

Full documentation: Programming at WRDS: https://wrds-www.wharton.upenn.edu/pages/grid-items/programming-guides/

Example: Connecting to WRDS Using Python

Preparation – Install WRDS Python Module

"All WRDS data is stored in a PostgreSQL database, and is available through Python through

In-house Python module, wrds, which is freely available on PyPI via a pip install.

```
Pip install wrds
```

1. Connecting to WRDS

1.1 Establish Connection

```
import wrds
conn = wrds.Connection(wrds_username= 'your_username')
```

1.2 List Libraries: This function displays all data libraries available through the institution's WRDS subscription.

```
conn.list_libraries()
```

1.3 List Tables Within a Library: Lists all datasets available under the boardex library, often used for executive data.

```
conn.list_tables(library='boardex')
```

1.4 Describe a Specific Table: Use this to understand the structure and variables in a given dataset (similar to proc contents in SAS).

```
conn.describe_table(library='boardex', table='na_wrds_org_summary')
```

2. Calling Data

2.1 Basic Method: This command extracts the first 1000 rows from the specified table.

```
Boardex_org_sum = conn.get_table(library='boardex', table='na_wrds_org_summary', obs=1000)
Print(boardex_org_sum.columns)
```

2.2 Advanced Method: This SQL-based method is useful for extracting filtered and conditional datasets.

Example: calling data while imposing a conditioning statement:

```
board_org_f = conn.raw_sql("""
SELECT directorname, rolename, rolestatus, gender, nationality
FROM boardex.na_wrds_org_summary
WHERE annualreportdate BETWEEN '2010-12-01' AND '2011-12-31'
  AND gender = 'F'
""")
```

3. Storing Output

3.1 Native "Pickle": The Python native "pickle" package is versatile for storing and retrieving data. The code below stores the data "board_org_f" extracted from the raw_sql method above.

```
import pickle as pkl
with open ("boardf.pkl", "wb") as f:
   pkl.dump(board_org_f, f)
```

Load the Pickle File Later:

```
with open("boardf.pkl", "rb") as f:
    board_fem = pkl.load(f)
```

3.2 Exporting as Flat Files: Store your data in a format that can be used by other tools like Excel, Stata, or R.

```
board_org_f.to_csv(r'your path\board_fem.csv') #csv
board_org_f.to_excel(r'your path\board_fem.xlsx') #excel
```

Further information can be found here:
https://www.fredasongdrechsler.com/data-crunching/connect-wrds

3. Using Web Query Forms with the web browser

All you need is a web browser and Internet access to query data from WRDS. No software to install, no specific programming language to learn. After you've decided on which dataset to use, you can use the web-based query form to extract and output data in the format of your choice.

Each dataset on WRDS includes a description of what's included as well as links to the pages for each query. The web-based version of WRDS is designed to work with any modern browser.

The web queries have the same general format for all databases in WRDS. When you submit your query, it runs on the powerful servers and is made available for you to download when it is complete.

Full documentation: Web Queries or Introduction to Web Queries

4. Connect Using the WRDS Cloud

WRDS provides a high-performance computing cluster, the WRDS Cloud. WRDS Cloud provides command-line access to a powerful research computing platform. It enables access to all of the data and includes support for many programming languages.

The WRDS Cloud is accessible via SSH, a way of securely connecting to WRDS. Once connected, you'll be able to access WRDS data directly in your favorite programming languages. Languages supported on the WRDS Cloud:

- **SAS**
- **Python**
- **R**
- **Stata** (requires active Stata license with StataCorp)
- **MATLAB** (requires active license with MathWorks)

Full documentation: The WRDS Cloud:
https://wrds-www.wharton.upenn.edu/pages/grid-items/accessing-wrds-cloud/

3 Section II : Free and Open Access Databases

- **10-X filing data from Notre Dame SRAF**
- **Institutional Holdings SEC 13F Data**
- **Master Dictionary from Notre Dame SRAF**

- **Other Asset Pricing open sources**
- **U.S. Bank Holdng Company Filing from National Information Center**
- **U.S. Regulation from Federal Register**
- **Volatility Analysis from NYU Volatility Lab**

3.1 10-X filing data from Notre Dame SRAF

3.1.1 Content

"10-X" refers to filings made to the U.S. Securities and Exchange Commission (SEC) that are variations of the 10-K report. These annual and quarterly filings are required for companies with securities registered under Section 12 or subject to Section 15(d) of the Securities Exchange Act of 1934 and subject to the periodic and current reporting requirements of Section 13 or 15(d). These include:

- 10-K filings: The annual reports that publicly traded companies are must file with the SEC. They include audited financial statements (such as the income statement, balance sheet, and cash flow statement), a summary of operations, risks, and management's analysis, and information about company structure, governance, and executive compensation, etc.
- 10-Q filings: The quarterly report that quarterly updates that contain similar information to the 10-K, but in a shorter and less detailed format.
- Amended filings, transition reports, and other filings that differ in format but are classified under the 10-K/10-Q categories.

3.1.2 Data source

The full-text versions of these filings, available on the SEC's EDGAR system, combine all components in the browser-friendly files. However, much of the content includes HTML formatting, embedded images, and other elements that are not typically of interest. To improve processing efficiency, these files can be cleaned and compressed to retain only the relevant text content.

Notre Dame Software Repository for Accounting and Finance (SRAF), which calls "Stage One Parsed" versions of these filings. All 10-X SEC complete text document filings are downloaded for each year/quarter, based on the original frequency provided by EDGAR. And in this stage, each document is cleaned to remove unnecessary coding and non-text elements, focusing only on the textual information. Additional markups are also added to the beginning of each parsed file to summarize its contents and structure.

3.1.3 Data Coverage

Notre Dame SRAF offers Stage One Parsed data for all 10-X filings from the third quarter of 1993 to the present. While data availability is more limited in the earlier years (especially before 2000), 10-K filings are available annually and 10-Q filings are available quarterly. Some gaps may exist due to missing data on the SEC servers, which are documented on the provider's website.

In addition to the Stage One Parsed data, Notre Dame SRAF also offers SEC/EDGAR related datasets, including raw versions of 10-X filings, and summary files that provide metadata for each

filing or the filing's header, and all SEC/EDGAR filings tabulation. More information about these additional resources is available through the official website.

3.1.4 Access

Official website for Notre Dame SRAF: https://sraf.nd.edu/data/stage-one-10-x-parse-data/

Where users can find detailed information and direct access to the Stage One Parsed 10-X filing data. For additional SEC/EDGAR-related datasets and resources, please visit: https://sraf.nd.edu/sec- edgar-data/ . The data is freely available for academic research, and no registration or license is required to download and use it. Users are kindly asked to cite the website when using the data in their research.

3.2 Institutional Holdings SEC 13F Data

3.2.1 Content

Form 13F filings include investment institutions, mutual funds, and hedge funds, which were disclosed quarterly. The U.S. Securities and Exchange Commission (SEC) website provides more detailed information about 13F filings.

This document provides instructions for downloading and processing SEC Form 13F datasets and is based on Li (2025). This guide helps to convert the data from SEC website in CSV form. The

following Python solution provides a way of working with SEC 13F data for academic research. This guide aims to support investment research at ESCP Center for
Finance.

3.2.2 Prerequisites

This section describes required Python libraries and directory structure. Code is tested in Python 3.9 via PyCharm.

Required Python Libraries

Required Python libraries can be installed with the following command:

```
pip install requests
pip install beautifulsoup4
pip install pandas
```

Directory Structure

sec_13f_data folder is necessary to store the results.

3.2.3 Script Overview

Python script ("sec_13f.py") processes SEC 13F data. Key functions include:

- *get_zip_links(start_year, end_year):* Fetches SEC ZIP file links for the specified years;
- *process_tsv_files(folder, output_csv)*: Parses and cleans TSV files, then save as CSV;
- *postprocess_csv_dates(csv_file)*: Converts date columns to YYYY-MM-DD format;
- *main()*: Orchestrates the workflow.

3.2.4 Step-by-Step Instructions

This section contains the necessary steps to process SEC Form 13F data.

Step 1: Downloading Raw Data

Download and extract quarterly ZIP files directly from https://www.sec.gov/data-research/sec-markets-data/form-13f-data-sets. Save every quarter in *sec_13f_data* folder.

Step 2: Process Files with Python

Save the following code as “sec_13f_main.py”.

```
from bs4 import BeautifulSoup
import pandas as pd
import requests
import shutil
import re
import gc

# Set folders for storing data
BASE_FOLDER = 'sec_13f_data'
ANNUAL_FOLDER = 'sec_13f_annual'
os.makedirs(BASE_FOLDER, exist_ok=True)
os.makedirs(ANNUAL_FOLDER, exist_ok=True)

# 1: Get SEC ZIP Links
def get_zip_links(start_year, end_year):
    url = "https://www.sec.gov/data-research/sec-markets-data/form-13f-data-sets"
    headers = {
        "User-Agent": "MyAppName/1.0 (myemail@example.com)",
        "Accept-Language": "en-US,en;q=0.9",
        "Referer": "https://www.sec.gov/",
    }
    try:
        response = requests.get(url, headers=headers, timeout=10)
        response.raise_for_status()
        soup = BeautifulSoup(response.content, 'html.parser')
        zip_links = []

        for year in range(start_year, end_year + 1):
```

```
            for link in soup.find_all('a', href=True):
                href = link['href']
                if href.endswith('.zip'):
                    if year >= 2024 and re.search(rf"{year}", href):
                        zip_links.append(f"https://www.sec.gov{href}")
                    elif f"{year}q" in href:
                        zip_links.append(f"https://www.sec.gov{href}")

        if not zip_links:
            raise Exception("[ERROR] No matching ZIP files found.")

        print(f"[INFO] Found {len(zip_links)} ZIP files and start to download and process.")
        return zip_links

    except requests.RequestException as e:
        raise Exception(f"[ERROR] Failed to fetch ZIP links: {e}")
    except Exception as e:
        raise Exception(f"[ERROR] An error occurred: {e}")

# 2: Process and Parsing TSV and convert to CSV
def process_tsv_files(folder, output_csv, chunksize=15000):
    # Step 1: Load all other TSVs (excluding INFOTABLE) into dicts of dataframes grouped
by ACCESSION_NUMBER
    merge_data = {}
    for file in os.listdir(folder):
        if file.endswith('.tsv') and 'INFOTABLE' not in file:
            path = os.path.join(folder, file)
            try:
                df = pd.read_csv(path, sep='\t', encoding='utf-8')
                if 'ACCESSION_NUMBER' in df.columns:
```

```
            merge_data[file] = df.set_index('ACCESSION_NUMBER')
            print(f"[INFO] Indexed {file}")
        except Exception as e:
            print(f"[ERROR] Could not load {file}: {e}")

# Step 2: Stream INFOTABLE
infotable_path = os.path.join(folder, 'INFOTABLE.tsv')
if not os.path.exists(infotable_path):
    raise FileNotFoundError("[ERROR] INFOTABLE.tsv not found.")

# Step 3: Define final column mapping
columns_to_keep = {
    'CIK': 'cik',
    'FILINGMANAGER_NAME': 'coname',
    'SUBMISSIONTYPE': 'form',
    'PERIODOFREPORT': 'rdate',
    'FILING_DATE': 'fdate',
    'NAMEOFISSUER': 'nameOfIssuer',
    'TITLEOFCLASS': 'titleOfClass',
    'VALUE': 'value',
    'SSHPRNAMT': 'sshPrnamt',
    'SSHPRNAMTTYPE': 'sshPrnamtType',
    'PUTCALL': 'putCall',
    'INVESTMENTDISCRETION': 'investmentDiscretion',
    'OTHERMANAGER': 'otherManager',
    'VOTING_AUTH_SOLE': 'Sole',
    'VOTING_AUTH_SHARED': 'Shared',
    'VOTING_AUTH_NONE': 'None'
}
```

```
    # Step 4: Write header once
    header_written = False

    for chunk in pd.read_csv(infotable_path, sep='\t', encoding='utf-8',
chunksize=chunksize):
        print(f"[INFO] Processing chunk of size {len(chunk)}")

        # Add CIK
        if 'ACCESSION_NUMBER' not in chunk.columns:
            raise ValueError("[ERROR] ACCESSION_NUMBER column missing in
INFOTABLE.")
        chunk['CIK'] = chunk['ACCESSION_NUMBER'].str.split('-').str[0]

        # Merge each additional file on ACCESSION_NUMBER
        for file, df in merge_data.items():
            merge_cols = [col for col in df.columns if col not in chunk.columns or col ==
'ACCESSION_NUMBER']
            if merge_cols:
                chunk = chunk.merge(df[merge_cols], left_on='ACCESSION_NUMBER',
right_index=True, how='left')

        # Drop duplicates in this chunk
        subset_cols = ['CIK', 'FILINGMANAGER_NAME', 'SUBMISSIONTYPE',
'PERIODOFREPORT',
                       'FILING_DATE', 'NAMEOFISSUER', 'TITLEOFCLASS', 'CUSIP', 'VALUE',
'SSHPRNAMT']
        subset_cols = [col for col in subset_cols if col in chunk.columns]
        chunk.drop_duplicates(subset=subset_cols, inplace=True)

        # Filter and rename columns
```

```
        final_chunk = chunk[[col for col in columns_to_keep if col in chunk.columns]].copy()
        final_chunk.rename(columns=columns_to_keep, inplace=True)

        # Downcast for memory savings
        if 'value' in final_chunk.columns:
            final_chunk['value'] = pd.to_numeric(final_chunk['value'], downcast='float')
        if 'sshPrnamt' in final_chunk.columns:
            final_chunk['sshPrnamt'] = pd.to_numeric(final_chunk['sshPrnamt'],
downcast='integer')

        # Append to CSV
        final_chunk.to_csv(output_csv, mode='a', index=False, header=not header_written)
        header_written = True

        del chunk, final_chunk
        gc.collect()

    # Step 5: Apply date processing after all chunks are written
    postprocess_csv_dates(output_csv)

    print(f"[INFO] Streaming complete. Output saved to {output_csv}")

# 3: Main Workflow
def main():
    start_year = int(input("Enter start year: "))
    end_year = int(input("Enter end year: "))
    zip_links = get_zip_links(start_year, end_year)
    # Ensure the user enters a valid range
    if start_year < 2013 or end_year > 2030 or start_year > end_year:
```

```
        print("[ERROR] Invalid year range. Please enter a valid range between 2013 and
2030.")
        return
    for zip_link in zip_links:
        zip_folder = os.path.join(BASE_FOLDER, os.path.basename(zip_link).replace('.zip',
''))
        print(zip_folder)
        # zip_path = download_zip(zip_link, zip_folder)
        # extract_zip(zip_path, zip_folder)
        process_tsv_files(zip_folder, os.path.join(BASE_FOLDER,
f"{os.path.basename(zip_link).replace('.zip', '')}.csv"))

# 4: Post-Process Dates in Final CSV
def postprocess_csv_dates(csv_file):
    """
    Ensures date columns (rdate and fdate) are formatted as YYYY-MM-DD after CSV
export.
    """
    try:
        # Explicitly define column types to handle mixed data types
        dtype_mapping = {'rdate': 'str', 'fdate': 'str'}
        df = pd.read_csv(csv_file, dtype=dtype_mapping, low_memory=False)
        # Explicitly format date columns
        for date_col in ['rdate', 'fdate']:
            if date_col in df.columns:
                df[date_col] = pd.to_datetime(df[date_col],
format='%d-%b-%Y',errors='coerce').dt.strftime('%Y-%m-%d')
        # Save the CSV again with updated date formats
        df.to_csv(csv_file, index=False)
    except Exception as e:
```

```
        print(f"[ERROR] Failed to format dates in final CSV: {e}")


# Run the main workflow
if __name__ == "__main__":
    main()
\end{lstlisting}
```

Please change the default name (MyAppName) and email address (myemail@example.com) to your own in # 1: Get SEC ZIP Links:

"User-Agent": "**MyAppName**/1.0 (**myemail@example.com**)"

After that, to execute, please run: sec_13f_main.py.

After running the code, there will be a request to input the start year and end year. In the first request, input the start year and enter. Then, input the end year and enter. The code will start processing Form 13F fillings.

3.2.5 Merge with CRSP

The CUSIP identifier is the basis to merge the data from the SEC website with CRSP. This identifier can have 6, 8, or 9 digits. The first 6 digits refer to the company, the next 2 refer to the security, and the 9th is a check digit that has no inherent meaning. CRSP uses 8-digit CUSIPs (for reference, Compustat uses 9-digit CUSIPs).

To merge 13F with CRSP using 8-digit CUSIPs you will need to access both data-frames through a programming language such as Python. Use a substring function to keep the first 8 digits.

```
import pandas as pd


# Keep first 8 characters (digits or letters)
df['CUSIP'] = df['CUSIP'].str[:8]
```

Dates are matched based on the period of report *rdate* (13F data) and *date* (CRSP data).

For more details, see:

Li, S. (2025). Guide for download and parsing sec 13f data in bulk: A python approach. *Available at SSRN 5101829*.

3.3 Master Dictionary from Notre Dame SRAF

3.3.1 Content

Words, including their inflected forms, are essential components of sentences, making accurate word identification a key part of textual analysis. To support this, various word lists are available online that include proper nouns and abbreviations (where a token refers to any group of characters). These word lists help identify whether a token is a valid word, which ensures consistency and provides a foundation for more advanced textual analysis.

3.3.2 Data source

The Loughran-McDonald Master Dictionary builds upon version 4.0 of the 2of12inf dictionary, described at: http://wordlist.sourceforge.net/12dicts-readme.html. To adapt the dictionary for financial applications, the authors expanded the word list by including commonly used terms and their inflections based on content from 10-K filings and earnings calls. The core dictionary provides data such as word counts, percentage of total words, average percentage per document, standard deviation across documents, and the number of documents in which each word appears. Further construction details are available in *When is a Liability not a Liability? (2011, JoF)*.

In addition to basic word information, the dictionary also includes sentiment and complexity indicators created by the authors. The sentiment categories include: negative, positive, uncertainty, litigious, strong modal, weak modal, and constraining. For further explanation, please refer to the JoF paper and Using 10-K Text to Gauge Financial Constraints (2015, JFQA). The complexity perspective provides a lexicon for measuring firm complexity. Details could be checked in Measuring Firm Complexity (2024, JFQA).

3.3.3 Data Coverage

The dictionary is updated on a periodic basis (yearly at least). The sentiment words are flagged with a number indicating the year in which they were added to the list. Details of additions and changes could be checked in the documentation and the related academic papers. The data is

free to download and use for academic research without registration or a license. Please remember to cite both the website and the relevant papers when using the data in your work.

3.3.4 Access

The dictionary can be accessed at https://sraf.nd.edu/loughranmcdonald-master-dictionary/. This page is directly linked to the description and access of the Master Dictionary. For detailed explanations of how the dictionary was created and structured, users should consult the documentation and the related academic papers by the authors. The data is freely available for academic research, and no registration or license is required to download and use it. Users are kindly asked to cite both the website and the relevant papers when using the data in their research.

3.4 U.S. Bank Holding Company Filing from National Information Center

3.4.1 Content

The FR Y-9 filing series is a set of regulatory reports required by the U.S. Federal Reserve Board (FRB) under the Bank Holding Company Act of 1956, as amended. These reports provide detailed financial and risk-related information about holding companies that control U.S. banks or savings and loan institutions. These include:

- FR Y-9C: Required for all domestic holding companies on a consolidated basis.
- FR Y-9LP: Required for all large domestic holding companies on an unconsolidated parent-only basis.
- FR Y-9SP: A simplified version of FR Y-9LP, required for all small domestic holding companies on an unconsolidated parent basis.

3.4.2 Data source

The National Information Center (NIC) provides downloadable versions of these reports. Each quarterly file is distributed as a compressed text file, organized by financial year and then quarter. It contains all reported variables at the time of the respective financial statements. When files are unzipped, the data will be in text files delimited by the caret symbol (^) because some values contain commas. Financial and some structural items for all three reports are contained in one row for each institution within each file.

The Federal Reserve System also maintains a data dictionary, known as the Micro Data Reference Manual (MDRM). This catalog describes each variable by its item name and number and identifies which reporting series (not limited to FR Y-9X) the variable appears in, along with the start and end dates of its use. Variables from each report type can be identified using specific prefixes: BHCK for FR Y-9C, BHCP for FR Y-9LP, and BHSP for FR Y-9SP.

In addition, the Federal Reserve Bank of Chicago previously provided a Holding Company List, which matches each holding company to a unique Federal Reserve identifier. However, this list is only updated through December 2020 due to changes in data sources (FRB Chicago to NIC).

3.4.3 Data Coverage

FR Y-9C and FR Y-9LP data are collected quarterly; FR Y-9SP is collected semiannually.

The data is available from 2000 Q1, and actively updated to date.

3.4.4 Access

https://www.ffiec.gov/npw/FinancialReport/FinancialDataDownload.

The data is freely available for academic research, and no registration or license is required to download and use it. Users are kindly asked to cite the website when using the data in their research.

The data dictionary MRDM could be checked here, along with a description of all reporting files and the mnemonics of corresponding series. The Holding Company List could be checked here.

The GitHub project here provides a convenient way to parse and select data from the downloaded files. The script processes the data for selected years, and then merges the files, and finally produces the output of selected variables.

3.5 U.S. Regulation from Federal Register

3.5.1 Content

The Federal Register is the official daily publication of the U.S. government that includes rules, proposed rules, notices from federal agencies and organizations, as well as executive orders and other presidential documents. It is produced by the Office of the Federal Register, a division of the National Archives and Records Administration, with the aim of improving public understanding- ing of the regulatory process.

Broadly speaking, although the content of the Federal Register can be extensive and complex, it reflects the flow of regulatory activity, including drafts and discussions, not just the final set of binding rules. When a federal agency issues a final rule and it is published in the Federal Register, it is eventually organized by topic and incorporated into the Code of Federal Regulations (CFR), which is updated annually.

3.5.2 Data source

Each daily issue of the printed Federal Register is organized into four categories:

- Presidential Documents, including executive orders and proclamations;
- Rules and Regulations, which include the full text of some final rules, along with supporting material such as background of regulatory implementation, the rationale for the final rule, rule-making analysis including the compliance with important existing requirements like the Paperwork Reduction Act, supporting information such as public comments and document statistics;
- Proposed Rules, including petitions to agencies from the public;
- Notices, such as scheduled hearings and meetings open to the public, grant applications.

3.5.3 Data Coverage

The Federal Register is published every weekday, except on federal holidays. It is accessible by

publication date or through CFR Indexing Terms, which is a classification that provides a common vocabulary for indexing the rule-making documents of all agencies.

3.5.4 Access

https://www.federalregister.gov/.

The data is freely available for academic research, and no registration or license is required to download and use it. Users are kindly asked to cite the website when using the data in their research.

3.6 Volatility Analysis from NYU Volatility Lab

3.6.1 Content

The Volatility Laboratory (V-Lab) of NYU Stern is established to advance the understanding of financial market dynamics through the focus on: Financial volatility, asset correlations, and systemic risk. Each application offers open-access models and real-time data to help society to understand the risks associated with financial markets and to support decision-making by academics, market participants, and policymakers.

3.6.2 Data Source

The key initiatives and research areas of V-Lab include:

- Volatility analysis: V-Lab produces forecasts of financial market volatility for various asset classes, such as equities, indices, currencies, and commodities. For each asset, users can access a volatility summary table, a graph showing one-step-ahead annualized volatility forecasts, and a model estimation table.
- Asset correlations: By analyzing time-varying correlations between asset returns, V-Lab helps track how relationships among assets change over time. These insights are useful for understanding contagion effects and the role of diversification in risk management. For each sub-topic, users can access a correlation matrix, time-series graphs of selected asset pairs, and correlation summary tables.
- Systemic risk: V-Lab provides indicators to evaluate vulnerabilities in the financial system and the potential impact of financial shocks. These measures capture the interconnected nature of financial institutions and markets. For a selected risk threshold, users can access SRISK graphs for specific institutions and a ranking table summarizing systemic risk metrics.

A wider array of financial risks are also analyzed in V-Lab, these include: Long-run Value at Risk, market liquidity measures, fixed income forecasting, climate risk indicators, and common volatility risk monitoring. All contents include detailed documentation.

3.6.3 Data Coverage

The datasets span from broad market-level indicators to individual firm/security-level or fund-level series. Well-established modules, such as volatility analysis, offer historical data dating back to the 1990s, while more recent areas, like climate risk, begin from the mid-2000s. Most data is available at a daily frequency.

3.6.4 Access

https://vlab.stern.nyu.edu/.

Visualization of analyses online requires no license. For downloading selected series for academic use, register process is required (free of charge). Users are kindly asked to cite the website when using the data in their research.

3.7 User guide links for Asset Pricing open sources

https://www.openassetpricing.com/

Ken French's Data library

https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

IPO Data – Jay Ritter

https://site.warrington.ufl.edu/ritter/ipo-data/

BEA - Bureau of Economic Analysis (GDP, Consumer Spending, and other national accounts)

https://www.bea.gov/data

BLS - Bureau of Labor Statistics (CPI, PPI, CES)

https://www.bls.gov/

Jordà-Schularick-Taylor Macrohistory Database

https://www.macrohistory.net/database/

Uncertainty indexes

https://www.sydneyludvigson.com/data-and-appendixes

CAY data

https://sites.google.com/view/martinlettau/data

Aswath Damodaran's datasets

https://pages.stern.nyu.edu/~adamodar/New_Home_Page/datacurrent.html

Will Rinehart's datasets

https://www.williamrinehart.com/datasets/#key-fred-data

4 Keywords