LEONARDO PALESTRA, MARCO GELSOMINI

INSIDER TRADING

Data Management Project

Agenda

1. Research question
2. Data sources
3. Storage
4. Data integration
5. Data quality
6. Visualizations
7. Future improvements

RESEARCH QUESTION

explore the relationships between the **stock market** and

the transactions of the **insider traders**

The insiders can make their

decisions of buying or selling anticipating a favorable move in the stock price?

Analyze different impacts and behaviors based on the role of the insider (CEO, CFO, ecc)

DATA SOURCES

NASDAQ OFFICIAL WEBSITE DOWNLOAD

1. Access the NASDAQ official website.
2. Retrieve the full list of listed companies, including ticker symbols and market capitalizazion.
3. Use Python to filter and retain only the **top 50** companies by market capitalization.
4. Store the filtered list in memory for subsequent API calls and web scraping.

ALPHA VANTAGE API

1. Register for an API key on the Alpha Vantage platform.
2. Iterate through the list of 50 selected companies, sending API requests for daily stock prices.
3. Retrieve and parse the JSON responses, extracting and elaborating relevant financial data.
4. Store the stock price data in memory for further analysis.

U.S SECURITY AND EXCHANGE COMMISSION

SEC WEBSITE SCRAPING

1. Iterate automatically through the list of 50 selected companies on the website. Visualize all the documents of the last 5 years.
2. Extract in json format all raw informations from each XML document (form-4) filling. This form is requested within 2 days when an insider executes a transaction.
3. Store the extracted transaction data in a json document along with the dates, the stock symbol and the entity who compiled the filling.

STORAGE

1. Why MongoDB?
2. NASDAQ company list storage
3. Stock Price Data storage
4. Insider Trading Data Storage

Why MongoDB?

- **JSON - based data structure**: Both the Alpha Vantage API and the SEC filings provide data in document format, which aligns naturally with MongoDB's document model. This eliminates the need for complex data transformations before storage.
- **Scalability for Large datasets**: The insider trading dataset consists of 18,914 documents, MongoDB efficiently handles large numbers of documents.
- **Flexible schema**: Insider trading data varies in structure depending on the content in each filling.
- **Best for integration**: The final analysis and integration of the data will also be performed within a MongoDB environment. It also usefull for handling non unique matches durign the lookup.

NASDAQ company list storage

The NASDAQ company list was used to define the scope of the analysis. Initially, all companies listed on the NASDAQ were retrieved from the official website.

Using a Python script, we filtered the dataset to retain only the top 50 companies by market capitalization.

This final list was then used to fetch stock prices and insider trading data.

The list itself was **not stored** in the database but was kept in memory during data retrieval.

Stock price data storage

The raw API response contained unnecessary fields and inconsistent field names, so before storing the data, the following transformations were applied:

- **Renamed API field** names to be more manageable (e.g., "1. close" → "close").
- Create a new field with the converted dates to **ISO date** format.
- Organized data so that each stock has a single document containing an **array** of daily price records. The array make it easier to retrive informations and reduce the memory required. The timestamps that were previously keys are now a field inside each daily data.

Before


```
{
    "Meta Data": {
        "1. Information": "Daily Prices
(open, high, low, close)     and Volumes",
        "2. Symbol": "IBM",
        "3. Last Refreshed": "2025-02-24",
        "4. Output Size": "Compact",
        "5. Time Zone": "US/Eastern"
    },
    "Time Series (Daily)": {
        "2025-02-24": {
            "1. open": "261.5000",
            "2. high": "263.8450",
            "3. low": "259.5800",
            "4. close": "261.8700",
            "5. volume": "4398107"
        },
        "2025-02-21": {
            "1. open": "263.8450",
            "2. high": "264.8300",
            "3. low": "261.1000",
            "4. close": "261.4800",
            "5. volume": "5667874"
        }, { ... } }
```


After


```
{
  "_id": {
    "$oid": "67b0db77ed9944851b01893d"
  },
  "Meta Data": {
    "Information": "Daily Prices (open, high, low, close) and
Volumes",
    "Symbol": "AAPL",
    "Last_Refreshed": "2025-02-14",
    "Output_Size": "Full size",
    "Time_Zone": "US/Eastern"
  },
  "Time Series (Daily)": [
    {
      "open": "241.2500",
      "high": "245.5500",
      "low": "240.9900",
      "close": "244.6000",
      "volume": "40896227",
      "date": "2025-02-14",
      "isodate": {"$date": "2025-02-14T00:00:00.000Z"}},
    {
      "open": "236.9100",
      "high": "242.3399",
      "low": "235.5700",
      "close": "241.5300",
      "volume": "53614054",
      "date": "2025-02-13",
      "isodate": {"$date": "2025-02-13T00:00:00.000Z"} ,
    { ... } ] }
```

Insider trading data storage

4 (Insider trading report)	2024-10-03	2024-10-01	Adams Katherine L. Apple Inc.
4 (Insider trading report)	2024-10-03	2024-10-01	WILLIAMS JEFFREY E Apple Inc.
4 (Insider trading report)	2024-10-03	2024-10-01	Maestri Luca Apple Inc.
4 (Insider trading report)	2024-10-03	2024-10-01	O'BRIEN DEIRDRE Apple Inc.
4 (Insider trading report)	2024-10-03	2024-10-01	COOK TIMOTHY D Apple Inc.

(Street) CUPERTINO CA 95014
(City) (State) (Zip)

4. If Amendment, Date of Original Filed (Month/Day/Year)

6. Individual or Joint/Group Filing (Check Applicable Line)
X Form filed by One Reporting Person
Form filed by More than One Reporting Person

Table I - Non-Derivative Securities Acquired, Disposed of, or Beneficially Owned

1. Title of Security (Instr. 3)	2. Transaction Date (Month/Day/Year)	2A. Deemed Execution Date, if any (Month/Day/Year)	3. Transaction Code (Instr. 8)		4. Securities Acquired (A) or Disposed Of (D) (Instr. 3, 4 and 5)			5. Amount of Securities Beneficially Owned Following Reported Transaction(s) (Instr. 3 and 4)	6. Ownership Form: Direct (D) or Indirect (I) (Instr. 4)	7. Nature of Indirect Beneficial Ownership (Instr. 4)
			Code	V	Amount	(A) or (D)	Price			
Common Stock	10/01/2024		M		127,282	A	(1)(2)	617,226	D(3)	
Common Stock(4)	10/01/2024		F		67,552	D	$226.21	549,674	D(3)	
Common Stock(5)	10/02/2024		S		51,674	D	$226.8(6)	498,000	D(3)	
Common Stock(5)	10/02/2024		S		8,056	D	$227.22(7)	489,944	D(3)	

Table II - Derivative Securities Acquired, Disposed of, or Beneficially Owned (e.g., puts, calls, warrants, options, convertible securities)

1. Title of Derivative Security (Instr. 3)	2. Conversion or Exercise Price of Derivative Security	3. Transaction Date (Month/Day/Year)	3A. Deemed Execution Date, if any (Month/Day/Year)	4. Transaction Code (Instr. 8)		5. Number of Derivative Securities Acquired (A) or Disposed of (D) (Instr. 3, 4 and 5)		6. Date Exercisable and Expiration Date (Month/Day/Year)		7. Title and Amount of Securities Underlying Derivative Security (Instr. 3 and 4)		8. Price of Derivative Security (Instr. 5)	9. Number of derivative Securities Beneficially Owned Following Reported Transaction(s) (Instr. 4)	10. Ownership Form: Direct (D) or Indirect (I) (Instr. 4)	11. Nature of Indirect Beneficial Ownership (Instr. 4)
				Code	V	(A)	(D)	Date Exercisable	Expiration Date	Title	Amount or Number of Shares				
Restricted Stock Unit	(1)(2)	10/01/2024		M			127,282	(8)(9)(10)(11)	(8)(9)(10)(11)	Common Stock	127,282	(1)(2)	0	D	

Explanation of Responses:

Open document Open filing Close

Example with Apple company:

The scraper goes throuh each document on the website and extract **all the raw XML content** as in the second image. Then construct a dictionary along with other key informations such as stock, date and entity and write it to the DB collection.

Each document is compiled by a single entity. Inside a single document **more transactions** can be registered for the entity.

Moreover **multiple documents** compiled by different people can refer to the **same transaction date** (see the red rectangle).

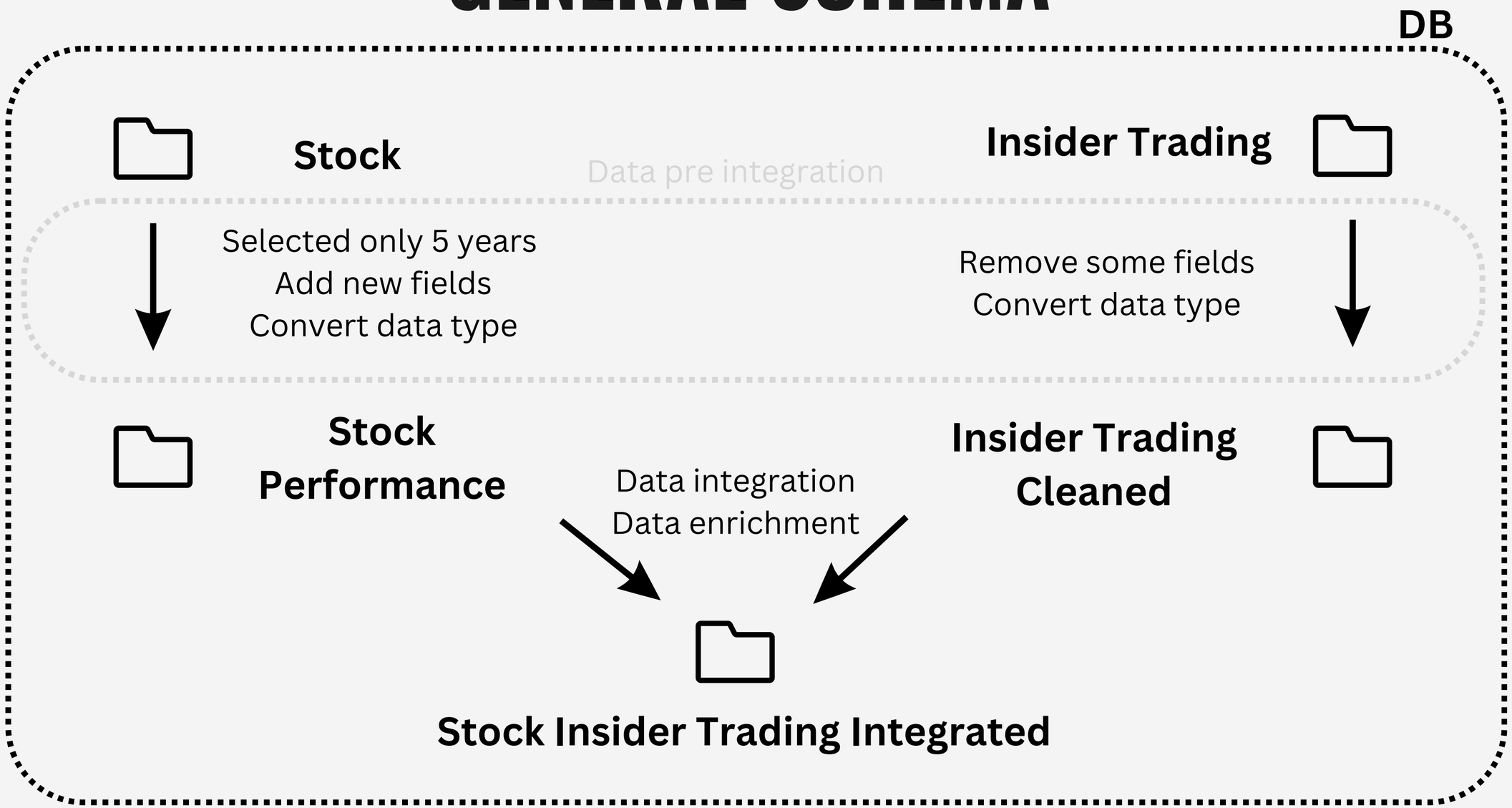
This resulted in a total of 18,914 documents in this collection. The structure of the document is quite complex due to the fact the XML behind the document is stored directly without processing. Filters will be applied in a second step.

Insider trading data storage

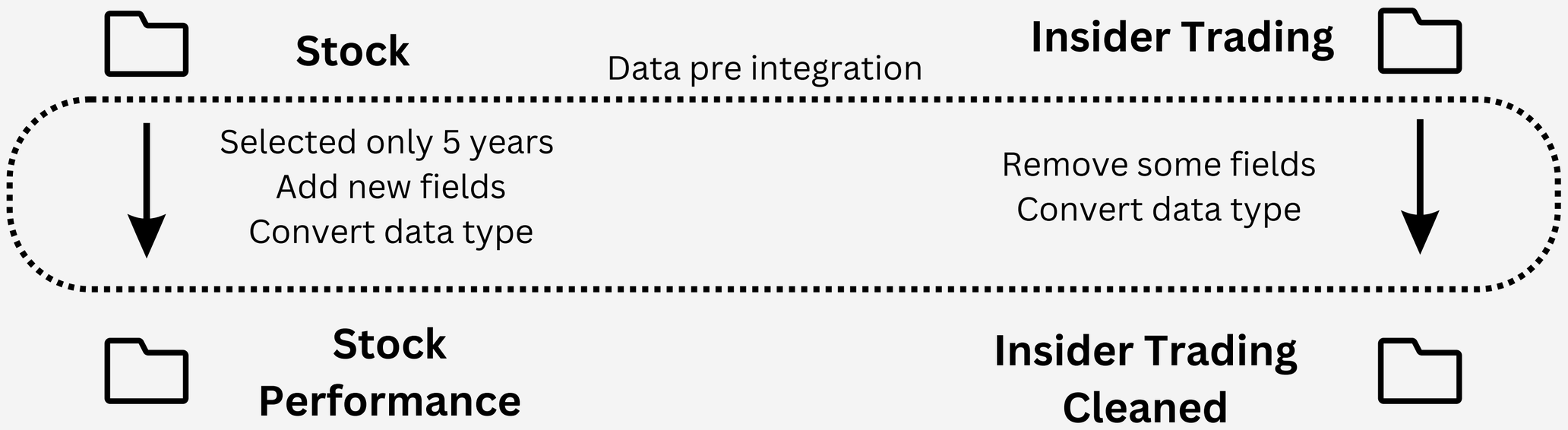

```
_id: ObjectId('67abdf7162076cdbfe412728')
Date_publication : "2024-10-03"
Date_transactions : "2024-10-01"
Entity_name : "WILLIAMS JEFFREY E
              Apple Inc."
Company : "AAPL"
▾ Filling_data : Object
  ▾ ownershipDocument : Object
      schemaVersion : "X0508"
      documentType : "4"
      periodOfReport : "2024-10-01"
      notSubjectToSection16 : "0"
    ▸ issuer : Object
    ▸ reportingOwner : Object
      aff10b5One : "1"
    ▾ nonDerivativeTable : Object
      ▸ nonDerivativeTransaction : Array (4)
    ▾ derivativeTable : Object
      ▸ derivativeTransaction : Object
    ▸ footnotes : Object
      remarks : null
    ▸ ownerSignature : Object
```


This is one of the 18k documents collected, in the raw version. The same in the previous example. The 4 transactions are stored automatically in an array.

GENERAL SCHEMA



Data pre integration



Added price return performances. Plus volume changes


```
"change_close_next_1_day": 0.75,
"change_close_next_5_days": -2.02,
"change_close_next_15_days": 2.17,
"change_volume_next_1_day": -4.56,
"change_volume_next_5_days": -16.06,
"change_volume_next_15_days": -58.3
```


Creation of Indexes → Company + Date

Preserving Raw data for future use

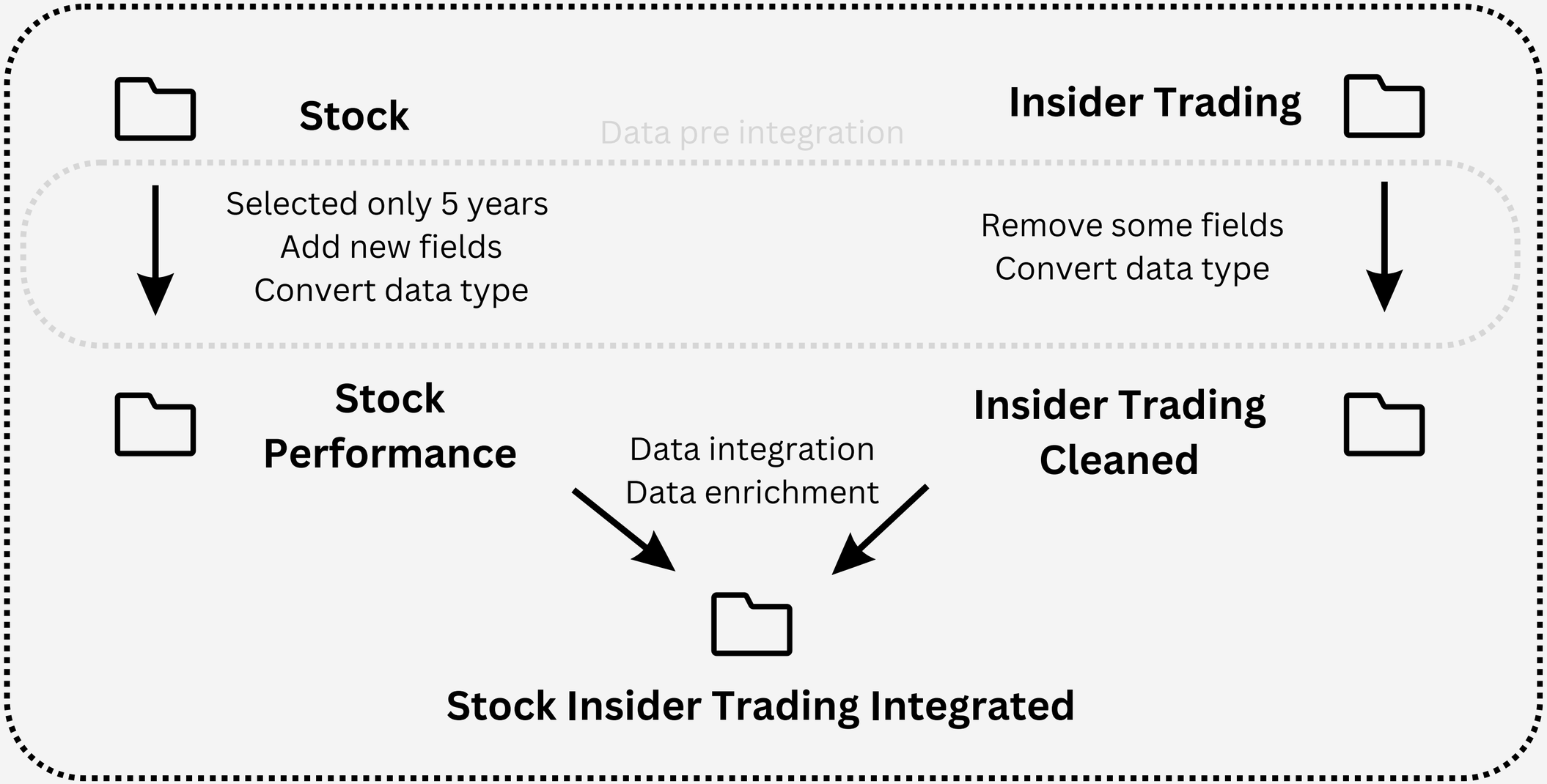
Converting data type → String -> Number

Insider trading data cleaning

After the storage of this raw data a more suitable and light collection is created:

- We keep only the documents that have form type = 4
- We remove all the documents that have no information in the NON-derivative table.
- Inside this tables we keep only the transactions that below to category S (sell) P (purchase).
- We extract only the entity name part of the string, separeting from its role.
- We make an the entity role field in each document.
- Conversion of date and values in respective type from strings.
- After these filters we uniform to keep all the object as an array (eventually composed only of one element).
- Finally we remove all documents with empty/null array of transactions.
- We reduce the collection to 7798 documents.

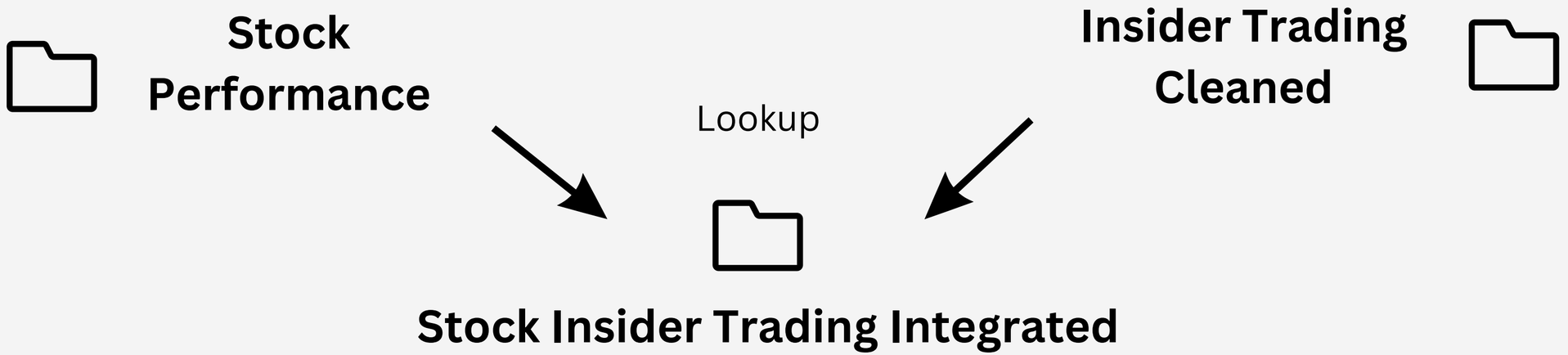
DATA INTEGRATION



Data integration

To merge the datasets, we performed a **lookup** operation between Insider Trader Cleaned and Stock Performance. The integration was based on the two commons key fields: **company and date.**

Setting these fields to **composite keys** in each sources enabled us to perform efficiently the lookup.
It is worth to mention that MongoDB lookup operation **automatically** deal with multiple match by **append** another element to the list of transaction for the same day+company. Keep in mind that multiple people (i.e. multiple documents) can execute transactions on the same day for the same company.

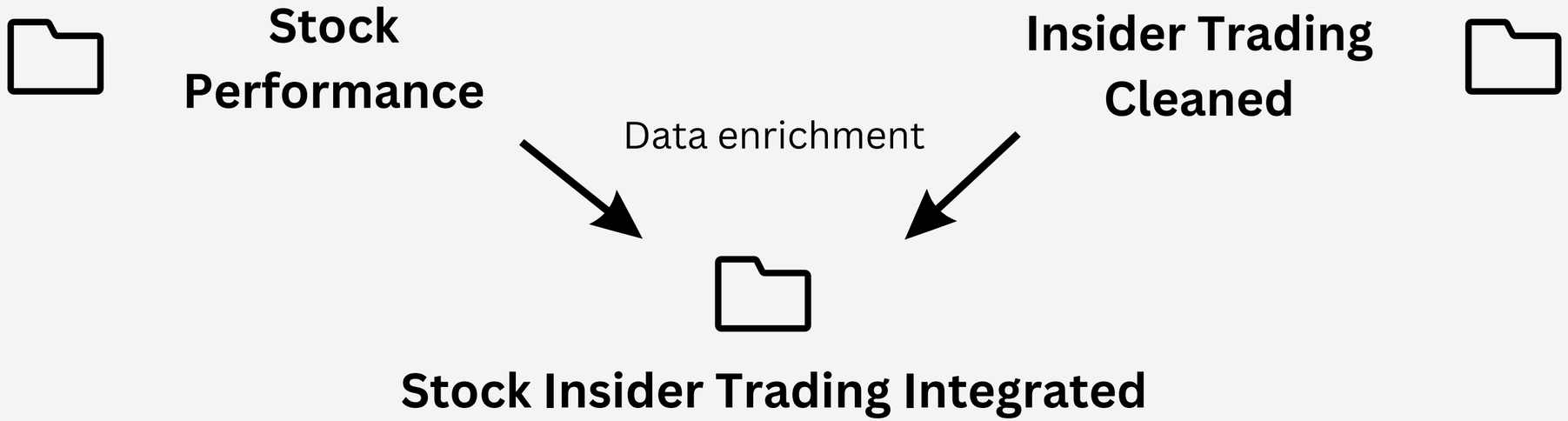


Data enrichment

These aggregated fields, added in the last phase of data integration provide further insights into the transactions and stock performance:

- **total entity transactions**: The total number of transactions per entity.
- **count acquired**: The number of transactions with the "A" code (acquisition) per entity.
- **count disposed**: The number of transactions with the "D" code (disposition) per entity.

- **total transactions**: The total number of transactions within the document, summing up the transactions for all entities involved in that date.



Transactions (**5**) made during 2021/04/05 to NVDIA

Stock Information and Performance

Transactions (**2**) made during 2021/04/05 to NVDIA by Perry Mark L

Transactions (**3**) made during 2021/04/05 to NVDIA by Maestri Luca


```
{
"_id": {"$oid": "67b34c5ee1025b242fd2522a"},
"symbol": "NVDA",
"$date": "2021-04-05T00:00:00.000Z","close": 559.5,"open": 554.7,
... ,
"change_close_next_1_day": -0.9, ... ,
"change_volume_next_1_day": -25, ... ,
"transactions_from_lookup": [
"$date": "2021-04-05T00:00:00.000Z",
"filtered_nonDerivativeTransaction": [ 0 [...],1 [...] ],
"Entity_name_cleaned": "PERRY MARK L",
"isDirector": "1",
"isOfficer": "0",
"isTenPercentOwner": "0",
"isOther": "0",
"officerTitle": null,
"otherText": null,
"total_entity_transactions": 2,
"count_acquired": 0,
"count_disposed": 2,
"$date": "2021-04-05T00:00:00.000Z",
"filtered_nonDerivativeTransaction": [ 0 [...],1 [...], 2[...] ],
"Entity_name_cleaned": "Luca Maestri",
"isDirector": "0",
"isOfficer": "1",
"isTenPercentOwner": "0",
"isOther": "0",
"officerTitle": null,
"otherText": null,
"total_entity_transactions": 3,
"count_acquired": 1,
"count_disposed": 2
],
"total_transactions": 5
}
```


Lookup “auto append”

DATA QUALITY

1. On Stock
2. On Insider Trading
3. On Integrated data

Data quality on Stock

Temporal completeness

This metric focused on checking for any missing **dates** in the stock price dataset.

For the 20-year period under analysis, **57.030** on 262.484 dates were found to be missing.
These missing dates align with weekends (Saturdays and Sundays), during which the stock market is closed.
No significant data quality issues were detected in this regard.

Data consistency

This metric checked whether the stock price values were consistent, specifically verifying that **the "high" value was greater than the "low" value** for each trading day.

```
"high": "245.5500",
"low": "240.9900",
```

All stock price data followed the expected pattern.

Data quality on Insider trading

Data transaction and publication consistency

The rule imposed by the SEC should implied that the **transaction date** must always precede the **publication date**

- Total checks performed: 18,914
- Valid checks: 18,246
- Failed checks: 668
- Checks with 'NA' dates: 2

This shows that **3.53%** of the checks failed, meaning there were inconsistencies where the transaction date was not before the publication date.
However, the number of valid checks is high, indicating that the majority of the data is consistent.

Data quality on Intregrated Data

Completeness

Counting how many fields have ‘null’ values

7972 values are null on 1282676 values evaluated, only **0. 02%.**

Price range consistency

Purchase price of insider traders (”transactionPricePerShare”) falls within the **high/low** range for the corresponding stock in that day.

```
"high": "245.5500",
“transactionPricePerShare”: “241.5590”
"low": "240.9900",
```

! This metric helped us to:

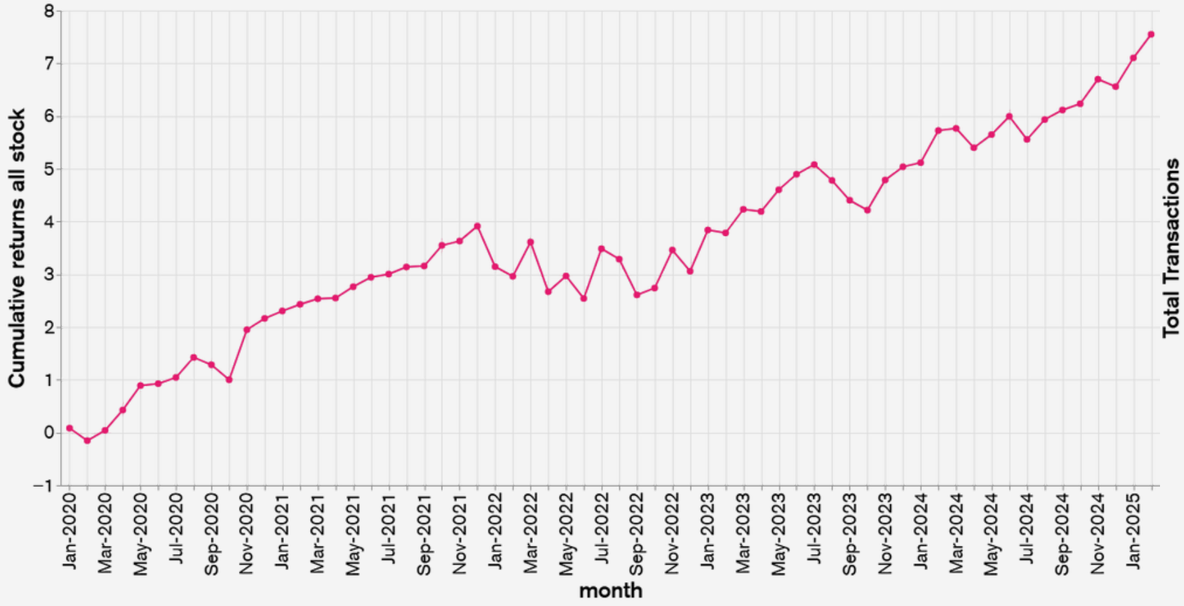
- Refine our initial filter selecting only relevant transaction codes for our analysis
- Spot a bug in the scraper as it was wrongly including some data

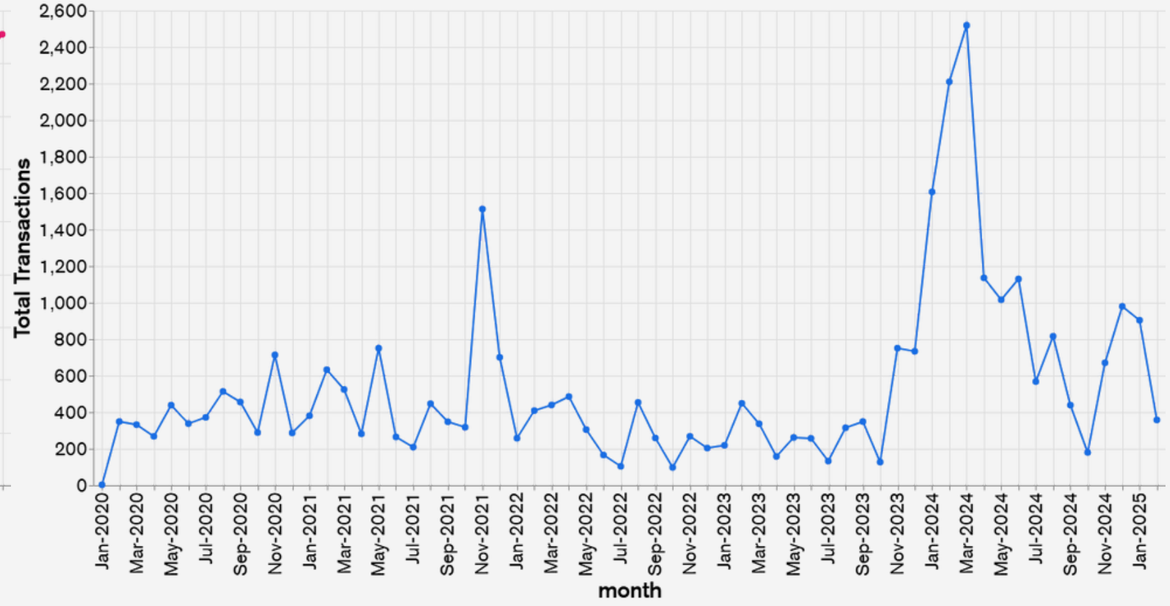
VISUALIZATIONS

In the end we used the cleaned data and the integrated dataset to discover the hidden structures of the data and find eventual relationships between the insider transactions and the stock price. Some results are in aggregate for all the stock, some are relevant for a particular stock.

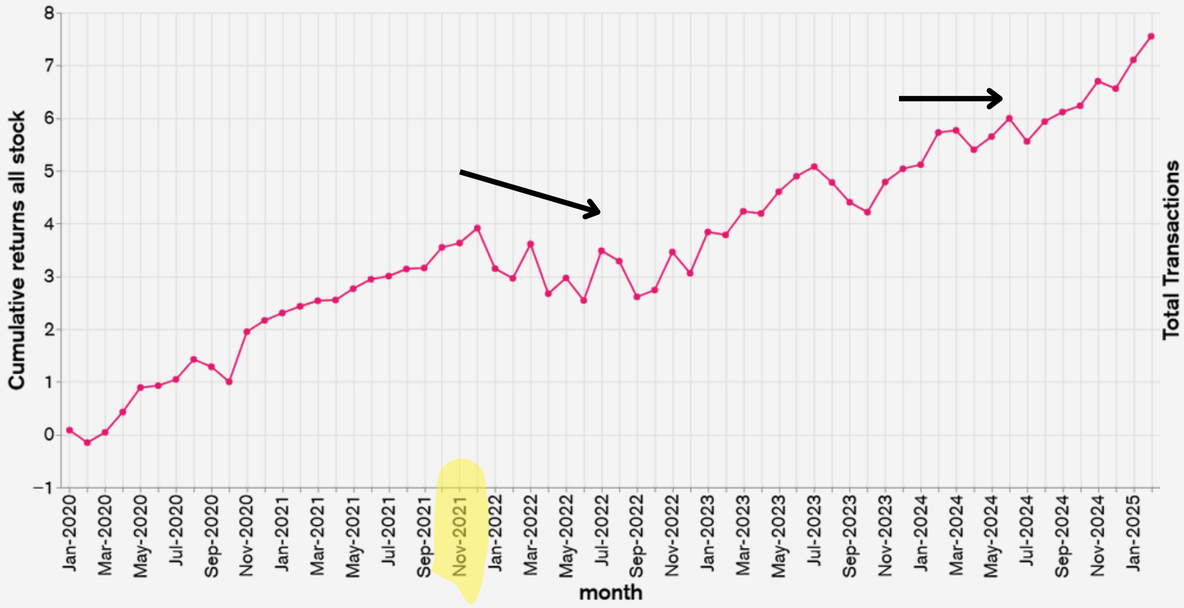
The insider can foreseen future price movements?

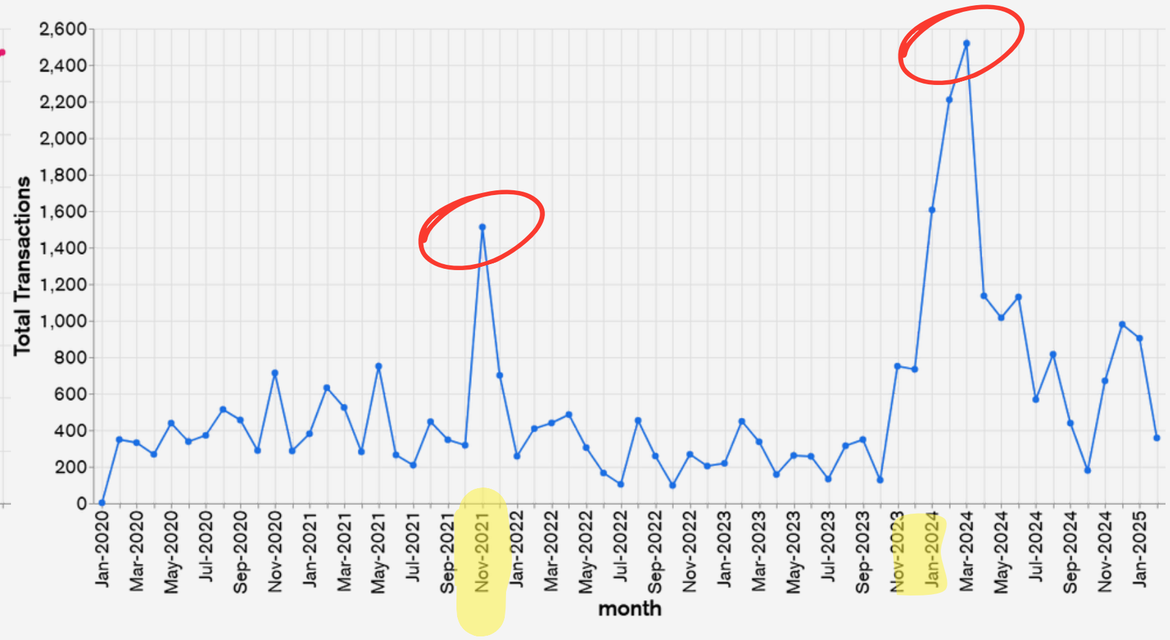
95% of the transactions are sell! So total transaction spike should implied a negative outlook on the price





The insider can foreseen future price movements?

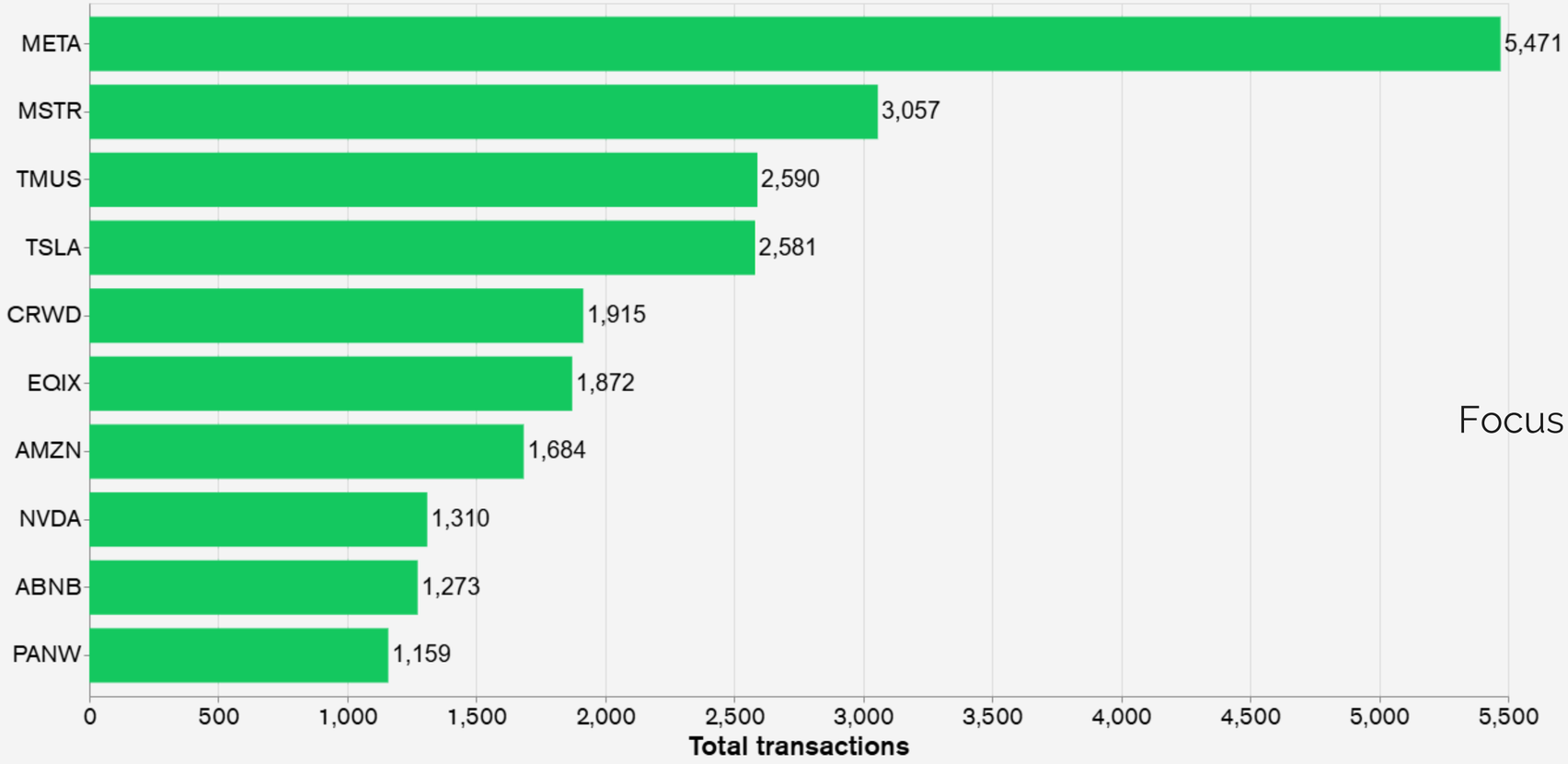




Focus on Crowdstrike and his CEO George Kurtz buys.

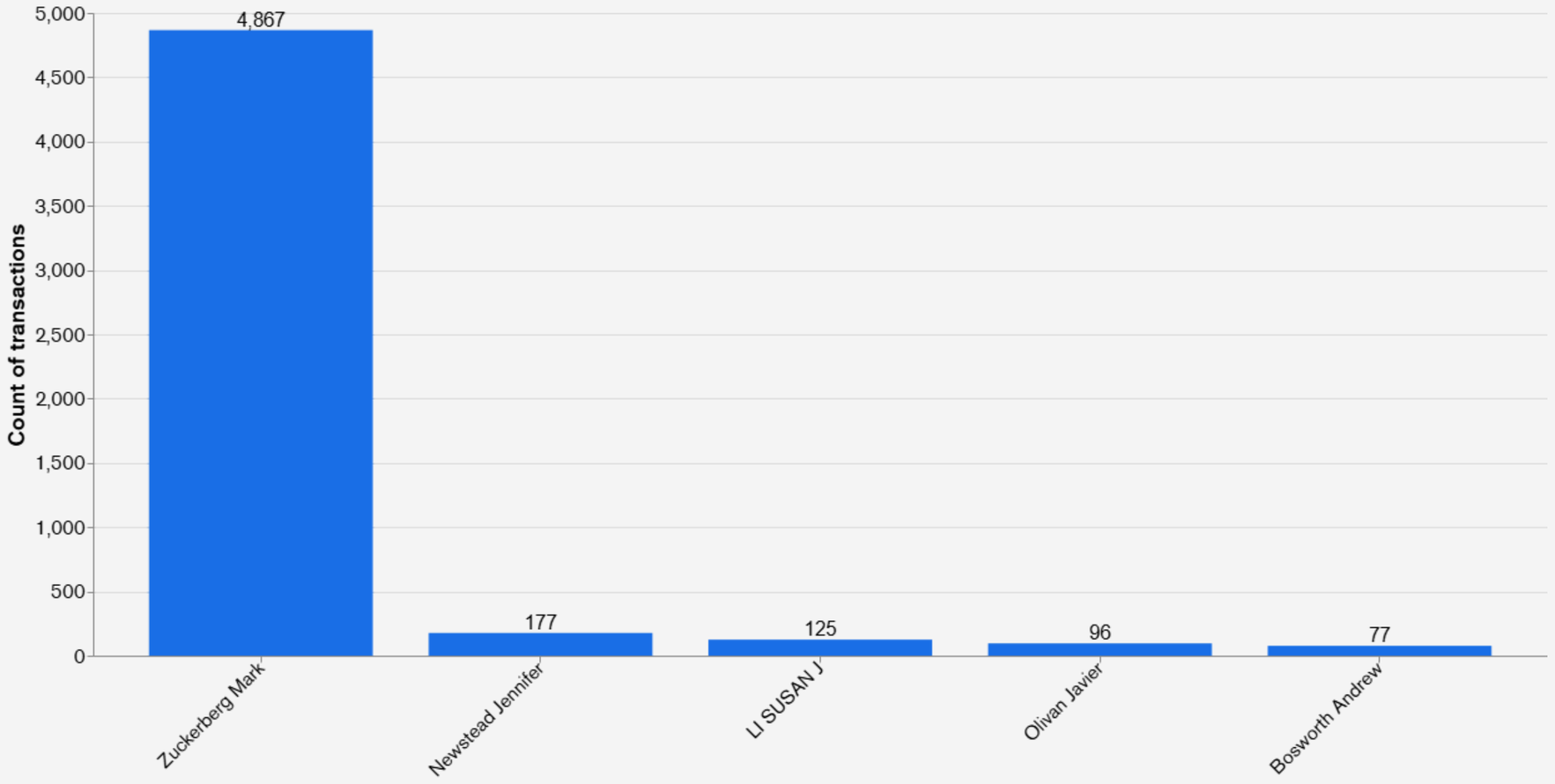
For this company, as we have some relevant **buy transactions**, we can inspect if the price and the buy transactions of its CEO are related. Especially at the start and at end of our period we can se some spike in the number of buy trade followed by an upward move of the price.

Which company made most transactions?



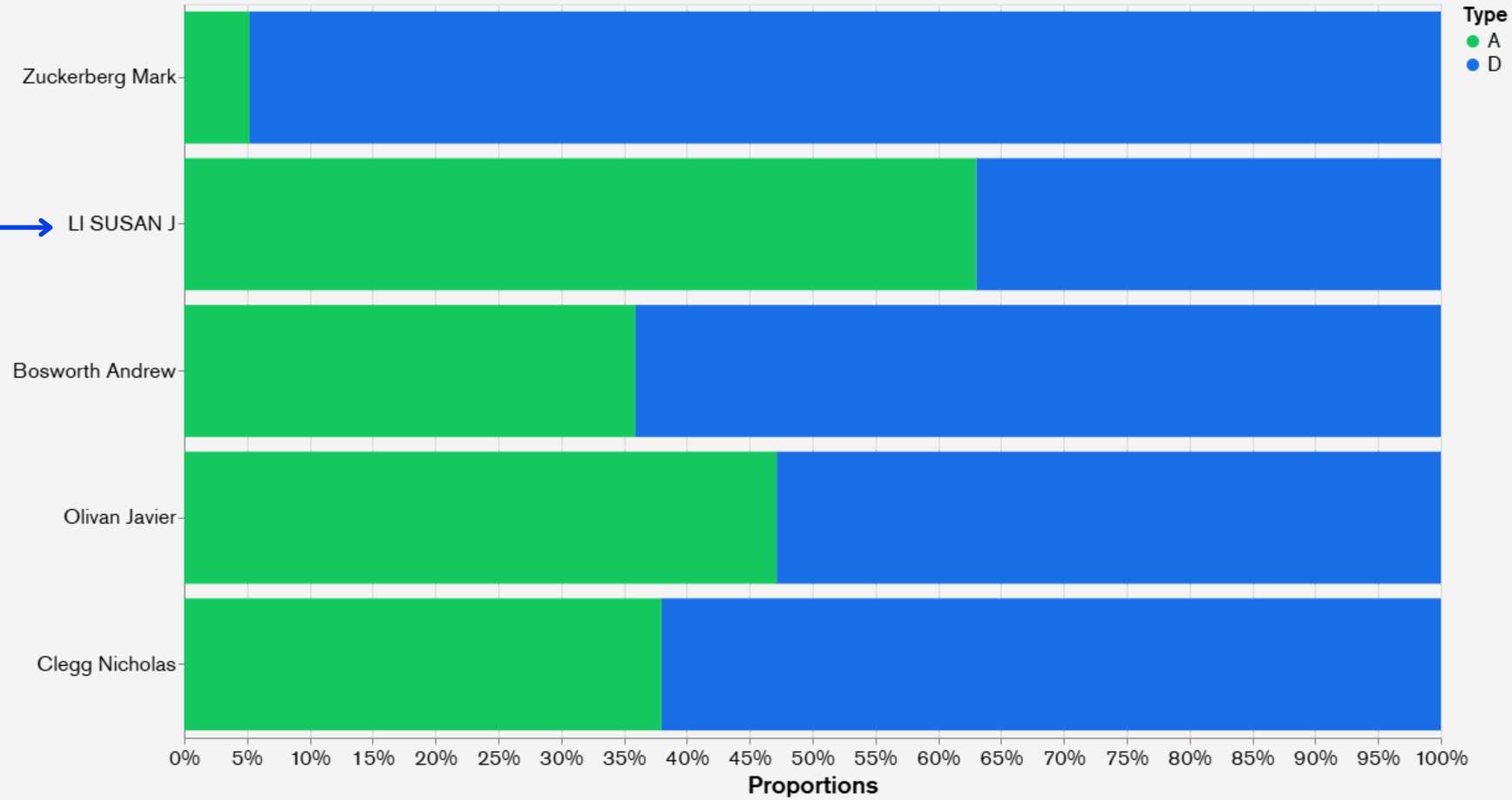
Focus on META

Who made most transactions in META?



META Insider roles - Which side of transactions?

mostly buys



CONCLUSIONS

From a broad perspective, our analysis suggests that insider trading may have a measurable impact on stock performance.

In particular, we observed that certain periods exhibit a correlation between insider transactions and significant stock price flat or downside direction.
We show also a specific case where buys from a CEO lead to an increase in the price.

This suggests that insider trading could be a leading indicator of market movements, potentially offering **valuable insights for investors and analysts**.

We also see that the total transaction differ a lot between different company. We also discover that different role inside a company can potentially have different views about the company.

FUTURE IMPROVEMENTS

CONSIDERING ADDITIONAL FACTORS

Such as market conditions, industry trends or major economic events. Or the transactions derived from option contracts.

EXPADING THE DATASET

By including more stocks or a longer period of time.

STATISTICAL TEST AND ML MODELS

Could provide a clearer picture of insider trading patterns and their potential impact on financial markets

Thank you!