NASDAQ 100 Investment Analytics

DATASCI 205

Gupta, Rini | Nguyen, Gia | Pham, Bao | Shin, Jaekwang

04/17/2024

Introduction

Objective/Use Case

- Identify clusters of companies that move together
- Provide investors with actionable insights through stock grouping

Importance

- NoSQL databases offer flexibility for dynamic grouping and analysis.
- Real-time processing capabilities

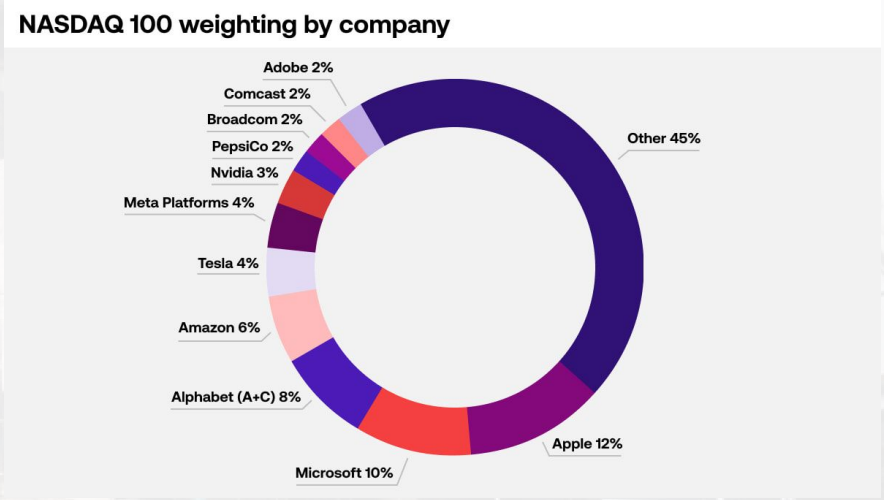
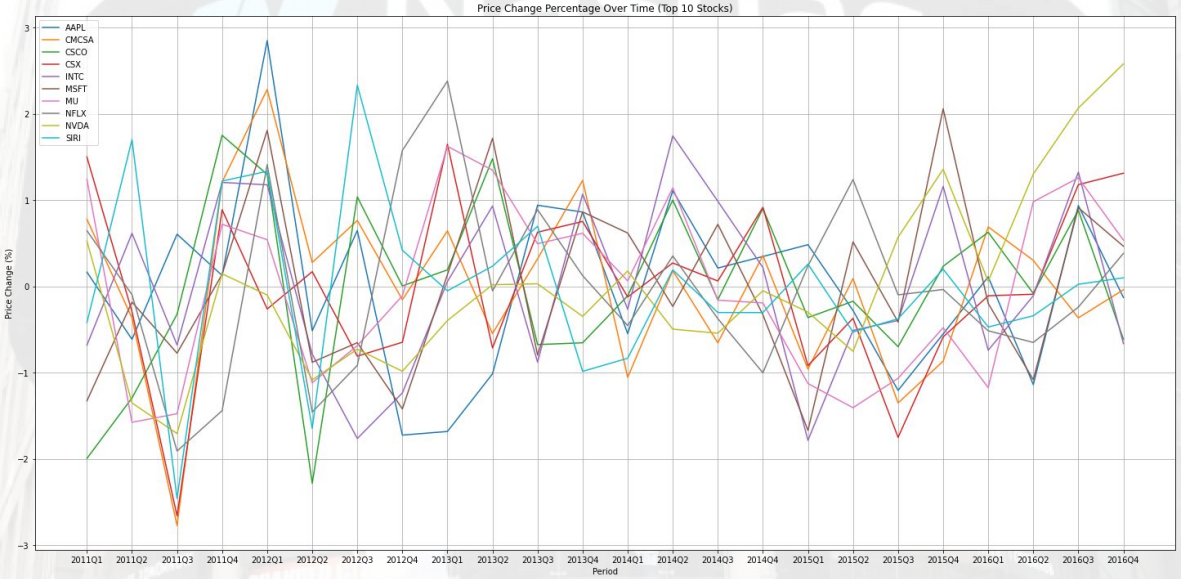


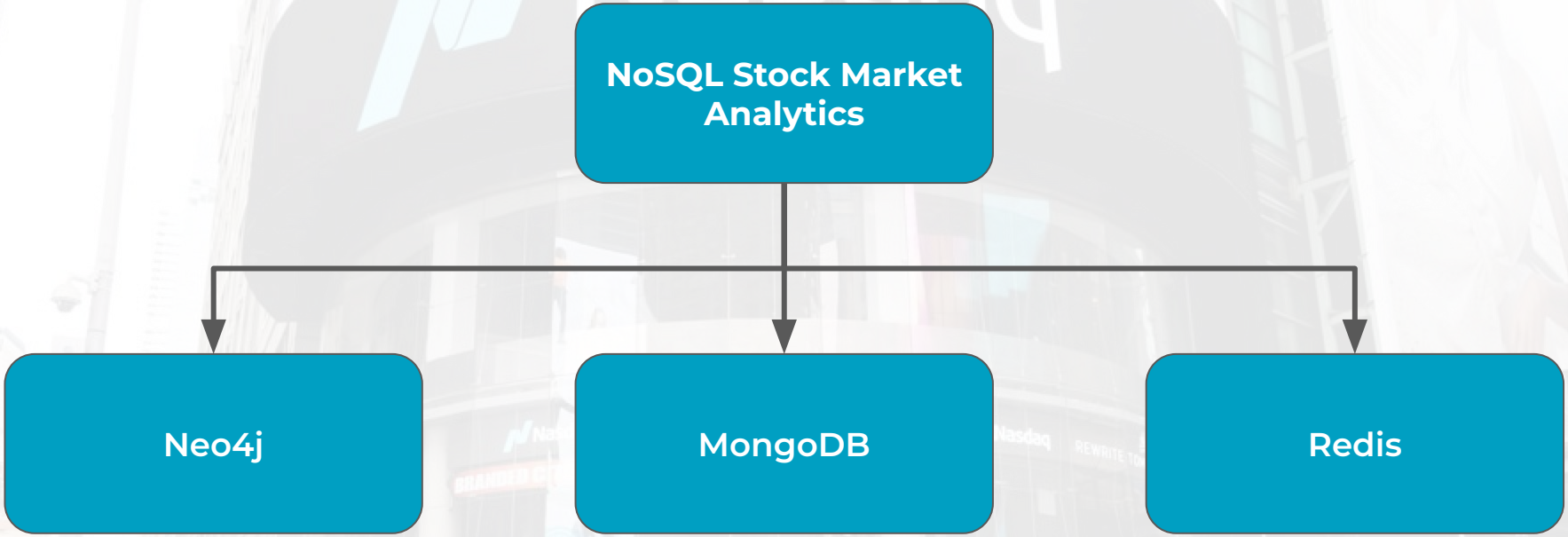
Image https://www.cityindex.com/en-uk/indices-trading/trading-the-nasdaq/

Limitations of Relational Database



- Visuals are messy
- Complex stock relationships
- Intensive data schema modifications
- Difficult horizontal scaling

Proposed Method



- Relationship Handling
- Graph-Based Queries
- Visual Data Exploration

- Flexible data ingestion
- High-Volume scalability

- Real-Time Performance
- High-Volume Scalability
- Efficient Data Processing

Dataset Overview

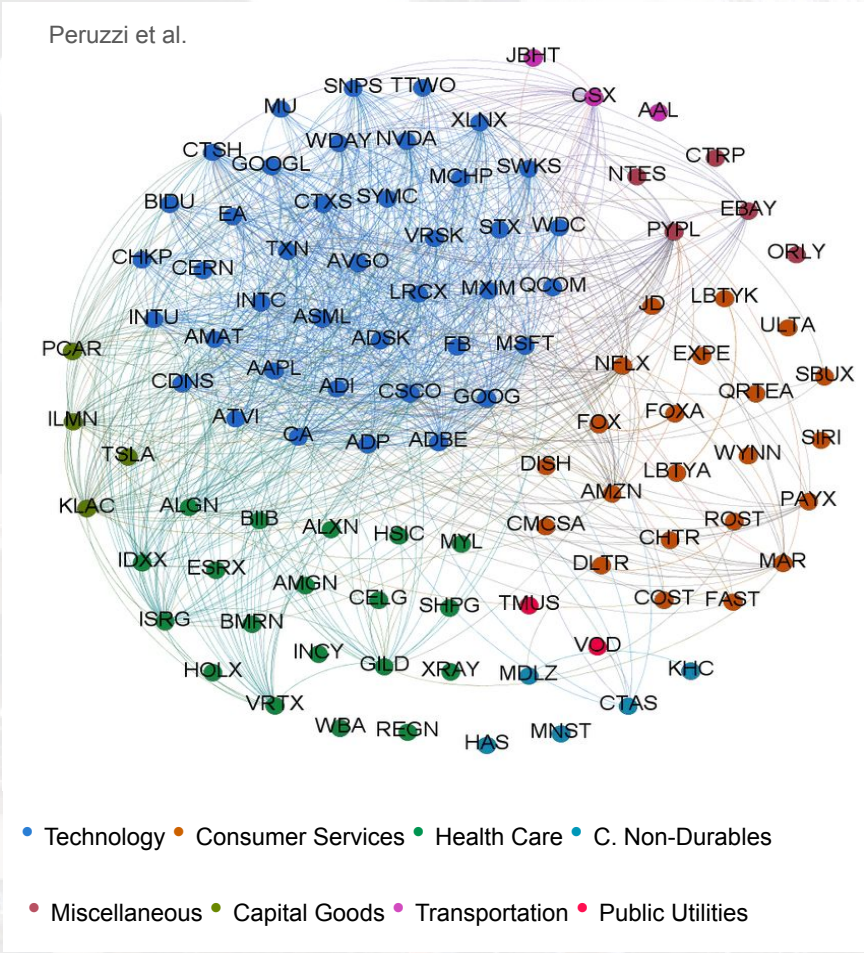
Focus: NASDAQ-100 Stock Price

Period: 2011 - 2016

Dimensions: 271680 observations, 8 columns

Data Example

Date	Open	High	Low	Close	Adj Close	Volume	Name
2010-01-04	7.62	7.66	7.59	7.64	6.56	493M	AAPL
2010-01-05	7.66	7.70	7.62	7.66	6.57	601M	AAPL
2010-01-06	7.66	7.69	7.53	7.53	6.47	552M	AAPL



Data Preprocessing & Analysis

Common Preprocessing

- Drop stocks that have missing data based on NASDAQ-100 holiday calendar
- Computed “Price Change ($)” and “Price Change (%)” as additional features
- Standardize each stock to ensure that the standardization is relative to each stock’s distribution of prices

Pearson Correlation

- Map each stock to a sector

Similarity Matrix

- Group dates by quarters to reduce dimensions

Data Preprocessing & Analysis

Cause	Observations for analysis	Removed number samples
Start	271,680	N/A
Filter data from 2011-2016	134,327	137,353
Missing data	128,350	5,977

Pearson Correlation Coefficient
w/ Exponential Weights

1. Emphasize recent trends – applying exponential weights put more emphasis on recent data points

2. Enhanced sensitivity to changes – quickly detect shifts in correlation due to recent events

3. Customizable time focus – flexibility in analyzing short-term, mid-term or long-term opportunities/strategies

$$\rho_{X,Y}^{w} = \frac{\mathrm{Cov}_w(X,Y)}{\sqrt{\sigma_{X,w}^2 \cdot \sigma_{Y,w}^2}}$$

	stock1	stock2	sector	correlation
0	AAPL	ADBE	Technology	0.575089
1	AAPL	ADI	Technology	0.487113
2	AAPL	ADP	Technology	0.494167
3	AAPL	ADSK	Technology	0.440092
4	AAPL	AMAT	Technology	0.520039

***NOTE:** Correlation is computed between pairs of stock within the same sector

Cosine Similarity

1. Insensitive to magnitude of vectors but focuses on directional change
2. Good at identifying stocks with similar price movements for grouping
3. Robust to noisy data/outliers

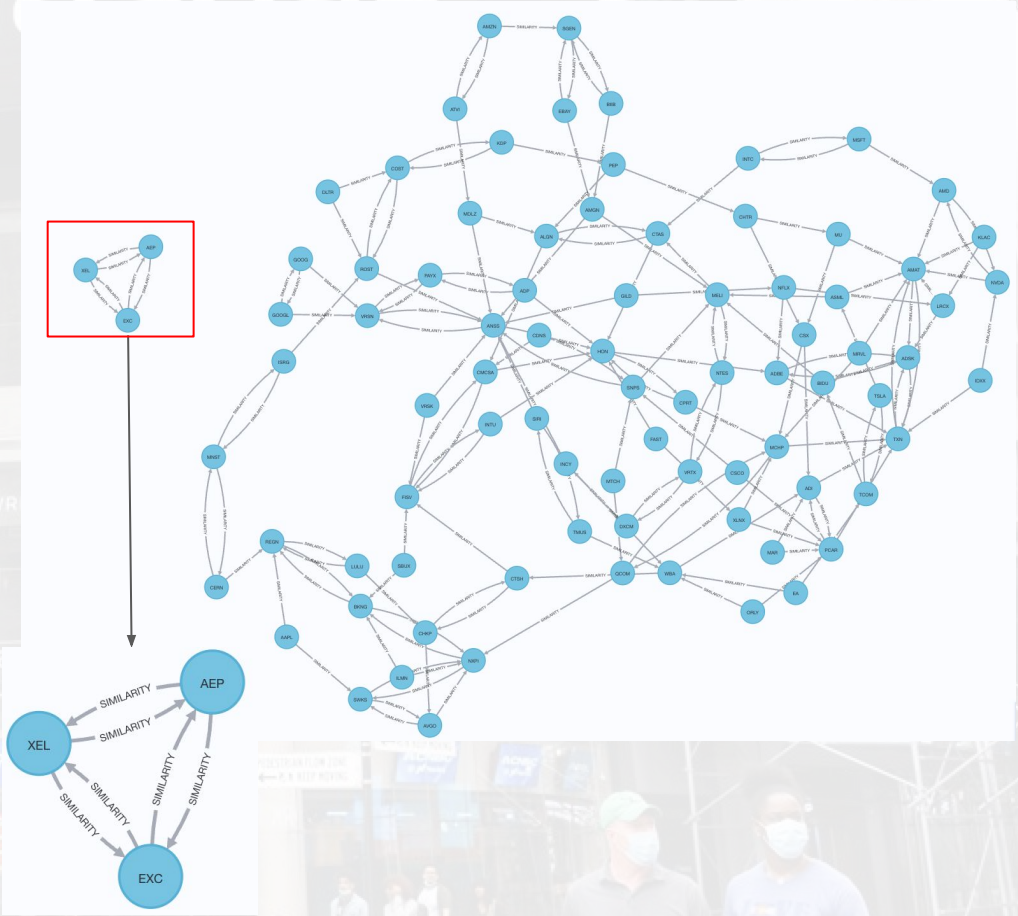
$$\cos(\mathbf{A}, \mathbf{B}) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Name	AAPL	ADBE	ADI	ADP	ADSK	AEP
AAPL	1.000000	0.124657	0.262914	-0.033264	0.410524	-0.222930
ADBE	0.124657	1.000000	0.637094	0.219216	0.763786	-0.039407
ADI	0.262914	0.637094	1.000000	0.470291	0.591171	0.023633
ADP	-0.033264	0.219216	0.470291	1.000000	0.295913	0.391271
ADSK	0.410524	0.763786	0.591171	0.295913	1.000000	-0.079660

Neo4j: Louvain Modularity (Cosine Similarity)

- Market segmentation
- Analyze risk profile of a group of stocks
- Higher potential for returns

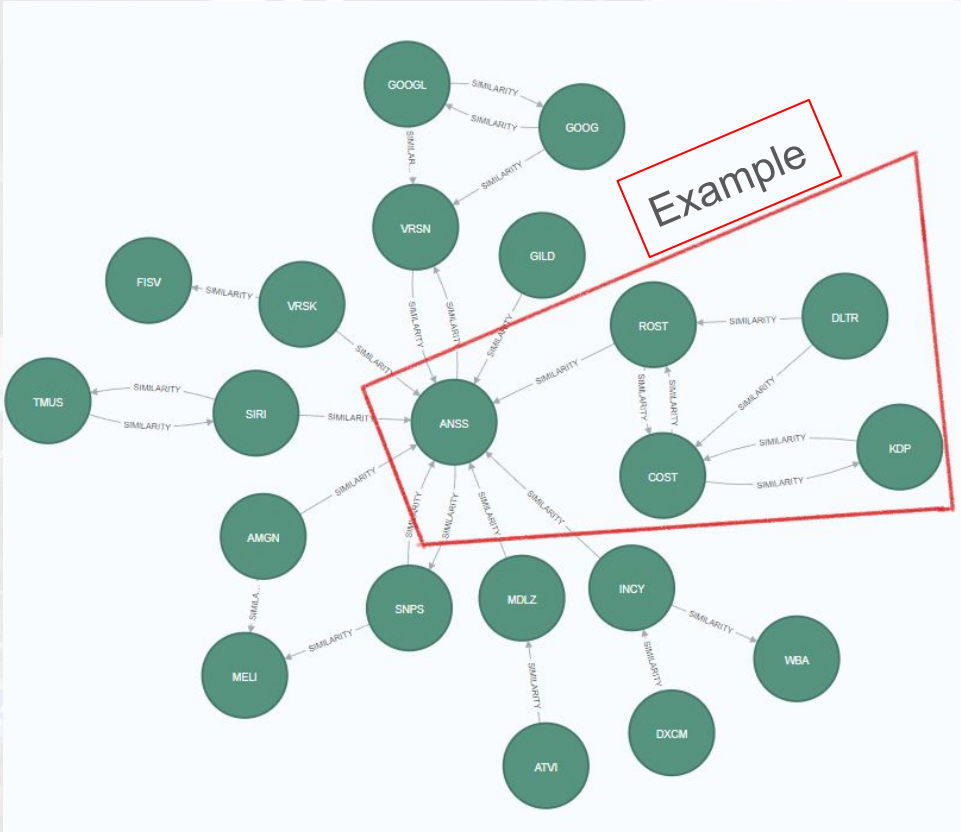
Community	Stocks
93	XEL, AEP, EXC
98	AMZN, ATVI
103	EA, TCOM, TSLA
104	BIIB, EBAY, SGEN
108	WBA, AAPL, BKNG, CHKP, CTSH, INCY, LULU, ORLY,...
112	ALGN, CDNS, CMCSA, CPRT, CTAS, FAST, FISV, HON...
126	GOOG, GOOGL
134	COST, DLTR, KDP, ROST
142	CERN, ISRG, MNST
149	AMD, IDXX, INTC, MSFT, NVDA
162	XLNX, AVGO, CHTR, CSX, ILMN, MCHP, MRVL, MU, N...
164	SIRI, TMUS
166	ADBE, ADI, ADSK, AMAT, KLAC, LRCX, MAR, NFLX, ...
168	ADP, ANSS, CSCO, GILD, MDLZ, MTCH, PAYX, VRSK,...
169	AMGN, ASML, BIDU, DXCM, MELI, NTES, SNPS, VRTX



Neo4j: PageRank (Cosine Similarity)

- Identify stocks that are most influential
- Quantifies importance of a stock

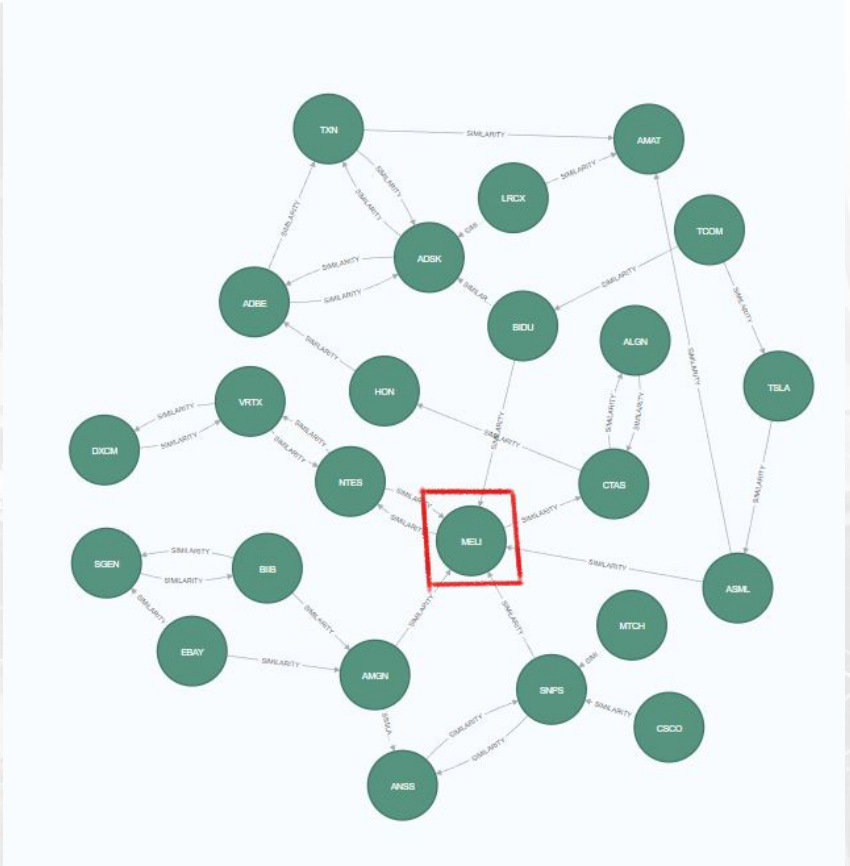
Name	PageRank
ANSS (#1)	1.173137
COST (#21)	1.023733
ROST (#22)	1.023718
KDP (#44)	0.975593
DLTR (#72)	0.950000



Neo4j: Betweenness (Cosine Similarity)

- Identify stocks that act as intermediaries or bridges between sectors
- With high betweenness, stock can cause ripple effects on market (momentum trading)

Name (Top 5)	Betweenness
MELI (#1)	567.0
HON (#2)	500.0
MCHP (#3)	405.0
NXPI (#4)	378.0
ANSS (#5)	342.0



MongoDB: Agile Financial Data Management

MongoDB Use Case: We plan to leverage document-oriented structure for agile data adaptation and analysis.

Limitation of Relational Database

- Inflexible schema
- Horizontal scaling challenges

High Volume Scalability

Flexible Data Ingestion

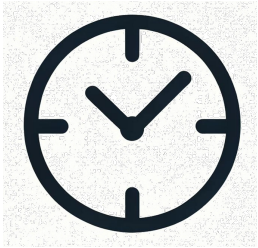
Redis: Enhancing Stock Trading Platforms

Redis Use Case: Stock trading platforms requiring real-time stock prices and fast transactional operations to provide info for traders and process trades quickly

Limitation of Relational Database

- Prioritize data integrity over speed - could lead to delays
- Difficult to scale horizontally - difficulty in handling large amounts of data

Real-Time Performance



High-Volume Scalability

Efficient Data Processing

Takeaways

- The **NASDAQ-100 stock** dataset (2011-2016) includes 271,680 observations with details on date, high, low, close, adjusted close, volume, and name.
- **Neo4j** excels in visualizing complex relationships and interconnected data points
- **Neo4j** supports real time data analysis, crucial for stock market environment since prices fluctuate rapidly
- **MongoDB** effectively handles different data formats and scales horizontally.
- **Redis** outperforms relational databases in fast-paced trading environments by offering superior real-time performance, scalability, and efficient data processing.

Q&A