NYU

Distributed Financial Risk Assessment

Group 007

TEAM

Anirudh Garg - ag9563
Nikhil Kommineni - nk3853
Sai Preetham Bojja - sb9509
Sujana Maithili Chindam - sc10648

12.10.24

Abstract

Objective: Develop a distributed framework for financial risk assessment.
Technology: Utilize Spark's distributed computing capabilities for efficient parallelization.
Scale: Process large-scale market data by simulating millions of scenarios.
Outcome: Generate robust and comprehensive risk metric VaR.

All computations are run on the NYU dataproc cluster hosted on GCP.

Motivation

Challenge: Traditional risk models lack scalability and real-time adaptability.

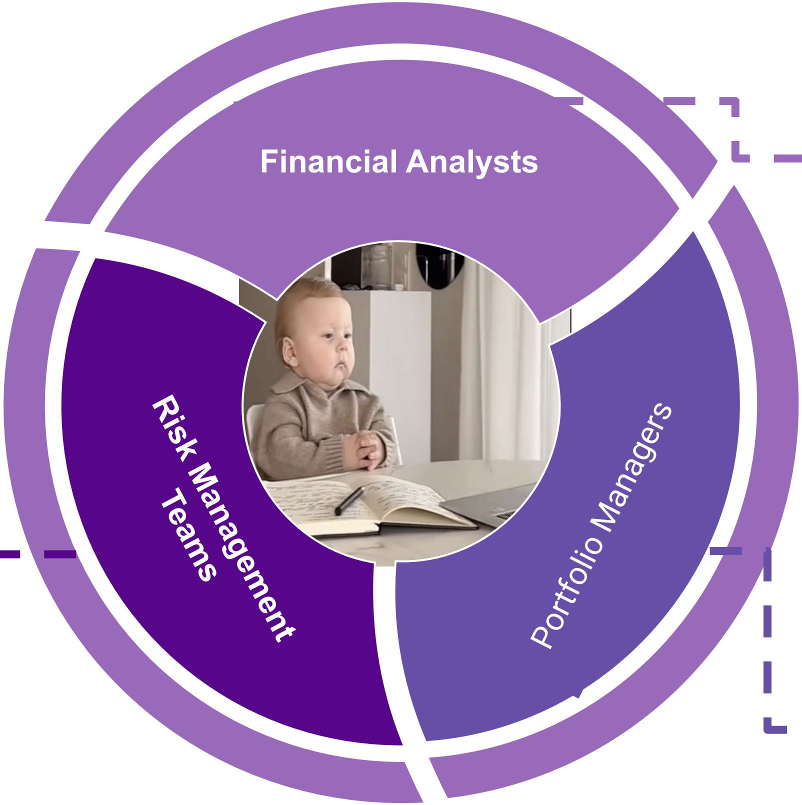
Need: Accurate, scalable tools for complex financial markets.

Our work aims to:

- Improve precision of risk analysis
- Provides actionable insights into portfolio performance and strategic decision-making
- Aid stakeholders to navigate uncertainties with confidence

NYU

Users and Intended Beneficiaries



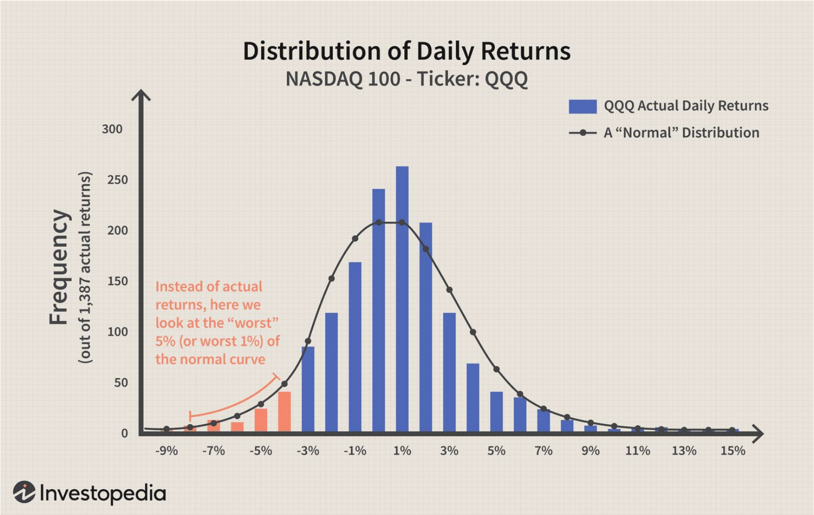
Utilize the system to assess portfolio risk and generate insights for investment strategies.

Use the framework to estimate capital reserves and design risk mitigation strategies.

Leverage VaR and CVaR metrics to optimize asset allocation and ensure risk-adjusted returns.

Value at Risk (VaR)

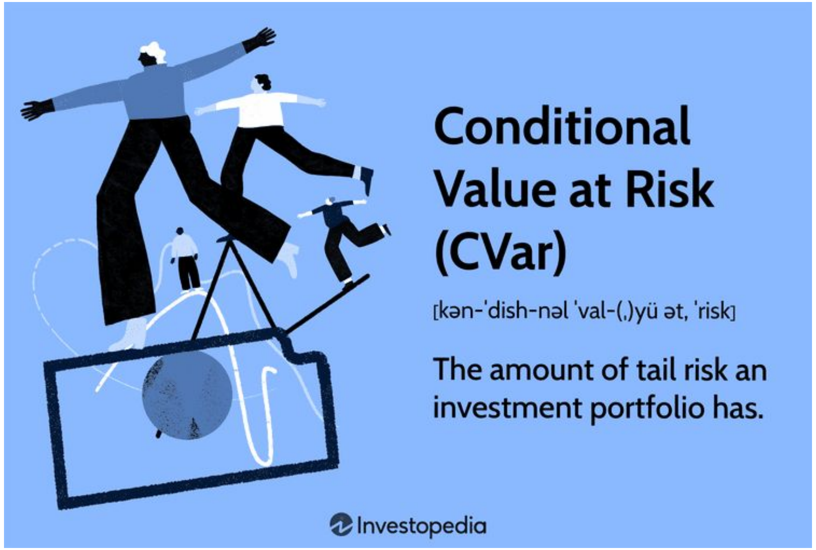
- **A financial risk metric that estimates the maximum potential loss of an investment or portfolio over a specified time period at a given confidence level.**
- **95% VaR of $1 million means there is a 95% chance that losses will not exceed $1 million within the defined time frame.**



Conditional Value at Risk (CVaR)

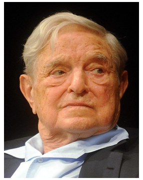
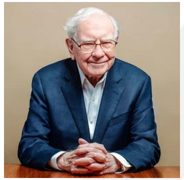
- Conditional Value at Risk (CVaR), also known as Expected Shortfall, measures the average loss in scenarios where the loss exceeds the Value at Risk (VaR) threshold.

- It provides a deeper insight into tail risk by focusing on the severity of extreme losses beyond the VaR, making it a crucial metric for assessing downside risk.



NYU

Our Approach

Data Collection	Data Cleaning	Analytics	Insights
We collect historical financial data from a reliable source like Yahoo finance using their official python library called “yfinance”.	We identify missing data points and any other inconsistencies in the collected data. Employ different extrapolation techniques to fill in missing values in the time series data.	We run Spark jobs to compute VaR and CVaR over selected stocks using a curated list of market factors. Run backtesting with different choices for market factors.	 VAR: 0.84%  VAR: 5.4%  VAR: -4.92%

NYU

DataSources

- We'll run a script for extracting data using **"wget"** command over stock symbols and google finance website.
- The stock symbols will correspond to stocks listed in US, India, Europe and Hong Kong stock markets. The list of the stocks in the countries can be extracted using the following links.
- All datasets have the same structure.

0 1 . NYSE

Size: 1.1 GB - 2800 symbols - 2000 to 2024

0 2 . NSE

Size: 870 MB - 2021 symbols - 2000 to 2024

0 3 . HKSE

Size: 693 MB - 1800 symbols - 2000 to 2024

0 4. LSE

Size: 601MB - 1720 symbols - 2000 to 2024

Datasets

Each dataset has two components:

a. Historical stock prices.
b. Historical Index prices.
   i. S&P 500
   ii. NASDAQ Composite
   iii. Treasury Yield 30-years (YYX)
   iv. Treasury Yield 5-years (FVX)
   v. Currency Exchange Value

Sample Stock Data

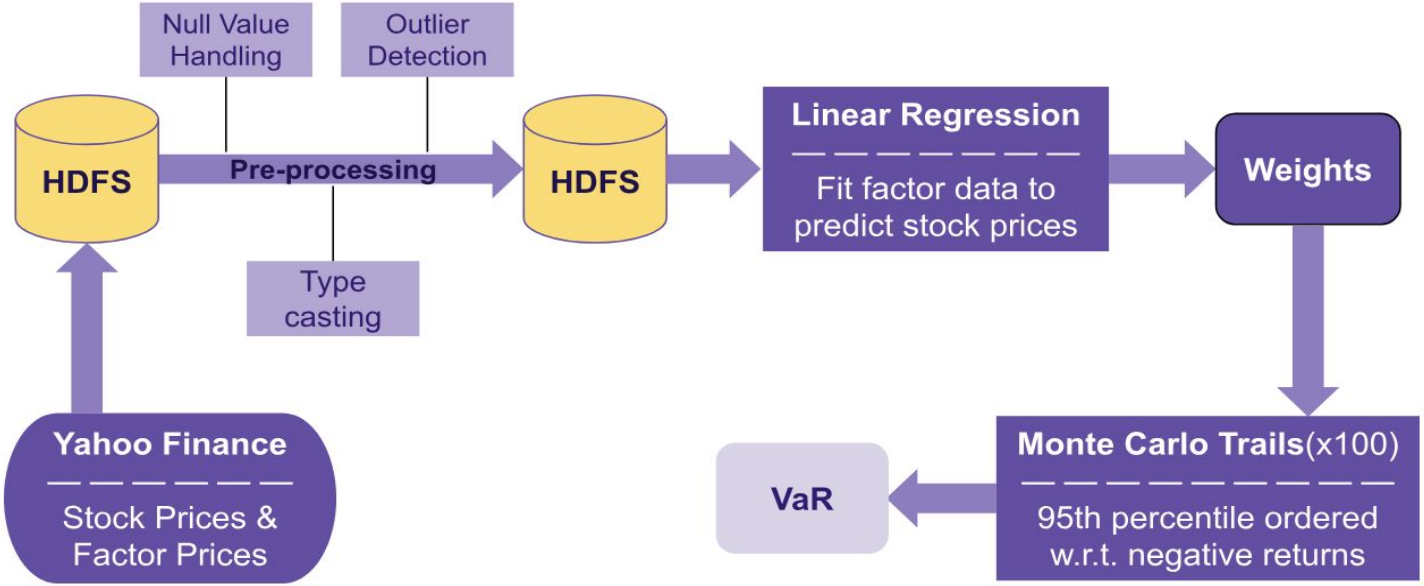
```
Date,Price,Adj Close,Close,High,Low,Open,Volume
2000-01-04,0.24795399606227875,1.2699999809265137,1.2699999809265137,1.1100000143051147,1.1100000143051147,17828000
2000-01-05,0.23819202184677124,1.2200000286102295,1.2699999809265137,1.100000023841858,1.100000023841858,20038000
2000-01-06,0.24990634620189667,1.2799999713897705,1.4700000286102295,1.2200000286102295,1.2599999904632568,38700000
2000-01-07,0.2596683204174042,1.3300000429153442,1.440000057220459,1.309999942779541,1.3200000524520874,22032200
2000-01-10,0.31824010610580444,1.6299999952316284,1.6699999570846558,1.399999976158142,1.399999976158142,58458000
2000-01-11,0.35338321328163147,1.809999942779541,2.049999952316284,1.7000000476837158,1.7000000476837158,75700600
2000-01-12,0.3397164046764374,1.7400000095367432,1.9500000476837158,1.6699999570846558,1.7300000190734863,50996200
2000-01-13,0.32409730553627014,1.659999966621399,1.840000033378601,1.6399999856948853,1.7799999713897705,25455000
2000-01-14,0.40512168407440186,2.075000047683716,2.0999999046325684,1.7599999904632568,1.7799999713897705,91712700
2000-01-17,0.3885263204574585,1.9900000095367432,2.25,1.9800000190734863,2.125,63926000
2000-01-18,0.3611926734447479,1.850000023841858,2.0250000953674316,1.809999942779541,2.0,21847200
2000-01-19,0.3709547221660614,1.899999976158142,1.9600000381469727,1.7999999523162842,1.8799999952316284,15783000
2000-01-20,0.34557363390922546,1.7699999809265137,1.909999966621399,1.75,1.899999976158142,13668000
2000-01-21,0.35338321328163147,1.809999942779541,1.9700000286102295,1.75,1.7699999809265137,25104000
2000-01-24,0.35338321328163147,1.809999942779541,1.8899999856948853,1.7999999523162842,1.850000023841858,9426000
2000-01-25,0.3611926734447479,1.850000023841858,1.8799999952316284,1.7899999618530273,1.809999942779541,12570000
2000-01-26,0.4490504562854767,2.299999952316284,2.375,1.8600000143051147,1.8899999856948853,69305200
```

Sample Index Data

```
Date,Price,Adj Close,Close,High,Low,Open,Volume
2000-01-03,6.456999778747559,6.456999778747559,6.4730000495910645,6.410999774932861,6.410999774932861,0
2000-01-04,6.395999908447266,6.395999908447266,6.449999809265137,6.388000011444092,6.434000015258789,0
2000-01-05,6.488999843597412,6.488999843597412,6.488999843597412,6.395999908447266,6.427000045776367,0
2000-01-06,6.449999809265137,6.449999809265137,6.4730000495910645,6.427000045776367,6.442999839782715,0
2000-01-07,6.396999835968018,6.396999835968018,6.489999771118164,6.396999835968018,6.442999839782715,0
2000-01-10,6.459000110626221,6.459000110626221,6.473999977111816,6.443999767303467,6.443999767303467,0
2000-01-11,6.567999839782715,6.567999839782715,6.567999839782715,6.498000144958496,6.50600004196167,0
2000-01-12,6.60699987411499,6.60699987411499,6.614999771118164,6.552999973297119,6.576000213623047,0
2000-01-13,6.513999938964844,6.513999938964844,6.599999904632568,6.513999938964844,6.576000213623047,0
2000-01-14,6.568999767303467,6.568999767303467,6.5929999351501465,6.4679999351501465,6.5229997634887695,0
2000-01-18,6.625,6.625,6.639999866485596,6.5929999351501465,6.5929999351501465,0
2000-01-19,6.593999862670898,6.593999862670898,6.602002365112305,6.585999965667725,6.602002365112305,0
2000-01-20,6.632999897003174,6.632999897003174,6.6570000648498535,6.593999862670898,6.64900016784668,0
2000-01-21,6.650000095367432,6.650000095367432,6.697000026702881,6.633999824523926,6.650000095367432,0
2000-01-24,6.572000026702881,6.572000026702881,6.635000228881836,6.572000026702881,6.626999855041504,0
2000-01-25,6.579999923706055,6.579999923706055,6.5960001945495605,6.548999786376953,6.557000160217285,0
```

NYU

Data Pipeline



Results

Stock Exchange	Value at Risk (VaR)
London Stock Exchange (LSE)	- 3.7%
New York Stock Exchange (NYSE)	+ 2.73%
National Stock Exchange (NSE)	- 4.2%
Hong Kong Stock Exchange (HKSE)	+ 2.11%

Indian Stock Exchange (NSE) is least risky in the short-medium term future.

NYU

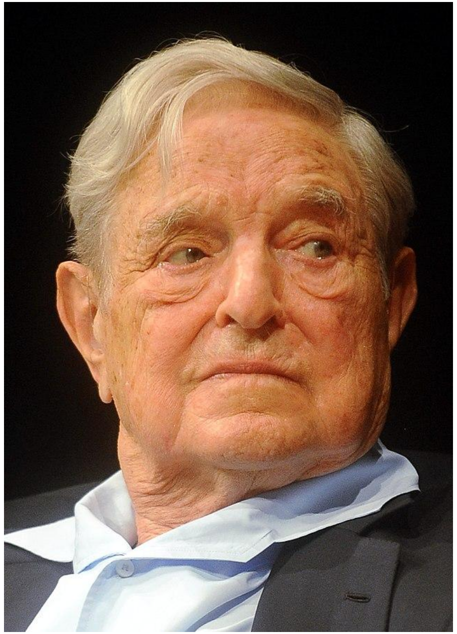
Warren Buffett

VaR: 4.92%

Bill & Melinda Gates

VaR: 5.4%

George Soros



VaR: 0.84%

Results

VaR of companies listed prior to 2000: 2.7%

VaR of companies listed in last 4 years: 5.2%

It is more risky to invest in young companies compared to the established ones.

Challenge

Aligning time series data with different trading calendars and handling gaps

Standardize Trading Calendars

Determine the trading calendars for each dataset and establish a unified timeline based on the most restrictive trading calendar.

Handle Non-Trading Days

Add rows for non-trading days and fill in missing values using Forward/Backward Filling and Interpolation.

Optimize for Large Datasets

Use Spark DataFrames and functions like window and lag/lead to handle missing values and align dates efficiently

Challenge

Curating the set of market factors can make or break a model

Focus on Relevant Factors

Use statistical methods and domain insights to identify factors that significantly impact portfolio performance.

Correlation and Multicollinearity

High correlation between factors can reduce model efficiency. Use PCA to select factors to reduce redundancy.

Iterative Testing

Continuously refine the set of factors based on simulation results and performance metrics.

NYU

Challenge

Analyzing private investment portfolio

01	Analyzing private investment portfolio	• Obtain their holdings and investment details • Obtain the respective weights of each stock in the portfolio
02	**Handle custom weights**	• Use custom weights instead of equal-weights

The VaR assesses the potential financial risk associated with individual investors

Goodness of our Results

- **Is the model internally consistent?**
  - Bootstrapping allowed us to get confidence intervals over the VaR by repeatedly sampling with replacement from the set of portfolio return results of our trials.
  - For HKSE, with high confidence we can say that VaR falls between **[0.021055, 0.021272]**.

- **How well does our model matches reality?**
  - The confidence interval does little to help us understand how well our model matches reality.
  - Kupiec's proportion-of-failures (POF) test considers how the portfolio performed at many historical time intervals and counts the number of times the losses exceeded the VaR.

Obstacles

01	Insufficient heap memory in Zeppelin	Switched to spark application mode on DataProc
02	Inconsistent behaviour during package creation	Needed to retry multiple times to overcome failures

ACKNOWLEDGEMENTS

- **Data Source:** We acknowledge Yahoo Finance for providing the financial data.
- **Computational Resources:** Our sincere thanks to the NYU High-Performance Computing (HPC) team for computational infrastructure
- **Academic Guidance:** We thank Prof. Yang Tang for his support throughout the course

Thank You