

# Case: Transforming data into insights

- ETL process of Financial data

# Introducing me

- Martin Birk Andreassen (M28)
- Data Scientist at STAR
- Interests:
  - NLP, ML and Econometrics
  - Data Science and Engineering
  - Running and CrossFit
- Economist (Cand.polit)
- Courses after graduation:
  - Deep Learning (DTU)
  - Data Engineering (udacity)
  - Machine learning (stanford)
  - Power BI (SuperUsers)

# Project: Finance ETL

## Outline

- Data is downloaded from yahoo finance and uploaded to a data lake on AWS S3
- - Create: IAM Role and Redshift cluster
- - Create PostgreSQL tables
- - \*E\*xtract data from S3 to staging tables on Redshift
- - \*T\*ransform data from staging tables to Fact and Dimension tables
- - Test tables are correct
- - \*L\*oad data for analysis

## Objective

- use the DWH for analytics e.g. forecasting model of price volatility
- Potentially combine with reddit sentiment analysis

## Data Warehouse Design

### Dimension tabel: DimCorp

Ticker	CorpName	CEO	Founded
TSLA	Tesla	Elon Musk	2003-07-01
AAPL	Apple	Tim Cook	1976-04-01

### Fact tabel: FactHist (sample)

ticker	date	adjclose
TSLA	2022-05-25	658.0
AAPL	2022-04-29	157.0

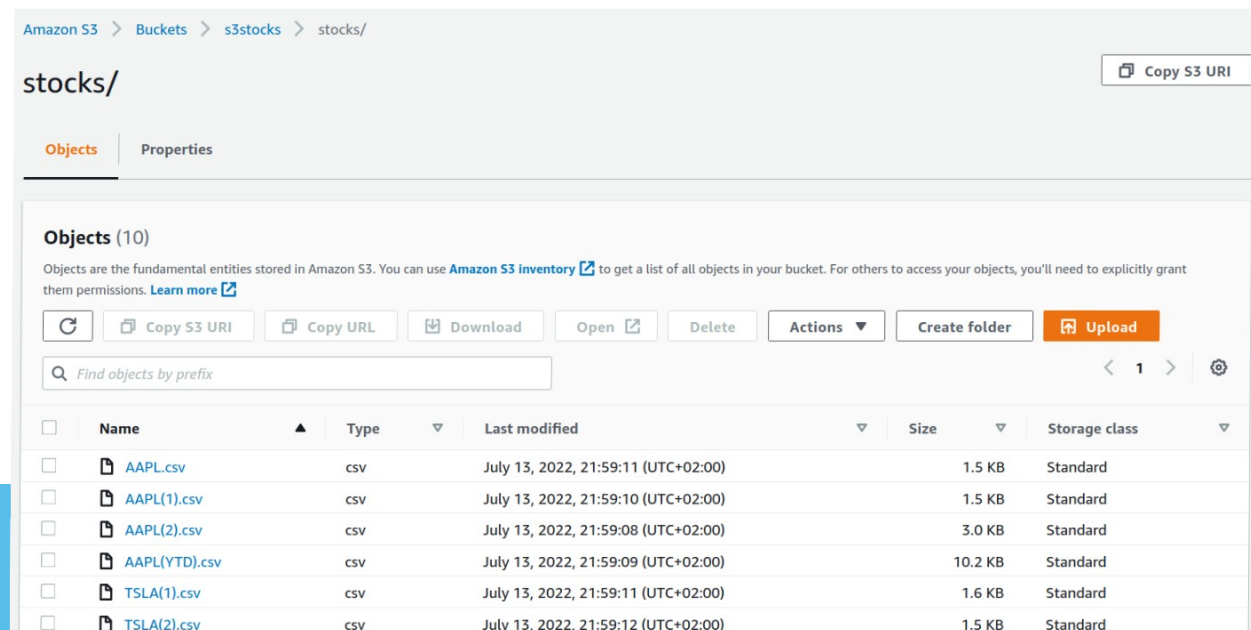
# AWS

## Data Lake: AWS Amazon Simple Storage Service (S3)

- Pros:
  - Cost-effective
  - Scalable
  - Automatic backup
  - Sharing/Access
- Cons:
  - Lack of total control

## Data Warehouse: AWS Redshift

- Pros:
  - Performance: Massively Parallel Processing (MPP)
  - Scalable
- Cons:
  - +/- price



# Data Lake: Approach

## Setting up AWS (General)

- 1) Create AWS user
- 2) Download and set up  
Command Line Interface (CLI)
- 3) Create IAM Role (permissions)
- 4) Create IAM user
  - programmatic actions
  - attach existing policies
  - AdministratorAccess
- 5) Create Security Group

## Setting up AWS S3

- A) Create S3 Bucket
- B) Set billing alarm!**
- C) Assign IAM Role permissions
- D) Upload (CSV/json) files
- E) Assign IAM users (clients)

```
(base) martin@martin-HP-ENVY-x360-Convertible-13-ag0xxx:~$ aws configure list --profile AdminMabiUser
Name                               Value                                Type    Location
----                               -
profile                            AdminMabiUser                        manual  --profile
access_key                         *****KHYC                         shared-credentials-file
secret_key                         *****QX/P                         shared-credentials-file
region                             us-east-1                           config-file  ~/.aws/config
```

# Let's explore

- Github
- AWS
- VS code
- Terminal

The screenshot shows the Yahoo Finance website for Apple Inc. (AAPL). The page includes a search bar at the top, navigation links, and a detailed view of the stock's performance. The current price is \$148.47, up \$2.98 (+2.05%) from the previous close. The historical data table shows prices from July 11, 2022, to July 14, 2022.

**Apple Inc. (AAPL)**  
NasdaqGS - NasdaqGS Real Time Price. Currency in USD

**148.47** +2.98 (+2.05%) **148.47** 0.00 (0.00%)  
At close: July 14 04:00PM EDT Pre-Market: 04:09AM EDT

[Add to watchlist](#) [Start Trading >>](#)  
Plus500 77% of retail CFD accounts lose money

[Summary](#) [Chart](#) [Conversations](#) [Statistics](#) [Historical Data](#) [Profile](#) [Financials](#) [Analysis](#) [Options](#) [Holders](#) [Sustainability](#)

Advertisement  
AD

Ads by Google  
[Send feedback](#) [Why this ad? ↗](#)

Time Period: Jul 15, 2021 - Jul 15, 2022 Show: Historical Prices Frequency: Daily [Apply](#)

Currency in USD [Download](#)

Date	Open	High	Low	Close*	Adj Close**	Volume
Jul 14, 2022	144.08	148.95	143.25	148.47	148.47	77,996,900
Jul 13, 2022	142.99	146.45	142.12	145.49	145.49	71,185,600
Jul 12, 2022	145.76	148.45	145.05	145.86	145.86	77,588,800
Jul 11, 2022	145.67	146.64	143.78	144.87	144.87	63,141,600

# Objective: Analysis

- Portfolio Overview

ticker	date	adjclose	corpname	ceo	founded
AAPL	2022-01-03	181.0	Apple	Tim Cook	1976-04-01
AAPL	2022-01-04	179.0	Apple	Tim Cook	1976-04-01
AAPL	2022-01-05	174.0	Apple	Tim Cook	1976-04-01
AAPL	2022-01-06	171.0	Apple	Tim Cook	1976-04-01
AAPL	2022-01-07	171.0	Apple	Tim Cook	1976-04-01

- Stock price volatility forecast

- Predict e.g. daily
- Retrain model monthly

ticker	date	adjclose	daily_return
AAPL	2022-01-03	181.0	NaN
AAPL	2022-01-04	179.0	-1.104972
AAPL	2022-01-05	174.0	-2.793296
AAPL	2022-01-06	171.0	-1.724138
AAPL	2022-01-07	171.0	0.000000

# Further development

## For Production

- Apache Airflow (ETL part)
- API for live update
- Add reddit sentiment
  - Requires EC2 w. GPU

## For speed (e.g. for trading)

- PySpark
- Distributed systems Hadoop
- AWS: EMR or EC2 Cluster Computing