Yahoo Finance MapReduce Checkpoint — Mark Chen (jc10691)

1. Source & metadata

- Purpose: produce daily equity, rate, and volatility indicators for the team's Fed-odds project.
- Extractor: `data_sources/yfinance.py` with a two-year window, daily interval. Writes raw CSVs to `market_data/yfinance/` before uploading to HDFS.
- Run metadata:


```
{
  "start": "2023-11-18",
  "end": null,
  "interval": "1d",
  "equity_tickers": {"^GSPC":"sp500_index","^IXIC":"nasdaq_composite","^DJI":"dow_jones_index","XLF":"financials_etf","XLK":"technology_etf","XLY":"consumer_discretionary_etf","XLP":"consumer_staples_etf","XLI":"industrial_etf","XLE":"energy_etf","XLU":"utilities_etf"},
  "rate_tickers": {"ZQ=F":"fed_funds_futures","^IRX":"3m_treasury_bill","^FVX":"5y_treasury_yield","^TNX":"10y_treasury_yield","^TYX":"30y_treasury_yield"},
  "vix_tickers": {"^VIX":"cboe_volatility_index","^VVIX":"cboe_vvix_index"}
}
```


2. Command log

`mark_yfinance_commands.log` captures every shell command executed on NYU Dataproc, including HDFS uploads, three profiling jobs, three cleaning jobs, sampling via `hdfs dfs -cat ... | head`, and the `getmerge` steps that produced the local CSV snippets.

3. Profiling summary

- Jobs: `yfinance_profile_mapper.py` + `yfinance_profile_reducer.py` via Hadoop Streaming per dataset.
- Example reducer output:

```
XLE {"row_count":500,"first_date":"2023-11-20","last_date":"2025-11-17","value":{"avg":88.3581,"min":76.44,"max":98.08},"volume":{"avg":15875207.058},"series_labels":{"energy_etf":500}}
ZQ=F {"row_count":419,"first_date":"2023-11-20","last_date":"2025-07-21","value":{"avg":95.111557},"volume":{"non_null":0},"series_labels":{"fed_funds_futures":419}}
```

4. Cleaning output (snippets)

Mapper `yfinance_clean_mapper.py` normalizes dates, ticker case, dataset type, and numeric fields. Reducer `yfinance_clean_reducer.py` de-duplicates by `(ticker, date, series_label)` and prints the canonical CSV schema.

```
# cleaned_yfinance_equity.csv (first 5 rows)
date,ticker,series_label,dataset_type,value,open,high,low,close,adj_close,volume,quality_notes
2023-11-21,XLE,energy_etf,equity,84.62000275,84.41000366,84.76999664,83.80999756,84.62000275,79.17155457,13486900,ok
...
```

```
# cleaned_yfinance_rates.csv (first 5 rows)
2023-11-21,ZQ=F,fed_funds_futures,rate,94.66999817,,,,,,ok
...
```

```
# cleaned_yfinance_vix.csv (first 5 rows)
2023-11-21,^VIX,cboe_volatility_index,equity,13.35000038,13.44999981,14.31000042,13.13000011,13.35000038,13.35000038,0,ok
...
```

5. Submission package

- `mark_yfinance_commands.log` — cleaned terminal transcript.

- `yfinance_profile_mapper.py`, `yfinance_profile_reducer.py`, `yfinance_clean_mapper.py`, `yfinance_clean_reducer.py` — raw MapReduce code.
- `sample_cleaned_yfinance_*.csv` — 20-line snippets of the cleaned outputs (full files withheld per assignment guidance).
- `yfinance_metadata.json` — extractor parameters.
- This markdown report — narrative + evidence, to be exported to PDF if requested.