Leidenschaft
fürs Leben.

Big Data and Business Analytics
Data Engineering 2
Project

Frank Schulz
13 May 2024

Data Engineering Project

Group work

- Groups of **3 participants**
- **Goa**l: Develop and describe a prototype of a **Data Pipeline**,
  addressing Data Ingestion, Data Storage and Data Processing/Visualization

Expected results

(1) Project report

- One report per group, around 20 pages (± 3 pages).
- Each participant shall contribute one or more sections,
  and authorship of each section has to be marked.

(2) Presentation with live demo on 7 June 2024 at 9:30

- 10 minutes, everybody from the group should present some part
- 10 minutes discussion

Project Report

Project report

- First page with title and abstract.
- Bibliography required: All used external sources have to be cited properly (books, papers, blog articles, web pages, Github repositories, ...)
- Nice layout (no distorted images, no screenshots with black background)

Goal

The project report should be a short instruction for repeating the work done in the prototype. Don't give any general information, please be specific to your prototype.

- Which tools have been chosen?
- Which steps have been taken (possibly with some selected snippets of commands or code)?
- Which difficulties have occurred, and how were they solved?
- Which decisions have been made (without theoretical justification)?

Project Report Details (1)

Abstract (1 page)

- Introduction (1-2 sentences)
- Application Problem (1 sentence)
- Summary of own approach (2 sentences: idea and methodology)
- Summary of own results (1-2 sentences)

Report

- Chapter 1: Introduction (2-3 pages)
  (description of application domain, problem on application level, benefits of a solution on application level, problem on technical level, technical solution idea)
- Chapter 2: Related Work (1-2 pages)
  (how did other people solve this issue or a similar issue already)
- Chapter 3: Dataset (description of data source) (1-2 pages)
- Chapter 4: Solution (all details of data processing and implementation) (11-15 pages)
- Chapter 5: Summary and Outlook (own results, future work) (1-2 pages)

Project Report Details (2)

Bibliography

- The bibliography must follow academic standards. It is mandatory that authors' surnames and initials, paper title and year are specified, and additional information depending on the type of reference (journal name and number, conference name, pages, URL for web article).
- Citation should consistently follow the ISO 690 numeric / IEEE style
- References in square brackets, for example [1] or [2,3,5].
- Bibliography example:

[1] D. E. Knuth. *The Art of Computer Programming, Vol. 1: Fundamental Algorithms*, 3rd edition. Addison-Wesley, 1997.

[2] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer New York, 2009.

[3] S. Hochreiter and J. Schmidhuber. *Long Short Term Memory*, Neural Computation, vol. 9, pp. 1735–1780, 1997.

[4] D. Becker. *Running Kaggle Kernels with a GPU*, 2018. URL: https://www.kaggle.com/code/dansbecker/running-kaggle-kernels-with-a-gpu/notebook (last accessed 28 November 2022)

Project Presentation

Project presentation should contain the following:

- Short introduction of the chosen data source
- Motivation from application perspective: which questions will be answered with help of the data
- Main part: Description and explanation of the data pipeline, **a live demo of the running pipeline is mandatory**

Goal

The project presentation should give an overview of the prototype that you have built.

- What are the components and how are they connected with each other?
- Did you use any specific configurations of the components?
- Which difficulties have occurred, and how were they solved?

- There will be challenges during the project. Addressing and resolving the challenges appropriately will be positively rewarded.

Project Tasks

Data Ingestion

- Either from API call
- Or from a data stream (Wikipedia live changes,...)

Data Storage

- Use Cloud Services for the pipeline and its components
- For example, Google Cloud Platform, Confluent Kafka
- Other platforms and services are possible (Azure, AWS, Snowflake, ...) and can be included in the live demo

Data processing and visualization

- Queries with Aggregation
- Display of results: as table and/or with simple visualizations
  (e.g. Jupyter Notebook, Streamlit, Google Cloud)

Data Source Proposals

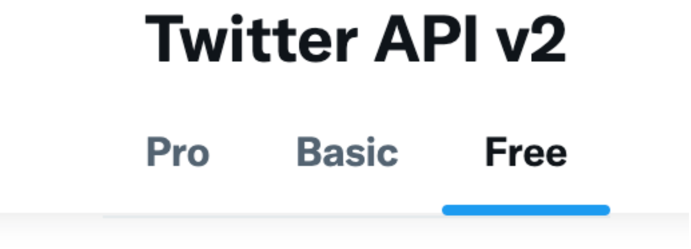
Data Source 1: Twitter

~~Twitter~~

- ~~Access to the Twitter live stream~~
- ~~Description see https://developer.twitter.com/en/docs/tweets/filter-realtime/api-reference/post-statuses-filter~~
- ~~Using endpoint https://stream.twitter.com/1.1/statuses/filter.json~~

Deprecated

- Twitter API v2 only allows very limited access in the free plan



Data Source 2: Other Social Media Platforms

Reddit

- https://www.reddit.com/wiki/api

Youtube

- https://developers.google.com/youtube/v3/

Instagram

- https://developers.facebook.com/docs/instagram-basic-display-api

Facebook

- https://developers.facebook.com/docs/graph-api/

Python client library for Facebook

- https://pypi.org/project/python-facebook-api/

Social Network Scraping

- https://github.com/JustAnotherArchivist/snscrape/blob/master/README.md

Data Source 3: Wikipedia

Wikipedia

- Data is available here (Download, API, Recent Changes Stream)
  https://meta.wikimedia.org/wiki/Research:Data

- Event stream of recent changes
  https://wikitech.wikimedia.org/wiki/EventStreams
  https://www.mediawiki.org/wiki/API:Recent_changes_stream

  Endpoint for reading data: https://stream.wikimedia.org/v2/stream/recentchange

- Examples
  http://rcmap.hatnote.com/#en
  http://listen.hatnote.com/

Data Source 4: New York Taxi

New York Cabs

- Data on the taxi trips in New York
  https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

- Monthly PARQUET files in three categories

Yellow	around 800-900 MB each
Green	around 100 MB each
FHV (For Hire Vehicle)	around 400-800 MB each

- Contains geographical data (Pickup / Dropdown Zone)
  => Possibility of visualization on a map

More open data from New York City Government

- https://opendata.cityofnewyork.us/

Data Source 5: Fuel Prices

Fuel prices

- Tankerkönig ("Fuel king") offers access to current fuel prices of all fuel stations in Germany. The fuel stations are obliged to report their prices to the "Markttransparenzstelle für Kraftstoffe" (MTS-K):
  http://www.tankerkoenig.de/

- Historical data for download
- API for accessing current data: https://creativecommons.tankerkoenig.de/
  - Name, address and geographical coordinates of the fuel station
  - Current prices for different types of fuel
  - Opening times and information whether currently open or closed

Data Source 6: News Articles

- **New York Times** offers extensive APIs for querying news articles and related information: https://developer.nytimes.com/apis
  - Article search
  - Most popular articles
  - Geographical information
  - User comments

- **FiveThirtyEight**
  - Opinion polls and other news
  - https://data.fivethirtyeight.com/

Data Source 7: Weather and Climate

- **Open Weather**
  - Free access to current weather and limited forecast, at most 1000 calls/day
  - https://openweathermap.org/api
  - https://openweathermap.org/api/one-call-api

- Weather API
  - https://www.weatherapi.com/

Data Source 8: Stock Market

- Alpha Vantage
  - Free real-time stock data
  - https://www.alphavantage.co

- IEX Cloud
  - Real-time data for financial applications
  - https://iexcloud.io/

- Twelve Data
  - https://twelvedata.com/docs#getting-started

- Quantopia
  - Quantitative finance data including real-time stock prices
  - https://www.quantopian.com

- Quandl
  - Economic and financial data
  - https://www.quandl.com/search

Data Source 9: Crypto

- Coincap
  - Free, no registration required (rate limiting of 200 queries / minute)
  - https://docs.coincap.io/

- Coingecko
  - Free, no registration required (rate limiting of 50 queries / minute)
  - https://www.coingecko.com/en/api

Data Source 10: Movie Data

- The Movie Database (TMDb)
  - Access to movie and series metadata
  - https://www.themoviedb.org/documentation/api

- Python client library
  - https://pypi.org/project/tmdbsimple/
  - and others

Data Source 11: Traffic Data

- Pedestrians in Germany cities
  - Free access to number of people passing a specific point
  - https://hystreet.com/ (requires registration)

- Bikesharing
  - NextBike: https://api.nextbike.net/maps/nextbike-live.json
  - Capital Bikeshare: https://www.capitalbikeshare.com/system-data

- German Mobilithek (Railway, ...)
  - https://mobilithek.info/

- German public transportation (ÖPNV)
  - https://www.opendata-oepnv.de/ht/de/willkommen

- Airplanes
  - FlightRadar24: https://www.flightradar24.com/
  - Only aggregate data can be downloaded for free
- Ships
  - MarineTraffic: https://www.marinetraffic.com/
  - No free download

More Data Sources

Various APIs for notifications and warnings of public interest:
https://github.com/bundesAPI

Further data sources
https://rapidapi.com/search/