MaincodeHQ / **mainpipe**

Code Issues Pull requests Actions Projects Security Insights

Take home for Senior Data Engineer

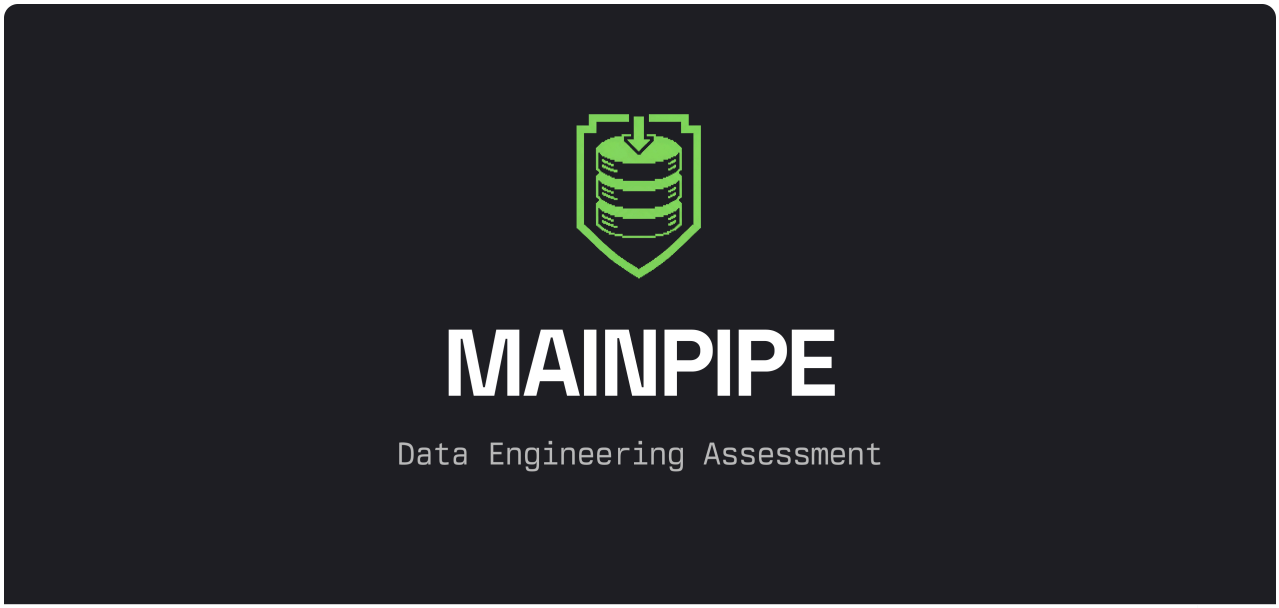
0 stars 0 forks 0 watching Branches Activity Custom properties
Tags

Public repository

1 Branch 0 Tags Go to file t Go to file Add file Code

maincode-dave Update readme.md dcb9533 · 2 months ago

docs/images	Delete docs/images/logo.png	2 months ago
readme.md	Update readme.md	2 months ago



README

Mainpipe - Senior Data Engineer Take-Home

Intro

This take-home is your chance to showcase your expertise as a Senior Data Engineer. You'll be building a simplified version of a real data pipeline used in LLM pre-training: we call it 'mainpipe'. We will evaluate your submission on system architecture, code quality, performance, and the clarity of your written report, which should clearly communicate your approach and results to the Maincode team.

Your Task

Build an end-to-end data pipeline focused on filtering & data preparation. We provide the starting point (sample of unprepared data) and the end result needs to be a an english language dataset ready for LLM pre-training.

The following elements are **must-haves** of the submission, and a good submission will have additional elements of your choice:

1. **Data acquisition**
2. **Data cleaning, normalisation and tokenisation**
3. **Training ready exports** (i.e. tokenised, mixtures, shards, etc.)
4. **Inspectability** (e.g. histograms of length, lang scores, dup markers, PII hit-rates, drop reasons)
5. **Conceptual plan for scaling**

The brief intentionally leaves room for interpretation - your choices and rationale are an important part of the evaluation. There are existing open-source pipelines for the preparation of LLM training datasets, and we encourage you to draw inspiration from them for best practices. **We expect you to leverage and combine existing building blocks rather than implementing everything from scratch** - focus on thoughtful integration and customization of proven tools and libraries.

Keep the solution self-contained, but feel free to explain what you would do differently at real scale. The take-home is designed to be completed in roughly four hours of focused work.

Ground Rules

- **Language:** Python 3.10+
- **Containerised pipeline** that runs end-to-end
- **Data:** Use the dataset provided below

Dataset Instructions

For this assignment, you'll work with a curated, multi-domain slice assembled from various sources. The raw dataset is available for download from:

https://s3.us-east-1.amazonaws.com/mainpipe.maincode.com/mainpipe_data_v1.jsonl

Deliverables

1. GitHub repository with your data pipeline and README explaining how to run it
2. A link to your fully processed dataset
3. A written report summarising your work and design decisions

How We Evaluate Your Take-Home Submission

- **Pipeline design** (containerisation, attributes, mixing)
- **Performance** (see below for details)
- **Scalability & Systems thinking** (Spark/Ray configs for scale up plan, partitioning, shuffle strategy, small-file mitigation, failure modes etc.)

- **Observability & reproducibility** (logs, metrics, deterministic seeds etc.)
- **Code quality and engineering hygiene**
- **Quality of your project report**
- **Creativity**

To evaluate the performance, we evaluate your pipeline and the processed dataset it produces against the following metrics:

- Deduplication
- Noise/Integrity: too-short, long repeats, non-printable, markup/boilerplate
- Linguistics: sample-based perplexity proxy with a small LM
- Safety: PII hits + toxicity (Detoxify)
- Coverage: language distribution
- Pipeline throughput

How to Submit

Please follow the submission instructions provided to you via email. A complete submission includes a link to your repository, your processed dataset and your written report.

Make sure the instructions in your README allow us to run the pipeline end-to-end without additional setup.

If we like your submission, we will invite you to a 30-minute call for an in-depth discussion of your work with our technical team.

Closing Note

We're excited to see your submission! This is your chance to show us your approach, creativity, and engineering craftsmanship. We're looking forward to reviewing your work and hope to talk to you soon.

The Maincode Team

Releases

No releases published

Packages

No packages published

Contributors 2


maincode-dave Dave Lemphers


lukas-maincode