Thinking Machines
Data Science

Data Engineer Take Home Exam

General instructions

You are given **72 hours** upon receiving the exam to complete the deliverables.

Remember : **do not upload any parts of this exam or your answers on publicly accessible sites** like the public repositories on Github. This is to ensure that the exam will not be leaked to future applicants.

Upon completion, send an email to your hiring manager with a link to your deliverables to stop the timer. Good luck!

16/F Twenty-Four Seven McKinley
24th Avenue corner 7th Avenue
Bonifacio Global City, Taguig City
Metro Manila, Philippines

http://thinkingmachin.es
info@thinkingmachin.es

Part 1: Algorithmic Thinking

Euler discovered the remarkable quadratic formula:

$$n^2 + n + 41$$

It turns out that the formula will produce 40 primes for the consecutive integer values $0 \leq n \leq 39$. However, when $n = 40, 40^2 + 40 + 41 = 40(40 + 1) + 41$ is divisible by 41, and certainly when $n = 41, 41^2 + 41 + 41$ is clearly divisible by 41.

The incredible formula $n^2 - 79n + 1601$ was discovered, which produces 80 primes for the consecutive values $0 \leq n \leq 79$. The product of the coefficients, −79 and 1601, is −126479.

Considering quadratics of the form:

> $n^2 + an + b$, where $|a| < 1000$ and $|b| \leq 1000$
>
> where $|n|$ is the modulus/absolute value of n
> e.g. $|11| = 11$ and $|-4| = 4$

Find the product of the coefficients, a and b, for the quadratic expression that produces the maximum number of primes for consecutive values of n, starting with $n = 0$.

Please code up a brute-force solution to this problem (it will not finish in a reasonable time). If you pass the exam stage, prepare to explain possible solutions to improve your current approach in relation to runtime during the panel interview.

Part 2: Soft(ware) skills

We live and breathe data. So naturally, we try to make our internal processes as data-driven as possible. Time-tracking allows us to determine our teams' bandwidth and throughput so we checkin religiously. We do this using a handful of tools that we can use without interrupting our work.

For this exam, you are given an anonymized dump of our checkins for the past year. Your goal is to:

- Clean the data
- Load it to your choice of database
- Create a web service that returns the checkin data associated to a given user

Additionally, answer the following questions:

- If the data is to be ingested periodically, what changes will you make to your current approach?
- Draw a data architecture showing different components of your ETL process.
- How will you verify the correctness of the ingested data?

Bonus points for:

- Using Python
- A publicly accessible deployment of your service
- Documentation
- Tests
- Diagrams
- Creating a web service that displays a per-user filtered view of the check-ins

We value:

- Communication
- Reproducibility
- Pragmatism
- Code hygiene

Submit your code by sharing a private Github repo with the following users:

- https://github.com/tm-jamie-macagba
- https://github.com/gab-tm
- https://github.com/tm-justin-beredo
- https://github.com/tm-erin-cheng
- https://github.com/tm-ramon-tonato
- https://github.com/tm-marco-francisco
- https://github.com/tm-tor-boonsri
- https://github.com/tm-peem-poomka

As a backup, please also upload your code to the Drive folder shared with you.