En esta parte, lo que haremos es un proyecto para hacer en casa. La idea es que nos hagas una presentación de 1 hora en donde podamos evaluar tus habilidades en ingeniería de datos. Para ello deberás exponer en base a un proyecto orientado a datos, motivado por algún análisis. El foco estará en los siguientes puntos:

- Las buenas prácticas que hayas implementado
- La elección de tecnologías y sus motivaciones
- Su interrelación con la parte de Data Analytics y Data Science

No nos enfocaremos en los resultados del análisis ni la metodología, salvo en casos muy puntuales. La presentación que se haga debe estar alineada en este sentido.

El tiempo de presentación debe contemplar las preguntas que podamos hacerte y la misma debe ser en inglés.

Como la idea no es rehacer la rueda, puedes elegir algún proyecto Open Source en el que hayas participado. En caso de no contar con ninguno de esta índole, puedes realizar algunas de las tareas de Kaggle.com, o bien mostrarnos un análisis sobre alguno de sus set de datos libres.

Para asegurarnos que el tema está bien alineado con nuestras expectativas, debes proveernos de tres opciones indicando título y un abstract de temas a tratar. Nosotros elegiremos de allí o te daremos más indicaciones para cerrar una mejor definición. La fecha de presentación se coordinará una vez concertado el tema.

El día previo a la presentación debes haber entregado:

- Descripción del problema funcional que resuelve la solución
- Los set de datos adicionales que has utilizado, si los hubiere
- El código fuente de tu solución e instrucciones para su reproducción
- Soporte digital (diapositivas)

Todo el proyecto no debe llevarte más de 24 horas de trabajo.

A continuación te dejo algunas preguntas frecuentes que hemos tenido y casos de ejemplo para motivarte. Cualquier duda, puedes contactarme sin problemas.

Q. ¿Estamos buscando específicamente set de datos cuantitativos o data relacionada con textos e imágenes también sirve?
A. Cualquier tipo de data está bien. Preferentemente, que esté abierta al público.

Q. ¿El tema del proyecto se enfoca más en procesamiento de datos o análisis de datos?
A. El proyecto se enfoca en ingeniería de datos de punta a punta. Es decir, en la articulación de una solución técnica para un problema dado. En consecuencia, algún tipo de mínimo de análisis de datos es requerido, pero no es suficiente ni el punto más importante a evaluar.

Q. ¿Cuál es la escala apropiada para el proyecto?
A. No esperamos que el candidato tarde más de tres días en el proyecto. La calidad es preferida a la cantidad.

Q. ¿Es esencial que el proyecto incluya procesamiento distribuido?
A. No hay necesidad de demostrar la solución en un ambiente distribuido, pero se espera que se encare el problema para escalar de manera distribuida o bien que se

indique claramente sus limitaciones durante la presentación. Es decir, una aplicación de Spark corriendo en una laptop está bien, si se prueba o se puede asumir con seguridad que escalará apropiadamente a set de datos más grandes u otros modelos.

Q. ¿Es necesario presentar un análisis exploratorio sobre el set de datos que contengan visualizaciones de tendencias generales y relativas?
A. El análisis exploratorio puede ser parte del proyecto, pero sólo como un paso previo a la solución técnica en concreto.

Q. ¿El proyecto debería incluir distintas capas de modelado de datos?
A. El assessment debe demostrar las capacidades ingenieriles del participante. Incluso si el problema funcional es pequeño, las propiedades de la solución deberían ser de carácter productiva o bien tener las limitaciones muy bien especificadas.

Proyectos ejemplo:

- Novel Coronavirus Dataset 2019
    - Dataset muy interesante ya que está bastante documentado y con buenas fuentes. En la descripción pone algunos insights que tal vez sean fáciles de explorar y que pueden servir de motivación. Otros insights válidos pueden ser:
        - ¿Qué países reaccionaron mejor a la infección? (más rápido, con mayor eficiencia, con menor cantidad de víctimas fatales, etc)
        - ¿Hay alguna correlación entre el primer infectado y el pico de la crisis?
        - ¿Qué tan infecciosa es la enfermedad?
- 2019 Coronavirus Dataset
    - Otro dataset sobre lo mismo. Este en particular parece estar más sucio y menos documentado, con lo cual requiere más trabajo de procesamiento previo al análisis. Las mismas preguntas que para el análisis anterior se aplican.
- Trump Impeachment Polls
    - Este dataset trata sobre un tema muy interesante de la política de EEUU y a su vez presenta varios desafíos en lo que análisis de lenguaje natural. En la descripción propone ciertas preguntas muy válidas para contestar. Yo también haría las siguientes preguntas:
        - ¿Las preguntas son buenas o son capciosas? ¿Guían al entrevistado?
        - ¿Existe parcialidad/*bias* en los entrevistadores? ¿Usaron muestras sesgadas?
- Bixi Montreal Bikeshare Data y Toronto Bikeshare Data
    - Estos datasets contienen datos de viajes de empresas de bikesharing de Canada. Cosas que podrían ser interesantes de analizar son:
        - Horas punta

        - Frecuencia de uso
        - Hotspots

    - A su vez, sería interesante hacer comparativa entre los dos datasets, ya que son muy parecidos pero levemente distintos. Las métricas podrían ser similares y hacer comparación entre ambas ciudades, como por ejemplo:

        - ¿Qué ciudad usa más la bici?
        - ¿Hay diferencia entre las horas pico de las dos ciudades?