PROCESO SELECCIÓN PERSONAL INGENIER@ DE DATOS

Prueba Técnica

Objetivo

Queremos evaluar tus competencias técnicas y tú conocimiento para aspirar al cargo de ingenier@ de datos. Evaluaremos los siguientes elementos:

CONOCIMIENTOS FUNDAMENTALES

1. Conocimientos básicos en gestión de bases de datos (MySQL, SQL Server, Postgre SQL, o alguna equivalente)
2. Entendimiento de archivos en formatos XLSX, XLS, CSV, JSON y XML
3. Conocimientos en estrategias ETL
4. Conocimientos en relacionamiento y visualización de datos

CONOCIMIENTOS DESEABLES

5. Manejo de archivos en formato Parquet, o uno equivalente de almacenamiento por columnas
6. Manejo de herramientas data en AWS (Kinesis, S3, Athena, Glue, Redshift)
7. Manejo Python, C# ó Java
8. Tecnologías de visualización: Tableu, Power Pivot, Power BI, Google Data Studio, AWS Quicksights.
9. Conocimientos básicos en gestión de bases de datos no relacionales (MongoDB, AWS DynamoDB)

Dinámica de la prueba

Encontrarás más adelante la descripción de un reto de desarrollo que involucra los elementos descritos en el objetivo de este documento. Te compartiremos también un contacto de alguien de nuestro equipo Tekus quien estará a disposición durante la ejecución de la prueba para resolver cualquier inquietud.

En la prueba se evaluarán los siguientes puntos:

1. Tiempo de ejecución de la prueba
2. Funcionalidad lograda
3. Organización (modularidad y comprensión del flujo desarrollado)

Esta prueba será tu principal carta de presentación. Te recomendamos leer detenidamente el problema y desarrollar cada requerimiento con todos los detalles solicitados.

Si deseas implementar alguna algoritmo o programa de software que involucre código te solicitamos crear un repositorio GIT al iniciar la ejecución de la prueba y compártalo con el contacto proporcionado, esto facilitará la revisión incluso sin finalizar la prueba.

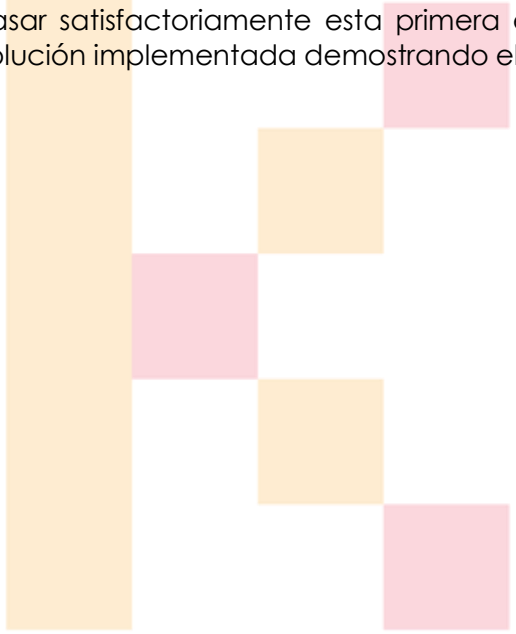
Si consideras que no alcanzarás a finalizar la prueba a tiempo, agrega una descripción de cómo resolverías el problema en un caso real.

Tiempo de desarrollo de la prueba

Aunque la prueba está diseñada para que la realices en un lapso de 4 a 5 horas, te otorgamos hasta **5 días (calendario)** en los que podrá realizar la prueba. Con esto esperamos que organices tu disponibilidad horaria como lo consideres conveniente y aproveches el tiempo restante para optimizar tu prueba (realizar pruebas, documentar, organizar el código, validar los requisitos solicitados)

Términos y condiciones

1. No tendrás ningún tipo de restricciones en las páginas que desees consultar, archivos de referencia, ni en el código de ejemplo que utilices siempre y cuando éste no incumpla ninguna ley de derechos de autor. En el caso de utilizar código de ejemplo deberás incluir las referencias al origen del código (URL, proyecto, autor)
2. Debes ser consciente que de pasar satisfactoriamente esta primera etapa serás cuestionado en una entrevista técnica acerca de la solución implementada demostrando el manejo y entendimiento de esta.



INSTRUCTIVO

Problema

La empresa **Colombiana de Ollas Inc** ha lanzado su nuevo y último producto: "La Olla arrocera **NOSEPEGA3000**". Entre sus más recientes innovaciones se encuentra un sistema que mide la cantidad de interacciones que tiene el usuario con su extraordinario panel interactivo de un solo botón. Así mismo este sensor mide cuando desplazaron la olla de lugar, lo que le permite identificar a la empresa que tan importante es fabricar Ollas móviles para un futuro cercano. Todos estos datos son consolidados y enviados cada 20 minutos al almacén de datos de la empresa para su posterior análisis.

Para simplificar la lectura de los datos hemos generado un archivo ZIP con los reportes de todas las ollas durante los años 2020 y 2021. Este archivo lo podrá descargar en la siguiente dirección:

https://tekus.s3.amazonaws.com/tests/data-engineer-sample-data-v2.zip

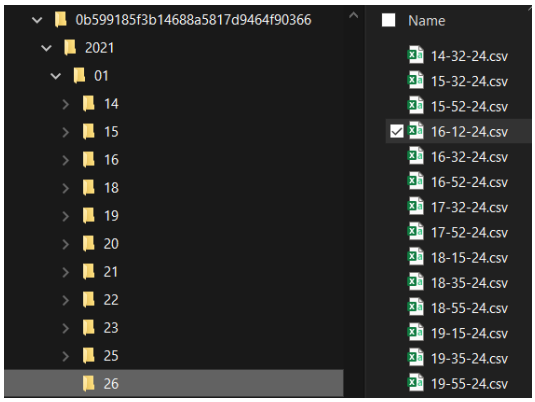
La estructura de los archivos contenidos en este ZIP es la siguiente:

IDENTIFICADOR_OLLA/YYYY/MM/DD/hh-mm-ss.csv

A continuación, se detalla cada una de las partes del sistema de archivos

Campo	Ejemplo	Descripción
IDENTIFICADOR_OLLA	0b599185f3b14688a5817d9464f90366	Es un código interno que permite identificar la olla
YYYY/MM/DD	2021/01/26	Describe la fecha en la que fue creado el registro. Cada parte corresponde a una carpeta
hh-mm-ss.csv	16-12-24.csv	Corresponde a un archivo cuyo nombre representa la hora a la que fue creado. Cada archivo puede contener varios registros

Al visualizar esta información en el explorador de Windows se vería algo cómo lo siguiente:



Cada archivo CSV contiene las siguientes columnas:

Campo	Ejemplo	Descripción
Duration	4	Describe la duración en segundos de la interacción que tuvo el usuario con la olla. Pudo ser debido a un movimiento o a la interacción con el panel de un solo botón
MovementDuration	3	Indica la duración en segundos del movimiento que registró la olla
MovementInteractions	2	Indica la cantidad de veces que movieron la olla
ArkboxInteractions	NO UTILIZADO	IGNORAR ESTE CAMPO
StandardInteractions	NO UTILIZADO	IGNORAR ESTE CAMPO
HardwareInteractions	3	Indica la cantidad de veces que el usuario interactuó con el extraordinario panel de un solo botón
TenantId	NO UTILIZADO	IGNORAR ESTE CAMPO
SessionId	NO UTILIZADO	IGNORAR ESTE CAMPO
Key	0b599185f3b14688a5817d9464f90366	Es un código interno que permite identificar la olla. Corresponde al campo IDENTIFICADOR_OLLA
Id	NO UTILIZADO	IGNORAR ESTE CAMPO
Date	2020-07-18T13:11:05.8800-0500	Fecha en formato ISO8601
DateInTicks	NO UTILIZADO	IGNORAR ESTE CAMPO
TTL	NO UTILIZADO	IGNORAR ESTE CAMPO

Los campos marcados en gris no se deberán tener presentes. Es importante aclarar que debido a una actualización que tuvieron algunos sensores en la etapa de fabricación es posible encontrar algunos registros que no contengan todas las columnas.

Los registros de las ollas los podrás consultar en una Base de datos de SQL Server a la que puedes acceder con la siguiente información:

Campo	Valor
Host	proyectos.tekus.co
Puerto	1433
Usuario	datatest
Contraseña	9cUQ*48AAX8Q
Base de datos	DataTest

Sólo hay dos tablas disponibles:

- **Pots**: Tabla que almacena la información de las ollas registradas
  - **PotKey**: Identificador interno de cada olla. Este campo corresponde al IDENTIFICADOR_OLLA relacionado en la carpeta dónde se encuentra el archivo CSV
  - **Serial**: Este serial corresponde a un código único impreso en la olla. Será el valor para mostrar en los reportes solicitados
  - **CityId**: Identificador de la ciudad en la que se encuentra la olla. Es una clave foránea a la tabla Cities
- **Cities:** Tabla que almacena la información de las ciudades en las que se encuentran las ollas
  - **CityId:** Identificador interno de la ciudad
  - **Name**: Nombre de la ciudad

Nota: El usuario *datatest* **no tiene permisos para listar las base de datos existentes pero podrá notar que al realizar una consulta sobre cualquier de las tablas se podría consultar la información sin problemas.**

Usted ha sido seleccionado como el encargado de desarrollar el flujo de procesamiento de datos que realice las siguientes acciones:

1. Extraer la data
2. Catalogar y filtrar la data
3. Agrupar y preparar la data

Así también deberá implementar la herramienta de visualización que permita resolver las siguientes preguntas:

1. ¿Cuál es el top 10 de las ciudades con más movimientos registrados por los usuarios?
2. ¿Cuál es el top 10 de las ollas con más interacciones de los usuarios con el extraordinario panel interactivo de un solo botón?
3. ¿Cuáles son los horarios entre semana y fines de semana en dónde se presentan más desplazamientos de ollas?

Te invitamos a que mantengas una comunicación continua a través de correo electrónico. Si tienes dudas, te pedimos que seas super conciso y te soportes en evidencia

Cómo ganar puntos extra?

- Si logras que la visualización de los datos se realice en una herramienta online.
- Evitando procesos manuales: ¡automatiza! Sea con código o herramientas.
- Sugiriendo una estrategia de limpieza de datos, algunos sensores podrían estar malos.
- Sustenta tus decisiones con análisis estadísticos cuando sea necesario
- Sugiera gráficas, tablas o análisis adicionales que le permitan al director de la empresa **Colombiana de Ollas Inc** tomar mejores decisiones.
- Hacer sugerencias sobre la idoneidad de esta prueba. Si algo está, explica por qué.

Entregables

1. Diagrama del flujo de datos dónde se detallen los procesos y etapas de la gestión de datos.
2. Gráficas que den respuesta a las preguntas del problema
3. Un video de no más de 3 minutos explicando el flujo de datos implementado.
4. Si se desarrolla algún tipo de algoritmo para facilitar el procesamiento de datos éste deberá estar versionado en una herramienta compatible con GIT (github, bitbucket, gitlab, etc)
5. Enviar un correo a jaime.marin@tekus.co y a william.guerra@tekus.co notificando la finalización de la prueba y cualquier consideración adicional que sea necesaria para facilitar la revisión.

¡Muchos Éxitos!

Tekus Team