ARGOS

Prueba técnica

Analista de modelos

A lo largo de esta prueba sus conocimientos y habilidades en temas técnicas serán evaluados por medio de una serie de ejercicios. Es posible que algunos de los conceptos requeridos sean desconocidos para usted; no obstante, se habilita la oportunidad de consultarlos en la web de modo que pueda contextualizarse y los pueda aplicar según convenga. En algunos casos, de acuerdo a la tarea, se le pedirá que justifique sus acciones en términos técnicos y que las defienda frente a otras posibles soluciones.

Challenge

> Responda a los ejercicios según los requerimientos usando el planteamiento de los mismos y apoyándose en sus conocimientos. Usted deberá documentar sus pasos y presentar los entregables según el caso. Se le sugiere elegir de forma estrategica aquellos ejercicios en los que mejor se desempeñe o tenga conocimiento. Esta prueba tiene una duración de **4** horas y sólo el primer ejercicio deberá ser expuesto en *inglés*.

Ejercicio 0

Según Wikipedia un ingeniero "es una persona que brinda el servicio de ingeniería o utilizan el ingenio para resolver problemas"; no obstante, es de conocimiento común, que un ingeniero en realidad es quien, mediante machetería, o sea, con lo que tenga al alcance y mucha creatividad, resuelve lo que asignen. Esta forma de ser de los ingenieros les permite actuar en muchos campos, así pues, describa en sus palabras:

- ¿Qué es un ingeniero de datos?
- ¿Hasta qué punto puede aportar un ingeniero de datos?
- ¿Qué relación tiene un ingeniero de datos con un arquitecto de infraestructura?
- ¿En qué es deficiente un ingeniero de datos?
- ¿Qué características, como persona, debe tener un ingeniero de datos para generar valor en un equipo?
- ¿Debe un ingeniero de datos saber como modelar?

Ejercicio 1

> El término *big data* es, hoy en día, ampliamente utilizado en la academia e industria para referirse a situaciones en las que las condiciones de los datos imposibilita tratarlos en una máquina local de forma eficiente.

A usted se le requiere explicar este término a una persona de negocio sin conocimientos técnicos que ha escuchado que la *big data* podría ser la solución a sus problemas de análisis. Por favor aclare el término en **4** minutos buscando satisfacer, adicionalmente, las siguientes condiciones

- Alcance
- Costos

- Plataforma
- Complejidad

Ejercicio 2

> Los equipos que manejan temas de ingeniería y arquitectura de datos se enfrentan a retos técnicos y conceptuales casi todos los días. Frente a la constante introducción de herramientas, servicios y plataformas en el mercado es fácil distraerse o atrasarse en la aplicación correcta y oportuna de las mismas.

A usted se le solicita construir, de principio a fin, una solución que responda a la pregunta "¿cómo se comportan mis clientes en el tiempo?". La siguiente lista contiene una serie origenes de información y una corta descripción de los mismos. Defina, como mejor considere, cuantas etapas y pasos por etapa (así como herramientas requeridas en cada paso) se requerirían para completar el requerimiento de acuerdo al tipo de origen (*Fíjese en el ejemplo*).

	Etapa 1		**Etapa 2**			**Etapa x**	
Origen	Paso	Herramienta	Paso	Herramienta	...	Paso	Herramienta
Archivo plano							
Fuente csv con datos, encoding UTF-8 y separador ";". Muestra única de datos							
Base de datos interna							
Base de datos estructurada (SQL Server, MySQL, Postgres, etc) en un servidor interno de la compañía, con diferentes esquemas y tablas. Datos se actualizan todos los días							
Base de datos externa							
Base de datos no estructurada (Cassandra, MongoDB, etc) en plataforma Azure. Datos con alta transaccionalidad							

Un ejemplo (**no necesariamente ideal**) del resultado esperado se muestra a continuación

	Etapa 1	**Etapa 2**

	Etapa 1		Etapa 2	
Origen	Paso	Herramienta	Paso	Herramienta
Archivo de Excel	Modelación		Entrega resultados	
Fuente .xlsx con datos en pestañas, todos formateados y dispuestos correctamente. Muestra única de datos	Crear tabla dinámica por métrica de cliente	Excel	Copiar y pegar los resultados en una pestaña final	Excel como pdf

Nota: su resultado puede ser expuesto en el formato que usted se sienta más cómodo, por ejemplo, como un diagrama.

Ejercicio 3

Los procesos de extracción, transformación y carga de datos (*ETL*) hoy en día se han facilitado gracias a arquitecturas que soportan no solamente procesos de movimiento de grandes volumenes de datos, sino también procesamiento en caliente, permitiendo generar insights al segundo sobre la información de interés.

La compañía enfrenta la siguiente situación y quisiera conocer su recomendación experta sobre arquitectura e ingeniería.

Condiciones

- Existe, al menos, una fuente de información externa a los sistemas de datos de la compañía que debe ser integrada a un proceso de tranformación de datos. Considere que la fuente es un servidor *linux* externo con sistema de autenticación por medio de credenciales
    - Sugiera, adicionalmente, un protocolo de seguridad que proteja a la compañía de datos maliciosos o reprocesamientos innecesarios
- Hay un conjunto de referencias a datos internos (en un servidor *SQL*) que deben ser agregados y cruzados con la fuente externa. Dichos datos se encuentran almacenados en un *data lake* en la nube
    - Suponga que tanto la fuente externa como la interna pueden cambiar en materia de días (u horas)
- Los resultados del cruce de información deben ingresar a un módulo de transformación cuyo output es una sábana de datos en un **formato óptimo** para su almacenamiento
- La salida del proceso debe ser dispuesta a un usuario final que pueda consumirla
    - ¿Dónde pondría usted a disposición dichos datos?
    - ¿Qué condiciones expondría usted al usuario final sobre éstos datos?

Requerimiento

- Plantee un esquema de arquitectura que responda a las condiciones expuestas. Dicho esquema puede ser plasmado a través de un diagrama de flujo (https://www.draw.io/) con sus respectivas anotaciones y separación de mundos
    - Se le recomienda, pero no es obligatorio, que se base en una herramienta como *AWS*, *Azure* o *Google*, y use (nombrando) los servicios que interactuarían en su proceso.

- Se le pide incluya la configuración de los disparadores correspondientes para que las actividades en su proceso se mantengan en pie y pueda ser llevado a un ambiente productivo

Exercise 4

Se cuenta con múltiples referencias de fuentes de datos (**6** en total) que hablan sobre los comportamientos macroeconomicos del país por medio de indicadores provenientes de fuentes de datos públicas como el *DANE*, La *Superfinanciera* y otros. El detalle de las fuentes es el siguiente:

- `DL_INDICADORES`: Fuente de hechos
- `DL_INDICMASTER`: Dimensión de indicadores
- `DL_INDICSOURC`: Dimensión de fuentes
- `DL_INDICPAIS`: Dimensión de países
- `DL_INDICREGIONS`: Dimensión de regiones (departamentos)
- `DL_INDICMPIOS`: Dimensión de municipios

Condiciones

- Las llaves de las fuentes son las siguientes:

Fuente	Campo	Referencia	Campo
DL_INDICADORES	ind_key	DL_INDICMASTER	KEY
DL_INDICMASTER	SOURCE	DL_INDICSOURC	ABR
DL_INDICADORES	COUNTRY	DL_INDICPAIS	COUNTRY_CODE
DL_INDICADORES	COUNTRY, REGION_CODE	DL_INDICREGIONS	COUNTRY_CODE, REGION_CODE
DL_INDICADORES	COUNTRY, REGION_CODE, CITY_CODE	DL_INDICMPIOS	COUNTRY_CODE, REGION_CODE, CITY_CODE

Nota: Es posible que deba aplicar algunos procesos de transformación y limpieza sobre las fuentes para que el cruce de información sea posible

Requerimiento

Construya un modelo de datos relacionales sobre éstas fuentes de modo que se genere una sábana de datos sobre la cual se puedan ejecutar consultas sobre el comportamiento de los indicadores y sus múltiples categorías en el tiempo. Más específicamente:

- Dibuje el modelo estrellado (https://www.draw.io/) con referencia a las *pk* y *sk* de las tablas
- Indique la cadena en lenguaje *SQL* que generaría la sábana solicitada haciendo referencia a las tablas

Ejercicio 5

Responda, de acuerdo a su experencia, búsqueda en la web, conocimiento general, o adivine

- ¿Cómo ejecuta uno un *request* o *llamada*?, ¿qué hay de por medio en dicha llamada?

- ¿Cómo se denomina al equipo que construye las bases de las soluciones digitales? ¿qué perfil de profesionales suelen encontrarse allí?
- ¿Qué equipo se encarga de exponer los resultados a los usuarios finales? ¿se requiere alguna habilidad especial en la etapa de exposición de resultados? ¿Qué formas de exponer resultados a los usuarios finales conoce usted? ¿Es importante el diseño en esta etapa?

Ejercicio 6

- Comente, con base a su experiencia y conocimiento, qué es *DevOps* y porqué es importante. En la medida de lo posible plasme el ciclo interno de la metodología y mencione diferentes herramientas en cada etapa
- Exponga su experencia con *git* y la importancia del versionamiento de código. ¿Con qué herramienta de *git* se siente usted más cómodo? ¿Qué forma de trabajo se requiere para que el desarrollo de software por medio de *git* sea eficiente?