ARGOS

Technical Proof Model Analyst

Throughout this test several of your technical and analitical skills will be evaluated to deal with a series of proposed exercises. Based on the initial context and the information provided, when required, it is expected that you apply your knowledge in descriptive analytics in order to present a business case. Depending on the task you might have to justify in technical terms your proposal and justify it against any other possible solution.

Exercise 1

- Explain in the most technical way what is *encoding* and why it is so important for any process related with data extraction and transformation.
- Have you ever had to work with *bash*? if so, what have you used it for? could you show an example of *bash* coding?

Exercise 2

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 3

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_INDICADORES	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

- 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

Exercise 4

- Comente, en 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
- Qué herramientas conoce usted de *integration services*?
- 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?

Exercise 5

En base al ejercicio exercise 3 responda a las siguientes preguntas de negocio:

- Para el PIB Total del país, según los datos, dónde se observa la mayor caída dada la tendencia histórica?
- Cuál de los segmentos tiene mayor peso para la región de Antioquia en el indicador de MERCADO CEMENTO DEPTO CANAL.
    - En qué período para el segmento encontrado tienden a haber menores ventas y para cuál mayores?
    - Dada una regresión lineal cuánto sería la estimación de ventas para Diciembre 2019?

Requerimiento

- Genere y comparta las sentencias SQL necesarias para responder a cada pregunta planteada
- Entregue gráficas que justifiquen su elección
- Indique el proceso de regresión lineal y los parámetros hallados

Exercise 6

Cargue los resultados previamente obtenidos en el exercise 5 (sábana de datos) a un GUI de *Python* (un jupyter notebook), *R* (R Studio), *Scala* (Java) o *PySpark* (Databricks Community)

Requerimiento

- Reproduzca las imágenes obtenidas en el ejercicio previo usando la librería que corresponda de acuerdo al lenguaje de predilección

Challenge

> Responda a los ejercicios según los requerimientos usando el planteamiento de los mismos y apoyándose en sus condiciones. 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 **2** horas y sólo el primer ejercicio deberá ser expuesto en *inglés*.