PRUEBA TECNICA ESPECIALISTA E INGENIERO INTEGRACION DE DATOS

INTEGRACION DE DATOS

Desarrollado por: Ing. Julian Andrés Quimbayo Castro

1. **¿Qué es una base de datos analítica?**

La base de datos analítica o más conocida como datawarehouse, es una base de datos que se conforma para el análisis y exploración de conocimiento que brinde inteligencia e insights significativos para el modelo de negocio que soporta. Algunas de las características más importantes son Homogeneidad y fiabilidad de la información, siempre la base es la consulta eficiente y jerarquía de esta. Se busca que la información se jerarquice y que se adecuan a distintos niveles de detalle y de diferentes usuarios.

Normalmente al realizar análisis de datos se tiene en cuenta datos de tipo presente con predicciones a futuro, pero el datawarehouse también apoya el hecho de datos históricos que permiten realizar mejores apuestas y tomar decisiones con base en información del pasado muy valiosa.

De igual forma se hace énfasis en una visión multidimensional que apalanca las diferentes áreas de la organización siendo fuente y respuesta de distintos procesos que deben tomar decisiones, generar dashboards, llevar control de proyectos y medir con base en indicadores.

1. **¿En qué se diferencia de una transaccional?**

La base de datos transaccional es un Sistema de Gestión de Base de Datos que se fundamenta en relaciones existentes entre las denominadas entidades, que son representaciones del mundo real. La diferencia mas significativa es el enfoque en la calidad y coherencia de datos ya que toma los datos de diferentes fuentes, los estandariza según restricciones uniformes (Datawarehouse). Esto posibilita que muchas áreas diferentes puedan consultar datos de otras áreas y así llegar a conclusiones en común o para debatir.

La correlación de la información es otro aspecto que las bases de datos relacionales carecen, ya que el datawarehouse permite potencializar el Bussiness intelligence tanto en procesos de venta, proveedores, clientes, estados financieros y estrategias de marketing y comerciales.

El alto retorno de inversión es otro factor ya que a medida que se integran sistemas BI, se reduce los tiempos de ejecución de dashboards, análisis y toma de decisiones esto supone en tiempo y hasta en cargos una mitigación sustanciable que se debe ver reflejado en la evolución del modelo de negocio.

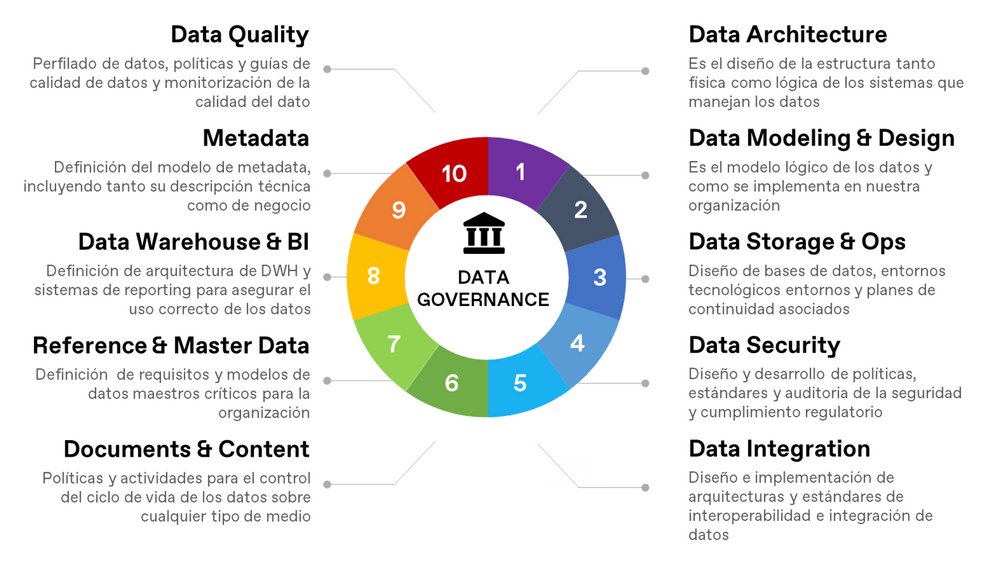
1. **¿Qué es una vista materializada?**

Una vista materializada es una tabla virtual, que muestra la representación de una consulta diseñada para arrojar cierto resultado y que puede ser modificado según las necesidades de reportes o análisis de datos.

1. **¿Qué es un hint?, ¿Cuáles conoce?**
2. Con base en el archivo adjunto diseñe un diagrama entidad relación que permita cargar la información normalizada.
3. Asumiendo que el archivo adjunto fue cargado en la tabla del schema de staging STG\_CO.STG\_ARCHIVO\_HISTORICO, escriba un código pl sql que cargue el modelo de datos del paso anterior, asuma que todos los campos de la staging son tipo varchar2(255).
4. Con base en la información cargada en las estructuras de datos generadas, escriba un query que genere un reporte con las personas con más de 2 siniestros en el año 2020, listando los certificados de esos siniestros.
5. Genere un query/vista en la que pueda consultar los clientes y sus certificados en estado vigente.
6. **¿Defina que es gobierno de datos? ¿Qué partes tiene?**

El gobierno de datos radica en la capacidad de que tiene una empresa para organizar su activo más preciado que son lo datos, conocer muy bien el origen y si están alineados con la política de la empresa.

Sus partes definidas son:



Es importante señalar que cada una de las partes del gobierno de datos implica la infraestructura, diseño, modelos, almacenamiento, seguridad, integridad, Calidad, documentación e inteligencia de negocio.

1. **¿Defina que es calidad de datos? ¿Cuáles son los criterios mínimos para tener un dato de calidad?**

La calidad de datos se define como el ajuste y perfilado que se realiza a los datos con base en las políticas, estándares y objetivos empresariales trazados, teniendo en cuenta un monitoreo constante de los mismos. Los criterios mínimos para que un dato se considere de calidad son:

* **Usabilidad:** La capacidad de poder utilizar el dato en diferentes momentos, presentaciones y procesos. De acuerdo con las necesidades de los colaboradores y clientes.
* **Confiabilidad:** Verificar el origen de los datos, fuentes, y que tan confiables son las fuentes de estos datos.
* **Pertinencia:** los datos deben establecerse en el flujo de tratamiento definido siendo acorde a varios de los procesos que se van a impactar, existirán algunos muy acertados para un proceso, pero debe procurarse que sea polifuncional y acorde a lo que se requiere para la toma de decisiones.
* **Disponibilidad:** 24/7 es la premisa de la infraestructura, modelos, diseños y tratamientos que permita acceder en cualquier momento y lugar siempre apalancando decisiones críticas.

1. **Defina la importancia de la calidad de los datos.**

La importancia de la calidad de datos radica en establecer un marco de trabajo que permita identificar, cargar, manipular, transformar, analizar y generar conocimiento significativo a partir de las políticas, objetivos y metas organizacionales trazadas. Esto permite pensar en diferentes procesos y organizar el flujo de tratamiento y monitoreo cosa que permita generar decisiones eficientes y conscientes de históricos de las fuentes recopiladas y así realizar proyecciones claras de tipo comercial, marketing, comercial, logísticas entre otras.

1. **Menciones tres atributos relacionados con la Calidad de los datos.**

Eficiencia: se balancea la eficiencia de los datos con base en el costo, esfuerzo, uso y límites de tiempo.

Validez: se define por tener consistencia en el tiempo y con los procesos que apalanca. Siempre aplicando el uso de estos hacia los requerimientos internos y externos.

Fiabilidad: consistencia para diferentes procesos y multipropósitos a través del tiempo.

Completo: evitar los datos inconsistentes, ruido en los datos y/o datos no válidos.

Relevancia: Revisión permanente durante el tiempo acorde a los requerimientos, objetivos y políticas claras de la organización.

Oportuno: para la toma de decisiones, es necesario determinar capturas de datos de manera rápida y a tiempo, que sea de fácil acceso y siempre disponible.

1. **Establezca un proceso técnico sencillo para la estandarización y mejora de una base de clientes para analizar los datos y darles remediación conforme a los atributos de data Quality.**

Propongo para el análisis de la base de clientes un proceso híbrido entre data quality y Team Data Science Process (TDSP) la cual es una metodología apoyada por Microsoft para el desarrollo de proyectos de ciencia de datos en equipo. Los pasos serían los siguientes:

* **Entendimiento del modelo de negocio**, en este paso se establecerá las metas para tener en cuenta en la mejora de base de datos del cliente. Indicadores de seguimiento.
* **Adquisición de la data y entendimiento de las fuentes**, en esta fase se determina si los datos son suficientes, consistentes, pertinentes y validar si cumplen con lo que el requerimiento necesita, sumado a las dimensiones para consolidar un dato de calidad.
* **Asignación de Tareas (Sprint)**, en este paso se determinan los criterios para tener en cuenta según los requerimientos definidos y con una combinación de agilismo se asigna el sprint, roles, pesos y recursos necesarios para pulir la base de datos.
* **Análisis de la base de datos**, en este paso se revisa porque hay datos anómalos, cuáles son las causas de que los datos no estén completos, si para dicha base de datos algunos campos son necesarios y oportunos.
* **Infraestructura para BI,** con base en los datos ya normalizados y estandarizados se definirá un Datawarehouse para múltiples propósitos tanto en reportes, informes establecidos, Procesos de Machine learning y futuras decisiones.
* **Control y feedback,** Se establecen políticas de control, seguimiento y retroalimentación para la mejora de infraestructura, procesos y soluciones.

1. **Mencione lo más importante a tener en cuenta en una herramienta de data Quality.**

Algunos aspectos a tener en cuenta para una herramienta de data quality es la necesidad de que cumpla con cuatro factores muy importantes los cuales son:

* **Gobernanza de Datos**, permita incorporar políticas, estándares y pilares para el manejo de los datos y cumplir con los objetivos propuestos.
* **Elaboración de perfiles de datos**, asignación de roles, seguimientos y control a los procesos de extracción, carga y transformación de la data con una disponibilidad 24/7.
* **Mantenimiento de datos**, esta herramienta debe permitir procesos periódicos de data cleansing, prevención, detección de errores y reparación de datos.
* **Integración de datos,** debe brindar la posibilidad de conectar los sistemas y bases de datos de diferentes áreas de la organización donde se sincronice de manera actualizada y accesible los datos.

1. **Mencione 3 herramientas informáticas posibles pata el seguimiento y presentación de indicadores de Calidad de Datos.**

Algunas de las mejores herramientas para el seguimiento y presentación de indicadores de calidad de datos son:

Oracle Cloud Infrastructure Data Catalog

SAS

SAP Master Data Governance