Ingeniería de Datos

Desbloqueando el potencial de tus datos

Presentación para el curso de Ingeniería de Características, Maestría en Ciencia de Datos, UNISON

¿Qué es la Ingeniería de Datos?

1

Diseño de Sistemas

Disciplina que diseña y construye sistemas a gran escala para recopilar, almacenar y preparar datos.

2

Calidad y Acceso

Garantiza que los datos estén limpios, consistentes y accesibles para científicos y analistas.

3

Base Fundamental

Es la base fundamental para una cultura de datos robusta en las empresas modernas.

La ingeniería de datos es el pilar que sostiene todas las iniciativas de análisis avanzado y toma de decisiones estratégicas basadas en información.

Ciclo de Vida de la Ingeniería de Datos

Generación

Captura de datos desde múltiples fuentes, como IoT, bases de datos o aplicaciones.

Almacenamiento

Guardar datos en sistemas escalables y confiables (relacionales, NoSQL, nube).

Ingesta

Integración y traslado de datos a plataformas centralizadas, preparando el camino para el análisis.

Transformación

Limpieza, normalización y aplicación de reglas de negocio para asegurar la utilidad del dato.

Servicio

Entrega de datos procesados para análisis y toma de decisiones informadas.

Rol del Ingeniero de Datos vs. Científico de Datos

Ingeniero de Datos

Construye y mantiene la infraestructura de datos, asegurando su calidad y disponibilidad.

- Diseño de bases de datos.
- Desarrollo de pipelines ETL.
- Optimización de sistemas de almacenamiento.

Científico de Datos

Analiza, modela y extrae conocimiento valioso usando los datos preparados.

- Desarrollo de modelos predictivos.
- Análisis estadístico.
- Visualización de resultados.

Ambos roles son interdependientes; la colaboración estrecha es clave para transformar los datos en valor empresarial real.

Tecnologías y Herramientas Clave

Ingesta

Apache Kafka: Plataforma de streaming distribuida de alto rendimiento. **Google Dataflow:** Servicio de procesamiento de datos unificado para batch y streaming.

Procesamiento

Apache Spark: Motor de análisis unificado para procesamiento de Big Data. **Airflow:** Plataforma para programar y monitorear flujos de trabajo de datos.

Almacenamiento

Amazon S3: Almacenamiento de objetos escalable y seguro. **Google BigQuery:** Almacén de datos totalmente gestionado y sin servidor. **MongoDB:** Base de datos NoSQL para datos no estructurados.

Visualización

Tableau: Herramienta líder en inteligencia de negocios. **Power BI:** Servicio de análisis de negocios de Microsoft. **Jupyter Notebooks:** Entorno interactivo para desarrollo de ciencia de datos.

Caso Práctico: Pipeline ETL para Análisis de Ventas

1. Extracción

Datos de ventas brutos obtenidos de sistemas ERP y plataformas de e-commerce.

2. Transformación

Limpieza de datos, unificación de formatos y cálculo de indicadores clave de rendimiento (KPIs).

3. Análisis Exploratorio de Datos

Identificación de patrones, anomalías y relaciones en los datos para guiar la limpieza y el modelado.

4. Carga

Datos procesados cargados en un Data Warehouse optimizado para consultas rápidas.

5. Contar Historias con Datos

Presentación de los hallazgos clave a través de visualizaciones interactivas, como dashboards en tiempo real, para facilitar la toma de decisiones ágiles y estratégicas.

Este ciclo completo transforma los datos crudos en inteligencia de negocio accionable, impulsando la estrategia de ventas.

Beneficios de una Arquitectura Sólida

Optimización de la Escalabilidad

Facilita la gestión de volúmenes masivos y una amplia diversidad de datos, permitiendo una adaptación eficiente al crecimiento exponencial y a las demandas cambiantes del entorno empresarial.

Fiabilidad de los Datos

Garantiza una mejora significativa en la calidad y la coherencia de los datos, lo que contribuye a la minimización de errores y a la mitigación de posibles sesgos inherentes a la información.

Adaptabilidad y Resiliencia

Promueve la integración expedita de nuevas fuentes de datos y la incorporación proactiva de tecnologías emergentes, fortaleciendo así la capacidad de respuesta organizacional ante escenarios dinámicos.

Catalizador de la Innovación

Proporciona un fundamento robusto para el desarrollo e implementación de iniciativas en el ámbito de la Inteligencia Artificial y el Aprendizaje Automático, posibilitando la identificación y capitalización de nuevas oportunidades estratégicas.

Calidad del Dato

Precisión

Asegura que los datos sean correctos, fiables y reflejen la realidad, minimizando errores y desviaciones.

Completitud

Garantiza que no falte información crítica, cubriendo todos los campos necesarios para un análisis exhaustivo.

Consistencia

Mantiene la uniformidad de los datos a través de diferentes sistemas y fuentes, evitando contradicciones.

Actualidad

Procura que los datos estén disponibles y sean pertinentes en el momento adecuado, facilitando decisiones oportunas.

Validez

Confirma que los datos cumplen con las reglas de formato, rango y tipo predefinidas, siendo coherentes con su propósito.

Una alta calidad del dato es crucial para obtener insights fiables y potenciar la inteligencia de negocio.

Gobernanza de Datos

Políticas y Estándares

Define las reglas y principios para la creación, uso, almacenamiento y eliminación de datos, asegurando la consistencia.

Gestión de Acceso y Seguridad

Establece quién tiene permiso para ver, modificar o usar los datos, protegiéndolos contra accesos no autorizados.

Cumplimiento Normativo

Garantiza que la organización cumpla con las leyes, regulaciones y estándares de la industria relacionados con los datos.

Auditoría y Monitoreo

Supervisa el uso de los datos para asegurar la adherencia a las políticas y detectar posibles anomalías o infracciones.

La gobernanza de datos es esencial para maximizar el valor de la información y minimizar los riesgos en el entorno empresarial actual.

Conclusión:

Ingeniería de Datos, el Pilar de la Ciencia de Datos

"Sin una arquitectura robusta y procesos eficientes, la ciencia de datos no puede prosperar."

Invertir en ingeniería de datos es invertir en decisiones basadas en datos confiables, que transforman información en impacto real.