Proyectos de datos masivos y ciencia de datos

**Introducción**

La gestión y análisis de grandes volúmenes de datos es crucial en el entorno digital actual. Entender las arquitecturas típicas, las estrategias de almacenamiento y las aplicaciones prácticas permite a las organizaciones tomar decisiones informadas y obtener ventajas competitivas. Además, conocer el perfil del científico de datos y sus habilidades complementa el entendimiento del proceso de transformación de datos en información valiosa. Esta actividad está diseñada para que los estudiantes trabajen de manera práctica en un proyecto que abarque todos estos temas de manera integral.

**Objetivo**

Aplicar de manera práctica e integrada los conocimientos sobre arquitecturas de proyectos de datos masivos, estrategias de almacenamiento, el perfil del científico de datos y las áreas de aplicación de la ciencia de datos.

**Objetivos específicos de la actividad**

* Comprender las **fuentes heterogéneas de datos:** identificar y clasificar diferentes tipos de fuentes de datos (estructurados, semiestructurados, no estructurados) y comprender sus características y desafíos.
* Diseñar **procesos ETL:** crear un esquema detallado del proceso ETL, describiendo métodos de extracción, transformación y carga de datos, y seleccionar herramientas ETL adecuadas para diferentes tipos de datos.
* Planificar la **arquitectura de almacenamiento:** diseñar una arquitectura de almacenamiento de datos que incluya soluciones como Data Lake y Data Warehouse, explicando las ventajas y aplicaciones de cada uno.
* Definir **estrategias de tratamiento de datos:** proponer estrategias para la limpieza, integración y análisis de datos, asegurando la calidad y coherencia de los datos para su posterior análisis.
* Desarrollar **planes de visualización de datos:** diseñar *dashboards* interactivos y planificar la visualización de datos utilizando herramientas como Tableau, Power BI o matplotlib para representar métricas clave y tendencias relevantes.
* Comprender el **perfil del científico de datos:** analizar las competencias necesarias en ciencias de la computación, matemáticas, comunicación y negocios que forman el perfil del científico de datos.
* Explorar **estrategias en almacenamiento masivo:** investigar y entender las diferencias entre Data Mart, Data Warehouse y Data Lake, y nuevas tendencias en almacenamiento masivo.
* Aplicar **estrategias de ciencia de datos y datos masivos:** proponer aplicaciones prácticas de la ciencia de datos en inteligencia de negocio, analítica de negocio, minería de datos, aprendizaje automático e inteligencia artificial en distintos sectores.

**Nota importante**

El objetivo de esta actividad no es desarrollar el proyecto completo, sino únicamente diseñarlo utilizando toda la creatividad posible. Los estudiantes deben enfocarse en conceptualizar y planificar cada etapa del proyecto, desde la obtención de datos hasta la visualización y análisis final. Las herramientas y habilidades necesarias para llevar a cabo el desarrollo real del proyecto se aprenderán en las asignaturas específicas del máster. Este ejercicio busca que los estudiantes apliquen y demuestren su comprensión teórica y su capacidad para estructurar un proyecto de datos masivos de manera integral.

Diseño de un proyecto de ciencia de datos

**Formación de equipos**

* Forma equipos de 4-5 estudiantes. Cada equipo asumirá el rol de un grupo de científicos de datos que trabajan en un proyecto para una empresa ficticia.

**Selección de un caso de estudio**

Cada equipo seleccionará uno de los siguientes sectores para desarrollar su proyecto:

* Comercio.
* Industria.
* Salud.
* Seguridad y ciberseguridad.
* Finanzas.
* Conducción autónoma.

**Diseño del proyecto**

1. Arquitecturas típicas de proyectos de datos masivos.

* **Fuentes heterogéneas.** Identifiquen y documenten al menos tres fuentes de datos heterogéneas relevantes para el sector seleccionado. Describan cómo se obtendrán los datos de cada fuente.
* **Extracción, transformación y carga (ETL).** Diseñen un flujo de proceso ETL para consolidar los datos de las fuentes identificadas. Elaboren un diagrama de flujo y seleccionen herramientas ETL adecuadas, justificando su elección.
* **Almacenamiento**. Elijan una solución de almacenamiento adecuada (Data Warehouse, Data Lake, NoSQL) para los datos transformados. Justifiquen su elección y describan cómo estructurarían el almacenamiento.
* **Tratamiento de los datos.** Describan los pasos necesarios para tratar los datos, incluyendo limpieza, integración y preparación para el análisis.
* **Visualización.** Seleccionen herramientas de visualización de datos y diseñen un *dashboard* básico que represente información clave del proyecto.

1. Perfil del científico de datos.

* **Ciencias de la computación.** Describan las habilidades técnicas necesarias para implementar el proyecto y cómo cada miembro del equipo contribuye con sus conocimientos técnicos.
* **Matemáticas.** Expliquen las técnicas estadísticas y matemáticas que se aplicarán en el análisis de los datos.
* **Comunicación.** Diseñen una estrategia para comunicar los hallazgos del proyecto a audiencias no técnicas, incluyendo un informe ejecutivo y presentaciones visuales.
* **Negocios.** Identifiquen los objetivos estratégicos del sector seleccionado y expliquen cómo el proyecto de datos contribuirá a lograr esos objetivos.

1. Estrategias en almacenamiento masivo.

* **Data Mart.** Diseñen un Data Mart específico para un área del negocio del sector seleccionado, describiendo las tablas y datos incluidos.
* **Data Warehouse.** Expliquen cómo se estructuraría un Data Warehouse para el proyecto y cómo se integraría con el Data Mart.
* **Data Lake.** Describan los beneficios de utilizar un Data Lake para el proyecto y cómo se gestionaría la ingesta de datos no estructurados.
* **Nuevas tendencias en almacenamiento masivo.** Investiguen y propongan una nueva tendencia en almacenamiento masivo (por ejemplo, almacenamiento en la nube) que podría mejorar el proyecto.

1. Estrategias de aplicación de la ciencia de datos y datos masivos.

* **Inteligencia de negocio.** Expliquen cómo se utilizará la inteligencia de negocio para mejorar la toma de decisiones en el proyecto.
* **Analítica de negocio.** Diseñen un análisis de negocio utilizando los datos disponibles para identificar oportunidades o problemas.
* **Minería de datos.** Apliquen técnicas de minería de datos para descubrir patrones o tendencias en los datos.
* **Aprendizaje automático.** Implementen un modelo de aprendizaje automático simple (utilizando una herramienta o librería) para predecir un aspecto relevante del sector seleccionado.
* **Inteligencia artificial.** Discutan cómo la inteligencia artificial podría aplicarse en el futuro para mejorar aún más el análisis de datos en el proyecto.

**Presentación del proyecto**

Cada equipo preparará una presentación que incluya:

1. Un resumen del proyecto y su objetivo.
2. El diagrama del flujo ETL.
3. La estructura de almacenamiento y tratamiento de datos.
4. El diseño del *dashboard* de visualización.
5. La aplicación de técnicas de ciencia de datos.
6. Un informe ejecutivo que resuma los hallazgos y recomendaciones del proyecto.

**Presentación oral en vídeo**

La presentación oral se realizará en un vídeo de 6 minutos.

Recomendaciones para el video:

* **Estructura.** Organiza el contenido de manera clara y lógica. Incluye una introducción, desarrollo y conclusión.
* **Claridad.** Habla de manera clara y pausada. Evita el uso de jerga técnica innecesaria.
* **Visuales.** Utiliza diapositivas, gráficos y diagramas para ilustrar tus puntos. Asegúrate de que los visuales sean legibles y relevantes.
* **Tiempo.** Practica tu presentación para asegurarte de que cumples con el límite de tiempo de 6 minutos.
* **Equipo.** Asegúrate de que todos los miembros del equipo participen en la presentación.
* **Calidad.** Graba el vídeo en un ambiente tranquilo y bien iluminado. Verifica la calidad del audio y vídeo antes de la entrega.

**NOTA:** No está permitido el plagio entre compañeros de este o anteriores cursos, esta es una actividad grupal. Tampoco está permitido el plagio de fuentes de Internet, libros, manuales o de cualquier otro tipo que no se citen apropiadamente. Entregar un ejercicio que incurra en los anteriores escenarios implicará automáticamente un 0 como calificación de la actividad, sin perjuicio de que se tomen medidas disciplinarias de acuerdo con la correspondiente normativa de la UNIR. Está permitido utilizar código, fracciones de código o ejemplos existentes en Internet u otras fuentes **siempre que se citen apropiadamente y su uso se limite a partes minoritarias de la actividad.**

**Extensión**

Extensión máxima de la actividad: un único informe de 15 páginas como máximo de extensión (Arial 11, interlineado 1,5). El informe no debe incluir anexos, debe llevar portada (esta no cuenta en la extensión del informe) y no debe repetir el enunciado o la rúbrica de la actividad. Cualquier página que incluyas será computada dentro de la extensión entregada. No se evaluará ninguna página más allá de las 15 páginas. No es posible adjuntar ningún archivo adicional.