

Planificación Docente Completa

Clase 1, Introduction to Analytics: Multidimensional Modeling:

- Comprender cómo se lleva a cabo un diseño multidimensional para el análisis.
- Conocer cómo se corresponden los elementos de un diseño multidimensional con las tablas del repositorio de datos.

Trabajo previo necesario:

Lectura de la presentación Data Analytics – Data Warehousing

Tarea 1: Creación de un esquema multidimensional en papel o en un editor de imágenes a preferencia del alumno a partir de una descripción de un caso de estudio real. Para esta tarea se facilitará tanto la descripción del caso de estudio como un conjunto de fuentes asociados al mismo que deberán ser tenidos en cuenta a la hora de diseñar el esquema. El alumno deberá entregar un PDF con el diseño que ha realizado tanto a nivel conceptual como lógico, describiendo cómo se reflejan los conceptos del caso de estudio en el esquema.

Clase 2, Multidimensional Modeling Implementation & ETL Processes:

- Implementar un esquema multidimensional en una solución tecnológica
- Iniciarse en el funcionamiento de los procesos de Extracción, Transformación y Carga (ETL) y su cometido en una arquitectura de almacenes de datos.

Trabajo previo necesario: Lectura de la Guía Básica de Mondrian Schema Workbench, instalación de Mondrian Schema Workbench

Tarea 2: Implementación de un esquema multidimensional utilizando Mondrian Schema Workbench. Para esta tarea se utilizará como entrada el diseño multidimensional realizado en la Tarea 1. El alumno deberá entregar un pdf en el que se describirá cómo quedan cubiertos los hechos, dimensiones y jerarquías del modelo mediante los elementos del esquema de Mondrian y el correspondiente modelo lógico en MySQL.

Clase 3, ETL Processes Implementation:

- Conocer los aspectos fundamentales a tener en cuenta a la hora de diseñar procesos ETL
- Conocer el funcionamiento de una herramienta de Extracción, Transformación y Carga (ETL) y como estructurar una solución ETL

Trabajo previo necesario: Lectura de la Guía de Instalación y Uso de Pentaho Data Integration

Tarea 3: Implementación de un proceso de transformación de datos a partir de un conjunto de fuentes operacionales a un almacén. Los alumnos deberán crear una serie de procesos ETL mediante procesos de Pentaho Data Integration que extraerán los datos de las fuentes y los cargarán en la base de datos de la Tarea 2. Los alumnos deberán entregar un pdf en el que describan la finalidad de cada paso de carga junto a los archivos de los procesos ETL y la base de datos.

Trabajo recomendado del lunes: Completar la tarea 3

Clase 4, Building a data warehouse solution:

- Comprender cómo se realiza un despliegue de los módulos creados hasta el momento en Pentaho y la responsabilidad de cada uno
 - Comprender el punto de vista del usuario en el uso de soluciones de almacenes de datos
 - Conocer otros módulos existentes y soluciones no vistas en clase
- Trabajo previo necesario: Lectura de la Guía de Instalación y Uso de Pentaho

Tarea 4: Despliegue y uso de un sistema analítico. Los alumnos deberán, utilizando los archivos creados hasta el momento, y los que fuesen necesarios añadir, desplegar un sistema de BI funcional. Para ello, deberán:

A) Arrancar el servidor de BI de Pentaho y crear una conexión a la base de datos del almacén. Una vez realizado, deberán crear y publicar el cubo de

Mondrian que representa el almacén en el servidor de BI. Finalmente, deberán llevar a cabo un análisis OLAP mediante la vista de análisis de Pentaho.

B) Arrancar Power BI Desktop y establecer una conexión con la base de datos del almacén. Una vez realizado, deberán crear una serie de cuadros de mando que permitan un análisis de los datos similar al que llevarían a cabo mediante análisis OLAP.

En cualquiera de los dos casos, los alumnos deben de proponer y responder a 4 preguntas que requieran información de al menos 2 dimensiones para poder responderse. Los alumnos deberán entregar un pdf con las preguntas propuestas y capturas de pantalla del análisis realizado para responderlas.

Clase 5, Analytics over Big Data sources: Data Science fundamentals:

- Conocer la problemática del análisis de Big Data y cómo se opera en un entorno de fuentes de Big Data: qué roles existen y qué pasos se siguen para llevar a cabo Data Science
- Comprender cuándo es necesario utilizar técnicas de Data Science o de almacenes de datos y las limitaciones de cada una

Tarea 5a: Búsqueda de información de Big Data. Los alumnos deberán de buscar al menos 2 fuentes de Open Data que estén relacionadas por su temática. Estas fuentes deberán ser descargadas y descritas: ¿De qué tratan? ¿Qué volumen de información aproximado tienen? ¿Por qué podrían resultar de interés? Finalmente, los alumnos deben de proponer al menos 4 preguntas que creen que se pudiesen responder a partir de las fuentes de forma individual o conjunta.

Clase 6, Big Data Analytics:

- Conocer las herramientas existentes para las distintas tareas necesarias en el contexto de Data Science y utilizarlas para preparación y transformación de datos.
- Saber llevar a cabo un proceso de perfilado y análisis sobre fuentes de Big Data Trabajo previo necesario: Lectura de la Guía de Apache Zeppelin, lectura de [este artículo](#).

Tarea 5b: Data profiling y visualización de datos mediante Apache Zeppelin. Los alumnos deberán cargar las fuentes de Open Data de la Tarea 6 mediante PySpark, y llevar a cabo un proceso de profiling para analizar las fuentes de datos. A continuación, deberán responder a las preguntas formuladas en la Tarea 6 mediante la creación de distintas visualizaciones en Apache Zeppelin. Los alumnos deberán entregar un pdf con las capturas de pantalla de Apache Zeppelin correspondientes a la carga, perfilado, visualizaciones, y las respuestas a las preguntas formuladas.

Trabajo recomendado del lunes: Tarea 6 (Challenge): Creación de un almacén de datos a partir del modelo de Big Data obtenido en la tarea 6. Para esta tarea, los alumnos se ayudarán del profiling realizado en cada fuente, de forma que puedan seleccionar hechos y dimensiones de interés para ser integrados en un posible almacén de datos. Los alumnos deberán entregar un pdf con el esquema multidimensional (a nivel conceptual) del almacén de datos, razonando qué datos se incluyen como hechos o dimensiones, cuáles no y por qué. Como apoyo a la justificación, los alumnos pueden incluir capturas del perfilado de datos realizado en Zeppelin que apoyen su selección.