Guía de actividad – Unidad 4
Entrega 2: Implementación del Proceso ETL y carga de datos.

Identificación

Asignatura: Base de Datos III.

Unidad: Desarrollo de un proyecto Datawarehouse.

Objetivo

- Obtener los conocimientos necesarios para la implementación de los Sistemas que requieren análisis y recopilación de datos basados en un Datawarehousing.
- Aplicar los conocimientos desarrollando un proyecto utilizando las tecnologías aprendidas.

Actividad

Realizar las siguientes actividades:

- Leer la Guía del Trabajo Práctico actualizada, para entender el contexto del trabajo total a ser desarrollado.
- Con el conjunto de datos asignado para cada grupo, como sistema origen, implementar el proceso para la carga de datos en la base de datos multidimensional. El conjunto de datos asignados es el siguiente:

Tema	Descripción	Grupos	Fuentes de Datos
Descarga de aplicaciones	Contiene datos de las descargas del Google Play Store como tipo, precio, género, entre otros.	1 y 7	Descarga csv
Asesinatos	Contiene datos sobre personas asesinadas como nombre, edad, sexo y raza, también información sobre como fueron asesinados, si éstos atacaron, portaban armas, mostraban alguna enfermedad mental, entre otros.	2 y 8	Descarga csv

Pre condiciones en pacientes COVID-19	Contiene información de los pacientes como historial y hábitos, además los resultados de los chequeos realizados.	3 y 9	Descargar csv
E-commerce de ropas	Contiene una lista de productos con sus calificaciones y datos de ventas.	4 y 10	Descargar csv
Airbnb en New York	Contiene información necesaria sobre los hostales en las diferentes ciudades.	5 y 11	Descargar csv
Trabajos de Analistas de datos	Contiene datos de las empresas que contratan analistas de datos.	6 y 12	Descargar csv

- Preparar el entorno de trabajo, descargando e instalando las herramientas a ser utilizadas.

Stack Tecnológico	Información General	Descargar e Instalar
	SQL Power Architect es una herramienta de modelado de datos creados por los diseñadores de almacenamiento de datos y tiene muchas características únicas, dirigidas específicamente para el arquitecto de almacenamiento de datos.	Descargar SQL Power Architect
 PostgreSQL	PostgreSQL, es un potente **sistema de base de datos** relacional de objetos de código abierto. PostgreSQL - Documentaciones.	Descargar PostgreSQL.
	Pentaho Data Integration (PDI) - Kettle, es una herramienta de la suite de Pentaho para el **desarrollo y ejecución del proceso ETL**. Se encarga de la Extracción de datos de una fuente, Transformación de esos datos y Carga de esos datos en otro sitio. Pentaho Data Integration - Sitio Oficial.	Instalar versión 8.0 o superior. (Recomendado 9.0). Descargar PDI Community Edition.

Power BI	Power BI Desktop, es una herramienta para crear sofisticados **informes interactivos con análisis visuales** de forma gratuita. Sistema operativo compatible sólo Windows. Power BI Desktop - Sitio Oficial.	Instalar versión 2.7 (2019) o superior. (Recomendado 2.8) Descargar Power BI Desktop.

- Para los grupos que aún no han instalado la herramienta Pentaho, pueden utilizar la guía de instalación que se encuentra en Educa.
- Crear un proyecto con Pentaho Data Integration.
- Crear la conexión al sistema origen para la extracción de los datos.
- Crear la conexión a la base de datos multidimensional para la carga de los datos.
- Implementar los procesos de extracción, transformación y carga de las dimensiones simples (tipo 1), utilizando los componentes de la herramienta.
- Implementar los procesos de extracción, transformación y carga de las dimensiones con lógica de carga (tipo 2 al 4), utilizando los componentes de la herramienta.
- Implementar los procesos de extracción, transformación y carga de la tabla de hechos, utilizando los componentes de la herramienta. En el caso de tener más de una tabla de hechos, implementar los procesos para todas las tablas.
- En caso de ser necesario, implementar el tratamiento de errores o excepciones.
- Validar que la carga realizada sea completa y correcta. En el caso de ser incorrecta o incompleta corregir los pasos anteriores y volver a cargar los datos.
- Presentar el proyecto implementado a los Docentes para la evaluación del mismo. La presentación se realizará a través de una actividad sincrónica con el grupo y el docente evaluador asignado. La fecha de presentación será el jueves **14/10** en el horario de clases, con horario definido y duración de **10 minutos** para cada grupo. *En el caso de que un alumno tenga inconvenientes con el horario, por temas laborales o de salud, comunicar a los docentes en la brevedad posible para evaluar alternativas.*
- Los horarios para las presentaciones del proyecto para el día jueves **14/10** serían:

Grupo	Tema	Horario	Docente evaluador

1	Descarga de aplicaciones	19:00 a 19:15 Hs	Romina Rojas y Delia Villasanti
2	Asesinatos	19:15 a 19:30 Hs	Romina Rojas y Delia Villasanti
3	Pre condiciones en pacientes COVID	19:30 a 19:45 Hs	Romina Rojas y Delia Villasanti
4	E-commerce de ropas	19:00 a 19:15 Hs	Ysacio Rejala y Natalia Barros
5	Airbnb en New York	19:15 a 19:30 Hs	Ysacio Rejala y Natalia Barros
6	Trabajo de analistas de datos	19:30 a 19:45 Hs	Ysacio Rejala y Natalia Barros
7	Descarga de aplicaciones	19:45 a 20:00 Hs	Romina Rojas y Delia Villasanti
8	Asesinatos	20:00 a 20:15 Hs	Romina Rojas y Delia Villasanti
9	Pre condiciones en pacientes COVID	20:15 a 20:30 Hs	Romina Rojas y Delia Villasanti
10	E-commerce de ropas	19:45 a 20:00 Hs	Ysacio Rejala y Natalia Barros
11	Airbnb en New York	20:00 a 20:15 Hs	Ysacio Rejala y Natalia Barros
12	Trabajo de analistas de datos	20:15 a 20:30 Hs	Ysacio Rejala y Natalia Barros

Consideraciones generales:

- En la plataforma EDUCA encontrarán una actividad que no requiere entrega o subida de archivo, la misma servirá para visualizar la calificación obtenida de la entrega.
- Puntaje: 100.
- Peso Total del Trabajo Práctico para el PP: 30%
- Peso de este trabajo para el Total del TP: 40 %
    - La entrega posterior a la fecha límite no califica.
    - El incumplimiento de las normativas será penalizado.

Plazo

La entrega 2 del proyecto estará disponible **desde el 27/09/2021 hasta el 14/10/2021** en el horario definido para cada grupo.

Evaluación

La Entrega 2.1 considera la entrega del proyecto de inserción de dimensiones para la evaluación previa por parte de los docentes y poder realizar las correcciones necesarias para la entrega 2.3. La calificación de esta entrega será de 100% y el criterio a ser considerado es la entrega puntual del proyecto de inserción de dimensiones. Para entregas fuera de plazo tendrá una penalización de 20%.

La Entrega 2.2 considera la entrega del proyecto de inserción de hechos para la evaluación previa por parte de los docentes y poder realizar las correcciones necesarias para la entrega 2.3. La calificación de esta entrega será de 100% y el criterio a ser considerado es la entrega puntual del proyecto de inserción de hechos. Para entregas fuera de plazo tendrá una penalización de 20%.

Los criterios y niveles de logro a considerar con los respectivos puntajes para la entrega final 2.3 son:

Aspecto	Peso %
Cumple con el diseño propuesto en el entregable 1 - Verificar que las dimensiones hayan sido poblados según tipos - Verificar que los hechos hayan sido poblados según tipos	10
ETL carga las dimensiones correctamente - Verificar cantidad de registros que se deben insertar en cada dimensión según conjunto de datos - Verificar que no duplique registros si se ejecuta por segunda vez sin cambios en los datos	10
ETL de dimensiones cumplen con el tipo propuesto en el Diseño (Slowly changing tipo 1 al 4, Rapid changing o Junk dimension) - Probar cambios de datos en conjunto de datos y verificar la actualización - Probar inserción de registros nuevos	10
ETL carga la tabla de hechos correctamente	10

- Verificar cantidad de registros que se deben insertar en cada hechos según conjunto de datos - Verificar que no duplique registros si se ejecuta por segunda vez sin cambios en los datos	
ETL de tabla de hechos presenta la granularidad propuesta en el diseño (transactional, period snapshot, accumulating snapshot) - Probar insertar una transacción, período o proceso nuevo	10
Las dimensiones presentan gestión correcta ante updates de registros en origen - Probar un cambio en una dimensión y luego la inserción de una nueva transacción, período o proceso asociada a dicho cambio	05
ETL incluye tratamiento de errores o excepciones - Probar inserción de nulos en dimensiones y en medidas	05
El orden de inserción respeta relaciones entre tabla de hechos y dimensiones - Ejecutar todos los pasos descritos anteriormente según metodología y orden correcto, en una base de datos totalmente vacía	10
Query sobre las dimensiones se ejecuta correctamente - Las verificaciones mencionadas anteriormente deben ser con querys en la base de datos para corroborar la inserción correcta	10
Query sobre la tabla de hechos se ejecuta correctamente - Las verificaciones mencionadas anteriormente deben ser con querys en la base de datos para corroborar la inserción correcta	10
Entrega puntual - Evaluación completa se realiza el jueves 14/10	10
TOTAL	100

La evaluación del proyecto final se realizará de la siguiente manera:

Actividad	Puntos	Peso para calcular el % de Trabajo Práctico
Entrega 1.1: Diseño multidimensional borrador	100	5 %
Entrega 1.2: Diseño multidimensional e implementación del modelo de datos	100	25 %

Entrega 2.1: Proceso ETL y carga de datos de dimensiones	100	5 %
Entrega 2.2: Proceso ETL y carga de datos de hechos	100	5 %
Entrega 2: Proceso ETL y carga de datos en el DWH.	100	30 %
Entrega 3: Creación de reportes y Dashboard. Informe final.	100	30 %
Puntaje Proyecto Final	**100%**	**100%**

Se considera entrega tardía hasta el martes **19 de octubre en el horario a definir con los docentes evaluadores asignados**, restando los puntos correspondientes a la entrega puntual. *Si ninguno de los integrantes del grupo puede presentar en el horario propuesto del jueves 14/10 o no están listos para entregar el trabajo, comunicar a los docentes hasta el miércoles 13/10 para coordinar el horario para el martes 19/10.*

La retroalimentación de los docentes se entregará hasta el jueves 21 de octubre.