UNIVERSIDAD PEDAGÓGICA Y TECNOLÓGICA DE COLOMBIA
FACULTAD DE INGENIERÍA
ESCUELA DE SISTEMAS Y COMPUTACIÓN
BASES DE DATOS II
PROYECTO

Descripción del proyecto:

El objetivo del proyecto es realizar la carga y manipulación de datos a partir de un archivo plano en formato CSV. Se requiere llevar a cabo un proceso de normalización y construcción de un modelo de datos que responda a una necesidad de negocio, de acuerdo con el tipo de dataset seleccionado.

El proyecto debe contemplar el modelado de datos siguiendo buenas prácticas, así como la implementación de funciones, bloques anónimos, procedimientos y packages. Además, se deberán incluir triggers que controlen parte de la lógica de negocio o funcionen como mecanismos de auditoría.

El desarrollo se realizará en grupos de dos estudiantes. Cada grupo podrá escoger el dataset de su preferencia, idealmente desde la plataforma kaggle. El único requisito es que el conjunto de datos contenga al menos una variable medible y cuantificable, preferiblemente con información histórica de más de tres años. En caso de no ser posible, se aceptarán datasets con datos mensuales, siempre que cuenten con un volumen significativo de información.

Actividades:

Descargar Datos:

- Descargar un set de datos desde la plataforma https://www.kaggle.com/datasets

Crear Usuario y Gestión de almacenamiento:

- Crear un usuario o esquema con permisos específicos para realizar operaciones CRUD, así como para crear vistas, procedimientos almacenados triggers, etc. Es importante otorgar únicamente los permisos necesarios, ni más ni menos. No está permitido asignar roles como ~~DBA~~ o ~~GRANT RESOURCE~~.
- Realizar la separación del almacenamiento de datos e índices: las tablas deben almacenarse en un tablespace exclusivo para datos, mientras que los índices deberán almacenarse en un tablespace diferente.

Cargar Archivo CSV:

- Utilizar SQL*Loader o herramientas como SQL Developer o PL/SQL Developer para cargar el archivo CSV en una tabla temporal. Esta tabla deberá contener las columnas originales del archivo plano. En caso de ser necesario, se deberá realizar un proceso de limpieza y depuración de datos antes de continuar con su transformación o integración al modelo final.

Poblado:

- Desarrollar los procesos de carga y transformación necesarios para poblar las tablas finales a partir de la tabla temporal, siguiendo la estructura definida en el modelo de datos propuesto y en función del dataset seleccionado. Para cada tabla final, se deberá implementar un procedimiento almacenado (stored procedure) independiente que realice la carga correspondiente
- Crear una tabla adicional de control para llevar el seguimiento de los procesos de carga realizados. Esta tabla debe registrar, al menos, los siguientes campos:
    - **nombre_tabla:** nombre de la tabla que fue poblada.
    - **filas_afectadas:** cantidad de registros insertados, actualizados o eliminados.
    - **operación:** tipo de instrucción realizada (INSERT, UPDATE o DELETE).
    - **fecha_proceso:** fecha y hora en la que se ejecutó la operación.
    - **Usuario_proceso:** Usuario que realizo el proceso

  Esta tabla permitirá auditar y validar la correcta ejecución de los procesos de carga, especialmente en el caso de operaciones repetidas o incrementalmente programadas, antes de iniciar el proceso esta table debe estar vacía.

- Desarrollar un **programa de limpieza de la base de datos**, cuya función sea **eliminar todos los registros** de las tablas del modelo, conservando únicamente la **estructura del esquema** (tablas vacías y sus relaciones referenciales). Es decir, el programa debe dejar la base de datos **lista para una nueva carga**, manteniendo intacta la definición del modelo y su integridad referencial.
- Es importante aclarar que el poblado de los datos podrá realizarse de manera individual (por cada tabla) o de forma general (para todas las tablas en un solo proceso). En ambos casos, se requiere que el usuario final proporcione una fecha de carga que actuará como parámetro de entrada para el proceso de carga de datos. Esta fecha será registrada en el sistema como la fecha del proceso.

Función de Validación:

- Se requiere implementar una función de validación que verifique que la fecha de carga de los datos cumpla con las siguientes condiciones:
    - La fecha debe corresponder a un horario laboral en un día hábil (lunes a viernes).
    - No se permitirá insertar fechas en el pasado ni fechas con más de tres días de antelación respecto a la fecha actual del sistema.

  En caso de que cualquiera de estas condiciones no se cumpla, se deberá generar una excepción que notifique el error relacionado con la fecha.

Tabla de Reporte:

Una vez se hayan poblado las tablas correspondientes al modelo relacional definido para el dataset seleccionado, cada grupo deberá crear una tabla de informes o resumen estadístico, diseñada para ofrecer una visión consolidada y organizada de los datos más relevantes.

Esta tabla deberá ser creada y alimentada mediante un procedimiento almacenado, el cual recibirá como parámetro de entrada un rango de fechas (fecha de inicio y fecha de fin). Con base en ese rango, el sistema deberá procesar la información existente, realizar los cálculos agregados necesarios (como sumas, promedios, conteos, etc.) y registrar los resultados en la tabla de informes.

Los datos deberán estar agrupados por una o varias dimensiones clave que permitan generar análisis útiles para el negocio, como por ejemplo por año, por región, por categoría, por tipo de evento, entre otros (dependiendo del contenido del dataset elegido por el grupo).

Este proceso debe ejecutarse una vez se hayan completado las cargas de las tablas principales, asegurando así que el flujo completo de información esté disponible para el cálculo de los indicadores.

Ejemplo:

Datos extraídos de https://data.who.int/dashboards/covid19/data?n=c

Datos iniciales

	A	B	C	D	E	F	G	H
1	Date_reported	Country_code	Country	WHO_region	New_cases	Cumulative_cases	New_deaths	Cumulative_deaths
99333	20/02/2021	JO	Jordan	EMR	867	357611	16	4528
99334	20/02/2021	IM	Isle of Man	EUR	2	445		35
99335	20/02/2021	IE	Ireland	EUR	896	213725	34	4397
99336	20/02/2021	IN	India	SEAR	13993	10977387	101	156212
99337	20/02/2021	IR	Iran (Islamic Republic of)	EMR	8017	1558159	77	59341
99338	20/02/2021	GP	Guadeloupe	AMR		9455		160
99339	20/02/2021	GA	Gabon	AFR	446	13553		75
99340	20/02/2021	FK	Falkland Islands (Malvinas)	AMR		49		0
99341	20/02/2021	DO	Dominican Republic	AMR	818	233598	17	3024
99342	20/02/2021	DM	Dominica	AMR		134		0
99343	20/02/2021	CW	Cura?ao	AMR	8	4666		22
99344	20/02/2021	CO	Colombia	AMR	4824	2212525	200	58334
99345	20/02/2021	KM	Comoros	AFR	15	3403	4	142
99346	20/02/2021	CK	Cook Islands	WPR	0	0	0	0
99347	20/02/2021	CA	Canada	AMR	3314	837495	65	21330
99348	20/02/2021	BB	Barbados	AMR	39	2647		29
99349	20/02/2021	BR	Brazil	AMR	51879	10030626	1367	243457
99350	20/02/2021	BD	Bangladesh	SEAR	350	543024	5	8342
99351	20/02/2021	AZ	Azerbaijan	EUR	145	232636	3	3190
99352	20/02/2021	AI	Anguilla	AMR		18		0
99353	20/02/2021	AE	United Arab Emirates	EMR	3140	365017	20	1093
99354	20/02/2021	TZ	United Republic of Tanzania	AFR		509		21
99355	20/02/2021	VE	Venezuela (Bolivarian Republic of)	AMR	462	134781	6	1303
99356	20/02/2021	TJ	Tajikistan	EUR		13714		91
99357	20/02/2021	SY	Syrian Arab Republic	EMR	49	15094	4	994
99358	20/02/2021	TH	Thailand	SEAR	82	25323	0	83
99359	20/02/2021	LK	Sri Lanka	SEAR	543	79480	2	435
99360	20/02/2021	SB	Solomon Islands	WPR	0	18	0	0
99361	20/02/2021	SN	Senegal	AFR	279	32378	6	787
99362	21/02/2021	AI	Anguilla	AMR		18		0
99363	21/02/2021	AZ	Azerbaijan	EUR	193	232829	3	3193
99364	21/02/2021	BD	Bangladesh	SEAR	327	543351	7	8349

Datos finales

PAIS	New_cases_2020	New_cases_2021	New_cases_2022	New_cases_2023	New_cases_2024	New_deaths_2020	New_deaths_2021	New_deaths_2022	New_deaths_2023	New_deaths_2024	TOTAL_CASOS	TOTAL_deaths
Afghanistan	2158	5194	493	128	9	51848	106054	49420	23053	1615	7982	231990
Albania	1134	2053	409	8	1	55380	151841	125449	1926	267	3605	334863
Algeria	2722	3507	652			97857	118519	54818	816		6881	272010
American Samoa			34				11	8255	93		34	8359
Andorra	83	56	19	1		7806	13924	25956	329		159	48015
Angola	400	1346	183	8		17149	53993	33928	1858	397	1937	107325
Anguilla		5	7			12	1634	2258			12	3904
Antigua and Barbuda	5	113	28			155	4074	4877			146	9106
Argentina	47507	70323	12294	348	76	1629908	3930008	4331223	153818	33828	130548	10078785
Armenia	2768	5175	765	67		157834	186647	101102	5843		8775	451426
Aruba	47	134	99	12		5228	11937	26587	472		292	44224
Australia	920	1457	15439	6528	368	28296	338311	10327434	1027494	63569	24712	11785104
Austria	7049	9538	4984	963		344732	907356	4426257	402942		22534	6081287
Azerbaijan	2416	5868	1722	347		211764	402355	211774	8361	531	10353	834785
Bahamas	169	544	120	11		7788	15842	13861	593		844	38084
Bahrain	351	1043	142			91518	188000	417096			1536	696614
Bangladesh	7452	20608	1379	38	9	509148	1074105	453771	9268	1595	29486	2047887
Barbados	7	253	308	25		347	27282	77886	2279	231	593	108025
Belarus	1376	4108	1634			184922	507679	301436			7118	994037
Belgium	19374	8818	5048	1099		638760	1409350	2631362	169273	10229	34339	4858974
Belize	228	364	96			10490	21013	38172	1107	17	688	70799
Benin	44	117	2			3205	21730	3050	51		163	28036
Bermuda	9	97	46	13		561	5503	12493	303		165	18860
Bhutan		3	18			597	2063	59871	166		21	62697
Bolivia (Plurinationa	9083	10447	2750	102		153590	421657	571494	51242	27	22382	1198010
Bonaire, Saint Eusta	3	19	17	2		182	3057	8477	206		41	11922
Bosnia and Herzegov	3923	9360	2939	158	5	109330	178386	113244	2548	77	16385	403585
Botswana	39	2400	347	14		13890	198592	115549	2386		2800	330417
Brazil	190488	427904	74351	9373		7448560	14782177	13893600	1395623		702116	37519960
British Virgin Islands	1	38	25			93	2978	4234	87		64	7392

Es importante destacar que el rango de fechas debe ser proporcionado como un parámetro de entrada al procedimiento almacenado. Por ejemplo, si se especifica el intervalo entre el 01/01/2021 y el 15/12/2022, la tabla de informes deberá incluir únicamente los datos correspondientes a ese período, realizando los cálculos agregados (*como sumas totales, promedios, etc.*) solo dentro del rango (*exacto*) proporcionado.

Además, se debe implementar un control previo para validar si existen datos disponibles dentro del rango indicado. En caso de que no se encuentren registros para procesar, el sistema deberá generar una alerta o lanzar una excepción controlada desde la base de datos, informando que no existen datos dentro del período solicitado.

Triggers:

Se deberán implementar triggers de auditoría sobre las tablas finales. Estos disparadores deberán activarse ante cualquier operación de modificación (UPDATE) o eliminación (DELETE) de datos.

Cada vez que ocurra una de estas operaciones, el sistema deberá registrar automáticamente la siguiente información en una tabla de auditoría:

- Nombre del usuario que realizó la operación.
- Nombre de maquina o SO desde la cual se ejecutó la acción
- Fecha y hora del proceso.
- Tipo de operación (UPDATE o DELETE).
- Valor anterior del dato afectado.
- Valor nuevo, en caso de una actualización.

Este mecanismo tiene como propósito garantizar el seguimiento y control de los cambios realizados sobre la información, y asegurar la trazabilidad de los datos en el sistema.

Package:

Toda la lógica de negocio desarrollada deberá ser empaquetada dentro de uno o varios packages PL/SQL, según la organización y el criterio del grupo. Cada package debe contener tanto las declaraciones (especificación) como las implementaciones (body) de los procedimientos, funciones, y cualquier otra unidad lógica que forme parte del proyecto.

El objetivo es promover una estructura modular, organizada y reutilizable, que facilite el mantenimiento del sistema, la trazabilidad de las operaciones y una mejor separación de responsabilidades dentro del código.

La distribución y agrupación de la lógica dentro de los packages podrá variar dependiendo de las necesidades y el diseño de cada solución, siempre que se garantice la claridad, cohesión y funcionalidad del sistema.

ENTREGABLES

Proporcionar documentación detallada sobre la ejecución del programa.

Incluir pasos para la carga inicial, ejecución de procedimientos y validaciones.

- Código Desarrollado (Fuentes):
  Proporcione todos los scripts y códigos desarrollados durante el proyecto. Esto incluye scripts SQL, PL/SQL y cualquier otro código relevante.
  Organice el código de manera lógica y coherente, facilitando su revisión y comprensión.

README (Paso a Paso):

- El archivo README debe contener instrucciones detalladas sobre cómo implementar y ejecutar el proyecto.
- Proporcione un resumen de cada paso, desde la creación de tablas hasta la ejecución de procedimientos almacenados y otros componentes críticos.
- Incluya cualquier configuración especial requerida, como permisos, tablespaces, etc.

SUSTENTACIÓN

Nota: No se limite únicamente a los requerimientos establecidos en este proyecto. Está permitido e incluso recomendado incluir funcionalidades adicionales que aporten valor, mejoren la experiencia de uso o fortalezcan el diseño técnico de la solución.