Evaluación Parcial 2

Trabajando en Big Data: herramientas y procesos esenciales (Datos Batch)

Sigla	Nombre Asignatura	Tiempo Asignado	% Ponderación
BIY7131	BIG DATA	5 h	40%

01. Agente evaluativo

X	Heteroevaluación		Coevaluación		Autoevaluación

02. Tabla de Especificaciones

Resultado de Aprendizaje	Indicador de Logro (IL)	Indicador de Evaluación (IE)*	Ponderación Indicador Logro	Ponderación Indicador de Evaluación
RA2_Utiliza herramientas de Big Data, con objetivo de ejecutar tareas de gestión de grandes volúmenes de datos en formato Batch para facilitar su análisis y visualización.	IL 2.1_Construye procesos de carga de datos (ingesta) en diferentes formatos y de diversas fuentes, de acuerdo a necesidades de la organización.	IE 1_Construye procesos de carga al DataLake, de acuerdo a disponibilidad de información desde la fuente origen, para preparación en caso de errores, de acuerdo a necesidades de la organización.	**25%**	**25%**
	IL 2.2_Construye procesos, con el fin de realizar limpieza, transformación y almacenamiento de grandes volúmenes de datos formato Batch. Presentaciones	IE 2_Construye procesos de transformación y limpieza de datos Batch, dejándolos en formato adecuado para capa de consumo, preparándolos para evitar duplicidad de datos y detección de errores.	**25%**	**45%**

		IE 3_Construye procesos orquestados, considerando la disponibilidad de información y dependencias de grandes volúmenes de datos en formato Batch.	**20%**	
		IE 4_Construye recursos (Informe y Presentación) que contiene todas las etapas solicitadas, con procedimientos relevantes y secuencia lógica de acuerdo a los requerimientos de la organización.	**10%**	**30%**
		IE 5_Presenta los resultados siguiendo una estructura lógica, considerando la información del informe.	**10%**	
		IE 6_Establece comunicación efectiva, utilizando lenguaje técnico requerido en la disciplina y contexto laboral.	**10%**	

03. Instrucciones para el/la estudiante

Esta es una evaluación que corresponde a un Encargo con presentación y tiene un **40%** de ponderación sobre la nota final de la asignatura.

El **tiempo** para presentar esta evaluación es de **15 minutos** y se realiza en grupos en **taller de PC.**

La evaluación consiste en **realizar y presentar un informe de gestión de grandes volúmenes de Datos,** mediante la carga histórica de todos los archivos disponibles, junto con información diaria reciente, la cual permita a los usuarios responder diversas preguntas de negocio relacionadas con el caso a resolver.

Consideraciones específicas:

Se asignará un caso cuyos antecedentes generales son los siguientes:

- **Datos**
- **Contexto**.
- **Información.**

La información que utilizaremos para desarrollar este examen proviene de varias fuentes. A continuación, se especifican dichas fuentes:

- **Datos Históricos:** Identificación de los datos para los casos asignados.
- **Datos Diarios:** Identificación de los datos para los casos asignados.

Los procedimientos específicos son los siguientes:

- Paso 1: Realizar las conexiones con la fuente de origen de datos (estas pueden ser bases de datos, archivos que deben descargar desde internet, etc.)
- Paso 2: Descargar y/o generar los archivos al dataLake.
- Paso 3: Construir los procesos de limpieza, transformación y carga al modelo de datos final.
- Paso 4: Construir los reportes y/o visualizaciones correspondientes.

Para lo anterior, deberá realizar lo siguiente:

- Construir procesos de carga en data Lake, considerando disponibilidad de la información desde de la fuente
- Construir procesos de transformación, limpieza de datos Batch.
- Construir procesos orquestados considerando disponibilidad de información y dependencias de grandes volúmenes de datos en formato Batch.

Para cada uno de estos pasos, debe considerar (si aplica) lo siguiente:

- **Control de errores:** todos los procesos pueden tener puntos de fallo, de acuerdo a lo identificado en la Etapa 1 (diseño), debe implementar los controles de errores correspondientes.
- **Control de duplicidad de archivos:** Los DataLake contienen múltiples archivos, debe considerar que los procesos se pueden ejecutar múltiples veces, por tanto, sus procesos deben determinar qué hacer si un fichero y/o datos ya existen (tome la decisión de acuerdo a lo visto durante el semestre).
- **Registro de actividad:** Como se señaló anteriormente, los procesos se podrían ejecutar varias veces, debe incorporar el control de ejecución (ej.: ¿si el proceso ya se ejecutó lo debo volver a ejecutar, lo debo bloquear o debo pedir autorización para volver a ejecutar?).
- **Validación de Datos y Procesos:** Según corresponda, debe considerar en su construcción la validación de los procesos y la validación de los datos a trabajar, incluyendo procesos de transformación, manteniendo la trazabilidad de los datos desde el origen. Tenga en cuenta que al ser datos Batch, los procesos deben permitir reprocesar datos históricos en alguna fecha en particular.

Entregable:

- Informe explicando las herramientas utilizadas y la estrategia de carga de datos seleccionada.
- Códigos desarrollados
- Instrucciones para instalar en ambiente
- Instrucciones para ejecución en ambiente
- Evidencia de las ejecuciones

ASPECTOS FORMALES DE LA PRESENTACIÓN:

Presentación y defensa:

La presentación y defensa se realizarán en un entorno simulado a una reunión de directorio de una empresa que requiera una solución en ámbitos Big Data.

La presentación será evaluada de forma individual, considerando los siguientes indicadores:

- Construye informe que contiene todas las etapas solicitadas, con procedimientos relevante y secuencia lógica de acuerdo a los requerimientos de la organización.
- Presenta los resultados siguiendo una estructura lógica, considerando la información del informe.
- Establece comunicación efectiva, utilizando lenguaje técnico requerido en la disciplina y contexto laboral.

Se recomienda la utilización de plantillas interactivas, en las cuales se priorice la organización de información en diagramas, flujo de procesos, tablas consolidadas.
La presentación no debe exceder las 10 láminas y debe ser enviada con 48 h de anticipación, de acuerdo al día de presentación.
Los estudiantes tienen un tiempo de 15 minutos para presentar sus resultados, este tiempo también incluye las consultas de la o él docente que asumen un rol de Jefatura/Gerencia de una empresa.

APP WEB Plantillas interactivas:

https://app.genial.ly/
https://www.canva.com/es_419/
https://prezi.com/es/

Pauta de Evaluación

Pauta tipo: Rúbrica

Categoría	% logro	Descripción niveles de logro
Muy buen desempeño	**100%**	Demuestra un desempeño destacado, evidenciando el logro de todos los aspectos evaluados en el indicador.
Desempeño aceptable	**60%**	Demuestra un desempeño competente, evidenciando el logro de los elementos básicos del indicador, pero con omisiones, dificultades o errores.
Desempeño incipiente	**30%**	Presenta importantes omisiones, dificultades o errores en el desempeño, que no permiten evidenciar los elementos básicos del logro del indicador, por lo que no puede ser considerado competente.
Desempeño no logrado	**0%**	Presenta ausencia o incorrecto desempeño.

Indicador de Evaluación	Categorías de Respuesta				Ponderación del Indicador de Evaluación
	Muy buen desempeño 100%	Desempeño aceptable 60%	Desempeño incipiente 30%	Desempeño no logrado 0%	
1.-Construye procesos de carga al DataLake, de acuerdo a disponibilidad de información desde la fuente origen, para preparación en caso de errores, de acuerdo a necesidades de la organización	Implementa el 100% de los procesos de acuerdo a la definición, al ejecutar funcionan y controlan los errores en caso de que los archivos no estén disponibles, incluye condiciones de re-ejecución y logs.	Implementa entre un 99% a un 51% de los procesos, considerando en ellos control de errores, disponibilidad de información en origen y logs.	Implementa < 50% y más de un 10% de los procesos, o gran parte de los procesos construidos NO consideran control de errores, logs ni disponibilidad de información en el origen.	Implementa < del 10% de Procesos y no son construidos o presentan errores en la ejecución no controlados, de acuerdo a necesidades de la organización.	**25%**
2.-Construye procesos de transformación y limpieza de datos Batch, dejándolos en formato adecuado para capa de consumo, preparándolos para evitar la duplicidad de datos y detección de errores.	Realiza el 100% de la transformación y limpieza de datos de acuerdo al modelo, dejando los datos en formato estructurado	Realiza entre el 99% y 51% de procesos de transformación y limpieza, dejando dichos datos en	Realiza < 50% y más de un 10% de procesos de transformación y limpieza, o los procesos que realiza	Realiza menos del 10% de procesos de limpieza y transformación de datos.	**25%**

	de acuerdo a la tecnología utilizada (parquet, columnares, etc).	formatos estructurados adecuados.	no deja los datos en un formato adecuado (por ejemplo, los almacena en TXT en vez de dejarlos en parquet o columnares).		
3.-Construye procesos orquestados, considerando la disponibilidad de información y dependencias de grandes volúmenes de datos en formato Batch.	Construye el 100% de las mallas de procesos considerando dependencias técnicas y funcionales, horarios y/o eventos de disponibilidad de información y control de errores.	Construye entre un 99% y 51% de las mallas de procesos, y considera en ellos las dependencias técnicas y funcionales, horarios y/o eventos de disponibilidad de información y control de errores.	Construye < 50% y más de un 10% de las mallas de procesos, o NO considera en la construcción las dependencias técnicas y funcionales, horarios y/o eventos de disponibilidad de información y control de errores.	Construye menos de un 10% de la orquestación de datos, los procesos se ejecutan manualmente	**20%**
4.-Construye recursos (Informe y presentación) que contiene todas las etapas solicitadas, con procedimientos relevantes y secuencia lógica de acuerdo a los requerimientos de la organización	Contiene todos los apartados solicitados, en el orden adecuado y con información relevante y oportuna que soporta adecuadamente la exposición con una secuencia lógica y existe interrelación entre todas y cada una de las partes	Contiene apartados solicitados, aunque el orden podría ser mejorado, falta alguna información relevante y los apartados tienen en general una secuencia lógica y existe interrelación entre la mayoría de las partes de los procesos.	Contiene apartados solicitados, aunque con falta alguna información relevante y los apartados tienen en general una secuencia lógica, pero con imprecisiones en la interrelación entre la mayoría de las partes de los procesos.	Hay elementos clave de los procedimientos que fueron omitidos en los recursos, siendo la estructura no adecuada y la información presentada tiene notables carencias	**10%**

5.-Presenta los resultados siguiendo una estructura lógica, considerando la información presentada.	La información es presentada de manera lógica (inductiva o deductiva) y coherente, de manera que la audiencia pueda seguir fácilmente.	La mayor parte de la información es presentada de manera lógica y generalmente bien organizada, pero hace falta mejores transiciones de una idea a otra.	Presenta con una organización adecuada, pero débil en aspectos claves, demostrando falta de conexiones claras entre las partes de la presentación, las partes parecen aisladas entre sí.	Presenta resultado de forma no estructurada y difícil de entender. Desorganizada, no evidenciándose hay secuencia lógica en la información.	**10%**
6.-Establece comunicación efectiva, utilizando lenguaje técnico requerido en la disciplina y contexto laboral.	Ejecuta apropiado y lenguaje técnico y con buena pronunciación comunicando efectivamente el proceso ejecutado.	En general, articula claramente y la pronunciación y lenguaje técnico, con una comunicación que en su mayoría correcta.	Habla en voz un poco baja y comete algunos errores de pronunciación y lenguaje técnico, pero es comprensible en general.	Presentación no es clara, con tono de voz demasiado bajo y no se le puede entender, comete errores de pronunciación y palabras técnicas que dificultan la comprensión	**10%**
				Total	**100%**