Examen Transversal Estudiante

Forma A

ÍTEM	PUNTAJE	% PONDERACIÓN
Competencias de Especialidad	91	100%

Porcentaje corte nota 4,0	60%

INSTRUCCIONES GENERALES:

La Entrega de Encargo consiste en el desarrollo y posterior entrega de un producto que ha sido solicitado con anticipación ("encargo") el cual puede ser presentado como parte de los requisitos del examen. Apropiado para evidencias de producto.

Tiempo: 0:15:0 y 3 semanas

La Entrega de Encargo con Presentación tiene un 100% del total del examen.

El tiempo para desarrollar la Entrega de Encargo con Presentación es de 0:15:0 y 3 semanas.

El ET BIY7131 Big Data, corresponde a un encargo con presentación y tiene un 40% de ponderación sobre la nota final de la asignatura.

El tiempo para desarrollar este encargo es de 3 semanas, comenzando la semana 15

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 la disponibilidad de servicios de transporte en ciertas zonas, horarios y frecuencia. Los estudiantes, adicionalmente, deberá diseñar y crear dos reportes que muestren información agregada.

El encargo debe cumplir con todas las instrucciones establecidas previamente, considerando aspectos asociados a contenido y forma de entrega. Recuerde respetar tanto el tiempo de entrega como la estructura propuesta, ya que de no cumplirlos incidirán negativamente en su evaluación.

Consideraciones específicas:
Se asignará un caso, el cual debe comenzar a ser desarrollado desde la semana 15, los antecedentes generales del caso son los siguientes;

Datos de Transporte Público;
Contexto: El transporte público es fundamental para la movilización de las personas en una ciudad, es por ello por lo que conocer los trayectos, paradas, horarios y duración de los trayectos es indispensable para planificar correctamente los viajes.
Este examen busca generar una plataforma de datos que contenga la información histórica de los viajes, de tal forma de identificar la cantidad de transportes disponibles por cada uno de los medios disponibles (buses, metro), ver qué horarios tiene mayor disponibilidad de transporte en una zona determinada, y cuáles de esas zonas han tenido la mayor variabilidad en los recorridos (señalando si han agregado o quitado recorridos). Por otro lado, es importante tener siempre la información lo más actualizada posible, es por ello que también será requerido en este examen la obtención diaria de los recorridos disponibles en Santiago.

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

Datos Históricos: Estos datos se pueden obtener desde la plataforma de datos abiertos del Gobierno de Chile, el link es el siguiente: https://datos.gob.cl/dataset/33245. Este set de datos contiene la información mensual de la planificación de los distintos medios de transporte en Santiago, y está disponible cerca del día 15 de cada mes. El examen requiere que la descarga de datos sea automatizada, para ello deberá utilizar la siguiente API **https://datos.gob.cl/api/action/package_show?id=33245**. Esta API devuelve los recursos disponibles, debe identificar la ULR en la respuesta que le permitirá descargar los archivos.

Datos Diarios: Para obtener la información de transportes diarios, primero deberá consultar una API que le devolverá todos los recorridos disponibles (**https://www.red.cl/restservice_v2/rest/getservicios/all**), luego, por cada uno de esos recorridos deberá obtener la información de su trayecto, horarios y paradas desde la siguiente API: **https://www.red.cl/restservice_v2/rest/conocerecorrido?codsint=101** (donde "101" es un código de servicio devuelto por la API anterior)

Los procedimientos específicos incluyen las siguientes etapas con sus respectivos requisitos:

ETAPA 1: Se deberá identificar la arquitectura que mejor se adapte a la problemática planteada, justificando su decisión desde una perspectiva técnica y funcional.
En esta etapa se debe realizar el diseño el o los modelos de datos finales optimizados para el consumo de usuarios finales y/o de herramientas de visualización, los procesos Batch y Streaming/api/near-real-time/real-time (según corresponda), junto con las mallas de ejecución y puntos de control de errores.

Esta etapa deberá:

- Seleccionar herramientas de procesamiento, transformación y visualización, justificando su aplicación.
- Incorporar ciclo de vida del dato en el proceso end to end.
- Analizar problemática planteada utilizando arquitecturas de referencia.
- Definir los procesos, flujos de información y orquestación de datos para iniciar la construcción de la solución.

ETAPA 2: Se deberá construir los procesos Batch, se sugieren los siguientes pasos:
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.

En esta etapa deberá:

- 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.

ETAPA 3: Deberán construir los procesos de BigData utilizando una estrategia de Real-Time, Streaming o API, según corresponda. Se sugieren los siguientes pasos:
Paso 1: Realizar las conexiones con la fuente de origen de datos.
Paso 2: Descargar y/o generar los archivos al dataLake o fuente de destino.
Paso 3: Construir los procesos de limpieza, transformación y carga al modelo de datos final, considerando la trazabilidad de información y ciclo de vida del dato.
Paso 4: Mejorar los reportes y/o visualizaciones correspondientes construidos previamente en la etapa 2.

En esta etapa deberá:

- Construir procesos de carga, considerando disponibilidad de la información desde fuente de origen, en caso de errores.
- Construir procesos de transformación y limpieza de datos, dejando los datos en formatos para capa de consumo, evitando duplicidad de datos real time/Streaming para detección oportuna de errores.

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 con lo identificado en la Etapa 1 (diseño), debe implementar los controles de errores correspondientes.
- Control de duplicidad de datos: Considerar que los procesos se pueden ejecutar múltiples veces, y que los datos desde el origen pueden cambiar, por tanto, sus procesos deben determinar qué hacer si una ejecución devuelve datos que ya existen (tome la decisión de acuerdo con lo visto durante el semestre). También debe considerar que parte de estos datos pueden haber sido cargados desde la etapa 2 de datos Batch.
- Registro de actividad: Los procesos se podrían ejecutar varias veces, debe incorporar el control de ejecución y considerar el ciclo de vida de los datos.
- 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.

Examen Transversal
Estudiante

Forma A

RESUMEN:

El ET considera los siguientes entregables:

- Informe Etapas 1, 2 y 3, que debe considerar:
  - Elección de la arquitectura.
  - Elección de las herramientas.
  - Definición de los procesos de carga Batch y Streaming/API, considerando:
    - Validaciones y controles de errores.
    - Conexión a las Fuentes de Origen, indicando nombres de servidores, archivos a leer, API a consultar, etc.
  - Definición de la orquestación de ejecución de procesos, dependencias, periodicidad de ejecución, etc.
  - Modelos de Datos y Diccionarios de datos, los cuales debe contener:
    - Tipos de datos de origen y destino.
    - Datos que considera anómalos y sugerencia de correcciones.
    - Datos que pueden fallar y sugerencias de control de errores.
    - Posibles transformaciones.
    - Posibles separaciones de campos.
  - Procesos Construidos Etapa 2: Desarrollo con etapas de gestión de volúmenes de datos en formato Batch para su análisis y visualización. Se deben incluir, además, las instrucciones para poder implementar y ejecutar.
  - Procesos Construidos Etapa 3: Desarrollo con etapas de gestión de volúmenes de datos en formato real-time/Streaming para su análisis y visualización. Se deben incluir, además, las instrucciones y código fuente para poder implementar y ejecutar.
  - Agregar una sección con el diseño de los procesos realmente implementados, identificando las diferencias vs el diseño inicial, y conclusiones/reflexiones del grupo en torno al proceso completo.
- PPT interactiva, que permita revisar la implementación de la solución en tiempo real.

PRESENTACION 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:
  a) Domina los procesos solicitados considerando estándares y procesos requeridos en la gestión de datos batch y streaming de acuerdo con las necesidades de la organización.
  b) Presenta los resultados siguiendo una estructura lógica, considerando la información del informe.
  c) 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 PPT no debe exceder las 15 láminas y debe ser enviada con 24h de anticipación, de acuerdo con el 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/