Proyecto de Diseño

(A realizar de manera individual o en grupos de 2-3 estudiantes.)

Un laboratorio de biotecnología necesita una plataforma eficiente y escalable que le permita almacenar y manipular la gran cantidad de datos generada por sus diferentes proyectos y medios de recolección de datos como sensores, imágenes de satélites, de secuenciación, etc. La amplia gama de arquitecturas y modelos de programación presenta oportunidades y desafíos para los científicos e ingenieros de análisis de datos biológicos. El laboratorio lo ha contratado a Ud. para que les proponga una arquitectura en la nube (con AWS o Azure o GCP) que cumpla con los siguientes requerimientos.

- Requerimientos funcionales:
    1. El sistema debe poder generar reportes visuales con los resultados de los análisis de los datos.
    2. El diseño debe incluir una o más soluciones de almacenamiento de datos, según Ud. considere conveniente.
    3. Se debe permitir realizar consultas de datos ad hoc o interactivas, así como soportar la generación de reportes diarios, semanales y mensuales.
    4. Debe haber una interfaz web para la administración de colecciones de especímenes.
    5. Debe haber una manera de realizar búsquedas en un registro de resultados de los análisis.

- Requerimientos no funcionales:
    1. Se debe poder procesar datos que llegan a gran velocidad y que por lo tanto, representan un gran volumen de datos (el lab produce cientos de TB de datos de secuencia por día).
    2. Se debe soportar aplicaciones con tiempos de ejecución extremadamente largos (una aplicación de análisis de datos biológicos puede correr durante días o incluso meses).

- Entregable: Un PDF que contenga los siguientes ítems, sobre los cuales Ud. será evaluado según el puntaje indicado.
    1. [20 pts] Diagrama del diseño (arquitectura) propuesto.
    2. [10 pts] Liste cada uno de los componentes mostrados en su arquitectura y explique por qué lo incluye, así como indique qué producto específico usaría y por qué (ej., qué base de datos). Justifique su respuesta.
    3. [10 pts] De los elementos que Ud. ha incluido en su diseño, diga cuáles pertenecen a cada uno de los tres componentes de un Pipeline de Big Data (según las diapositivas del vídeo 2 del tema de Plataformas de Procesamiento Distribuido). NOTA: Es posible que alguno(s) de los componentes de su diseño no pertenezcan a ninguna de las tres categorías.
    4. [20 pts] Haga un presupuesto para un año, usando los precios reportados por el proveedor en la nube de su elección.
    5. [20 pts] De las siguientes preguntas, seleccione dos (2) y contéstelas en su reporte. Seleccione las que su grupo se sienta más preparado para contestar. Cuáles seleccione no afectará su nota.
        - [10 pts] ¿Cómo su diseño asegura la tolerancia a fallos para conservar la **integridad** y **durabilidad** de los datos almacenados?
        - [10 pts] El rendimiento general del sistema es crítico para el laboratorio, especialmente para lograr baja latencia de las interacciones de los usuarios. ¿Cómo su diseño asegura una baja latencia al momento de consultar datos?
        - [10 pts] Es importante para el laboratorio que se conserve la integridad de los datos, especialmente cuando nuevas colecciones de especímenes o taxonomías son incluidas. ¿Cómo su diseño asegura que las actualizaciones de usuarios y aplicaciones no afecte la consistencia de los datos?
        - [10 pts] ¿Su diseño permite que nuevos reportes sean fáciles de incluir en la plataforma? Explique.
        - [10 pts] Para el laboratorio es importante que sus sistemas mantengan un buen nivel de **confiabilidad** y sean tolerantes a fallos. Describa las posibles fuentes de falla en su diseño propuesto, y cómo se recuperaría de estas fallas.