Laboratorio Cluster para el análisis de datos con Spark

Objetivo del Laboratorio:
El objetivo de este laboratorio avanzado es enseñar a los estudiantes cómo configurar un clúster de Apache Spark utilizando Docker Compose, conectar un Jupyter Notebook, y realizar un análisis de datos utilizando un conjunto de datos de su elección.

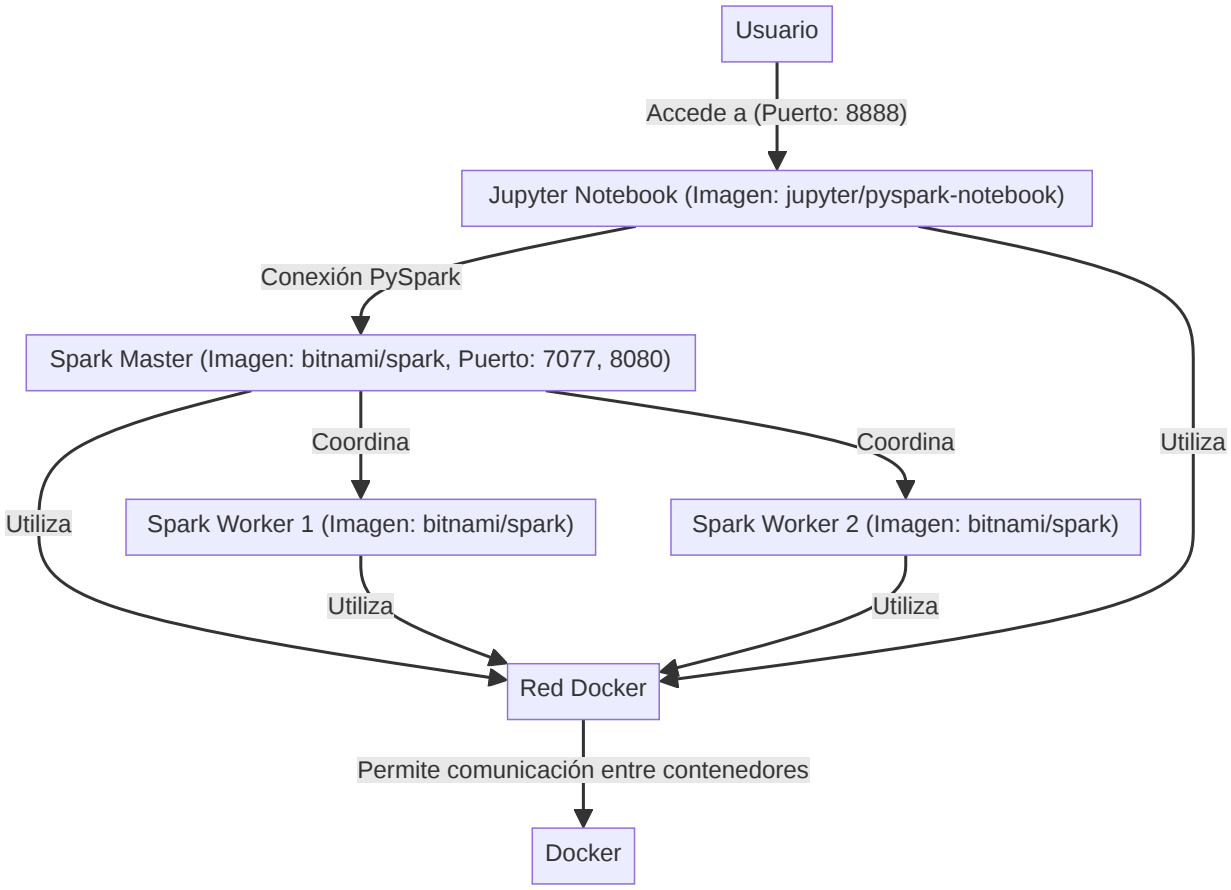
Descripción de la Tarea:
Los estudiantes utilizarán Docker Compose para definir y ejecutar un clúster de Spark con dos workers y un servicio de Jupyter Notebook. Posteriormente, seleccionarán un conjunto de datos, formularán una pregunta de investigación y utilizarán PySpark en el Jupyter Notebook para responder a dicha pregunta.

Pasos del Laboratorio:

1. **Preparación del `docker-compose.yml`:**
    - Crear un archivo `docker-compose.yml` que defina los servicios de Spark Master, Spark Workers y Jupyter Notebook.
2. **Despliegue del Clúster de Spark y Jupyter Notebook:**
    - Iniciar todos los servicios utilizando Docker Compose.
    - Verificar el estado del clúster a través de la interfaz web de Spark Master.
3. **Selección y Análisis de Datos:**
    - Buscar y seleccionar un conjunto de datos apropiado para el análisis.
    - Formular una pregunta de investigación relevante.
    - Conectar el Jupyter Notebook al clúster de Spark y utilizar PySpark para analizar los datos y responder a la pregunta formulada.
4. **Limpieza Post-Laboratorio:**
    - Detener y eliminar todos los servicios utilizando Docker Compose.

Arquitectura del Laboratorio:

El diagrama de la arquitectura del laboratorio con Docker Compose es el siguiente:



Instrucciones Detalladas:
Se proporcionarán en el material del curso, incluyendo el ejemplo de `docker-compose.yml` y guías para la selección y análisis de datos.

Entregables:
Los estudiantes deberán entregar:

- Un archivo `docker-compose.yml` para desplegar el clúster y el servicio de Jupyter.
- Un informe que incluya:
  - La descripción del conjunto de datos seleccionado y la pregunta de investigación.
  - La explicación de cada servicio definido en el `docker-compose.yml`.
  - Capturas de pantalla del clúster y del Jupyter Notebook en funcionamiento.

    - Un notebook de Jupyter con el código de PySpark ejecutado, los resultados del análisis y la respuesta a la pregunta de investigación.

Evaluación:
La tarea será evaluada en base a la correcta configuración del clúster, la originalidad y relevancia de la pregunta de investigación, la adecuada selección del conjunto de datos, y la calidad del análisis y del informe entregado.

Fecha de Entrega:
La tarea deberá ser entregada a más tardar el 13 de noviembre, antes de las 23:00.

Nota:

- Hay libertad en usar la imagenes que estime conveniente para los containers.