SUPUESTO TEÓRICO-PRÁCTICO

D2: PROGRAMACIÓN Y COMPUTACIÓN CIENTÍFICA

CASO 1: GENERAR CONOCIMIENTO DESDE LA EXPERIMENTACIÓN

Introducción:

Se nos ha encargado la implementación de un experimento para evaluar la estructura interna de un modelo biomimético diseñado para calibrar ecógrafos. Básicamente consiste en una pieza dopada con micropartículas uniformemente distribuidas que hacen de reflectores, a la que se le han hecho hasta 6 inserciones de otros bloques con diferente densidad y tamaño de reflectores respecto al bloque principal.

Se espera que como resultado del experimento se obtengan un conjunto de imágenes 3D que representen la estructura interna para estos dos parámetros (densidad, reflectividad) y que identifique los diferentes bloques. En total se tienen que analizar 16 piezas distintas y, para evaluar su degradación a largo plazo, cada pieza se analizará al menos 9 veces en un periodo de un año.

La instrumentación se compone por: un transductor ultrasónico, un osciloscopio programable para la recepción, un generador de señal para la emisión y un sistema de movimiento cartesiano (plano XY) controlado. Sobre la cara superior de la pieza se diseña una rejilla de muestreo (plano XY) en los cuales tenemos que tomar medidas ultrasónicas en pulso-eco (eje Z). De forma que tras la inspección el modelo físico se representa como un volumen de 1024x1024x2048 puntos (XYZ)

El procesado que se aplica a la envolvente de los datos es el siguiente:

1) Preprocesado: filtrado paso bajo del volumen para eliminar ruido. Se pasan los datos a valor absoluto.
2) Identificar los reflectores. Se umbraliza el volumen y los valores que superen este umbral se consideran reflectores y se identifican con su posición (x,y,z), su valor de amplitud máxima y su extensión (suponemos una forma esférica)
3) Evaluar la densidad de reflectores en el volumen con una resolución de .1mm^3 (32x32x32).
4) Identificación de las regiones por agrupación entre densidad y reflectividad

Preguntas para Evaluación:

BLOQUE 1: Control e integración de instrumentación científica. Algoritmos y estructuras de datos. Workflows científicos en entornos en la nube.

- Considerar las opciones (herramientas de programación, así como señales y protocolos de comunicación) para automatizar el proceso, sincronizar la instrumentación y recoger y almacenar las adquisiciones.

- Diseñe una estructura de datos que almacene toda la información relevante del experimento y que permita su seguimiento a lo largo del tiempo. Proponga una solución de almacenamiento que la haga accesible a varios usuarios que sobre esta puedan añadir nuevas capas de datos numéricos anotadas (resultados de análisis sobre la adquisición y las anotaciones que los usuarios consideren pertinentes).

BLOQUE 2: Algoritmos y estructuras de datos. Programación en entornos HPC. Programación de GPUs y otros aceleradores. Uso de librerías científicas.

- Usando pseudocódigo y los diagramas que considere necesarios, plantee como implementaría el proceso de análisis de los datos a nivel algorítmico (pasos del 1 al 3). Las operaciones matemáticas pueden darse a partir de funciones conocidas (como por ejemplo funciones de CUDA, matlab o scipy). Si lo ve razonable y posible priorice la paralelización de todo o parte del proceso.

- Justifique su estrategia de implementación y considere cual es la arquitectura que mejor se adapta a sus propuestas y que herramientas de programación usaría.

BLOQUE 3: Gráficos y técnicas de visualización. Uso de librerías científicas.

- Proponga la operación de clustering más apropiada para identificar las diferentes regiones (paso 4). Justifique su respuesta e indique librerías y herramientas para llevar a cabo este trabajo

- Proponga un modo de visualización que combine los resultados de reflectividad y densidad con el resultado de clustering, que sea claro y lo más informativo posible. Puede usar la combinación de gráficos (2D y 3D) que considere conveniente. Complete la representación anterior incluyendo el factor tiempo en esta visualización. Puede ayudarse de texto y dibujos en su explicación e indique que librerías y herramientas emplearía para su implementación.

SUPUESTO TEÓRICO-PRÁCTICO

D2: PROGRAMACIÓN Y COMPUTACIÓN CIENTÍFICA

CASO 2: Adaptación de un entorno de computación y programación para una actividad de análisis en física de partículas

Introducción:

Un grupo de investigación de un centro de un OPI, por ejemplo, en Física Experimental de Partículas, está formado por 4 personas y tiene un problema determinado consistente en:

Los datos originales procedentes del experimento son muy básicos (1 y 0) y hay varias etapas hasta llegar a los datasets abreviados (Ntuplas) que pueden analizarse con un programa de análisis de fase final. Como tenemos un gran volumen de datos, unido a las diferentes etapas de reducción de datos, etc necesitaremos algún sistema de recopilación organizada de dichos datos (Base de Datos) que sea versátil y eficaz.

Al final del proceso de reducción de datos, van a disponer de una serie de Ntuplas con Datos Reales procedente de la toma de datos de un experimento. Y con el fin de verificar modelos teóricos disponen también de Ntuplas de Datos Simulados de las mismas variables que describen el proceso a estudiar.

El formato de los datos en su etapa final puede definirse como de Ntupla y se trata de los siguiente: en el experimento se producen colisiones de dos haces de protones y en el estado final tenemos una serie de partículas salientes de forma que a cada colisión le vamos a llamar suceso (evento) y es la información suministrada en cada una de las filas, y consta de un número *M de sucesos* y los valores de *N variables* con significado físico (V_(i), i:1,…N) obtenidas a partir de las magnitudes que definen cada una de las partículas. La Ntupla tiene la estructura siguiente:

Num_event	V_(1)	V_(2)	V_(3)	V_(4)	…..	V_(N-1)	V_(N)
#1	4.5	124.5	345.1	-1.8		4	-0.3
#2	3.4	86.9	24.5	2.5		8	1.2
#3	-2.0	103.7	560.3	-0.9		7	-4.8
-----	….	…..	…..	…..	…… …….. ……. …..	…..	….
#M-1	-0.7	33.5	456.8	3.2		9	0.4
# M	-1.2	412.0	112.3	1.6		7	5.9

Siendo este formato común tanto a las Ntuplas de datos reales como de datos simulados y los valores de las variables son totalmente académicos.

Por otro lado, este grupo de 4 personas están escribiendo código para su análisis y además colaboran con tres grupos más en centros (nacionales o internacionales). Como ejemplo de

actividad de análisis de la física que nos sirve de caso de uso podemos coger el siguiente: queremos leer los datos, sistematizarlos, representarlas en Histogramas 1D y 2D, estudiar correlaciones entre las variables y luego intentar ajustar a diferentes posibles funciones de forma que consigamos el mejor chi2 de ajuste y al final comparar datos reales y de simulación.

Se plantean las preguntas en los siguientes 3 bloques.

Preguntas para evaluación:

Bloque 1: Integración de los datos

Identificar los principales puntos a resolver en aspectos relacionados con los datos: plantear un sistema de recopilación de los datos

- ¿Cómo podría minimizarse el tiempo de acceso a los datos?
- ¿Cómo se podría gestionar la proliferación de datasets en diferentes etapas del análisis? Comentar el tipo de solución según que el volumen de datos sea de centenares de TB hasta decenas de PB; utilización de Bases de Datos (Relacionales y NoSQL) para agilizar el acceso a los datos.
- ¿Qué formatos de ficheros abreviados se podrían utilizar?

Bloque 2: Desarrollo de Software

- ¿Qué lenguaje de programación elegiría, R o Python? Pros y contras de la utilización de estos lenguajes. Si cree que puede utilizarse otro lenguaje con mejores prestaciones o funcionalidades puede añadirlo y discutirlo.
- Identificar los principales aspectos a resolver del supuesto planteado, entre ellos, encontrar mecanismos o herramientas eficientes para que las personas del grupo colaborativo puedan compartir el código que desarrollan
- ¿Qué herramientas de depuración emplearía (debugging)? ¿Qué métricas de calidad del software utilizaría?

Bloque 3: Librerías e interactividad

En la fase final del análisis, los investigadores tienen la necesidad de trabajar de forma interactiva, por ejemplo, visualizar datos y resultados obtenidos a partir de los cálculos sobre esos datos, y compartir documentos, con fórmulas, visualizaciones y texto, con estos resultados.

- ¿Qué librerías serían necesarias para analizar los datos de forma gráfica y hacer ajustes a determinadas funciones, etc?
- ¿Qué herramientas/plataformas de computación interactiva se podrían utilizar para cubrir las necesidades aludidas encaminadas, por ejemplo, a las tareas de análisis de física que hemos descrito anteriormente?