La investigación científica en Física de Altas Energías (HEP) se caracteriza por desafíos computacionales complejos, que durante décadas tuvieron que ser abordados mediante la investigación de técnicas informáticas en paralelo a los avances en la comprensión de la física. Uno de los principales actores en el campo, el CERN, alberga tanto el Gran Colisionador de Hadrones (LHC) como miles de investigadores cada año que se dedican a recopilar y procesar las enormes cantidades de datos generados por el acelerador de partículas. Históricamente, esto ha proporcionado un terreno fértil para las técnicas de computación distribuida, conduciendo a la creación de Worldwide LHC Computing Grid (WLCG), una red global de gran potencia informática para todos los experimentos LHC y del campo HEP. Los datos generados por el LHC hasta ahora ya han planteado desafíos para la informática y el almacenamiento. Esto solo aumentará con futuras actualizaciones de hardware del acelerador, un escenario que requerirá grandes cantidades de recursos coordinados para ejecutar los análisis HEP. La estrategia principal para cálculos tan complejos es, hasta el día de hoy, enviar solicitudes a sistemas de colas por lotes conectados a la red. Esto tiene dos grandes desventajas para el usuario: falta de interactividad y tiempos de espera desconocidos. En años más recientes, otros campos de la investigación y la industria han desarrollado nuevas técnicas para abordar la tarea de analizar las cantidades cada vez mayores de datos generados por humanos (una tendencia comúnmente mencionada como "Big Data"). Por lo tanto, han surgido nuevas interfaces y modelos de programación que muestran la interactividad como una característica clave y permiten el uso de grandes recursos informáticos.

A la luz del escenario descrito anteriormente, esta tesis tiene como objetivo aprovechar las herramientas y arquitecturas de la industria de vanguardia para acelerar los flujos de trabajo de análisis en HEP, y proporcionar una interfaz de programación que permite la paralelización automática, tanto en una sola máquina como en un conjunto de recursos distribuidos. Se centra en los modelos de programación modernos y en cómo hacer el mejor uso de los recursos de hardware disponibles al tiempo que proporciona una experiencia de usuario perfecta. La tesis también propone una solución informática distribuida moderna para el análisis de datos HEP, haciendo uso del software llamado ROOT y, en particular, de su capa de análisis de datos llamada RDataFrame. Se exploran algunas áreas clave de investigación en torno a esta propuesta. Desde el punto de vista del usuario, esto se detalla en forma de una nueva interfaz que puede ejecutarse en una computadora portátil o en miles de nodos informáticos, sin cambios en la aplicación del usuario. Este desarrollo abre la puerta a la explotación de recursos distribuidos a través de motores de ejecución estándar de la industria que pueden escalar a múltiples nodos en clústeres HPC o HTC, o incluso en ofertas serverless de nubes comerciales. Dado que el análisis de datos en este campo a menudo está limitado por E/S, se necesita comprender cuáles son los posibles mecanismos de almacenamiento en caché. En este sentido, se investigó un sistema de almacenamiento novedoso basado en la tecnología de almacenamiento de objetos como objetivo para el caché.

En conclusión, el futuro del análisis de datos en HEP presenta desafíos desde varias perspectivas, desde la explotación de recursos informáticos y de almacenamiento distribuidos hasta el diseño de interfaces de usuario ergonómicas. Los marcos de software deben apuntar a la eficiencia y la facilidad de uso, desvinculando la definición de los cálculos físicos de los detalles de implementación de su ejecución. Esta tesis se enmarca en el esfuerzo colectivo de la comunidad HEP hacia estos objetivos, definiendo problemas y posibles soluciones que pueden ser adoptadas por futuros investigadores.