Resumen

La utilización de sistemas para el tratamiento eficiente de grandes volúmenes de información ha crecido en popularidad durante los últimos años. Esto conlleva el desarrollo de nuevas tecnologías, métodos y algoritmos, que permitan un uso eficiente de las infraestructuras. El tratamiento de grandes volúmenes de información no está exento de numerosos problemas y retos, algunos de los cuales se tratarán de mejorar. Dentro de las posibilidades actuales debemos tener en cuenta la evolución que han tenido los sistemas durante los últimos años y las oportunidades de mejora que existan en cada uno de ellos.

El primer sistema de estudio, el Grid, constituye una aproximación inicial de procesamiento masivo y representa uno de los primeros sistemas distribuidos de tratamiento de grandes conjuntos de datos. Participando en la modernización de uno de los mecanismos de acceso a los datos se facilita la mejora de los tratamientos que se realizan en la genómica actual. Los estudios que se presentan están centrados en la transformada de Burrows-Wheeler, que ya es conocida en el análisis genómico por su capacidad para mejorar los tiempos en el alineamiento de cadenas cortas de polinucleótidos. Esta mejora en los tiempos, se perfecciona mediante la reducción de los accesos remotos con la utilización de un sistema de caché intermedia que optimiza su ejecución en un sistema Grid ya consolidado. Esta caché se implementa como complemento a la librería de acceso estándar GFAL utilizada en la infraestructura de IberGrid.

En un segundo paso se plantea el tratamiento de los datos en arquitecturas de Big Data. Las mejoras se realizan tanto en la arquitectura Lambda como Kappa mediante la búsqueda de métodos para tratar grandes volúmenes de información multimedia. Mientras que en la arquitectura Lambda se utiliza Apache Hadoop como tecnología para este tratamiento, en la arquitectura Kappa se utiliza Apache Storm como sistema de computación distribuido en tiempo real. En ambas arquitecturas se amplía el ámbito de utilización y se optimiza la ejecución mediante la aplicación de algoritmos que mejoran los problemas en cada una de las tecnologías.

El problema del volumen de datos es el centro de un último escalón, por el que se permite mejorar la arquitectura de microservicios. Teniendo en cuenta el número total

de nodos que se ejecutan en sistemas de procesamiento tenemos una aproximación de las magnitudes que podemos obtener para el tratamiento de grandes volúmenes. De esta forma, la capacidad de los sistemas para aumentar o disminuir su tamaño permite un gobierno óptimo. Proponiendo un sistema bioinspirado se aporta un método de autoescalado dinámico y distribuido que mejora el comportamiento de los métodos comúnmente utilizados frente a las circunstancias cambiantes no predecibles.

Las tres magnitudes clave del Big Data, también conocidas como V's, están representadas y mejoradas: velocidad, enriqueciendo los sistemas de acceso de datos por medio de una reducción los tiempos de tratamiento de las búsquedas en los sistemas Grid bioinformáticos; variedad, utilizando sistemas multimedia menos frecuentes que los basados en datos tabulares, y por último, volumen, incrementando las capacidades de autoescalado mediante el aprovechamiento de contenedores software y algoritmos bioinspirados.