Título: Serverless Computing Strategies on Cloud Platforms
Autor: Diana María Naranjo Delgado
Directores: Dr. Ignacio Blanquer Espert
Dr. Germán Moltó Martínez

Con el desarrollo de la Computación en la Nube, la entrega de recursos virtualizados a través de Internet ha crecido enormemente en los últimos años. Las Funciones como servicio (FaaS), uno de los modelos de servicio más nuevos dentro de la Computación en la Nube, permite el desarrollo e implementación de aplicaciones basadas en eventos que cubren servicios administrados en Nubes públicas y locales. Los proveedores públicos de Computación en la Nube adoptan el modelo FaaS dentro de su catálogo para proporcionar computación basada en eventos altamente escalable para las aplicaciones.

Por un lado, los desarrolladores especializados en esta tecnología se centran en crear marcos de código abierto serverless para evitar el bloqueo con los proveedores de la Nube pública. A pesar del desarrollo logrado por la informática serverless, actualmente hay campos relacionados con el procesamiento de datos y la optimización del rendimiento en la ejecución en los que no se ha explorado todo el potencial.

En esta tesis doctoral se definen tres estrategias de computación serverless que permiten evidenciar los beneficios de esta tecnología para el procesamiento de datos. Las estrategias implementadas permiten el análisis de datos con la integración de dispositivos de aceleración para la ejecución eficiente de aplicaciones científicas en plataformas cloud públicas y locales.

En primer lugar, se desarrolló la plataforma CloudTrail-Tracker. CloudTrail-Tracker es una plataforma serverless de código abierto basada en eventos para el procesamiento de datos que puede escalar automáticamente hacia arriba y hacia abajo, con la capacidad de escalar a cero para minimizar los costos operativos.

Seguidamente, se plantea la integración de GPUs en una plataforma serverless local impulsada por eventos para el procesamiento de datos escalables. La plataforma admite la ejecución de aplicaciones como funciones severless en respuesta a la carga de un archivo en un sistema de almacenamiento de ficheros, lo que permite la ejecución en paralelo de las aplicaciones según los recursos disponibles. Este procesamiento es administrado por un cluster Kubernetes elástico que crece y decrece automáticamente según las necesidades de procesamiento. Ciertos enfoques basados en tecnologías de virtualización de GPU como rCUDA y NVIDIA-Docker se evalúan para acelerar el tiempo de ejecución de las funciones.

Finalmente, se implementa otra solución basada en el modelo serverless para ejecutar la fase de inferencia de modelos de aprendizaje automático previamente entrenados, en la plataforma de Amazon Web Services y en una plataforma privada con el framework OSCAR. El sistema crece elásticamente de acuerdo con la demanda y presenta una escalado a cero para minimizar los costes. Por otra parte, el front-end proporciona al

usuario una experiencia simplificada en la obtención de la predicción de modelos de aprendizaje automático.

Para demostrar las funcionalidades y ventajas de las soluciones propuestas durante esta tesis se recogen varios casos de estudio que abarcan diferentes campos del conocimiento como la analítica de aprendizaje y la Inteligencia Artificial. Esto demuestra que la gama de aplicaciones donde la computación serverless puede aportar grandes beneficios es muy amplia. Los resultados obtenidos avalan el uso del modelo serverless en la simplificación del diseño de arquitecturas para el uso intensivo de datos en aplicaciones complejas.