Técnicas de Recuperación y Procesamiento de Grandes Volúmenes de Datos Sísmicos. Un Repositorio Público Basado en Serverless.

[1]María Murazzo, [1,2]Marcelo Moreno, [1]Nelson Rodríguez, [2]Ricardo Sifón, [2]Valeria Nicolia, [3]Ignacio Benemerito, [3]Leonardo Celador
[1] Departamento e Instituto de Informática - F.C.E.F. y N. - U.N.S.J.
[2] INPRES (Instituto Nacional de Prevención Sísmica)
[3]Alumno Avanzado Licenciatura en Sistemas de Información y Cs. de la Computación - F.C.E.F. y N. - U.N.S.J.
Complejo Islas Malvinas. Cereceto y Meglioli. 5400. Rivadavia. San Juan, 0264 4234129
marite@unsj-cuim.edu.ar, nelson@iinfo.unsj.edu.ar, mmoreno@inpres.gob.ar, rsifon@inpres.gob.ar, vnicolia@inpres.gob.ar, ignacio.benemerito@gmail.com, leonardomiguelcelador@gmail.com

Resumen

El Instituto Nacional de Prevención Sísmica (INPRES), por Ley 19616, es el organismo público en argentina encargado, del monitoreo sísmico de todo el territorio Argentino. Como tal, es poseedor de un vasto e importante catálogo sísmico y de formas de onda, único en Argentina. Estos datos están disponibles en servidores locales, y cualquier usuario, puede solicitarlos. Cada pedido de datos supone un importante esfuerzo, ya que los requerimientos de ese tipo generalmente se refieren a grandes volúmenes de datos, y requiere de una también muy importante capacidad de cómputo, los que se deben realizar y satisfacer desde el mismo Centro de Datos dedicado a tareas de rutina. Dada esta problemática es que se plantea el uso de serverles computing con el objeto de usar una infraestructura cloud pública para alojar los datos y FaaS para implementar las técnicas de manipulación de los datos

Palabras clave: *serverless computing, cloud computing, grandes volumenes de datos*

Contexto

El presente trabajo se encuadra dentro del área de I/D Sistemas Distribuidos y Paralelos y es una de las líneas de investigación internas, del proyecto: Soporte Serverless para aplicaciones móviles de nueva generación, cuya propuesta está en etapa de evaluación para el período 2023-2024 Asimismo el grupo de investigadores viene trabajando en proyectos relacionados con la computación móvil y distribuida desde hace más de 22 años. Como continuación del proyecto anterior: Computación Serverless para tratamiento de datos provenientes de dispositivos de IoT, se continúa el trabajo con investigadores de otras universidades, lo cual favorece notablemente a todos las instituciones participantes.

Introducción

El Instituto Nacional de Prevención Sísmica (INPRES), por Ley 19616, es el organismo público en argentina encargado, entre otras tareas, del monitoreo sísmico de todo el territorio Argentino. Como tal, es poseedor de un vasto e importante catálogo sísmico y de formas de onda, único en Argentina. Todos los datos están disponibles en servidores locales de INPRES, y cualquier usuario, generalmente de organismos científicos, puede solicitar sus datos. Cada pedido de datos supone un importante esfuerzo, ya que los requerimientos de ese tipo generalmente se refieren a grandes volúmenes de datos, y requiere de una también muy importante capacidad de cómputo, los que se deben realizar y satisfacer desde el mismo Centro de Datos dedicado a tareas de rutina.

Además, dos factores de reciente irrupción, han aumentado la carga de trabajo del Centro de Datos de INPRES. Primero, la Ley 26.899 de creación de repositorios digitales institucionales de acceso abierto, que fue establecida en 2013, y que representa un enorme aporte para el avance del conocimiento científico-tecnológico en nuestro país, pero que a la vez, ha sumado presión a las capacidades de INPRES de poder responder a todos los requerimientos que la Ley ha

generado. En segundo lugar, los avances en la ciencia de datos ha alimentado el interés de terceros en el análisis de grandes conjuntos de datos Sísmicos. Temas tales como detección de eventos en volúmenes históricos, Machine Learning, estudios de réplicas, reproceso o refinamiento de localizaciones, etc. Todos ellos requieren disponibilidad de datos de forma de onda continua que abarcan meses o años e incluyen registros de decenas de estaciones.

En INPRES los datos de forma de onda se almacenan en un formato estándar denominado miniSEED, que es a la vez, un subconjunto del estándar para el intercambio de datos de terremotos (Standard for the Exchange of Earthquake Data - SEED), que es un formato de datos destinado al archivo e intercambio de datos de series temporales sismológicas y metadatos relacionados. El formato está definido por la Federación Internacional de Redes de Sismógrafos Digitales (FDSN) y está documentado en el Manual de SEED [1]. En miniSEED se incluyen metadatos muy limitados para la serie temporal, más allá de la identificación de la serie temporal. En particular, no se incluyen las coordenadas geográficas, la información de respuesta instrumental y otra información necesaria para interpretar los datos.

Las series de tiempo se almacenan como registros de datos de longitud fija independientes, cada uno de los cuales contiene un pequeño segmento de valores de series contiguos. Las longitudes de registro comunes son 512 bytes (para flujos en tiempo real) y 4096 bytes (para archivo). Además, existen numerosas librerías de código abierto para facilitar la lectura y escritura de datos miniSEED sin necesidad de conocer los detalles del formato.

La Red Nacional de Estaciones Sismológicas (RNES) [2] de INPRES se compone de cincuenta y cinco (55) estaciones sismológicas, de uno y tres canales cada una, transmitiendo en tiempo real al Centro de Datos de INPRES. Cada registro de una hora de duración para una estación típica a 100 muestras por segundo, tiene un tamaño aproximado de 3 Megabyte, con lo cual un día completo representan 72 Megabytes, y 26 Gigabyte al año, para cada canal. Si la estación registra para tres canales (Vertical, Norte-Sur y Este-Oeste), se tiene que calcular que cada estación sísmica genera aproximadamente 80 Gigabyte de datos al año. Para las 55 estaciones hace aproximadamente 1.5 Terabytes anuales.

El INPRES comenzó sistemáticamente a registrar sus datos de onda en formato miniSEED desde el año 2010 con un comienzo de 20 estaciones, hasta llegar a las 65 actuales. El tamaño total de datos de forma de onda en formato miniSEED en INPRES asciende a 12 Terabytes. Muchos de esos datos están almacenados en dispositivos offline. La siguiente tabla muestra la evolución de estaciones y cantidad de datos almacenados en INPRES a febrero de 2023.

Año	N° de Estaciones	Terabytes/Año
2010	20	0,52
2011	22	0,572
2012	19	0,494
2013	22	0,572
2014	20	0,52
2015	25	0,65
2016	30	0,78
2017	30	0,78
2018	38	0,988
2019	42	1,092
2020	42	1,092
2021	47	1,222
2022	50	1,3
2023	55	1,43
Tamaño total (en TB)		**12,012**

Infraestructura de soporte

En función de lo expresado en párrafos anteriores, se hace necesario contar con una solución que provea el acceso a los datos y al análisis de los mismos en forma eficiente, con mínimo esfuerzo por parte del usuario y con escasa administración y gestión de la plataforma de soporte.

El fenómeno de los grandes volúmenes de datos no es nuevo y ha sido abordado en forma transversal por múltiples áreas del

conocimiento. El tratamiento de estos datos en área de la sismología y en particular la problemática que la presente línea de investigación abordara confluye con cloud computing.

Cloud Computing se ha caracterizado por ser una tecnología centrada en ofrecer cómputo bajo demanda como cualquier otro servicio. Esto es una ventaja para montar aplicaciones donde es necesario el procesamiento intensivo, tales como aquellas aplicaciones que procesen y extraigan información de grandes volúmenes de datos [1].

Los proveedores cloud alegan muchas ventajas en la migración de aplicaciones para el tratamiento de grandes volúmenes de datos, como el acceso rápido a los recursos, costos más bajos y flexibilidad en la contratación y el aprovisionamiento de recursos. Un punto adicional es la seguridad, la cual, afirman es de alto nivel y en muchos casos muy difícil de implementar en la mayoría de los laboratorios, ya que tener personal de TI especializado en seguridad no es común.

Un aspecto más que lleva a la adopción del cloud como plataforma de despliegue de aplicaciones para grandes volúmenes de datos es, mejorar la colaboración científica, es decir, facilitar la investigación colaborativa y la innovación; este aspecto ha sido el foco de este proyecto desde hace varios años al incorporar investigadores de otras universidades del país.

Otro tema es el potencial ahorro de tiempo que ofrece el cloud a los usuarios finales. Estos usuarios no necesitan preocuparse por actualizaciones de software, compatibilidad o parches de seguridad, pues todo esto es aprovisionado de forma transparente [2].

Sin embargo, el cloud tiene dos grandes desventajas, la primera es la degradación de la performance de las aplicaciones al montarlas sobre arquitectura virtualizada, debido a que genera overhead en la contextualización de las máquinas virtuales; la segunda desventaja cuando se despliegan aplicaciones en el cloud, es que es responsabilidad de la organización mantener funcionando de forma correcta la infraestructura que se necesite para el despliegue de las aplicaciones, lo cual lleva a cargar costos sobre el presupuesto para su mantenimiento y soporte [3].

A pesar de esto, en los últimos años, las necesidades de rendimiento y escalabilidad para el procesamiento de grandes volúmenes de datos se han realizado en forma exitosa mediante el uso de frameworks de infraestructuras distribuidas aprovisionados en el cloud. Sin embargo, los costos de personal operativo y los de infraestructura presentan factores económicos significativos para el procesamiento de datos a escala. Además, el procesamiento de grandes volúmenes de datos en el cloud requiere un amplio conocimiento para definir y ejecutar trabajos y para implementar, configurar y mantener la plataforma de infraestructura requerida para su ejecución. [4]

En este sentido, la aparición del Serverless Computing [5] y su core Function-as-a-Service (FaaS), logra que los desarrolladores no tengan que preocupar por el aprovisionamiento y escalado de la infraestructura, por lo que se pueden centrar en la lógica de sus aplicaciones. De esta forma es posible lograr la abstracción de la gestión de servidores (aprovisionamiento, configuración, escalado, etc.) para que los usuarios, en este caso desarrolladores, puedan enfocarse en la lógica de sus aplicaciones [6]. Esto se debe a que las funciones no cuentan con ningún tipo de dependencia de la infraestructura, por lo que se puede cargar en el entorno cloud y ser ejecutadas en cualquier momento. Una de las grandes ventajas de FaaS es que el negocio únicamente se preocupa de la lógica y desarrollo de la funcionalidad de las aplicaciones, mientras que la ejecución, seguridad y mantenimiento de la aplicación es tarea del proveedor cloud [7].

Líneas de Investigación, Desarrollo e Innovación

La presente línea de investigación se enmarca dentro del proyecto mencionado en el contexto y surge como resultado de la cooperación obtenida con el INPRES gracias a que un docente investigador perteneciente al grupo de trabajo desempeña tareas en dicha

institución. Además, se contara con la colaboración de dos profesionales del INPRES

Gracias a esta cooperación el grupo de trabajo contará con una importante fuente de datos para poder realizar las tareas necesarios para trabajar con un back end serverless.

Es por ello que la metodología a seguir será experimental deductiva, lo cual permitirá analizar cómo se comportan las aplicaciones en diferentes entornos de ejecución.

Objetivos

Los objetivos del grupo de investigación en esta línea de conocimiento son los siguientes:
Desde el punto de vista institucional:

- Proveer un entorno en el cloud que permita dar cumplimiento con la Ley 26.899 de Creación de Repositorios Digitales Institucionales de Acceso Abierto, para los datos de forma de onda sísmica disponibles en INPRES.
- Proveer alta disponibilidad y escalabilidad de datos de forma de ondas sísmicas en la nube, que facilite investigaciones que utilizan procesamiento computacionalmente intensivo y grandes volúmenes de datos sísmicos.

Desde el punto de vista científico:

- Analizar la factibilidad de manipular grandes volúmenes de datos con FaaS.
- Estudiar las técnicas de analítica de datos sobre FaaS
- Evaluar la performance de serverless cuando procesa grandes volúmenes de datos.

Formación de Recursos Humanos

El equipo de trabajo de esta línea de investigación está compuesto de seis investigadores que figuran en este trabajo de las universidades Nacional de San Juan y un alumno de grado. Además, el proyecto marco donde se está desarrollando esta propuesta ha establecido vínculos con investigadores de la Nacional de San Luis, de la Universidad Champagnat y de la Universidad Nacional de Salta y dos alumnos de grado.

Se está desarrollando una tesis doctoral sobre paralelismo híbrido y Big Data, una tesis de maestría en áreas afines y dos tesinas de grado en el área de Serverless computing, Concurrencia y Computación distribuida. Además se espera aumentar el número de publicaciones. Por otro lado también se prevé la divulgación de varios temas investigados por medio de cursos de postgrado y actualización o publicaciones de divulgación y asesoramiento a empresas y otras instituciones públicas y privadas.

Referencias

[1] Bahrami, M., & Singhal, M. (2015). The role of cloud computing architecture in big data. Information granularity, big data, and computational intelligence, 275-295.

[2] Amin, R., Vadlamudi, S., & Rahaman, M. M. (2021). Opportunities and challenges of data migration in cloud. Engineering International, 9(1), 41-50.

[3] Sarmah, S. S. (2019). Cloud Migration-Risks and Solutions. Science and Technology, 9(1), 7-11.

[4] Liu, X., Liu, Y., & Fei, Y. (2022). Computer big data analysis and cloud computing network technology. In The 2021 International Conference on Machine Learning and Big Data Analytics for IoT Security and Privacy: SPIoT-2021 Volume 1 (pp. 517-522). Springer International Publishing.

[5] Eismann, S., Scheuner, J., Van Eyk, E., Schwinger, M., Grohmann, J., Herbst, N., ... & Iosup, A. (2020). A review of serverless use cases and their characteristics. arXiv preprint arXiv:2008.11110.

[6] Añel, J. A., Añel, J. A., Montes, D. P., Iglesias, J. R., & Romano. (2020). Cloud and Serverless Computing for Scientists. Springer International Publishing.

[7] Fox, G. C., Ishakian, V., Muthusamy, V., & Slominski, A. (2017). Status of serverless computing and function-as-a-service (FaaS) in industry and research. arXiv preprint arXiv:1708.08028.