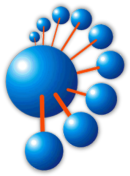
Universidad de Concepción
Facultad de Ingeniería
Ingeniería Civil Informática



Estudio empírico del uso de estructuras de datos compactas en el procesamiento distribuido de datos.

Propuesta de Memoria de Título

Autor: **Leonardo Aravena Cuevas**
Patrocinante: **Dr. José Fuentes Sepúlveda**

1 de septiembre de 2022, Concepción

Introducción

Resulta evidente cómo el avance tecnológico de las últimas décadas ha permitido la proliferación de dispositivos conectados a internet, que van desde pequeños aparatos y poco potentes como cámaras de video, electrodomésticos o sensores hasta las grandes granjas de servidores y *data centers* que forman el núcleo de la red mundial. La abundancia y diversidad de estos dispositivos, junto con la enorme cantidad de datos generados, ha fomentado el desarrollo de distintos paradigmas de procesamiento de estos datos, como el *cloud computing* y el *edge computing*, cuya relación se observa en la figura 1.

Mientras que en el *cloud computing* la información generada por los nodos más alejados de la red es transmitida hacia los núcleos (la *nube*) para ser procesada, en el *edge computing* la información es procesada, parcial o totalmente, más cerca de los nodos en donde se genera esta información (el *edge*), para luego ser enviada a la nube de ser necesario. Este último enfoque tiene la ventaja de reducir la cantidad de información que se transmite y obtener respuestas más rápidas [1].

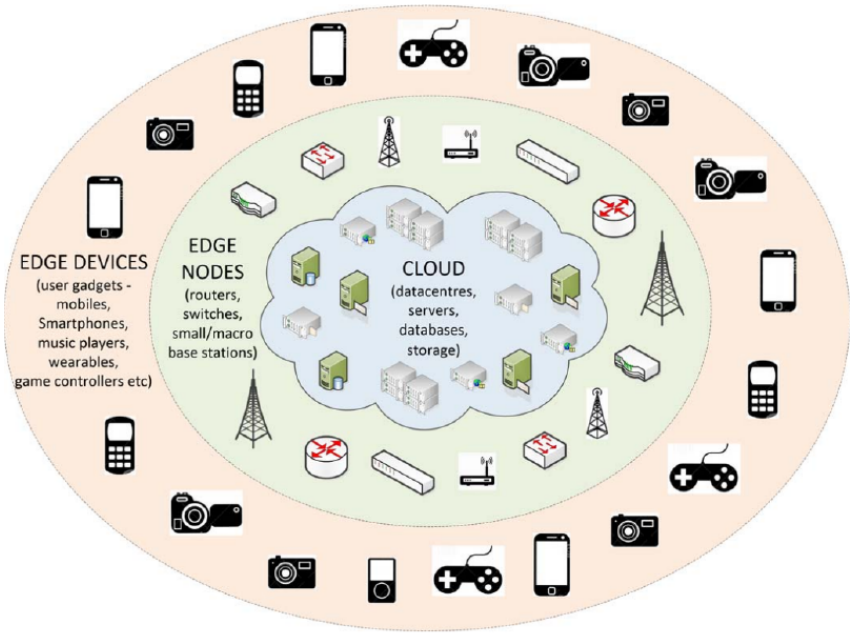


Figura 1: Dispositivos *Edge* y Nodos *Edge* en relación con el *Cloud*. [2]

Una gran parte de los dispositivos que forman parte del *edge* tienen limitaciones de recursos, ya sea de memoria, almacenamiento, velocidad de transmisión y capacidad de procesamiento, por lo que reducir el espacio utilizado por los datos a procesar puede aportar una mejora en los tiempos de respuesta.

La forma más común de reducir el espacio utilizado por los datos es usando algoritmos de compresión, pero se requiere que estos datos sean descomprimidos para su procesamiento. Como alternativa surgen las estructuras de datos compactas (CDS por sus siglas en inglés), las que

permiten una disminución del espacio usado por los datos y permiten realizar operaciones de consultas y manipulación directamente en su forma compacta [3]. Aunque las CDS ya cuentan con un cuerpo sólido de conocimiento, el estudio de su uso en el edge computing es reciente y no se ha abordado en profundidad [4], por lo que en esta memoria se propone evaluar de manera experimental el uso de estructuras de datos compactas en un ambiente distribuido de datos con recursos limitados y así obtener conclusiones sobre si la idea de usar CDS en el edge computing es prometedora.

Objetivos

General

El objetivo general de la memoria de título propuesta es evaluar de manera experimental el uso de estructuras de datos compactas en datos tabulares sobre un ambiente distribuido de procesamiento de datos con recursos limitados para obtener conclusiones que puedan contribuir en esta área de investigación.

Específicos

1.- Implementar un clúster sobre nodos con recursos limitados, el cual servirá como ambiente de pruebas.

2.- Desarrollar un software que sirva de *baseline* para el procesamiento de datos tabulares en el clúster.

3.- Modificar el baseline utilizando estructuras de datos compactas para el procesamiento de datos tabulares en menos espacio.

4.- Evaluar de manera experimental el rendimiento del baseline y la alternativa basada en CDS para el procesamiento distribuido de datos tabulares utilizando el clúster.

5.- Analizar los resultados para obtener conclusiones acerca del uso de CDS en el procesamiento distribuido de datos.

Metodología

La metodología a utilizar durante el desarrollo de esta memoria de título será la siguiente:

* Se realizarán reuniones semanales con el profesor patrocinante para obtener retroalimentación sobre el avance del estudio.
* Se implementará el ambiente de pruebas donde se realizará el estudio experimental de la aplicación de estructuras de datos compactas a los datos tabulares, como se describe a continuación:

Ambiente de Pruebas

El ambiente de pruebas consistirá en:

* Un clúster de cuatro nodos (un nodo *master* y tres *workers*) compuestos por Raspberry Pi versión 4 *model B*, que poseen las siguientes características de hardware:
  - *Procesador:* ARM Quad core de 64 bits a 1.5 GHz.
  - *Ram:* 4 GB para los nodos workers y 8 GB para el nodo master.
  - *Red:* Ethernet Gigabit y WiFi 802.11ac de 2.4 GHz y 5.0 GHz.
  - *Almacenamiento:* 16 GB para los nodos workers y 32 GB para el nodo master en MicroSD.
* El sistema operativo usado en los nodos del clúster será, inicialmente, Ubuntu 18.04.5 LTS. Podría utilizarse otro SO dependiendo del resultado obtenido por los nodos.
* La comunicación entre los nodos del clúster se hará mediante una red local cableada.

Herramientas software

Para el procesamiento distribuido de datos se utilizará el framework Apache Hadoop [5] que permite la programación de aplicaciones distribuidas en el lenguaje Java. Tanto el baseline como la variante que utiliza CDS se implementarán en Hadoop para ejecutarse de manera distribuida y que permita obtener resultados concluyentes del comportamiento del clúster.

La aplicación de CDS a los datos tabulares se realizará utilizando bibliotecas disponibles como la desarrollada por Yauheni Shahun [6].

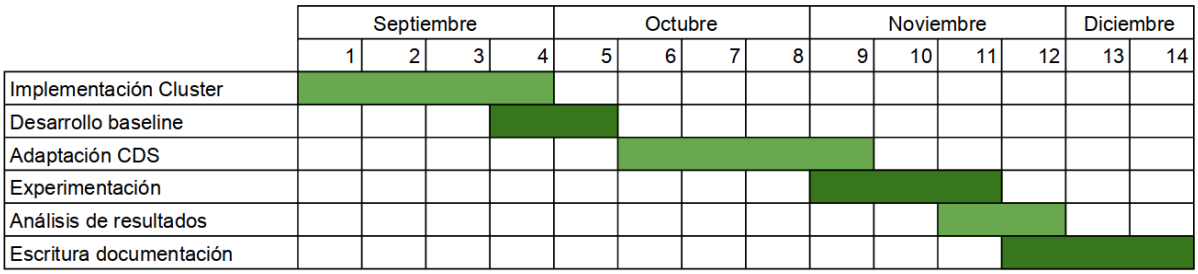
Evaluación

Los datos de prueba que se utilizarán en este estudio serán datos tabulares, los cuales podrán ser sintéticos u obtenidos desde el estado del arte, como el registro de precios del servicio *Spot* de Amazon Web Services que es utilizado en [4].

Se aplicarán distintas configuraciones en el clúster, limitando la cantidad de nodos, cambiando la memoria RAM disponible y la velocidad de transmisión de la red de los nodos para simular las distintas condiciones que pueden encontrarse en los dispositivos *edge*.

Planificación

La planificación para la realización de esta memoria es de 14 semanas, distribuidas como se indica a continuación:



Referencias

[1] N. Mor, "Research for practice: Edge computing," *Commun. ACM*, vol. 62, p. 95, mar 2019.

[2] B. Varghese, N. Wang, S. Barbhuiya, P. Kilpatrick, and D. S. Nikolopoulos, "Challenges and opportunities in edge computing," in *2016 IEEE International Conference on Smart Cloud (SmartCloud)*, pp. 20–26, 2016.

[3] G. Navarro, *Compact Data Structures – A practical approach*. Cambridge University Press, 2016. ISBN 978-1-107-15238-0. 536 pages.

[4] Z. Li, D. Seco, and J. Fuentes-Sepúlveda, "When edge computing meets compact data structures," in *2021 IEEE Cloud Summit (Cloud Summit)*, pp. 29–34, 2021.

[5] Apache, "Apache hadoop." https://hadoop.apache.org/.

[6] Y. Shahun, "Yshahun/succinct-util: Java library for the succinct data structures." https://github.com/yshahun/succinct-util.