Log clustering tool

# Resumen

Contenido

[Resumen i](#_Toc519523891)

[1. Introducción 1](#_Toc519523892)

[2. Bibliografía 2](#_Toc519523893)

# 1. Introducción

Hoy en día se genera una gran cantidad de archivos de logs en campos como la industria y la ciencia, ya que estos son una fuente de información muy importante y útil en muchas situaciones, Sin embargo, a medida que aumenta la complejidad de los sistemas, el análisis de los archivos de logs es cada vez más exigente y dificultoso, ya que hay que realizar un gran esfuerzo para recopilar, almacenar e indexar una gran cantidad de logs que se agrava más cuando estos logs no son heterogéneos **(1)**.

Ccc

La tasa de producción de registros puede alcanzar varios TeraBytes (TB) o PetaBytes (PB) por día. Por ejemplo Facebook lidió con 130 TB de registros diarios en 2010 (https://www.facebook.com/notes/facebook- engineering / scaling- facebook- to- 500- million- users- and- beyond / 409881258919) y en 2014 han almacenado 300 PB de registros (https://code.facebook.com/posts/229861827208629/scaling-the- facebook- data- warehouse- to-300- pb). Debido al tamaño de estos conjuntos de datos, las soluciones de bases de datos convencionales no pueden utilizarse para el análisis, en su lugar se consideran más apropiadas las bases de datos virtuales combinadas con los sistemas de procesamiento distribuido y paralelo (Kumar et al., 2015).

Aunque todavía hay desafíos abiertos, la nube o incluso los sistemas en la nube interconectados (Moschakis y Karatza, 2015) satisfacen las necesidades del análisis de registros. Cloud es una solución probada en el campo que muchas empresas grandes como Facebook, Amazon, eBay, etc. utilizan durante años para analizar los registros. También en la academia hay muchos estudios que investigaron la computación en la nube (principalmente Hadoop) para analizar troncos (Kotiyal et al., 2013; Wang et al., 2014; Narkhede y Baraskar, 2013; Yu y Wang, 2012; Wei et al. ., 2011; Yang et al., 2013; LIN et al., 2013; Liu et al., 2010; Vernekar y Buchade, 2013; Kumar y Hanumanthappa, 2013; Kathleen y Abdelmounaam, 2013). Con lo anterior, podemos suponer que el análisis de registros es un caso de uso de big data y, como resultado, muchos de los desafíos importantes del análisis de archivos de registro basados en la nube son, en realidad, desafíos relacionados con los grandes volúmenes de datos. Estos desafíos, como la variedad de datos, el almacenamiento de datos, el procesamiento de datos y la gestión de recursos, se estudian en muchos trabajos (Assunoa et al., 2015).

Los archivos de registro son una fuente de información muy importante, útil en muchos casos. Sin embargo, a medida que aumentan la escala y la complejidad de los sistemas, el análisis de los archivos de registro es cada vez más exigente. El esfuerzo de recopilar, almacenar e indexar una gran cantidad de registros se agrava aún más cuando los registros son heterogéneos.

ss (1)

# 2. Bibliografía

1. *Performance evaluation of cloud-based log file analysis with Apache Hadoop and Apache Spark.* **Ilias Mavridis y Helen Karatza.** 2017, The Journal of Systems and Software, págs. 133-151.

2. **[En línea] https://spark.apache.org/.**