INFORMACIÓN VERAZ PARA LA REACTIVACIÓN ECONÓMICA EN EL CONTEXTO PANDÉMICO

Pedro Donis
Universidad Mariano Gálvez, Guatemala
https://orcid.org/0000-0003-0844-9796
pdonisd1@miumg.edu.gt

INTRODUCCIÓN

Los grandes volúmenes de datos tienen como consecuencia un nuevo paradigma de cómputo donde múltiples trabajadores o nodos de trabajo necesitan estar orquestados. La investigación analiza los factores que pueden interferir o contribuir para que la calidad de la información se mantenga dentro de los diferentes sistemas de información o conjuntos de datos del universo de datos en determinada industria.

El método principal consiste en estudiar el rastro o huella que los sistemas de *Big Data* dejan y que contribuyen como evidencia para poder medir las tasas, probabilidad y relaciones por duplicidad, ausencia total o parcial de información, para que no afecte en los procesos donde se utiliza la misma. Esta puede ser para un análisis forense, monetización, seguridad de la red y otras categorías que la industria requiera.

Entre los resultados principales se observan cuatro grupos de información completa, dos de ellos de forma similar, y los otros dos tienen características extremas; o sea, un grupo muy reducido y, en el extremo derecho, un grupo muy grande, donde el proceso se ha mantenido consistente durante un tiempo. Se analiza un conjunto de ocho muestras de una población total de cincuenta y tres variables indicadoras de calidad de información para un conjunto de datos *big data*. Estas variables han presentado errores importantes en la muestra investigada, por lo que se determina que la fuente no es 100 % confiable dadas las tasas de errores. Esto se debe a problemas en el proceso que genera el conjunto de datos o la fuente primaria operativa.

Objetivos

- Determinar la calidad de la información por medio de un método que clasifica los errores para poder establecer un nivel de impacto en los resultados de decisiones basadas o dirigidas en datos.
- Determinar sus implicaciones en los análisis generados a partir de estos datos, es decir, si ocasionan resultados inconsistentes a los analistas de la información.
- Analizar la relación existente entre las variables que fueron medidas en la investigación, sus relaciones y la influencia en el mantenimiento de la información confiable y exacta.

METODOLOGÍA

La investigación es cuantitativa, porque se analizan variables relacionadas con la consistencia de los datos en base al historial de las ejecuciones del proceso Spark por medio de ejecutores y haciendo uso del disco por medio de HDFS. Este sistema permite procesar los datos de manera que todos los nodos (50) de trabajo procesen parte del conjunto de *big data*.

Cabe mencionar que la captura de datos y las mediciones de tamaño informático -es decir, bytes y sus diferentes escalas de medición: kB, MB, GB, etc.— se hizo mediante marcas.

De los cincuenta y tres elementos de población, se tomaron ocho muestras para que existiera continuidad de los datos en la gráfica, por lo que se considera una investigación explicativa del flujo de datos históricos.

RESULTADOS

Para que la calidad mantenga el valor de la variable se debe mantener la tendencia y nunca acercarse al valor cero, dado que esta variable representa la variación de la medición de los datos y puede cambiar respecto a la marca anterior. Estas marcas temporales pueden ser observadas en el gráfico.

Cuando el valor tiende a cero es porque no se aprecian datos. Esto puede ser consecuencia de las siguientes causas: 1. No hay datos o no se produjeron, lo cual es muy improbable. 2. El sistema de *big data* ha fallado, por lo que no fue posible procesar la información. 3. La operación o el proceso de toma de las mediciones ha fallado. Esto último tampoco es muy probable.

Las sesiones de Spark generan ejecutores de trabajo los cuales acceden a información de datos en disco. Es importante hacer notar que el procesamiento se lleva en memoria para evitar latencias de lectura y escritura en el trabajo que se está analizando. Por ende, la lectura y escritura solo se hace una vez y para el proceso de investigación se realizan anotaciones en ficheros de texto en los resultados de las variables a medir.

CONCLUSIONES

Se puede determinar la calidad de la información en ambientes de *big data* analizando las variables presentadas en esta investigación. Igualmente, analizando las variables se pudieron segmentar cuatro grupos categóricos espe-

ciales para los cuales el tiempo útil del trabajo ha representado un reto principal. Por último, se recomienda replicar el estudio para más nodos de trabajo donde se puedan hacer variar las capacidades de los ejecutores en el trabajo y se pueda repartir de diferentes maneras y bajo varios escenarios.

REFERENCIAS

Andrade, C. (2022). Designing Monitoring Systems for Complex Event Processing in BigData Contexts (DOI: 10.1007/978-3-030-95947-0J>)

Andrade, C. (2022). A BigData Perspective on Cyber-Physical Systems for Industry 4.0: Modernizing and Scaling Complex Event Processing

Roriz, M. et al. (2019). An introduction to data stream processing: a complex event processing approach (DOI: 10.1145/3323503.3345028)