CAPÍTULO 2

CALIDAD EN BIG DATA ECOSISTEMA TIC: BIG DATA COMO UNA PROPUESTA DE INVESTIGACIÓN APLICADA A LA CALIDAD DEL DATO PARA EL CORRECTO GOBIERNO DE UN ECOSISTEMA TIC

Data de aceite: 03/04/2023

Pedro Elizardo Donis del Cid

RESUMEN: El término "dato" deriva del latín DATUM y su significado es lo que se da en el sentido de lo que acontece. Big Data es la terminología utilizada para designar a grandes volúmenes de datos que se generan con una determinada velocidad en la generación de la información y variedad en el formato en que se guarda. El objeto del artículo es determinar los factores que definen la calidad global de la información en términos de variables de unidades de información por medio del marcaje según las tipologías de los errores presentados en la metodología de esta investigación. Se concluye que para efectos de un adecuado gobierno del dato es necesario considerar los pequeños errores en las fuentes de datos para que los datos puedan ser utilizados por la industria en diferentes aplicaciones a través de los profesionales capacitados para el efecto. Siendo la hipótesis de la investigación válida donde, existe la siguiente relación proporcional: cuando el valor tiende a 1 existe alta calidad de la información y cuando tiende a 0 existe baja calidad, según las observaciones realizadas, siendo esta una relación directamente proporcional, entre las variables medibles y la categórica de calidad.

PALABRAS CLAVE: Big Data, Hadoop, MapReduce, NoSQL, Análisis y modelado de datos.

INTRODUCCIÓN

El término Big Data se ha utilizado en diferentes contextos tecnológicos debido a los avances que se tenido, sin embargo, en los últimos años ha tenido un gran valor en términos de investigaciones relacionadas con este enfoque. Big Data excede el alcance de los entornos de hardware de uso común y herramientas de software para capturar, gestionar y procesar los datos dentro de un tiempo transcurrido tolerable (Teradata Magazine, 2011) para su población de usuarios. Se trata de datos que crecen de manera exponencial y que los sistemas tradicionales, a pesar de los avances en tecnología, no logran soportarlo (almacenar, procesar y visualizar) para poder obtener el valor efectivo en tiempo del análisis de los datos. Big Data se refiere

a los conjuntos de datos cuyo tamaño está más allá de las capacidades de las herramientas típicas de software de bases de datos para capturar, almacenar, gestionar y analizar. Esta definición es, según McKinsey Global Institute (2011). Según Gartner, Big Data, son los grandes conjuntos de datos que tiene tres características principales: volumen (cantidad), velocidad (velocidad de creación y utilización) y variedad (tipos de fuentes de datos no estructurados, tales como la interacción social, video, audio, cualquier cosa que se pueda clasificar en una base de datos).

En este artículo se estudia la veracidad como propiedad del Big Data en base a su calidad como dato y disponibilidad en un sistema con una velocidad de integración constante, con marcas tanto de tiempo como de volumen o unidades de información que son guardados en dispositivos electrónicos.

El método utilizado es el uso de herramientas Hadoop, MapReduce, NoSQL y BigData para generar indicadores importantes que minimicen la cantidad de errores en base a tipologías de error relacionadas con variables de entrada y salidas de un procesamiento específico en Big Data en diferentes períodos o casos de estudio. Con variables que miden la relación del volumen de entrada medidos en bytes con los volúmenes de salida con la misma unidad de medida y, la relación que existe entre cada medición y su antecedente.

El proceso trabaja directamente sobre el motor de procesamiento de datos de código abierto Apache Spark creado por Matei Zaharia en la Universidad de Berkeley. Programando en el lenguaje Scala.

APARTADO TEÓRICO Y CONCEPTOS GENERALES

Grupo IGN (2017) Erick Larson utilizó en primera ocasión el concepto Big Data y lo hizo una publicación de un tema relaciona con marketing y datos de clientes. La palabra dato deriva del latín DATUM y significa lo que se da en el sentido de lo que acontece, es decir, antecedentes necesarios para llegar al conocimiento exacto de una cosa. Existe un estándar ISO para la calidad de datos se trata de la norma ISO 8000.

Industrias importantes toman decisiones en base a datos masivos, entre las cuales se encuentra transporte, alimentación, comunicación, banca, etc. Así pues, las plataformas tecnológicas y los datos que se pueden guardar en almacenes históricos logran realizar análisis de tipo correlacional, predictivo e histórico a través de métodos estadísticos representados por algoritmos y herramientas para comparar efectividad. Según la investigación de Réda, et al. (2020), en la Industria 4.0, la generación de código se puede obtener de procesos estadísticos dinámicos a partir de datos sin la necesidad de que estas acciones sean previamente programadas.

Algunas de las áreas que utilizan el Big Data que pueden mencionarse son la de bancos, salud, educación, comercio, gobierno, entre otros. De acuerdo con da Silva (2021) se confirma que la empresa de vestuario y calzado deportivo Nike ofrece servicios de

prendas que monitorean datos de la actividad deportiva mientras se tienen puestas las prendas, lo cual mediante la evaluación de la cantidad de datos que los consumidores se ingresan por segundo, pueden determinar cuántos consumidores hacen uso de la prenda con el servicio que ofrecen. También determinan en qué países se da más uso de estas prendas mediante el ingreso que los consumidores hacen en sus redes sociales y a la vez pueden determinar en qué horas del día es donde más se comparte la información en tiempo real. Y al mismo tiempo se verifican que los datos sean verdaderos según los datos y las prendas que se han vendido.

En el ámbito educativo, por ejemplo, se produce un gran volumen de datos a través de cursos en línea, actividades de enseñanza y aprendizaje (Oi, Yamada, Okubo, Shimada y Ogata, 2017). Con la llegada de los grandes datos, ahora los profesores pueden acceder al rendimiento académico de los estudiantes, los patrones de aprendizaje y proporcionar comentarios instantáneos (Black & Wiliam, 2018). La retroalimentación oportuna y constructiva motiva y satisface a los estudiantes, lo que repercute positivamente en su desempeño (Zheng & Bender, 2019). Los datos académicos pueden ayudar a los docentes a analizar su pedagogía de enseñanza y afectar los cambios de acuerdo con las necesidades y requisitos de los estudiantes. Se han diseñado muchos sitios educativos en línea y se han introducido múltiples cursos basados en las preferencias individuales de los estudiantes (Holland, 2019). La mejora en el sector educativo depende de la adquisición y la tecnología. Los datos administrativos a gran escala pueden desempeñar un papel tremendo en la gestión de diversos problemas educativos (Sorensen, 2018). Por lo tanto, es esencial que los profesionales entiendan la efectividad de los grandes datos en la educación para minimizar los problemas educativos.

Para el 2021 el 8.5% de las empresas españolas utilizan Big Data. Según los datos del INE el 8.5% corresponde a industria, servicios y construcción principalmente. Además, el interés por el Big Data según Google Trends ha tenido crecimientos desde el 2004. Pero en estos últimos años se ha disparado drásticamente. Dentro de las tecnologías que potencializa este término tenemos: Internet de las cosas, robots industriales, robots de servicios, análisis de grandes volúmenes de datos, chat bots. Pero existe una necesidad de profesionales especializados, en España, por ejemplo, solo el 18.4% cuentan con especialistas en TIC y solo 8.5% utilizan el análisis de Big Data. Especialistas en TIC Big Data por sectores y regiones:

Gráfico 1: Imagen adaptada de: https://bigdatamagazine.es/2021-el-ano-del-big-data#comment-10566

Industrialización

Galvão et al. (2022) Osmana y Ghirana, 2019 definen a la automatización como la base de la industria 4.0 combinado con Big Data. Para lo cual se necesitan sensores y otras tecnologías especializadas para la comunicación máquina a máquina. Qaffas et al. (2021) el IoT constituye un avance importante para este proceso. Franke et al., (2016) Qaffas et al. (2021) Tang et al. (2019) van Evert et al. (2017) conceptualizan al Big Data como una tecnología emergente dado exponenciales crecimientos en volumen de datos y limitaciones de Hardware. Varios nodos, núcleos y memora trabajan en una misma tarea. Qaffas, et al. (2021) IoT recolecta datos de manera automática para investigaciones solamente configurando equipos y parametrizando sistemas.

Galvão et al (2022) ha utilizado la automatización de un proceso industrial usando técnicas de Big Data con Spark y Python. Modelando datos, funciones de agregación y objetos JVM con mapas y reductores. La aplicación de API's de este tipo son utilizadas para la adquisición, procesamiento y almacenamiento de datos con el fin de lograr la comunicación entre máquinas.

Es posible ahora predecir (Galvão el al., 2022) usando variedad de datos de tecnologías de Big Data y IoT, por medio de ETL. La informática hace posible recolectar, almacenar y procesar la información que llega desde máquinas como es el caso de IoT. Dada la tecnología actual. Lo más importante son los datos porque sirven para toma de decisiones, por medio de la transformación en información y luego en conocimiento. La retrospectiva y descriptiva avanzada (Mohamed et al.,2019) utiliza información con datos no estructurados y estructurados para generar correlaciones con mayor objetividad.

Qaffas et al (2019) The Internet of Things and Big Data Analytics for Chronic Disease Monitoring in Saudi Arabia, The Steps of Support Vector Machine Classifier, muestra un caso de studio donde existen variables de entrada las cuales son definidas como Hypertension disease data set y variables de salida que comprende una clasificación el SVM. SVM o support vector machine muestra como se puede aplicar por medio del método: Cargada del conjunto de datos, agregación de manera aleatoria un conjunto de entrenamiento del

20% y un conjunto de datos de prueba del 80%, en el método del caso de estudio, se incluye una generación de clasificadores en base al conjunto de datos de entrenamiento, posteriormente, se entrena el clasificador, se construye la predicción aplicada a la muestra por medio del clasificador que se ha entrenado.

El caso de estudio incluye una evaluación de los clasificadores basados en los parámetros, así como una selección de sus características según sus pesos; para poder aplicar este tipo de biotecnología en la resolución de problemas asistidos por ordenadores en la clasificación de los parámetros o factores que inciden en la hipertensión, es necesario hacer uso de IoT.

Al final de la investigación se determinó que los datos basados en IoT son muy efectivos y prometedores, sin embargo, es necesario realizar alguna optimización; dentro de los resultados relevantes del estudio se tiene que las personas mayores con diabetes tienen más probabilidad de desarrollar hipertensión. El tabaquismo juega un papel menos importante que la diabetes y las personas mayores deben consumir menos sal. SVM produce mejores resultados que el algoritmo C4.5.

Kumar et al (2021) realizó un estudio utilizando un enfoque escalable de detección de intrusos haciendo uso de un Framework de Big Data utilizando métodos de clasificación como K-Means, RUSBoost y DT.

En la industria se han utilizado los marcos de trabajo como Spark y Storm, los cuales son utilizados en combinación con ATCS. En estos marcos de trabajo existen muchos parámetros de configuración en el caso de Spark más de 180 parámetros para aplicaciones, por lo que en algunos casos resulta ineficaz o no disponible por el manejo de tantos parámetros configurables para diferentes escenarios. Por lo cual surge la necesidad de autoajuste para marcos generales de procesamiento de Big Data. Después del estudio, de sintonización automática de parámetros se mejora el rendimiento del sistema para buscar las configuraciones óptimas Li et al (2020).

Kibria et al (2018) describe los elementos de la red en un entorno interconectado donde participan Micro BS, Drone BS, Macro BS, Massive MIMO, D2D, Cloud-RAN, BBU Pool, MTC, Multi-Access Connectivity, Ultra-Dense Network, V2X, 3D Beamforming, AP. Todas estas redes interconectan servidores de datos, personas, dispositivos como automóviles conectados a la red, teléfonos inteligentes, redes Wi-Fi, empresas y múltiples dispositivos de Smart City. Por ejemplo, el Massive MIMO sirve para conectar industrias y Ultra-Dense Network para conectar a múltiples dispositivos reunidos en un entorno pequeño como un concierto. V2X puede utilizar una red de dispositivos en postes de alumbrado eléctrico para poder controlar automóviles autónomos o manejados sin intervención humana, por medio de sensores y redes.

Qi and Tao (2018) hicieron un estudio de la industria 4.0 donde se integran sistemas ciberfísicos para la fabricación. En el estudio se hace diseño, planificación de la producción, fabricación y mantenimiento predictivo por medio de Big Data y Digital Twin. Tomando

en cuenta que la automatización industrial ya es una realidad, es posible, ahora generar información de estos, tanto de producción y de consumo que ayuden a los fabricantes a tomar mejores decisiones respecto al producto.

Explotación

Abdessatera et al (2020) ha utilizado un análisis multivariado de datos mediante una regresión logística mediante una encuesta con una duración de tres días para poder determinar la salud física y psicológica de jóvenes, fueron encuestados el 55.5% de los 495 miembros, de los cuales el 90% respondieron que se sintieron más estresados durante la pandemia. Lo cual tuvo un impacto importante en la calidad de su trabajo.

Van Evert et al (2017) realizó un estudio de control de malezas haciendo usos de sistemas de Big Data y dentro de los métodos utilizados se hizo uso de ambientes NoSQL ('no solo' Structured Query Language) porque los sistemas tradicionales resultan difíciles de manejar cuando se particiona el sistema en varias máquinas. Se hizo uso de técnicas SVMs y el conjunto de datos respaldados por Big Data para el control de amenazas incluye incluye información espacial, la posición del paisaje, las características del suelo y del clima.

Dentro de los resultados obtenidos se obtuvo que, si se puede controlar mediante los factores de tiempo, severidad y ubicación, además, se determinó que en las temporadas de crecimiento de la maleza es cuando se deben de capturar estos datos en tiempo real para obtener avances en la ciencia agrícola en TICs para la captura, almacenamiento de datos, análisis y los aportes en la misma, para implementar un sistema de agricultura de precisión es necesario tener formas automáticas para la obtención o inyección de datos, de los cuales los más importantes son: sensores remotos, cartográficos, datos de análisis de suelo, datos de exploradores de campo; esto último se puede llevar a cabo mediante la tecnología móvil; adicional a esta información se puede obtener información en tiempo real del tiempo, si esta soleado, nublado, lluvioso, etc.

Luego de estos datos importantes el sistema de evaluación debe de considerar una fase de procesamiento de datos, es aquí donde el cómputo juega un papel muy importante, por medio de operaciones estadísticas automatizadas, se construyen los mapas temáticos de prescripción. Y para la implementación en el campo por medio de los datos de entrada hacia VRA, se utilizó una construcción de predictores de rendimiento y de calidad de datos que se almacenan en bases de datos que soportan una gran cantidad de información histórica de varios períodos y además se construyan las bibliotecas de referencia. En resumen, este caso de estudio de grandes volúmenes de datos necesita una correcta adquisición de los datos para su posterior procesamiento.

La implementación en el campo obtuvo importantes hallazgos: se debe controlar especialmente las malezas invasoras, parasitarias y las resistentes a herbicidas, pero es necesario un esfuerzo de expertos en informática, ciencia de datos y expertos agrícolas

para que todo el sistema sea automático, funcional y eficiente en la producción agrícola. También se notó que existen otros aspectos organizativos, éticos y legales para la correcta administración de los datos. Para finalizar, este caso: adquisición de datos, procesamiento de datos e implementación; los repositorios de datos históricos Big Data o bases de datos NoSQL necesitan de metodologías de Big Data para que sea independiente de la plataforma, escalable y de código libre, es decir se pueda replicar a muchos sectores demográficos; siempre en el contexto de la agricultura como parte del modelado de las necesidades para implementar sistemas de Big Data.

Lavalle (2021) describe a la analítica visual de los grandes volúmenes de datos como una parte crucial para la toma de decisiones sobre datos que requieren un modelo visual de datos acompañado de requerimientos y descripción detallada de las fuentes de datos. Dentro de los agentes involucrados los requisitos para el modelo otorgados por el usuario, las especificaciones de visualización, el modelo de visualización de datos, las fuentes de datos y su modelo correspondiente y el proceso de detención de sesgos. Todos estos pasos implementados conllevan a la implementación y al monitoreo periódico de los resultados obtenidos, que básicamente son visualizaciones de datos que pueden ser interpretadas por el usuario. Los elementos visuales serán el resultado de un sistema controlado de procesamiento e ingesta de información a las fuentes de datos, modelado, procesamiento y especificaciones.

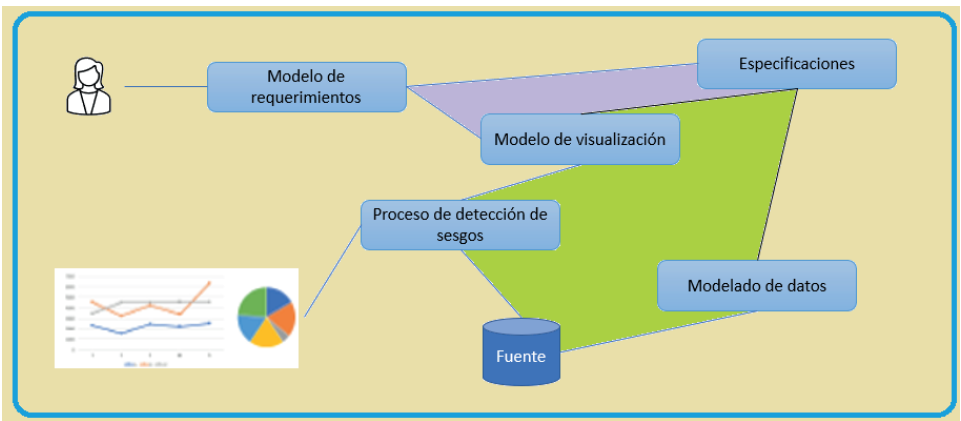

Gráfico 2: Imagen adaptada de: http://hdl.handle.net/10045/119626

Para poder entender la visualización de estos datos es necesario entender el contexto de los datos y del modelo de este. Contar con una fuente fiable de datos de Big Data y los procesos que faciliten la disponibilidad periódica de la información.

Ciencia de Datos

Jamshidi et al (2020) han utilizado DL para el diagnóstico y tratamiento de la

enfermedad COVID19. Además, han utilizado técnicas de GAN, ELM y LSTM. Usando un enfoque bioinformático donde los datos no estructurados son abundantes y requieren ser usados por médicos e investigadores. La principal ventaja de la plataforma es acelerar el diagnóstico y tratamiento. El Big Data es utilizado como repositorio de información para análisis. Después de utilizar RNN, LSTM, GAN y ELM se han logrado identificar patrones de propagación además se mejora la velocidad y precisión del diagnóstico. Ayuda a los expertos en la materia en el desarrollo de enfoques terapéuticos nuevos eficaces incluyendo el descubrimiento de rasgos genéticos y fisiológicos que hacen vulnerable a una persona. Es importante hacer notar que el sistema es capaz de recolectar datos de diferentes fuentes como los sistemas de emisión de boletos de avión lo que habilita la propiedad de volumen del Big Data. Los conjuntos de datos son tanto clínicos como no clínicos.

Principalmente se estudia la insuficiencia cardiaca por la inflamación originada por la enfermedad. Sobre todo, en paciente con problemas cardiovasculares, dentro de los principales resultados y por medio de las técnicas RNN, LSTM, GAN y ELM; se tiene una correcta identificación de personas de alto riesgo analizando grandes volúmenes de datos y colocando esta información a disposición de los expertos, sin embargo, como todo, tiene sus ventajas y limitaciones por lo que resulta importante abastecerse de un arsenal de plataformas, métodos, enfoques y herramientas para lograr los objetivos.

Los operadores de red tienen acceso a grandes volúmenes de datos de la red y de los suscriptores, por ello lo que pretende este estudio es hacer la gestión de red auto adaptativo, proactivo y prescriptivo para los sistemas de comunicación de próxima generación para las BS (macro, micro, femto, pico) dados los desafíos actuales en aplicar coberturas con escasez tanto de capital como el espectro. En este estudio al ampliar la variedad de fuentes de datos se ha utilizado NWDA lo que requiere más esfuerzo de optimización. Esto logra mejoras en la planificación y control de la red, así se otorga valor a la analítica. Con el análisis de macrodatos se puede transmitir una mejor intuición y comprensión por las múltiples fuentes para revelar patrones y correlaciones. Logrando QoE. Kibria et al (2018).

METODOLOGÍA DE LA INVESTIGACIÓN

Varl es la variación del conjunto de entrada de datos para el proceso identificado con valores porcentuales y que responde al nombre de variación de lectura, se divide la cantidad de unidades de información de la marca anterior con respecto a la actual.

Vare es la variación del conjunto de salida de datos para el proceso identificado con valores porcentuales y que responde al nombre de variación de escritura, se divide la cantidad de unidades de información de la marca anterior con respecto a la actual.

Var es la variación del conjunto de la salida de datos respecto a la entrada de datos. El proceso está identificado con valores porcentuales y responde a la variación escritura y

lectura, se divide la cantidad de unidades de información de la escritura con respecto a la lectura. Esta última depende directamente de varl y vare.

Además de las variables se debe de comprender la forma de tipificar (clasificar) los errores, para lo cual tenemos como valores iniciales:

- Los problemas operativos, error400 datos faltantes, 401 datos duplicados.
- Errores en la calidad identificados como moderados: 402 bajas (valores muy bajos) y 403 altas (valores por encima de lo normal).
- Errores de calidad severa: 404 valores atípicos, 405 variación con valor cero.
- Errores no tipificados, falta de datos de muestra Error406 los cuales se excluyen del análisis de la investigación porque están implícitos.

Este mismo criterio aplica para los errores con el rango 300 al 306 y 200 al 206; los cuales corresponden a otras variables medibles de diferentes perspectivas. En los anexos uno y dos se muestra una descripción de los prefijos para entender mejor las variables y contenido operativo.

Entonces el problema radica cuando un trabajo se puede levantar por tiempo indefinido, para lo cual puede tener dos estados finales suspendido o removido del ecosistema, en el caso la herramienta computacional de Big Data lanza una excepción al darse de baja o cancelarse un trabajo, porque se dio un suceso inesperado.

Entonces el objetivo de la investigación es estudiar, clasificar y determinar tanto las razones del suceso inesperado como la calidad global del conjunto de datos del experimento que tuvo a bien el caso de estudio.

El conjunto de datos resultante de procesamiento en Big Data genera algunos indicadores importantes de información de la red entre ellos se destaca el cliente identificado por su identificador único en la red móvil, este indicador surge otro como la cantidad de clientes distintos para un determinado período de tiempo a lo que se le llama parque. Obtener esta información directamente de la fuente principal sería un proceso con alto consumo de los recursos del Big Data, dada la frecuencia de uso. El indicador más importante es la cantidad de tráfico de navegación y sus variaciones operativas. Entre ellas la aplicación, el servidor y su identificación, los grupos clasificadores, reglas adicionales, el protocolo de comunicación.

La operación de la variable y definición paramétrica del caso de estudio para varl su rango es [<=0.8] errores con valores bajos y [>=1.2] para los errores con valores altos es decir la variable se encuentra en este rango de aceptación [0.8<varl<1.2]. Para la otra variable (variación de escritura) tenemos los rangos: vare [<=0.8] [>=1.2].

DISCUSIÓN DE LA RUTA METODOLÓGICA

La hipótesis indica que los valores que representan un error en calidad de la

información están representados por los valores de varl, vare y var que tienden a cero. Es decir, cuando las variables tienden a cero es un error en el motor de procesamiento, y de forma invertida cuando, tienden a 1 los datos presentan mayor fidelidad.

Para iniciar el análisis de casos de estudio tenemos el siguiente: (502), este estudio trata de determinar cuáles son los experimentos con mayor incidencia de errores durante las cargas en un sistema de información basado en tecnologías de Big Data en un ambiente de archivos distribuidos (hdfs). Para ello se analizaron cincuenta y tres experimentos (representados por semanas del año calendario anual) donde, cada experimento tiene asignados siete grupos dimensionales semejantes (días de la semana de domingo a sábado) y un conjunto de jobs, procesos o trabajos, que se corren en el ambiente de Big Data cada día e interactúan con los conjuntos de datos de entrada y generan un conjunto de datos de salida. Estos experimentos (Jobs) se ejecutan de forma secuencial, en total veinticuatro por cada grupo dimensional del experimento, cada grupo dimensional sigue patrones semejantes en base al día de la semana y cada uno representa una medición que corresponde a datos de entrada y de salida para un día determinado siguiendo la descripción metodológica. Para tal efecto, siguiendo la metodología de Big Data, cada job cuenta con ejecutores que llevan a cabo tareas con un conjunto de registro de entrada y salida cada una. Esto lo hace por medio de tareas (etapas) del job.

En el sistema de Big Data analizando los logs de errores para determinar la calidad de la información; en el presente caso de estudio se analiza el comportamiento de la variable que determina el nivel de medición o relación entre una entrada y una salida de datos en un sistema de información. El sistema de información comprende un conjunto de datos medidos en bytes de entrada y otro de salida, esta variable consiste en el valor de razón de estos y se filtran algunos casos donde aparentemente se tienen errores. El conjunto de datos se identifica con el año, experimento, día de la experimento y hora de la ocurrencia del error.

Se realiza el análisis del siguiente caso de estudio, este, es parte de la tipología de errores en el sistema de Big Data que consiste en un conjunto de datos de entrada y otro de salida, para lo cual se analizan varias variables que influyen en el comportamiento de este. Estas variables tienen que ver con el volumen de los datos, es decir, la cantidad de bytes, mb, gb o tb que corresponden. Estas mediciones de volumen de datos están relacionadas entre sí, para construir las variables de razón que se analizan en este artículo.

Cuando hablamos de experimento nos referimos a un espacio temporal donde se estudia la variable con la cual se puede evaluar la magnitud o medición del error o posible error. Estos espacios temporales se dan por ciclos, son cincuenta y tres ciclos por año. Este caso de estudio (502) tiene los siguientes rangos de errores y de aceptación: varl[<=0.8][>=1.2] vare[<=0.8][>=1.2]. Es decir, el rango de aceptación es el siguiente 0.8<=varl<=1.2 y 0.8<=vare<=1.2.

En este gráfico podemos ver el historial de errores para cada experimento temporal

realizado, todos los valores que se dan al alza corresponden a rangos donde la variable es tipificada en este error, no necesariamente tiene que ser error, pero la hipótesis afirma que se ha presentado un error, sin embargo, entre menor sea el valor de esta variable mayor será la posibilidad de que sea un error efectivo y haya pérdida importante de información en el job por las causas presentadas en esta investigación.

RESULTADOS Y DISCUSIÓN DE LA HIPÓTESIS

A continuación, se muestra una tabla de muestra de las mediciones que se toman en consideración. Siendo las más importantes, entrada y salida, fecha, y el identificador del experimento; con otras variables derivadas (x, y, z)

CA	SU	EX	ENTRADA	SALIDA	FC_CARG	VAR(X)	INICIO_EXPER	FIN_EXPERIME	VAR_L(Y)	VAR_E(Z)
5	1	1	493053813252	31101966718	20220102	0.06308026808040	3/11/2022 12:17	3/11/2022 14:44	0.93936734174157	1.00763616064813
5	1	2	493765998187	32269493369	20220103	0.06535381838257	3/11/2022 14:44	3/11/2022 17:48	1.00144443652165	1.03753867598104
5	1	3	495239547843	32588672327	20220104	0.06580385687884	3/11/2022 17:48	3/11/2022 19:54	1.00298430767086	1.00989104335634
5	1	4	497166219799	32547872257	20220105	0.06546678145220	3/11/2022 19:54	3/11/2022 22:00	1.00389038388471	0.99874802908230
5	1	5	493463712292	32023808723	20220106	0.06489597497303	3/11/2022 22:00	4/11/2022 00:15	0.99255277740210	0.98389868530078
5	1	6	495717471775	31915963111	20220107	0.06438337345004	4/11/2022 00:15	4/11/2022 02:13	1.00456722435077	0.99663233024738
5	2	1	485570052987	28253496467	20220109	0.05818624170333	4/11/2022 04:11	4/11/2022 06:42	1.00938719122806	0.92581973408349
5	2	2	491109472341	31641914687	20220110	0.06442945304266	4/11/2022 06:42	4/11/2022 08:50	1.01140807452998	1.11992916430565
5	2	7	481054304242	30517276125	20220108	0.06343831840999	4/11/2022 02:13	4/11/2022 04:11	0.97042031324719	0.95617594301837

Tabla 1: Elaboración propia [2022], Fuente de datos: Trabajo de campo [2022]

Para poder garantizar la calidad de la salida sin depender de la entrada de datos, esta última representa un volumen mucho mayor de información, es necesario verificar las condiciones que originan las fallas en el experimento. Para el caso [5,1] experimento [1] según las estimaciones los datos se encuentran completos. Esta calidad se puede observar a través de las tendencias para esta observación.

En el caso de estudio 4 y el experimento 3 se muestra una incidencia en el job de este experimento. El problema se produce por un colapso en el sistema de Big Data. Producido entre las 0 horas 14 minutos y las 03 horas 25 minutos. Se mantuvo el proceso durante un tiempo 191.5 minutos. Al final se concluye lo siguiente: existía una saturación en la cola de procesamiento esto es verificable por medio del log o historial de sucesos (H10. log). Por su parte, el caso de estudio 2 y experimento 7, y el caso de estudio 3 experimento 1, han fallado por las siguientes razones: La cola (t) estaba saturada entre las 20 horas 46 minutos y las 01 horas 06 minutos; y las 01 horas 06 minutos y 05 horas 26 minutos.

(https://netty.io/wiki/reference-counted-objects.html)

Al agotarse el tiempo de espera el job es eliminado y provoca ausencia de información. Es decir, la aplicación ha sido finalizada. Se registraron problemas en el administrador de bloques.

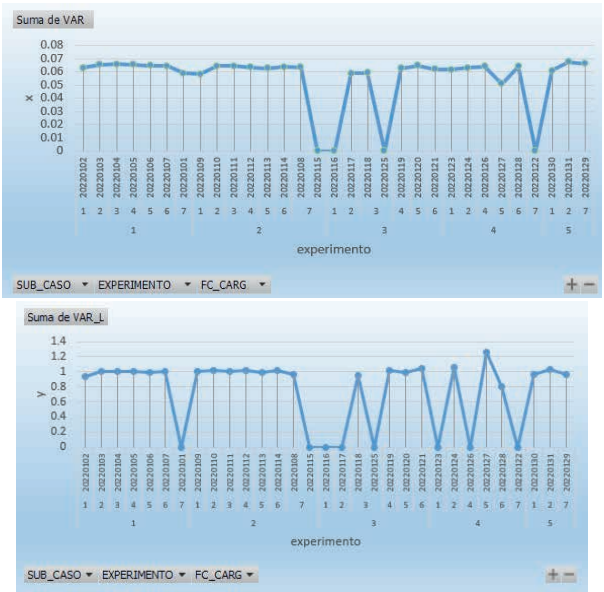

Gráfico 3: Elaboración propia [2022],

Fuente de datos: Trabajo de campo [2022]

La marca 20220122 que corresponde al experimento 07 del caso 04 tiene un error de operación el cual fue limpiado antes de iniciar el análisis. Quedando de la siguiente manera:

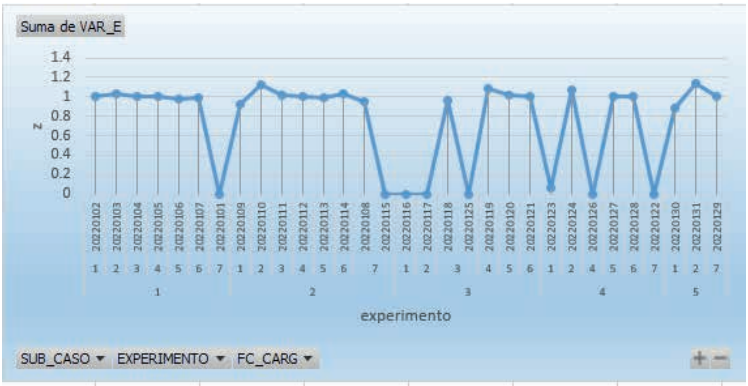

Gráfico 4: Elaboración propia [2022],

Fuente de datos: Trabajo de campo [2022]

Se pueden apreciar las bajas en las mediciones realizadas, estas originan los casos típicos de errores que se están estudiando y clasificando, para poder medir la calidad de datos.

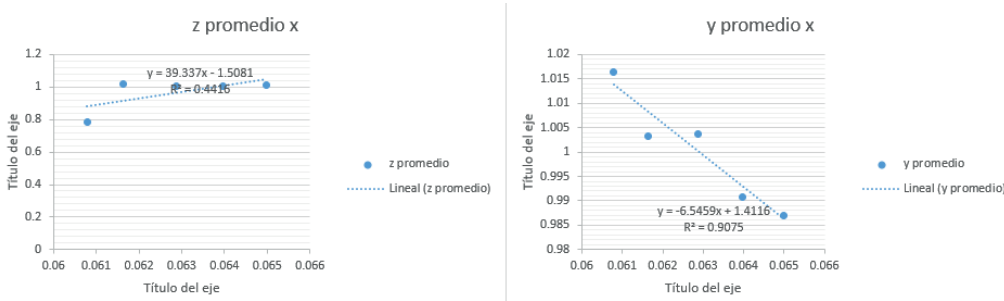

Gráfico 5: Elaboración propia [2022],

Fuente de datos: Trabajo de campo [2022]

En el gráfico del lado izquierdo de muestra un R cuadrado de **0.4416**, menor, en relación al R cuadrado del lado derecho **0.9075**, el primero está relacionado con las relaciones que existen entre la variación de escritura-lectura y las variaciones de lectura en relación a la marca anterior. El segundo está relacionado siempre con la variación escritura-lectura y la variación de escritura en relación a la marca anterior.

A continuación, se presenta un resumen de los hallazgos encontrados para los errores que fueron categorizados en cada variable (x, y, z) en cada caso (p, q).

	X	Y	Z
1	7	6	6
2	7	7	7
3	5	4	4
4	5	4	4
5	6	5	5
6	7	7	7
7	7	7	7
8	4	2	5
9	7	6	7
10	7	7	7
11	7	7	7
12	6	6	7
13	7	6	7
14	6	5	5
15	7	7	7
16	6	5	7
17	7	7	7
18	7	7	7
19	5	5	5

20	5	5	5
21	7	7	7
22	7	7	7
23	7	7	7
24	7	7	7
25	7	7	7
26	6	6	6

Tabla 2: Elaboración propia [2022].

Fuente de datos: Trabajo de campo [2022]

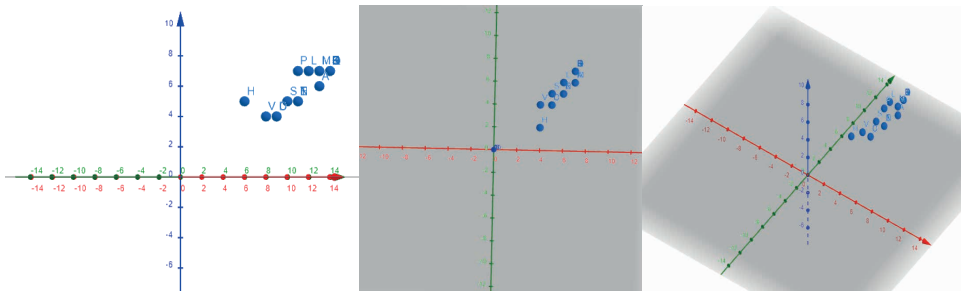

Gráfico 6: Elaboración propia [2022].

Fuente de datos: Trabajo de campo [2022] Rojo **x** Verde **y** Azul **z**

En cada caso de estudio donde el valor de la variable es 7 no se presenta ningún error, en los demás casos si lo hay, según la comprobación de la hipótesis y los resultados, se concluye que las variables si son representativas para poder determinar las tipologías de errores presentadas en este estudio para la calidad de la información.

Experimentos efectivos por cada caso de estudio.

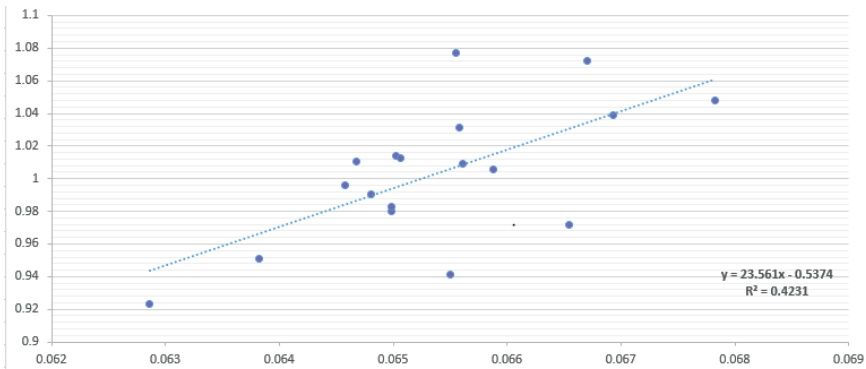

Gráfico 7: Elaboración propia [2022].

Fuente de datos: Trabajo de campo [2022]

En el gráfico anterior en eje x se coloca el valor de la variable [var]. Y en eje y la variable [varl] para este caso de estudio (p = {502}) donde los sub casos son q = {9,10,11} fechas comprendidas entre marzo 01 y marzo 18 (muestra parcial). Cuando se realiza en análisis completo se presenta el siguiente gráfico:

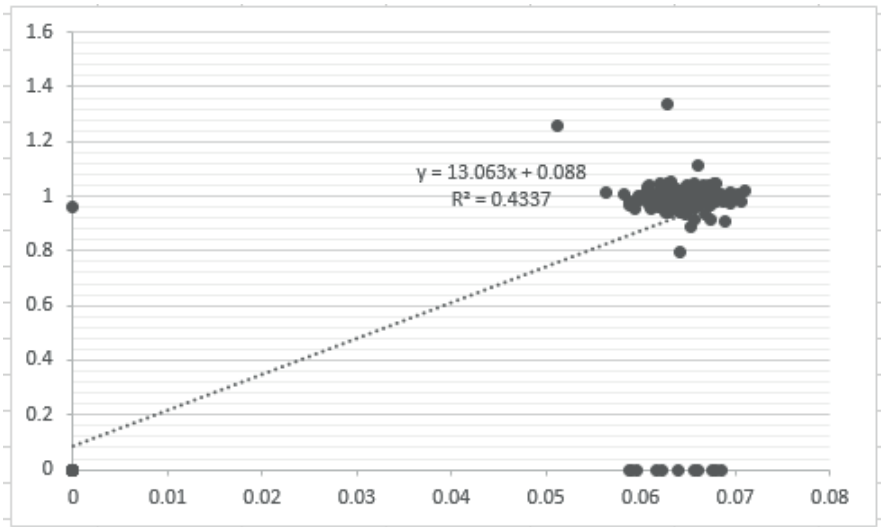

Gráfico 8: Elaboración propia [2022].

Fuente de datos: Trabajo de campo [2022]

Los puntos que se encuentran agrupados son los que tienen la tendencia a mantener la calidad de la información y los dispersos por el contrario representan errores típicos clasificados en el apartado metodológico, tanto de la variable var (x) y de la variable varl (y).

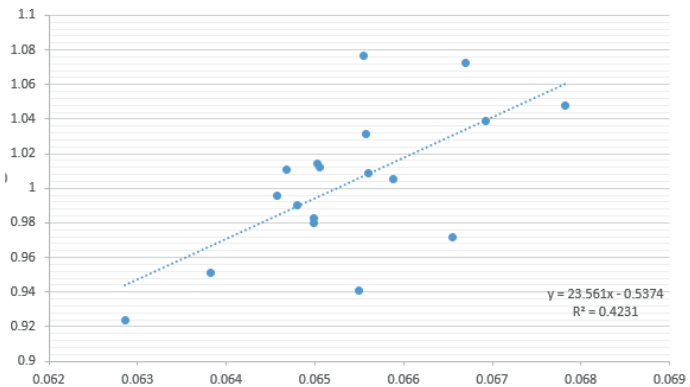

Gráfico 9: Elaboración propia [2022].

Fuente de datos: Trabajo de campo [2022]

El gráfico anterior muestra en el eje x la variable [var] y, en eje y [vare]. La ecuación de la recta esta denotada por y=23.561x-05374 con un R cuadrado de 0.4231. Tomando en cuenta el caso de estudio p = {502} y sub casos q= {9,10,11}, fechas del 01 de marzo al 18 de marzo 2022. Luego de realizar un análisis general se presenta el siguiente gráfico:

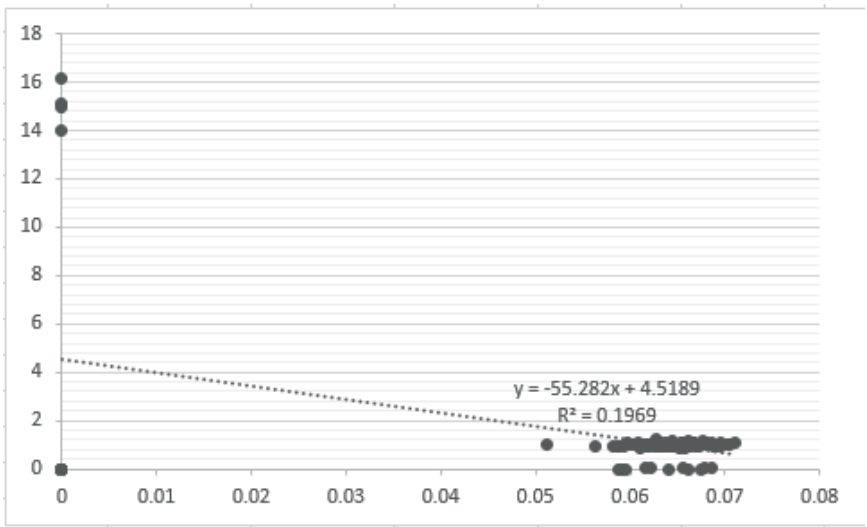

Gráfico 10: Elaboración propia [2022].

Fuente de datos: Trabajo de campo [2022]

Un coeficiente de correlación cercano al 0 indica que no existe correlación entre variables. Esto debido a que los valores de vare son mucho más atípicos de que los de varl, esto sucede de manera generalizada en todo el caso de estudio para este período de tiempo.

CONCLUSIÓN

Toda investigación se basa en hechos o acontecimiento los cuales han sido registrados o colocados en una fuente fiable, por ello, es importante verificar la calidad de estos. Por ende, en el presente estudio se ha clasificado calidad de la información por medio de una variable dicotómica, esta manifiesta una falla importante que desencadena una saturación en las colas del Yarn (Big Data), este administrador intenta asignar los recursos, pero no es posible porque ha llegado al límite. En consecuencia, la saturación, liberación de recursos forzados, eliminaciones de procesos, perdida es casi inevitable.

El aprendizaje de máquina y otros análisis basados en datos requieren de grandes volúmenes de información que sean confiables, por ello es necesario contar con herramientas de este tipo.

Se comprueba la hipótesis: Para este caso de estudio tenemos ochocientos noventa y un errores, los cuales, según el gráfico, un gran número corresponden al experimento cuarenta y siete.

Los experimentos que han fallado han sido el dos, diecisiete (en varias ocasiones), dieciocho, veintiuno, veinticinco, treinta y tres, cuarenta y uno, cuarenta y cuatro y cuarenta y siete. Algunos con valores muy cercanos al cero, y otros casi en el umbral normal. En este caso la variable presenta múltiples variaciones al momento de presentarse a lo largo de los periodos (semanas en los que se hizo el experimento).

REFERENCIAS

Black, P., & Wiliam, D. (2020). **Classroom assessment and pedagogy. Assessment in Education: Principles, Policy & Practice**, 25(6), 551–575. Recuperado de: https://doi.org/10.1080/0969594X.2018.1441807.

Da Silva (2021) **¿Qué es el big data y para qué sirve?** Zendesk, México. Recuperado de: https://www.zendesk.com.mx/blog/big-data-que-es/

Franke, B., Plante, J.-F., Roscher, R., Lee, E. A., Smyth, C., Hatefi, A., Chen, F., Gil, E., Schwing, A., Selvitella, A., Hoffman, M. M., Grosse, R., Hendricks, D., & Reid, N. (2016). **Statistical Inference, Learning and Models in Big Data. International Statistical Review**, 84(3), 371–389. https://doi.org/10.1111/insr.12176

Galvão J.; Ribeiro, D.; Machado, I.; Ferreira, F.; Gonçalves, J.; Faria, R.; Moreira, G.; Costa, C.; Cortez, P.; Santos, M. (2022) **Bosch's Industry 4.0 Advanced Data Analytics: Historical and Predictive Data Integration for Decision Support.**

Holland, A. (2020). **Effective principles of informal online learning design: A theory-building metasynthesis of qualitative research**. Computers & Education, 128, 214–226. Recuperado de: https://doi.org/10.1016/j.compedu.2018.09.026.

Kumar, S.; Prasad, D.; Kumar, J.; Sagar, K.; and Ashish Kr. (2021). **An Ensemble-Based Scalable Approach for Intrusion Detection Using Big Data Framework**. Big Data. Aug 2021.303-321. Volume: 9 Issue 4: August 16, 2021 Recuperado de: https://doi.org/10.1089/big.2020.0201

Mohamed, A.; Nahafabadi, M.; Wah, Y.; Zaman, E; Maskat, R. (2019). **The state of the art and taxonomy of big data analytics: View from the new big data framework.**

Oi, M., Yamada, M., Okubo, F., Shimada, A., & Ogata, H. (2020). **Reproducibility of findings from educational big data. In Paper presented at the proceedings of the Seventh International Learning Analytics & Knowledge Conference**, (pp. 536–537). New York: ACM. Recuperado de: https://doi.org/10.1145/3027385.3029445.

Qaffas A.; Hoque R.; Almazmomi N. (2021). **The Internet of Things and Big Data Analytics for Chronic Disease Monitoring in Saudi Arabia**

Sorensen, L. C. (2020). **"Big data" in educational administration: An application for predicting school dropout risk. Educational Administration Quarterly**, 45(1), 1–93. Recuperado de: https://doi.org/10.1177/0013161x18799439.

Tang, Y., Tang, Y., Peng, Y. et al.(2020). **Automated abnormality classification of chest radiographs using deep convolutional neural networks. npj Digit**. Med.3,70. https://doi.org/10.1038/s41746-020-0273-z.

Wang, L., y Wong, A. (2020). COVID-Net: **A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images**, 1-12.

van Evert, F. K., Fountas, S., Jakovetic, D., Crnojevic, V., Travlos, I., & Kempenaar, C. (2017). **Big Data for weed control and crop protection. Weed Research**, 57(4), 218–233. https://doi.org/10.1111/wre.12255

Zheng, M., & Bender, D. (2020). **Evaluating outcomes of computer-based classroom testing: Student acceptance and impact on learning and exam performance**. Medical Teacher, 41(1), 75–82. Recuperado de: https://doi.org/10.1080/0142159X.2018.1441984