

Características de las Bases de Datos para el Cómputo en la Nube: Un Estudio Secundario

Raúl A. Aguilar Vera¹, Naomi García Sánchez¹, Julio C. Díaz Mendoza¹, Brenda L. Flores Ríos²

averá@correo.uady.mx; A17016302@alumnos.uady.mx; julio.diaz@correo.uady.mx;
brenda.flores@uabc.edu.mx

¹ Facultad de Matemáticas, Universidad Autónoma de Yucatán, C.P. 97000, Mérida, Yucatán, México.

² Instituto de Ingeniería, Universidad Autónoma de Baja California, Mexicali, Baja California, México.

DOI: 10.17013/risti.49.115–130

Resumen: El crecimiento exponencial de las Bases de Datos a principios del presente siglo generó la necesidad de mecanismos innovadores para almacenamiento y gestión de información. Uno de estos paradigmas emergentes es el cómputo en la nube, un modelo que brinda acceso a un conjunto de recursos informáticos a través de la red de una manera sencilla. El presente estudio tiene como objetivo elaborar un estado del arte sobre aspectos vinculados con los esquemas de persistencia utilizados para el cómputo en la nube. Para ello se especificaron 5 preguntas de investigación que sirvieron para guiar una Revisión Sistemática de Literatura con la que se analizaron 72 Estudios Primarios. Entre los hallazgos se encontraron ocho técnicas específicas para el modelado de Bases de Datos, siendo la más citada el cifrado de datos; así mismo, se pudo identificar que Amazon es el proveedor corporativo más popular para dicho tipo de sistemas de gestión.

Palabras-clave: Cómputo en la Nube; Estudio Secundario; Modelado de Bases de Datos.

Characteristics of Databases for Cloud Computing: A Secondary Study

Abstract: The exponential growth of Databases at the beginning of this century generated the need for innovative mechanisms for information storage and management. One of these emerging paradigms is cloud computing, a model that provides access to a group of computing resources through the network in a simple way. The aim of this study is to elaborate a state of the art on aspects related to Databases used for cloud computing. For this, 5 research questions were specified that served to guide a systematic review of the literature with which 72 Primary Studies were analyzed; among the findings, eight specific techniques for Database modeling were found, the most cited being data encryption; likewise, it was possible to identify that Amazon is the most popular corporate provider for this type of management systems.

Keywords: Cloud Computing; Database Modeling; Secondary Study.

1. Introducción

A poco más de medio siglo de su concepción como disciplina ingenieril, la Ingeniería de Software (IS) ha acumulado un cuerpo de conocimientos relacionado con los procesos de desarrollo y gestión del software (Bourque & Fairley, 2014); sin embargo, el desarrollo acelerado y la evolución de los sistemas de información, ha dado pie en el nuevo milenio, a un conjunto de paradigmas emergentes (Aguilar, Oktaba y Juárez, 2019). Entre estos paradigmas, uno de singular importancia para la IS, es un modelo innovador conocido como Cómputo en la Nube (CN), el cual permite habilitar el acceso ubicuo bajo demanda a un grupo compartido de recursos informáticos configurables (Mell & Grance, 2011). La comprensión del constructo CN resulta importante para poder describir de forma apropiada la gestión de la información al interior de dicho ambiente; su éxito radica, en parte, por la disminución en tamaño para aquellas empresas que lo han incorporado, lo cual resulta un factor crítico en el beneficio económico de las mismas. El modelo del CN ha dado paso a la evolución de centros de datos que brindan a los clientes la infraestructura física requerida para alojar sus sistemas de información, incluidas las fuentes de alimentación redundantes, capacidades de comunicación de gran ancho de banda, monitoreo del entorno, así como servicios de seguridad (Zhao et al., 2014) permitiendo de esta manera la administración de sus procesos desde la nube (Peralta et al., 2016).

El presente estudio es una Revisión Sistemática de Literatura (RSL) con la cual se pretende sintetizar la investigación disponible sobre el paradigma del CN, en particular, sobre aspectos vinculados con las técnicas de modelado, tipos de datos, aspectos de calidad; así como proveedores y herramientas más comunes y/o populares para dichos sistemas.

El documento se encuentra organizado de la siguiente manera: la sección 2 presenta antecedentes y trabajos previos relacionados con el CN; la sección 3 describe la metodología utilizada en el presente estudio secundario. Las secciones 4 y 5 describen la planificación y ejecución del estudio, de acuerdo con la metodología descrita. En la sección 6 se describen los resultados del estudio de acuerdo con las Preguntas de Investigación planteadas. Finalmente, las conclusiones y trabajos futuros se presentan en la sección 7.

2. Antecedentes

Al inicio del presente siglo, los sistemas de información comenzaron a gestionar grandes cantidades de información para las aplicaciones de escritorio, con la expansión de los sistemas al incorporarse al internet, así como con la aparición de una gran diversidad de tipos de datos que empezaron a ser utilizados, la necesidad de mayor capacidad de persistencia en las Bases de Datos (BD) se incrementó de manera exponencial, por lo que el almacenamiento de la información pasó de medios locales, a medios gestionados mediante el CN. Para 2012 en internet ya existían 2.27 miles de millones de usuarios que requerían almacenamiento en algún dispositivo, y para el año 2020, de acuerdo con el informe de *Data Never Sleeps*, se generaban al día más de 2,500 millones de bytes, en particular, en plataformas como *Netflix*, *YouTube*, *Facebook* y otras aplicaciones de uso cotidiano (DOMO, 2021) en las que la variedad de tipos de datos ya no resultaba

compatible con los tradicionales esquemas de persistencia vinculados al modelo relacional.

El CN generalmente se rige por cinco características fundamentales: (1) ser un servicio automático por demanda, (2) ofrecer amplio acceso a la red, (3) disponer de un conjunto de recursos en común, (4) brindar elasticidad rápida, y (5) ser un servicio con medida. Con las características antes citadas, podemos decir que la era del CN inicia formalmente en 2006 con la aparición del CN elástico, un servicio de almacenamiento simple y novedoso introducido por la Web de Amazon (AWS). No obstante, independientemente de cuál haya sido la primera empresa en dar pauta al desarrollo de dicha tecnología, la principal razón por la que se destacó, es debido a que tanto el procesamiento de la información, como su almacenamiento, puede ser más eficiente si se cuenta con un ambiente más grande y accesible como lo es internet (Marinescu, 2018).

En el caso de la comunidad de Ciencia y la Tecnología, el paradigma del CN ha venido demostrando mayor interés de investigación; de hecho, en el contexto de la IS ha dado como resultado sistemas y aplicaciones orientadas a servicios que se ofrecen a través de internet, con un conjunto limitado de aspectos de calidad como son: la escalabilidad, portabilidad, así como la capacidad de integración con otras aplicaciones o servicios. Así pues, el CN puede ser descrito como un modelo que brinda acceso de red a un conjunto de recursos informáticos configurables, los cuales pueden ser accedidos por un usuario con esfuerzos mínimos y sin necesidad de terceros (Jain, Deshpande & Khanaa, 2017).

Diversos estudios secundarios (Bhatti & Rad, 2017; Karataş et al., 2017; Mostajabi, Safaei & Sahafi, 2021) han documentado que una de las problemáticas del CN gira en torno al desarrollo de las BD utilizadas por dicho modelo, en parte, debido al avance y crecimiento exponencial de los servicios que ofrecen, lo cual obliga a que sean diversos los elementos y aspectos de calidad que deben ser verificados.

3. Metodología de Investigación

El presente estudio utilizó como referencia la guía para realizar RSL propuesta por Kitchenham & Charters (2007). Las principales actividades de que constan las tres fases consideradas por la metodología son las siguientes:

1. Planificar la RSL:

- Identificar la necesidad de la revisión.
- Formular las preguntas de investigación.
- Definir la estrategia de búsqueda.
- Establecer los criterios para selección de los estudios primarios.
- Definir la estrategia de evaluación de la calidad de los estudios seleccionados.

2. Realizar la RSL

- Seleccionar los estudios primarios relevantes.
- Evaluar la calidad de los estudios primarios seleccionados.
- Extraer la información de los estudios primarios seleccionados.
- Sintetizar los hallazgos.

3. Reportar RSL

- Elaborar un informe técnico de la revisión.
- Evaluar el informe de la revisión.

Para incrementar el número los estudios relevantes, se incorporó la estrategia denominada bola de nieve hacia adelante (Wohlin, 2014) la cual sería aplicada al primer conjunto de Estudios Primarios Seleccionados (EPS).

4. Planificación del Estudio

4.1. Justificación de la necesidad del estudio

La evolución de manera acelerada de las Tecnologías de la Información y Comunicaciones dio origen, en el nuevo milenio, a un novedoso conjunto de paradigmas en la IS; en particular, el CN, el cual brinda acceso ubicuo bajo demanda a un conjunto compartido de recursos informáticos configurables, con los cuales es posible gestionar grandes volúmenes de información, de tipos de datos diversos. Su desarrollo en las últimas dos décadas ha sido gracias a numerosos estudios empíricos que se encuentran dispersos en diversas fuentes de información, para poder identificar nichos de oportunidad y continuar con el avance de la investigación en este innovador paradigma, es necesario realizar una selección, análisis e integración del conocimiento hasta ahora generado, por lo cual, se propone el presente estudio secundario, particularmente, mediante una RSL.

4.2. Formulación de las Preguntas de Investigación

Con base en los antecedentes del novedoso paradigma del CN, así como del análisis de algunos estudios secundarios previos (Bhatti & Rad, 2017; Karataş et al., 2017; Mostajabi, Safaei & Sahafi, 2021) los autores diseñaron un grupo de Preguntas de Investigación (PI), las cuales sirvieron para guiar la presente RSL:

PI-1 ¿Cuáles son las técnicas, procedimientos, o prácticas más utilizadas en el modelado de BD para el CN?

PI-2 ¿Cuáles son los principales modelos de datos que se usan en las BD para el CN?

PI-3 ¿Cuáles son los principales aspectos de calidad considerados por las arquitecturas y/o modelos propuestos para las BD utilizados para el CN?

PI-4 ¿Cuáles son los principales proveedores de herramientas para el CN?

PI-5 ¿Cuáles son las principales herramientas vinculadas con la gestión de BD para el CN?

4.3. Definición de la Estrategia de Búsqueda

Se realizó una búsqueda no sistemática de información sobre la temática del CN con el fin de identificar, tanto los conceptos clave vinculados con las PI, así como las fuentes donde suelen publicarse los estudios primarios vinculados con el tema.

Con base en la exploración del tema, se seleccionaron como fuentes de información: *IEEE Xplore, ACM Digital Library, ScienceDirect y SpringerLink*; así mismo, se decidió

agregar un repositorio —*Google Scholar*— que pudiera contener estudios relevantes de otras fuentes no consideradas. Así mismo, con base en las PI acordadas, se seleccionó un conjunto limitado de palabras clave que pudieran servir para identificar estudios primarios pertinentes con la temática del estudio: *cloud database, design, modeling, architecture*. Los conceptos considerados, junto con un conjunto de operadores lógicos, sirvieron para integrar una cadena de búsqueda genérica la cual será configurada en la fase de ejecución, en función de las características particulares de cada una de las 5 fuentes seleccionadas:

“Cloud database AND (modeling OR design)”

4.4. Criterios de Inclusión y Exclusión

Para discernir entre los estudios vinculados con la temática y los que realmente proporcionan información relevante, de acuerdo con las PI, se establecieron un conjunto de criterios de inclusión y de exclusión. Los criterios definidos para la selección de los estudios primarios son:

Criterios de Inclusión (CI):

- Artículos publicados en revistas y memorias de congresos especializados.
- Estudios empíricos relacionados con el diseño de BD para el CN.
- Publicaciones en idioma inglés.
- Estudios en la línea de tiempo de 2010 a 2021.

Criterios de Exclusión (CE):

- Estudios duplicados (manteniendo los de la fuente original).

4.5. Técnica de Bola de Nieve hacia adelante

Para incrementar y mejorar el conjunto de estudios pertinentes para la RSL, se decidió por el grupo de investigadores, utilizar la técnica denominada bola de nieve hacia adelante (Wohlin, 2014). Para la selección de un segundo conjunto de estudios con base en la aplicación de la técnica antes citada, se seleccionó *Google Scholar*, el cual es un motor de búsqueda de contenido y bibliografía científico-académica provista por *Google*, el cual, aparte de ser un repositorio, permite identificar, con cierta vigencia, los artículos que han citado a una publicación.

4.6. Estrategia de Evaluación de la Calidad de los EPS

El proceso de evaluación de la calidad en una RSL sirve para garantizar la validez y confiabilidad de los hallazgos descritos en los EPS, ponderando la importancia de cada uno de los n estudios (Kitchenham, Budgen & Brereton, 2015). Sin embargo, debido a la diversidad de metodologías utilizadas en los estudios primarios, no es posible disponer de un instrumento estandarizado, por lo que, en el caso del estudio, se procedió a adaptar —en un total de 10 ítems— el instrumento propuesto por Dyba & Dingsøyr (2008). Dicho instrumento considera los criterios: (1) Rigor, (2) Calidad del informe, (3) Relevancia y (4) Credibilidad. En función del grado de cumplimiento de cada uno de los criterios antes citados, el investigador asignará un valor cuantitativo —en una escala ordinal— con los siguientes valores:

- 1.0 – Para indicar que el artículo cumple con el mecanismo en su totalidad.
- 0.5 - Para indicar que el artículo cumple con el mecanismo en cierto grado.
- 0.0 – Para indicar que el artículo no cumple con el mecanismo en lo absoluto.

Luego de la evaluación del conjunto de EPS, cada estudio poseerá una valoración numérica en el intervalo de 0 - 10 puntos bajo una escala de tipo ordinal, con alguno de los valores cualitativos de calidad: Baja, Media y Alta. Para los autores, la valoración de cada uno de los EPS será utilizada para ponderar la importancia de los estudios individuales al momento de realizar, tanto el análisis, como la síntesis/agregación de los estudios seleccionados.

5. Ejecución del Estudio

5.1. Selección de los estudios primarios

Tomando como base la cadena genérica, se configuró en cada una de las BD, así como en el repositorio seleccionado. Las cadenas generadas se listan en la Tabla 1.

Fuente	Cadena Configurada
IEEE Xplore	cloud database AND (design OR architecture OR modeling)
ACM-DL	“cloud database” AND (“design” OR “architecture” OR “modeling”)
ScienceDirect	“cloud database AND (cloud computing OR modeling OR design OR architecture)”
SpringerLink	“cloud AND database AND (modeling OR architecture, OR design)”
Google Scholar	“cloud database AND (modeling OR design)” “cloud database AND (modeling OR architecture)”

Tabla 1 – Cadenas generadas para las 5 fuentes de información

Con las Cadenas Configuradas (CC) se procedió al proceso de búsqueda en cada una de las cinco fuentes de información, posteriormente se procedió a la aplicación de los CI, y finalmente, en un tercer momento, fueron aplicados los CE obteniendo al final un primer conjunto de 51 EPS. Dicho proceso se realizó entre los meses de noviembre y diciembre de 2021. En la Tabla 2 se puede observar la cantidad de estudios seleccionados por fuente, en cada una de las tres etapas del proceso.

Fuente	E1-CC	E2-CI	E3-CE
IEEE Xplore	6, 020	9	7
ACM-DL	423	15	15
ScienceDirect	914	17	16
SpringerLink	33	10	9
Google Scholar	14, 600	10	4

Tabla 2 – Número de estudios por fuente en cada etapa del proceso

Luego de realizar las tres etapas de la estrategia búsqueda, se utilizó la técnica de bola de nieve hacia adelante sobre el primer conjunto de 51 EPS, dicha actividad se realizó en las últimas dos semanas del mes de enero de 2022; con dicho proceso se incorporaron 21 nuevos estudios, teniendo al final del proceso, un total de 72 EPS para analizar.

5.2. Evaluación de la calidad de los EPS

La evaluación de la calidad para cada uno de los 72 EPS con el instrumento adaptado de Dyba & Dingsøyr (2008) permitió clasificar a: (1) 43 estudios con una valoración > 8 pts., (2) 25 estudios una valoración menor a 8 y mayor a 5, y (3) 4 estudios con una valoración igual o menor a 5 pts. La distribución cualitativa —de acuerdo con su calidad— de los 72 EPS se ilustra en la Figura 1.

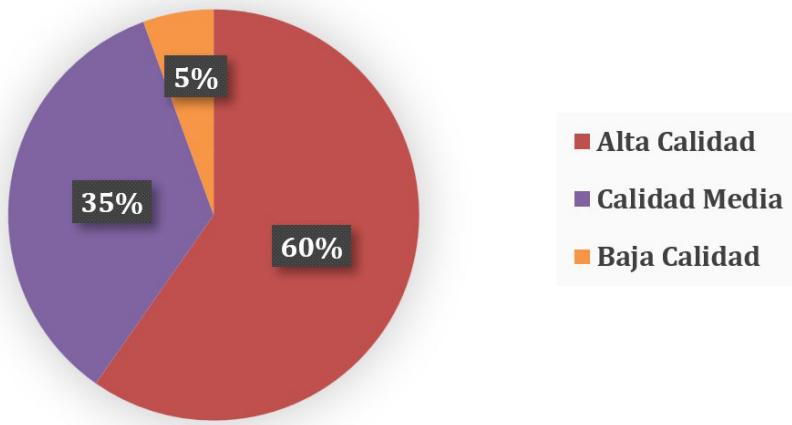


Figura 1 – Porcentaje de EPS de acuerdo con su calidad

6. Resultados del Estudio

La Figura 2 ilustra la distribución de los EPS en la ventana de tiempo considerada en el estudio, se puede identificar que no existe una tendencia clara en investigación sobre la temática del CN; no obstante, las aportaciones al cuerpo de conocimientos han sido constantes, y en particular, de 2019 a 2021 se ha tenido en promedio 10 publicaciones por año, lo cual es una señal de la vigencia en el interés de mantener la investigación sobre la temática que venimos analizando.

Luego de realizar las tres etapas de la estrategia búsqueda, se utilizó la técnica de bola de nieve hacia adelante sobre el primer conjunto de 51 EPS, dicha actividad se realizó en las últimas dos semanas del mes de enero de 2022; con dicho proceso se incorporaron 21 nuevos estudios, teniendo al final del proceso, un total de 72 EPS para analizar.

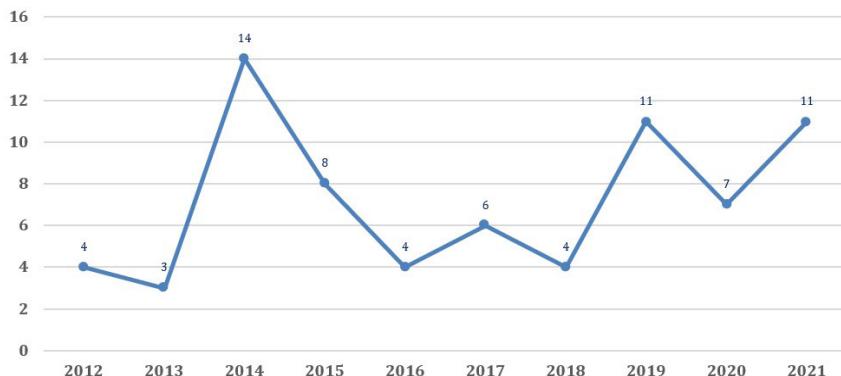


Figura 2 – Frecuencia de los EPS por año de publicación

Por otro lado, también resulta de interés para los investigadores conocer el tipo de investigación empírica realizada, en la Figura 3 se puede observar que el Estudio de Caso es la técnica más utilizada; así mismo, en función del propósito de la investigación, el CN está siendo estudiado empíricamente en todos los niveles: exploratoria, descriptiva, causa-efecto y de mejora (Robson, 2011).

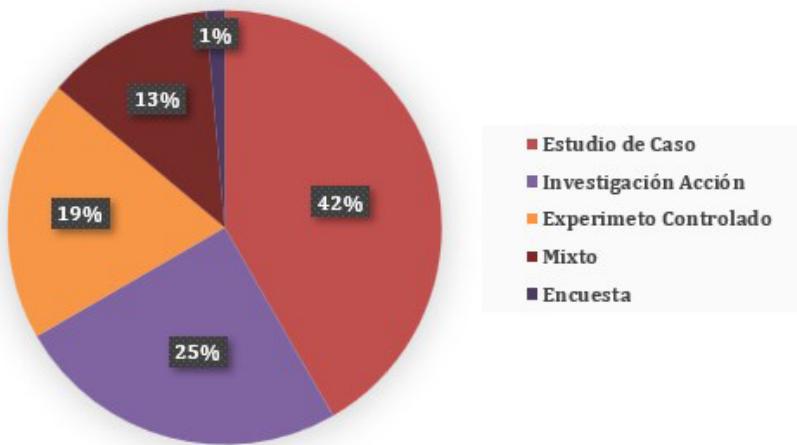


Figura 3 – Porcentaje de EPS por método empírico

Los hallazgos en función de las 5 PI se describen a continuación:

PI-1 ¿Cuáles son las técnicas, procedimientos, o prácticas más utilizadas en el modelado de BD para el CN?

Con la revisión a los 72 EPS podemos comentar que existen diversas técnicas, procedimientos y prácticas utilizadas para proveer de calidad las BD para el CN, siendo las más citadas, las descritas en la Tabla 3.

Fuente	Descripción/Referencias
Cifrado de datos	Consiste en cifrar los conjuntos de datos mediante algoritmos de encriptación; mediante este proceso los datos se codifican en texto cifrado ilegible para ayudar a mantenerlos seguros. Normalmente debe apoyarse en otros procedimientos (p.e. intercambio secreto). (Ferretti et al., 2014B), (Awadallah & Samsudin, 2021), (Ferretti, Colajanni, & Marchetti, 2014B), (Wong et al., 2014), (Wu et al., 2021), (Kavin & Ganapathy, 2019).
Fragmentación	La Fragmentación de un elemento E, consiste en dividir sus atributos en diferentes fragmentos E_1, \dots, E_n , de tal manera que solo los atributos en el mismo fragmento son visibles en asociación, permitiendo de esta manera optimizar la eficiencia y la seguridad de las BD. (Branco et al., 2017), (Hwang & Fu, 2016)
Balanceo de cargas	Es un mecanismo mediante el cual se distribuye la carga de trabajo dinámica de manera uniforme entre todos los nodos. El balanceo o equilibrio de carga en la nube también se conoce como equilibrio de carga como servicio (LBaaS). (TengJiao et al., 2012), (Zhang et al., 2021)
Mecanismo de Reversión	El mecanismo de reversión es aplicado cuando se detectan anomalías consecutivas en los esquemas de persistencia o cuando el tiempo de respuesta se ralentiza luego de aceptar un determinado número de consultas; en ambos casos, las últimas transacciones son eliminadas y el esquema de persistencia regresa a un estado previo. (Tan et al., 2019), (Ferretti, Colajanni, & Marchetti, 2012).
Migración y replicación	Cuando las frecuencias de acceso de los elementos de datos conducen a una carga de trabajo desequilibrada en los nodos del sistema, la migración y replicación de arrendatarios/datos para distribuir la carga de trabajo a un conjunto flexible de sitios, es recomendable, lo anterior, con el fin de evitar violaciones a los acuerdos de servicio. (Abdel, Abo-Alian, & Badr, 2021), (TengJiao et al., 2012), (Xiao et al., 2021)
Algoritmos	El desarrollo y modificación de algoritmos específicos es una práctica común para el desarrollo de técnicas y procedimientos para crear u optimizar modelos que proveen aspectos vinculados con la calidad del esquema de persistencia de datos para el CN. (Mulani et al., 2015), (Kavin & Ganapathy, 2019), (Kayed & Omar, 2019), (Alomari & Noaman, 2019), (Tan et al., 2019), (Depoutovitch et al., 2020), (Vainshtein & Gudes, 2021), (Abdel, Abo-Alian, & Badr, 2021), (Antonopoulos et al., 2019), (Cao et al., 2018), (Kim & Song, 2014), (Dokeroglu, Bayir, & Cosar, 2015)
Desagregación de recursos	La desagregación de recursos se ejecuta en los centros de datos desagregados, en éstos los recursos están ubicados en diferentes nodos conectados a través de una red de alta velocidad, lo cual permite que cada tipo de recurso mejore su tasa de utilización y de recuperación a fallos, expanda su volumen y se actualice de forma independiente. (Cao et al., 2021)
DevOps	El propósito de utilizar DevOps para el CN consiste en facilitar el uso de metodologías de desarrollo propias de la IS. Las plataformas Cloud DevOps actualmente permiten a los equipos de desarrollo aprovechar los procesos DevOps para la integración de las herramientas necesarias. La comunidad de BD ha aceptado esta nueva dinámica de trabajo y se encuentra en proceso de adaptación. (Airaj, 2016)

Tabla 3 – Técnicas/Procedimientos más citadas en los EPS

PI-2 ¿Cuáles son los principales modelos de datos que se usan en las BD para el CN?

Con la revisión de los EPS se encontraron únicamente dos estudios en los que explícitamente se describen modelos de datos particulares: (1) *HashMap* y (2) Un Modelo Basado en Columnas; no obstante, aunque los demás estudios no describieron los modelos de datos que utilizaron, en dichos EPS se habla de estructuras de datos primitivas de manera indirecta o simplemente se omiten, los autores del presente estudio consideran, que posiblemente en dichos estudios se da por hecho que las personas interesadas en el área de BD conocen el estándar o están familiarizadas con los modelos que comúnmente son usados. Cabe destacar que, de la revisión de los estudios previos, se pudo identificar que los autores hacen referencia a modelos de datos conocidos y usados de manera cotidiana, tales como los arreglos y apuntadores, lo cual sustenta la reflexión a la que se llegó en nuestro análisis.

El primer modelo citado es el *HashMap* (Obiniyi, Dzer, & Abdullahi, 2015), dicho modelo se deriva de las *hashtable* y se forma mediante la transformación de un rango de valores clave a un rango de valores de índice de matriz empleando la función que lleva su nombre (*hash function*). Un modelo *HashMap* tiene dos enfoques para su uso: cerrado o abierto; en el caso de estudio primario analizado, los autores utilizan el enfoque abierto, ya que indican, asegura un rendimiento óptimo. Esta estructura de datos es preferida porque el aleatorizar el orden de los datos, provee un acceso eficiente a la información con base en las llaves correspondientes del usuario. El segundo modelo es el descrito en (Wong et al., 2014); los autores mencionan el uso de un modelo de datos basado en columnas, el cual brinda suficiente flexibilidad como para manipular dichas columnas y aplicar las operaciones necesarias en los datos durante las peticiones del usuario, sin dejar de lado la seguridad.

PI-3 ¿Cuáles son los principales aspectos de calidad considerados por las arquitecturas y/o modelos propuestos para las BD utilizadas para el CN?

Derivado del análisis de los EPS se pudo observar que los estudios no se enfocan en de manera particular en aspectos de calidad específicos, sino más bien en las necesidades que tienen las aplicaciones, servicios o las aplicaciones de software en la nube en general.

Los autores comúnmente recomiendan distribuir la arquitectura de la nube por capas (Alam et al., 2013; Wong et al., 2014) de manera que cada una conlleve responsabilidades específicas y provea garantías para aquellos aspectos clave del CN, como seguridad, consistencia, privacidad, garantías ACID —Atomicidad, Consistencia, Aislamiento y Durabilidad— y garantizando el control de acceso (John & Ramesh, 2017; Zhang et al., 2019; Chitra & Rani, 2014); sin embargo, también se reconoce que las arquitecturas pueden contribuir a la mejora del rendimiento de las BD al apoyar en tareas como la configuración de ajustes o el *buffer* (Tan et al., 2019), el control de cargas (TengJiao et al., 2012), entre otros aspectos que también resultan vitales para el funcionamiento óptimo del entorno.

Por otro lado, con el avance de la tecnología, los autores han tendido a enfocarse en arquitecturas lógicas con marcos de trabajo para mejorar/reforzar las características anteriormente mencionadas (Sahri, Moussa, Long, & Benbernou, 2014) empleando estrategias novedosas que incluyen el uso del aprendizaje máquina (Xiao et al., 2021) o adoptando estructuras basadas en sistemas distribuidos y BD en tiempo real en la

memoria (Stoja et al., 2014) por mencionar algunas innovaciones. También se identificó que los autores consideran que, por ejemplo, los modelos NoSQL poseen grandes ventajas sobre los SQL y ofrecen solución a múltiples problemáticas que surgen con las nuevas necesidades del software en la nube, aunque se habla también de integrar trabajos previos en el área de las BD en la nube para mejorar aquellos aspectos “flojos” en los modelos SQL, pues una gran mayoría del software en la nube emplea este modelo y es imposible que quede obsoleto, al menos en la actualidad.

Independientemente de los criterios/aspectos de calidad considerados, es necesario ofrecer toda la ayuda posible al ingeniero de software que diseña esquemas de persistencia para el CN, a fin de aumentar su productividad y contribuir a la generación de diseños pertinentes a dichos criterios (Blas, Leone y Gonnet, 2019).

PI-4 ¿Cuáles son los principales proveedores de herramientas para el CN?

Con la revisión y análisis a los EPS se pudo observar que, si bien los autores de los estudios presentan herramientas que contribuyen a la gestión de la información en el cómputo en la nube, la gran mayoría se enfoca principalmente en describir las herramientas en función de las necesidades de cada estudio particular, para facilitar tareas que se relacionan con el propósito de su investigación, o probar sus propuestas, por ejemplo, si se habla de la seguridad de un sistema, los autores presentan herramientas que permiten probar la privacidad de los datos al ejecutar peticiones sobre una base de datos cifrada, o herramientas por el estilo, pero siempre relacionadas con la temática y con el objetivo de facilitar las pruebas o las actividades a realizar. Un ejemplo de esto se da en Malhotra et al (2016), donde los autores emplean *CryptDB* como una herramienta para probar peticiones sobre datos cifrados.

En cuanto a los proveedores para el cómputo en la nube, la Figura 4 ilustra la distribución de las 44 menciones identificadas con el análisis de los EPS; cabe mencionar que la categoría denominada “Otros”, hace referencia a entidades como Apache y *Yahoo!*, que proveen herramientas “independientes” sin una entidad corporativa. Cabe destacar, que de entre los proveedores de entidades corporativas, el más citado fue Amazon con 12 menciones.

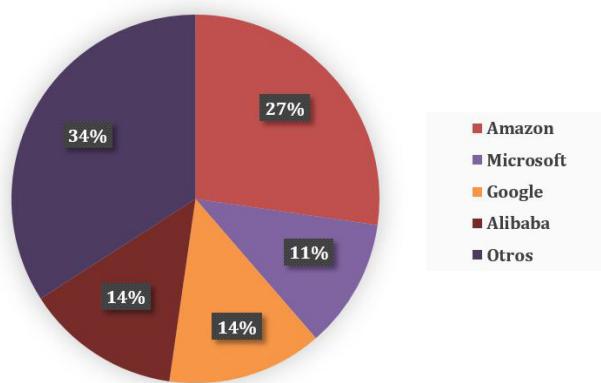


Figura 4 – Proveedores de herramientas para el CN

PI-5 ¿Cuáles son las principales herramientas vinculadas con la gestión de BD para el CN?

El análisis de los EPS permitió identificar que solamente en siete estudios los autores citan herramientas específicas (ver Tabla 4) que se emplean para la gestión de las BD en la nube, sin embargo, resulta importante resaltar que: (1) las herramientas de las que hablan son las más populares o, al menos, las que más se repiten en los estudios previos analizados y (2) los autores hacen hincapié en ellas porque, tal y como mencionan otros estudios, dichas herramientas tienen características clave como: buen rendimiento, alta disponibilidad, elasticidad, seguridad y consistencia.

Herramienta	Referencias
Microsoft Windows Azure, Amazon Elastic Compute Cloud.	(Stoja et al., 2014)
Big Table de Google, Dynamo de Amazon.	(Sakr, 2013)
Apache Cassandra.	(Silva-Muñoz, Franzin, & Bersini, 2021)
Alibaba DB System.	(Li, 2019)
Amazon AWS, Microsoft Azure, Google's Cloud Platform y Alibaba.	(Cao et al., 2021)
Alibaba Cloud RDS.	(Xiao et al., 2021)
Big Table de Google, Apache Cassandra, y Dynamo de Amazon.	(Cheng et al., 2014)
Microsoft Windows Azure, Amazon Elastic Compute Cloud.	(Stoja et al., 2014)

Tabla 4 – Herramientas para gestión de BD en la nube citadas en los EPS

Si bien estas características pueden no estar 100% aseguradas en todas las herramientas, se pueden seguir procedimientos o realizar mejoras para asegurar que el funcionamiento y la integridad de los datos no sean comprometidos.

7. Conclusiones y Trabajos Futuros

De acuerdo con el conjunto de 72 EPS analizados, es posible concluir que los aspectos vinculados con los esquemas de persistencia para el cómputo en la nube representan un paradigma aún vigente en el contexto de la investigación en IS.

En cuanto a las técnicas y/o procedimientos para el modelado de BD en la nube, el análisis a los EPS permitió identificar como las más citadas: Cifrado de Datos, Fragmentación, Balanceo de Cargas, Mecanismos de Reversión, Técnicas que hacen uso de Algoritmos, Procedimientos relacionados con la Migración y Replicación, Desagregación de Recursos, así como DevOps. Respecto de los modelos de datos utilizados, al parecer siguen siendo utilizados estructuras de datos primitivas y solo se pudo identificar dos modelos particulares: *HashMap* y un esquema basado en columnas. En relación con los aspectos de calidad, se pudo identificar que si bien los estudios no se enfocan de manera

particular en aspectos de calidad específicos, a través de las necesidades descritas para las aplicaciones y/o servicios, es posible identificar como aspectos de calidad que deben estar presentes en las BD para el cómputo en la nube: la disponibilidad, escalabilidad, elasticidad, rendimiento y seguridad, aunque aunado a éstos, es relevante tomar en cuenta otros elementos cuya influencia en la nube es alta, como por ejemplo los principios ACID que están estrechamente relacionados con dichas BD y con los acuerdos de nivel servicio, los cuales se relacionan con el funcionamiento de los sistemas de gestión de bases de datos. Por su parte, en cuanto a los proveedores de servicios/herramientas para el cómputo en la nube, si bien se identificaron diversas herramientas que están siendo utilizadas para actividades de investigación, buena parte de éstas son herramientas “independientes” sin una entidad corporativa, siendo Amazon, Microsoft, Google, y Alibaba, los 4 proveedores corporativos más citados o conocidos (populares). Así mismo, en cuanto a herramientas específicas para la gestión de esquemas de persistencia para el cómputo en la nube, solo en 7 estudios analizados fueron citadas de manera particular un limitado conjunto de herramientas, varias de éstas provistas por los proveedores corporativos denominados populares.

Finalmente, con la RSL se pudo identificar que el estudio del CN sigue en constante evolución, y en lo que compete a la IS, resulta necesario realizar mayor investigación empírica para comparar bondades y condiciones de uso de las diversas técnicas y procedimientos utilizados en el modelado de los esquemas de persistencia para aplicaciones en la nube, así como profundizar en propuestas de mejora —con base en aspectos de calidad— de las herramientas para la gestión de BD utilizadas en los entornos del CN.

Referencias

- Abdel R., Abo-Alian, A., & Badr, B. (2021). A Predictive Multi-Tenant Database Migration and Replication in the Cloud Environment. *IEEE Access*.
- Aguilar, R., Oktaba, H. & Juárez R. (2019). Introducción a la Ingeniería de Software. En Aguilar, R. (ed.) Ingeniería de Software en México: Educación, Industria e Investigación. Academia Mexicana de Computación, (pp. 11-14).
- Airaj, M. (2016). Enable cloud DevOps approach for industry and higher education.
- Alam, B., Doja, M. N., Alam, M., & Mongia, S. (2013). 5-Layered Architecture of Cloud Database Management System. *AASRI Procedia*, (pp. 194-199).
- Alomari, E., & Noaman, A. (2019). SeCloudDB: A Unified API for Secure SQL and NoSQL Cloud Databases. *Proceedings of the 2019 3rd International Conference on Cloud and Big Data Computing - ICCBDC*.
- Antonopoulos, P., Byrne, P., Chen, W., Diaconu, C., Kodandaramaih, R. T., Kodavalla, H. & Venkataramappa, G. M. (2019). Constant Time Recovery in Azure SQL Database. *Proceedings VLDB Endow*, (pp. 2143-2154).
- Awadallah, R., & Samsudin, A. (2021). Using Blockchain in Cloud Computing to Enhance Relational Database Security. *IEEE Access*..

- Bhatti, H. J., & Rad, B. B. (2017). Databases in Cloud Computing: A Literature Review. *Information Technology and Computer Science*, 9(4), pp. 9-17.
- Blas, M., Leone, H. & Gonnet, S. (2019) Modelado y Verificación de Patrones de Diseño de Arquitectura de Software para Entornos de Computación en la Nube. *RISTI – Revista Ibérica de Sistemas e Tecnologias de Informação*, (35).
- Bourque, P. & Fairley, R. (2014). Guide to the Software Engineering Body of Knowledge (SWEBOK V3.0). IEEE Computer Society.
- Branco Jr, E. C., Monteiro, J. M., Reis, R., & Machado, J. C. (2017). A New Mechanism to Preserving Data Confidentiality in Cloud Database Scenarios. *Lecture Notes in Business Information Processing*, (pp. 261-283).
- Cao, W., Zhang, Y., Yang, X., Li, F., Wang, S., Hu, Q. & Gao, Y. (2021). PolarDB Serverless: A Cloud Native Database for Disaggregated Data Centers. *Proceedings of the International Conference on Management of Data*.
- Chitra, K., & Rani, J. B. (2014). DES: Dynamic and Elastic Scalability in Cloud Computing Database Architecture. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 5(1), pp.173-175.
- Cheng, X., Li, H., Wang, Y., Wang, T., & Yang, D. (2014). BF-Matrix: A Secondary Index for the Cloud Storage. *WAIM 2014. Lecture Notes in Computer Science*, (pp. 384-396).
- Depoutovitch, A., Chen, C., Chen, J., Larson, P., Lin, S., Ng, J. & He, Y. (2020). Taurus Database: How to be Fast, Available, and Frugal in the Cloud. *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, (pp. 1463-1478).
- Dokeroglu, T., Bayir, M. A., & Cosar, A. (2015). Robust heuristic algorithms for exploiting the common tasks of relational cloud database queries. *Applied Soft Computing*, (pp. 72-82).
- DOMO. (2021). Domo Resource. Recuperado el 15 de Octubre de 2021, de <https://www.domo.com/learn/infographic/data-never-sleeps-8>
- Dyba T. & Dingsøyr, T. (2008). Strength of evidence in systematic reviews in software engineering. *Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement*.
- Ferretti, L., Colajanni, M., & Marchetti, M. (2012). Supporting Security and Consistency for Cloud Database. *Lecture Notes in Computer Science*.
- Ferretti, L., Pierazzi, F., Colajanni, M., & Marchetti, M. (2014a). Performance and cost evaluation of an adaptive encryption architecture for cloud databases.
- Ferretti, L., Pierazzi, F., Colajanni, M., & Marchetti, M. (2014b). Scalable architecture for multi-user encrypted SQL operations on cloud database services. *IEEE Transactions on Cloud Computing*, (pp. 448-458).
- Hwang, G.-H., & Fu, S.-K. (2016). Proof of Violation for Trust and Accountability of Cloud Database Systems. *16th IEEE/ACM IS on CCGrid*, (pp. 425-433)

- Jain, N., Deshpande, A. V., & Khanaa, V. (2017). The Effectiveness Of Cloud Databases. *International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies*.
- John P, A. J., & Ramesh K, B. S. (2017). Proposed Architecture for Improving Security and Consistency of Data Transactions in Cloud Database using Tree-Based Consistency Approach. *IJCSIS*, 15(12), 118-126.
- Karataş, G., C. F., Doğan, G., Konca, C., & Akbulut, A. (2017). Multi-tenant architectures in the cloud: A systematic mapping study. *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*.
- Kavin, B. P., & Ganapathy, S. (2019). A secured storage and privacy-preserving model using CRT for providing security on cloud and IoT-based applications. *Computer Networks*, (pp. 181-190).
- Kayed, A., & Omar, S. (2019). Periodical key change for cloud mutable security protocol.
- Kim, T., & Song, S. (2014). Dynamic partition lock method to reduce transaction abort rates of cloud database. *Cluster Computing*, (pp. 233-242).
- Kitchenham, B. A., Budgen, D., & Brereton, P. (2015). *Assessing Study Quality*. In Evidence-Based Software Engineering And Systematic Reviews, (pp. 79-88). Florida: CRC Press.
- Kitchenham, B. & Charters, S. (2007). Guidelines for Performing Systematic Literature Reviews in Software Engineering. *Technical Report EBSE-2007-01, School of Computer Science and Mathematics*. Keele University.
- Li, F. (2019). Cloud-Native Database Systems at Alibaba: Opportunities and Challenges. *Proceedings VLDB Endow*, (pp. 2263-2272).
- Malhotra, S., Doja, M. N., Alam, B., & Alam, M. (2016). Cloud Database Management System Security Challenges and Solutions: an analysis. *CSI Transactions on ICT*, (pp. 199-207).
- Marinescu, D. (2018). Cloud Computing: Theory and Practice. MK Publishers.
- Mell, P. & Grance, T. (2011). The NIST definition of cloud computing.
- Mostajabi, F., Safaei, A. A., & Sahafi, A. (2021). A Systematic Review of Data Models for the Big Data Problem. IEEE Access.
- Mulani, N., Pawar, A., Mulay, P., & Dani, A. (2015). Variant of COBWEB Clustering for Privacy Preservation in Cloud DB Querying. *Procedia Computer Science*.
- Obiniyi, A. A., Dzer, R. M., & Abdullahi, S. (2015). Balancing Query Performance and Security on Relational Cloud Database: An Architecture. *IJCA*, 34-39.
- Peralta, M., Salgado, C., Montejano, G. & Riesco D. (2016). Las Redes Sociales y la Nube: un nuevo Paradigma para los Procesos de Negocio. *RISTI- Revista Ibérica de Sistemas e Tecnologia de Informação*, (8).
- Robson, C. (2011). *Real World Research: A Resource for Social Scientists and Practitioner-Researchers*, 3a Ed, Wiley.

- Sahri, S., Moussa, R., Long, D. D., & Benbernou, S. (2014). DBaaS-Expert: A Recommender for the Selection of the Right Cloud Database. *Foundations of Intelligent Systems*, (pp. 315-324).
- Sakr, S. (2013). Cloud-hosted databases: technologies, challenges, and opportunities. *Cluster Computing*, (pp. 487-502).
- Silva-Muñoz, M., Franzin, A., & Bersini, H. (2021). Automatic configuration of the Cassandra database using irace. *Peerj Computer Science*.
- Stoja, S., Jelačić, B., Vukmirović, S., Čapko, D., & Dalčeković, N. (2014). Architecture of Real-Time Database in Cloud Environment for Distributed Systems. *2nd International Conference on AI, Modelling and Simulation*,
- Tan, J., Zhang, T., Li, F., Chen, J., Zheng, Q., Zhang, P. & Zhang, R. (2019). iBTune: Individualized Buffer Tuning for Large-scale Cloud Database. *VLDB Endo*.
- TengJiao, W., ZiYu, L., BiShan, Y., Jun, G., Allen, H., DongQing, Y., & JinZhong, N. (2012). MBA: A market-based approach to data allocation and dynamic migration for cloud database. *Science China Information Sciences*.
- Vainshtein, Y., & Gudes, E. (2021). Use of Blockchain for Ensuring Data Integrity in Cloud Databases. *Lecture Notes in Computer Science*, (pp. 325-335).
- Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. *Proceedings of the 18th ICE & A in SE*.
- Wong, W. K., Kao, B., Lok, C. D. W., Li, R., & Yiu, S. M. (2014). Secure Query Processing with Data Interoperability in a Cloud Database Environment. *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data - SIGMOD '14*, (pp. 1395-1406).
- Wu, W., Xian, M., Parampalli, U., & Lu, B. (2021). Efficient privacy-preserving frequent itemset query over semantically secure encrypted cloud database. *World Wide Web (2021)*, pp. 607-629.
- Xiao, W., Yang, C., Wang, J., Zhu, X., Bao, W., Feng, X. & Liu, L. (2021). YISHAN: Managing Large-scale Cloud Database Instances via Machine Learning. *IEEE Transactions on Services Computing*.
- Zhao, L., Sakr, S., Liu, A., & Bouguettaya, A. (2014). Administración de Datos en la Nube. New York: Springer.
- Zhang, X., Hong, W., Chang, Z., Jin, S., Tan, J., Li, F. & Cui, B. (2021). ResTune: Resource Oriented Tuning Boosted by Meta-Learning for Cloud Databases. *Proceedings of the 2021 International Conference on Management of Data*.
- Zhang, J., Liu, Y., Zhou, K., Li, G., Xiao, Z., Cheng, B. & Li, Z. (2019). An End-to-End Automatic Cloud Database Tuning System Using Deep Reinforcement Learning. *Proceedings of the 2019 International Conference on Management of Data - SIGMOD '19*, (pp. 415-432).