Performance Limitations of a Text Search Application Running in Cloud Instances

J. I. Zablah, I. X. Corrales, J. M. Aguilar, A. Garcia, F. Gomez and M. T. Medina

***Abstract*— This article analyzes the performance of MySQL in clouds based on commodity hardware in order to identify the bottlenecks in the execution of series of scripts developed on the SQL standard. The developed scripts were designed in order to perform text search in a considerable amount of records. Two types of platforms were employed: a physical machine that serves as host and an instance within a cloud infrastructure. The results show that the intensive use of a relational database presents a greater loss of performance in a cloud instance due limitations in the primary storage system that was employed in the cloud infrastructure.**

***Keywords*—Virtualization, regular expressions, benchmark, Cloud, MySQL.**

I. INTRODUCCIÓN

EN la actualidad, se han realizado varias evaluaciones del rendimiento de diversas aplicaciones científicas con el fin de estudiar la necesidad de una infraestructura computacional de altas prestaciones para el manejo y procesamiento de datos de la Universidad Nacional Autónoma de Honduras (UNAH) [1-3]. Para esto se ha considerado como solución el paradigma de la computación en la nube (cloud), debido al constante desarrollo en software y hardware [4].

La computación en la nube se define como una tecnología de sistemas distribuidos ensamblado sobre máquinas virtuales interconectadas [5], que proveen de forma dinámica recursos computaciones unificados conforme a un acuerdo de nivel de servicio (SLA – Service Level Agreement). Este paradigma computacional aporta e integra las ventajas que ofrece la arquitectura orientada a servicios (SOA) y la tecnologías de virtualización [6, 7]. Esta última es la base del modelo cloud y es conocido como infraestructura como servicio (IaaS), destacándose por una gestión más eficiente, flexible y escalable de los recursos de hardware. Adicionalmente, este tipo de esquemas aportan beneficio en la gestión de los recursos computaciones, ya que se pueden incrementar según la necesidad de procesamiento a través de una gestión centralizada que permite el uso de perfiles de recursos conforme las necesidades típicas [8, 9], asimismo este tipo de arquitecturas son idóneas para las necesidades de procesamiento crecientes que se consideran como BigData [18,19].

Son múltiples los trabajos realizados para conocer y cuantificar las ventajas de la computación en la nube en entornos científicos, académicos y administrativos [10-12]; sobre todo buscando compararla con las tecnologías existentes. Se vuelve necesario, debido a lo anterior, evaluar las soluciones a los requerimientos computacionales de la UNAH en un entorno cloud, ya que hasta el momento no hay una implementación sobre la misma. Por lo tanto, es necesario identificar las ventajas y desventajas entre el entorno tradicional frente a la nube.

Ya que existen requerimientos computacionales temporales y de ejecución planificada, el paradigma cloud provee la facilidad de reutilizar infraestructuras, ya que se puede contar con una gran capacidad computacional a un menor costo al ser comparado con soluciones dedicadas [13]. Este trabajo hace uso de una nube privada para proveer una infraestructura como servicio donde se ha ejecutado la aplicación evaluada. Se ha comparado esta ejecución frente a un ordenador con las mismas características del anfitrión físico de las instancias cloud y la instancia virtual.

Los scripts son ejecutados dentro del entorno gestor de bases de datos relacional, teniendo como especial característica que los datos usados se integran en un esquema a partir de un volcado o “dump”. Los scripts se han desarrollado como funciones y procedimientos almacenados, con la finalidad de procesar los datos, desde la normalización, clasificación, búsqueda de palabras y raíces semánticas previamente definidas. Se han utilizando expresiones regulares sobre cadenas de texto que se originan a partir de una dada. Lo anterior se lleva a cabo con la finalidad de conocer los tiempos de ejecución de cada plataforma y estimar de esa manera la conveniencia de cada una en este de tipo de aplicaciones.

Para los autores es evidente que el uso de discos duros con tecnología SSD – *Solid State Drive*, deberá arrojar mejores resultados en el acceso al sistema de ficheros frente a un sistema de almacenamiento de red, pero la ventaja del paradigma cloud en cuanto a flexibilidad y reusabilidad del hardware, representa un valor agregado en aplicaciones administrativas que requieren concurrencia; de forma que medir sus diferencias en el procesamiento entre ambas

J. I. Zablah, Universidad Nacional Autónoma de Honduras, Honduras, jose.zablah@unah.edu.hn
M. T. Medina, Universidad Nacional Autónoma de Honduras, Honduras, marco.medina@unah.edu.hn
I. X. Corrales, Universidad Nacional Autónoma de Honduras, Honduras, iris.corrales@unah.edu.hn
J. M. Aguilar, Universidad Nacional Autónoma de Honduras, Honduras, jaguilar@unah.edu.hn
A. García, Universidade Santiago de Compostela, España, antonio.garcia.loureiro@usc.es
F. Gomez, Universidade Santiago de Compostela, España, fernando.gomez.folgar@usc.es

plataformas (física y virtual) es el objetivo final de este artículo.

Este trabajo está organizado en varias secciones, siendo esta introducción la primera. La segunda describe los trabajos desarrollados que están relacionados a este. La tercera parte describe los detalles experimentales en los que se detalla la aplicación y los scripts empleados, donde se dan los detalles de ejecución y el sentido de su funcionamiento; así mismo en esta sección se describe la infraestructura computacional utilizada para las mediciones. La cuarta sección describe los resultados obtenidos, clasificados según la infraestructura empleada. Finalmente, en la quinta parte se presentan las conclusiones.

II. ESTUDIOS RELACIONADOS

Hasta la fecha, se ha evaluado el rendimiento de múltiples aplicaciones en la UNAH, con la finalidad de conformar una solución computacional flexible que utilice el modelo en la nube para resolver problemas de orígenes diversos que requieren el uso de capacidad computacional bajo demanda.

Un caso exitoso de una aplicación bajo demanda portada a un entorno cloud, viene a ser la renderización, que consiste en un proceso computacional intensivo para integrar recursos en la conformación de multimedios. En ello, se empleó Cinelerra [23] para una solución que aprovechaba nodos de procesamiento desplegados de forma virtualizada para completar contenidos para difusión televisiva [2].

Por otra parte, se ha evaluado sobre la nube aplicaciones con fines científicos que han sido una prioridad, como es la elaboración de modelos cosmológicos de la evolución de galaxias, donde se evaluó la aplicación Gadget2 [24] sobre un entorno virtual [3]. Asimismo, se ha evaluado la simulación numérica de nanodispositivos usando recursos en la nube [25], entre otros trabajos con resultados alentadores.

Este trabajo representa el primer estudio orientado a satisfacer las necesidades computacionales intensivas de las tareas de administración de la UNAH, incorporando la tecnología en la nube en el proceso.

III. DETALLES EXPERIMENTALES

En esta sección efectuaremos la descripción de la aplicación que se ha usado en la evaluación del desempeño, efectuaremos la descripción tanto de la infraestructura hardware empleada, así como la descripción de la plataforma cloud y la metodología llevada a cabo.

A. Aplicación de valoración del desempeño docente

Como resultado de las necesidades identificadas en la comunidad universitaria de la UNAH, se ha desarrollado e implementado un instrumento para la Valoración del Desempeño Docente (VDD) accesible por medio de un interfaz web. Gracias a ello, los estudiantes, al finalizar cada período académico, tienen la opción de evaluar a los profesores que les han impartido algún curso. Este aplicativo está compuesto de dos partes: la *primera*, consta de una evaluación cuantitativa a través de preguntas cerradas diseñadas en una escala positiva, mientras que la *segunda* está conformada por preguntas abiertas donde el estudiante puede libremente detallar su experiencia con sus docentes.

Con el fin de procesar millones de registros, se han desarrollado una serie de scripts del tipo procedimiento/función almacenada en MySQL. Estos hacen, entre otras cosas, un conteo de muestras por docentes y una búsqueda de palabras/raíces semánticas agrupadas en patrones de hábitos que se espera que se encuentren en el texto libre introducido como respuestas del instrumento VDD.

Dentro de los scripts evaluados, se encuentra el llamado *muestreo*, que crea una tabla en la que introduce los registros únicos de los docentes que recibieron una evaluación y calcula el valor global de respuestas que recibió. Posteriormente, se ejecuta el script llamado *patrón* que se encarga de buscar los aciertos de las palabras y raíces. Para ello, emplea una búsqueda de cadenas de caracteres (*strings*) usando expresiones regulares. En este artículo, se han empleado cinco patrones con una longitud media entre siete y ocho caracteres por palabra, como se describe en la Tabla I. Para los fines de esta aplicación, un patrón debe entenderse como un conjunto de palabras y raíces semánticas agrupadas y relacionadas entre sí, que en conjunto pueden ser usadas para describir comportamientos de los docentes que son evaluados.

TABLA I. NÚMERO DE PALABRAS Y RAÍCES SEMÁNTICAS AGRUPADAS POR PATRÓN

Patrón	Palabras - Raíces	Media de Caracteres
1	41	8
2	55	7
3	68	8
4	27	7
5	41	7

Se ha preferido el uso de expresiones regulares para facilitar la implementación y migración entre plataformas de diversos fabricantes de los gestores de base de datos basados en el estándar SQL. Es importante mencionar que se ha preferido esta solución sobre las técnicas de búsqueda de texto completa (*full text search*) [20] porque requieren de la creación de índices optimizados que permitan hacer consultas basadas en el sentido (significado) de las palabras utilizadas, uso de adjetivos, adverbios, verbos y sus conjugaciones; que permiten hacer pesquisas extendidas a otros elementos de la muestra tomando los primeros resultados como nuevos parámetros de los subsiguientes; pero su implementación difiere en los gestores de bases de datos, limitando así su portabilidad a otras plataformas.

El gestor de base de datos empleado ha sido MySQL Community Edition v5.6 64bit [17], con el cual se creó una base de datos con el cotejamiento UTF-8 y del tipo MyISAM. Una única tabla sirvió como fuente de datos para análisis. La misma es resultado de restaurar un volcado de datos de otro sistema, que contiene para este artículo la cantidad de 902,051 registros. Los scripts evaluados procesan los datos con el fin de normalizarlos, clasificarlos y, posteriormente, hacer una búsqueda de palabras y raíces semánticas definidas a través de expresiones regulares sobre cadenas de texto con el contenido de interés.

Los scripts desarrollados proporcionan como resultado una tabla en la que se resume el número de aciertos por patrón de comportamiento y la muestra global de las respuestas por docente. Lo anterior, representa una conversión de lo cualitativo a dato cuantitativo que es el objetivo final de toda la aplicación.

B. Descripción de la plataforma física

En esta subsección se describen las características del hardware que ha sido empleado tanto en la plataforma cloud como en las pruebas sobre infraestructura no virtualizada. El *Anfitrión Físico***,** es un ordenador que cuenta con un microprocesador Intel Core i7 Q740 con una velocidad de reloj de 1.73Ghz, 4GB de RAM, disco duro de estado sólido (SSD – *Solid State Drive*) de interfaz SATA. Las pruebas del anfitrión físico se realizaron con dos núcleos activos, esta configuración se cambió en el BIOS. Debido al hyperthreading el sistema operativo puede acceder a cuatro núcleos. *Instancia Cloud***,** es una instancia computacional virtualizada que se desplegó con las mismas características de hardware que el anfitrión físico descrito. Esta instancia se ejecutó siempre sobre el mismo nodo, ya que el cloud utilizado cuenta con varios de ellos. Está gestionado por Apache Cloudstack y accede a un sistema de archivos en red tipo NFS.

Como sistema de almacenamiento primario del cloud, se empleó un sistema de almacenamiento de red dedicado que implementa el protocolo NFS, su hardware fue fabricado por Synology y el modelo DS411, configurado con dos discos duros de 2TB de 7500RPM e interfaz SATA2; ambos configurados en un arreglo RAID 0. Todos los dispositivos que se utilizan, están interconectado a través de una red Ethernet gigabit que provee un switch Cisco Catalyst 500 Series. La instancia virtual, se ha desplegado con 4GB de memoria RAM, un disco duro virtual de 10GB y se le ha deshabilitado la interfaz gráfica.

C. *Descripción de la plataforma cloud*

Los experimentos de ejecución se han realizado empleando dos infraestructuras, una física y una cloud. La instancia virtualizada dentro de un entorno de nube privada es gestionada por Apache CloudStack v4.0.1 [14], implementando el hipervisor KVM [15].Tanto la plataforma física como la virtual se ejecutaron empleando recursos computacionales similares, con el fin de evitar introducir sesgo en los resultados.

Apache CloudStack es una arquitectura de software de código abierto que permite la construcción de varios tipos de nubes: públicas, privadas e híbridas. El servidor de administración de Apache CloudStack controla toda la infraestructura de la nube y asigna las máquinas virtuales a los anfitriones (hosts).

Típicamente, la administración de la nube está compuesta por seis tipos de componentes: la zona de disponibilidad, pods, clústeres, nodos de computación, sistema de almacenamiento primario y el sistema de almacenamiento secundario. Una zona de disponibilidad se puede definir como un único centro de datos, compuesto por uno o más pods con un sistema de almacenamiento secundario. Las zonas de disponibilidad son visibles para el usuario final quien debe seleccionar una de ellos para iniciar una máquina virtual. Un pod es equivalente a un gabinete (*rack*) de hardware, este incluye un conmutador de capa 2 (*switch*) y varios clústers. Un clúster se compone de varios anfitriones con sus respectivos hipervisores y el sistema de almacenamiento primario. Un CN (nodo cloud) es un anfitrión con un hipervisor que se incluye dentro de un clúster de Apache CloudStack; los hipervisores soportados son KVM, Xen, VMware vSphere y Citrix Xen Server. El CN ejecuta máquinas virtuales. Los CN se pueden añadir en cualquier momento para aumentar la capacidad de cálculo de la infraestructura en la nube. Los anfitriones no son visibles para los usuarios y no pueden determinar qué CN se les ha sido asignado para ejecutar su VM (máquina virtual). El sistema de almacenamiento primario se asocia con el clúster, que guarda los volúmenes de disco para las máquinas virtuales ejecutadas. El sistema de almacenamiento secundario se asocia con una zona de disponibilidad, siendo el encargado de almacenar las plantillas de las máquinas virtuales, imágenes ISO y las instantáneas del volumen de disco asociadas a las VM. Lo anterior se muestra en la Fig. 1.

Apache CloudStack soporta tres roles de usuario: administrador root, administrador de dominio y usuario sin privilegios. El administrador root, puede administrar la plataforma Apache CloudStack por completo. El administrador de dominio, puede realizar y completar labores de gestión de la infraestructura para los usuarios que pertenecen a un único dominio específico. Los usuarios sin privilegios pueden administrar sus propias VM y acceder a ellas a través de una conexión de red.

Con el fin de tener una medición del tiempo de ejecución lo más precisa posible, se ha hecho una sincronización del reloj del ordenador anfitrión y de la instancia cloud, por medio del uso del protocolo NTP [21], implementado al usar el

comando *ntpdate*. Finalmente, el sistema operativo común para todos los equipos fue el Linux en su distribución CentOS v6.6 64bit [16].

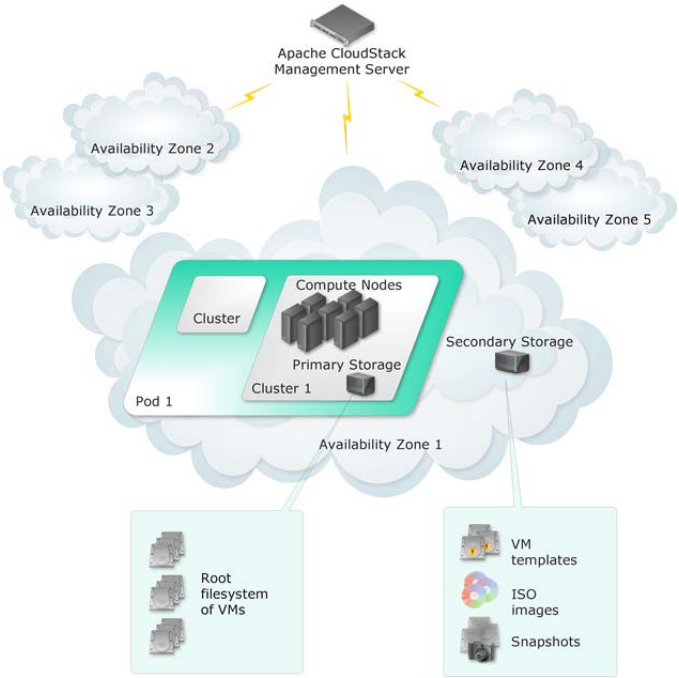

Figura 1. Infraestructura de operación de Apache CloudStack.

D. Metodología

Con la finalidad de conocer de forma unificada el rendimiento de las diferentes plataformas en lo que respecta al sistema de archivos, se hizo uso de IOzone v3.424 64bit [22], que es un benchmark sintético que se utiliza para conocer la medida del rendimiento del sistema de archivos. IOzone fue ejecutado evitando la influencia de la cache en el acceso al sistema de ficheros. Para esto, se usó un tamaño de archivo que fuese equivalente al doble de la memoria RAM disponible por cada plataforma, evaluando de esta manera la velocidad de escritura, re-escritura, lectura, re-lectura del tipo secuencial y aleatoria.

Posteriormente, se han desarrollado una serie de scripts del tipo procedimiento/función almacenada en MySQL con la intención de evaluar el desempeño de este gestor de base de datos en cloud. El primer tipo de script desarrollado, denominado *muestreo,* se encarga de crear una tabla que contiene un registro único con los datos del docente y la frecuencia de las respuestas asociadas a él. El segundo tipo de script, llamado *patrón,* calcula el número de aciertos de las palabras y raíces semánticas encontradas en la muestra específica por docente. Se repitió la ejecución de los scripts en diferentes ocasiones, teniendo el cuidado de borrar los resultados y reiniciar las plataformas al finalizar cada ciclo.

Se realizaron doce mediciones con cada patrón en cada plataforma, pero se descartaron las dos primeras para limitar el efecto de calentamiento (*warm up*) del gestor de base de datos. Entre cada ciclo de mediciones, se eliminaba la estructura de datos y se restauraba de nuevo a partir de un respaldo que contiene los datos crudos y se reiniciaban las medidas las medidas.

También se midió el uso de la memoria RAM y de la memoria de intercambio SWAP previo al inicio y posterior a la ejecución, para considerar con ello el efecto del uso de la memoria y en qué medida esta interferiría en las mediciones obtenidas.

IV. RESULTADOS

En esta sección efectuaremos la descripción de los resultados obtenidos para MySQL en la plataforma cloud, en la máquina física empleando discos SSD, así como la comparación de sus resultados. Además, se efectúa la descripción de los resultados obtenidos para IOzone en el cloud, en la máquina física con discos SSD y, también, en máquina física sobre NFS, así como la comparación de sus resultados.

A. MySQL en cloud

La evaluación inició con el *muestreo*, que en el caso de la instancia en la nube, requirió en promedio 6919.34 segundos para completar y reducir los datos en una tabla con los registros y frecuencias de respuestas. Las mediciones presentaron una desviación estándar de 502.06 segundos.

La ejecución del script patrón, tuvo un comportamiento heterogéneo entre las diferentes series en la instancia cloud. El cálculo medio del tiempo requerido por los aciertos del patrón(1) fue de 6212.21 segundos, el patrón(2) utilizó una media de 6023.78 segundos, siendo el más eficiente. El patrón(3) utilizó 6314.30 segundos, el patrón(4) 6290.50 segundos y finalmente el patrón(5) utilizó 6635.32 segundos el cual representó el mayor tiempo de ejecución.

TABLA II. TIEMPO DE EJECUCIÓN REQUERIDO POR LOS DIFERENTES PATRONES EN LA INSTANCIA CLOUD. A MENOR TIEMPO MEJOR RENDIMIENTO

Patrón	Tiempo Medio (s)	Desviación Estándar (s)	Número de Palabras	Tiempo Medio por Palabra (s)
1	6212.21	237.46	41	151.52
2	6023.78	91.18	55	109.52
3	6314.30	155.52	68	92.86
4	6290.50	183.24	27	232.98
5	6635.32	118.53	41	161.84

En cuanto a los tiempos medios de ejecución por palabra el patrón(3) requirió una media de 92.86 segundos por palabra, siendo el que usó menos tiempo de todos; en cambio el patrón(4) fue el que más tiempo empleó con una media de 232.98 segundos. Se hace notar que el patrón(1) y el patrón(5) tienen el mismo número de palabras, pero este último conjunto requirió 423.11 segundos más para completarse; muy

probablemente este sea efecto de la moda estadística de las palabras que conforman al patrón(1), que tienden a ser más cadenas de mayor tamaño. El promedio general requerido por palabra fue de 135.67 segundos entre todos los patrones. El comportamiento general de la aplicación evaluada en la instancia cloud se describe en la Tabla II.

Los tiempos de ejecución más dispersos fueron los del patrón(1) con una desviación de 237.46 segundos, el patrón(2) tuvo la menor desviación estándar entre sus medidas con un valor de 91.18 segundos; el patrón(3) presentó 155.52 segundos, el patrón(4) de 183.24 segundos y el patrón(5) de 118.53 segundos de desviación estándar respectivamente.

El uso de memoria RAM en la instancia cloud antes de la ejecución fue en promedio 676.30MB, con una desviación estándar de 1.95MB, con un uso máximo de 680MB y el mínimo de 674MB en toda la serie. Posterior a la ejecución, el uso promedio fue 1341.5MB, con una desviación estándar de 20.87MB, con un uso máximo de 1381MB y un mínimo de 1327MB. Entre antes y después de la ejecución la diferencia de uso oscila en 665.20MB en promedio entre todas las mediciones realizadas.

B. MySQL en máquina física SSD

El *muestreo*, en la máquina física requirió en promedio 4859.08 segundos para completar y reducir los datos, pero presentó una desviación estándar de 255.82 segundos entre todas las mediciones.

TABLA III. TIEMPO DE EJECUCIÓN POR PATRÓN EN LA MÁQUINA FÍSICA. A MENOR TIEMPO MEJOR RENDIMIENTO

Patrón	Tiempo Medio (s)	Desviación Estándar (s)	Número de Palabras	Tiempo Medio por Palabra (s)
1	4159.53	106.21	41	101.45
2	4299.23	126.17	55	78.17
3	4324.69	165.42	68	63.60
4	4369.01	70.34	27	161.81
5	4878.44	267.67	41	118.99

En la máquina física con disco SSD, se mantuvo una menor desviación entre las medidas del tiempo de ejecución de cada patrón, pero entre ellos no fue completamente homogéneo. El patrón(1) utilizó una media de 4159.53 segundos para completarse, siendo el más eficiente. El patrón(2) usó 4299.23 segundos, el patrón(3) empleó 4324.69 segundos, el patrón(4) requirió 4369.01. El patrón(5) usó 4878.44 segundos el cual representó el mayor tiempo de ejecución. Es importante mencionar que el patrón(5), tiene el mismo número de palabras que el patrón(1), pero presentaron una diferencia significativa de ejecución de 718.91 segundos.

El tiempo medio requerido por palabra por el patrón(1) fue de 101.45 segundos, el patrón(2) usó 78.17 segundos; el patrón(3) sólo requirió una media de 63.60 segundos, siendo el más eficiente. En cambio, el patrón(4) fue el que más tiempo empleó con una media de 161.81 segundos y, finalmente, el patrón(5) requirió 118.99 segundos.

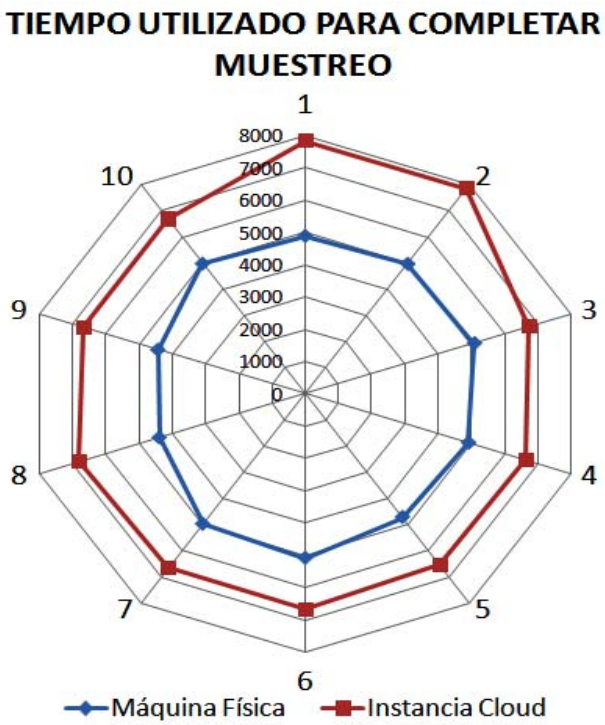

Figura 2. Tiempo (en segundos) utilizado por las infraestructuras para completar el script muestreo, en una serie de diez pruebas consecutivas de ejecución. Su comportamiento es homogéneo en toda la serie.

En cuanto a la dispersión de las mediciones, el patrón(4) presentó la menor desviación estándar entre sus medidas, con un valor de 70.34 segundos. En cambio, los más dispersos fueron los del patrón(5), con una desviación de 267.67 segundos. El promedio general requerido por palabra fue de 94.96 segundos entre todos los patrones. Gráficamente, el comportamiento de la infraestructura física se describe en la Tabla III.

El uso de memoria RAM en la máquina física antes de la ejecución fue en promedio1193.4MB, con una desviación estándar de 23.56MB, con un uso máximo de 1240MB y el mínimo de 1170MB en toda la serie. Posterior a la ejecución el uso promedio fue 1910.3MB, con una desviación estándar de 43.28MB, con un uso máximo de 2003MB y mínimo de 1858MB. Entre antes y después de la ejecución la diferencia de uso oscila en 716.90MB en promedio en todas las mediciones realizadas.

C. Comparación de resultados MySQL

La instancia en la nube requirió 42.4% más tiempo de ejecución que la máquina física para completar el muestreo. Ello equivale a 2060.27 segundos de diferencia promedio para finalizar el procedimiento almacenado. A grandes rasgos, ambas infraestructuras mantuvieron un comportamiento

homogéneo entre si durante las diversas mediciones en el conjunto de pruebas efectuadas, su comportamiento se describe gráficamente en la Fig. 2.

En cuanto a los patrones, el comportamiento entre infraestructuras fue similar entre todas las series. En todos los casos, la instancia en la nube requirió más tiempo para completar las tareas de ejecución frente a la instancia física. La mayor diferencia fue en el patrón(1), que requirió un 49% de tiempo adicional, en cambio el patrón(5) sólo necesitó un 36% más. El patrón(2) necesitó 40%, el patrón(3) un 46% y el patrón(4) un 44% de tiempo adicional respectivamente. En general, la instancia cloud requirió en promedio de un 43% más tiempo para completar todos los patrones.

En cuanto al uso de memoria, ésta se mantuvo sin variaciones importantes entre cada ciclo de ejecución. Es importante mencionar y hacer notar que la memoria de intercambio no fue utilizada en ninguna plataforma, lo que indica que el gestor de base de datos no escaló de forma que requiriera esos recursos. Las mediciones se detallan en la Tabla IV.

TABLA IV. USO DE MEMORIA RAM ENTRE LAS INFRAESTRUCTURAS. VALORES ESTÁN DADOS EN MEGABYTES

Serie	INSTANCIA CLOUD			MÁQUINA FÍSICA		
	Antes	Después	Uso	Antes	Después	Uso
1	675	1329	*654*	1171	1858	*687*
2	680	1340	*660*	1171	1926	*755*
3	677	1330	*653*	1199	1926	*727*
4	674	1381	*707*	1225	1910	*685*
5	678	1334	*656*	1181	1926	*745*
6	676	1327	*651*	1240	2003	*763*
7	675	1330	*655*	1199	1889	*690*
8	676	1334	*658*	1186	1858	*672*
9	674	1380	*706*	1192	1931	*739*
10	678	1330	*652*	1170	1876	*706*

En cuanto al uso de memoria RAM, antes de la ejecución el ordenador físico usaba en promedio 1193MB y la instancia 677MB con una diferencia de 43.25% menos de uso inicial de memoria por parte del cloud. Esto se debe principalmente, a que el sistema operativo en la instancia carece de interfaz gráfica y se está ejecutando en modo consola, condición que se mantiene en todo momento. Posterior a la ejecución el uso del ordenador físico fue de 1931MB y la instancia empleó 1340MB, siendo un 30.60% menos. Por otro lado, el uso promedio de la memoria RAM previo a la ejecución por parte de la instancia cloud fue de 665.20MB siendo menor la infraestructura física que usó 716.90MB; esto es un incremento de un 7.77% por parte de la última.

D. IOzone en cloud

En la ejecución de IOzone en esta plataforma, en cuanto a sus rendimientos promedio en escritura reportó 11352Kbit/s, en re-escritura presentó 11241Kbit/s, lectura 11686Kbit/s, re-lectura 12964Kbit/s, lectura aleatoria 9282Kbit/s y, finalmente, en escritura aleatoria 11190Kbit/s.

En general, las mediciones mantuvieron un comportamiento similar en todas las ejecuciones, excepto en la lectura que presentó una desviación estándar de 1317.69 segundos. De igual manera hubo dispersión en las medidas de re-lectura con una desviación estándar de 1622.45 segundos y en la lectura aleatoria con 669.11 segundos respectivamente. Esta dispersión en las medidas se origina por la tecnología de plato rotatorio y cabezales que compone al arreglo de discos que provee el NAS, donde se realizan las pruebas del cloud. Adicionalmente al ser un sistema virtualizado es afectado por el controlador de disco que provee el hipervisor.

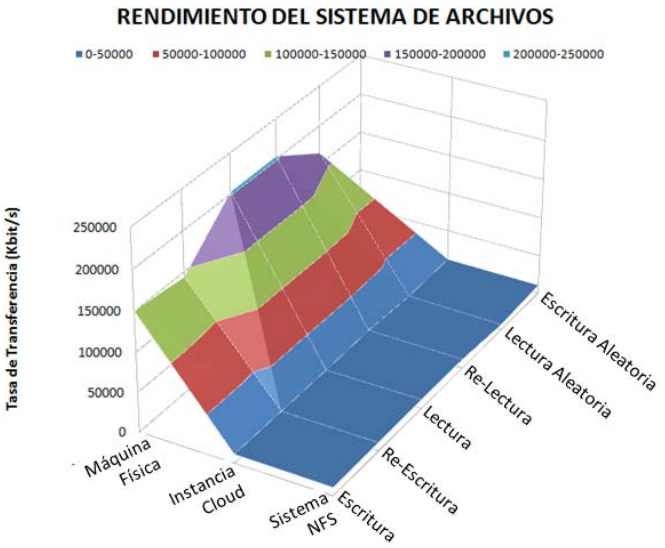

Figura 3. Rendimiento del sistema de ficheros, los valores mayores representan los mejores resultados.

E. IOzone en máquina física SSD

En esta serie de mediciones se obtuvieron los mejores resultados. Su rendimiento promedio en escritura reportó 150630Kbit/s, en re-escritura presentó 143570Kbit/s, lectura 202387Kbit/s, re-lectura 203521Kbit/s, lectura aleatoria 164737Kbit/s y finalmente en escritura aleatoria 54095Kbit/s.

En esta serie de mediciones se obtuvo una dispersión muy baja, siendo la de mayor proporción el proceso de escritura que presentó una desviación estándar de 1263.40Kbit/s. Las otras mediciones no presentaron dispersión de importancia.

F. IOzone en máquina física sobre NFS

En la ejecución de IOzone en esta plataforma, en cuanto a su rendimiento promedio en escritura reportó 10467Kbit/s, en re-escritura presentó 10614Kbit/s, en lectura 11331Kbit/s, en re-lectura 11330Kbit/s, en lectura aleatoria 6648Kbit/s y finalmente en escritura aleatoria 10529Kbit/s. En general, las mediciones mantuvieron un comportamiento similar en todas las ejecuciones sin presentar una dispersión de importancia.

G. Comparación de resultados IOzone

Los mejores resultados los presentó la máquina física, la cual contaba con una unidad de almacenamiento secundario del tipo SSD. En cambio, la instancia física con almacenamiento NFS presentó el peor rendimiento; pero muy similar a la instancia cloud, siendo esta mejor en re-lectura en un 14% y lectura aleatoria en un 39%. El sistema de almacenamiento en red (NAS) fue accedido sobre una red Ethernet gigabit con el fin de limitar el efecto de la red, pero la misma no fue causa del bajo rendimiento; sino que la tecnología que utilizan los discos que la conforma que es la clásica de platos rotatorios y cabezales de lectura.

El sistema de archivo que aprovechó el de almacenamiento SSD, presentó su mayor eficiencia en las tareas de lectura y re-lectura siendo hasta diez y siete veces superior al NFS. El comportamiento en las diferentes modalidades de la prueba de rendimiento puede observarse en la Fig. 3.

V. CONCLUSIONES

En este artículo se ha evaluado el desempeño de un gestor de base de datos MySQL en una plataforma cloud y se ha efectuado la comparación con la ejecución sobre una plataforma física. Para ello, se han desarrollado una serie de scripts del tipo procedimiento/función almacenada en MySQL con la intención de evaluar el desempeño de este gestor de base de datos en cloud. El primer tipo de script desarrollado, denominado *muestreo,* se encarga de crear una tabla que contiene un registro único con los datos del docente y la frecuencia de las respuestas asociadas a él. El segundo tipo de script, llamado *patrón,* calcula el número de aciertos de las palabras y raíces semánticas encontradas en la muestra específica por docente. Se repitió la ejecución de los scripts en diferentes ocasiones, teniendo el cuidado de borrar los resultados y reiniciar las plataformas al finalizar cada ciclo.

Al ejecutar las pruebas de rendimiento en ambas plataformas, el ordenador físico mostró un rendimiento superior respecto a la instancia cloud, ya que esta última emplea un sistema de almacenamiento basado en disco de platos rotatorios siendo menos eficiente. La tecnología SSD provee una interesante alternativa para ser considerada como sistema de almacenamiento en una infraestructura cloud, siendo la única limitante de esto la relación de precio y capacidad. En cambio, al evaluar el rendimiento de la ejecución de los scripts dentro del gestor de base de datos, los resultados se ven sacrificados en un tercio en la instancia cloud frente a la alternativa física.

Con el fin de identificar a qué es debido esta pérdida de rendimiento, se ha ejecutado IOzone bajo tres condiciones diferentes: sobre la plataforma cloud, sobre la máquina física empleando como almacenamiento el SSD, y sobre la máquina física empleando como almacenamiento un directorio exportado desde el sistema de almacenamiento primario del cloud exportado por medio de NFS. Los resultados han mostrado que el cuello de botella de la infraestructura cloud proviene del sistema de almacenamiento primario empleado en el cloud, el sistema NAS en RAID 0, que es el que provee los discos raíz de las máquinas virtuales.

A corto plazo, se requerirá utilizar métodos y herramientas para una búsqueda más compleja en la que se incluya el significado de las palabras de cada patrón dentro del contexto de la respuesta del instrumento de Valoración del Desempeño Docente; en este sentido una infraestructura de altas prestaciones ofrece la flexibilidad para desplegar diversas herramientas en instancias virtuales en la nube, para que de esta manera se pueda realizar ejecución simultánea de tareas de análisis sin sacrificar demasiado el rendimiento frente a una infraestructura dedicada.

El cambio del sistema de almacenamiento a uno de estado sólido puede aportar mucho beneficio en el tiempo de ejecución de aplicaciones en una instancia cloud, este deberá ser un punto de estudio a futuro para conocer las ventajas y poderlas cuantificar.

REFERENCIAS

[1] UNAH. Universidad Nacional Autónoma de Honduras. 2015. URL: http://www.unah.edu.hn.

[2] ZABLAH, I.; GARCIA-LOUREIRO, A.; GOMEZ-FOLGAR, F.; PENA, T.F. *Render on the cloud: Using Cinelerra on virtualized infraestructures*. [ed.] Alexander Mikroyannidis, Rocael Hernandez Rizzardini and Hans-Christian Schmitz. Antigua: WCLOUD 2012 Proceedings, 2012. Proceedings of the 1st International Workshop on Cloud Education Environments (WCLOUD2012). pp. 28-32.

[3] ZABLAH, JOSÉ ISAAC; GARCÍA LOUREIRO, ANTONIO. *Análisis del rendimiento de una aplicación cosmológica sobre máquinas virtuales*. Actas de las XXIII Jornadas de Paralelismos. Elche, Alicante-España : Universidad de Elche, 2012. pp. 406-411.

[4] BUYYAA, RAJKUMAR; SHIN CHEE, YEO; VENUGOPALA, SRIKUMAR; BROBERG, JAMES; BRANDIC, IVONA. *Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility.* Future Generation Computer Systems, Vol. 25, pp. 599-616.

[5] MELL, PETER; GRANCE, TIMOTHY. *The NIST Definition of Cloud Computing. Computer Security Division*. United States Department of Commerce. Gaithersburg : s.n., 2011. Special Publication 800-145.

[6] BEN LETAIFA, ASMA; HAJI, AMEL; JEBALIA, MAHA; TABBANE, SAMI. *State of the Art and Research Challenges of new services architecture technologies: Virtualization, SOA and Cloud Computing.*2010. International Journal of Grid and Distributed Computing, Vol. 3.

[7] ZHANG, LIANG-JIE; ZHOU, QUN.*CCOA: Cloud Computing Open Architecture*. New York,USA. 2009 IEEE International Conference on Web Services.

[8] BÔAVENTURA, R. S.; YAMANAKA, K.; OLIVEIRA, G.P. - *Performance Analysis of Algorithms for Virtualized Environments on Cloud Computing*. IEEE Latin America Transactions, Vol. 12, No. 4, Junio - 2014.

[9] SANTANA, P.C.; GONZALES, F.J.; GARCIA M.A.; MAGAÑA, M.A. - *Social Cloud Computing: an Opportunity for Technology Enhanced Competence Based Learning*. IEEE Latin America Transactions, Vol. 13, No. 1, Enero - 2015.

[10] MARSTON, SEAN; LI, ZHI; BANDYOPADHYAY, SUBHAJYOTI; ZHANG, JUHENG; GHALSASI, ANAND. *Cloud computing — The business perspective*. 2011. Decision Support Systems, Vol. 51.

[11] SULTAN, NABIL. *Cloud computing for education: A new dawn?*. 2010, International Journal of Information Management, Vol. 30, pp. 109-116.

[12] DIKAIAKOS, MARIOS D.; PALLIS, GEORGE; KATSAROS, DIMITRIOS; MEHRA, PANKAJ; VAKALI, ATHENA. *Distributed Internet Computing for IT and Scientific Research*. SEPTEMBER-OCTOBER 2009. IEEE Internet Computing, pp. 10-13.

[13] ERCAN, TUNCAY. *Effective use of cloud computing in educational institutions*. 2010. Procedia Social and Behavioral Sciences, Vol. 2, pp. 938-942.

[14] CLOUDSTACK, APACHE. 2015. URL: https://cloudstack.apache.org/

[15] KVM, Kernel Virtual Machine. 2015. URL: http://www.linux-kvm.org/page/Main_Page.

[16] PROJECT CENTOS. 2015. URL: http://www.centos.org.

[17] MySQL. 2015. URL: http://dev.mysql.com/.

[18] STUART WARD, JONATHAN; BARKER, ADAM. *Undefined By Data: A Survey of Big Data Definitions*. 2013. arXiv:1309.5821v1.

[19] NIST Big Data Working Group (NBD-WG). 2015. URL: http://bigdatawg.nist.gov/home.php..

[20] ORACLE MySQL. *MySQL 5.6 Reference Manual*. 2015. URL: https://dev.mysql.com/doc/refman/5.6/en/innodb-fulltext-index.html.

[21] NETWORK TIME FUNDATION. *Network Time Protocol*. 2015 URL: http://www.ntp.org/.

[22] IOZONE. *Benchmark Filesystem*. 2015. URL: http://www.iozone.org/.

[23] CINELERRA. *The GNU Video Editor*. 2015. URL: http://www.cinelerra.org/.

[24] GADGET2. *Code for cosmological simulations of structure formation*. 2015. URL: http://wwwmpa.mpa-garching.mpg.de/gadget/.

[25] SEOANE, N.; VALIN, R.; GARCÍA LOUREIRO, A.; PENA, T.F.; ZABLAH, I. *Performance of numerical simulations on the cloud*. Actas de las XXIII Jornadas de Paralelismos. Elche, Alicante-España : Universidad de Elche, 2012. pp. 395-399.

José Isaac Zablah Avila es Profesor de la Facultad de Ciencias Médicas - Universidad Nacional Autónoma de Honduras. Sus áreas de interés en investigación son la computación y redes de altas prestaciones, gobierno electrónico, reducción del dividendo digital, seguridad de la información entre otros.

Iris Xiomara Corrales es Directora de Carrera Docente - Universidad Nacional Autónoma de Honduras y tiene más de treinta años de experiencia docente en el Departamento de Matemáticas. En el ámbito privado posee catorce años de experiencia en el sector financiero hondureño en áreas de banca comercial y bolsa de valores.

José Medardo Aguilar es Especialista II en Sistemas de Información de la Dirección de Carrera Docente - Universidad Nacional Autónoma de Honduras. Sus áreas de interés en desarrollo de sistemas de información cliente servidor y ambiente web, bases de datos, servidores y certificados digitales entre otros.

Antonio de Jesús García Loureiro es profesor permanente del Departamento de Electrónica y Computación de la Universidad de Santiago de Compostela. Actualmente coordinador del Máster Interuniversitario de Investigación en Tecnologías de la Información y del Programa de Doctorado Interuniversitario en Tecnologías de la Información, Profesor visitante en las Universidades Politécnica de Cataluña, Granada, Edinburgh, Glasgow y Swansea. Ha desarrollado su investigación en el campo de computación de altas prestaciones, implementación de herramientas de simulación y despliegue de infraestructuras de computación.

Fernando Gómez Folgar es investigador doctoral en El Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS) – Universidad Santiago de Compostela. Entre sus intereses en investigación se encuentran la computación de altas prestaciones, el paradigma de computación en la nube y sus aplicaciones, entre otras.

Marco Tulio Medina Hernández es el Decano de la Facultad de Ciencias Médicas de la Universidad Nacional Autónoma de Honduras. Médico especialista en neurología y sub-especialista en neurofisiologia clínica y epileptología pediátrica y del adulto. Director actual de neurología para América Latina de la Federación Mundial de Neurología (WFN) y es Co-fundador de la Federación Panamericana de Sociedades Neurológicas (PAFNS). Entre sus áreas de interés en investigación se encuentran las aplicaciones de supercomputación aplicadas a las ciencias médicas, entre otras.