UNIVERSIDAD TÉCNICA DE AMBATO

FACULTAD DE INGENIERÍA EN SISTEMAS, ELECTRÓNICA E INDUSTRIAL

CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES E INFORMÁTICOS

TEMA

ANÁLISIS DE MÉTODOS DE DEDUPLICACIÓN DE DATOS APLICADOS EN REPOSITORIOS LINUX PARA LA FACULTAD DE INGENIERÍA EN SISTEMAS ELECTRÓNICA E INDUSTRIAL

Trabajo de Graduación. Modalidad: Proyecto de Investigación, presentado previo la obtención del título de Ingeniero en Sistemas Computacionales e Informáticos.

SUBLÍNEA DE INVESTIGACIÓN: Redes de Computadoras

AUTOR: Morey Solís David Giovanni

PROFESOR REVISOR: Ing. Ernesto Pérez Estévez, Mg.

Ambato – Ecuador

Agosto – 2015

APROBACIÓN DEL TUTOR

En mi calidad de Tutor del Trabajo de Investigación sobre el Tema: **"Análisis de métodos de Deduplicación de datos aplicados en repositorios Linux para la Facultad de Ingeniería en Sistemas Electrónica e Industrial.",** del Señor David Giovanni Morey Solis, estudiante de la Carrera de Ingeniería en Sistemas Computacionales e Informáticos, de la Facultad de Ingeniería en Sistemas, Electrónica e Industrial, de la Universidad Técnica de Ambato, considero que el informe investigativo reúne los requisitos suficientes para que continúe con los trámites y consiguiente aprobación de conformidad con el Art. 16 del Capítulo II, del Reglamento de Graduación para Obtener el Título Terminal de Tercer Nivel de la Universidad técnica de Ambato.

Ambato, Agosto de 2015

EL TUTOR

_____

Ing. Ernesto Pérez Estévez, Mg.

AUTORÍA

El presente trabajo de investigación titulado: **"ANÁLISIS DE MÉTODOS DE DEDUPLICACIÓN DE DATOS APLICADOS EN REPOSITORIOS LINUX PARA LA FACULTAD DE INGENIERÍA EN SISTEMAS ELECTRÓNICA E INDUSTRIAL.".** Es absolutamente original, auténtico y personal, en tal virtud, el contenido, efectos legales y académicos que se desprenden del mismo son de exclusiva responsabilidad del autor.

Ambato, Agosto de 2015

_____

David Giovanni Morey Solís

CC: 1804518643

APROBACIÓN COMISIÓN CALIFICADORES

La Comisión Calificadora del presente trabajo conformada por los señores docentes Ing. David Guevara Aulestia e Ing. Franklin Mayorga Mayorga, revisó y aprobó el Informe Final del trabajo de graduación titulado **ANÁLISIS DE MÉTODOS DE DEDUPLICACIÓN DE DATOS APLICADOS EN REPOSITORIOS LINUX PARA LA FACULTAD DE INGENIERÍA EN SISTEMAS ELECTRÓNICA E INDUSTRIAL**, presentado por el señor David Giovanni Morey Solís de acuerdo al Art. 17 del Reglamento de Graduación para obtener el título Terminal de tercer nivel de la Universidad Técnica de Ambato.

_____

Ing. José Vicente Morales Lozada. Mg.

PRESIDENTE DEL TRIBUNAL

_____ _____

Ing. David Guevara Aulestia. Mg. Ing. Franklin Mayorga Mayorga. Mg

DOCENTE CALIFICADOR DOCENTE CALIFICADOR

DEDICATORIA:

Esta Tesis se la dedico en primera instancia a Dios por darme la fuerza suficiente para seguir día a día,

A mis padres por siempre estar a mi lado apoyándome en cada una de las decisiones acertadas o no acertadas que he tomado en mi vida académica,

A mi hermano por quien me esfuerzo cada día más para ser un espejo en el cual él pueda verse reflejado,

A mi novia por su apoyo incondicional durante cada etapa universitaria,

Se la dedico también a toda mi familia por siempre estar pendientes de mí andar durante la carrera Universitaria.

A mis amigos y compañeros por siempre apoyarme cada día en las aulas,

A mis maestros por confiar en mis habilidades y destrezas como estudiante y como persona.

A todas las personas que mencione… mil gracias.

Giovanni Morey Solis.

AGRADECIMIENTO:

Mi agradecimiento sincero a todos quienes hacen la Universidad Técnica de Ambato, Autoridades, Docentes y Personal Administrativo por darme la oportunidad de ser parte del Alma Mater Ambateña, de manera muy especial a la Facultad de ingeniería en Sistemas Electrónica e Industrial por permitirme formar parte de sus valiosos estudiantes.

Al Ingeniero Ernesto Pérez Estévez, Profesor Tutor, por brindarme su conocimiento y confianza para poder llevar a cabo este proyecto de investigación.

A todos quienes directa o indirectamente contribuyeron al desarrollo de esta Tesis, gracias por su ayuda y su apoyo

Giovanni Morey Solis.

ÍNDICE

ÍNDICE ..... **vii**

ÍNDICE DE TABLAS ..... **x**

ÍNDICE DE FIGURAS ..... **xi**

RESUMEN ..... **xiii**

ABSTRACT ..... **xv**

GLOSARIO DE TÉRMINOS Y ACRÓNIMOS ..... **xvii**

INTRODUCCIÓN **1**

CAPÍTULO I **2**

EL PROBLEMA **2**

1.1 Tema 2

1.2 Planteamiento del problema 2

1.3 Delimitación 4

1.4 Justificación 4

1.5 Objetivos 6

1.5.1 Objetivo General 6

1.5.2 Objetivos Específicos 6

CAPÍTULO II **7**

MARCO TEÓRICO **7**

2.1 Antecedentes Investigativos 7

2.2 Fundamentación teórica 9

2.2.3 Deduplicación 9

2.2.4 Funcionamiento de la Deduplicación de datos 10

2.2.5 Tipos de duplicación de datos 11

Enfoques por bloque 11

Deduplicación por bytes de datos 12

Deduplicación por Algoritmo 13

Deduplicación por el nivel de análisis (Archivo) 13

2.2.6 Inicios del almacenamiento de la información y su administración 14

2.2.7 Sistema Integral de Información 15

Breve Historia de la Deduplicación 16
2.2.8 La Deduplicación y su importancia en la actualidad 16
2.2.9 Compresión de Datos vs. Deduplicación 17
2.2.10 Parámetros a evaluar antes de realizar una Deduplicación de Información. 18
Análisis de Software 19
Hardware que se posee 20
Funcionamiento en conjunto de software y hardware con el fin de deduplicar 21
2.2.11 Deduplicación en Linux 22
Linux 22
Software Libre 22
¿Qué es el software libre? 22
Reglamentos que definen al Software Libre 23
Beneficios del Software Libre y de Código Abierto 24
2.3 Propuesta de solución 25
CAPÍTULO III 26
METODOLOGÍA 26
3.1 Tipo de investigación 26
3.2 Modalidad 26
3.3 Población y muestra 26
3.4 Recolección de la información 27
3.5 Procesamiento y análisis de datos 27
3.6 Desarrollo del proyecto 27
CAPÍTULO IV 29
DESARROLLO DE LA PROPUESTA 29
4.1 Análisis de la situación Actual de la Facultad de Ingeniería en Sistemas Electrónica e Industrial. 29
4.2 Análisis de Técnicas 31
LessFS 31
SDFS 32
ZFS 33
4.3 FS Nativo en CentOS vs SDFS y ZFS. 37
4.4 Descripción de Hardware y Software. 38
4.5 Creación e instalación de máquinas virtuales 40

4.6 Instalación y configuración de servidor prototipo de Mirror (HTTP Service) 43
4.6.1 Procedimiento de configuración de Shorewall 45
4.7 Instalación y configuración de servidor prototipo de Mirror (FTP Service) ... 48
4.8 Configuración de Directorio raíz del repositorio 48
4.9 Descarga, Instalación y configuración de los sistemas para Deduplicación de datos 50
4.9.1 LessFS 50
4.9.2 SDFS 52
4.9.3 ZFS 57
4.10 Pruebas de funcionamiento de Deduplicación en FS ZFS y SDFS 61
4.10.1 SDFS 61
4.10.2 ZFS 64
4.10.3 Pruebas Funcionamiento Prototipo de Mirror 70
4.10.4 Análisis de Resultados 72
Pruebas de Funcionamiento: Verificación de ahorro de espacio de almacenamiento en Discos Duros. 72
Prueba de Funcionamiento: Rendimiento de CPU, RAM, Red TX al realizar procesos de deduplicación 75
CAPÍTULO V 79
CONCLUSIONES Y RECOMENDACIONES 79
5.1 Conclusiones 79
5.2 Recomendaciones 80
BIBLIOGRAFÍA 81
ANEXOS 84
Anexo 1. 84
Anexo 2. 86
Anexo 3. 87

ÍNDICE DE TABLAS

Tabla 1: Especificaciones técnicas del equipo de cómputo usado como servidor para montar las máquinas virtuales..... 38
Tabla 2 : Especificaciones técnicas de las máquinas virtuales. 38
Tabla 3: Tabla comparativa de resultados SDFS vs. ZFS. 78

ÍNDICE DE FIGURAS

Figura 1: Gráfica de Almacenamiento usando Deduplicación. 9
Figura 2: Funcionamiento de la Deduplicación de datos. 10
Figura 3: Reducción de espacio al usar técnicas de Deduplicación. 11
Figura 4: Deduplicación Nivel de Bloque. 11
Figura 5: Deduplicación Nivel de Byte. 12
Figura 6: Deduplicación Nivel de Byte. 13
Figura 7: Esquema general de virtualización. 39
Figura 8: Descarga de paquetes para entorno gráfico GNOME. 40
Figura 9: Instalación de paquetes para entorno gráfico GNOME. 41
Figura 10: Configuración de Usuario y contraseña bajo el entorno gráfico GNOME. .. 42
Figura 11: Escritorio de trabajo entorno gráfico GNOME en CentOS 6.6. 42
Figura 12: Habilitación de Servicios HTTP y HTTPS. 43
Figura 13: Habilitación del puerto 80 en el archivo de configuración iptables. 44
Figura 14: Formato de archivo de configuración Shorewall. 45
Figura 15: Instalación de Java mediante línea de comandos 52
Figura 16: Descarga e instalación del Sistema de archivos SDFS. 53
Figura 17: Verificación de la descarga e instalación de SDFS. 53
Figura 18: Detención de servicios de Firewall. 54
Figura 19: instalación del Sistema para manejo y administración de Discos Duros GParted. 55
Figura 20: Funcionamiento de la herramienta GParted. 55
Figura 21: Creación del Volumen con un Sistema de archivos SDFS. 56
Figura 22: Volumen SDFS montado en nuestro servidor. 56
Figura 23: Cambio de la configuración de SELinux. 57
Figura 24: Instalación del paquete de librerías "Development Tools". 58
Figura 25: Carga de módulos ZFS en el sistema. 58
Figura 26: Visualización de Discos duros con direcciones físicas. 59
Figura 27: Creacion de piscina con el FS ZFS. 59
Figura 37: Direccionamiento al directorio del volumen con FS SDFS. 61
Figura 38: Verificación del proceso de Deduplicación SDFS con archivos iguales. 62
Figura 39: Visualización de espacio físico disponible y usado. SDFS. 62

Figura 40: Proceso de Deduplicación en archivos distintos SDFS..... 63
Figura 41: Verificación del proceso de Deduplicación con archivos distintos SDFS. ... 63
Figura 28: Propiedades del directorio usado para las pruebas de funcionamiento de la Deduplicación. 64
Figura 29: Direccionamiento al directorio del volumen con FS ZFS. 64
Figura 30: Funcionamiento comando zpool list. 65
Figura 31: Visualización del funcionamiento de ZFS. 65
Figura 32: Deduplicación ZFS – Duplicación de archivos. 66
Figura 33: Verificación de proceso de Deduplicación en ZFS FS. 66
Figura 34: Funcionamiento de zpool list después de duplicar información. 67
Figura 35: Proceso de Deduplicación en archivos distintos ZFS. 68
Figura 36: Verificación del proceso de Deduplicación con archivos distintos ZFS. 68
Figura 42: Prueba de descarga de archivos desde el servidor FTP bajo SDFS. 70
Figura 43: Prueba de descarga de archivos desde el servidor FTP bajo ZFS. 71
Figura 44: Uso de la CPU en procesos de escritura y Deduplicación. 76
Figura 45: Uso de la RAM en procesos de escritura y Deduplicación. 76
Figura 46: Uso de la Red txKB por sistema de archivos y tiempo de descarga bajo SDFS y ZFS..... 77

RESUMEN

El presente trabajo de Tesis describe la investigación sobre la Deduplicación de datos, procedimiento de análisis de contenidos que realiza un procedimiento de no almacenamiento de datos cuando estos se encuentren repetidos, las técnicas y métodos basados en Software Libre existentes en el mercado, su funcionamiento en entornos de servidores de almacenamiento, su implementación en un prototipo de repositorio virtual dedicado para el almacenaje de Distribuciones de Sistemas Operativos Linux, repositorio en donde se verifica la Deduplicación, recuperando espacio de almacenamiento en discos duros mediante la eliminación y no almacenamiento de información que se consideren redundantes o duplicados.

Se logró realizar un análisis en tiempo real de los procesos de Deduplicación, los sistemas de archivos para deduplicar se encargan de revisar y analizar la información que contienen cada uno de los procesos, realizando un procedimiento transparente para el usuario, en el mismo busca la información a nivel de archivo comparando contenido y verificando que sea único en el directorio, al existir duplicidad físicamente no almacena el archivo en el disco, pero virtualmente en el entorno del usuario se muestra tal y como si el archivo este almacenándose con normalidad.

Para llevar a cabo esta investigación práctica se utilizó un computador portátil el mismo que será considerado servidor de nuestras máquinas virtuales que poseen las mismas características virtuales en cuanto a hardware y software.

Cada uno de los sistemas de archivos analizados tienen un funcionamiento similar, la Deduplicación, sin embargo uno de los 3 sistemas de archivos (File System) escogidos para su análisis no puede ser instalado debido a problemas de compatibilidad, en primera instancia con sus paquetes y librerías de dependencia, así también con las versiones de Java y finalmente el problema más relevante recae en el Sistema Operativo usado para el proyecto, CentOS.

Los 2 restantes seleccionados para el análisis, SDFS y ZFS, son instalados, configurados y puestos en marcha. Su instalación conlleva a analizar los requerimientos técnicos tales como librerías, paquetes dependientes, los mismos que son descargados e instalados para no tener complicaciones durante la instalación y configuración de los FS. Cada una de los

servidores virtuales ejecuta a su vez servicios tales como HTTP y FTP, los que son usados para el funcionamiento del prototipo de Mirror.

Las pruebas de funcionamiento fueron satisfactorias, SDFS y ZFS llegaron a funcionar en los servidores, cada una realizando su trabajo de Deduplicación que es similar, sin embargo las velocidades de lectura, escritura y descarga son las que definen las diferencias entre cada uno de ellos, ZFS por un lado realiza una Deduplicación en tiempo real más rápida que la de su rival SDFS, los márgenes de tiempo a pesar de que son cortos nos da una idea de que herramienta es la mejor para ser aplicada a futuro en los repositorios virtuales de SO Linux instalados a la Facultad de Ingeniería en Sistemas Electrónica e Industrial de la Universidad Técnica de Ambato.

ABSTRACT

This thesis work describes a research on data deduplication, content analysis method that takes a no-data storage when they are repeated, techniques and based on existing Software Libre in the market methods, their operation storage server environments, its implementation in a prototype dedicated virtual repository for storing Linux operating system distributions, repository where deduplication takes place, recovering storage space by removing hard drives and no storage files deemed redundant or duplicated.

It was achieved to analyze real-time systems to deduplicate files are responsible for reviewing and analyzing the information contained in each of the processes, making a transparent process for the user, on the same level information searches Photo comparing and verifying content that is unique to the directory, to be physically duplicity does not store the file on disk, but virtually in the user environment shown as if the file is stored normally.

A laptop was used, these equipment will be considered server of virtual machines that have the same characteristics in terms of virtual hardware and software to perform the fieldwork.

Each file systems analyzed have similar performance, deduplication, but nevertheless 1 of the 3 FS chosen for analysis can't be installed because of compatibility issues in the first instance with their packages and libraries dependence well with versions of Java and finally the most important problem lies with the operating system used for the project, CentOS.

The remaining 2 selected for analysis, SDFS and ZFS are installed, configured and implemented. Its installation leads to analyze the technical requirements such as libraries, dependent packages, etc., they are downloaded and installed to avoid complications during the installation and configuration of the FS. Each of the virtual servers running at the same time services such as HTTP and FTP, which are used for the operation of our prototype Mirror.

Performance tests were satisfactory, SDFS and ZFS came to work on servers, each doing their Deduplication jobs is similar, however the speeds of reading, writing and download are those that define the differences between each of them, ZFS makes a deduplication faster real time than its rival SDFS, the timeframes although they are short gives us an idea of which tool is the best to apply to future virtual Linux OS repositories installed at the Faculty of Engineering in Systems Electronics and Industrial at the Technical University of Ambato.

GLOSARIO DE TÉRMINOS Y ACRÓNIMOS

Máquina virtual: es un software que simula a una computadora y puede ejecutar programas como si fuese una computadora real.

Servidor: es una aplicación en ejecución (software) capaz de atender las peticiones de un cliente y devolverle una respuesta en concordancia. Los servidores se pueden ejecutar en cualquier tipo de computadora, incluso en computadoras dedicadas a las cuales se les conoce individualmente como "el servidor".

Deduplicación: técnica de respaldo que elimina los datos redundantes almacenados, guardando una única copia idéntica de los datos, y reemplazando las copias redundantes por indicadores que apuntan a esa única copia.

Servidor HTTP Apache: es un servidor web HTTP de código abierto, para plataformas Unix (BSD, GNU/Linux, etc.), Microsoft Windows, Macintosh y otras, que implementa el protocolo HTTP/1.1 y la noción de sitio virtual.

HTTP: HyperText Transfer Protocol (Protocolo de transferencia de hipertexto) es el método más común de intercambio de información en la world wide web, el método mediante el cual se transfieren las páginas web a un ordenador.

FTP: File Transfer Protocol, "Protocolo de Transferencia de Archivos", es un protocolo de red para la transferencia de archivos entre sistemas conectados a una red TCP (Transmission Control Protocol), basado en la arquitectura cliente-servidor. Desde un equipo cliente se puede conectar a un servidor para descargar archivos desde él o para enviarle archivos, independientemente del sistema operativo utilizado en cada equipo.

Virtualización: es la creación -a través de software- de una versión virtual de algún recurso tecnológico, como puede ser una plataforma de hardware, un sistema operativo, un dispositivo de almacenamiento u otros recursos de red.

Sistema de Archivo: frecuentemente, FS del inglés **F**ile **S**ystem: es el componente del sistema operativo encargado de administrar y facilitar el uso de las memorias periféricas, ya sean secundarias o terciarias.

Repositorio: depósito o **archivo** es un sitio centralizado donde se almacena y mantiene información digital, habitualmente bases de datos o archivos informáticos.

Distros: es una distribución de software basada en el núcleo Linux que incluye determinados paquetes de software para satisfacer las necesidades de un grupo específico de usuarios, dando así origen a ediciones domésticas, empresariales y para servidores. Por lo general están compuestas, total o mayoritariamente, de software libre, aunque a menudo incorporan aplicaciones o controladores propietarios.

Firewall: es una parte de un sistema o una red que está diseñada para bloquear el acceso no autorizado, permitiendo al mismo tiempo comunicaciones autorizadas.

Puerto: es una interfaz a través de la cual se pueden enviar y recibir los diferentes tipos de datos.

Directorio: es un contenedor virtual en el que se almacenan una agrupación de archivos informáticos y otros subdirectorios, atendiendo a su contenido, a su propósito o a cualquier criterio que decida el usuario. Técnicamente, el directorio almacena información acerca de los archivos que contiene: como los atributos de los archivos o dónde se encuentran físicamente en el dispositivo de almacenamiento.

Root: es el nombre convencional de la cuenta de usuario que posee todos los derechos en todos los modos (mono o multi-usuario). Normalmente esta es la cuenta de administrador.

Montar: acción de integrar un sistema de archivos alojado en un determinado dispositivo dentro del árbol de directorios de un sistema operativo.

Archivo comprimido: es el resultado de tratar un archivo, documento, carpeta, etc., con un programa específico para comprimir, cuyo objetivo principal es reducir su peso para que ocupe menos espacio**,** pero con este proceso no perdemos la información original.

Paquete: es una serie de programas que se distribuyen conjuntamente. Algunas de las razones suelen ser que el funcionamiento de cada uno complementa a o requiere de otros, además de que sus objetivos están relacionados como estrategia de mercadotecnia.

Mirror: es un sitio web que contiene una réplica exacta de otro. Estas réplicas u *espejos* se suelen crear para facilitar descargas grandes y facilitar el acceso a la información aun cuando haya fallos en el servicio del servidor principal.

SO.: frecuentemente, OS del inglés **O**perating **S**ystem, es un programa o conjunto de programas de un sistema informático que gestiona los recursos de hardware y provee servicios a los programas de aplicación, ejecutándose en modo privilegiado respecto de los restantes.

GNU: es un sistema operativo de tipo Unix desarrollado por y para el Proyecto GNU y auspiciado por la Free Software Foundation. Está formado en su totalidad por software libre, mayoritariamente bajo términos de copyleft. *GNU* es un acrónimo recursivo de "GNU"s Not Unix" (en español: GNU no es Unix), elegido porque GNU sigue un diseño tipo Unix y se mantiene compatible con éste, pero se diferencia de Unix en que es software libre y que no contiene código de Unix.

GNOME: es un entorno de escritorio e infraestructura de desarrollo para sistemas operativos GNU/Linux, Unix y derivados Unix como, BSD o Solaris; compuesto enteramente de software libre.

Iptables: es un espacio de usuario del programa de aplicación que permite a un administrador del sistema para configurar las tablas proporcionadas por el núcleo de Linux firewall (implementado como diferentes Netfilter módulos) y las cadenas y lo gobierna tiendas. Los diferentes módulos y programas del núcleo se utilizan actualmente para diferentes protocolos; *iptables* aplica a IPv4, ip6tables a IPv6, arptables de ARP , y ebtables de tramas Ethernet.

SElinux: Security-Enhanced Linux es un módulo de seguridad para el kernel Linux que proporciona el mecanismo para soportar políticas de seguridad para el control de acceso, incluyendo controles de acceso obligatorios como los del Departamento de Defensa de Estados Unidos.

Kernel: es un software que constituye una parte fundamental del sistema operativo, y se define como la parte que se ejecuta en modo privilegiado (conocido también como modo núcleo). Es el principal responsable de facilitar a los distintos programas acceso seguro al hardware de la computadora o en forma básica, es el encargado de gestionar recursos, a través de servicios de llamada al sistema.

INTRODUCCIÓN

El presente Trabajo Estructurado de Manera Independiente denominado: "**ANÁLISIS DE MÉTODOS DE DEDUPLICACIÓN DE DATOS APLICADOS EN REPOSITORIOS LINUX PARA LA FACULTAD DE INGENIERÍA EN SISTEMAS ELECTRÓNICA E INDUSTRIAL**", está compuesto por los siguientes capítulos:

CAPÍTULO I: "EL PROBLEMA", en donde se detalla la situación actual y la importancia de los Mirror de Distros Linux y la manera de cómo podrá optimizar y ahorrar espacios de disco de almacenamiento de los mismo aplicando técnicas de Deduplicación adecuadas. Además se detalla la justicación y los objetivos que se van a cumplir en la presente investigación.

CAPÍTULO II: "MARCO TEÓRICO", muestra las investigaciones previas que sirven de soporte para el desarrollo de la investigación, información encontrada acerca de las técnicas de Deduplicación, además de la sustentación de la importancia del uso de software libre en el desarrollo de la investigación, finalmente para llegar a determinar los FS que serán analizados y obtención de resultados.

CAPÍTULO III: "METODOLOGÍA", define el tipo de investigación que ha sido desarrollada, el tratamiento de los procesos que señala la modalidad de investigación, así mismo se presenta el tipo de análisis de los datos según el tipo de investigación.

CAPÍTULO IV: "DESARROLLO DE LA PROPUESTA", describe el tema investigado en donde se detalla procesos y métodos de instalación y la manera de configurar cada una de las herramientas, además de la puesta en funcionamiento y las pruebas realizadas en cada uno de los aspectos definidos para la investigación.

CAPÍTULO V: "CONCLUSIONES Y RECOMENDACIONES", expone de forma clara y concisa las consideraciones más relevantes que se han obtenido al realizar el proyecto, además se indican recomendaciones que servirán de apoyo para el desarrollo del mismo y futuras implementaciones de ser el caso.

CAPÍTULO I

EL PROBLEMA

1.1 Tema

“ANÁLISIS DE MÉTODOS DE DEDUPLICACIÓN DE DATOS APLICADOS EN REPOSITORIOS LINUX PARA LA FACULTAD DE INGENIERÍA EN SISTEMAS ELECTRÓNICA E INDUSTRIAL”.

1.2 Planteamiento del problema

Las distros de software libre tienden a crecer en espacio en disco, varias de estas se han convertido en los SO instalados por defecto de cada uno de los ordenadores que se comercializan [1], además gracias a la importancia que hoy en día se ha dado al uso del Software Libre siendo utilizados muy a menudo, no solo como software de investigación sobre su funcionamiento, sino como software de uso diario dentro de cada institución educativa como comercial.

Estos SO han tenido acogida desde hace algún tiempo atrás, lo que ha permitido que estas distribuciones tengan varios formatos de descarga: archivos comprimidos, imágenes de disco, paquetes de instalación y hasta cada uno de los datos o archivos que conforman el paquete de instalación estén disponibles por separado, sin embargo estas bondades de descarga son las que crearían redundancia de información trayendo problemas que a simple vista no tendrían mayor inconveniente, pero que con el paso del tiempo, pueden recaer en grandes dificultades.

Las grandes cantidades de información almacenadas en los distintos Repositorios de software libre han permitido que la capacidad almacenamiento de los discos duros se reduzcan de manera acelerada, lo que conlleva muchos aspectos positivos y negativos según sea la perspectiva [2].

El incremento de costos no previstos por la adquisición de nuevos dispositivos de almacenamiento puede causar problemas en cualquier ámbito ya sea comercial, empresarial o institucional, así también la mala administración de los recursos tecnológicos, específicamente de Red, podrían llevar a un consumo inadecuado de los anchos de banda gracias al transportamiento de información que puede ser repetida, llegando a un consumo imprevisto y exagerado de recursos informáticos.

La sobrecarga de métodos y procedimientos por la disminución acelerada de espacio de almacenamiento dentro de un entorno informático puede ser perjudicial, la reducción acelerada de espacio de almacenamiento produce defectos en cuanto al rendimiento de los ordenadores, llevando problemas desde el procesamiento de datos, pasando por una disminución de memoria acelerada hasta la trasmisión de la información través de la red, ocasionando con estos niveles bajos de fluidez de datos a través de una red que a pesar de tener todas las características necesarias para un funcionamiento óptimo, sea considerada como subutilizada.

La complejidad de conocer la información que se considera repetida dentro de un espacio de almacenamiento de un ordenador podría llevar a un desorden total en almacenamiento y administración de datos dentro de un servidor, por lo cual para poder ser considerada como información que “sobra” esta debería pasar un número determinado de procesos y métodos que detecten que en verdad podría considerarse duplicada y procederá un proceso de eliminación responsable, esto generando la posibilidad de recuperar espacio de almacenamiento.

La Universidad Técnica de Ambato se ha convertido en una de las primeras Academias en utilizar en la mayoría de sus Carreras software Libre, teniendo distribuciones de Linux como Sistemas Operativos, y herramientas de Ofimática como Libre Office, aunque el

uso de dicho Software es a diario dentro de nuestras aulas clase, aún no se ha implementado la opción de que exista un repositorio virtual local lo que conlleva a que se utilicen redes externas a la Universidad para la obtención del material necesario.

En Ecuador la implantación de Repositorios de Software Libre en este caso de distribuciones, se podría considerar como nuevo, existiendo organizaciones como el Consorcio Ecuatoriano para el Desarrollo de Internet Avanzado (CEDIA) y en el ámbito educativo la Escuela Superior Politécnica del Chimborazo (ESPOCH) quienes han sido los pioneros en la puesta en marcha de Repositorios, y en este caso enfocándonos a los Mirror's de varias Distribuciones de Software Libre contribuyendo con recursos de almacenamiento y red al momento de realizar descargas y actualizaciones del software a nivel local y regional [3].

1.3 Delimitación

Área Académica: Hardware y Redes.

Línea de Investigación: Tecnologías de la Información.

Sub líneas de Investigación: Redes de Computadoras.

Delimitación Espacial:

La presente investigación se desarrollará en la Facultad de Ingeniería en Sistemas Electrónica e Industrial de la Universidad Técnica de Ambato.

Delimitación Temporal:

La presente investigación se desarrollará en los 6 meses posteriores a la aprobación del proyecto por parte del H. Consejo Directivo.

1.4 Justificación

Las técnicas de Deduplicación son usadas por organizaciones quienes ven en el almacenaje ordenado de la información una manera óptima de mejorar procesos de búsqueda, a pesar de ser una técnica extremadamente confiable, aún no se la ha puesto funcionar dentro de Repositorios Virtuales de Distribuciones de SO basados en Linux.

Acorde a la Seguridad de la Información como son los respaldos, copias de seguridad, entre otros, se puede ver a la técnica de Deduplicación como una alternativa positiva, ya que desde este punto de vista, el duplicar la información es beneficiosa ya que si dicha

información es vulnerable, ya sea a ataques cibernéticos, robos, destrucción de datos o daños relativos al hardware de almacenamiento (Discos Duros), se vería positivo el realizar una copia o duplicado de la información permitiendo realizar recuperación de datos, asegurando la integridad de la misma, sin embargo desde el punto de vista de espacio de almacenamiento, es conocido que el realizar copias de un mismo archivo sin tener una lógica o política de seguridad, producen efectos negativos en consumo de recursos computacionales, la reducción significativa del espacio del almacenamiento conlleva a la adquisición e instalación de nuevo hardware de almacenamiento llevando a un gasto económico extra que a largo plazo podría convertirse en un problema dentro de la organización.

Dependiendo de los aspectos analizados anteriormente, en el presente proyecto de investigación, se tratará de llevar un enfoque directo hacia la posibilidad de reducir el espacio de almacenamiento de archivos que se consideren repetidos dentro de un Mirror de distros GNU/Linux implantado en la Facultad de Ingeniería en Sistemas Electrónica e Industrial y analizar las ventajas o desventajas que puede darse con la aplicación de Técnicas de Deduplicación, sin embargo estas técnicas deberán ser aplicadas y analizadas acorde a políticas de seguridad en cuanto a recuperación de datos se trate.

Actualmente la FISEI posee una red de comunicaciones de excelente rendimiento, que a pesar de que funciona correctamente, está catalogado como subutilizado, la transmisión de datos y de información por este medio aún es mínima para su carga total, lo que nos lleva a pensar en nuevas maneras de utilizar estas capacidades que se encuentran desperdiciadas, y es por esta razón que se ha pensado en Implantar un Mirror de distros GNU/Linux, el mismo que funcionará como fuente de descarga y actualización local de software además de poseer de una técnica de Deduplicación apropiada para poder optimizarlo de mejor manera permitiendo ahorros no solo en el ámbito de mantenimiento lógico y físico del Repositorio sino también beneficioso en cuanto a la reducción en adquisiciones de hardware innecesarias.

En la Facultad de Ingeniería en Sistemas Electrónica e Industrial la implantación de un repositorio Linux tendrá impacto no solo a nivel regional, sino a nivel nacional, sin

embargo dicha implantación tendrá que ser supervisada desde su inicio, la selección de los SO, que ofrecerá deberá estar ligada con los métodos de Deduplicación para que no existan datos redundantes generando un ahorro en cuanto a recursos informáticos se refiera.

La utilización de una técnica de Deduplicación será beneficiosa desde cualquier ámbito informático, no solo dentro de SO. Linux sino también en cualquier rama informática en donde la optimización de recursos informáticos y almacenamiento responsable y confiable sean de gran importancia.

La investigación podrá ser llevada a cabo ya que a pesar de no contar con Libros Físicos sobre el tema, existe gran cantidad de Información almacenada en la Red (Internet), que servirá como fuente de investigación, de igual manera la creación de un Prototipo de Mirror's de SO Linux podrá ser realizada ya que se posee los equipos necesarios para su implantación y funcionamiento.

1.5 Objetivos

1.5.1 Objetivo General

- Analizar los métodos de Deduplicación de datos aplicables en repositorios Linux a ser implantados en la Facultad de Ingeniería en Sistemas Electrónica e Industrial.

1.5.2 Objetivos Específicos

- Analizar la situación Actual de la Facultad de Ingeniería en Sistemas Electrónica e Industrial en cuanto a usos de Servidores de Almacenamiento de Información y la disponibilidad para implantar un Mirror Dedicado de Sistemas Operativos basados en Linux.

- Estudiar y determinar las técnicas de Deduplicación y algoritmos utilizados en el funcionamiento de cada una de ellas que sean aplicables a una distribución Linux.

- Crear un Prototipo de Mirror con la técnica de Deduplicación más adecuada.

CAPÍTULO II

MARCO TEÓRICO

2.1 Antecedentes Investigativos

Debido al enorme valor que la información albergada en los sistemas puede llegar a poseer, se hace necesario establecer una serie de técnicas y disciplinas orientadas a la ejecución de operaciones de copia de seguridad de datos que permitan recuperar la información en caso de ocurrir alguna contingencia que afecte a su disponibilidad por mal funcionamiento del almacenamiento donde se aloja o por algún error operacional humano o software [4].

A pesar que la tecnología de Deduplicación lleva varios años en el entorno informatico, tiene vocación de permanencia y está evolucionando con rapidez, así logrará ser realmente efectiva en el mundo empresarial de hoy, como en el del futuro, las organizaciones deben buscar soluciones que se adhieran a tres principios básicos: ser transparentes al usuario final y a las aplicaciones; ser sensibles a industrias específicas, con más y mejores algoritmos y principios lógicos; y proyectarse sobre todo el ciclo de vida de los datos para garantizar la optimización de toda la infraestructura tecnológica [5].

La mayoría de productos de Deduplicación se han diseñado y se venden como soluciones de software/hardware combinados. La mayor parte de los casos el hardware por sí solo ha sido difícil de justificar debido a su alto costo. Para ilustrar esto último, según Acronis, organización enfocada al almacenamiento y copias de respaldo, el hecho que un proveedor conocido redujo el costo de uno de sus dispositivos de Deduplicación de datos

de gama alta en marzo de 2009 por más de un tercio. Pero en unos $ 130,000 12 TB de capacidad de almacenamiento, todavía es una propuesta costosa. Obstáculos como éstos han limitado las promesas de Deduplicación a la mayor de las organizaciones [6].

La información que se genera debe ser almacenada, manteniendo estándares de control de seguridad, copias de respaldo y cifrado de datos si lo amerita, este almacenamiento hoy en día se está volviendo común en las organizaciones que generan software, como son las diversas Distribuciones de SO libre que se basan en Linux, en teoría se podría decir que cada versión nueva que sale de una anterior contiene nuevos datos, nuevos códigos, si bien es cierto existe cambio de datos en este caso líneas de código que modifican aspectos en las distribuciones, apariencia, seguridad del Sistema Operativo, sin embargo cabe mencionar que no todas las características que traen consigo las nuevas versiones son recientes o creadas desde cero, muchas de ellas posean características iguales a las de sus versiones predecesoras, y al ser similares cabe una pregunta: ¿Es recomendable tener la información "repetida" consumiendo espacio de almacenamiento y recursos computacionales?, la misma que responderemos durante el desarrollo del tema [7].

Las versiones como: "Live" ya sea CD o DVD, paquetes de instalación o archivos comprimidos, etc., de los Sistemas Operativos con licencias GNU son hoy en día muy utilizadas no solo por su facilidad de descarga, sino también porque cada una de estas versiones tiene sus "ventajas" según sea el caso, desde el tamaño de archivos de descarga así como su uso final, sin embargo esta variedad conlleva a la mayor redundancia de información que se puede encontrar en los Repositorios en donde se encuentran alojadas. [8] Cada una de estas versiones contiene información que se puede considerar repetida, desde los mismos archivos de configuración, kernels, e instalación que muy a menudo son los mismos y forman parte distinta una vez que están ubicadas en cada instalador por separado, permitiendo que se genere un problema de duplicación [9].

2.2 Fundamentación teórica

2.2.3 Deduplicación

La Deduplicación de datos busca la redundancia de secuencias de bytes en ventanas de comparación de gran tamaño. Las secuencias de datos (de más de 8 KB de longitud) se comparan con el historial de otras semejantes. Se hace referencia a la primera versión almacenada de forma exclusiva de una secuencia, en vez de almacenarla de nuevo. Este proceso queda completamente oculto para los usuarios y las aplicaciones, de modo que todo el archivo es legible después de su escritura (Figura 1) [7].

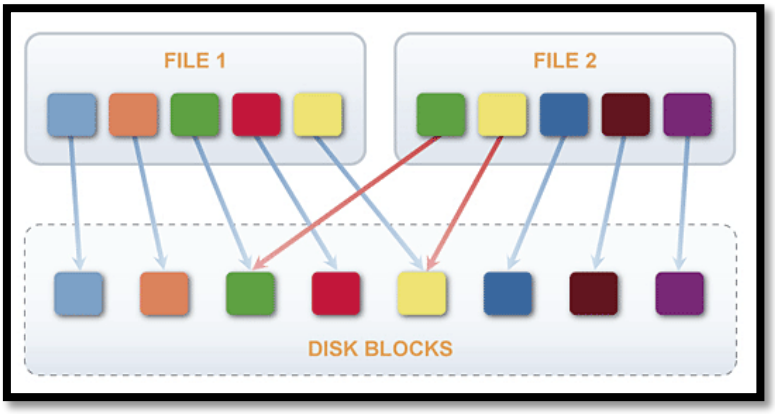

Figura 1: Gráfica de Almacenamiento usando Deduplicación.

Fuente: http://elastic-security.com/2013/12/10/cloudedup-secure-deduplication/ - Pasquale Puzio – Elastic Security.

Según el Organismo Internacional SNIA (Storage Networking Industry Association) define Deduplicación como el proceso de identificación de dato redundante en un volumen de información y su posterior sustitución por referencias a instancias únicas [10].

Independientemente de la técnica subyacente, todos los sistemas de Deduplicación constan de los siguientes componentes:

- Repositorio. El repositorio es el conjunto de segmentos únicos de datos. Se aloja en disco.

- Índice. El índice es un mapa binario de la localización en el repositorio de aquellos segmentos únicos, también nombrado como tabla lookup [10].

La Deduplicación de datos identifica los datos duplicados, elimina las redundancias y reduce el volumen global de datos transferidos y almacenados [11].

2.2.4 Funcionamiento de la Deduplicación de datos

La Deduplicación de datos segmenta un flujo de datos entrante, identifica los segmentos de datos de manera exclusiva y, luego, los compara con los datos almacenados anteriormente. Si el segmento es único, se almacena en el disco. Sin embargo, si un segmento de datos entrante es un duplicado de uno ya almacenado, se crea una referencia a este y el segmento no se almacena nuevamente (Figura 2).

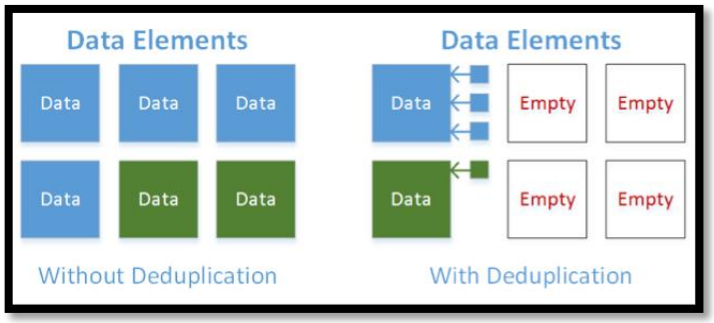

Figura 2: Funcionamiento de la Deduplicación de datos.

Fuente: http://www.enterprisestorageguide.com/how-data-deduplication-works - Scott D. Lowe. Enterprise Storage Guide.

Por ejemplo, un archivo o un volumen del que se hace un respaldo todas las semanas crea una cantidad significativa de datos duplicados. Los algoritmos de Deduplicación analizan los datos y almacenan solo los segmentos exclusivos comprimidos de un archivo. Este proceso puede reducir los requisitos de capacidad de almacenamiento en un promedio de 10 a 30 veces, en un contexto de políticas de retención de respaldo estándares para datos empresariales normales. Esto significa que las empresas pueden almacenar de 10 TB a 30 TB de datos de respaldo en 1 TB de capacidad física de disco, lo que proporciona enormes beneficios económicos [12].

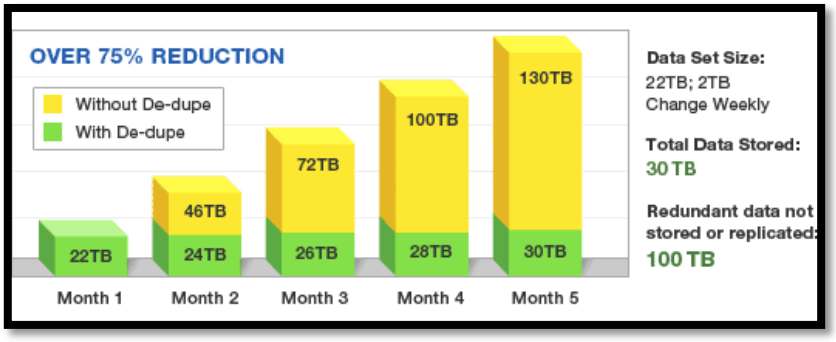

Figura 3: Reducción de espacio al usar técnicas de Deduplicación.

Fuente: http://www.reliant-technology.com/services/deduplication/ - Reliant Technology.

2.2.5 Tipos de duplicación de datos

Enfoques por bloque

La Deduplicación por bloques de datos segmenta las corrientes de datos en bloques, e inspecciona los bloques para determinar si ya han aparecido antes, generando una huella digital o un identificador único a través de un algoritmo de Hash. Si el bloque es único, se escribe en el disco y su identificador se incluye en un índice; de lo contrario sólo se almacena un indicador que remite al bloque original. Al sustituir los bloques repetidos por indicadores mucho más pequeños en lugar de volver a guardar el bloque, se ahorra espacio (Figura 4) [13].

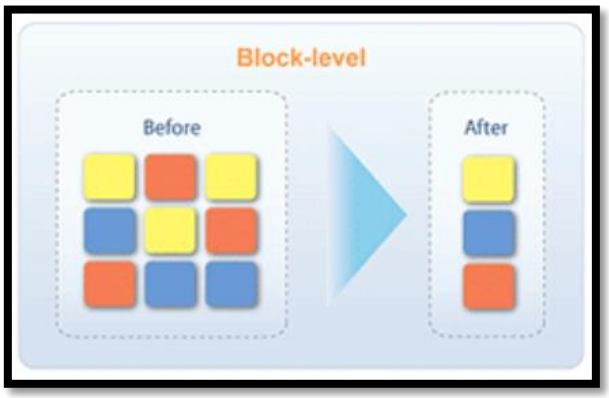

Figura 4: Deduplicación Nivel de Bloque.

Fuente: https://www.starwindsoftware.com/file-level-vs-block-level-vs-byte-level-deduplication - Star Wind Software.

La Deduplicación a nivel de bloque tiene algo de mayor sobrecarga que la Deduplicación a nivel de archivo cuando se duplican los archivos enteros, pero a diferencia de Deduplicación a nivel de archivo, que maneja los datos a nivel de bloque, tales como imágenes de máquinas virtuales extremadamente bien. La mayor parte de la imagen de una máquina virtual duplica datos, es decir, una copia del sistema operativo invitado pero algunos bloques son únicos para cada máquina virtual. Con Deduplicación a nivel de bloque, sólo los bloques que son únicos para cada máquina virtual consumen espacio de almacenamiento adicional. Todos los otros bloques son compartidos. [14]

Deduplicación por bytes de datos

Otro enfoque de Deduplicación consiste en analizar las corrientes de datos por bytes. Al realizar una comparación byte a byte de las corrientes de datos nuevos con los almacenados previamente, se puede conseguir un mayor nivel de precisión. Los productos de Deduplicación que utilizan este método tienen un rasgo en común: es posible que la corriente de datos de safeguard entrante ya se haya visto antes, de modo que se revisa para ver si coincide con datos similares recibidos con anterioridad (Figura 5) [11].

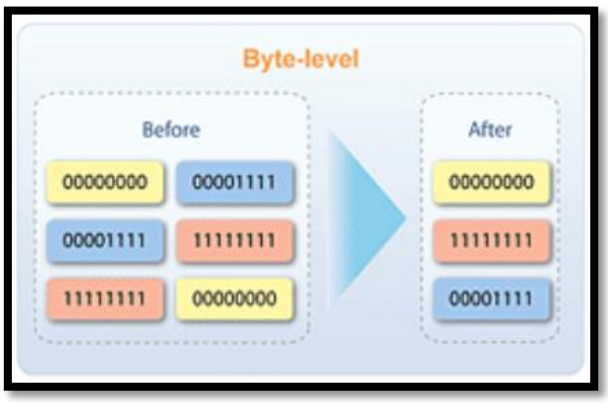

Figura 5: Deduplicación Nivel de Byte.

Fuente: https://www.starwindsoftware.com/file-level-vs-block-level-vs-byte-level-deduplication - Star Wind Software.

A nivel de byte es, en principio, la técnica más general, pero también es la más costosa porque el código a deduplicar debe calcular "puntos de anclaje" para determinar dónde las regiones del duplicado vs. datos únicos comienzan y terminan. Sin embargo, este enfoque es ideal para ciertos servidores de correo, en el que un archivo adjunto puede

aparecer muchas veces, aunque no siempre sea necesario que este en la bandeja de entrada de cada usuario. Este tipo de Deduplicación es generalmente mejor dejar a la aplicación (por ejemplo, servidor de Exchange) que actue, ya que la aplicación entiende los datos es la gestión y puede eliminar fácilmente duplicados internamente en lugar de confiar en el sistema de almacenamiento para encontrarlos después de los hechos. [14]

Deduplicación por Algoritmo

En cuanto al algoritmo matemático subyacente sobre el que se basan, las técnicas de Deduplicación de la actualidad pueden dividirse en:

- Deduplicación usando técnicas basadas en hashing.
- Deduplicación usando técnicas basadas en diferencias binarias (delta differencing).

Deduplicación por el nivel de análisis (Archivo)

Existe una serie de técnicas que buscan redundancia a nivel de ficheros. Estas técnicas analizan el conjunto de datos (filesystems) calculando la clave hash de cada fichero y detectando duplicación cuando detectan una clave hash existente. Esta técnica se conoce también como SIS (Single Instance Store) o almacenamiento de instancia única.

Figura 6: Deduplicación Nivel de Byte.

Fuente: https://www.starwindsoftware.com/file-level-vs-block-level-vs-byte-level-deduplication - Star Wind Software.

A nivel de archivo asigna una firma de hash para un archivo completo permitiendo una menor sobrecarga cuando se efectúa Deduplicación de datos de archivos enteros, pero también tiene importantes limitaciones: cualquier cambio a cualquier bloque en el archivo, debe volver a calcular la suma de control de todo el archivo, lo que significa que si incluso un bloque cambia, cualquier ahorro de espacio se pierde porque las dos versiones del archivo ya no son idénticas. Esto está bien cuando la carga de trabajo constituye archivos como JPEG o MPEG, pero es totalmente ineficaz en la gestión de cosas como imágenes de máquinas virtuales, que son casi idénticas pero difieren en unas pocas cuadras [14]

Otras técnicas de Deduplicación trabajan a nivel inferior al de fichero (subfile) y dividen el conjunto de datos en segmentos de longitud fija o variable pero en cualquier caso a nivel inferior de fichero [15].

2.2.6 Inicios del almacenamiento de la información y su administración

El almacenamiento de la información ha sido el talón de aquiles para cualquier organización que necesite resguardar o respaldar los datos que a diario se generen, lo que conlleva a cuantificar dicha información, y lógicamente necesita grandes capacidades por parte de los dispositivos de almacenamiento para poder ser guardadas y poder disponer de la misma cuando se amerite, sin embargo el almacenamiento a pesar de ser una manera de "respaldo o backup", genera en muchas de las ocasiones una duplicidad de datos que al ser de magnitudes grandes (Gigabytes –Terabytes - Petabytes) [16] se vuelve en el peor de los casos inmanejable por parte de los usuarios o administradores de los centros de datos [17] .

Hoy en día, las redes computacionales con su distribución de contenidos, almacenamiento de copias de seguridad en línea, difusión de noticias, blogs y redes sociales están centradas en esencialmente en datos. Cientos de millones de usuarios de estos servicios generan petabytes de información al día, un ejemplo claro es lo realizado por Dropbox a partir Abril del 2011, quienes al brindar un proceso de intercambio y gestión de archivos, además de servicios de copia de seguridad en línea, generó más de 25 millones de

“dropboxes” (Usuarios que contaban en sus inicios con una cantidad de 2GB de espacio para almacenamiento en línea) lo que al final almacenó un total de 50 petabytes. [16] .

La gran parte de los datos almacenados mediante los servicios de Internet se consideran redundantes por dos razones importantes. En primer lugar, debido a la disminución significativa en el costo de almacenamiento por GB, la gente tiende a almacenar múltiples copias de un único archivo de datos por seguridad o conveniencia. En segundo lugar, mientras que las copias de seguridad de datos incrementales (o diferenciales) o archivos de imagen de disco para escritorios virtuales no suelen ser duplicados en su totalidad del archivo origen, todavía hay una gran parte de la porción de datos que se encuentra duplicada debido a las modificaciones y revisiones que se hacen a dichos archivos [17] .

2.2.7 Sistema Integral de Información

La evolución del almacenamiento de la información está ligada al continuo aparecimiento y evolución de las Tecnologías de Información y Comunicaciones (TIC) las mismas que al trabajar en conjunto integran o centralizan la gestión de la información dentro de una organización, ofreciendo varias herramientas enfocadas a manejar varios procesos, tales como el almacenamiento de datos, sistemas de gestión de conocimiento, gestión documental, entre otros; herramientas que tienen su importancia ya que en años posteriores la información dentro de las organización se encontraba dispersa lo que generaba conflictos al momento de recuperar datos ocasionando duplicidad y desactualización de la misma.

Ante los antecedentes mencionados se deciden crear los Sistemas Integrales de información los mismo que desde sus inicios estuvieron enfocados a mantener la información almacenada de una manera adecuada y en niveles para facilitar su localización y recuperación, posteriormente dichos sistemas incorporarían varias alternativas extras a su funcionamiento inicial, permitiendo crear nuevas versiones de sistemas que Integran Información llegando así a lo que hoy conocemos como herramientas ERP las mismas que han sido desarrolladas para mantener un orden en entornos empresariales.

Breve Historia de la Deduplicación

La Deduplicación de datos como tecnología inicia desde 2007, cuando esta tecnología llega a ser conocida por grandes cantidades de personas y obteniendo una aceptación a nivel mundial, su primera aparición fue en noviembre de 1997, cuando Microsoft lanzó Exchange Server 5.5, tecnología que mantiene en su arquitectura una instancia de almacenamiento individual (SIS) de los mensajes en los que se puede ver la forma original de la Deduplicación de datos, es decir, una especie de fragmentación de archivos como es conocida en la actualidad. [18]. Es también en Microsoft Exchange 2000 y Microsoft Exchange Server 2003, donde se conforma la idea básica de que si se envía un mensaje a un destinatario, y si el mensaje se copia a otros 20 beneficiarios que residen en el mismo almacén de buzones, Exchange Server mantiene sólo una copia del mensaje en su base de datos y dichos mensajes son enviados a los beneficiarios haciendo uso de punteros que son creados por Exchange Server. [19].

Estos punteros enlazan tanto el destinatario original y los 20 destinatarios adicionales al mensaje original. Si el destinatario original y los 20 destinatarios adicionales se mueven a otro almacén de buzones, se mantiene sólo una copia del mensaje en el nuevo almacén de buzones. El nuevo almacén de buzones puede estar en otro servidor en el mismo sitio o en un grupo administrativo. Si el servidor se encuentra en otro sitio, el almacenamiento de instancia única se conserva sólo si se utiliza el asistente "Mover buzón" en Microsoft Exchange Server 2003 Service Pack 1 (SP1) o versiones posteriores. SIS ya ha sido una configuración estándar desde Windows Storage Server2003 R2. [18].

2.2.8 La Deduplicación y su importancia en la actualidad

En los últimos años se han venido analizando técnicas y desarrollando software que permita de una manera rápida, eficaz y segura poder realizar comprobaciones de duplicidad de información, técnicas o métodos que en pocas palabras analizan comparan y eliminan datos duplicados redundantes, permitiendo al final de este proceso almacenar una sola "copia" del archivo en las unidades de almacenamiento.

Los casos de duplicidad de información son muy a menudo desconocidos, ya que se tiene una mala administración al momento de almacenar los datos y generar las copias de los

mismos, y son varios los ejemplos que se pueden dar a conocer acerca de cómo se genera una duplicidad de información sin darnos cuenta, copias de un archivo, correo electrónicos enviados a distintos destinatarios pero con el mismo contenido, versiones de programas almacenadas con distintas fases de desarrollo, etc., que sonaría poco si en tamaño de archivos estaríamos hablando pero que al compararlos con meses y años de realizar las mismas actividades de manera repetitiva podrían llevar a que esos tamaños pequeños de datos se conviertan en cantidades cuestionables si relacionamos a la cantidad con el costo que conlleva el mantener almacenada esa información.

"La Deduplicación es una tecnología que permite un ahorro drástico en cuanto a almacenamiento se refiere. Esta tecnología ha ido implantándose en las soluciones que los fabricantes de almacenamiento y backup ofrecen, de manera que hoy en día es imprescindible plantearse su uso en una estrategia adecuada de backup". [20]

El concepto anterior refleja la importancia que tiene estas técnicas en la actualidad, y por qué estas técnicas en la mayoría de los casos están ligadas a los fabricantes de dispositivos de almacenamiento, sin embargo existen ya en la actualidad grandes cantidades de información que necesita ser depurada y revisada para ver su estado real, si se encuentra o no redundante y la posibilidad de ser llevada a un análisis de Deduplicación.

2.2.9 Compresión de Datos vs. Deduplicación

La compresión de datos o conocida técnicamente como la reducción de velocidad binaria, implica decodificar a la información usando menos bits que la representación inicial. La compresión de datos tradicional es una especie de extensión o aplicación del código Morse`s con la misma idea basándose en que las secuencias de bits más largos pueden ser representados por secuencias más cortas. La idea de esta tecnología es tratar de disminuir el tamaño de algunos archivos mediante la eliminación de los datos redundantes en el archivo sin cambiar el contenido del mismo, ahorrando mucho espacio comprimiéndolos.

La Deduplicación de datos también proporciona más espacio al igual que la compresión, la Deduplicación está relacionada con la compresión en algunos niveles conceptualmente pero en general poseen una diferencia total. La idea de compresión es de modificar

secuencias de bits más largos por más cortos, si cambiamos los bits en un archivo, podemos tener una idea similar de Deduplicación. Así que podemos encontrar fácilmente que en comparación con la tecnología de Deduplicación a la de compresión de datos tradicional, no sólo puede eliminar la redundancia de los datos dentro del archivo, sino también eliminar la redundancia de datos entre los conjuntos de datos compartidos dentro del archivo. Sin embargo, no es totalmente correcto entender que la Deduplicación de datos es sólo para eliminar copias redundantes de archivos, de hecho, hay otras formas de esta tecnología.

Las ideas de cada una de estas tecnologías deben estar bien distinguidas. Así que en la compresión, intenta hacer que el tamaño de un archivo determinado sea más pequeño reemplazando algunas secuencias de bits a algunos otros de menor tamaño sin cambiar el contenido del archivo. En cuanto a la Deduplicación, ahorra espacio de almacenamiento manteniendo la única copia y eliminar todos los que se han visto y almacenado antes, y las copias podríamos hacer referencia a archivos, trozos o bytes, dependiendo del método que está utilizando [21].

La tecnología de compresión es muy fácil de aplicar, pero limitado a conjuntos de datos relativamente cortos. La Deduplicación de datos es más eficaz en la práctica, pero va a llegar a ser difícil de implementar todas las formas ya que definitivamente hay más datos para procesar. Hay que admitir que la Deduplicación ha demostrado sus logros en diferentes tipos de almacenamiento referentes a entornos empresariales y sistemas de copias de seguridad, pero eso no quiere decir que es mejor que la tecnología de compresión, entonces, la mejor opción a usarse dependería de Qué tipo de datos se posee?, si se tiene tantos datos duplicados, la Deduplicación de datos sería la mejor opción, y la compresión sería una respuesta agradable si se tiene un muy buen conocimiento de los datos almacenados.

2.2.10 Parámetros a evaluar antes de realizar una Deduplicación de Información.

Antes de poner en marcha alguna técnica o software específicamente diseñado para Deduplicación de datos, se deben tomar en cuenta varios parámetros, analizarlos y

encontrar las razones para determinar que es necesaria la aplicación de la Deduplicación en cualquier servidor de almacenamiento de datos [22].

Entre los parámetros más importantes a ser analizados destacan los siguientes: software a utilizarse, el hardware que se posee, y la combinación de software y hardware para obtener resultados eficientes después de aplicar las técnicas.

Análisis de Software

Existen aplicaciones listas para realizar acciones de Deduplicación, una de ellas constituye en una Deduplicación Basada en la Fuente, la misma que realiza una compresión de la información en el origen antes de generar una copia de seguridad (backup) [22], aunque esta técnica sería constituida como Deduplicación, aunque es muy práctica puede traer pequeñas sobrecargas en cuanto al uso del CPU y un trabajo acelerado en el Disco Duro.

Técnicamente este software se incrusta en la capa del cliente, servidor o la aplicación para servir como repositorio para los datos de copia de seguridad ubicando una especie de agente de copia de seguridad de peso ligero en el servidor virtual o físico y sólo generando copias de seguridad de los segmentos de datos únicos que han cambiado desde el trabajo previo.

Las ventajas de la Deduplicación basada en la fuente son: Backup rápidos y una gran reducción en el volumen de tráfico entre redes LAN / WAN, en lugar de empujar una copia de seguridad completa a un servidor de medios, los segmentos de dicha copia de seguridad llegan de a poco (después de ser deduplicados) a los hosts de aplicación y la red de almacenamiento de Deduplicación backend. Las transferencias de datos son extremadamente eficientes como incrementales gracias al aumento de velocidad en cada transferencia de información [23].

En general, la Deduplicación basada en fuente es una gran solución para entornos con una baja tasa de cambio de datos diarios.

Sin embargo la Deduplicación en la Fuente no es la única técnica aplicable, algunos de los productos de software con dichas finalidades también realizan Deduplicación en el Destino [22], que no es más que realizar el proceso de evaluacion de la información duplicada en el destino, en lugar del origen.

Deduplicación en el destino permite que no se cambien radicalmente la forma en la que se hacen las copias de seguridad. En lugar de almacenar su copia de seguridad en la misma ubicación origen, pueden simplemente volver a apuntar sus copias de seguridad en un aparato "objetivo" donde toda la Deduplicación de datos se lleva a cabo de una manera segura, no hay necesidad de cambiar las aplicaciones de copia de seguridad, hacer grandes modificaciones a la infraestructura de red subyacente o cambiar los procesos operativos. Los usuarios también tienen la opción de respaldar su información desde el dispositivo de Deduplicación a una biblioteca objetivo. Esto es útil para aquellos que tienen como requisito primordial o estándar el mantener una copia de los datos de copia de seguridad durante largos períodos de tiempo [23].

Algunos proveedores de tecnología han integrado Deduplicación basada en fuente y en Objetivo en sus aplicaciones de copia de seguridad precargadas en sus dispositivos de almacenamiento, sin embargo dichos benéficos que en parte funcionan y de buena manera, no siempre son consideradas como aplicables a todos los entornos en donde se almacenen grandes cantidades de información, lo que las vuelven limitadas, y para esto existen técnicas específicamente diseñadas para poder trabajar en estos ámbitos antes mencionados que brindan resultados eficientes y al mismo tiempo permiten considerar la sustitución de su aplicación de copia de seguridad embebida en el producto por otras técnicas que posean características extendidas para realizar la Deduplicación.

Hardware que se posee

Actualmente los proveedores de sistemas de almacenamiento (servidores dedicados - NAS) ofrecen grandes beneficios al momento de adquirir alguna de sus herramientas, desde grandes capacidades de almacenamiento, estaciones robustas y diseños que se adaptan a cualquier tipo de ambiente ya sea en entornos empresariales o de uso personal, hasta la posibilidad de que dichas herramientas posean técnicas de Deduplicación

previamente instaladas en los dispositivos permitiendo realizar los análisis de copias de seguridad en tiempo real, lógicamente cada una de estas técnicas preinstaladas tienen sus propias características las que las hacen diferentes pero comparten la finalidad del caso [22].

Funcionamiento en conjunto de software y hardware con el fin de deduplicar

A menudo se realiza la pregunta si el usar las técnicas Basadas en software son las más aptas para realizar una Deduplicación, o si las herramientas por medio de Hardware permiten obtener mejores resultados al momento de analizar la información duplicada.

Cada una de las Técnicas mencionadas anteriormente tiene puntos fuertes y débiles, y es por eso que se han visto las maneras de obtener resultados mucho más significativos al unir al Software y al Hardware para crear una solución óptima de Deduplicación.

Una explicación muy válida para fusionar las dos opciones sería al momento de generar copias de seguridad de servidores a través de la WAN, el software de copia de seguridad que realiza la Deduplicación de tipo fuente suele ser la mejor opción, ya que los datos se pueden deduplicar antes de que tenga que atravesar la WAN, lo que disminuye el consumo de ancho de banda del proceso de copia de seguridad; y al ser tratados de manera local, es probable que tenga más sentido llevar a cabo la Deduplicación de destino en un dispositivo de copia de seguridad ,ya que al hacerlo le permitirá a la organización evitar sobrecargar los servidores de producción con discos adicional de E / S y los ciclos de CPU, estos dos esquemas son comúnmente vistos en organizaciones que manejan grandes cantidades de datos, recibiéndolos desde distintos puntos y siendo almacenados en su totalidad como concepto de seguridad [22] [23].

Finalmente la decisión de implantar y hacer uso de cada una de las técnicas dependerá del esquema organizacional de almacenamiento y respaldos de información, las mismas que definirán que es lo más adecuado, a nivel de método de Deduplicación o combinación de métodos.

2.2.11 Deduplicación en Linux

Linux

A partir de la aparición de la Internet, y gracias a la difusión que esta brindaba se empezó una acelerada y amplia difusión de actividades relacionadas al código abierto. El volumen de las contribuciones y la diversidad de contribuyentes expandidos fuertemente, y numerosos nuevos proyectos de código abierto surgieron, especialmente Linux.

Linux es un sistema operativo tipo Unix libre y de código abierto que está diseñado originalmente por Linus Benedict Torvalds en 1991. Linux es un sistema líder que es capaz de ejecutarse en servidores y otras plataformas, como los mainframes y superordenadores.

Linux es ampliamente utilizado en sistemas embebidos, tales como teléfonos inteligentes, computadoras, tablet"s y en equipos de comunicaciones tales como routers, Después de la Licencia Pública General de GNU, cualquier persona puede hacer uso de todo el código fuente subyacente de Linux y modificarlo o distribuirlo libremente. Normalmente, Linux está empaquetado en una distribución Linux para el uso de computadoras personales y servidores, incluyendo Debian (y sus versiones derivadas de Ubuntu, Linux Mint), Fedora (y sus versiones relacionadas de Red Hat Enterprise Linux CentOS) y openSUSE entre otros.

Software Libre

¿Qué es el software libre?

Richard M. Stallman lidera una de las piedras angulares del desarrollo de programas computaciones, el software libre o también conocido como software con licenciamiento GNU [24] se basa en su contexto como sinónimo de libertad, ya sea en su ejecución, su análisis y estudio de funcionamiento, adaptabilidad a cualquier entorno mediante la modificación de su código fuente, distribución del mismo en cualquier ámbito comercial sin fines de lucro y el compromiso de mejora continua de manera que sea beneficioso para la comunidad que lo usa [25].

Actualmente se encuentra regido bajo la Fundación de Software Libre (FSB).

El software no libre también es llamado *software propietario*. El software libre no debe confundirse con el software gratuito. Hay que distinguir entre libre y gratuito, uno no implica lo otro.

Sin embargo se podría realizar una clasificación de este tipo de software enfocándonos en los siguientes aspectos:

- **OSS** son las iniciales de **Open Source Software** (*Software de Código Abierto*).
- **FOSS**, iniciales de **Free Open Source Software** Para referirse al software de código abierto que también es software libre.
- **FLOSS**, que añade la L inicial de *Libre*, creando un juego de palabras al incluir tanto la raíz anglosajona *Free* como los términos latinos de raíz *Libre*.

Reglamentos que definen al Software Libre

La licencia no debe contaminar a otros programas. La licencia no debe poner restricciones sobre otros programas que se distribuyan junto con el software licenciado.

Los distribuidores de software de código abierto tienen derecho a tomar sus propias decisiones acerca de su propio software.

En los últimos años se ha visto un crecimiento sin precedentes de código abierto software. Al mismo tiempo, el movimiento se ha enfrentado a una serie de desafíos, la "bifurcación" de los proyectos (el desarrollo de las variaciones de la competencia) y el desarrollo de productos para usuarios de gama alta. [25].

Una cuestión que ha surgido en una serie de proyectos de código abierto es el potencial que brindan los mismos al momento de “mutar” en diversas variantes.

En algunos casos, las controversias apasionadas sobre el diseño de productos han llevado a la fragmentación de proyectos de código abierto en diferentes variantes.

Beneficios del Software Libre y de Código Abierto

- **Arreglar el software**

Se puede corregir los errores que existen en el software, o se puede modificar el software para adaptarlo a sus necesidades propias, o incluso arreglar los problemas de seguridad que pudieran existir.

- **Compartir**

El software libre te permite compartir el software y así ayudar a sus amigos y vecinos sin necesidad de licencias restrictivas molestas.

- **Conocer y controlar lo que está pasando**

Con el software propietario no se puede saber lo que un determinado programa hace realmente. Algunos programas propietarios muy conocidos han sido sorprendidos espiando a los usuarios y enviando información sobre su comportamiento y otros datos personales. Con el software libre tiene acceso al código fuente y puede estudiar lo que hace el programa y cambiarlo si no es del agrado de usuario final.

- **Beneficios técnicos**

El Software Libre permite a más personas ver el código y arreglarlo, se puede desarrollar más rápido y mejorarlo de una manera más eficiente y rápida. Este sistema de revisión a pares ("peer review" en inglés) se puede comparar con la forma en la que trabaja investigación científica.

- **Beneficios económicos**

El coste económico es un tema que podría ser analizado aparte, ya que el tema lo permite, sin embargo se hará un corto análisis del mismo como beneficio ya que también es una forma en que las empresas pueden compartir los costes de desarrollo. Por ejemplo, Novell y Red Hat son competidores pero ambos gigantes del software

desarrollan muchos programas conjuntamente y por lo tanto se pueden ayudar mutuamente. IBM y HP también podrían ser vistas como competidoras, sin embargo, ambas contribuyen al desarrollo del kernel de Linux, etc, compartiendo así los costes de desarrollo.

2.3 Propuesta de solución

La aplicación de metodologías de Deduplicación de Datos en Repositorios de SO Linux permitirá realizar una administración coherente y oportuna de los recursos informáticos, ya sean procesamiento de datos, optimización de redes, así también el almacenamiento de la información de una manera óptima sin que esta sea redundante.

CAPÍTULO III

METODOLOGÍA

3.1 Tipo de investigación

El presente trabajo será de tipo investigativo y experimental.

3.2 Modalidad

- Se considera la modalidad Bibliográfica de tipo documental debido a que se recurre a diferentes fuentes obtenidas de libros, artículos, Tesis desarrolladas en Universidades para profundizar enfoques con respecto al tema de la investigación, como también el uso de textos, revistas y artículos alojados en Internet.

- También se maneja la modalidad de campo, la que permite realizar una visita al establecimiento para conocer el estado actual y se realizará una entrevista al personal encargado del área a investigar.

- Además la investigación será experimental, de tal manera que se montará un ámbito de pruebas mediante la implantación de un Mirror Virtual, en el cual se realizarán las pruebas necesarias y con eso obtener resultados sobre el tema indicado

3.3 Población y muestra

La presente investigación no requiere de población ni muestra debido a que su realización es un análisis experimental de la duplicidad de datos en los servidores de almacenamiento.

3.4 Recolección de la información

Se recolectará información detallada en artículos de interés alojados en Internet, tales como publicaciones científicas y documentos técnicos especializados en el área.

Además se hace uso de la información obtenida mediante una entrevista realizada a las personas relacionadas con el tema.

3.5 Procesamiento y análisis de datos

Para el procesamiento de la información se realizarán las siguientes actividades:

- Recolección de la información mediante la investigación en documentos electrónicos referentes al tema.
- Revisión de la información recogida.
- Análisis de los datos.
- Lectura de artículos relacionados con la investigación presentada.
- Interpretación de los resultados mediante gráficos, cuadros para analizar e interpretar y por último redactar una síntesis de los resultados.

3.6 Desarrollo del proyecto

Análisis de la situación Actual de la Facultad de Ingeniería en Sistemas Electrónica e Industrial en cuanto a usos de Servidores de Almacenamiento de Información y la disponibilidad para implantar un Mirror Dedicado de Sistemas Operativos basados en Linux.

- Estudio de los mecanismos utilizados en la Facultad para almacenar la Información.
- Análisis de la existencia de servidores de almacenamiento dedicados.

Estudio y determinación de las técnicas de Deduplicación y algoritmos utilizadas en el funcionamiento de cada una de ellas que sean aplicables a una distribución Linux.

- Investigación de la Deduplicación de la Información en Repositorios Linux
- Identificación los tipos de técnicas y metodologías de Deduplicación.
- Selección de una de las distros GNU/Linux existentes.

Creación un Prototipo de Mirror con la técnica de Deduplicación más adecuada.

- Análisis de ventajas y desventajas en el uso de las técnicas.
- Generación de un prototipo de Mirror Virtual de SO Linux.
- Aplicación de la técnica de Deduplicación que mejor se adapte al Prototipo.
- Aplicación de las técnicas en repositorios de Software Libre.
- Obtención y análisis de resultados.

CAPÍTULO IV

DESARROLLO DE LA PROPUESTA

4.1 Análisis de la situación Actual de la Facultad de Ingeniería en Sistemas Electrónica e Industrial.

Mediante una entrevista realizada al, Administrador de Redes y Coordinador de los Laboratorios de Computación de la Facultad, se pudo recabar la siguiente información acerca de la situación actual de la Facultad.

La FISEI cuenta con un servidor HP Proliant DL380 G8, dicho servidor comparte distintas funciones, está configurado bajo CentOS 7 y en el mismo se encuentra funcionando y corriendo varios servidores virtuales para Proxy, DNS, y Enrutamiento, mas no está siendo considerado como servidor dedicado para almacenamiento de información.

El servidor posee 8 TB de espacio de almacenamiento los que están configurados en Raid (0 – 1) que finalmente otorgan un espacio de almacenamiento final de 4 TB, en este espacio se almacena información correspondiente a registros diarios y actividades realizadas por parte de los docentes.

Se indica que en los datos almacenados en el servidor no existe duplicidad de información, y que de haberla, esta es controlada paulatinamente mediante mantenimientos preventivos de los equipos, verificando su funcionamiento y eliminando la información que se considere innecesaria o duplicada, cabe recalcar que este procedimiento de eliminación se lo realiza de forma manual.

Anteriormente en el servidor se ejecutaba un Server FOG, el que integraba en su funcionamiento varios servicios de red (proxy, DHCP, DNS y servicios FTP), el funcionamiento del FOG de alguna manera controlaba la duplicidad de datos, sin embargo este ya no está siendo usado ya que en su lugar se posee servidores independientes para cada servicio de red mencionado.

Los procedimientos y técnicas de deduplicación no son conocidos dentro de la Facultad, sin embargo después de una explicación al acerca del funcionamiento y los beneficios que aporta en los entornos de almacenamiento, el Administrador cree que sería importante el uso de estas técnicas dentro del servidor y considera la Instalación de una de estas técnicas para poder administrar de mejor manera el almacenamiento y contribuir con el ahorro y optimización del espacio en los discos duros de los equipos en general.

En la entrevista con el Administrador, se comentó sobre las distribuciones GNU/LINUX que son más usadas en los equipos de cómputo y por parte de docentes y estudiantes, Fedora, Ubuntu y CentOS, son las distros más usadas, las dos primeras tienen gran acogida por parte de los docentes ya que las mismas manejan un entorno grafico amigable además de su fácil manejo, por otra parte, CentOS es usado por pocos docentes pero es muy solicitado por parte de los estudiantes, ya que es un SO en el cual se aprende a conocer el funcionamiento de los sistemas basados en UNIX, lo que contribuye con su aprendizaje, y es por esta razón por la cual se decide tomar a CentOS como distro para realizar la instalación y configuración de las técnicas de deduplicación.

4.2 Análisis de Técnicas

En el mercado, existen gran cantidad de herramientas diseñadas específicamente para la Deduplicación de datos, aunque varias de ellas tienen en su estructura el mismo o similar código de programación de los algoritmos para deduplicar, son pocas las que muestran más del 99% de efectividad al momento de ser puestas en marcha en el entorno final.
Aquí mostramos un análisis de las herramientas que han sido seleccionadas a ser evaluadas, de las cuales se obtendrán las más aptas para ser aplicadas en nuestro repositorio de SO libres.

Hasta donde se conoce, sólo hay tres sistemas de Deduplicación que están libres, de código abierto y ampliamente utilizados: LESSFS, SDFS, ZFS. En este análisis se va a omitir varios prototipos de investigación ya que suelen ser inadecuados para nuestra necesidad.

LessFS

LessFs es un sistema de archivos de Deduplicación de datos en línea de alto rendimiento por escrito para Linux y actualmente está licenciado bajo la Licencia Pública General GNU versión 3. También es compatible con LZO, QuickLZ y bZIP compresión (entre un par de otros), y el cifrado de datos, además cumple con el estándar POSIX, y es muy útil para realizar copias de seguridad, así como proporcionar almacenamiento para imágenes de máquinas virtuales [26]. Aunque LessFS es un sistema de archivos que se implementa en el espacio de usuario con FUSE, que ofrece un rendimiento decente. LessFS es capaz de manejar velocidades de datos de hasta 350MB / seg, además es compatible con el cifrado del sistema de archivos.

LessFS es una solución práctica para las necesidades de almacenamiento real, se implementa en el espacio de usuario, además es desarrollado bajo lenguaje C.

Un ejemplo de cómo LessFS se puede poner a trabajar es utilizarlo para archivar los datos de la matriz de almacenamiento preferido. Esto es mucho más eficiente utilizando instantáneas o la replicación. Una instantánea requiere al menos 100% de almacenamiento adicional cuando uno quiere estar seguro de que la copia de seguridad

es siempre válido. Si desea que sus datos sean copiados a otro lugar de almacenamiento, su necesidad será de lo menos 105% de almacenamiento adicional. LessFS se puede utilizar para hacer el trabajo mucho más eficiente del espacio. La Deduplicación y compresión que LessFS proporciona comprime los datos de la mayoría de los casos al menos 30% para la primera copia [26]. Cuando se necesita una copia de seguridad diaria LessFS requerirá muy poco espacio extra de almacenamiento para guardar una copia de seguridad completa.

Frescura del Código

Desde su creación en el 2009, LessFS, ha tenido 98 archivos de descarga compartidas entre sus versiones y modificaciones de cada una de ellas, todas, permitiendo tener un balance de frescura de código entre alrededor de semanas y meses, llegando así a la última versión de este software lanzada en noviembre del 2013, acogiendo la versión 1.7.0 [26].

Estas versiones de descarga están disponibles en la página de SourceForge.net, servidor de alojamiento de software muy conocido en el entorno Web y que ha sido escogido desde sus inicios por parte de los programadores de LessFS para la descarga más fácil, segura y rápida a los usuarios finales, el link de descargas es: http://sourceforge.net/projects/lessfs/files/lessfs/.

SDFS

Opendedup SDFS es un sistema de archivos que tiene como característica primordial su funcionamiento de soporte en línea y el trabajo por lotes en sistemas Linux y Windows, junto con entornos virtualizados VMware. Este sistema de archivos afirma que puede reducir la utilización del almacenamiento entre un 90 a 95%, además que es capaz de deduplicar más de un petabyte de datos a una velocidad de más de 1 GB / s, y puede hacer este proceso de deduplicación, ya sea a nivel local, en la red, o en la nube [27]. De hecho, SDFS es particularmente adecuado para la nube bajo entornos virtualizados en VMware, Xen y KVM.

SDFS también es compatible con archivos y carpetas instantáneas. Estas afirmaciones son bastante interesantes, especialmente desde un insólito proyecto de código abierto que en sus inicios fue posteado en comentarios en la red social Twitter [27].

Frescura del Código

El hecho de que sea una herramienta libre y de código abierto permite que su actualización y mejoramiento entre versiones este entre periodos mensuales hasta diarios, en su cuenta en Twitter existen comentarios que datan desde mayo del 2011, dando a conocer en los mismos la herramienta, sus beneficios y técnicas que se utilizan en su funcionamiento, sin embrago su primera versión (0.8.14) data de abril de 2010, teniendo hasta la fecha alrededor de 48 versiones de la aplicación lanzadas en la web para su descarga [27]. Actualmente la herramienta está en su versión 2.0.11 lanzada en marzo del 2015. El link para poder acceder a las mismas es: http://www.opendedup.org/download.

ZFS

ZFS es un combinado de sistema de archivos y gestor de volúmenes lógicos diseñada por Sun Microsystems. Las características de ZFS incluyen la protección contra la corrupción de los datos , el apoyo a grandes capacidades de almacenamiento, la compresión de datos eficiente, la integración de los conceptos de sistema de archivos y administración de volúmenes , instantáneas y clonación en escritura, la integridad continua comprobación y reparación automática, RAID-Z y nativos NFSv4 ACL.

ZFS se implementó originalmente como software de código abierto , licenciado bajo la Common Development and Distribution License (CDDL). El nombre de ZFS se ha registrado como marca comercial de Oracle Corporation.

Una ventaja de copia en escritura es que, cuando ZFS escribe nuevos datos, los bloques que contienen los datos antiguos pueden ser retenidos, lo que permite una instantánea de la versión del sistema de archivos que se mantenga, estas instantáneas se crean muy rápidamente, ya que todos los datos que componen la instantánea ya están almacenados, permitiendo la administración eficiente de los espacios ya que cualquier dato sin cambios

es compartida entre el sistema de archivos y sus instantáneas, sin embrago también se pueden crear instantáneas de datos de manera general y global, dando lugar a dos sistemas de archivos independientes que comparten un conjunto de bloques. A medida que se realizan cambios en cualquiera de los sistemas de archivos clónicos, nuevos bloques de datos se crean para reflejar esos cambios, pero todos los bloques sin cambios continúan siendo compartidos, no importa el cómo sean generados los clones. Se trata de una aplicación de la copia en escritura principio [14].

Otros proveedores de almacenamiento tales como GreenBytes y Tegile utilizan versiones modificadas de ZFS para alcanzar muy altas relaciones de compresión En mayo de 2014, Oracle compró GreenBytes para su Deduplicación ZFS y tecnología de replicación [28].

Frescura del Código

La Deduplicación de datos ha sido añadida al repositorio de fuentes ZFS a finales de octubre de 2009 a pesar de que la herramienta tiene sus inicios como código nativo para Linux (2004), teniendo así alrededor de 38 versiones, cada una editada y mejorad acorde a los sistemas en donde será puesta en marcha. Actualmente los paquetes de desarrollo de OpenSolaris ZFS pertinentes han estado disponibles desde diciembre del 2009, haciendo que su mejoramiento vaya de la mano con el del sistema operativo, permitiendo una mejor eficiencia en el funcionamiento de cada versión nueva lanzada al mercado ya que cuenta con el respaldo y el desarrollo comprobado de Sun ya perteneciente al gigante Oracle [14].

Beneficios de las técnicas de deduplicación

La Deduplicación de datos conlleva algo más que la simple comodidad, puede incrementar la eficiencia de la utilización del espacio en disco hasta un 50%, lo cual es mucho, sobre todo en el caso de grandes empresas.

La eliminación de los datos redundantes puede disminuir significativamente los requisitos de almacenamiento y mejorar la eficiencia del ancho de banda. Dado que el almacenamiento primario se ha abaratado con el tiempo, las empresas suelen almacenar muchas versiones de la misma información, de modo que los nuevos empleados puedan reutilizar el trabajo ya hecho. Algunas operaciones como el respaldo almacenan información extremadamente redundante [12].

La Deduplicación reduce los costos de almacenamiento, ya que se necesitan menos discos. También mejora la recuperación de desastres porque hay muchos menos datos que transferir. Los datos de archivo y respaldo suelen incluir un montón de datos duplicados. Esto crea una cadena de ineficiencias de costos y recursos dentro de la organización [12] [6].

Un sistema de copias de seguridad que utilice la técnica de la Deduplicación guarda sólo una vez el archivo y reemplaza las demás por un enlace a dicho archivo, o un indicador que apunta a esta única copia. Este sistema consigue ahorrar espacio ocupado por las copias de seguridad lo que nos ayudará a ahorrar costes en discos duros y cintas, así como recuperar con mayor rapidez los datos desde la copia [12].

Depende del sistema que tengamos montado podemos optar por distintos tipos de soluciones de Deduplicación:

- Deduplicación en destino o a nivel de fuente de datos a los que estamos aplicando el backup. Con esto ahorramos espacio fundamentalmente en las cintas de seguridad y tiempo en la restauración de las copias.
- Deduplicación en fuente cuando lo aplicamos en el servidor, con lo cual ahorramos espacio también en los discos del servidor y tendremos más ordenados

los datos. Además del ahorro en cintas añadimos también ahorro en ampliación de discos duros en el servidor [12].

Dentro de este segundo nivel se pueden dividir también en el momento en el que se produce el procesamiento de Deduplicación de datos, que lógicamente llevará un tiempo. Hay soluciones que calculan y buscan archivos comunes en segundo plano o fuera de banda, dirigiendo estos cálculos hacia la memoria intermedia y ejecutándose de forma que no interfieren en el proceso del servidor y la copia. Otras soluciones lo realizan en tiempo real, lo que puede provocar cierto retardo.

Todas estas copias de información que ocupan espacio en el disco suelen venir en los paquetes de backups de seguridad, aunque existe distinto software específico para el servidor para evitar estos problemas de duplicación de datos. Utilizando estas técnicas se consiguen ahorros de espacio en torno a 3:1 y 500:1. Estas soluciones son muy recomendables si se realiza backup en la nube donde el espacio es más limitado y caro que en local [29].

4.3 FS Nativo en CentOS vs SDFS y ZFS.

En el mundo de alas distros GNU/Linux existen gran cantidad de sistemas de archivos, en este caso al realizar la practica con CentOS, que trae como FS nativo EXT4, realizamos una tabla comparativa entre este sistema ante los FS aptos para Deduplicación.

Cabe recalcar que en la tabla están siendo tomados parámetros en base a las diferencias en cuanto a mejoras en evolución de reconocimiento de tamaños de discos, además de las ventajas en cuanto a realizar procesos de deduplicación de datos [30] [31].

Tabla 1: Tabla comparativa entre FS EXT4 y SDFS - ZFS

	Sistema de Archivos Normal (Sin DEDUP)	Sistemas de Archivos para DEDUP	
Nombre	**EXT4**	**SDFS**	**ZFS**
Máxima dimensión de archivo	Hasta Exabyte (10^{18}) 16 TB	Sobre el Petabyte (10^{15})	Ilimitada (16 exabytes referencial)
Máximo número de archivos	4 mil millones ($4x10^{9}$) (especificado en el tiempo de creación del sistema de archivos)	N/E	2^{48}
Tamaño máximo del volumen	1 Exabyte	8 TB	16 Exabytes
Compresión de datos transparente	No soportado	Configurable	SI (Nativo)
Deduplicación de Datos	No soportado	Si	SI (Nativo)
Tamaño Maximo de archivo a deduplicar	No soportado	Sobre el Petabyte	Ilimitado
Pools Almacenamiento Dedup	No soportado	Ilimitada	Ilimitada
Instantáneas (Snapshots)	No soportado	Si	SI (Nativo)
Soporte Discos SSD	Configurable	Configurable	SI (Nativo)

Fuente: Elaborado por el investigador.

Como se observa en la tabla anterior (Tabla:1) existen diferencias significativas entre EXT4 con SDFS y ZFS, estas parámetros que resaltan serán significativos y dependerán del uso y los ambientes en el que sean puesto en marcha, pero si el tema principal es la

optimización y correcta administración de espacio de almacenamiento de Discos duros, SDFS y ZFS son los FS que lideran temas de deduplicación.

4.4 Descripción de Hardware y Software.

El desarrollo de este Trabajo de Investigación se realiza utilizando un Computador portátil HP Pavilion dv6 Notebook PC con las siguientes especificaciones (Tabla 2):

Tabla 2: Especificaciones técnicas del equipo de cómputo usado como servidor para montar las máquinas virtuales.

	Marca	Modelo	Velocidad
Procesador	Intel	Core i7 310QM	2.30 Ghz
RAM	ADATA	PC3-12800 DDR3 / 16 GB	800 Mhz
HDD	Seagate	ST750LX003 / 750GB Hybrid	7200 RPM
Coprocesador Video	NVIDIA	GeForce GT 650M	2048 MB GDDR5

Fuente: Elaborado por el investigador.

Para la aplicación de cada una de las Herramientas mencionadas anteriormente, se hizo uso de máquinas virtuales, las que se instalarán y configurarán con las siguientes características (Tabla 3) bajo el software de virtualización VMware en su versión 11.

Tabla 3 : Especificaciones técnicas de las máquinas virtuales.

	Especificaciones
Procesador	4 núcleos
RAM	8 GB
HDD 1	100 GB.
HDD 2	100 GB.

Fuente: Elaborado por el investigador.

Esquema general de virtualización en el servidor y servicios que ejecutan en cada una de las máquinas (Figura 7).

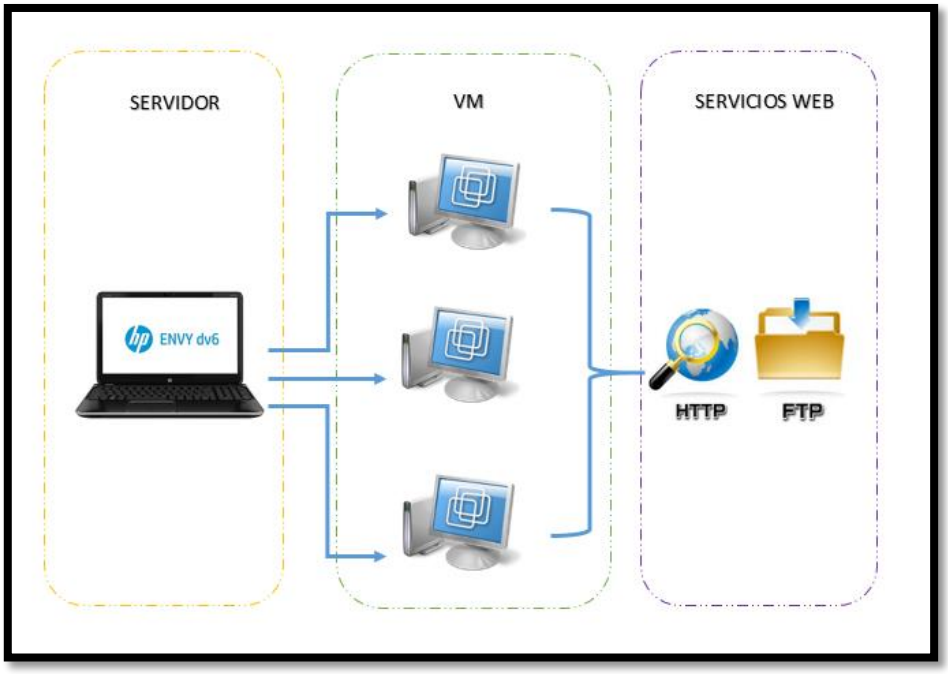

Figura 7: Esquema general de virtualización.
Fuente: Elaborado por el investigador.

En el esquema antes mostrado se instalarán y configurarán 3 máquinas virtuales, las mismas que usarán la Distribución de CentOS como Sistema Operativo Linux, SO que ha sido escogido bajo requerimientos técnicos por parte de la plataforma actualmente instalada en la FISEI, además es el SO. que más uso tiene por parte de docentes y estudiantes en cuanto a investigación y desarrollo se refiere (información recolectada mediante la entrevista realizada al Administrador de red de la Facultad), la Distro a instalarse será en su versión 6.6.

Con el fin de crear un prototipo similar de Mirror de SO Linux, se instala y configura en las máquinas virtuales los siguientes servicios Web: Protocolo de Transferencia de Hipertexto (HTTP) y Protocolo de transferencia de Archivos (FTP), servicios que están siendo usados actualmente el Repositorio de Software Libre de la Facultad de Ingeniería en Sistemas Electrónica e Industrial de la UTA.

4.5 Creación e instalación de máquinas virtuales

Se instala en cada una de las máquinas virtuales el Sistema Operativo CentOS 6.6, realizando una instalación mínima, lo que permitirá que el sistema ejecute solo los servicios y software necesarios para su funcionamiento, mejorando de esta manera el rendimiento de lo que posteriormente serán nuestros servidores dedicados para el Mirror de SO Linux.

Para facilidad de visualización también se instaló un entorno gráfico para nuestro CentOS, existen varias opciones tales como KDE, XFCE, GNOME, escogiendo a GNOME por su instalación y configuración sencilla [32].

Para su instalación debemos estar logueados como usuario con todos los privilegios (root) y ejecutamos los siguientes comandos (Figura 8):

```
yum -y groupinstall "Desktop" "Desktop Platform" "X Window System" "Fonts"
```

```
Transaction Summary
================================================================================
Install     478 Package(s)

Total download size: 286 M
Installed size: 888 M
Downloading Packages:
(1/478): ConsoleKit-0.4.1-3.el6.x86_64.rpm                |  82 kB     00:00
(2/478): ConsoleKit-libs-0.4.1-3.el6.x86_64.rpm           |  17 kB     00:00
(3/478): ConsoleKit-x11-0.4.1-3.el6.x86_64.rpm            |  20 kB     00:00
(4/478): DeviceKit-power-014-3.el6.x86_64.rpm             |  91 kB     00:00
(5/478): GConf2-2.28.0-6.el6.x86_64.rpm                   | 964 kB     00:00
(6/478): GConf2-gtk-2.28.0-6.el6.x86_64.rpm               |  23 kB     00:00
(7/478): ModemManager-0.4.0-5.git20100628.el6.x86_64.rpm  | 177 kB     00:00
(8/478): NetworkManager-0.8.1-66.el6.x86_64.rpm           | 1.1 MB     00:00
(9/478): NetworkManager-glib-0.8.1-66.el6.x86_64.rpm      | 230 kB     00:00
(10/478): NetworkManager-gnome-0.8.1-66.el6.x86_64.rpm    | 463 kB     00:00
(11/478): ORBit2-2.14.17-3.2.el6_3.x86_64.rpm             | 168 kB     00:00
(12/478): PackageKit-0.5.8-21.el6.x86_64.rpm              | 526 kB     00:00
(13/478): PackageKit-device-rebind-0.5.8-21.el6.x86_64.r  |  95 kB     00:00
(14/478): PackageKit-glib-0.5.8-21.el6.x86_64.rpm         | 221 kB     00:00
(15/478): PackageKit-gtk-module-0.5.8-21.el6.x86_64.rpm   |  95 kB     00:00
```

Figura 8: Descarga de paquetes para entorno gráfico GNOME.
Fuente: Elaborado por el investigador.

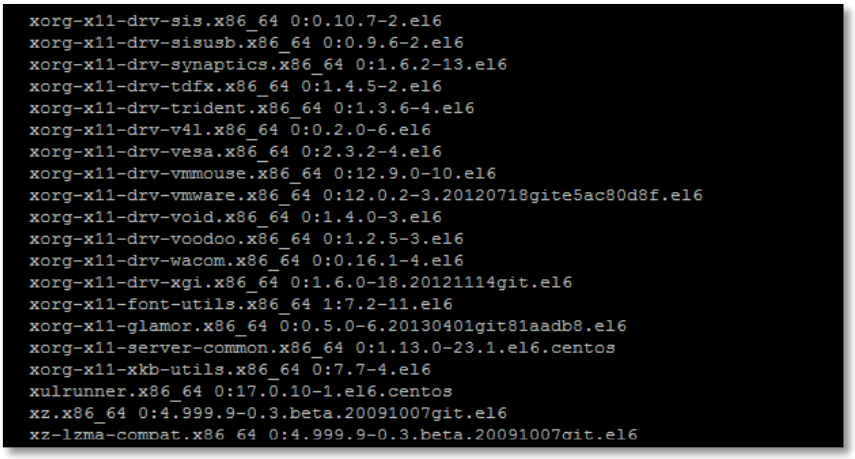

Figura 9: Instalación de paquetes para entorno gráfico GNOME.
Fuente: Elaborado por el investigador.

Esto permite descargar e instalar los paquetes necesarios para el funcionamiento del entorno de escritorio GNOME (Figura 8) & (Figura 9).

Una vez instalado el entorno, se configura el archivo inittab, el mismo que está dentro del directorio raíz en la carpeta **`/etc/.`**

```
vi /etc/inittab
```

En el archivo inittab en donde cambiaremos el valor de 3 por 5 en la siguiente línea:

```
id:3:initdefault:
```

por

```
id:5:initdefault:
```

Permitiendo que después de un reinicio de la máquina nuestro entorno gráfico GNOME se ejecute al iniciar nuestro servidor no solo bajo el usuario root sino a nivel multiusuario.

Posteriormente se realiza una configuración de usuarios, pero en este caso haciendo uso del entorno gráfico GNOME (Figura 10).

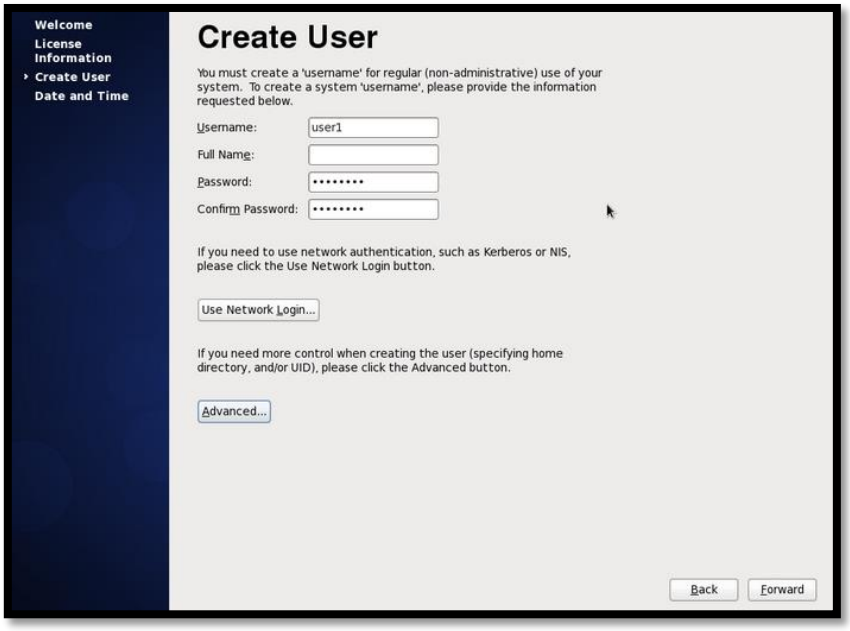

Figura 10: Configuración de Usuario y contraseña bajo el entorno gráfico GNOME.
Fuente: Elaborado por el investigador.

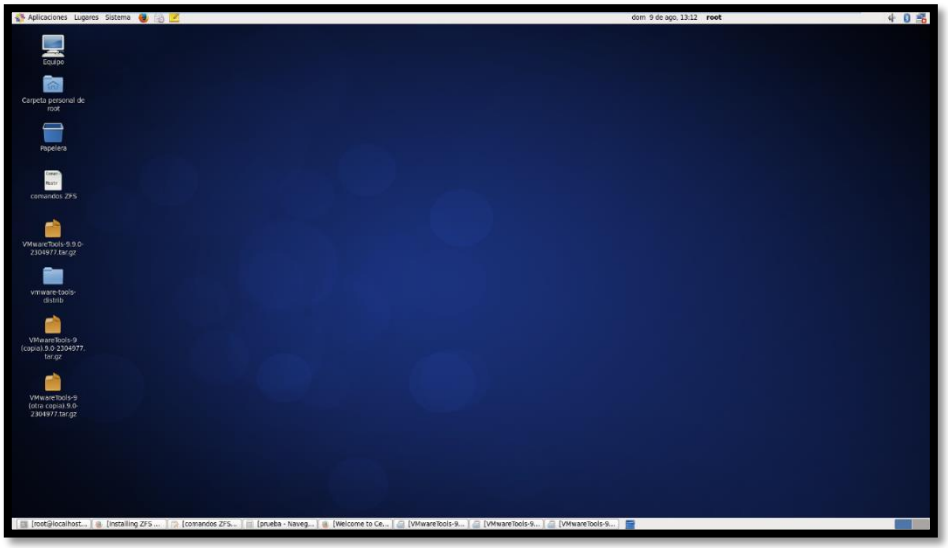

Figura 11: Escritorio de trabajo entorno gráfico GNOME en CentOS 6.6.
Fuente: Elaborado por el investigador.

Una vez iniciada la sesión tendremos un escritorio con entorno gráfico como el que se muestra en la siguiente figura. (Figura 11).

4.6 Instalación y configuración de servidor prototipo de Mirror (HTTP Service)

Haciendo uso de la pantalla Terminal (Figura 12), procedemos a realizar la descarga e instalación de los paquetes necesarios para el funcionamiento del servicio Web HTTP.

`yum -y install httpd`

Instala los paquetes soporte para HTTP

Instalados los módulos anteriores se cambia la configuración en el Firewall del sistema, habilitando el funcionamiento del puerto # 80 para HTTP.

Para realizar esta modificación hacemos uso de los siguientes comandos:

`system-config-firewall-tui`

Ejecutado el comando antes mencionado se muestra la siguiente pantalla en la que habilitamos el servicio HTTP (Figura 12).

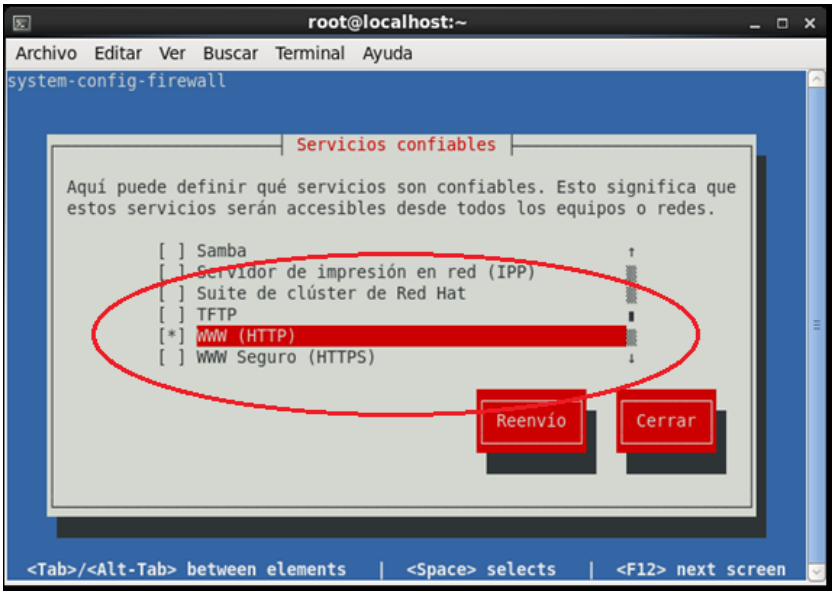

Figura 12: Habilitación de Servicios HTTP y HTTPS.
Fuente: Elaborado por el investigador.

A continuación se procede a modificar el archivo: `estables` ubicado en el directorio `/etc/sysconfig/iptables` (Figura 13):

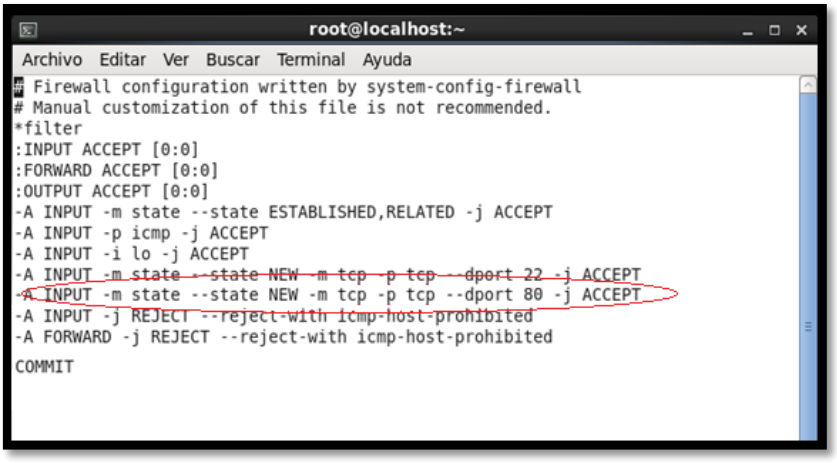

Figura 13: Habilitación del puerto 80 en el archivo de configuración iptables.
Fuente: Elaborado por el investigador.

Reiniciamos el servicio iptables:

```
service iptables restart
```

Para continuar con la configuración del servicio HTTP debemos instalar y configurar el paquete Shorewall para el cual procedemos de la siguiente manera:

```
yum -y install shorewall
```

Realizamos la siguiente configuración para que Shorewall funcione de manera correcta. Creamos el siguiente el directorio:

```
mkdir /usr/share/selinux/packages/shorewall
```

En este directorio creamos un archivo con el siguiente nombre: `shorewall.te`, el mismo que contiene la siguiente información (Figura 14):

Figura 14: Formato de archivo de configuración Shorewall.
Fuente: Elaborado por el investigador.

Creamos el archivo de módulo **shorewall.mod** a partir del archivo **shorewall.te**:

```
checkmodule -M -m -o shorewall.mod shorewall.te
```

De la misma manera creamos el archivo de política **shorewall.pp** a partir del archivo **shorewall.mod**

```
semodule_package -o shorewall.pp -m shorewall.mod
```

Incluimos la siguiente política al sistema con los siguientes comandos:

```
semodule-i
/usr/share/selinux/packages/shorewall/shorewall.pp
```

Aplicamos los cambios con el siguiente comando:

```
sysctl –p
```

4.6.1 Procedimiento de configuración de Shorewall

Se modificarán los siguientes archivos:

- `/etc/shorewall/shorewall.conf:`

Activamos el servicio shorewall:

```
vi /etc/shorewall/shorewall.conf
```

Localizamos la línea **STARTUP_ENABLED** y cambiamos la palabra **No** por **Yes**

- /etc/shorewall/zones

Editamos el archivo vi /etc/shorewall/zones

Definimos una zona (**net**) tipo **ipv4:**

```
Fw      firewall
net     ipv4
```

- /etc/shorewall/interfaces:

Se define los dispositivos de red que corresponden a cada zona del muro cortafuegos, en este caso realizamos la siguiente configuración:

Editamos el archivo vi/etc/shorewall/interface

```
# Shorewall version 4 - Interfaces File
#
# For information about entries in this file, type "man
shorewall-interfaces"
#
# The manpage is also online at
#            http://www.shorewall.net/manpages/shorewall-
interfaces.html
#
###############################################################
#####
#ZONE   INTERFACE       BROADCAST       OPTIONS
net     eth0    detect  dhcp
```

En este caso solo lo dejaremos con las líneas que están por defecto ya que estamos manejando solo una interfaz de red.

- /etc/shorewall/policy:

Editamos el archivo `vi /etc/shorewall/policy` y definimos las siguientes políticas:

```
fw   all   ACCEPT
net  all   DROP  info
```

- `/etc/shorewall/rules:`

Editamos el archivo `vi/etc/shorewall/rules` y agregamos la siguiente línea:

```
ACCEPT     all   fw    tcp   80
```

Permitiendo el acceso al puerto 80.

Iniciamos el servicio con los siguientes comandos:

`service shorewall start`

`systemctl start shorewall` (si es la primera vez que se ejecuta el servicio).

Configuramos que el servicio de Shorewall se inicie con el arranque del sistema:

`chkconfig shorewall on`

Una vez terminada la instalación y configuración del paquete Shorewall, podemos continuar con la configuración del Protocolo HTTP.

Procedemos configurar HTTP para que se inicie conjuntamente con el arranque del sistema:

`chkconfig httpd on`

y procedemos a iniciar el servicio:

`service httpd start`

Permitimos el envío de correo electrónico y lectura de contenidos localizados en directorios definidos por los usuarios:

```
setsebool -P httpd_can_sendmail 1
setsebool -P httpd_read_user_content 1
```

4.7 Instalación y configuración de servidor prototipo de Mirror (FTP Service)

Así como lo hicimos con el servicio de HTTP, procedemos a descargar e instalar el servicio FTP con la siguiente línea de comandos:

```
yum -y install vsftpd
```

Al finalizar la instalación procedemos a iniciar el servicio:

```
service vsftpd start
chkconfig vsftpd on
```

Las configuraciones de cortafuegos y shorewall no son necesarias, ya que las mismas ya fueron efectuadas durante la configuración de ftp.

En este caso solo procedemos a configurar las direcciones del `path` en donde iniciará el servicio ftp. Esta configuración esta en el archivo vsftpd.conf ubicado en el directorio `/etc/vsftpd/.`

4.8 Configuración de Directorio raíz del repositorio

Por lo general el Servicio HTTP trae como directorio raíz la siguiente dirección: **`/var/www/,`** permitiendo que todo el contenido que sea alojado en la carpeta **`www`** sea visualizada en el entorno web, sin embargo se puede cambiar de dirección para el almacenamiento que utilizaremos en nuestro repositorio con el siguiente comando:

Creamos el directorio en donde se almacenara el contenido para nuestro repositorio:

```
mkdir /Mirror/centos/www/
```

A continuación configuramos que el servicio se ejecute y visualice el contenido del nuevo directorio:

```
Chcon -t httpd_sys_content_t
/Mirror/centos/www/tesis/paginainicio_html
```

Como nuestro servidor solo será de descarga de archivos lo configuramos solo con permisos de lectura:

```
chcon -t httpd_sys_script_ro_t
/Mirror/centos/www/tesis/paginainicio_html
```

En este caso toda la instalación y configuración será replicada en cada una de las máquinas virtuales.

Para el Servicio FTP, en el archivo de configuración `vsftpd.conf` agregamos la línea:

`local_root=/Mirror/centos/www/` en el caso de usar el volumen SDFS

o

`local_root=/storage`, en el caso de usar ZFS.

4.9 Descarga, Instalación y configuración de los sistemas para Deduplicación de datos

En esta investigación se pone a prueba a los 3 sistemas de Deduplicación más eficientes: LessFS, SDFS y ZFS los mismos que serán descargados, instalados y configurados de manera independiente en cada una en las distintas Máquinas Virtuales.

4.9.1 LessFS

Una vez que se tiene instalado y configurado el Servicio HTTP para nuestro prototipo de repositorio empezaremos con la instalación de la primera herramienta (Sistema de archivos) para Deduplicación, LessFS.

En este caso mencionamos que este FS a pesar de ser eficiente y uno de los primeros Sistemas de archivos para realizar Deduplicación, ha quedado obsoleto en el contexto de soporte para las versiones de CentOS, la información que se encuentra acerca de la instalación y configuración de LessFS es solo para versiones anteriores a la versión 6.6 de la Distro que se usa en esta investigación, lamentablemente las librerías y paquetes no son compatibles y a pesar de que se las instaló por separado y de manera independiente se llegó al problema final que tiene relación con la base de datos con la que trabaja LessFS.

En sus primeras instancias LessFS utilizaba una versión de Fuse 2.8.5 que es compatible con TokyoCabinet en la versión 1.4.47, estos dos paquetes son completamente compatibles con la versión de LessFS 1.5.4, sin embargo debido a la actualización de CentOS entre una versión a otra se encuentran los siguientes problemas:

TokyoCabinet a pesar de haber llegado a la versión 1.1.48 y de haber sido lanzada el 5 de agosto de 2010, ya no es compatible con CentOS 6 debido a cambios de actualización en el Kernel de esta versión.

Según foros y la página de soporte oficial de LessFS, TokioCabinet quedo obsoleta y fue reemplazada por las bases de Datos: Hámster DB y Berkeley DB. [33] [34]

Fuse también tuvo actualización en su código llegando a la versión 2.9.4 lanzada el 22 de mayo de 2015, lo mismo ocurrió con LessFS, que llegó el 16 de noviembre de 2013 a su versión 1.7.0.

Debido a que TokioCabinet ya no es compatible con las versiones más recientes de Fuse y LessFS, se intenta instalar y configurar la primera alternativa, Hámster DB, paquete que llego a su versión 2.1.10, lamentablemente su software a pesar de ser actualizado recientemente presentó varios problemas al momento de instalarlo en el sistema, problemas debidos a la versión de Java que utiliza CentOS 6.

A pesar de haberse topado con los inconvenientes mencionados anteriormente se procede a descargar e instalar la última opción de DB a usarse con Fuse y LessFS, Berkeley DB la misma que no trae características referentes a la Deduplicación bajo LessFS debido nuevamente a la versión del JRE y JDK de Java que a pesar de estar instaladas y en funcionamiento, no son reconocidas en la instancia de instalación de Berkeley.

Según varios portales en la Web [14] el inconveniente podría deberse a una razón, Berkeley DB fue adquirida por Oracle, quien a su vez también ahora es propietaria de Sun Microsystems y mediante esta adquisición han descartado el soporte que tenía Berkeley para Dedup presumiblemente ya que ahora Oracle tiene su propio sistema para realizar Deduplicación el cual es ZFS [27], FS que venía incorporada en las versiones de Solaris, esta información esta descrita en comentarios de foros de personas que utilizan la Deduplicación a diario y a pesar de que no hay información veraz de esta aseveración de incompatibilidad, es la práctica e intento de instalación y configuración la que nos da posibles respuestas [28]. Foros y páginas [24] [26] especializadas en Deduplicación mencionan que LessFS aún tiene soporte en versiones de Debían, lamentablemente este Sistema Operativo no es el sistema definido para realizar esta investigación, razón por la cual se decide dejar de lado la instalación, configuración y pruebas de funcionamiento de LessFS, documentando lo mencionado anteriormente sobre el Sistema de archivos y los problemas de instalación e incompatibilidad que se generan en CentOS específicamente en versiones superiores a la 6.

4.9.2 SDFS

La instalación y configuración de SDFS (Opendedup) es sencilla y rápida, la misma que se encuentra disponible en la página oficial del proveedor [27].

Se utiliza una máquina virtual que al igual que la anterior tiene instalado y configurado los servicios de HTTP y FTP para simular el comportamiento real de un repositorio.

SDFS utiliza Java, en este caso se realiza la instalación de la siguiente manera:

Para la instalación del JRE en la versión 1.7.0 de Java debemos estar logueados con todos los privilegios de usuario (root) y mediante los siguientes comandos descargamos e instalamos lo necesario (Figura 15):

```
yum install java-1.7.0-openjdk
```

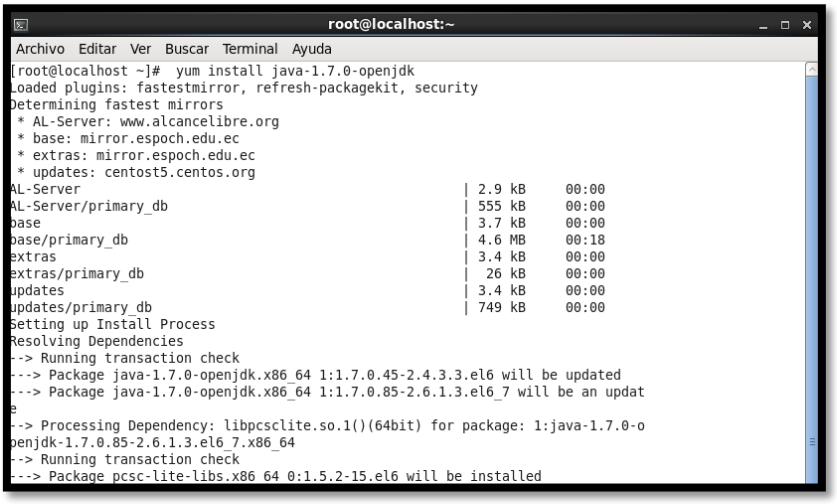

Figura 15: Instalación de Java mediante línea de comandos
Fuente: Elaborado por el investigador.

Una vez instalado Java procedemos a instalar el sistema de archivos SDFS desde el servidor de la aplicación (Figura 16).

```
wget http://www.opendedup.org/downloads/SDFS-2.0.11-2.x86_64.rpm
```

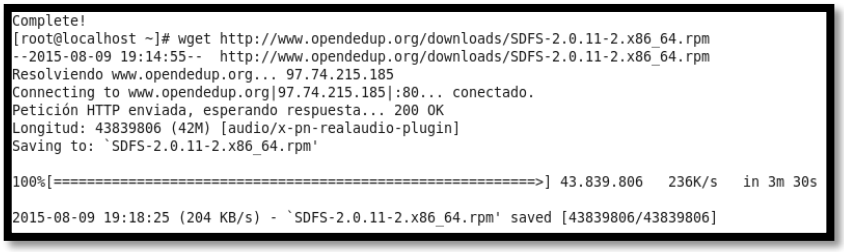

Figura 16: Descarga e instalación del Sistema de archivos SDFS.
Fuente: Elaborado por el investigador.

Verificamos que se haya instalado el paquete en su versión más reciente con los siguientes comandos (Figura 17):

```
rpm -iv SDFS-2.0.11-2.x86_64.rpm
```

Figura 17: Verificación de la descarga e instalación de SDFS.
Fuente: Elaborado por el investigador.

Con las siguientes líneas de código podemos cambiar le máximo de archivos abiertos que pueden mostrase usando SDFS. En el instructivo dela herramienta muestra un valor por defecto para esta instancia:

```
echo "* hardnofile 65535" >> /etc/security/limits.conf

echo "* soft nofle 65535" >> /etc/security/limits.conf

exit
```

Continuando con la instalación y configuración, procedemos a desactivar momentáneamente el Firewall mediante la detención del servicio Iptables, para eso ejecutamos las siguientes líneas de comandos (Figura 18):

```
service                    iptables                    save
service                    iptables                    stop
chkconfig iptables off
```

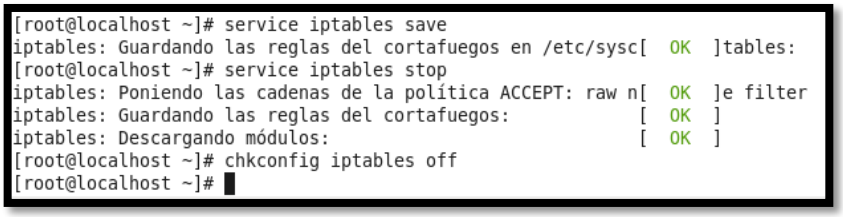

Figura 18: Detención de servicios de Firewall.
Fuente: Elaborado por el investigador.

Para continuar en este caso con la configuración del FS realizamos un reinicio del servidor para que en su nuevo arranque cargue y verifique el sistema lo instalado anteriormente.

Volvemos a iniciar sesión como root y en esta ocasión procedemos configurar el sistema de archivos en uno de los discos duros virtuales que están conectados a nuestro servidor.

Realizaremos una configuración para un Volumen que tenga un sistema de archivos SDFS para Deduplicación por Bloques.

Algunas aplicaciones no están disponibles en los repositorios que vienen definidos por defecto en CentOS, por lo que se recomienda agregar uno de los repositorio más completos para esta Distro de Linux, EPEL Repository, para su instalación ejecutamos la siguiente línea de comando en una ventana de terminal (Figura 20):

```
yum intall epel-release
```

Verificamos que el sistema este reconociendo a nuestro disco duro de 100 GB que esta creado virtualmente y conectado. Una herramienta que tiene un manejo sencillo es GParted la misma que también la instalamos para poder usarla (Figura 19).

```
yum install gparted
```

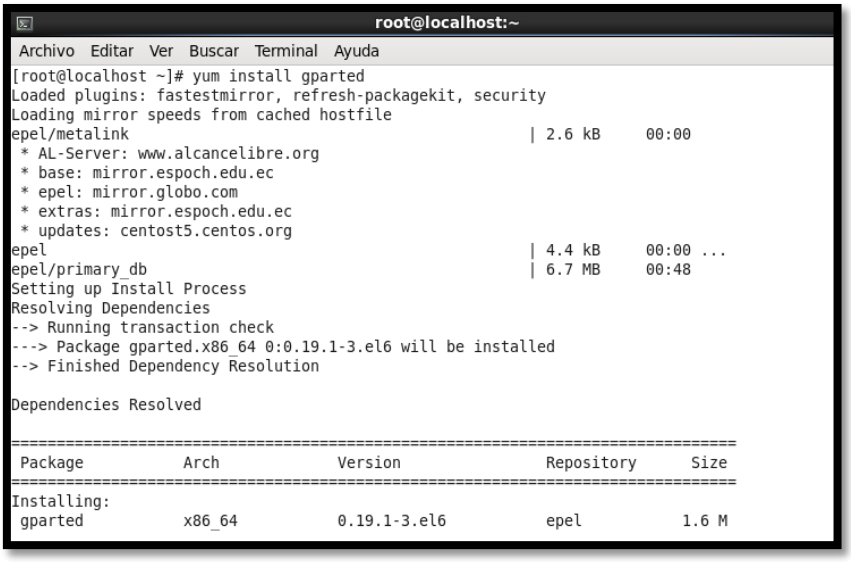

Figura 19: instalación del Sistema para manejo y administración de Discos Duros GParted.
Fuente: Elaborado por el investigador.

Haciendo uso de la herramienta GParted podemos verificar la dirección real de nuestro disco duro virtual (Figura 20):

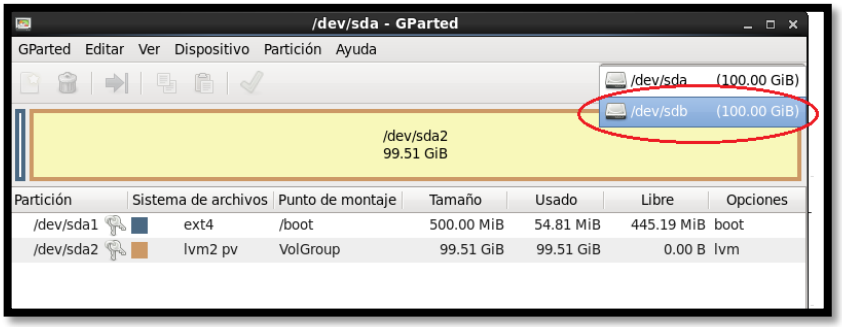

Figura 20: Funcionamiento de la herramienta GParted.

Como se puede observar nuestro disco Duro de 100GB tiene una dirección física:

```
/dev/sdb/
```

La siguiente línea de comandos crea un Volumen de 100 GB de capacidad el mismo que usara un tamaño de bloques de 4 kilobytes (Figura 21).

```
mkfs.sdfs --volume-name=pool1 --volume-capacity=100GB /dev/sdb
```

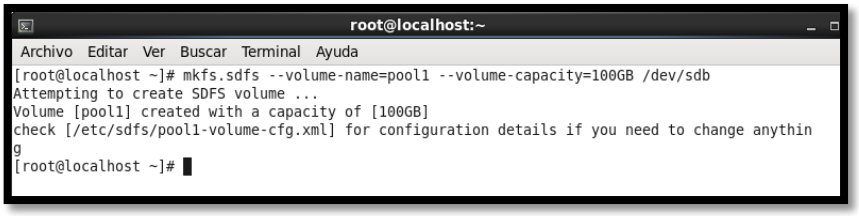

Figura 21: Creación del Volumen con un Sistema de archivos SDFS.
Fuente: Elaborado por el investigador.

El volumen creado tiene el nombre pool1, el mismo que lo direccionamos virtualmente al directorio creado para nuestro servidor HTTP:

`/Mirror/centos/www/`

Esto lo realizamos con la siguiente línea de configuración:

```
mount.sdfs pool1 /Mirror/centos/www/
```

Una vez direccionado virtualmente nuestro volumen tendremos la siguiente ventana (Figura 22):

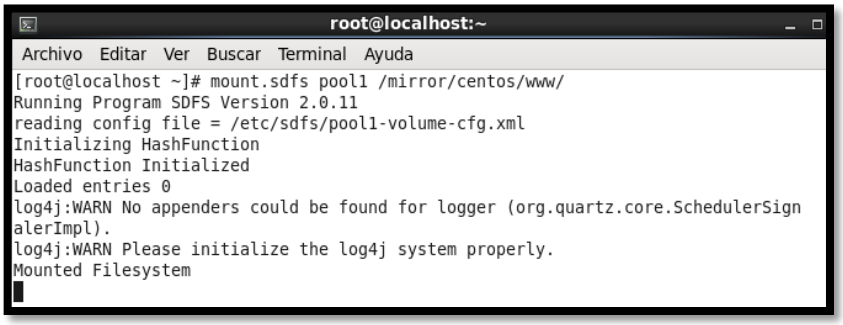

Figura 22: Volumen SDFS montado en nuestro servidor.
Fuente: Elaborado por el investigador.

Finalmente nuestro Volumen con una capacidad de 100 GB se encuentra montado en el directorio `/Mirror/centos/www/` funcionando bajo el Sistema de archivos SDFS listo para deduplicar.

4.9.3 ZFS

Antes de Iniciar con este Sistema de archivos desactivamos el SELinux de nuestro SO, esto se lo hace editando el archivo ubicado en el directorio `/etc/sysconfig/selinux` y en este cambiamos la palabra **`enforced`** por **`disabled`** (Figura 23). También se realiza una actualización limpia al sistema (`yum -y update`) y posterior a esto un reinicio del mismo para que los cambios tengan efecto.

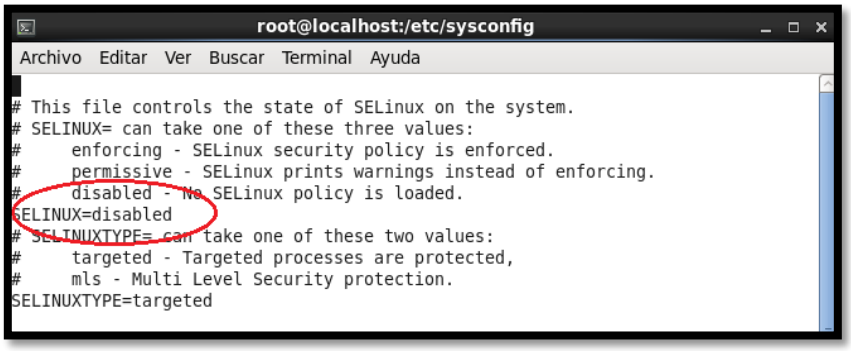

Figura 23: Cambio de la configuración de SELinux.
Fuente: Elaborado por el investigador.

Para poder instalar y configurar el sistema de archivos ZFS se debe tener instalado el repositorio EPEL, y los módulos de ZFS, esto lo realizamos con las siguientes líneas de código:

```
yum localinstall --nogpgcheck
http://archive.zfsonlinux.org/epel/zfs-release.el6.noarch.rpm
yum install kernel-devel zfs
```

Además también se debe descargar e instalar los paquetes **"`Development Tools`"** que consiste en un pack completo de librerías previas para uso de sistemas de archivos compatibles con lenguajes de programación y servicios WEB (Figura 24).

```
yum -y groupinstall "Development Tools"
```

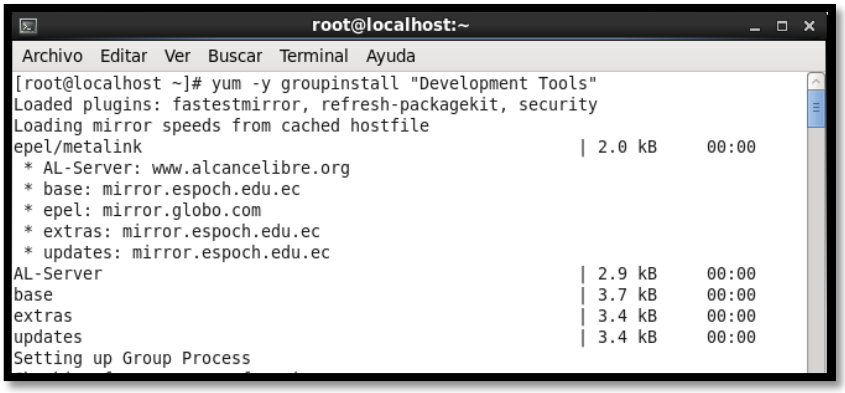

Figura 24: Instalación del paquete de librerías "Development Tools".
Fuente: Elaborado por el investigador.

En este momento ya están instaladas las librerías necesarias para poder instalar ZFS, este sistema de archivos necesita descargar de la misma manera un conjunto de paquetes, los mismos que se descargarán e instalarán con la siguiente línea de comandos (Figura 24):

Una vez instaladas todos los paquetes necesarios procedemos con la configuración y puesta en marcha del FS ZFS.

Iniciamos cargando los módulos de ZFS (Figura 25):

```
modprobe zfs
lsmod | grep -i zfs
```

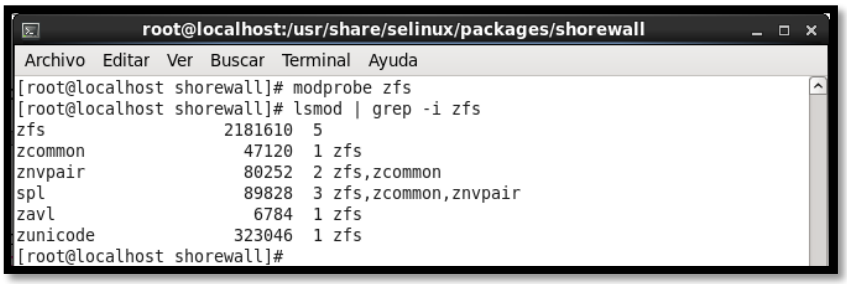

Figura 25: Carga de módulos ZFS en el sistema.
Fuente: Elaborado por el investigador.

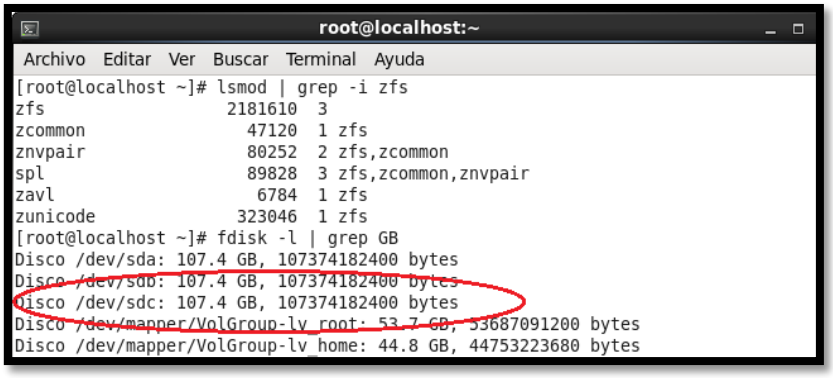

Figura 26: Visualización de Discos duros con direcciones físicas.
Fuente: Elaborado por el investigador.

Ahora usamos los comandos `fdisk-l | grep GB` para poder mostrar las unidades físicas que se encuentran en nuestro sistema (HDD) (Figura 26), ZFS será aplicado en la unidad `/dev/sdc` que tiene una capacidad de 100 GB, esto lo realizamos con los siguientes comandos:

```
zpool create storage -f sdc
```

Esto creará de igual manera una piscina (directorio) en donde `/sdc/` se convertirá virtualmente en el directorio `/storage/` (Figura 27), así mismo podemos visualizar el estado de nuestras piscinas con el comando `zpool status`.

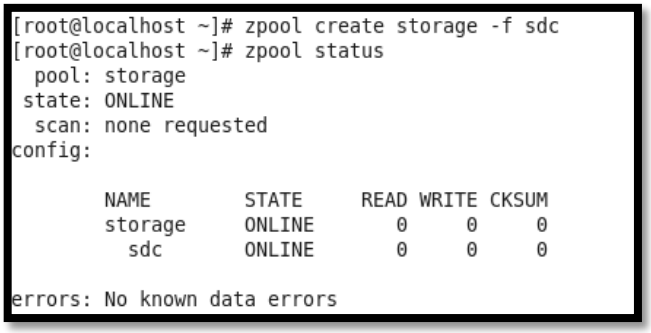

Figura 27: Creacion de piscina con el FS ZFS.
Fuente: Elaborado por el investigador.

ZFS trae 2 opciones importantes tales como la Deduplicación y la compresión, en este caso al ser utilizado para Deduplicación desactivamos la compresión, esto lo hacemos con los siguientes comandos:

```
zfs set compression=off storage
zfs set dedup=on storage
```

Activando solo ZFS para deduplicar con eso tenemos listo el directorio para empezar las pruebas de funcionamiento y rendimiento.

4.10 Pruebas de funcionamiento de Deduplicación en FS ZFS y SDFS

Para comprobar el funcionamiento se realizarán 2 tipos de pruebas en las cuales utilizaremos un archivo perteneciente a una imagen de una Distribución de CentOS, (CentOS-5.11-x86_64-bin-DVD-1of2), el mismo que tiene un peso de 4.9 GB en los que están contenidos 3662 archivos. Además de una nueva ISO de una Distribución CentOS en su versión mínima (CentOS-6(copia)5-x86_64_minimal.iso) que tiene un peso de 398 MB.

EN la primera prueba se realizan copias iguales del archivo CentOS-5.11-x86_64-bin-DVD-1of2 en el mismo volumen con el fin de verificar que la deduplicación funcione, y finalmente se copia al mismo directorio el archivo CentOS-6(copia)5-x86_64_minimal.iso para verificar el funcionamiento con archivos distintos y su comportamiento en cuanto al almacenamiento físico y lógico de la información.

4.10.1 SDFS

Accedemos a nuestro volumen SDFS montado en el directorio `/mirror/centos/www/`, y verificamos el contenido y espacio usado. (Figura 37).

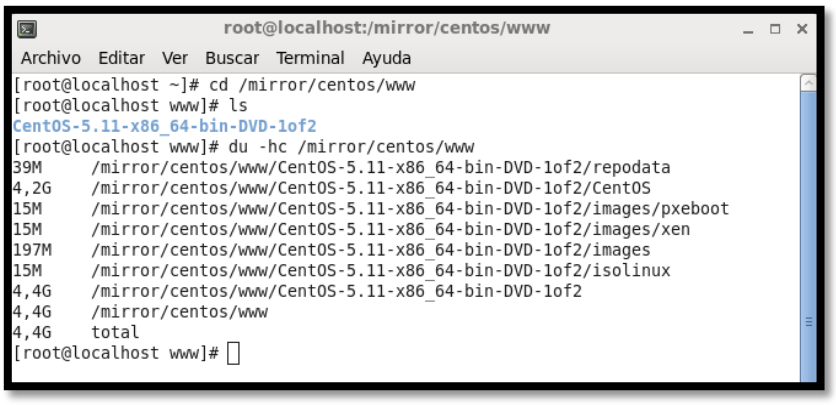

Figura 28: Direccionamiento al directorio del volumen con FS SDFS.
Fuente: Elaborado por el investigador.

Realizamos nuevamente en nuestro directorio una copia del mismo archivo alojado en nuestro volumen, nuevamente procedemos a verificar el tamaño, haciendo el del comando `du -hc` en el terminal, podemos observar que el tamaño se ha duplicado, sin embargo al ejecutar el comando `df -h` se visualiza el uso físico de nuestro disco, en el que se verifica que solo ha sido ocupado 4.3 GB, tamaño original del archivo, concluyendo que al igual que ZFS también está realizando una Deduplicación de contenido (Figura 38).

Realizamos la última prueba de funcionamiento, copiamos nuevamente un archivo en este caso uno diferente a los anteriores (Figura 39), y verificamos los valores gráficos y físicos de los tamaños de nuestro volumen SDFS (Figura 40), verificando que SDFS también funciona deduplicando.

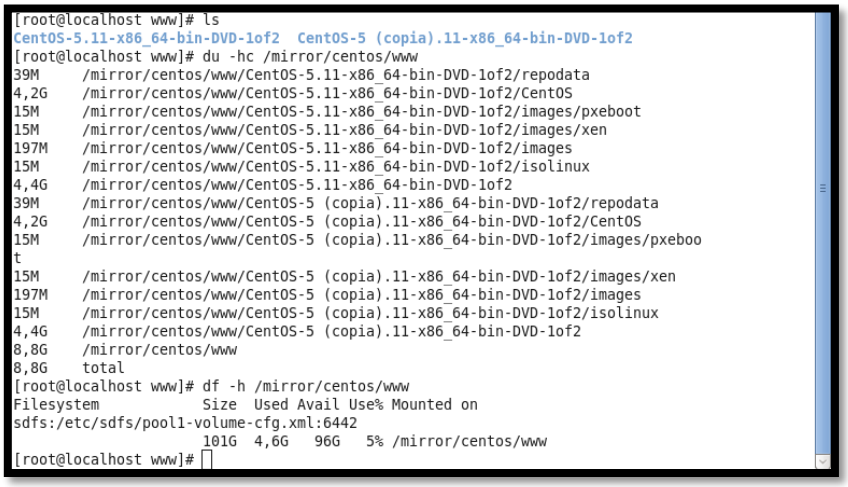

Figura 29: Verificación del proceso de Deduplicación SDFS con archivos iguales.
Fuente: Elaborado por el investigador.

Figura 30: Visualización de espacio físico disponible y usado. SDFS.
Fuente: Elaborado por el investigador.

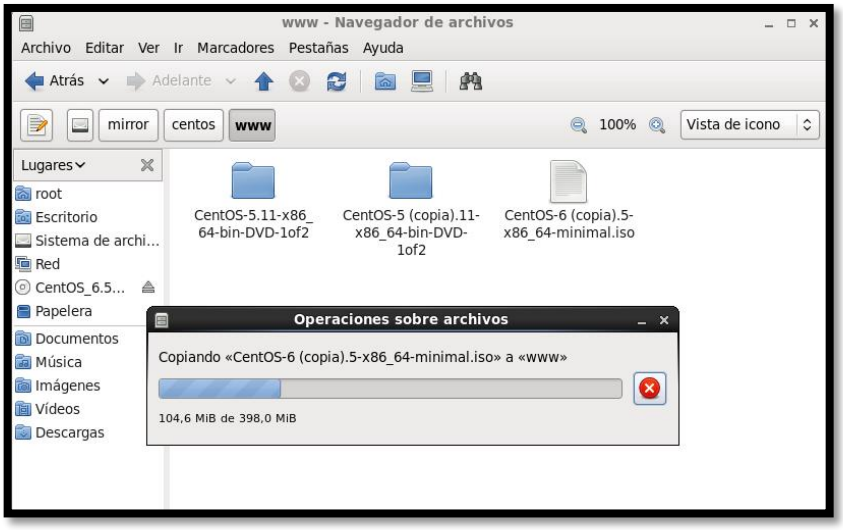

Figura 31: Proceso de Deduplicación en archivos distintos SDFS.
Fuente: Elaborado por el investigador.

```
sdfs:/etc/sdfs/pool1-volume-cfg.xml:6442
                      101G  4,6G   96G   5% /mirror/centos/www
[root@localhost www]# ls
CentOS-5.11-x86_64-bin-DVD-1of2          CentOS-6 (copia).5-x86_64-minimal.iso
CentOS-5 (copia).11-x86_64-bin-DVD-1of2
[root@localhost www]# du -hc /mirror/centos/www
39M     /mirror/centos/www/CentOS-5.11-x86_64-bin-DVD-1of2/repodata
4,2G    /mirror/centos/www/CentOS-5.11-x86_64-bin-DVD-1of2/CentOS
15M     /mirror/centos/www/CentOS-5.11-x86_64-bin-DVD-1of2/images/pxeboot
15M     /mirror/centos/www/CentOS-5.11-x86_64-bin-DVD-1of2/images/xen
197M    /mirror/centos/www/CentOS-5.11-x86_64-bin-DVD-1of2/images
15M     /mirror/centos/www/CentOS-5.11-x86_64-bin-DVD-1of2/isolinux
4,4G    /mirror/centos/www/CentOS-5.11-x86_64-bin-DVD-1of2
39M     /mirror/centos/www/CentOS-5 (copia).11-x86_64-bin-DVD-1of2/repodata
4,2G    /mirror/centos/www/CentOS-5 (copia).11-x86_64-bin-DVD-1of2/CentOS
15M     /mirror/centos/www/CentOS-5 (copia).11-x86_64-bin-DVD-1of2/images/pxeboot
15M     /mirror/centos/www/CentOS-5 (copia).11-x86_64-bin-DVD-1of2/images/xen
197M    /mirror/centos/www/CentOS-5 (copia).11-x86_64-bin-DVD-1of2/images
15M     /mirror/centos/www/CentOS-5 (copia).11-x86_64-bin-DVD-1of2/isolinux
4,4G    /mirror/centos/www/CentOS-5 (copia).11-x86_64-bin-DVD-1of2
9,2G    /mirror/centos/www
9,2G    total
[root@localhost www]# df -h /mirror/centos/www
Filesystem            Size  Used Avail Use% Mounted on
sdfs:/etc/sdfs/pool1-volume-cfg.xml:6442
                      101G  5,0G   96G   5% /mirror/centos/www
[root@localhost www]#
```

Figura 32: Verificación del proceso de Deduplicación con archivos distintos SDFS.
Fuente: Elaborado por el investigador.

4.10.2 ZFS

Una vez montados los volúmenes virtuales cada uno con los sistemas de archivos diferentes, en este caso ZFS, realizamos las pruebas del funcionamiento, para estas hacemos uso de un directorio perteneciente a una Distro de CentOS 5.11, estos archivos que vamos a utilizar para nuestros test tienen un peso aproximado de 4.3 GB, y están compuestos de 3662 elementos (Figura 28).

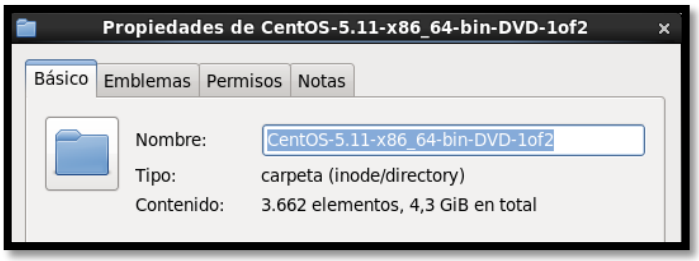

Figura 33: Propiedades del directorio usado para las pruebas de funcionamiento de la Deduplicación.
Fuente: Elaborado por el investigador.

Como observamos primero nos ubicamos en nuestro directorio montado con el FS ZFS, el procedimiento es el siguiente (Figura 29).

```
cd /storage
```

Listamos el contenido de nuestro directorio, en este caso como se observa en la figura presentada a continuación (Figura 29) este ya tiene los archivos pertenecientes a la Distro de CentOS.

Figura 34: Direccionamiento al directorio del volumen con FS ZFS.
Fuente: Elaborado por el investigador.

Para verificar el funcionamiento de nuestra piscina ZFS, introducimos el siguiente comando (Figura 30):

```
zpool list
```

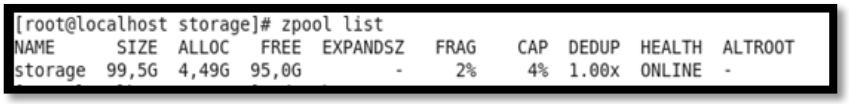

Figura 35: Funcionamiento comando zpool list.
Fuente: Elaborado por el investigador.

Se visualiza el nombre, tamaño total, espacio usado, espacio libre de nuestro volumen ZFS, así también lo hacemos utilizando `du -hc /storage`, para verificar los datos dentro del volumen (Figura 31).

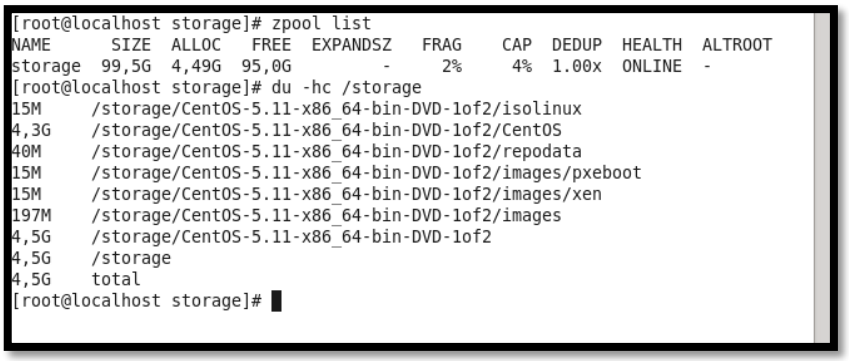

Figura 36: Visualización del funcionamiento de ZFS.
Fuente: Elaborado por el investigador.

Haciendo uso de nuestro entorno gráfico realizamos la primera prueba de funcionamiento, realizamos una copia del mismo contenido en el mismo directorio (Figura 32), y a continuación volvemos a listar el contenido de `/storage/`, (Figura 33) en donde observamos que ya se encuentran alojados los dos archivos, un original y una copia.

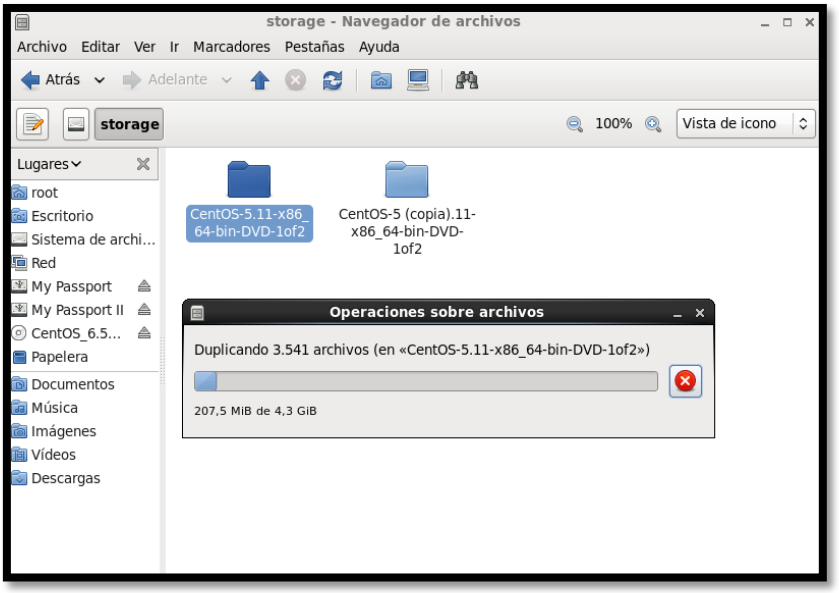

Figura 37: Deduplicación ZFS – Duplicación de archivos.
Fuente: Elaborado por el investigador.

```
[root@localhost storage]# ls
CentOS-5.11-x86_64-bin-DVD-1of2  CentOS-5 (copia).11-x86_64-bin-DVD-1of2
[root@localhost storage]# zpool list
NAME      SIZE  ALLOC   FREE  EXPANDSZ   FRAG    CAP  DEDUP  HEALTH  ALTROOT
storage  99,5G  4,64G  94,9G         -     3%     4%  1.90x  ONLINE  -
[root@localhost storage]# du -hc /storage
15M     /storage/CentOS-5.11-x86_64-bin-DVD-1of2/isolinux
4,3G    /storage/CentOS-5.11-x86_64-bin-DVD-1of2/CentOS
40M     /storage/CentOS-5.11-x86_64-bin-DVD-1of2/repodata
15M     /storage/CentOS-5.11-x86_64-bin-DVD-1of2/images/pxeboot
15M     /storage/CentOS-5.11-x86_64-bin-DVD-1of2/images/xen
197M    /storage/CentOS-5.11-x86_64-bin-DVD-1of2/images
4,5G    /storage/CentOS-5.11-x86_64-bin-DVD-1of2
15M     /storage/CentOS-5 (copia).11-x86_64-bin-DVD-1of2/isolinux
15M     /storage/CentOS-5 (copia).11-x86_64-bin-DVD-1of2/images/pxeboot
15M     /storage/CentOS-5 (copia).11-x86_64-bin-DVD-1of2/images/xen
194M    /storage/CentOS-5 (copia).11-x86_64-bin-DVD-1of2/images
4,1G    /storage/CentOS-5 (copia).11-x86_64-bin-DVD-1of2/CentOS
39M     /storage/CentOS-5 (copia).11-x86_64-bin-DVD-1of2/repodata
4,4G    /storage/CentOS-5 (copia).11-x86_64-bin-DVD-1of2
8,8G    /storage
8,8G    total
[root@localhost storage]#
```

Figura 38: Verificación de proceso de Deduplicación en ZFS FS.
Fuente: Elaborado por el investigador.

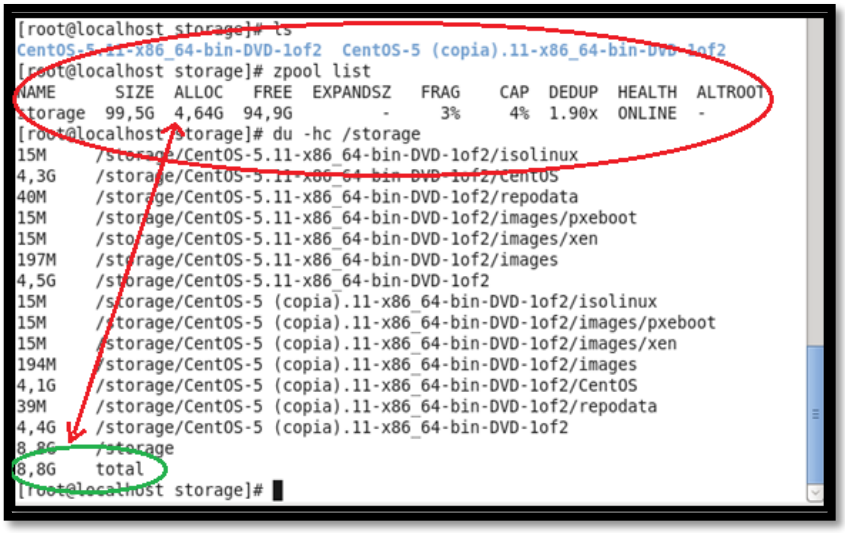

Figura 39: Funcionamiento de zpool list después de duplicar información.
Fuente: Elaborado por el investigador.

Al igual que el procedimiento anterior, haciendo uso del comando `zpool list` (Figura34) podemos observar que fisicamente el espacio de disco inicial no ha cambiado, lo que nos da a entender que el proceso de Deduplicación en el volumen funciona, por otra parte usando `du -hc /storage` verificamos el tamaño actual de nuestro `/storage`, el mismo que ahora es el doble de la cantidad en GB iniciales, cabe recalcar que este comando muestra la información transparente para el usuario en donde se observa de manera gráfica que existen los 2 archivos iguales, cuando el uno solo es un link que apunta al archivo original, sin guardar información repetida en el disco. (Figura 33).

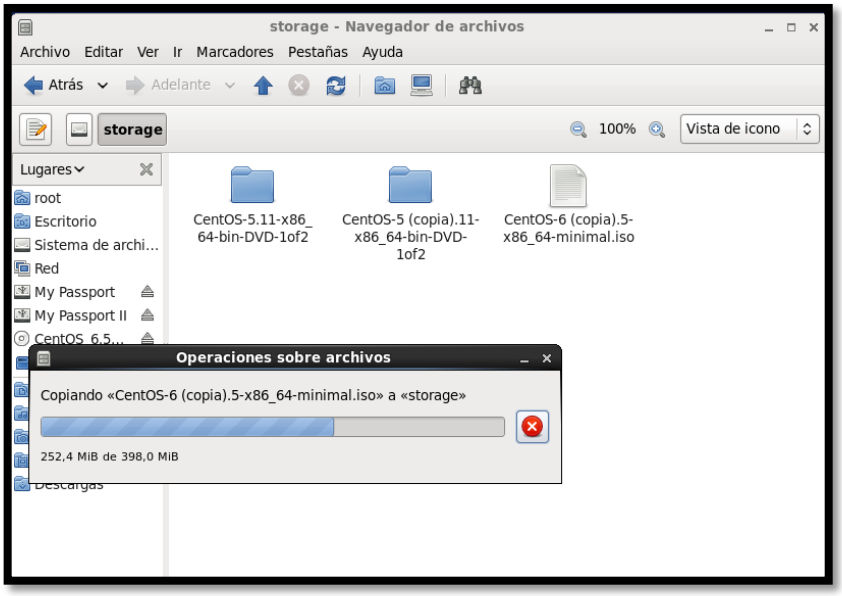

Figura 40: Proceso de Deduplicación en archivos distintos ZFS.
Fuente: Elaborado por el investigador.

Ahora procedemos a copiar un nuevo archivo en `/storage`, este es diferente a los archivos que se encuentran en nuestro volumen ZFS, con el fin de verificar si se realizan cambios tanto físicos como gráficos para el usuario. (Figura 35).

```
[root@localhost storage]# ls
CentOS-5.11-x86_64-bin-DVD-1of2          CentOS-6 (copia).5-x86_64-minimal.iso
CentOS-5 (copia).11-x86_64-bin-DVD-1of2
[root@localhost storage]# zpool list
NAME      SIZE  ALLOC   FREE  EXPANDSZ   FRAG    CAP  DEDUP  HEALTH  ALTROOT
storage  99,5G  5,02G  94,5G         -     3%     5%  1.83x  ONLINE  -
[root@localhost storage]# du -hc /storage
15M     /storage/CentOS-5.11-x86_64-bin-DVD-1of2/isolinux
4,3G    /storage/CentOS-5.11-x86_64-bin-DVD-1of2/CentOS
40M     /storage/CentOS-5.11-x86_64-bin-DVD-1of2/repodata
15M     /storage/CentOS-5.11-x86_64-bin-DVD-1of2/images/pxeboot
15M     /storage/CentOS-5.11-x86_64-bin-DVD-1of2/images/xen
197M    /storage/CentOS-5.11-x86_64-bin-DVD-1of2/images
4,5G    /storage/CentOS-5.11-x86_64-bin-DVD-1of2
15M     /storage/CentOS-5 (copia).11-x86_64-bin-DVD-1of2/isolinux
15M     /storage/CentOS-5 (copia).11-x86_64-bin-DVD-1of2/images/pxeboot
15M     /storage/CentOS-5 (copia).11-x86_64-bin-DVD-1of2/images/xen
194M    /storage/CentOS-5 (copia).11-x86_64-bin-DVD-1of2/images
4,1G    /storage/CentOS-5 (copia).11-x86_64-bin-DVD-1of2/CentOS
39M     /storage/CentOS-5 (copia).11-x86_64-bin-DVD-1of2/repodata
4,4G    /storage/CentOS-5 (copia).11-x86_64-bin-DVD-1of2
9,2G    /storage
9,2G    total
[root@localhost storage]#
```

Figura 41: Verificación del proceso de Deduplicación con archivos distintos ZFS.
Fuente: Elaborado por el investigador.

Una vez realizado este proceso verificamos el funcionamiento de nuestro FS, se observa que el valor físico del uso de disco ha variado en 400 MB, tamaño del nuevo archivo

agregado al `/storage`. Nuevamente verificamos los dos entornos, el físico y el gráfico (Figura 36) llegando a la conclusión de que la configuración y montaje de nuestro volumen `/storage` bajo el Sistema de archivos ZFS funciona permitiendo deduplicar la información en tiempo real y bajo los parámetros de Deduplicación conocidos.

4.10.3 Pruebas Funcionamiento Prototipo de Mirror

A continuación probamos el funcionamiento de nuestro prototipo de Mirror direccionado a cada uno de los directorios de los volúmenes montados con los sistemas de archivos diferentes, y conectados desde otro computador con la ayuda de una navegador web accedemos a la dirección `ftp://127.1.1.1/mirror/centos/www/` para poder visualizar nuestro directorio y proceder a realizar una descarga de archivos, con el fin de realizar una medición de uso de la red y tiempo de descarga bajo el FS SDFS (Figura 41).

Al igual que con SDFS, vamos a realizar el mismo procedimiento pero ahora conectándonos a nuestro otro volumen con FS ZFS, para esto primero debemos re direccionar el `path` que mostrará nuestro servicio FTP, esta configuración se la cambia en el archivo vsftpd.conf alojado en el directorio `/etc/vsftpd/`.

Posterior a este cambio ingresamos nuevamente a nuestro volumen, en este caso bajo ZFS con la dirección `ftp://127.1.1.1/storage/` y realizamos nuevamente la descarga del mismo archivo para poder medir el rendimiento de cada uno de los FS en cuanto a uso de red al realizar una descarga (Figura 42).

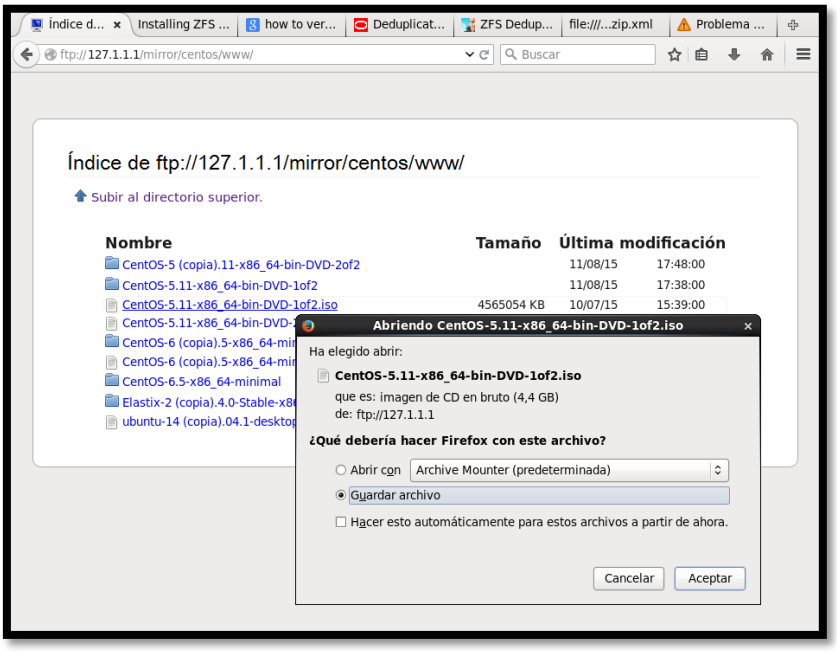

Figura 42: Prueba de descarga de archivos desde el servidor FTP bajo SDFS.
Fuente: Elaborado por el investigador.

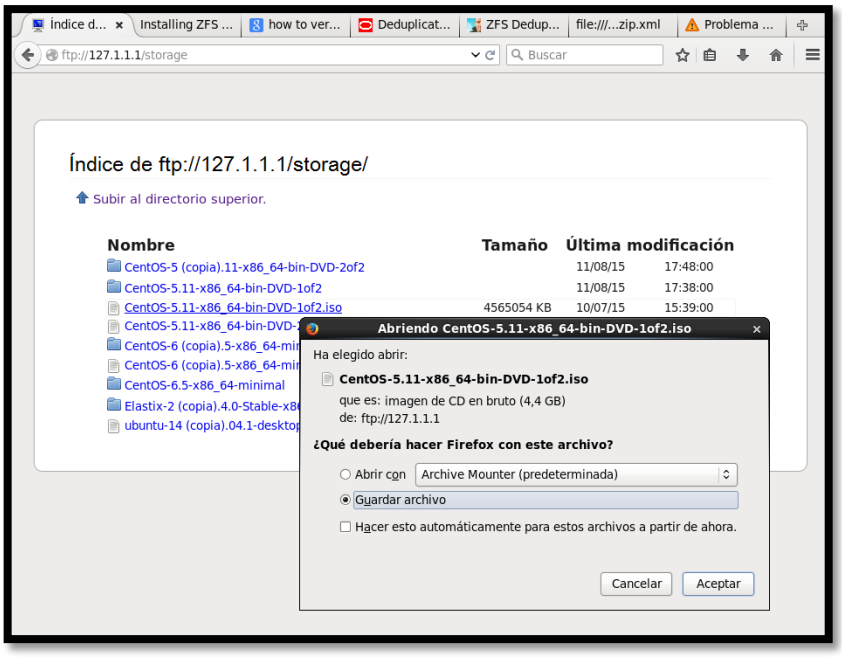

Figura 43: Prueba de descarga de archivos desde el servidor FTP bajo ZFS.
Fuente: Elaborado por el investigador.

4.10.4 Análisis de Resultados

Pruebas de Funcionamiento: Verificación de ahorro de espacio de almacenamiento en Discos Duros.

Las pruebas de funcionamiento se realizan haciendo uso de la herramientas SysStat, la misma que utiliza el comando SAR que recoge información en tiempo real del comportamiento de cada uno de los recursos de hardware tomarse en cuenta en las pruebas, referentes a CPU, RAM, RED TX,(Anexo 2, Anexo 3).

Los siguientes resultados son basados en las Pruebas de funcionamiento efectuadas en la puesta en marcha de cada uno de los sistemas de archivos para deduplicar, SDFS por su parte arroja resultados al deduplicar durante la escritura en el volumen de 2 archivos iguales, mostrando en la figura (Figura 29) un consumo real de 4.4 GB de HDD antes los 8,8 GB que se muestran al usuario, mostrando en sus resultados un uso de 5% del tamaño global del disco usado (100 GB). Por otro lado, la prueba realizada bajo ZFS entrega resultados más exactos, al realizar el mismo procedimiento de escritura en el volumen de los mismos archivos iguales dan como resultado una escritura que ocupa 4.64 GB, la diferencia es que ZFS muestra un uso de HDD del 4% con la misma cantidad de información en su volumen que SDFS. (Figura 44).

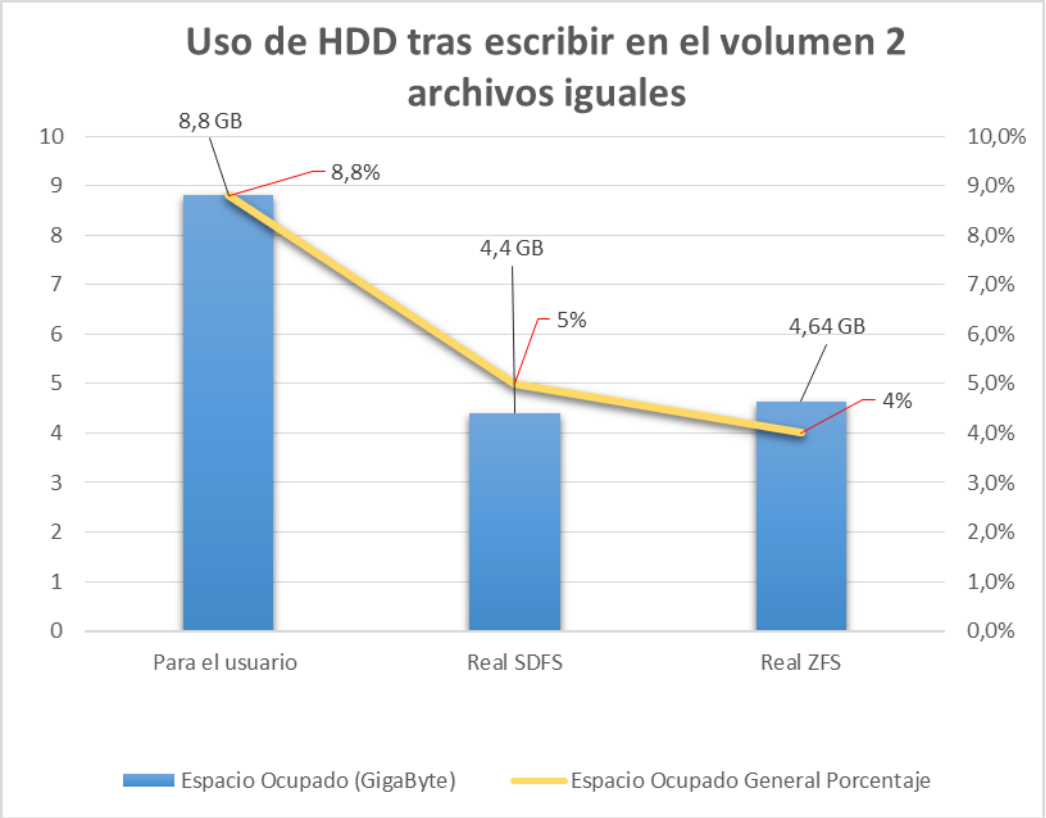

Figura 44: Comparativa de resultados tras escribir 2 archivos iguales en volúmenes con SDFS y ZFS
Fuente: Elaborado por el investigador.

Al realizar las pruebas de funcionamiento con 2 archivos iguales y uno diferente, SDFS entrega resultados de un uso del HDD de 5% , el mismo valor que se mostró con la prueba anterior, y un uso real en cuanto a escritura en el volumen de 5 GB ante los 9.2 GB que se muestran transparentemente al usuario como espacio usado, por otra parte ZFS a realizar la misma operación muestra un uso de 5% del tamaño total del disco, y un espacio ocupado en su volumen de 5.02GB. (Figura 45).

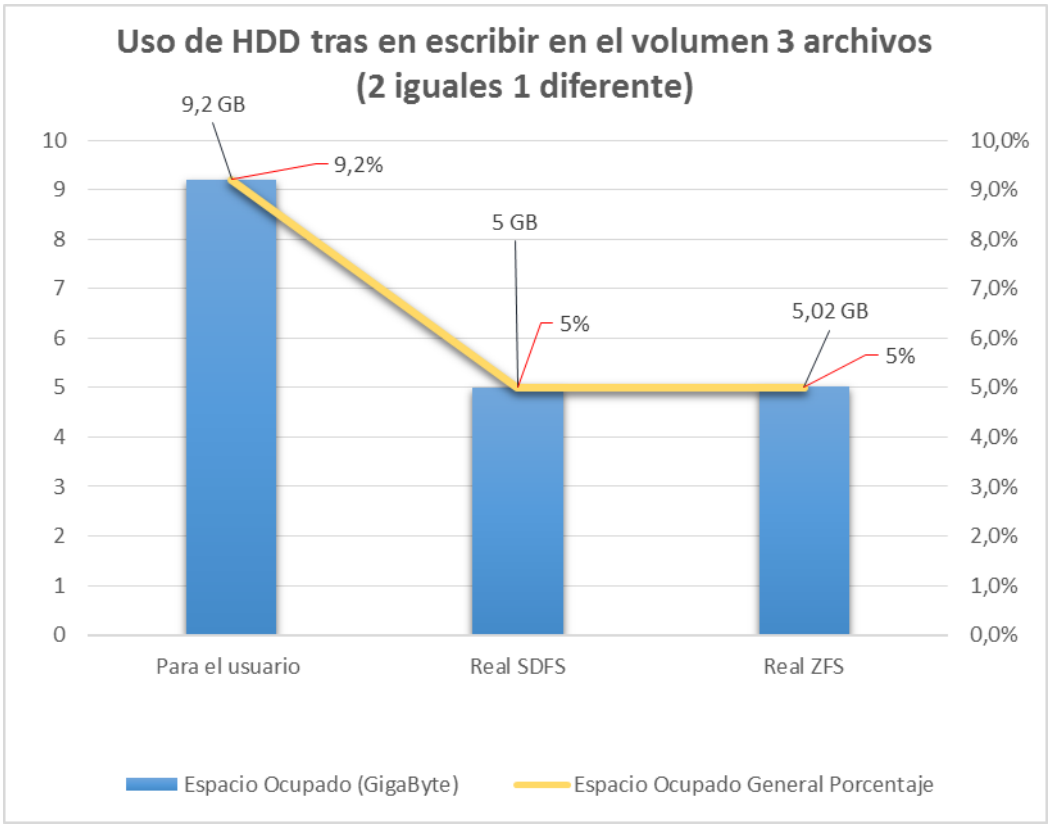

Figura 45: Comparativa de resultados tras escribir 3 archivos (2 iguales y 1 diferente) en volúmenes con SDFS y ZFS
Fuente: Elaborado por el investigador.

Prueba de Funcionamiento: Rendimiento de CPU, RAM, Red TX al realizar procesos de deduplicación.

Para las pruebas de funcionamiento de rendimiento se realizaron 30 tomas diferentes de datos, recogiendo valores de consumo de CPU, RAM y velocidades de descarga de archivos, así mismo consumo de ancho de banda de red durante cada descarga, valores que al ser promediados dieron los resultados en cada una de las pruebas que muestran a continuación.

Los siguientes gráficos muestran el comportamiento de cada una de los sistemas de archivos probados anteriormente, cada uno está enfocado al consumo de CPU (Figura 43) y Memoria RAM (Figura 44) durante la operación de escritura y procesos de Deduplicación de archivos en nuestros volúmenes con FS diferentes, además de la gráfica en cuanto al consumo de red y tiempo de descarga de archivos (imágenes de disco) cada una bajo SDFS y ZFS (Figura 45).

Para la recolección de la información se utilizó el comando SAR, el que recopila valores de rendimiento de uso de hardware de nuestras máquinas virtuales realizando cada uno de los procesos.

Los comandos (Anexo 2) utilizados para llevar a cabo esta recopilación están disponibles en el repositorio global de la plataforma office365, archivos que pueden ser usados y modificados para uso general. [35]

Como se observa en el gráfico (Figura 44), se muestra un comportamiento de uso de la CPU elevado al realizar escritura en el volumen formateado con el FS SDFS, llegando a usar el 24,67 % de procesador, mientras que la escritura en el volumen ZFS, el uso en este llega a 23,59%, se puede deducir que la mejor alternativa en cuanto al uso y consumo de procesador durante la tarea de escritura en un volumen con FS para deduplicar, es ZFS.

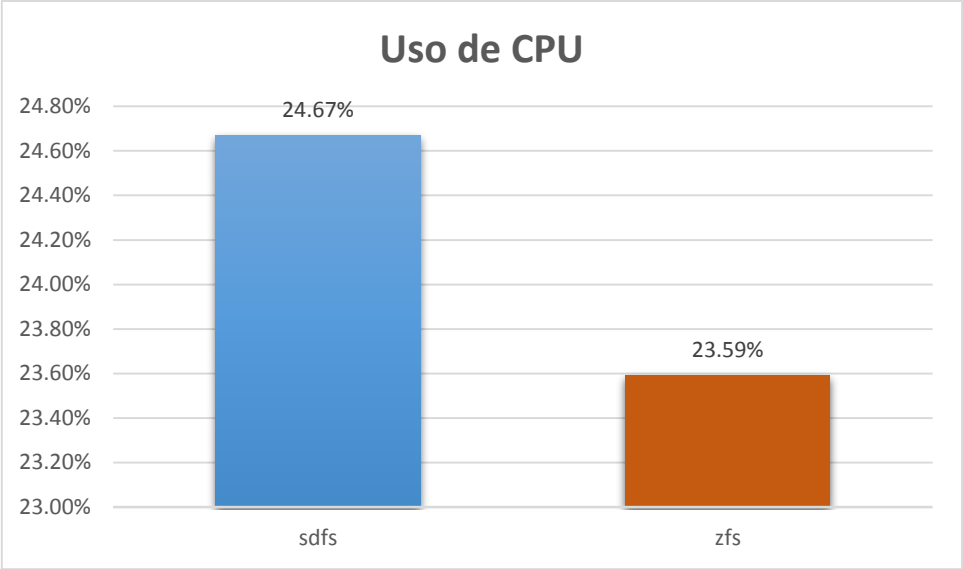

Figura 46: Uso de la CPU en procesos de escritura y Deduplicación.
Fuente: Elaborado por el investigador.

La grafica (Figura 46) en cuanto al consumo de Memoria RAM, muestra también una ventaja significativa de ZFS ante SDFS, la primera consumiendo 94.99% de RAM ante los 96.85% de su rival, ZFS muestra un leve carga a este valor además de ser más eficiente que SDFS que consume mucho más que la anterior.

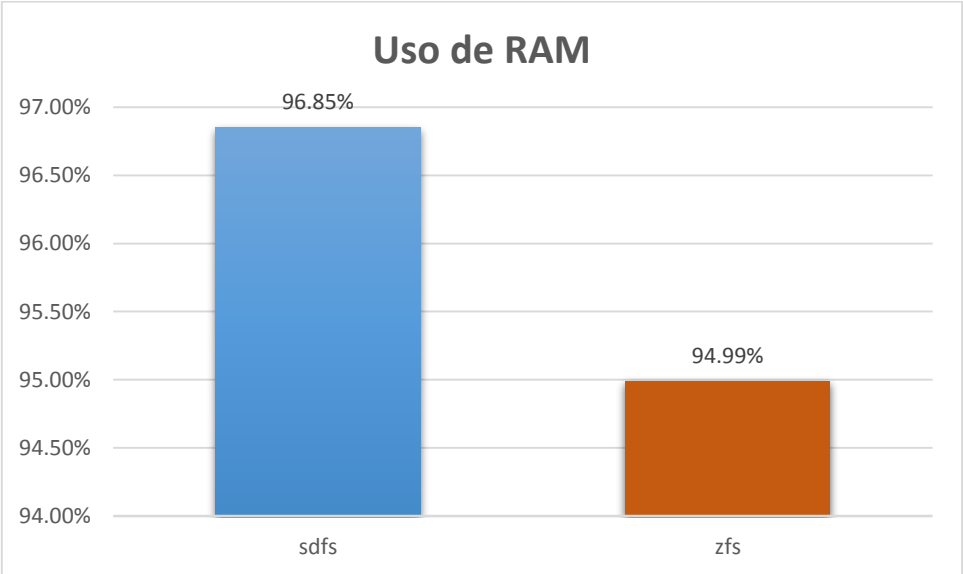

Figura 47: Uso de la RAM en procesos de escritura y Deduplicación.
Fuente: Elaborado por el investigador.

Otro parámetro a ser evaluado corresponde al tiempo que tomaría realizar la descarga de un archivo que este alojada en volúmenes tanto con FS ZFS y SDFS, se observa en la gráfica (Figura 47) que el tiempo de descarga desde el volumen con SDFS tarda 8 minutos, ZFS por su parte permite la descarga del mismo archivo en 5 minutos, con estos datos llegamos a verificar que ZFS tiene ventaja sobre SDFS en cuanto a descarga de archivos.

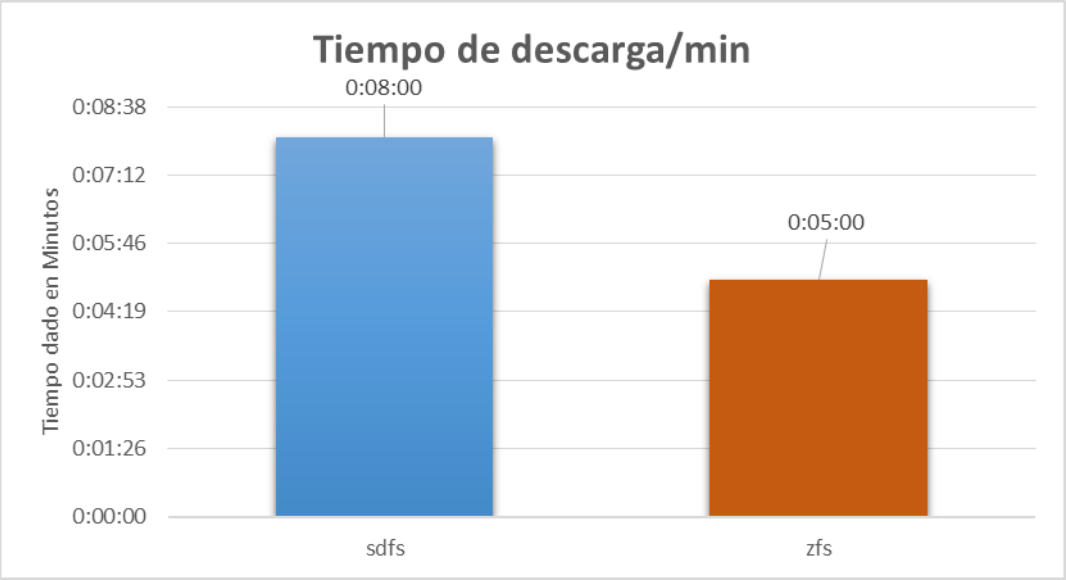

Figura 48: Tiempo de descarga bajo SDFS y ZFS.
Fuente: Elaborado por el investigador.

La última prueba de rendimiento realizada corresponde a uso de La Red TX al momento de realizar la descarga de un archivo alojado en cada uno de los volúmenes, SDFS durante la descarga consume un 48.37 % de red TX, no así su rival ZFS que consume mucho menos, llegando a ocupar el 29.45% de ancho de banda haciendo el mismo procedimiento (Figura 49).

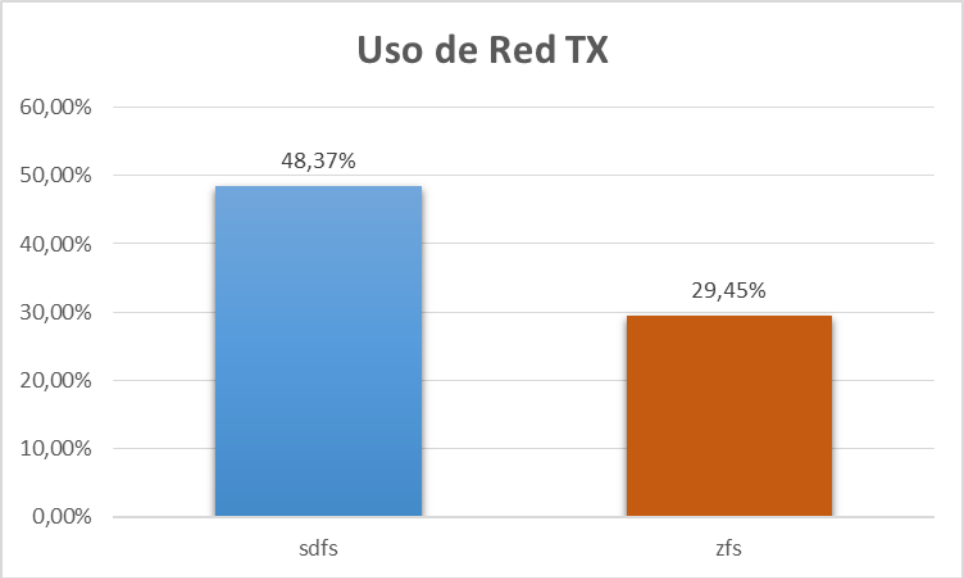

Figura 49: Consumo de red TX (Descarga de archivos desde volúmenes SDFS y ZFS)
Fuente: Elaborado por el investigador.

A continuación mostramos una tabla comparativa con los valores obtenidos de cada una de las pruebas de escritura y deduplicación tanto de SDFS como ZFS (Tabla 3) (Tabla 4).

Tabla 3: Tabla comparativa de resultados SDFS vs. ZFS (Escritura en Volumen).

ESCRITURA EN VOLUMEN	EXT4 (PREDETERMINADO)	SDFS	ZFS
	SIN DEDUP	CON DEDUP	CON DEDUP
USO CPU	18,14%	24,67%	23,59%
USO RAM	93,26%	96,85%	94,99%
TIEMPO:	4 MINUTOS	8 MINUTOS	5 MINUTOS

Fuente: Elaborado por el investigador.

Tabla 4: Tabla comparativa de resultados SDFS vs. ZFS (Descarga de archivo desde el Volumen).

	SDFS		ZFS	
DESCARGA DE ARCHIVO	USO	TIEMPO	USO	TIEMPO
USO RED Tx	48,37%	8 MINUTOS	29,45%	5 MINUTOS

Fuente: Elaborado por el investigador.

CAPÍTULO V

CONCLUSIONES Y RECOMENDACIONES

5.1 Conclusiones

Tanto ZFS como SDFS deduplican información en un directorio en tiempo real, las pruebas funcionamiento y rendimiento muestran que ZFS tiene un mejor comportamiento ante SDFS, mejor optimización de recursos tales como: CPU, RAM, Red, y tiempos de descargas web más cortos, convierten a ZFS en la primera opción a utilizarse si se desea tener un FS que administre de manera eficiente un entorno de almacenamiento de información usando deduplicación de datos.

Dentro de un Repositorio virtual de distros GNU/Linux, la deduplicación es aplicable en directorios o archivos que contengan exactamente el mismo contenido puesto que la deduplicación no solo actúa verificando nombres iguales entre los ficheros, sino a nivel de contenido que sea exactamente igual en número de archivos y en tamaño de cada uno de ellos.

Los sistemas de archivos usados en GNU/Linux (ext3, ext4) actualmente tienen un funcionamiento ligeramente más rápido que los analizados en este Proyecto de Investigación, sin embargo hay que tener en cuenta que estos FS convencionales no traen en sus características de funcionamiento métodos y técnicas para realizar deduplicación de datos.

5.2 Recomendaciones

En entornos de almacenamiento de datos a gran escala se recomienda el uso de un FS que tenga en su funcionamiento técnicas para deduplicar basadas en software libre, la decisión de utilizar ZFS o SDFS dependerá del criterio de la organización, sin embargo bajo los resultados obtenidos el FS para deduplicación de información es ZFS.

Para utilizar técnicas de deduplicación en repositorios con distribuciones derivadas de otras se recomienda que el contenido de cada una de estas derivaciones se encuentre almacenado en un solo directorio para que nuestra deduplicación pueda llevarse a cabo, existen versiones modificadas de distribuciones que comparten paquetes iguales, CentOS posee versiones derivadas tales como: Elastix, ClearOS, TripBox, Blue Quartz, etc, estas son modificaciones, por lo tanto comparten en algunos casos los mismos paquetes e información.

La Deduplicación no solo está disponible para realizar actividades dentro de Mirror's o repositorios, las técnicas efectuadas por cada uno de los sistemas de archivos expuestos en esta investigación podrían tener varios fines por lo que se recomienda utilizar esta información dependiendo del entorno en donde se desee aplicar.

BIBLIOGRAFÍA

[1] S. J. Vaughan-Nichols, «ZDNet,» Linux and Open Source, 29 Julio 2011. [En línea]. Available: http://www.zdnet.com/blog/open-source/the-top-five-linux-desktop-vendors/9313.

[2] C. Álvarez, «NetApp,» 2007. [En línea]. Available: http://www.netapp.com/es/communities/tech-ontap/es-tot-bb-depublication.aspx.

[3] E. P. Estevez, «Exposición sobre implementación de mirror de repositorios de Linux y Software Libre,» 24 Abril 2014. [En línea]. Available: http://ernestoperez.com/2014/04/exposicion-sobre-implementacion-de-mirror-de-repositorios-de-linux-y-software-libre/.

[4] Jimenez, Francisco Javier Patricio, Tecnicas de Deduplicacion de Datos y Aplicacion en Librerias Virtuales de Cintas, Madrid: Universidad Politecnica de Madrid, 2009.

[5] S. Ramos, «La nueva generación de la deduplicación y compresión de datos,» Enterprise Brand Marketing Manager for Spain, Italy & Switzerland Social Media and Community Professional de Dell, 07 Febrero 2012. [En línea].

[6] A. B. &. Recovery, «How Dedupplication Benefits Companies of all sizes,» Massachusetts, 2009.

[7] M. Studio, «Maravento Studio,» 21 Mayo 2014. [En línea]. Available: http://www.maravento.com/2014/05/deduplicacion.html.

[8] C. Pardo, «Distribuciones libres de GNU/Linux,» 31 Octubre 2014. [En línea]. Available: http://www.gnu.org/distros/free-distros.es.html. [Último acceso: 2003].

[9] R. Martinez, «El rincón de Linux,» 2014. [En línea]. Available: http://www.linux-es.org/distribuciones.

[10] M. Brisse, Data deduplication. Methods for achieving data efficiency, Quantum & Gideon Senderon, NEC, 2008.

[11] L. Whitehouse, «Search Data Center,» 24 Noviembre 2008. [En línea]. Available: http://searchdatacenter.techtarget.com/es/consejo/Metodos-de-deduplicacion-de-datos-por-bloques-o-por-bytes.

[12] EMC, «EMC.com,» 2014. [En línea]. Available: http://mexico.emc.com/corporate/glossary/data-deduplication.htm.

[13] D. B. &. L. Freeman, «Understanding Data Deduplication,» SNIA Technical Tutorials, 2009.

[14] J. Bonwick, «https://blogs.oracle.com/bonwick/entry/zfs_dedup,» Oracle Bogs, 01 Noviembre 2009. [En línea]. Available: https://blogs.oracle.com/bonwick/entry/zfs_dedup. [Último acceso: 2009].

[15] D. Hamilton, «Deduplication Methods for Achieving Data Efficiency,» SNIA Technical Tutorials, 2008.

[16] G. Lu, «Libraries Digital Conservancy Minnesota University,» 01 2012. [En línea]. Available: http://purl.umn.edu/120894.

[17] Erik Kruus, Cristian Ungureanu, Cezary Dubnicki, «Bimodal content defined chunking for backup streams,» *FAST'10 Proceedings of the 8th USENIX conference on File and storage technologies* , p. 18, 2010.

[18] S. Microsoft, «Exchange single-instance storage and its effect on stores when moving mailboxes,» 26 05 2011. [En línea]. Available: http://support.microsoft.com/en-us/kb/175481/es.

[19] S. Software, «Sherpa Software Blog,» A MIcrosoft Exchange history lesson & fun factoids, [En línea]. Available: http://www.sherpasoftware.com/blog/microsoft-exchange-history-lesson/.

[20] M. Domínguez, «OpenExpo,» 23 06 2014. [En línea]. Available: http://www.openexpo.es/blog/seguridad-blog/la-deduplicacion-en-el-backup.

[21] I. Casajús, «Another vSolutions.es Blog…,» 28 10 2009. [En línea]. Available: https://murchan.wordpress.com/2009/10/28/compresion-vs-deduplicacion-compress-or-dedupe-in-primary-storage/.

[22] B. M. Posey, «Tech Target,» 2013. [En línea]. Available: http://searchdatabackup.techtarget.com/feature/Deduping-101-What-you-must-know-before-buying-a-deduplication-product.

[23] C. Keegan, «Source vs Target Based Data Deduplication,» 02 01 2013. [En línea]. Available: http://www.storage-switzerland.com/Articles/Entries/2013/1/2_Source_vs_Target_Based_Data_Deduplication.html.

[24] O. S. Org., «Software libre y de código abierto,» [En línea]. Available: https://es.opensuse.org/Software_libre_y_de_c%C3%B3digo_abierto.

[25] Josh Lerner, Jean Tirole, «Some Simple Economics of Open Source,» *The Journal of Industrial Economics,* vol. 50, nº 2, p. 234, 2010.

[26] «Open source data de-duplication & data tiering for less,» [En línea]. Available: http://www.lessfs.com/wordpress/.

[27] G. Dedup, «Global inline deduplication for Block Storage and Files,» [En línea]. Available: http://www.opendedup.org/.

[28] D. T. Oracle, «Haga realidad la relación superior ventaja/precio de ZFS Storage Appliance de Oracle,» Octubre 2013. [En línea]. Available: http://www.oracle.com/us/try/business-value-wp-2292382.pdf.

[29] R. Carlos, «Tecnologia Pyme,» 28 Junio 2009. [En línea]. Available: http://www.tecnologiapyme.com/hardware/que-es-la-deduplicacion-de-datos.

[30] J. Schiff, «Open Source Deduplication: Ready for Enterprises?,» Enterprise Storage Forum, 13 Mayo 2010. [En línea]. Available: http://www.enterprisestorageforum.com/continuity/features/article.php/3882106/Open-Source-Deduplication-Ready-for-Enterprises.htm.

[31] Varios, «FileSystems,» ArchLinux, 29 Agosto 2015. [En línea]. Available: https://wiki.archlinux.org/index.php/File_systems_%28Espa%C3%B1ol%29.

[32] T. G. project, «GNOME,» 2015. [En línea]. Available: http://www.gnome.org.

[33] J. R. David H. Rhodes Clymer - Gruher, «Sourceforge.net,» 27 04 2013. [En línea]. Available: http://sourceforge.net/p/lessfs/mailman/lessfs-users/.

[34] «LessFS & Btier,» 09 2011. [En línea]. Available: http://www.lessfs.com/wordpress/?p=649.

[35] G. M. Solis, «OneDrive,» 17 08 2015. [En línea]. Available: https://onedrive.live.com/redir?resid=B56EE7ED4C74CDE2%21516.

[36] L. Whitehouse, «Enterprise Strategy Group and covers data protection technologies,» Analista, 2008.

[37] C. Gonzalez, «Constant Thinking,» Julio 2011. [En línea]. Available: http://constantin.glez.de/blog/2011/07/zfs-dedupe-or-not-dedupe.

ANEXOS

Anexo 1.

Entrevista realizada al administrador de red y laboratorios de la FISEI para recopilación de información

ANÁLISIS DEL ESTADO ACTUAL DE LOS SERVIDORES DE ALMACENAMIENTO DE LA FACULTAD DE INGENIERÍA EN SISTEMAS ELECTRÓNICA E INDUSTRIAL DE LA UNIVERSIDAD TÉCNICA DE AMBATO

Entrevista N.1

Nombre del Encuestado:.....

Departamento/Área:.....

Cargo que desempeña:.....

Fecha:.....

Preguntas:

1.- En la FISEI se dispone de servidores de almacenamiento de datos dedicados? Cuales son y qué características tienen?

Marca:	
Modelo:	
Procesador:	
Memoria:	
Disco (rpm) Velocidades:	
Almacenamiento Capacidad:	
Tecnología Raid usada:	

2.- Estime un valor aproximado del uso del espacio del almacenamiento global disponible en los servidores actualmente.

.....%

3.- Qué tipo de dado se considera que es el más guardado en los servidores?

Docentes: .....

Estudiantes:.....

Otros:.....

4.- Cree Ud., que en los servidores de almacenamiento existan datos duplicados o redundantes?.

Si ..... **No**.....

5.- Utiliza técnicas de depuración para contrarrestar la duplicidad de información?

Si ..... **No**.....

6.- Conoce alguna técnica adecuada para contrarrestar la duplicidad de datos?

Si..... **Cual?** ..... **No**.....

7.- Tiene conocimiento de lo que es la De duplicación de Datos.

Si........ **No**.....

8.- Referente a S. O de código abierto, que Distribución de Linux es la más usada por el personal de la Facultad.

Docentes:.....

Estudiantes:.....

Otros:.....

Firma del Encuestado:

.....

Gracias por haber participado con la encuesta.

Anexo 2.

Código que permite la obtención de valores de rendimiento de funcionamiento de hardware mediante el comando SAR.

```
#!/bin/bash
Path_toma='/media/tesiszfs/'
g/'
Fecha=$(date +%Y-%m-%d-%H:%M)
Nombre_archivo1='CPU'
Nombre_archivo2='Mem'
Nombre_archivo3='RED'

echo ...... Inicia sar
sleep 5s

sar -u 1 120 > "$Path_toma$Nombre_archivo1$Fecha" &
sar -r 1 120 > "$Path_toma$Nombre_archivo2$Fecha" &
sar -d 1 120 > "$Path_toma$Nombre_archivo3$Fecha" &
sar -n ALL 1 120 > "$Path_toma$Nombre_archivo4$Fecha" &


echo "$Path_toma$Nombre_archivo$Fecha"


sleep 1s
```

Anexo 3.

Código que permite recuperar valores y realizar la gráfica estadística delos valores recuperados por el comando SAR, utilizando la herramienta GNU Plot.

```
#!/bin/bash
function plot {
cat $PLT_TEMPLATE | sed "s/OUTPUTFILE/$2/g" | sed
"s/INPUTFILE/$1/g" | sed "s/TITLE/$3/g" | sed "s/UNITS/$4/g" |
sed "s/MAXY/$5/g"  > tmp.metric.plt

gnuplot  tmp.metric.plt
ps2pdf $2.ps
rm -f $2.ps
rm -f tmp.metric.plt
}

PLT_TEMPLATE="metric_graph.plt"

plot "cpu.dat" "cpu" "Uso de la CPU por sistema de archivos"
"Uso de la CPU (%)" ""

plot "mem.dat" "mem" "Uso de la Memoria por sistema de archivos"
"Uso de la Memoria (%)" ""

plot "net_tx.dat" "net_tx" "Uso de la Red txKB por sistema de
archivos y tiempo de retardo" "Uso de la red (%)Kbits\/second"
```