

UNIVERSIDAD TÉCNICA DEL NORTE



Facultad de Ingeniería en Ciencias Aplicadas
Carrera de Ingeniería en Sistemas Computacionales

DISEÑO DE UNA ARQUITECTURA TECNOLÓGICA QUE PERMITA FORTALECER EL PROCESO DE ANALÍTICA WEB EN EL STATUS DE PORTALES EDUCATIVOS DE ALTO TRÁFICO.

Trabajo de grado previo a la obtención del título de Ingeniero en Sistemas Computacionales

Autor:
Mario Alexander Palacios Acosta

Tutor:
MSc. Alexander Guevara

Ibarra – Ecuador

2022



UNIVERSIDAD TÉCNICA DEL NORTE
BIBLIOTECA UNIVERSITARIA
AUTORIZACIÓN DE USO Y PUBLICACIÓN
A FAVOR DE LA UNIVERSIDAD TÉCNICA DEL NORTE

IDENTIFICACIÓN DE LA OBRA

En cumplimiento del Art. 144 de la Ley de Educación Superior, hago la entrega del presente trabajo a la Universidad Técnica del Norte para que sea publicado en el Repositorio Digital Institucional, para lo cual pongo a disposición la siguiente información.

Datos del contacto	
CEDULA DE IDENTIDAD:	0401303326
APELLIDOS Y NOMBRES:	Palacios Acosta Mario Alexander
DIRECCIÓN:	San Pedro de Huaca, calle Juan Montalvo y 8 de diciembre
EMAIL:	mapalaciosa@utn.edu.ec
TELÉFONO MÓVIL:	0995213046

DATOS DE LA OBRA					
TÍTULO:	DISEÑO DE UNA ARQUITECTURA TECNOLÓGICA QUE PERMITA FORTALECER EL PROCESO DE ANALÍTICA WEB EN EL STATUS DE PORTALES EDUCATIVOS DE ALTO TRÁFICO.				
AUTOR (ES):	PALACIOS ACOSTA MARIO ALEXANDER				
FECHA:	29/03/2022				
PROGRAMA:	<table border="1"><tr><td><input checked="" type="checkbox"/></td><td>PREGRADO</td><td><input type="checkbox"/></td><td>POSGRADO</td></tr></table>	<input checked="" type="checkbox"/>	PREGRADO	<input type="checkbox"/>	POSGRADO
<input checked="" type="checkbox"/>	PREGRADO	<input type="checkbox"/>	POSGRADO		
TÍTULO POR EL QUE OPTA:	INGENIERO EN SISTEMAS COMPUTACIONALES				
DIRECTOR:	MSc. ALEXANDER GUEVARA				

1. CONSTANCIAS

El autor manifiesta que la obra la obra objeto de la presente autorización es original y se desarrolló, sin violar derechos de auto de terceros, por lo tanto, la obra es original y que es el titular de los derechos patrimoniales, por lo que asume la responsabilidad sobre el contenido de esta y saldrá en defensa de la Universidad Técnica del Norte en caso de reclamación por parte de terceros.

Ibarra, 31 de mayo del 2022

EL AUTOR:



.....

Mario Alexander Palacios Acosta
0401303326



UNIVERSIDAD TÉCNICA DEL NORTE
FACULTAD DE INGENIERÍA EN CIENCIAS APLICADAS

Ibarra, 29 de marzo de 2022

CERTIFICACIÓN DEL DIRECTOR

El Sr. Mario Alexander Palacios Acosta portador de la cedula de ciudadanía número 0401303326, ha trabajado en el desarrollo del proyecto de grado **"DISEÑO DE UNA ARQUITECTURA TECNOLÓGICA QUE PERMITA FORTALECER EL PROCESO DE ANALÍTICA WEB EN EL STATUS DE PORTALES EDUCATIVOS DE ALTO TRÁFICO."**, previo a la obtención del Título de Ingeniero en Sistemas Computacionales realizado con interés profesional y responsabilidad que certifico con honor de verdad.

Es todo en cuanto puedo certificar a la verdad

Atentamente

.....
MSc. ALEXANDER GUEVARA

DIRECTOR DE TRABAJO DE GRADO



UNIVERSIDAD TÉCNICA DEL NORTE

Resolución No 173-SE-33- CACES – 2020

DEPARTAMENTO DE DESARROLLO TECNOLÓGICO E INFORMÁTICO

DIRECTOR DE LA DIRECCIÓN DE DESARROLLO TECNOLÓGICO E INFORMÁTICO

CERTIFICA

QUE: El señor **MARIO ALEXANDER PALACIOS ACOSTA** con cédula identidad 0401303326 estudiante de la Facultad de Ingeniería en Ciencias Aplicadas – de la Carrera de Ingeniería en Sistemas Computacionales, ha desarrollado con los datos entregados de la Dirección de Desarrollo Tecnológico e Informático, el Proyecto de Tesis **"DISEÑO DE UNA ARQUITECTURA TECNOLÓGICA QUE PERMITA FORTALECER EL PROCESO DE ANALÍTICA WEB EN EL STATUS DE PORTALES EDUCATIVOS DE ALTO TRÁFICO"**.

QUE: El señor Palacios, entrega el proyecto, manuales el informe técnico de investigación, a la ingeniera Gabriela Cárdenas. – funcionario de la Dirección de Desarrollo Tecnológico e Informático el 25 de marzo del 2022.

Es todo cuanto puedo certificar, facultando al interesado hacer uso de este certificado como estime conveniente, excepto para trámites judiciales.

Ibarra, 28 de marzo del 2022

Atentamente
CIENCIA Y TÉCNICA AL SERVICIO DEL PUEBLO


Ing. Juan Carlos García
DIRECTOR



DEDICATORIA

Dedico el presente trabajo de titulación, a todos los que me apoyaron en todo momento, especialmente a mis padres y hermanos que han sido un pilar fundamental en mi formación personal y profesional, por lo cual este logro fue posible.

Muchas gracias.!

Mario Alexander Palacios Acosta

AGRADECIMIENTOS

El primer lugar agradecer a Dios por cuidar de todos mis allegados y por darme valor y fortaleza para seguir adelante.

A mis padres, por todo el apoyo brindado en este arduo camino de formación profesional.

A los profesionales que conforman la Dirección de Desarrollo Tecnológico e informático de la UTN, principalmente a los ingenieros, Ing. Juan Carlos García, Ing. Gabriela Cárdenas y MSc. Vinicio Guerra, por toda la colaboración prestada en este proceso.

A la Universidad Técnica del Norte, a todos los que conforman parte de la Carrera de Ingeniería en Sistemas Computacionales, quienes fueron parte de esta formación, ya que mediante sus enseñanzas y guías se ha podido llegar a la culminación de la carrera.

Un agradecimiento especial a mi tutor MSc. Alexander Guevara y asesores MSc. Carpio Pineda y MSc. Antonio Quiña, quienes gracias a su guía y apoyo se logró culminar el presente trabajo de titulación.

A mis compañeros y amigos que con el tiempo compartido en las, formaron a ser una parte fundamental en mi vida estudiantil

Mario Alexander Palacios Acosta

Índice de Contenido

IDENTIFICACIÓN DE LA OBRA.....	I
DEDICATORIA	V
AGRADECIMIENTOS.....	VI
Índice de Contenido	VII
Índice de Figuras	XI
Índice de Tablas.....	XIII
RESUMEN.....	XIV
ABSTRACT.....	XV
INTRODUCCIÓN.....	1
Antecedentes.....	1
Situación Actual	3
Prospectiva.....	3
Planteamiento del Problema	3
Objetivo General.....	4
Objetivos Específicos:.....	4
Alcance.....	5
Justificación	6
CAPÍTULO I	7
1. Marco Teórico	7
1.1. Introducción a la minería de Datos Web.	7
1.1.1. Minería de Datos Web.....	7
1.1.2. Web Usage Mining o Minería de Uso Web.	9
1.1.2.1. Archivos .log.	10
1.2. Técnicas empleadas en la Minería de Uso de Web	12
a. Agrupamiento y clasificación	12
b. Reglas de asociación.....	13
c. Secuencias Frecuentes	13
1.3. Arquitectura tecnológica	14
1.4. Proceso de Descubrimiento KDD.	15
1.5. Etapas del proceso KDD	17
1.5.1. Fase de integración y recopilación.....	17
1.5.2. Fase de selección, limpieza y transformación.....	18
1.5.3. Fase de minería de datos	21
1.5.4. Fase de evaluación e interpretación	22

1.6.	Tareas y modelos predictivos	24
1.6.1.	Clasificación	24
1.6.2.	Regresión	25
1.7.	Norma ISO/IEC/IEEE4210:2011- Sistemas de ingeniería de software - Descripción Arquitectura	27
1.7.1.	Historia y evolución de la norma ISO/IEC/IEEE4210:2011	28
1.7.2.	Términos y definiciones de Norma ISO/IEC/IEEE4210:2011.....	29
1.7.3.	Organigrama ISO/IEC/IEEE 42010.....	30
1.7.4.	Estructura General	31
CAPÍTULO II	32
2.	Desarrollo.....	32
2.1.	Portal web institucional	32
2.1.1.	Definición.....	32
2.1.2.	Funciones.....	32
2.1.3.	Descripción del portal Web UTN.....	33
2.2.	Gestión del proyecto utilizando ISO/IEC/IEEE 42010	34
2.3.	Fase de diseño de la arquitectura.....	34
2.4.	Open Group Architecture Framework (TOGAF).....	35
2.4.1.	Fases de TOGAF	35
2.5.	Fase preliminar.....	36
2.5.1.	Principios de Arquitectura.....	36
2.5.1.1.	Resumen de Principios de arquitectura.....	36
2.5.1.2.	Principios de Negocio	37
2.5.1.3.	Principio de Datos.....	38
2.5.1.4.	Principio de Aplicaciones	39
2.5.1.5.	Principio de Tecnología.....	41
2.6.	Petición de trabajo de arquitectura	43
2.6.1.	Limitaciones Financieras	43
2.6.2.	Descripción de la situación actual del Negocio	43
2.6.3.	Proceso de Estimación	43
2.6.4.	Proceso de Ejecución	43
2.7.	Descripción de la situación actual de la analítica web.....	43
2.8.	Visión de arquitectura	44
2.8.1.	Declaración de trabajo de arquitectura	44
2.8.2.	Descripción del proyecto de arquitectura y alcance	45
2.8.3.	Roles responsabilidades y entregables	46
2.8.4.	Descripción del problema	47

2.8.4.1.	Interesados y sus preocupaciones	47
2.8.4.2.	Lista de asuntos y escenarios que deben abordarse.....	47
2.9.	Arquitectura (AS IS / TO BE)	48
2.9.1.	Documento de definición de arquitectura.....	48
2.9.1.1.	Alcance	48
2.9.1.2.	Metas, objetivos y limitaciones	49
2.9.1.3.	Principios de arquitectura.....	49
2.10.	Arquitectura línea base.....	50
2.10.1.	Arquitectura del negocio.....	50
2.10.2.	Arquitectura de datos	54
2.10.3.	Arquitectura de aplicaciones	55
2.10.4.	Arquitectura tecnológica.....	55
2.10.5.	Especificaciones de hardware y red	57
2.10.6.	Fundamentos y justificación del enfoque arquitectónico.....	58
2.11.	Arquitectura de destino	60
2.11.1.	Arquitectura de negocio	60
2.11.2.	Mapa de procesos de negocio	60
2.11.3.	Diagrama de actividades (solución).....	60
2.11.4.	Arquitectura de datos	61
2.11.5.	Arquitectura de aplicaciones	62
2.11.6.	Arquitectura Tecnológica.....	63
2.11.7.	Especificación de hardware y red.....	64
2.12.	Análisis de brechas.....	64
2.12.1.	Arquitectura de Negocio.....	64
2.12.2.	Arquitectura de aplicaciones	65
2.12.3.	Arquitectura tecnológica.....	66
2.13.	Oportunidades y soluciones.....	67
2.13.1.	Plan de implementación y migración	67
2.14.	Solución (Metamodelo).....	69
2.15.	Fase de evaluación de la arquitectura	69
2.16.	Fase de implementación de la arquitectura	71
2.17.	Instalación y prueba de herramientas	71
2.17.1.	Descripción de las herramientas de análisis de archivos .log.....	72
2.18.	Proceso de analítica Web	73
2.18.1.	Herramienta RStudio.....	73
2.18.1.1.	Fase de integración y recopilación de Datos.....	74
2.18.2.	Herramienta WebLog Expert	78

2.18.3.	Herramienta Screaming Frog SEO Log File Analyser	84
2.18.4.	Herramienta LogFile Analyzer/ Semrush	92
2.19.	Evaluación de funcionalidad de herramientas	99
CAPÍTULO III	101
3.	Validación de resultados	101
3.1.	Sección 1. Datos informativos	102
3.2.	Sección 2. Variables del modelo de éxito de los sistemas de información de Delone y Mclean	103
3.2.1.	Calidad del Software	104
3.2.2.	Calidad de la información	104
3.2.3.	Calidad del Servicio.....	105
3.2.4.	Uso – intención de Uso	106
3.2.5.	Satisfacción de usuario	107
3.2.6.	Beneficios obtenidos	108
3.3.	Calificación a la herramienta WebLog expert.....	108
	CONCLUSIONES	110
	RECOMENDACIONES	111
	BIBLIOGRAFÍA	112
	ANEXOS	114

Índice de Figuras

Fig. 1	Diagrama de causas y efectos.....	4
Fig. 2	Diagrama proceso KDD.....	5
Fig. 3	Clasificación minería web.....	9
Fig. 4	Etapas técnicas de desarrollo.....	14
Fig. 5	Fases del Proceso KDD.....	15
Fig. 6	Naturaleza cíclica del proceso de descubrimiento.....	16
Fig. 7	Extraction-Transformation-Load.....	17
Fig. 8	Tareas de minería de datos.....	21
Fig. 9	Fase de evaluación e interpretación.....	22
Fig. 10.	Línea de tiempo ISO/IEC/IEEE 42010:2011.....	28
Fig. 11	Organizador gráfico ISO/IEC/IEEE 42010.....	30
Fig. 12	Descripción de la Arquitectura.....	31
Fig. 13	Implementación servidor Web.....	33
Fig. 14	Estructura Capítulo 2.....	34
Fig. 15	Diseño de arquitectura.....	34
Fig. 16.	Fases Descripción de Arquitectura.....	35
Fig. 17.	Diagrama de Análisis del portal web UTN.....	45
Fig. 18.	Estructura Organizacional UTN.....	50
Fig. 19.	Organigrama Departamento Informático UTN.....	50
Fig. 20.	Mapa de procesos UTN.....	51
Fig. 21.	Modelo de Datos Análisis web.....	52
Fig. 22.	Diagrama del proceso de analítica web.....	53
Fig. 23.	Diagrama Lógico.....	54
Fig. 24.	Diagrama de Arquitectura de Aplicaciones.....	55
Fig. 25.	Diagrama de componentes.....	55
Fig. 26.	Oracle Cloud Infraestructura.....	56
Fig. 27.	Diagrama de hardware y red.....	57
Fig. 28.	Diagrama de Actividades.....	61
Fig. 29.	Modelo Lógico (solución).....	61
Fig. 30.	Diagrama de Arquitectura de Aplicaciones.....	62
Fig. 31	Arquitectura Tecnológica.....	63
Fig. 32.	Estructura Archivos .log.....	66
Fig. 33.	Estructura y desglose del trabajo.....	67
Fig. 34.	Plan de implementación.....	68
Fig. 35.	Metamodelo analítico web.....	69
Fig. 36	Gráfico de cumplimiento de principios del estándar ISO/IEC/IEEE 42010.....	71
Fig. 37	Herramientas de Análisis de archivos .log.....	72
Fig. 38.	Descripción herramientas de análisis de archivos .log.....	73
Fig. 39	Selección de datos a analizar.....	76
Fig. 40	Identificación de columnas a utilizar.....	77
Fig. 41	Limpieza de columnas.....	77
Fig. 42	Filtrado de atributos.....	78
Fig. 43	Crear nuevo perfil.....	79
Fig. 44	Ingreso de perfil y dominio.....	80
Fig. 45	Carga de archivos .log.....	80
Fig. 46	Selección de archivos .log.....	81
Fig. 47	Carga archivos .log.....	81
Fig. 48	Selección de rango de hora y fecha.....	82
Fig. 49	Análisis de datos.....	82

Fig. 50 Sumario de análisis	83
Fig. 51 Importar archivos log	85
Fig. 52 Selección archivos log	86
Fig. 53 Asignación de nombre y selección de uso horario	86
Fig. 54 URL -UTN.....	86
Fig. 55 Importar archivos log	87
Fig. 56 Visión general.....	87
Fig. 57 Información de URLs encontradas.....	88
Fig. 58 Códigos de respuesta.....	88
Fig. 59 Agentes de Usuario	89
Fig. 60 Referencias	89
Fig. 61 Directorios de acceso	90
Fig. 62 IPs Host remoto.....	90
Fig. 63 Eventos	91
Fig. 64 Eventos 2	91
Fig. 65 Importar datos de URL	92
Fig. 66. Pantalla Principal Log File Analyzer/Semrush.....	93
Fig. 67. Botón Buscar archivos log	93
Fig. 68. Crear cuenta.....	94
Fig. 69. Selección archivos .log	94
Fig. 70. Selección de archivos desde ruta especificada.....	95
Fig. 71. Barra de proceso de subida de archivos log	95
Fig. 72 Botón Start Log File Analyzer	96
Fig. 73. Proceso de análisis de archivos .log.....	96
Fig. 74. Reporte de análisis de archivos .log	97
Fig. 75. Vistas por pagina.....	98
Fig. 76 Edad.....	102
Fig. 77. Género	102
Fig. 78. Relación con el proyecto	103
Fig. 79. Calidad del Software.....	104
Fig. 80. Calidad de la información	105
Fig. 81. Calidad del servicio.....	106
Fig. 82. Uso e intención de uso	106
Fig. 83. Satisfacción del usuario.....	107
Fig. 84. Beneficios obtenidos.....	108
Fig. 85. Calificación de la herramienta.....	108
Fig. 86. Encuesta	114

Índice de Tablas

Tabla 1 Matriz de confusión.....	23
Tabla 2 Términos y definiciones norma ISO/IECE/IEEE42010	29
Tabla 3. Resumen de Principios de Arquitectura	36
Tabla 4. Alineación entre TI y negocio.....	37
Tabla 5. Enfoque al cliente	37
Tabla 6. Enfoque a largo plazo.....	37
Tabla 7. Información Relevante.....	38
Tabla 8. Información accesible	38
Tabla 9. Seguridad de la Información	39
Tabla 10. Copia de Seguridad de los datos	39
Tabla 11. Seguimiento de Estándares.....	39
Tabla 12. Independencia de la Tecnología	40
Tabla 13. Aplicaciones Fáciles de usar	40
Tabla 14. Tecnología madura.....	41
Tabla 15. Infraestructura escalable.....	41
Tabla 16. Reevaluar la seguridad.....	42
Tabla 17. Seguimiento.....	42
Tabla 18. Matriz Roles responsabilidades y entregables	46
Tabla 19. Descripción de Hechos y problemas.....	47
Tabla 20. Principios de Arquitectura	49
Tabla 21. Roles y partes interesadas	51
Tabla 22. Descripción componentes de diagrama lógico	54
Tabla 23. Descripción de componentes principales.....	58
Tabla 24. Requerimientos	59
Tabla 25. Descripción Modelo Lógico.....	62
Tabla 26. Análisis de Brechas proceso de analítica web	64
Tabla 27. Arquitectura de Aplicaciones	65
Tabla 28 Checklist de evaluación de cumplimiento de principios del estándar ISO/IEC/IEEE 42010.....	69
Tabla 29 Descripción herramientas análisis archivos .log.....	72
Tabla 30 Requisitos mínimos de instalación RStudio.	73
Tabla 31. Estructura archivos .log	74
Tabla 32 Requisitos mínimos de instalación Weblog expert	79
Tabla 33 Códigos de estados de respuesta http encontrados	83
Tabla 34 Requisitos de instalación Screaming Frog Log Analyzer.....	84
Tabla 35 Comparativa de funcionalidad de las herramientas.....	99

RESUMEN

Mediante el presente proyecto se tiene como objetivo diseñar una arquitectura tecnológica institucional bajo los principios de la norma ISO/IEC/IEEE 42010, que permita obtener un modelo arquitectónico estandarizado de acuerdo con los requerimientos institucionales, que permita fortalecer el proceso de analítica web, para determinar el status del portal Web de la Universidad Técnica del Norte.

Capítulo I: Este capítulo está conformado por la revisión bibliográfica correspondiente a la minería de datos web, técnicas utilizadas en la minería de uso web, conceptos de arquitectura tecnológica, proceso y etapas de KDD, tareas y modelos predictivos y la conformación y descripción del a norma ISO/IEC/IEEE 4210:2011

Capítulo II: Se realizó el proceso de analítica web que se sigue para determinar el estado de salud del portal Web de la UTN.

Siguiendo los principios y requerimientos del estándar ISO/IEC/IEEE 42010:2011, se realizó el diseño del modelo de arquitectura tecnológica, basado en el marco de trabajo TOGAF 9.2, luego de esto se puso a prueba varias herramientas que nos permiten realizar el análisis de archivos .log, provenientes del portal web de la UTN.

Capítulo III: Después de haber realizado las pruebas correspondientes con las diferentes herramientas de análisis de archivos .log, mediante la aplicación de las variables del modelo de éxito de los sistemas de información de Delone y Mclean se logró determinar cuál de ellas cumplen con los requerimientos de los stakeholders.

Finalmente, se describe las conclusiones y recomendaciones que se pudo llegar al realizar este proyecto.

Palabras clave: arquitectura de sistemas, analítica web, log, ISO/IEC/IEEE 42010, Web portal, TOGAF 9.2

ABSTRACT

Through this project, the objective is to design an institutional technological architecture under the principles of the ISO/IEC/IEEE 42010 standard, which allows obtaining a standardized architectural model under institutional requirements, which allows strengthening the web analytics process, to determine the status of the Web portal of the Universidad Técnica del Norte.

Chapter I: This chapter is made up of the bibliographic review corresponding to web data mining, techniques used in web use mining, concepts of technological architecture, process and stages of KDD, tasks and predictive models, and the conformation and description of an ISO/IEC/IEEE 4210:2011 standard.

Chapter II: The web analytics process followed to determine the health status of the UTN Web portal was conducted.

Following the principles and requirements of the ISO/IEC/IEEE 42010:2011 standard, the design of the technological architecture model was carried out, based on the TOGAF 9.2 framework, after which several tools were tested that allow us to carry out the analysis of .log files, from the UTN web portal.

Chapter III: After having carried out the corresponding tests with the different .log file analysis tools, by applying the variables of the success model of the DeLone and McLean information systems, it was possible to determine which of them meets the requirements of the stakeholders.

Finally, the conclusions and recommendations that could be reached when conducting this project are described.

Keywords: system architecture, log, web analytics, ISO/IEC/IEEE 42010, Web portal, TOGAF 9.2

INTRODUCCIÓN

Antecedentes

Gracias a la irrupción de las tecnologías de la información y comunicación (TIC), la confluencia de la Sociedad de la Información Comunicación y la Sociedad de la Imagen, la humanidad vive por y para la imagen y la información. El motor de este engranaje es el Internet.

Con la profunda transformación social que ha provocado la integración de las TIC en nuestro día a día, los modos de comunicación en el mundo educativo han evolucionado hacia el internet y a un universo en red. Este hecho ha convertido a las páginas web de los centros educativos, entre otros elementos, en herramientas catalizadoras del panorama escolar.

El gran desarrollo tecnológico producido recientemente ha propiciado lo que algunos autores denominan la nueva "revolución" social, con el desarrollo de "la sociedad de la información". Con ello, se desea hacer referencia a que la materia prima "la información" sea el motor de esta nueva sociedad, y en torno a ella, surgirán profesiones y trabajos nuevos, o se readaptarán las profesiones existentes.

El impacto de las nuevas tecnologías en la educación constituye un desafío en los procesos de enseñanzaaprendizaje en todos los niveles educativos, desde el manejo tecnológico hasta su uso en el intercambio, búsqueda y difusión de la información. Contar con los conocimientos necesarios para su adecuada gestión garantiza un proceso educativo interactivo y dinámico apegado a las competencias y exigencias actuales (Maldonado, 2015).

El crecimiento explosivo del Internet y particularmente de la World Wide Web (www), ha hecho cada vez más necesario para las empresas utilizar herramientas automatizadas para encontrar, extraer, filtrar y evaluar los recursos de información disponibles. Unido a ello y con la transformación de la Web, como la herramienta primaria para el acceso a cualquier tipo de información, se hace indispensable para las empresas que basan su negocio en Internet poder rastrear y analizar modelos de acceso de usuarios con el fin de cumplir sus objetivos y sus metas.

En los últimos años se ha desarrollado un enorme crecimiento en la capacidad de generación y almacenamiento de información, debido a la creciente automatización de procesos, en general, y a los avances tecnológicos radicados en la capacidad de almacenamiento de la información. Conjuntamente, las herramientas de software también han desarrollado un fuerte crecimiento, en el proceso de descubrir conocimiento o como es conocida, la minería de datos (data mining) ha sido definida como la identificación de patrones no triviales válidos, nuevos, comprensibles y potencialmente útiles de un conjunto enorme de datos definidos así por (Usuma Fayyad, 1996) (Fernando Berzal, 2001). Los factores antes mencionados dan lugar a la necesidad de crear sistemas inteligentes, tanto del lado cliente y del servidor, que puedan hacer búsquedas o minería en la Web para obtener conocimiento.

Web mining (minería Web) puede definirse como el descubrimiento y análisis de información útil en la World Wide Web. (IONOS, 2018) En su artículo menciona que data mining, también conocida como minería de datos, se describen los procedimientos algorítmicos para la evaluación de datos aplicados a bloques de datos con un tamaño y una complejidad determinados. Su función es la de extraer la información oculta en grandes volúmenes de datos, especialmente en las masas de información conocidas como big data, y además reconocer tendencias, relaciones y patrones ocultos en ellas. Para que este proceso se pueda llevar a cabo se recurre a la data mining tools.

De acuerdo a la investigación de (Ming-Syan Chen, 1997), existen varias técnicas para hacer minería de datos, como son: las reglas de asociación, reglas de extracción, clustering, algoritmos genéticos y redes neuronales. Cada una de ellas se aplica con menor o mayor grado de dificultad a las bases de datos relacionales, en que se ha visto que la primera técnica mencionada es más exitosa que las demás, debido a que su aplicación es inmediata en un lenguaje de cuarta generación, por otra parte, tiene la limitante de ser estrictamente predictiva y no de búsqueda. En el caso de la aplicación de inteligencia artificial es más complejo aterrizar o encontrar un camino que lleve a algoritmos de búsqueda inteligentes debido a que no existe una completa conexión entre las bases de datos relacionales y la inteligencia artificial (Hongjun Lu, 1996).

Situación Actual

El portal Web institucional de la UTN en los últimos años ha tenido un crecimiento exponencial por la gran cantidad de información que se genera diariamente sin control alguno de contenidos de los archivos, es por ello por lo que esto lleva al deterioro del estado de salud del portal.

Del mismo modo los datos que se encuentran almacenados en los repositorios de bases de datos de la UTN en ciertos casos puntuales presentan inconsistencias, tales como documentación duplicada, documentación dañada, entre otros. Por este motivo, se vuelve indispensable establecer patrones o metodologías de análisis, limpieza y en muchos casos reestructuración de los datos que se generan, para así mejorar el rendimiento de la página web institucional. Para llevar a cabo la propuesta se trabajará con los archivos log generados en el servidor Web de la Universidad Técnica del Norte que se encuentra desplegado en la nube.

Prospectiva

Con la presente investigación se pretende realizar un análisis de los datos generados en el portal web de la Universidad Técnica del Norte(UTN), al mismo tiempo saber el estado de salud de la página Web, además de analizar la Data generada, este proceso se lo podrá realizar mediante el análisis de los log files que se generan cada vez que se realiza una transacción dentro del portal, con esto se pueden recopilar datos para posteriormente generar estadísticas sobre el acceso a la página o cifras clave sobre el uso de una web o un servidor web.

Planteamiento del Problema

Existe un alto índice de saturación en el tráfico de red de los portales web educativos, caso de estudio UTN; mediante la utilización de técnicas de analítica web se quiere llegar a determinar un proceso preventivo y correctivo que nos permita determinar un status del portal web, además nos ayude a descubrir la información oculta.

En la UTN, se ha podido evidenciar que no existen métodos para analizar y controlar los tipos de archivos que se manejan dentro de la red, el estado de salud de la plataforma, esto surge por la falta de una arquitectura tecnológica basada en procesos analíticos acerca

del descubrimiento, del uso y aprovechamiento de la información guardada en los archivos .log.

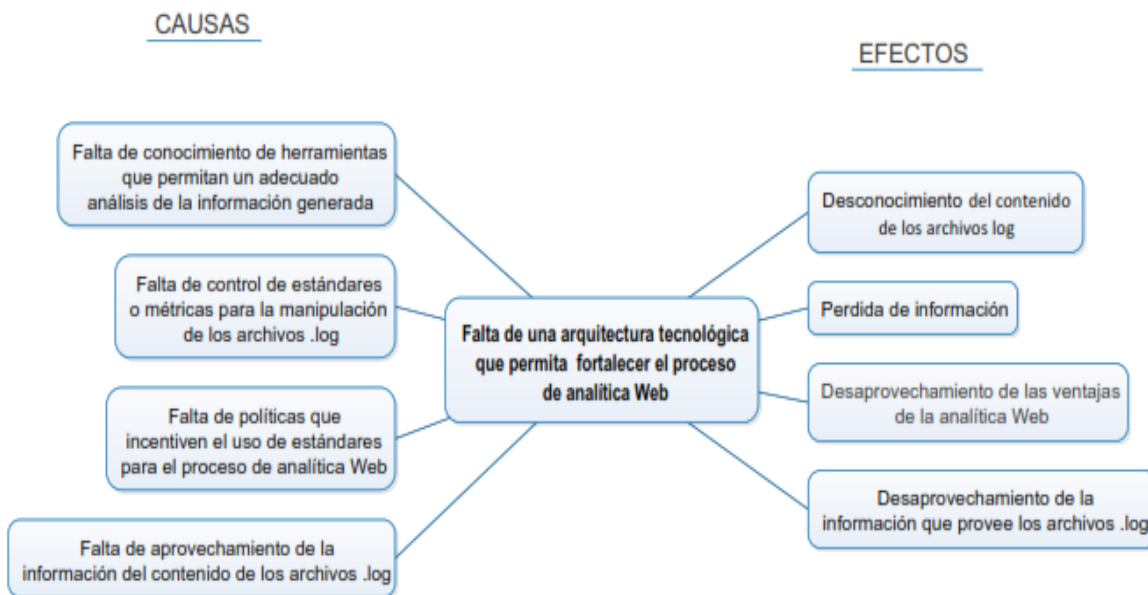


Fig. 1 Diagrama de causas y efectos
Fuente: Propia

Objetivo General

Diseñar una arquitectura tecnológica que permita fortalecer los procesos analíticos en Uso de Portales Web Educativos de alto tráfico, basados en el estándar ISO /IEC 42010 y la aplicación de Técnicas de Web Mining, caso Portal Web UTN.

Objetivos Específicos:

- Construir un marco teórico sobre la arquitectura tecnológica, para el análisis de los datos provenientes de Portales Web Educativos de alto tráfico, utilizando técnicas de Web Mining.
- Diseñar e implementar una arquitectura tecnológica para el análisis de uso del Portal Web UTN, basados en el estándar ISO /IEC 42010.
- Aplicar técnicas de Web Mining en base a la arquitectura tecnológica implementada para el análisis de los archivos LOG provenientes del Portal Web UTN enfocado a Web Usage Mining.
- Validar los resultados obtenidos de la investigación propuesta.

Alcance

El presente trabajo de investigación pretende implementar una arquitectura tecnológica, basada en el estándar ISO/IEC/IEEE 42010, la cual define diferentes características a cumplir para establecer una arquitectura estandarizada y robusta, que permita extraer patrones de grandes volúmenes de información utilizando diferentes técnicas para así fortalecer los procesos analíticos del Portal Web de la UTN.

Las técnicas utilizadas ayudarán a determinar el estado de salud del portal web de la UTN, además de permitir realizar el análisis de la estructura y contenido mediante el estudio de los archivos generados por la web (logs), aplicando procesos de descubrimiento del conocimiento KDD y técnicas de Web Mining; la validación de los resultados de la investigación se la realizará aplicando métodos formales.

La aplicación de herramientas inteligentes utiliza técnicas de extracción de conocimiento para descubrir información útil que nos permita tener indicios de acceso a la información dentro del portal Web, que permitan mejorar la seguridad del servidor Web en producción de la UTN.

La aplicación de técnicas de minería de datos se puede contemplar desde dos perspectivas complementarias:

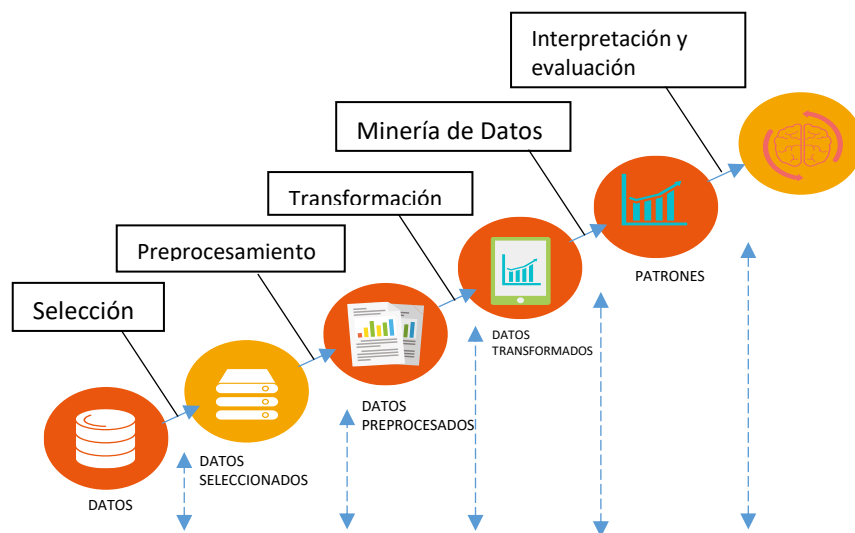


Fig. 2 Diagrama proceso KDD
Fuente: Propia, adaptada (Ordoñez, Vázquez, and grass boada 2011)

Justificación

La realización de este proyecto está enfocada principalmente en el Objetivo de Desarrollo Sostenible (ODS) 9, principalmente en los puntos 9.5 y 9b.

9.5. Aumentar la investigación científica y mejorar la capacidad tecnológica de los sectores industriales de todos los países, en particular los países en desarrollo, entre otras cosas fomentando la innovación y aumentando considerablemente, de aquí a 2030, el número de personas que trabajan en investigación y desarrollo por millón de habitantes y los gastos de los sectores público y privado en investigación y desarrollo (CEPAL/ONU, 2016).

9.b. Apoyar el desarrollo de tecnologías, la investigación y la innovación nacionales en los países en desarrollo, incluso garantizando un entorno normativo propicio a la diversificación industrial y la adición de valor a los productos básicos, entre otras cosas. (CEPAL/ONU, 2016).

Debido a la facilidad de utilización y disponibilidad de las herramientas para navegar por la Web, así como de la facilidad en el desarrollo y mantenimiento de los recursos Web. Su desarrollo ha tenido un crecimiento constante durante los últimos años y esto también ha motivado la aplicación de técnicas de minería de datos o descubrimiento de conocimiento que ya se han aplicado con éxito en sistemas de comercio electrónico o e-commerce para comprender el comportamiento de clientes en línea con la finalidad de poder incrementar las ventas.

CAPÍTULO I

1. Marco Teórico

1.1. Introducción a la minería de Datos Web.

Con la evolución, aumento y utilización de páginas web en la actualidad ha dado lugar al aumento de cantidad de datos generados que se almacenan particularmente en formato de archivos temporales(.txt,.log), que pueden proporcionar información útil para el giro de negocio de las pequeñas, medianas y grandes empresas; es por esto por lo que el análisis de la data generada se ha vuelto uno de los más grandes desafíos de la informática.

Con el continuo crecimiento y proliferación del comercio electrónico, los servicios web y sistemas de información basados en web, los volúmenes de clickstream¹ (datos de vistas) y usuario, los datos recopilados por las organizaciones basadas en la Web en sus operaciones diarias tienen alcance de proporciones astronómicas. Analizar dichos datos puede ayudar a estas organizaciones a determinar el valor de los clientes a lo largo de la vida, diseñar el marketing cruzado estrategias a través de productos y servicios, evaluar la efectividad de las campañas promocionales, optimizar la funcionalidad de las aplicaciones basadas en la Web, Proporcionar contenido más personalizado a los visitantes y encontrar el más eficaz(Mendoza 2011).

1.1.1. Minería de Datos Web

La minería de datos refiere al proceso de extracción de información en forma de modelo que se puede tomar como una característica extrema de datos(Leskovec, Rajaraman, and Ullman 2014).

De acuerdo con los estudios realizados se dice que la minería de datos utiliza apropiadamente algoritmos de aprendizaje automático. Los practicantes de aprendizaje automático usan los datos como un conjunto de entrenamiento, para entrenar un algoritmo, de los muchos existentes, como redes de Bayes, máquinas de vectores de soporte (SVM), árboles de decisión, redes neuronales artificiales (ANN), ocultos Modelos de Markov, y muchos otros(Leskovec et al. 2014).

La informatización y la recopilación automatizada de datos dieron como resultado depósitos de datos extremadamente grandes; por los problemas de escalabilidad y el deseo

de una mayor automatización hacen que las técnicas más tradicionales sean menos efectivas. (Warf 2018).

La minería de datos nos permite realizar un análisis más profundo acerca de los datos en bruto generados, siguiendo diferentes patrones, por lo tanto, aumentan el conocimiento acerca de los tipos de datos que se generan en la base de datos que hospeda una página web.

La minería web o Web Mining comprende una serie de técnicas encaminadas a obtener inteligencia a partir de datos procedentes de la web. Aunque las técnicas utilizadas tienen su raíz en las técnicas de data mining o minería de datos, presentan características propias debido a las particularidades que presentan las páginas webs.(Intelligent 2015).

a. Proceso de la Minería Web

El proceso de minería Web tiene 4 procesos para seleccionar y transformar los resultados:

La Web como ecosistema contiene y genera un universo de datos, tanto provenientes del propio contenido de sus páginas y la estructura de sus enlaces como de su uso por parte de las personas. Estos datos tienen una importancia crucial para el mejoramiento de esta, desde un punto de vista social y también comercial. Por esta razón la minería de datos de la Web ha crecido rápidamente y es una herramienta vital para entenderla y dar valor económico a los datos que obtenemos de ella(Baeza-Yates 2009).

Encontrar una definición exacta acerca de la minería web (MW) es una tarea difícil, por lo general este término se utiliza para describir o catalogar tres tipos de actividades considerablemente diferentes. Todas estas actividades se enmarcan dentro de la minería de Datos y, además, están relacionadas con la web, pero los datos que son objeto de la minería son diferentes(Martín 2004). Estas actividades son las siguientes:

- **Minería de contenido:** Engloba aquellos procesos que tienen la finalidad de extraer información a partir de información contenida en un determinado portal. Esta información puede estar en formato de texto, imágenes, etiquetas metadatos (text, images, records).

- **Minería de estructura:** Se enmarcan aquí los procesos cuyo objeto es extraer información sobre la topología de la web, es decir, de los enlaces entre las páginas (hyperlinks, tags).
- **Minería de uso:** En este caso se trata de extraer información acerca del comportamiento de los usuarios en la web, interacción de las personas con la Web analizando principalmente los archivos (http logs, app server logs, etc).

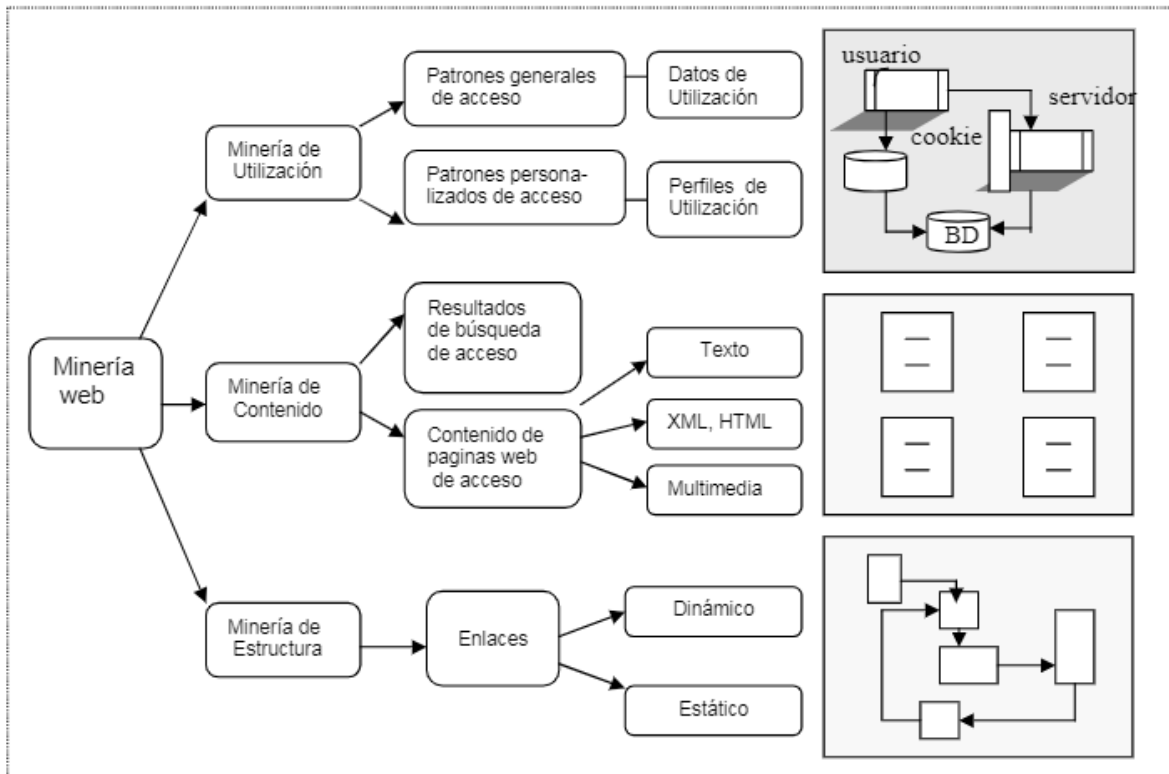


Fig. 3 Clasificación minería web
Fuente: (Ulises Román 2005)

1.1.2. Web Usage Mining o Minería de Uso Web.

Debido a la cantidad de páginas web que existen en la actualidad así mismo el enorme número de usuarios que las visitan e interactúan con ellas, existe un gran volumen de datos generados disponibles para poder ser analizados(Mendoza 2011).

La minería de datos web, es uno de los procesos que permiten registrar las transacciones informáticas que se realizan en la web, permite la personalización de archivo clientes, uso de perfiles, detección de intrusos, paginas agentes, entre otras(Mendoza 2011).

Los patrones de usuarios pueden ser obtenidos a diferentes niveles, desde la secuencia de clicks realizada por los usuarios dentro de una sesión por un usuario individual hasta el conjunto de patrones que registran las compras de determinados productos por un conjunto de usuarios durante un período de tiempo (Martín 2004). Una vez que los datos de logs están organizados es posible ejecutar(Mendoza 2011).

Los patrones detectados permiten identificar modelos de navegación de los usuarios, así como el tipo recursos disponibles en el sitio o páginas web que pueden ser de mayor interés. Esta información es de gran valor, permitiendo detectar grupos de usuarios con intereses comunes. La información obtenida podrá ser usada para automáticamente modificar el diseño del sitio de acuerdo a las necesidades de grupos específicos, lo que se conoce como personalización del sitio(Mendoza 2011).

1.1.2.1. Archivos .log.

Cada vez que el usuario accede a un sitio web este se registra en los archivos logs. Los accesos registrados corresponden a peticiones sobre recursos específicos del sitio (requests). Los requests son realizados usando un protocolo específico, que generalmente en el caso de la Web es el HTTP. Cada línea del archivo de logs almacena una petición, en la cual además se guardan otros datos, como el host desde el cual se solicitó el recurso, la fecha y hora de la solicitud y el resultado de la petición(IONOS 2016). Los campos que generalmente permiten describir una petición son:

“host_remoto logname_remoto logname_local [fecha] "request" estado bytes”

- ***host_remoto*** indica el nombre del host desde donde se ha solicitado la petición esta también puede estar determinado por la IP si el DNS no está disponible.
- ***logname_remoto*** indica el nombre que el usuario utilizo para acceder remotamente.
- ***logname_local*** indica el nombre que el usuario utilizado para realizar la autenticación.
- ***Fecha*** corresponde a la fecha y hora en la cual se realizó la petición o solicitud del servicio.
- ***request*** indica exactamente lo que el usuario solicitó.
- ***estado*** indica el código HTTP retornado al usuario
- ***bytes*** indica la cantidad de información en bytes del contenido transferido al usuario.

Si alguno de los campos no puede ser determinado, se inserte un signo menos (-).

Las peticiones realizadas a un sitio son registradas en el archivo de logs de acuerdo con la fecha y hora en la cual fueron realizadas. Al campo fecha y hora se le denomina timestamp. El timestamp de una petición indica el punto en el tiempo exacto en el cual la petición fue realizada(Mendoza 2011).

El Consorcio de la World Wide Web (W3C) ha especificado una extensión al formato common log conocida como “Extended Log File Format” (ELF), motivada principalmente por la necesidad de proveer más campos a aplicaciones Web que lo requieran. Entre los campos más relevantes destacan el campo referrer, el cual corresponde a la URL que el usuario estaba visitando cuando realizó la petición, browser, que indica desde cual browser HTML el usuario accedió al sitio, S.O., que indica cual sistema operativo era usado, y cookie, el cual se usa para indicar si el sitio visitado utilizaba cookies. Frecuentemente los campos sistema operativo y browser se expresan en un solo campo denominado agente(Mendoza 2011).

Una vez recolectados los datos de los archivos log se procede a realizar una tarea de limpieza aplicando los pasos descritos a continuación: filtrado, identificación de usuarios y determinación de sesiones.

a. Filtrado

En este paso se procede a eliminar los registros que no son realizados, como por ejemplo los log de solicitudes de imágenes que son solamente parte de la página HTML, todos los archivos que tengan extensiones “.jpg, .gif., png” son eliminados(Martín 2004). En una primera instancia el filtrado da como resultado datos a nivel de página web.

b. Identificación de usuarios

En este proceso se lleva dos niveles de identificación, en el primer nivel se identifican las peticiones de páginas realizadas por el mismo usuario durante una visita del sitio. El segundo nivel radica en reconocer a un usuario dentro de sus múltiples visitas a un determinado sitio web, con la finalidad de poder analizar el comportamiento del usuario a lo largo del tiempo, pueden ser días meses o años(Martín 2004).

Una estrategia para la identificación de usuario sería mediante un “nombre de usuario” y “contraseña”, pero como se conoce la navegación web se lleva a cabo normalmente de forma anónima, por lo resulta bastante complicado reconocer a un mismo usuario entre los diferentes servicios a los que accede dentro de una misma sesión y mucho más compleja resulta cuando se tiene en cuenta la evolución temporal.

c. Determinación de sesiones

Se determina por una serie de servicios solicitados por un mismo usuario a una única visita al sitio o portal web. Tomando en cuenta que las sesiones conforman un factor importante que a través de ellas se puede conocer la percepción del usuario con respecto a su visita al portal(Martín 2004).

La mejor solución para realizar esta actividad es mediante una aplicación que cree un identificador de sesión la primera vez que un determinado usuario acceda al portal.

1.2. Técnicas empleadas en la Minería de Uso de Web

Dentro de la minería de uso web existen técnicas de las cuales se mencionará tres de las comunes:

a. Agrupamiento y clasificación

Las técnicas de agrupamiento o clustering distribuyen comportamientos de individuos similares en grupos con características comunes. Esto es, dos elementos con características parecidas pertenecerán al mismo grupo, y las características de un grupo (definidas por el elemento prototipo o ideal) serán diferentes a los de otro grupo. Dependiendo de la información almacenada en los ficheros log, es posible detectar grupos de usuarios como:

- Aquellos usuarios que visitan gran cantidad de páginas con un tiempo de permanencia similar en todas ellas.
- Los que visitan un número reducido de sitios web en sesiones cortas.
- Los que visitan un número pequeño – mediano de sitios web con tiempo variable en cada una de ellas.
- Luego de haber descubiertos los prototipos o perfiles a cada grupo, se pueden usar las características de cada uno de ellos para realizar clasificación.

Dentro de la Minería de Usos Web, las técnicas de agrupamiento y clasificación nos permiten crear perfiles para cada uno de los usuarios o clientes tomando en cuenta sus patrones de acceso, todo esto puede ser utilizado por las empresas para desarrollar y ejecutar estrategias de mercadeo futuro. Esto puede ser aplicado tanto on – line como off – line; tales como envío de correo automáticos a aquellos clientes / usuario que se encuentren dentro de un cierto grupo, reasignación dinámica de servidor para un cliente (por ejemplo: menos sobrecargado, para darle un mejor servicio), o presentación de contenidos específicos según el tipo de cliente (Ulises Román 2005).

b. Reglas de asociación

Las reglas de asociación permiten capturar patrones referentes a los *itemsets* sin distinción en la que ocurren en una transacción de datos.

A través del análisis de asociaciones podemos descubrir las relaciones sin que exista intervención alguna por parte del operador. El descubrimiento de estas reglas ayuda a las organizaciones dedicadas a e-commerce a definir sus estrategias de mercados efectivos. El aprendizaje Reglas de Asociación se divide normalmente en 2 fases: Fase 1. La extracción de los conjuntos de los items que cumplen con la cobertura requerida desde los datos y Fase 2. La generación de las reglas a partir de estos documentos (Ulises Román 2005).

c. Secuencias Frecuentes

La minería de secuencias puede ser considerada como asociación minando sobre los datasets temporales y una secuencia de lista ordenada (con el paso del tiempo) de itemsets no vacíos.

El objetivo de esta técnica es descubrir el tiempo de las secuencias ordenadas de URLs que ha sido seguido por usuarios, para predecir a futuros. En general en las bases de datos Transaccionales, se tienen disponibles los datos en un periodo de tiempo y se cuenta con la fecha en que se realizó la transacción.

El descubrimiento de patrones de secuencia – *sequential patterns* en el log puede ser utilizado para predecir las futuras visitas y así poder organizar mejor los accesos y publicidades para determinados periodos de tiempo.

1.3. Arquitectura tecnológica

Dentro del desarrollo de sistemas de software se llevan a cabo distintas actividades técnicas que se describen a continuación (Cervantes Maceda, Velasco-Elizondo, and Luis 2016):

- a. **Requerimientos.** - Se refiere a las especificaciones y requerimientos del cliente de cómo va a estar estructurado el sistema.
- b. **Diseño.** – en esta etapa se transforma los requerimientos en un diseño en donde se construye un modelo, el cual represente el funcionamiento del sistema.
- c. **Construcción.** – Se refiere a la creación del sistema mediante el desarrollo del software y prueba individual de todas las partes que lo componen
- d. **Pruebas.** - Etapa en las que se realiza las pruebas correspondientes al sistema y permite determinar los posibles fallos que esté presente.
- e. **Implantación.** – En esta etapa se realiza la transacción del sistema desde el entorno de desarrollo hacia el entorno en donde se ejecutará definitivamente.

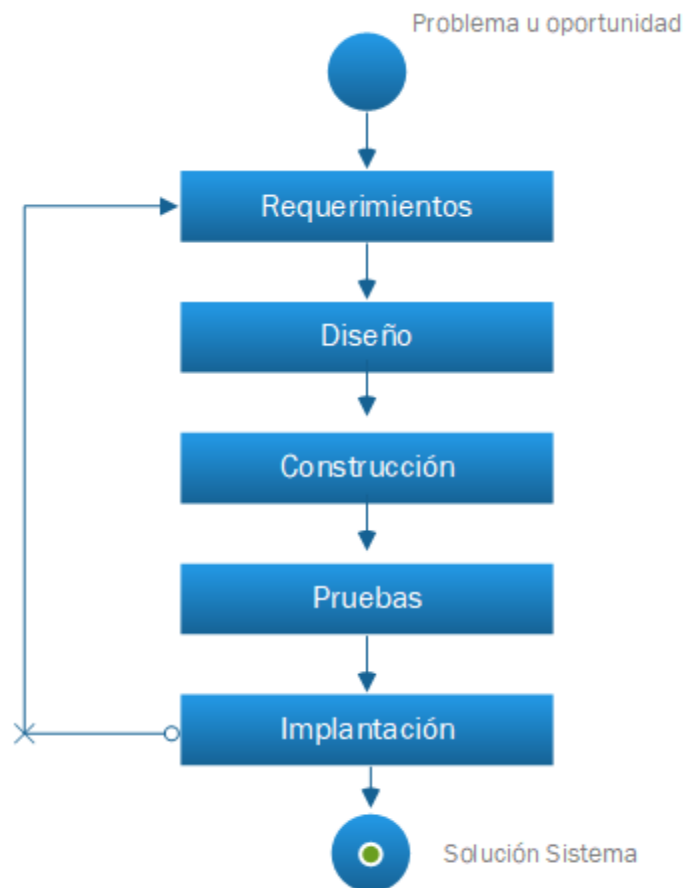


Fig. 4 Etapas técnicas de desarrollo
Fuente: (Cervantes Maceda et al. 2016)

1.4. Proceso de Descubrimiento KDD.

La minería de datos es una parte fundamental para el proceso de descubrimiento del conocimiento KDD (Knowledge Discovery in Databases) este término se utiliza para referirse al proceso de extracción automatizada del conocimiento a partir de grandes volúmenes de Datos. Este proceso integra varias etapas hasta llegar a obtener el conocimiento para la toma de decisiones.

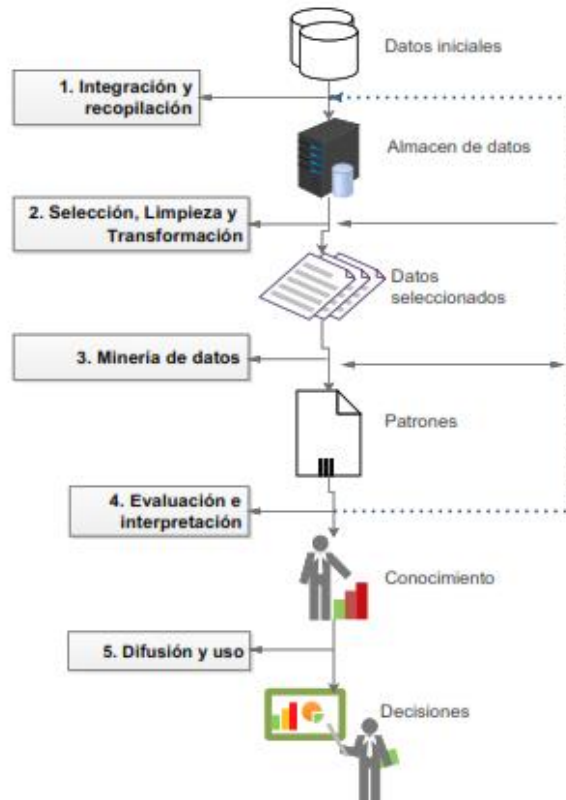


Fig. 5 Fases del Proceso KDD
Fuente:(Ordoñez and Grass 2015)

El proceso KDD empieza con la recopilación e integración de la información a partir de datos iniciales de los diferentes orígenes de datos (data warehouse), estos pueden ser de bases de datos relacionales, temporales, multimedia, texto, etc., estos almacenes de datos pueden ser de origen interno o externo esta etapa es mejor conocida como la etapa de selección de datos y es el pilar para que el proceso completo sea exitoso (Pérez & Santín, 2007). La siguiente fase es el procesamiento de datos; no se puede realizar directamente una data mining sobre los datos recopilados en el almacén de datos, debido a que los datos no pueden estar limpios estos pueden contener atributos irrelevantes,

dentro de esta fase se realiza una selección de los datos integrados en la dataWarehouse. El resultado de esta fase es denominada **vista minable**, esto es un subconjunto limpio y transformado de los datos sobre el que ya se puede aplicar las técnicas de data mining(Lara 2014).

La etapa más importante del proceso KDD es la minería de datos, ya que, aplicando las diferentes tareas de minería de datos, predicción y clasificación, en este caso, se obtiene varios modelos que representarán a los datos analizados, para llegar a la etapa final que consiste en evaluar e interpretar la información obtenida para obtener el conocimiento, dentro de esta etapa final los modelos obtenidos en la fase anterior de data mining pueden ser evaluados. Una vez comprobada la calidad de los mismos, estos son interpretados, y a partir de ellos, se obtiene el conocimiento necesario(Lara 2014).

Después de haber aplicado el proceso KDD y se obtiene el conocimiento, el mismo que se aplica para resolver el problema inicial propuesto de la empresa y por ello se miden los resultados luego de haber aplicado dicho conocimiento. Si tenemos el caso que los resultados no sean satisfactorios, el proceso KDD se plantea y se vuelve a aplicar realizando los cambios necesarios en las diferentes fases, en busca de conocimiento que permita resolver el problema inicial, tal como se muestra en la Fig. 6, se puede apreciar la naturaleza cíclica de la extracción de conocimiento a partir de grandes volúmenes de datos(Lara 2014).

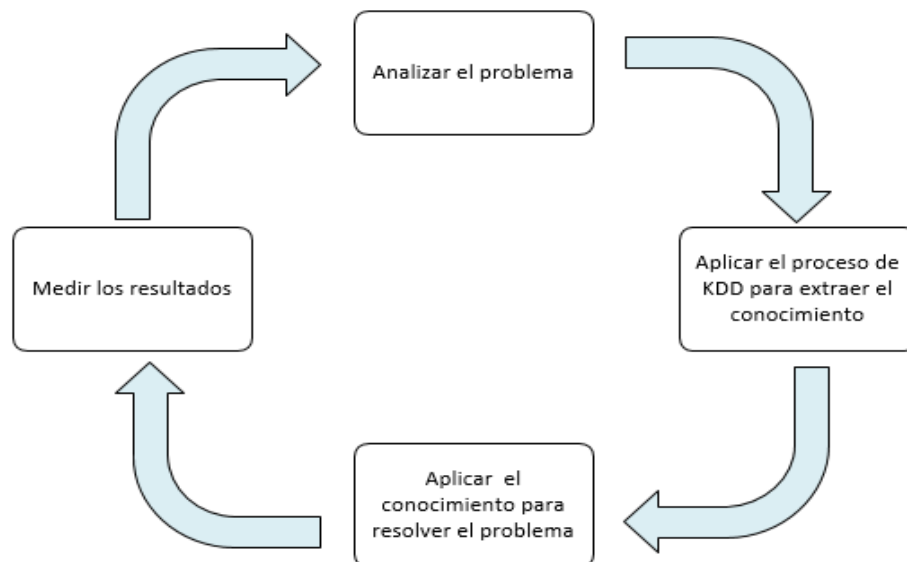


Fig. 6 Naturaleza cíclica del proceso de descubrimiento
Fuente: (Lara 2014)

1.5. Etapas del proceso KDD

1.5.1. Fase de integración y recopilación

En esta fase lo primero que se debe realizar es tomar los datos que se desea analizar e integrarlos en un repositorio del cual partirán las siguientes fases. Al realizar minería de datos es posible que la institución es decir el propietario de los datos disponga múltiples fuentes de datos por separado. Existe la posibilidad que, para el análisis se utilicen fuentes de datos públicas como, por ejemplo, datos de censo de población de una determinada región(Lara 2014).

Para poder realizar cualquier operación con los datos de diferentes fuentes, es fundamental aglutinarlos en un mismo almacén de datos o data warehouse(Lara 2014). Antes de integrar las fuentes de datos en un mismo almacén, es necesario realizar un proceso que lea los datos de diferentes fuentes, los limpie y adecue a la estructura que tiene el data warehouse para su almacenamiento, este proceso se lleva a cabo mediante un sistema conocido como sistema **ETL (Extraction - Transformation-Load)**(Lara 2014).

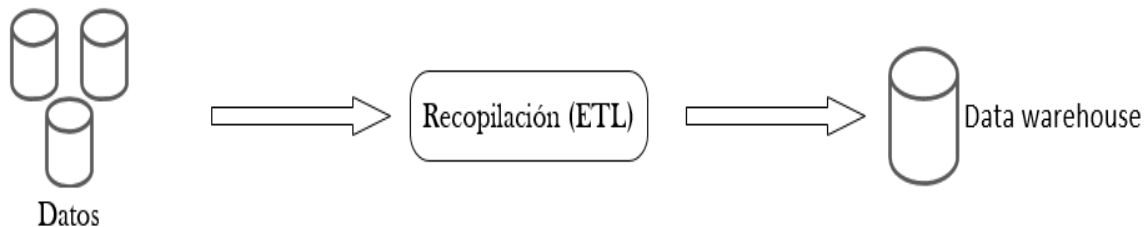


Fig. 7 Extraction-Transformation-Load
Fuente: (Lara 2014)

El análisis posterior de la información se tornará más sencillo, si la fuente de datos a ser analizados es unificada, accesible y sobre todo que se encuentre separada del trabajo transaccional (Pérez & Santín, 2007). La idea de integrar múltiples bases de datos, con sus respectivos formatos, identificadores, etc., es un reto significativo que ha dado lugar a los conocidos almacenes de datos o data warehouse, que según (PowerData, 2018) son los que permiten crear un repositorio de bases de datos transaccionales provenientes de diferentes fuentes para la toma de decisiones. Para crear un almacén de datos se pueden aplicar un sinnúmero de técnicas, una de las más comunes es integrar y almacenar la información en un nuevo esquema.

Los almacenes de datos se emplean para integrar información de manera sofisticada, por esta razón los datos se modelan con una estructura de bases de datos multidimensional, en el cual cada dimensión corresponde a un atributo o grupo de atributos en el esquema en torno a hechos que almacenan el valor de acuerdo con una medida agregada, por ejemplo, la cantidad de estudiantes que aprobaron una materia en un año en concreto de una carrera.

De esta forma se da paso a que los almacenes de datos sean adecuados para el procesamiento analítico en línea (on-line analytical processing, OLAP), que es un análisis de datos superior al clásico SQL (Structured Query Language), ya que permite presentar la información a diferentes niveles de abstracción, dependiendo de las necesidades del usuario (Hernández et al., 2004).

Para el presente trabajo de titulación, se recopilará los archivos .log que se genera en la base de datos transaccionales en Oracle (Oracle, 2018) los cuales se generan automáticamente a partir del acceso y realización de peticiones al portal web de la Universidad Técnica del Norte.

1.5.2. Fase de selección, limpieza y transformación

Para asegurar la calidad del conocimiento que se pretende obtener, no solo se deben aplicar las técnicas de minería de datos adecuadas, sino también que los datos a minar deben ser de calidad; por este motivo después de la fase de integración y recopilación de la información es indispensable seleccionar, limpiar y transformar la información, de esta forma se obtiene una vista minable de los datos más importantes y relevantes (Hasperue 2013).

Para realizar la selección de datos es necesario realizar el proceso de filtrado que se puede ejecutar a varios niveles (Lara 2014).

- **Filtrado de atributos**

La selección de los atributos es uno de los pasos más relevantes del proceso KDD, ya que los atributos que se van a considerar deben ser relevantes para el estudio (Hernández Orallo et al., 2004), por ejemplo, para identificar los desertores estudiantiles de

una universidad uno de los principales datos irrelevantes es el nombre del estudiante y su pasaporte o cédula de ciudadanía, ya que no aportan al estudio.

- **Filtrado de registros**

La selección de registros en muchas ocasiones depende de la naturaleza del problema que se pretende solucionar, por ejemplo, es posible que en el caso anterior se desee tomar en cuenta únicamente a los estudiantes que se encuentran cursando una carrera afín a la informática y por ende se eliminarían el resto de los registros.

b. Limpieza

La limpieza de datos consiste en eliminar el mayor número posible de datos erróneos o inconsistentes y ausencia de valores. En esta etapa del proceso, generalmente, se emplean herramientas de consulta de información y herramientas estadísticas (Pérez López & Santín González, 2007).

En el caso de que exista ausencia de valores, por ejemplo, puede ser que en el data warehouse exista un campo que haga referencia al lugar de nacimiento de la persona y dicho campo en un registro se encuentre vacío, es indispensable que caso de faltantes se realice un análisis minucioso, ya que puede ser que arrojen información interesante, tal como cuando una persona no quiere dar a conocer su información o se reserva el derecho a divulgarla, o dado el caso se puede pasar por alto la ausencia y continuar con el análisis o filtrar el registro (Lara, 2014).

Cuando se habla de datos inconsistentes o erróneos, se dice que pueden representar ruidos o excepciones, sin embargo, otros son muy relevantes y el resultado se puede alterar por ello, por ejemplo, se puede dar el caso en el cual el estudiante tenga como año de nacimiento 2203, lo cual sería un dato erróneo. En muchos casos no es conveniente eliminarlos, ya que en ciertos casos como detecciones de fraudes pueden ser más interesantes que los datos regulares (Hernández et al., 2004).

c. Transformación

La transformación de datos es la etapa en la cual se preparan los datos para facilitar el uso de las diferentes técnicas de minería de datos que requieren los diferentes datos;

existen varias técnicas de transformación de datos, entre las principales se encuentran (Lara 2014):

- **Numerización**

La numerización consiste en transformar un atributo de tipo cualitativo a cuantitativo (Hasperue 2013) por ejemplo, al momento de almacenar si un estudiante aprobó o no, se puede colocar para el valor verdadero 1 o 0 para falso respectivamente(Lara 2014).

- **Discretización**

Es el proceso inverso a la numerización, en el cual los valores numéricos son transformados en discretos o nominales(Lara 2014). Un ejemplo claro es el peso de una persona que puede pasar de > de 18.5 a deficiente, de 18.5 a 24.9 a normal, de 25.0 a 29.9 a sobrepeso y de <30 a obesidad («Calculadora del índice de masa corporal (IMC)», s. f.).

- **Creación de características**

Consiste en crear nuevos atributos en función a atributos existentes, que son las variaciones de estos (Hernández et al., 2004). En el caso del estudiante se puede crear el promedio general sumando el promedio de cada semestre cursado y dividiéndolo para el número de semestres.

- **Normalización**

La normalización consiste en transformar el rango de valores que toma un determinado atributo. Generalmente se emplea la normalización lineal uniforme, que transforma los valores de un atributo a una escala entre 0 y 1 mediante la Ecuación 1(Lara 2014)a:

Ecuación 1

$$\text{ValorNormalizado} = \frac{\text{ValorInicial} - \text{ValorMínimo}}{\text{ValorMáximo} - \text{ValorMínimo}}$$

Para seleccionar, limpiar y transformar la información se emplearán las herramientas estadísticas que nos brinda la herramienta ofimática de Microsoft Excel (Microsoft, 2018), la herramienta RStudio.

1.5.3. Fase de minería de datos

Una vez obtenida y consolidada la vista minable, el siguiente paso es aplicar las diferentes técnicas de minería de datos para obtener los modelos que representan dichos datos. Para poder iniciar con la fase de minería de datos es importante tomar decisiones que afectarán a la calidad del conocimiento que se pretende obtener (Lara 2014) entre ellas:

- Determinar la tarea de minería de datos que es la más apropiada para el análisis, por ejemplo, si se desea predecir cierta información.
- Escoger el modelo de acuerdo con la forma que se pretende que la información sea presentada, por ejemplo, en reglas de decisión.
- Elegir el algoritmo más eficiente al momento de resolver la tarea y que devuelva el modelo que se está buscando.

1.5.3.1. Tareas de minería de datos

En la etapa de minería de datos, se aplican diferentes técnicas para resolver los diferentes problemas que se presenten. A los diferentes tipos de problemas que se pueden resolver por medio de las técnicas de minería de datos se conocen como tareas. Las tareas de minería de datos pueden ser predictivas o descriptivas (Lara 2014).

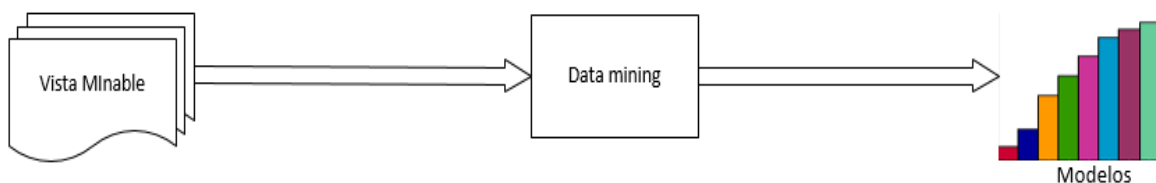


Fig. 8 Tareas de minería de datos
Fuente:(Lara 2014)

a. Tareas predictivas

Según (Hasperue 2013), las tareas predictivas se tratan principalmente de problemas y tareas en los que hay que predecir uno o más valores desconocidos, de uno o varios atributos para varios registros de la vista minable, entre las principales tareas predictivas se encuentran la clasificación y la regresión (Lara 2014).

- **Clasificación**

La tarea de clasificación es la más utilizada, consiste en utilizar un conjunto de entrenamiento para construir un modelo que a futuro se empleará para clasificar elementos

o individuos desconocidos en base a una variable (atributo) de clase de tipo cualitativo(Lara 2014).

- **Regresión**

La regresión es similar a la clasificación, sin embargo, en este caso el atributo a predecir no es cualitativo, sino más bien cuantitativo que es la variable de clase.(Lara 2014).

b. Tareas descriptivas

Las tareas descriptivas generan modelos que de una forma u otra describen un grupo de datos.

- **Agrupamiento**

El objetivo principal de esta tarea es obtener grupos homogéneos a partir de los datos heterogéneos, ya que en este caso se habla de grupos y no de clases, puesto que no analiza los datos a partir de una etiqueta conocida, sino que analiza los datos para obtener dicha etiqueta(Hernandez Orallo, Ramirez Quintana, and Ferri Ramirez 2004).

- **Asociación**

El objetivo de la tarea de asociación es buscar las relaciones que no se encuentran explícitas entre los atributos, por medio de reglas de asociación(Lara 2014).

- **Detección de atípicos**

La detección de atípicos consiste en encontrar objetos que presentan un comportamiento significativamente diferente al resto de los objetos del conjunto de registros dentro de una vista minable (Lara 2014).

1.5.4. Fase de evaluación e interpretación

Finalmente, después de obtener los modelos de minería de datos el siguiente paso es evaluar la calidad de los modelos e interpretarlos, con la finalidad de obtener el conocimiento deseado(Lara 2014).

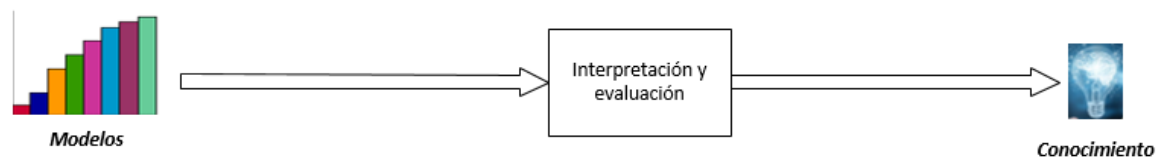


Fig. 9 Fase de evaluación e interpretación
Fuente: (Lara 2014)

a. Evaluación

Validar la bondad de un modelo predictivo es lo que se conoce como evaluar y sirve principalmente para medir su capacidad de predicción de nuevas instancias. Habitualmente la validación se realiza en base a las siguientes consideraciones (Sierra 2006):

- **Matriz de confusión:**

La matriz de confusión es una herramienta que contiene información acerca de valores reales y las clasificaciones y es fundamental para evaluar el desempeño de un algoritmo de clasificación.

Tabla 1 Matriz de confusión

		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Fuete: (Sierra 2006)

a. Verdaderos Positivos: Es el número de predicciones correctas de un caso sea positivo.

b. Falsos Negativos: es el número de predicciones incorrectas de que un caso es negativo, o sea la predicción es negativa cuando realmente el valor tendría que ser positivo. A estos casos también se les denomina errores de tipo II.

c. Falsos Positivos: es el número de predicciones incorrectas de que un caso es positivo, es decir, la predicción es positiva cuando realmente el valor tendría que ser negativo. A estos casos también se les denomina errores de tipo I.

d. Verdaderos Negativos: es el número de predicciones correctas de que un caso es negativo.

La validación proviene de la generalización, de dividir la muestra total en K grupos de aproximadamente el mismo tamaño, en el cual K -1 constituye el grupo de entrenamiento y el resto el grupo de testeo (Lara 2014), un caso particular de validación cruzada es el método de 'dejar uno fuera', en donde K es igual al número de instancias que son

empleadas para entrenar el modelo, ya que el proceso se repite varias veces para obtener el promedio de error cometido(Sierra 2006).

Cuando ya se ha evaluado el modelo, es indispensable expresar el conocimiento en términos que conoce el usuario final; por ello es importante que se relacione la minería de datos con técnicas de visualización, para que dichos modelos sean comprendidos e interpretados por los expertos de cada área; así los expertos podrán comparar el conocimiento obtenido con la realidad que ellos perciben para emplearlo en la toma de decisiones estratégicas(Lara 2014) ,

1.6. Tareas y modelos predictivos

Las tareas predictivas de minería de datos aportan diferentes modelos predictivos, que tienen el objetivo de obtener un modelo válido para tratar futuros casos (Sierra 2006), los cuales se dividen en clasificación y regresión.

1.6.1. Clasificación

Entre las principales técnicas de clasificación se encuentran:

- **Algoritmos de clasificación por vecindad**

Los algoritmos de vecindad exigen una definición de una cierta medida de distancia entre los elementos del espacio en representación. Una de las principales ventajas de esta técnica, es la simplicidad de su concepto, ya que la clasificación de un nuevo punto del espacio de representación se calcula en función a las clases, de los puntos más próximos a él (Sierra 2006) Por ejemplo, el algoritmo K-NN.

- **Árboles de clasificación**

Los árboles de decisión son un conjunto de condiciones que tienen una estructura jerárquica, de tal manera que la predicción se puede realizar siguiendo el camino de las condiciones hasta una de sus hojas. Una de las principales ventajas de los árboles de decisión es que se pueden expresar en grafo o en reglas de decisiones (Hernandez Orallo et al. 2004). Por ejemplo, el algoritmo Random Tree, Random Forest, entre otros. El algoritmo Random Tree considera un número dado de características aleatorias en cada nodo sin realizar ninguna poda, mientras que Random Forest construye bosques aleatorios mediante el empaquetamiento de conjuntos de Random Tree(Frank, Hall, and Witten 2017).

- **Redes bayesianas**

Las redes bayesianas modelan un hecho mediante un conjunto de variables y relaciones entre ellas, basadas en el teorema de Bayes; esta técnica predictiva permite estimar la probabilidad futura de las variables desconocidas en base a las conocidas(Sierra 2006).

Thomas Bayes expresa que, el teorema de Bayes es un resultado que expresa la probabilidad condicionada de un evento aleatorio dado otro evento(Lara 2014).

Por ejemplo, el algoritmo Naive Bayes, que implementa la teoría bayesiana para generar las probabilidades, también puede usar estimaciones de densidad del kernel, que mejora el rendimiento del algoritmo en caso de que el supuesto de normalidad sea muy incorrecto(Frank et al. 2017).

- **Redes neuronales artificiales**

Las redes neuronales se basan en el aprendizaje humano, es decir en las neuronas cerebrales. Gracias a ellas un valor de entrada se transforma en salida mediante una función no lineal. Las redes neuronales poseen las siguientes características(Lara 2014):

- La exactitud usualmente es muy alta
- Trabajan de manera correcta incluso con datos erróneos
- La salida puede ser un valor real o un conjunto de reales
- Evolucionan rápidamente de la función de entrenamiento

Por ejemplo, el algoritmo perceptrón multicapa.

Máquinas de vectores de soporte (SVM)

El desarrollo del aprendizaje supervisado ha dado paso a la creación de algoritmos denominados métodos de Kernel, que han tomado un gran éxito, por su método denominado Máquinas de Vectores Soporte (support vector machines). Entre sus principales ventajas se encuentra la aplicabilidad a cualquier tipo de datos, ya que las funciones Kernel sirven como mecanismo de transformación y representación de información de entrada al algoritmo.

1.6.2. Regresión

Entre las principales técnicas de regresión se tienen las siguientes:

- **Regresión lineal**

La regresión lineal es la forma más sencilla de regresión, puesto que los datos se modelan usando una línea recta; mediante una variable aleatoria que se denomina variable respuesta, que es la función lineal de otras variables, a_i ($0 \leq i \leq k$), denominadas variables predictoras, como se aprecia en la Ecuación (Lara 2014):

Ecuación 2

$$y = w_0 a_0 + w_1 a_1 + \dots + w_k a_k$$

El principal objetivo de la regresión es obtener el valor de una serie de pesos, a partir de los datos de un conjunto de entrenamiento. Los pesos se calculan mediante la técnica de los mínimos cuadrados, con el objetivo de minimizar la expresión de la Ecuación 10, donde y_i representa a la variable de respuesta y para el objeto i (Lara 2014).

Ecuación 3

$$\sum (y_i - \sum w_j a_{ji})^2, (1 \leq i \leq n), (0 \leq j \leq K)$$

- **Regresión logística**

La regresión logística se denomina así puesto que es un modelo más generalizado de regresión, denominada también discriminación logística; este tipo de tarea predictiva obtiene una estimación de probabilidades para variables categóricas, es decir una variable que puede adoptar un número limitado de categorías, y cuando se habla de dos variables de clase se habla de ranking (Hernandez Orallo et al. 2004).

La regresión logística es un tipo de análisis discriminante predictivo, ya que su objetivo principal es brindar procedimientos sistemáticos para clasificar una observación cuyo origen se desconoce empleando los valores que toman las variables clasificadoras, que generalmente son variables de tipo cuantitativo. El objetivo de esta técnica es estimar probabilidades a posteriori $\{P(G_i|X); i = 1, \dots, M\}$, en este estudio el número de grupos a discriminar es $M = 2$, para la clase desértico que toma los valores Si o No. Constituyen modelos de la forma que se presenta en la Ecuación 11 (Sierra 2006):

Ecuación 4

$$P(G_1|X) = F(X'\beta); P(G_2|X) = 1 - P(G_1|X)$$

Donde F es la función de distribución de probabilidad acumulada, denominada función de enlace, particularmente si $F(x) = \varphi x$, la función es estándar y por ende el modelo es de Regresión Binomial tal como se aprecia en la Ecuación 12:

Ecuación 5

$$f(x) = \frac{\exp(x)}{1 + \exp(x)}$$

En el presente trabajo se va a ejecutar el análisis de los portales web, esencialmente se va a realizar el estudio y análisis de los archivos .log, para ello utilizaremos la categoría de minería de uso Web (Web Usage Mining).

1.7. Norma ISO/IEC/IEEE4210:2011- Sistemas de ingeniería de software - Descripción Arquitectura

La ISO / IEC / IEEE 42010: 2011 define los requisitos en la descripción del sistema, de software y de la empresa arquitecturas. Se persigue el objetivo de estandarizar la práctica de la descripción de la arquitectura mediante la definición de los términos estándar, presentando una base conceptual para la expresión, la comunicación y la revisión de arquitecturas y especificar los requisitos que se aplican a las descripciones de la arquitectura, los marcos de arquitectura y descripción de la arquitectura idiomas(Hilliard 2011) .

Esta norma internacional especifica también disposiciones que hacen cumplir las propiedades deseadas de los marcos de arquitectura y lenguajes de descripción de la arquitectura (AVD), con el fin de aportar una contribución útil al desarrollo y uso de las descripciones de la arquitectura. Esta norma internacional proporciona una base sobre la cual comparar e integrar los marcos de arquitectura y AVD, proporcionando una ontología común para especificar su contenido. Esta norma internacional se puede utilizar para establecer una práctica coherente para desarrollar las descripciones de la arquitectura, marcos de arquitectura y descripción de la arquitectura lenguas dentro del contexto de un ciclo de vida y sus procesos (no definido por esta norma internacional). Esta Norma Internacional puede ser utilizado además para evaluar la conformidad de una descripción

de la arquitectura, de un marco de arquitectura, de un lenguaje de descripción de la arquitectura o de una arquitectura punto de vista de sus disposiciones(Hilliard 2011).

1.7.1. Historia y evolución de la norma ISO/IEC/IEEE4210:2011

La norma ISO/IEC/IEEE 42010:2011, ha tenido varios antecesores y cambios a lo largo del tiempo como se muestra en la Fig. 10

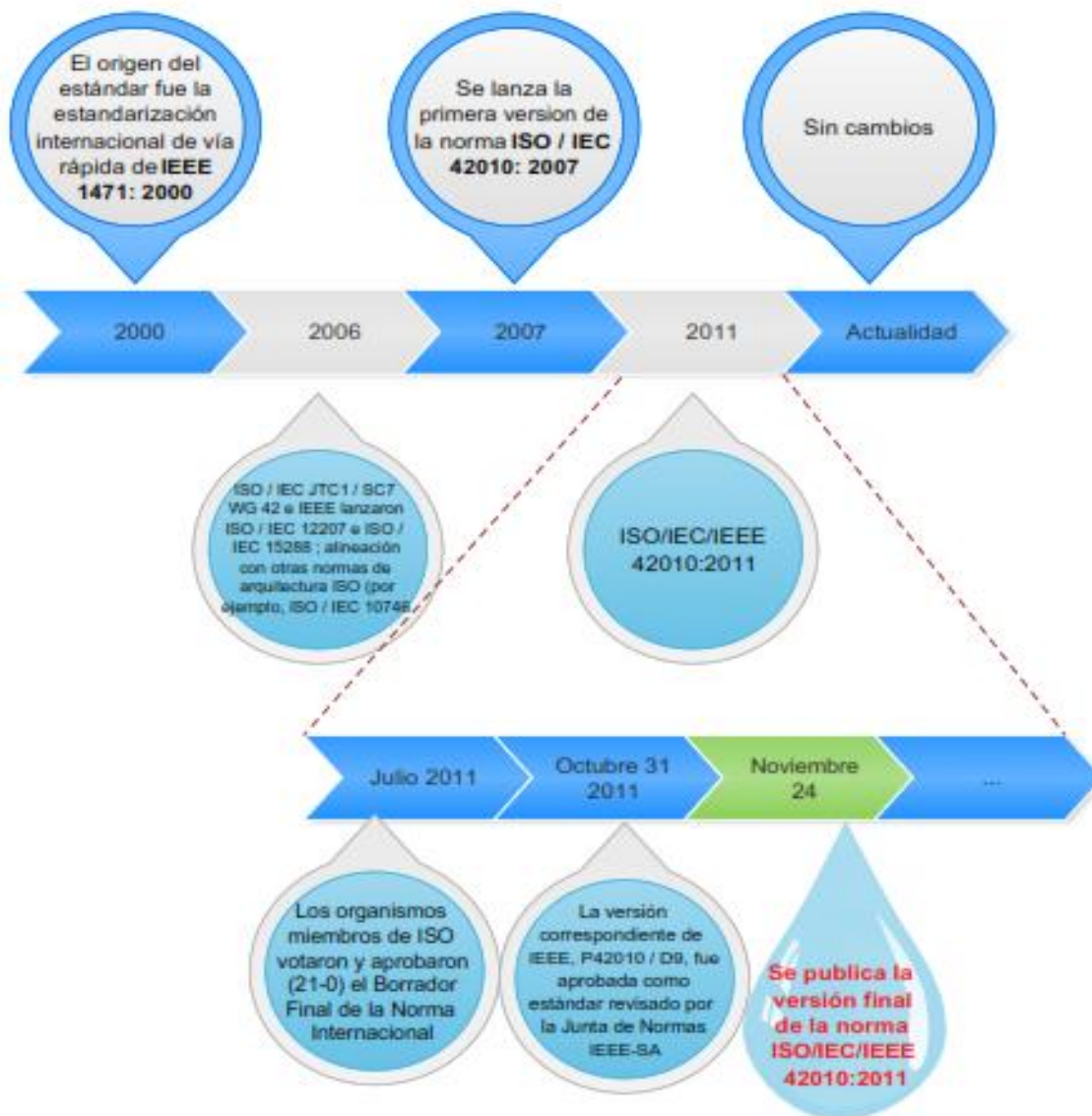


Fig. 10. Línea de tiempo ISO/IEC/IEEE 42010:2011
Fuente: Propia adaptada a (Hilliard 2011)

1.7.2. Términos y definiciones de Norma ISO/IEC/IEEE4210:2011.

Tabla 2 Términos y definiciones norma ISO/IECE/IEEE42010

Término	Definición
Architecting	Es el proceso de concebir, definir, comunicar, certificar la implementación adecuada y mejorar una arquitectura a través de su ciclo de vida.
System Architecture	Conceptos fundamentales o propiedades de un sistema corporizado en sus elementos, relaciones y los principios de su diseño y evolución.
Architecture description (AD)	Producto usado para expresar una arquitectura
Architecture framework	Convenciones, principios y prácticas para la descripción de arquitecturas que están establecidas en un dominio específico.
Architecture view	Producto que expresa la arquitectura de un sistema desde la perspectiva de un <i>concern</i> específico del mismo
Architecture viewpoint	Producto que establece las convenciones para la construcción, interpretación y uso de las vistas arquitectónicas para enmarcar un <i>concern</i> específico
Concern	Interés en un sistema que es relevante a uno o más de los participantes
Environment	contexto que determina las circunstancias de todas las influencias sobre un sistema

Fuete: Propia adaptada a (Anon 2011).

Los términos anteriormente descritos hacen referencia a conceptos relacionados para la aplicación de esta norma, que es útil al realizar una descripción de la arquitectura que se expresa para un sistema de interés. En esta Norma Internacional, el término “Sistema de interés”, hace referencia al sistema cuya arquitectura está bajo consideración en la preparación de una descripción de la arquitectura.

1.7.3. Organigrama ISO/IEC/IEEE 42010

La norma está estructurada por los siguientes apartados y anexos, como se muestra en la Fig. 10:

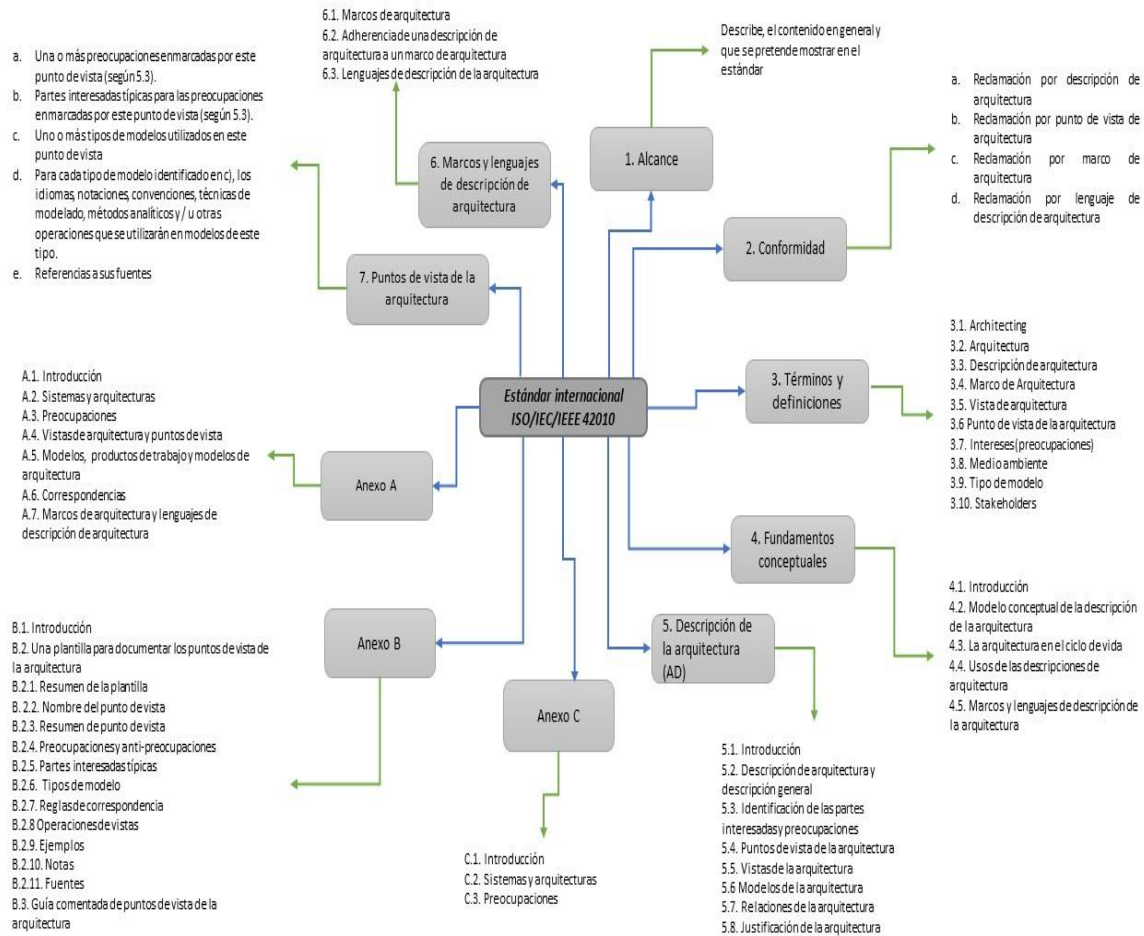


Fig. 11 Organizador gráfico ISO/IEC/IEEE 42010
Fuente: Elaboración propia

1.7.4. Estructura General

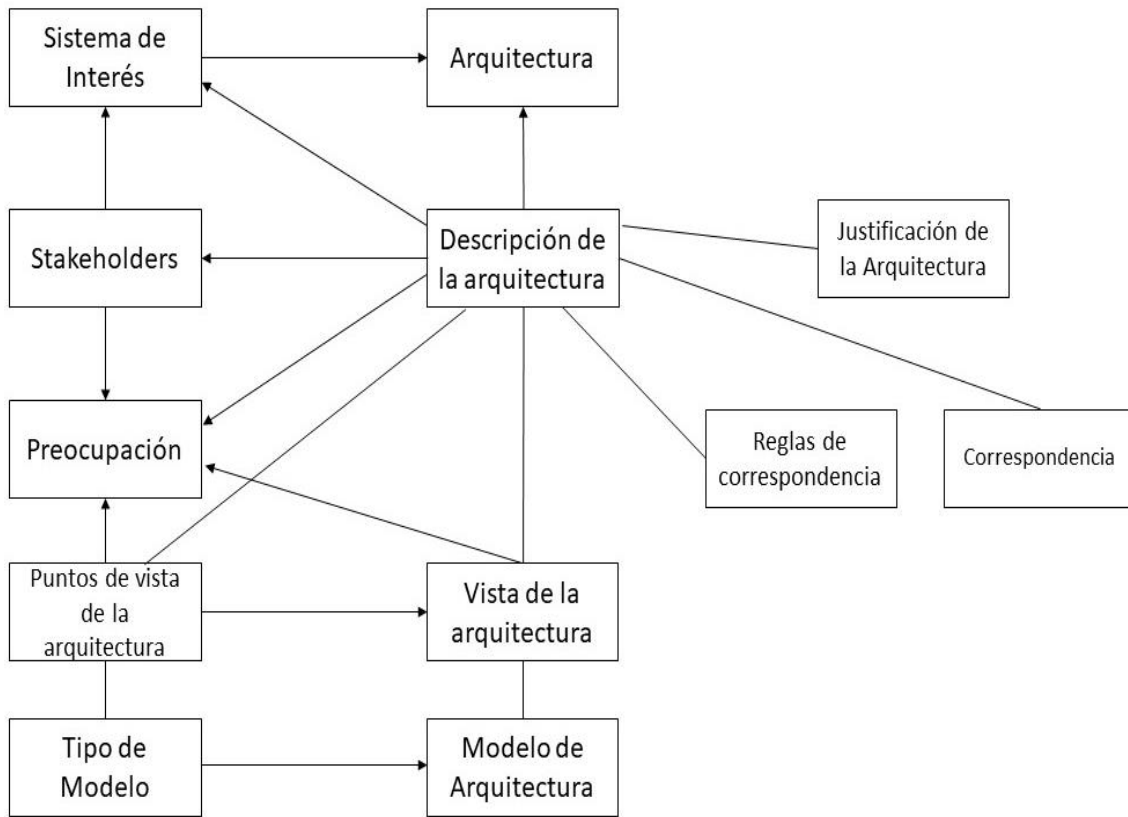


Fig. 12 Descripción de la Arquitectura
Fuente:(Anon 2011)

CAPÍTULO II

2. Desarrollo

La presentación del presente proyecto de minería de datos web, proporcionara una información clara del análisis de los archivos .log provenientes del portal de la Universidad Técnica del Norte, además del levantamiento de una arquitectura tecnológica que facilitara el entendimiento del comportamiento, y los patrones que se puede obtener, se realizara mediante la aplicación de la Norma ISO/IECE/IEEE 42010, para el análisis de los datos obtenidos se utilizara el Proceso de analítica web, conjuntamente con el proceso KDD.

A continuación, se procederá a explicar cada una de las fases desarrolladas para la obtención de patrones y elaboración de la arquitectura tecnológica.

2.1. Portal web institucional

2.1.1. Definición

Son sitios web avanzados que se crean utilizando plataformas denominadas Sistemas de Administración de Contenidos (ó CMS por sus siglas en inglés). Estas plataformas de software permiten manejar múltiples tipos de contenidos clasificados de acuerdo a sus funcionalidades y que se almacenan en bases de datos, lo que permite que crezcan de acuerdo a lo que se necesite(Liferay n.d.).

2.1.2. Funciones

Los portales Web institucional están calificados como de gran importancia ya que ofrece a los usuarios un sitio dónde encontrar gran cantidad de información, servicios, actividades comerciales y recursos relacionados con la organización, aprovechando al máximo la información que brinda el Internet.

Los portales institucionales sirven como una vía de acceso a otra web que estén directamente relacionadas con la entidad, como: departamentos, oficinas, servicios y otras reparticiones además permite:

- Proporcionar información relevante para la comunidad universitaria
- Suministrar recursos didácticos de todo tipo, que aportan a la formación profesional.
- Centralizar los contenidos científicos y de valor para la comunidad universitaria
- Organizar las fuentes informativas
- Orientar la navegación de los usuarios.

2.1.3. Descripción del portal Web UTN

El portal Web UTN provee información de eventos, anuncios, ofertas académicas, además de acceso a varios recursos académicos, para mantener una buena gestión y control del portal Web posee una distribución de blogs o micrositos asignados a cada una de las áreas que requieran desplegar información de interés universitario (Arcos 2017).

El portal Web UTN, así como los blogs se encuentran desarrollados bajo la plataforma WordPress 4.7, desplegados bajo un servidor Web Apache 2.2 y MySQL 5.6.13 como gestor de base de datos. Cada vez que un usuario realice una petición o el portal detecta un fallo en el mismo, esto se almacena en el servidor, en la carpeta /var/log/httpd, estos pueden ser error_log, acces_log y sys_log, de los cuales se toma en cuenta error_log y acces_log.

Mediante un diagrama de proceso se representa el modelo de gestión Web de un servidor apache como se muestra en la Fig. 13

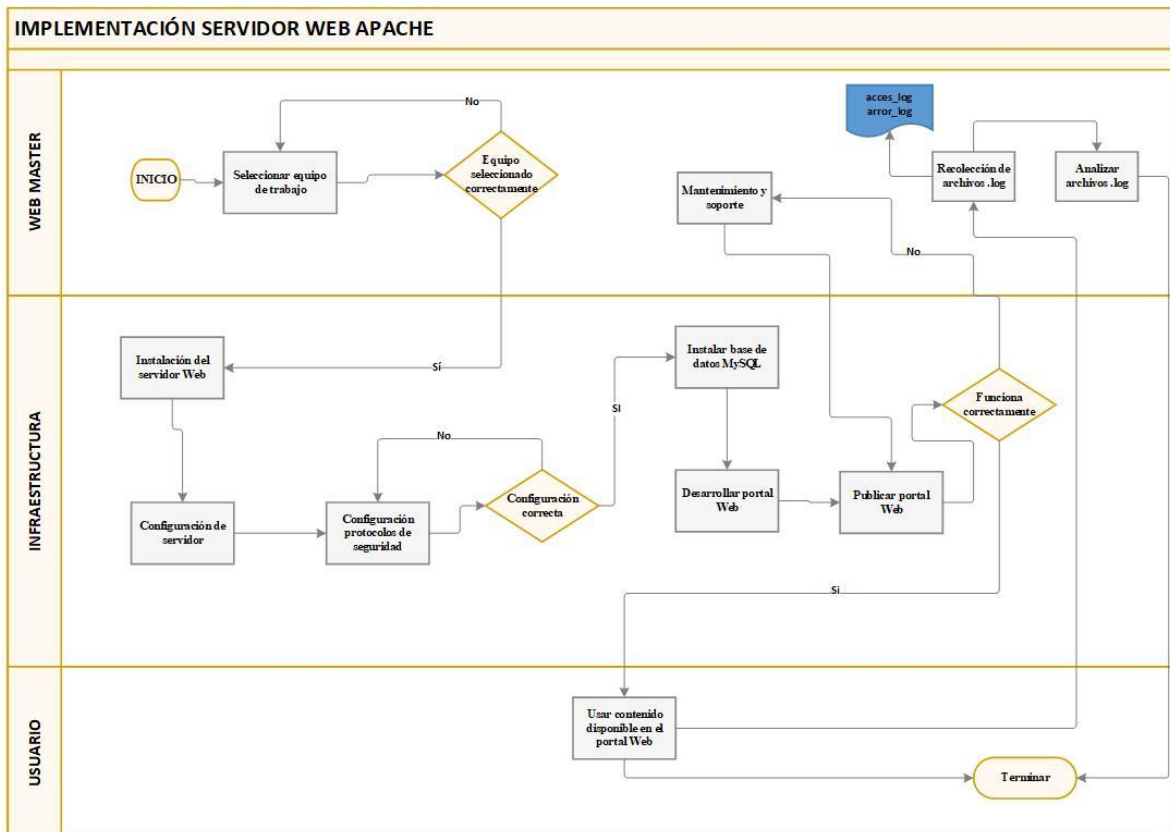
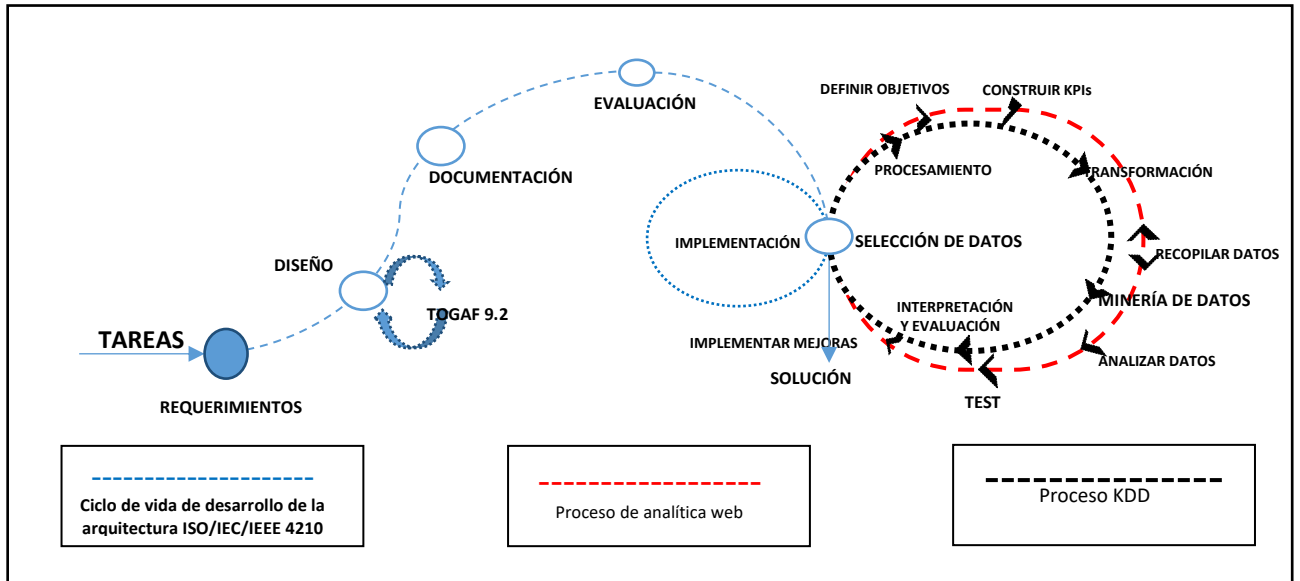


Fig. 13 Implementación servidor Web.
Fuente: Elaboración propia

2.2. Gestión del proyecto utilizando ISO/IEC/IEEE 42010

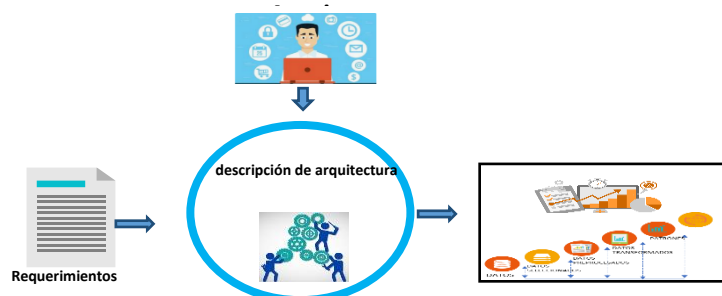
En la Fig. 14 se muestra la estructura del proyecto



*Fig. 14 Estructura Capítulo 2
Fuente: Propia*

2.3. Fase de diseño de la arquitectura

Según, (Cervantes Maceda et al. 2016), la fase de diseño de arquitectura es la transformación de los requisitos, que buscan alcanzar los objetivos para satisfacer una serie de requerimientos.



*Fig. 15 Diseño de arquitectura
Fuente: Propia*

Para la elaboración del diseño de la solución y descripción de la arquitectura se toma en cuenta el marco de trabajo TOGAF en su versión 9.2.

2.4. Open Group Architecture Framework (TOGAF)

Se define como una arquitectura empresarial que ofrece un marco de alto nivel para el desarrollo de software empresarial. Ayuda a organizar el proceso de desarrollo a través de un enfoque sistemático para reducir los errores, mantener los plazos, mantenerse dentro del presupuesto y alinear la TI con las unidades de negocios para producir resultados de calidad(The Open Group 2018).

2.4.1. Fases de TOGAF

En la Fig. 16, se muestra las fases de descripción de arquitectura propuesto en el marco de trabajo TOGAF 9.2.

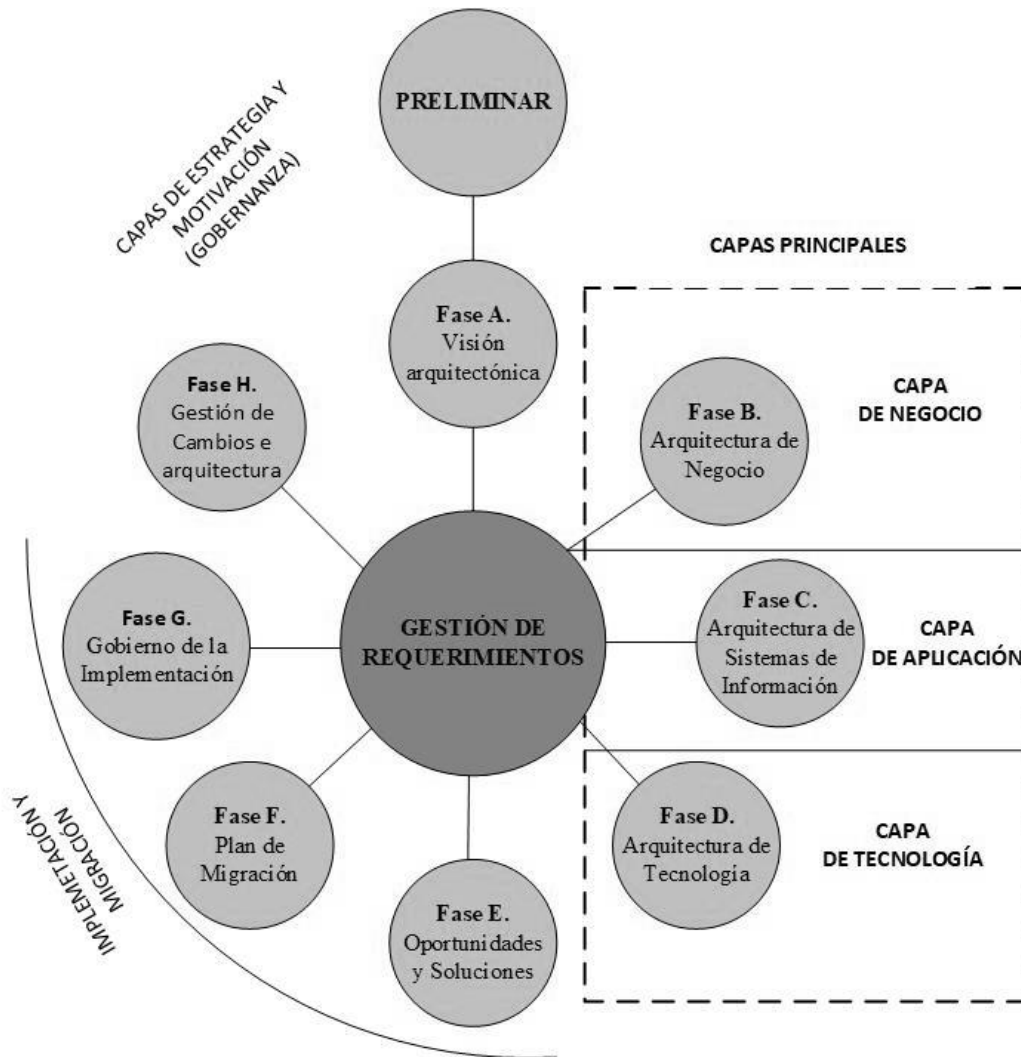


Fig. 16. Fases Descripción de Arquitectura

Fuente: Propia adaptada a (The Open Group 2018)

2.5. Fase preliminar

2.5.1. Principios de Arquitectura

Esta sección detalla los principios de la arquitectura a la que el DDTI se adhiere. Su propósito es definir los principios de la arquitectura para los cuatro dominios de la arquitectura, enunciados por TOGAF tales como: negocio, aplicaciones, datos y tecnología; con el objetivo de informar y dar soporte a la forma en que la organización se ajusta sobre el cumplimiento de su misión.

2.5.1.1. Resumen de Principios de arquitectura

En la Tabla 5 se muestra el resumen de los principios de arquitectura según TOGAF.

Tabla 3. Resumen de Principios de Arquitectura

DOMINIO	PRINCIPIO
Negocio	Alineación entre TI y el negocio
	Enfoque en el cliente
	Enfoque a largo plazo
Datos	Información relevante
	Información accesible
	Seguridad de la información
	Copia de seguridad de datos
Aplicaciones	Seguimiento de estándares
	Independencia de la tecnología
	Aplicaciones fáciles de usar
	Reutilización y simplicidad
Tecnología	Tecnología madura
	Infraestructura escalable
	Reevaluar la seguridad
	Seguimiento

Fuente: Propia

2.5.1.2. Principios de Negocio

En la **tabla 4**, se describe la alineación o la relación que tiene el TI y el negocio.

Tabla 4. Alineación entre TI y negocio

Nombre	Alineación entre TI y el negocio
Referencia	PDN01
Enunciado	Los proyectos de TI deben estar relacionados con el objeto de negocio y las estrategias.
Fundamento	Este principio permite que la tecnología que se vaya a implementar se adapte a las necesidades de la institución, asegurando así que los cambios sirvan de soporte a las operaciones del negocio.
Repercusiones	Los cambios o mejoras se aceptarán, cuando estos estén justificados por las necesidades del negocio.

Fuente: Propia

En la **Tabla 5**, se muestra el enfoque del proyecto hacia el cliente, a satisfacer sus necesidades.

Tabla 5. Enfoque al cliente

Nombre	Enfoque en el cliente
Referencia	PDN02
Enunciado	Las decisiones arquitectónicas deben buscar la satisfacción del cliente
Fundamento	Este principio asegura que los productos y servicios estén orientados a la satisfacción del cliente
Repercusiones	Se revisarán las peticiones y funcionamiento de manera que este enfocado a lo que el usuario espera. Se dedicará esfuerzos al control de calidad de la solución.

Fuente: Propia

En a la **Tabla 6** se fundamenta el enfoque de la solución a largo plazo

Tabla 6. Enfoque a largo plazo

Nombre	Enfoque a largo plazo
Referencia	PDN03
Enunciado	Las decisiones se basan en estrategias aplicables a corto y largo plazo.

Fundamento Este principio promueve que el trabajo se enfoque principalmente en la solución a corto y largo plazo, al mismo tiempo ir mejorando con el uso de los datos obtenidos.

Repercusiones Se analizarán los procesos de mejor manera para así ir corrigiendo si existe errores

Fuente: Propia

2.5.1.3. Principio de Datos

En la **Tabla 7** se describe los fundamentos del principio de datos con respecto a la información relevante.

Tabla 7. Información Relevante

Nombre	Información relevante
Referencia	PDD01
Enunciado	La información debe generar valor al negocio
Fundamento	Este principio evita mantener costos innecesarios que se producen al mantener información que no tiene valor para la institución. Se tendrá datos fáciles de mantener y de analizar
Repercusiones	Se deberá revisar la base datos y los archivos temporales que se generan con la finalidad de identificar los que realmente son útiles

Fuente: Propia

En la **Tabla 8** se describe los fundamentos del principio de datos con respecto a la información accesible.

Tabla 8. Información accesible

Nombre	Información accesible
Referencia	PDD02
Enunciado	La información debe ser accesible para apoyar la productividad e innovación
Fundamento	Este principio permite al profesional tener acceso a los datos a fin de tomar decisiones de manera óptima.
Repercusiones	Se revisará el acceso de los usuarios a los datos y se proporcionaran la información útil para sus tareas

Fuente: Propia

En la **Tabla 9** se describe los fundamentos del principio de datos, con respecto a la seguridad de la información.

Tabla 9. Seguridad de la Información

Nombre	Seguridad de la información
Referencia	PDD03
Enunciado	Los datos son un activo que debe protegerse
Fundamento	Este principio permitirá que la información valiosa para la institución solo pueda ser accedida por el personal autorizado.
Repercusiones	Se realizará una revisión de la configuración de seguridad para identificar si posee vulnerabilidades que así poder corregirlas.

Fuente: Propia

En la **Tabla 10** se describe los fundamentos del principio de datos, con respecto a la copia de seguridad de los datos.

Tabla 10. Copia de Seguridad de los datos

Nombre	Copia de seguridad de datos
Referencia	PDD04
Enunciado	Todos los datos deben tener una copia de seguridad
Fundamento	Este principio garantiza que la información se mantenga a salvo por si un caso se genera algún tipo de error, ataque informático o desastre.
Repercusiones	Se deberá implementar cronogramas de creación de copias de seguridad de información de la base de datos. Se deberá mantener los servidores en funcionamiento óptimo para así mantener los respaldos en buenas condiciones.

Fuente: Propia

2.5.1.4. Principio de Aplicaciones

En la **Tabla 11**, se describe los fundamentos del principio de aplicaciones, con respecto al seguimiento de estándares.

Tabla 11. Seguimiento de Estándares

Nombre	Seguimiento de estándares
Referencia	PDA01

Enunciado	Los programas y aplicaciones para utilizar deberán cumplir con los estándares y procesos establecidos.
Fundamento	Este principio permite seleccionar de mejor manera el programa o aplicación que cumpla con lo requerido, siendo estos fácil de entender y mantener
Repercusiones	Se deberá capacitar al equipo con el uso y seguimiento de estándares para que si es el caso cambiar la adaptabilidad del software a usar

Fuente: Propia

En la **Tabla 12**, se describe los fundamentos del principio de aplicaciones, con respecto a la independencia tecnológica.

Tabla 12. Independencia de la Tecnología

Nombre	Independencia de la tecnología
Referencia	PDA02
Enunciado	El software usado debe permitir ser utilizada en diferentes hardware.
Fundamento	Este principio debe asegurar que los softwares utilizados, puedan implementarse en cualquier dispositivo, reduciendo así los costos de implementación.
Repercusiones	Las aplicaciones que solo funcionan con un fabricante se deberán migrar hacia el uso de software libre.

Fuente: Propia

En la **Tabla 13**, se describe los fundamentos del principio de aplicaciones, con respecto a la facilidad de uso de la aplicación.

Tabla 13. Aplicaciones Fáciles de usar

Nombre	Aplicaciones fáciles de usar
Referencia	PDA03
Enunciado	El Software o ampliación para utilizar debe ser amigable con el usuario
Fundamento	Este principio asegura que el software seleccionado, sea intuitivo es decir fácil de usar e implementar.

Repercusiones	Se debe invertir un mayor esfuerzo el entendimiento y entrenamiento del software a utilizar para que el usuario final, no tenga problemas del entendimiento de este.
----------------------	--

Fuente: Propia

2.5.1.5. Principio de Tecnología

En la **Tabla 14**, se describe los fundamentos del principio de tecnología, con respecto a la madurez de la tecnología.

Tabla 14. Tecnología madura

Nombre	Tecnología madura
Referencia	PDT01
Enunciado	Las tecnologías tempranas no se probarán a menos que sea realmente necesario.
Fundamento	Este principio reduce el riesgo de utilizar tecnologías que aún no hayan sido probadas en el mercado.
Repercusiones	Se deberá identificar los softwares que no cumplan con ciertos parámetros para el uso de estos.

Fuente: Propia

En la **Tabla 15**, se describe los fundamentos del principio de tecnología, con respecto si la tecnología es escalable.

Tabla 15. Infraestructura escalable

Nombre	Infraestructura escalable
Referencia	PDT02
Enunciado	La infraestructura deberá de ser capaz de soportar la adición de dispositivos.
Fundamento	Este principio permite que los softwares utilizados, no presenten problemas al momento de realizar algún cambio
Repercusiones	Se deberá probar los softwares a utilizar, para que al momento de la ejecución no presenten problemas al realizar un incremento de dispositivos

Fuente: Propia

En la **Tabla 16**, se describe los fundamentos del principio de tecnología, con respecto a la reevaluación de la seguridad.

Tabla 16. Reevaluar la seguridad

Nombre	Reevaluar la seguridad
Referencia	PDT03
Enunciado	La seguridad de los datos debe realizarlo de manera periódica
Fundamento	Este principio ayuda a mantener constantemente una seguridad alta para así poder determinar si se ejecuta algún ataque o determinar nuevas amenazas.
Repercusiones	Se deberá revisar y evaluar la seguridad de manera continua

Fuente: Propia

En la **Tabla 17**, se describe los fundamentos del principio de tecnología, con respecto al seguimiento que se le da al software.

Tabla 17. Seguimiento

Nombre	Seguimiento
Referencia	PDT04
Enunciado	Toda acción realizada debe ser rastreada
Fundamento	Este principio permite que toda acción realizada tenga un seguimiento ordenado para así no afectar la integridad de a información
Repercusiones	Se revisará los archivos logs generados en el portal web para determinar el estado de salud.

Fuente: Propia

2.6. Petición de trabajo de arquitectura

Esta sección es la petición de trabajo de arquitectura tecnológica para el proyecto de arquitectura empresarial que permitirá fortalecer el proceso de analítica web para determinar el status de los portales web de la UTN, mediante la aplicación de minería de datos web. Su propósito es describir un plan de como las soluciones serán abordadas a través del proceso de arquitectura.

2.6.1. Limitaciones Financieras

La UTN, es una institución educativa publica de tercer nivel que cuenta con una sólida condición financiera ya que en su mayor parte los recursos financieros son entregados por el estado ecuatoriano además de otros factores determinantes.

2.6.2. Descripción de la situación actual del Negocio

El proceso de analítica web dentro de la UTN, es uno de los más importante ya que permite mantener al profesional al tanto de los fallos que se generan en el portal de la institución hacia los usuarios finales.

2.6.3. Proceso de Estimación

Se encarga de estimar el tiempo y los recursos necesarios para llevar a cabo un requerimiento. Este se inicia cuando el profesional encargado del portal web, requiere una información complementaria del estado del portal Web.

2.6.4. Proceso de Ejecución

Comprende la planificación y ejecución del análisis. Se inicia cuando se requiere saber el estado en el que está el portal Web y su incidencia de errores.

2.7. Descripción de la situación actual de la analítica web

El software utilizado para ayudar al profesional a dar seguimiento del estado actual del portal web es el siguiente:

Similarweb

Es una herramienta de análisis de aplicaciones y sitios web que ayuda a conocer el mercado y monitorear a los competidores. La clasificación en la búsqueda, el número de visitas y las fuentes de tráfico son informaciones que puedes recopilar con el uso de esta herramienta(Anon n.d.).

Esta aplicación soporta varios procesos tal como se muestra a continuación:

- Ranking Global
- Clasificación por país
- Rango de categoría
- Total, de visitas
- Duración media de todas las visitas
- Audiencia (Principales países desde donde se realizan las consultas)
- Intereses
- Competidores y sitios similares
- Descripción general de los canales de marketing
- Tráfico de búsqueda de palabras clave
- Tráfico de referencia
- Trafico de redes sociales
- Enlaces salientes

Los principales problemas encontrados son los siguientes:

- Este proceso no cuenta con subprocesos o pasos detallados para la realización la analítica web.
- No se utilizan los archivos .log para en análisis del status del portal web UTN.
- No se cuenta con una arquitectura empresarial tecnológica estandarizada para la realización de procesos.

2.8. Visión de arquitectura

2.8.1. Declaración de trabajo de arquitectura

La Universidad técnica del Norte fue creada en el año de 1987, siendo un referente de educación superior en el norte del país, desde ahí ha tenido un gran avance científico, educativo y tecnológico.

Con la llegada de las nuevas tecnologías se ha ido implementando nuevos procesos, nuevas herramientas para mantener la información segura de todos los que conformamos la UTN, cada vez es más imprescindible conocer el estado de salud de los

portales web que la institución maneja, con el aumento de elementos tales como estudiantes y docentes también aumenta el peligro de saturación en las bases de datos.

El mejoramiento constante de la institución ha dado lugar a una nueva categorización del ranking, nacional, regional y mundial, que cada vez aumenta, por lo tanto, el número de consultas que se genera minuto a minuto en el portal web (blog, página oficial, biblioteca), también aumenta notablemente.

2.8.2. Descripción del proyecto de arquitectura y alcance

El presente proyecto de arquitectura es una propuesta de arquitectura tecnológica para el proceso de minería de datos web del portal web de la UTN, utilizando técnica de Web Usage Mining es decir el análisis de los archivos. logs generados por el portal. Por ello se solicita el presente proyecto de arquitectura cuyo objetivo es optimizar el proceso de analítica web, específicamente el análisis de archivos .log, con el propósito de dar un mejor uso de los recursos tecnológicos de la UTN y presentar de mejor manera un análisis de status del mismo.

En la Fig. 17, se muestra el diagrama de procesos de la analítica web, como se realiza en la actualidad:

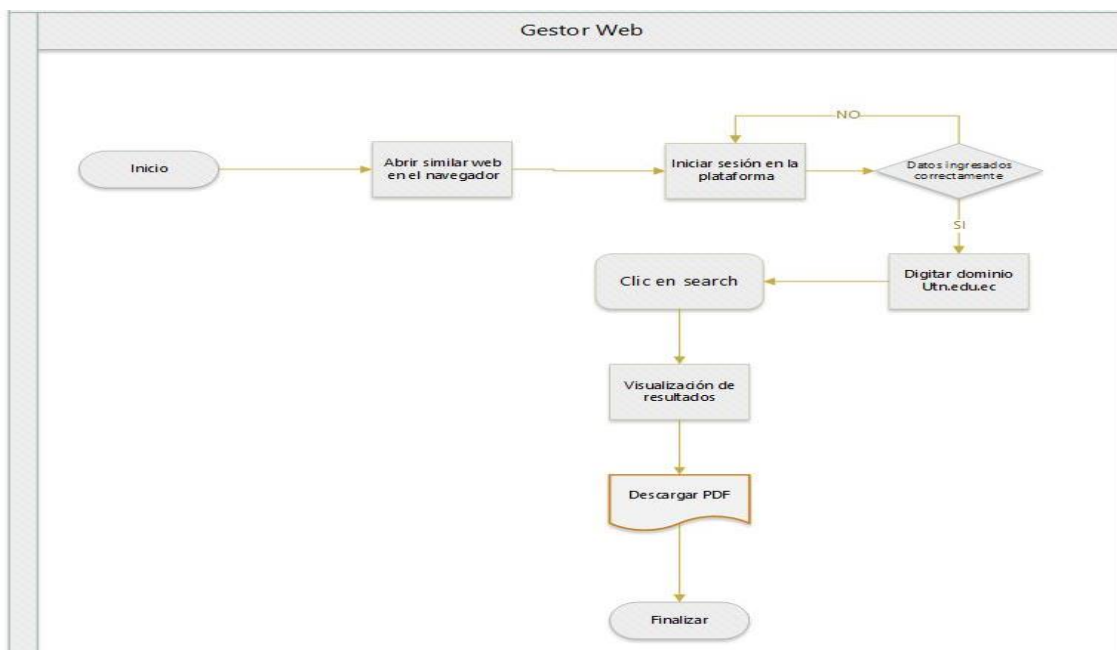


Fig. 17. Diagrama de Análisis del portal web UTN
fuente: Propia

2.8.3. Roles responsabilidades y entregables

En la siguiente matriz se muestra la relación entre los roles, partes interesadas que intervienen en el proyecto y los entregables.

Tabla 18. Matriz Roles responsabilidades y entregables

Entregables	Roles			
	Director de DDTI UTN	Encargado gestión web UTN	Administrador de redes y telecomunicaciones	Miembros del Proyecto
Principios de Arquitectura	C	R	I	R
Petición de Trabajo de Arquitectura	I	R	I	C
Declaración de trabajo de arquitectura	I	R		
Visión de Arquitectura	I	R	C	C
Documento de definición de Arquitectura				R
Arquitectura de Negocio	I	C	I	R
Arquitectura de aplicación	I	C	I	R
Arquitectura de datos	I	C	C	R
Arquitectura tecnológica	I	C	C	R
Plan de implementación y Migración	I	C	C	R

Fuente: Propia

Leyenda: R= Responsable; E= Encargado; C= Consultado; I= Informado

Criterios de aceptación

- Los criterios de aceptación de los entregables son los siguientes:
- Se debe garantizar que exista trazabilidad entre los elementos considerados, actividades, propuestas, etc.; contra los principios de arquitectura definidos los cuales a su vez deberán estar alineados con los objetivos estratégicos del proyecto.

- Desarrollares las secciones obligatorias de la documentación propuesta, según el marco de trabajo TOGAF.
- Cumplir con las buenas prácticas definidos por el proceso de analítica web, específicamente Web Usage Mining, análisis de archivos .log.

Los entregables deben contener:

- Diagrama de los procesos para la arquitectura de negocio.
- Diagrama de componentes para la arquitectura de aplicaciones.
- Diagrama de modelo lógico y físico para la arquitectura de datos.
- Diagrama de Hardware y red para la arquitectura Tecnológica.

2.8.4. Descripción del problema

2.8.4.1. Interesados y sus preocupaciones

Los principales Interesados (Stakeholders) son, MSc. Gabriela Cárdenas, MSc. Juan Carlos Garcia y el MSc. Vinicio Guerra, profesionales que están al frente del área tecnológica de la UTN.

Su preocupación principal es fortalecer el proceso de analítica web de los archivos provenientes del portal web de la UTN.

2.8.4.2. Lista de asuntos y escenarios que deben abordarse

El propósito principal de la arquitectura propuesta es proporcionar principios y prácticas que ayuden a mejorar el proceso de analítica web, específicamente aplicando, principios de minería de datos web, en este caso Web Usage Mining, es decir el análisis de los archivos .log.

A continuación, se presentan los principales hechos y problemas encontrados en el proceso de analítica web, a los cuales se pretende dar solución con la presente propuesta de arquitectura.

Tabla 19. Descripción de Hechos y problemas

Hechos	Problemas
Se realiza la comprobación del portal, utilizando herramienta de comprobación de tráfico y ranking de esta.	En la practica el uso de esta herramienta no nos permite visualizar la incidencia de errores, en el momento que haya ocurrido esta.
La información del recurso utilizado para realizar la comprobación de trafico de red, solo es manejada por el	Este procedimiento puede resultar en parte un poco perjudicial ya que, sin la presencia del profesional a cargo, no habría una

profesional que está a cargo del funcionamiento del portal web	respuesta rápida de solución de cualquier incidencia. La información solo permite la visualización sin guardar ningún reporte que pueda servir para un posterior análisis.
No se utilizan herramientas para un análisis de archivos .log, generados por el portal web de la UTN.	Al no tener una herramienta especializada para el análisis complementario de los archivos .log, se deja mucha información importante sin utilizar para el correspondiente análisis del status y la incidencia de errores que estas pueden guardar.

Fuente: Propia

Según los problemas descritos, se tiene una principal causa es de ellos es que no existe procedimientos claramente establecidos para la realización del proceso de analítica web, utilizando los archivos .log, ni tampoco una guía de lo que se puede encontrar realizando en análisis de este tipo de archivos que son generados cada vez que un usuario hace una petición dentro del portal. Esta información puede ser de mucha utilidad para determinar la cantidad de visitas, la cantidad de incidencia de errores que genero el portal, al mismo tiempo si el portal en uno de sus procesos sufrió algún ataque o perdida repentina de la información.

Para evitar estos problemas, la propuesta de arquitectura actual deberá revisar y proporcionar procedimientos actuales a fin de proponer mejoras. Además, debe plantear una solución informática que funcione como un soporte para los procedimientos y permita la integración de nuevas herramientas que ayuden al proceso analítica web.

2.9. Arquitectura (AS IS / TO BE)

2.9.1. Documento de definición de arquitectura

2.9.1.1. Alcance

El alcance del presente proyecto deberá abarcar los cuatro dominios de arquitectura:

Arquitectura de Negocios, mediante la cual se define la estructura de la organización, el mapa de procesos y procesos claves para la organización objetivo(The Open Group 2018).

Arquitectura de Datos, mediante la cual se describe la estructura de los datos lógicos de la organización.

Arquitectura de aplicaciones, mediante la cual se describe cada una de las soluciones de software utilizados en los procesos.

Arquitectura tecnológica, la cual describe la estructura física de los componentes que dan soporte a las soluciones en la organización.

El proceso que se abarcará en este análisis es el proceso de analítica web, el cual se encarga de proporcionar la información necesario acerca del estado de salud del portal web de la UTN, control de incidencia de errores mediante el análisis de los archivos .log.

2.9.1.2. Metas, objetivos y limitaciones

- Fortalecer el proceso de analítica web.
- Establecer pasos detallados del uso de herramientas para el análisis de los archivos .log
- Brindar reportes claros acerca del proceso de análisis del contenido de los archivos .log

2.9.1.3. Principios de arquitectura

Tabla 20. Principios de Arquitectura

DOMINIO	PRINCIPIO
	Alineación entre TI y el negocio
Negocio	Enfoque en el cliente
	Enfoque a largo plazo
	Información relevante
Datos	Información accesible
	Seguridad de la información
	Copia de seguridad de datos
	Seguimiento de estándares
Aplicaciones	Independencia de la tecnología
	Aplicaciones fáciles de usar
	Reutilización y simplicidad
	Tecnología madura
Tecnología	Infraestructura escalable
	Reevaluar la seguridad
	Seguimiento

Fuente: Propia

2.10. Arquitectura línea base

2.10.1. Arquitectura del negocio

2.10.1.1. Estructura Organizacional

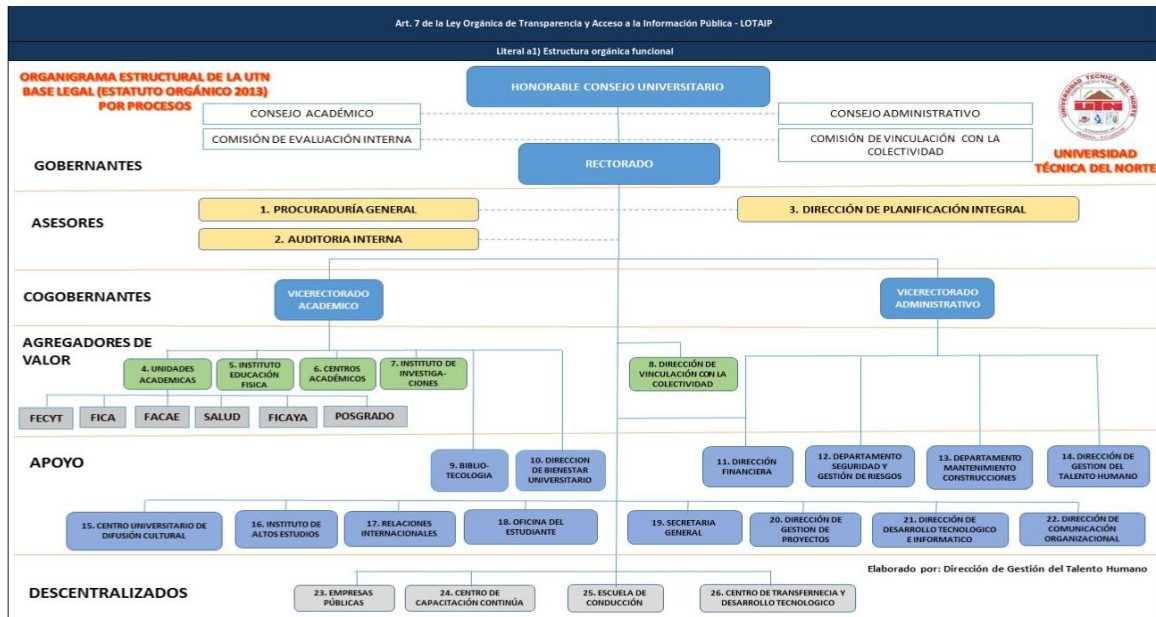


Fig. 18. Estructura Organizacional UTN

Fuente: (Anon n.d.)

2.10.1.2. Organigrama dirección Informática UTN

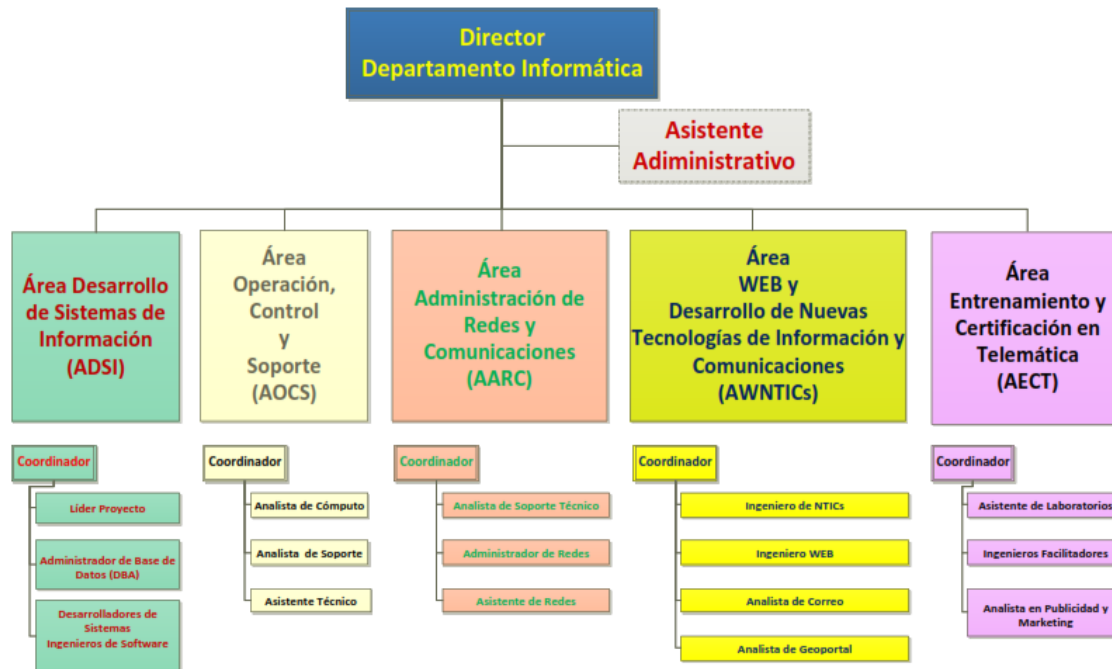


Fig. 19. Organigrama Departamento Informático UTN

Fuente: (Anon n.d.)

2.10.1.3. Mapa de procesos del Negocio



Fig. 20. Mapa de procesos UTN

Tomando en cuenta el mapa de procesos de la institución, se pretende fortalecer el proceso de analítica web que esta, dentro delo proceso de soporte, gestión tecnológica el cual lo realiza el departamento de TI de la UTN, más específicamente el profesional que está a cargo de la gestión de recursos web.

2.10.1.4. Roles y partes interesadas

En la **Tabla 21**, se muestra los roles y las partes interesadas(stakeholders) que conforman el proyecto.

Tabla 21. Roles y partes interesadas

ID	Cargo	Descripción	Persona encargada
S1	CIO	Director de tecnologías de la información	Ing. Juan Carlos García
S2	Web master	Encargado del correcto funcionamiento del sitio web.	Ing. Gabriela Cárdenas
S3	Redes comunicación	de Se encarga de organizar, instalar y brindar soporte al sistema informático de una organización asegurándose el correcto funcionamiento de las redes informáticas	Ing. Vinicio Guerra

		internas y externas, según los niveles de servicio operacional y de seguridad que se establezcan		
S4	Miembros de proyecto	Personas encargadas de desarrollar la arquitectura	MSc, Guevara	Alexander

Fuente: Propia

2.10.1.5. Modelo de datos

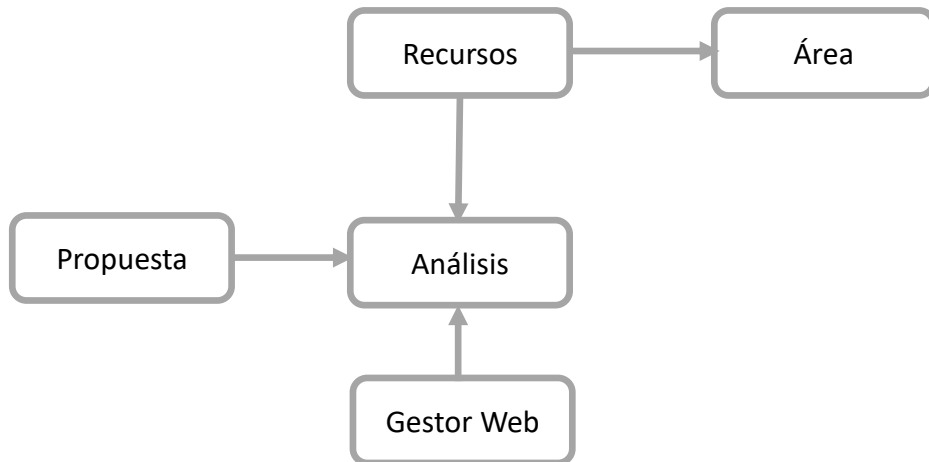


Fig. 21. Modelo de Datos Análisis web
Fuente: Propia

2.10.1.6. Diagrama de actividades

En la Fig. 22 se muestra como actualmente se realiza el proceso de análisis del status del Portal Web de la UTN.

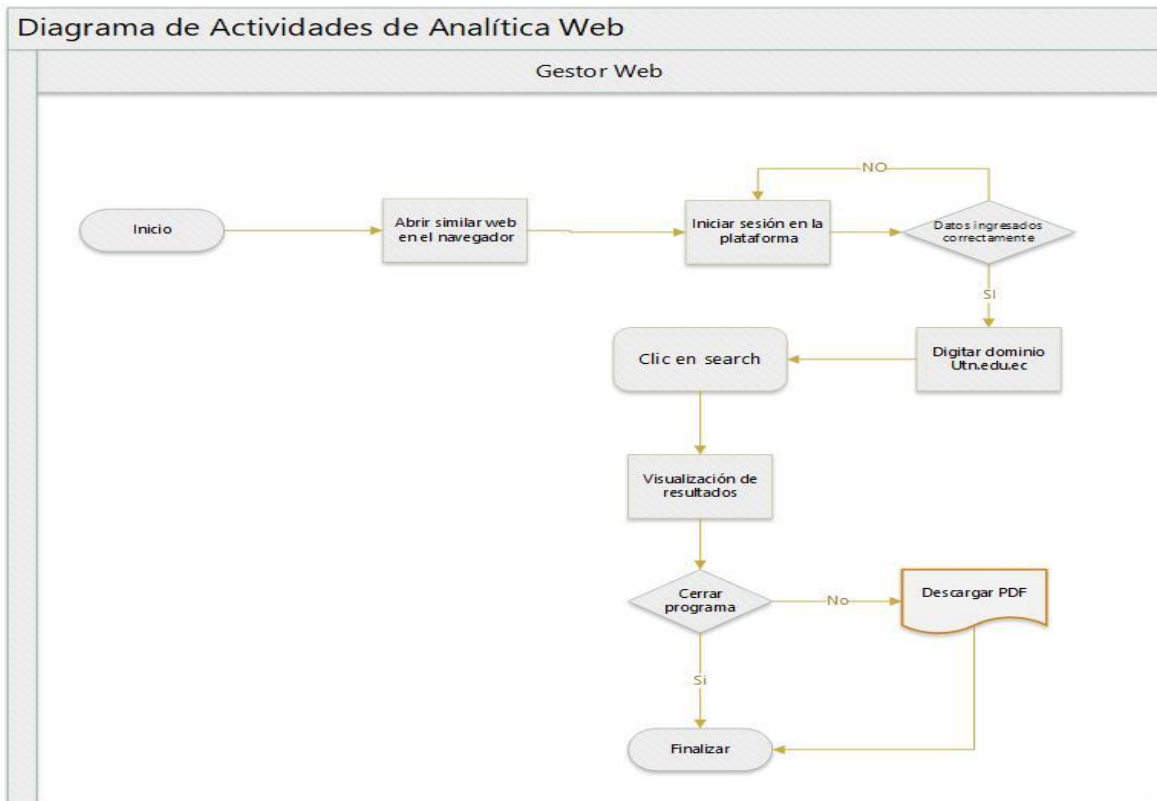


Fig. 22. Diagrama del proceso de analítica web
Fuente: Propia

2.10.2. Arquitectura de datos

Para representar la arquitectura de datos se realiza una vista lógica del proceso de analítica web representado mediante diagrama de clases como se muestra en la Fig. 22.

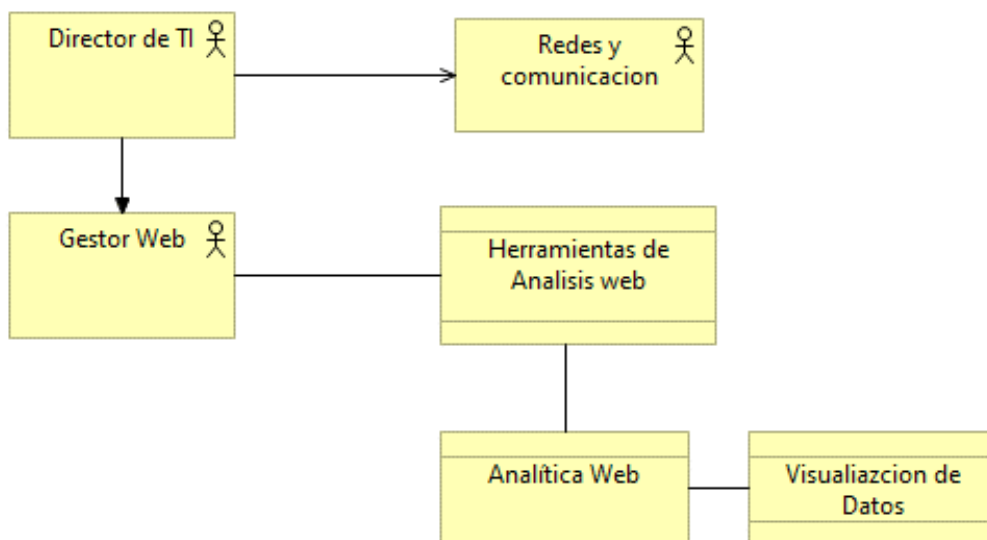


Fig. 23. Diagrama Lógico
Fuente: Propia

En la **Tabla 22**, se describe cada uno de los componentes del diagrama lógico

Tabla 22. Descripción componentes de diagrama lógico

ID	ENTIDAD	DESCRIPCIÓN
E01	Director TI	Director del departamento informático
E02	Redes y comunicación	Determina persona encargada de mantener en buen funcionamiento la conectividad interna
E03	Gestor web	Quien realiza el análisis
E04	Herramientas de análisis	Determina la herramienta utilizada para el proceso de analítica web
E05	Analítica web	Proceso de análisis del portal
E06	Visualización de datos	Visualización de datos obtenidos mediante el análisis con la herramienta utilizada.

Fuente: Propia

Se puede determinar que la responsabilidad de la analítica web recae sobre el profesional encargado de la gestión del Portal Web de la UTN, es quien realiza el análisis del status de este.

2.10.3. Arquitectura de aplicaciones

Listado de aplicaciones

Se menciona que el proceso de análisis del status, del portal web UTN, se utiliza la herramienta tecnológica SIMILARWEB, la cual permite conocer diferentes puntos clave del portal web, tales como el Ranking local y mundial, de todas las aplicaciones utilizadas por la UTN, el tráfico de red, desde que país se realiza la consulta, el interés de la audiencia es decir lo más buscado en este caso educación, sus principales competidores como entidad educativa, entre otras cosas.



Fig. 24. Diagrama de Arquitectura de Aplicaciones
Fuente: Propia

2.10.4. Arquitectura tecnológica

Se detalla los componentes de tecnología utilizado y la relación con el sistema de información.

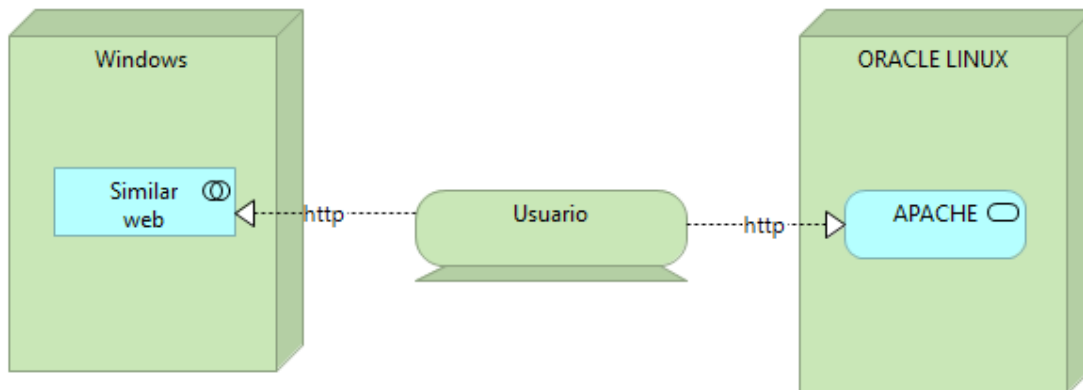


Fig. 25. Diagrama de componentes
Fuente: Propia

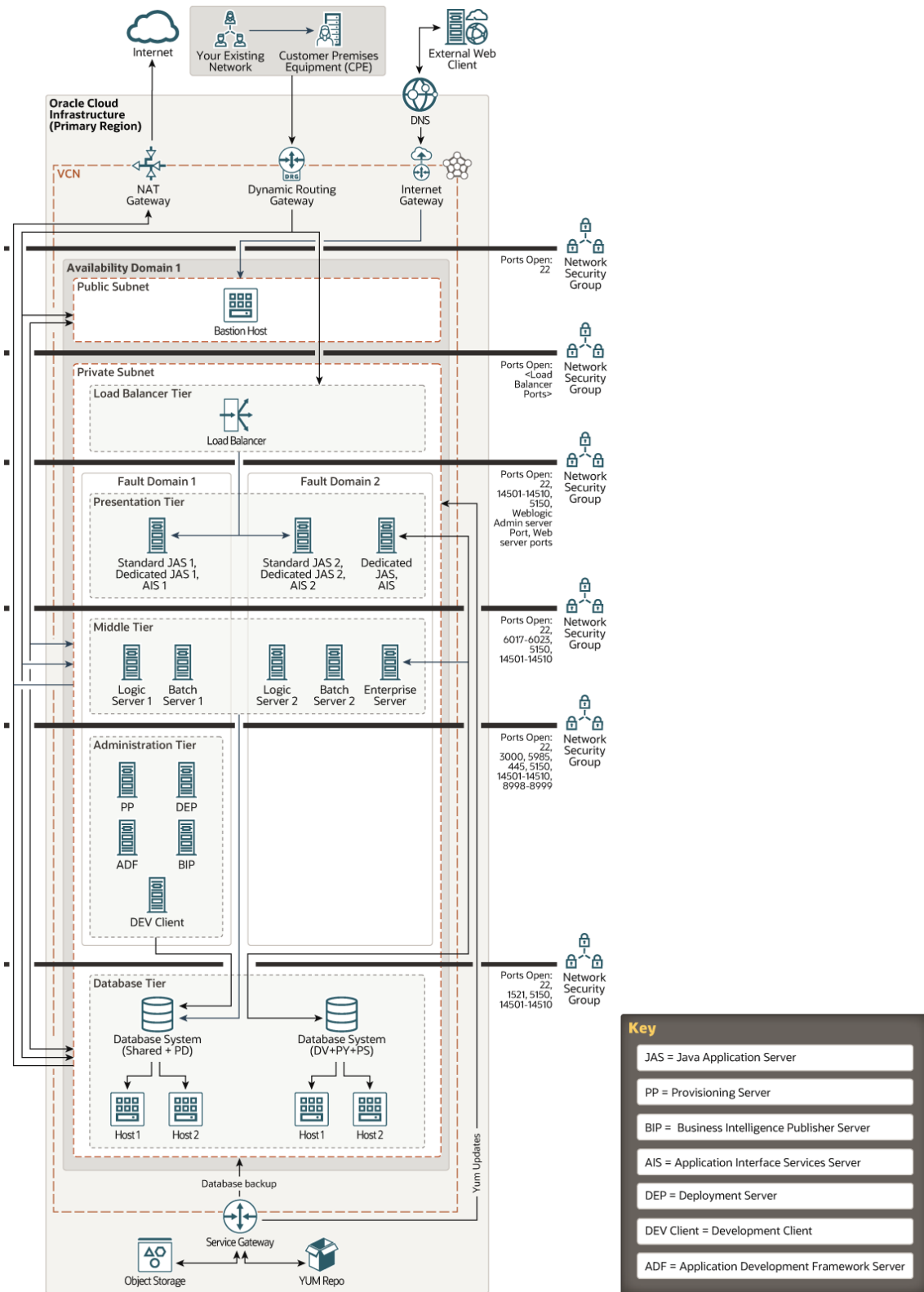


Fig. 26. Oracle Cloud Infraestructura
Fuente:(Marinescu 2013)

2.10.4.1. Descripción de la plataforma tecnológica

Para el proceso de análisis o visualización del estado del portal web el usuario realiza la petición, ingresando al dominio <https://www.similarweb.com/>.

Esta página web de verificación de Rankin de aplicaciones y portales, puede ser ejecutado bajo cualquier sistema operativo, ya que solo es necesario abrir el navegador e ingresar el dominio u objeto de análisis que corresponda, en este caso utn.edu.ec.

En conformidad, la UTN, cuenta con servidores en la nube como es Oracle Cloud, el cual se ejecuta bajo el sistema operativo Oracle Linux, la capacidad de almacenamiento y velocidad depende específicamente lo que el proveedor asigne dependiendo el tipo de contrato.

Servidor apache nos permite mostrar el contenido de las aplicaciones o páginas web bajo su dominio, igualmente es aquí donde se guardan los registros de errores, registros de navegación.

2.10.5. Especificaciones de hardware y red

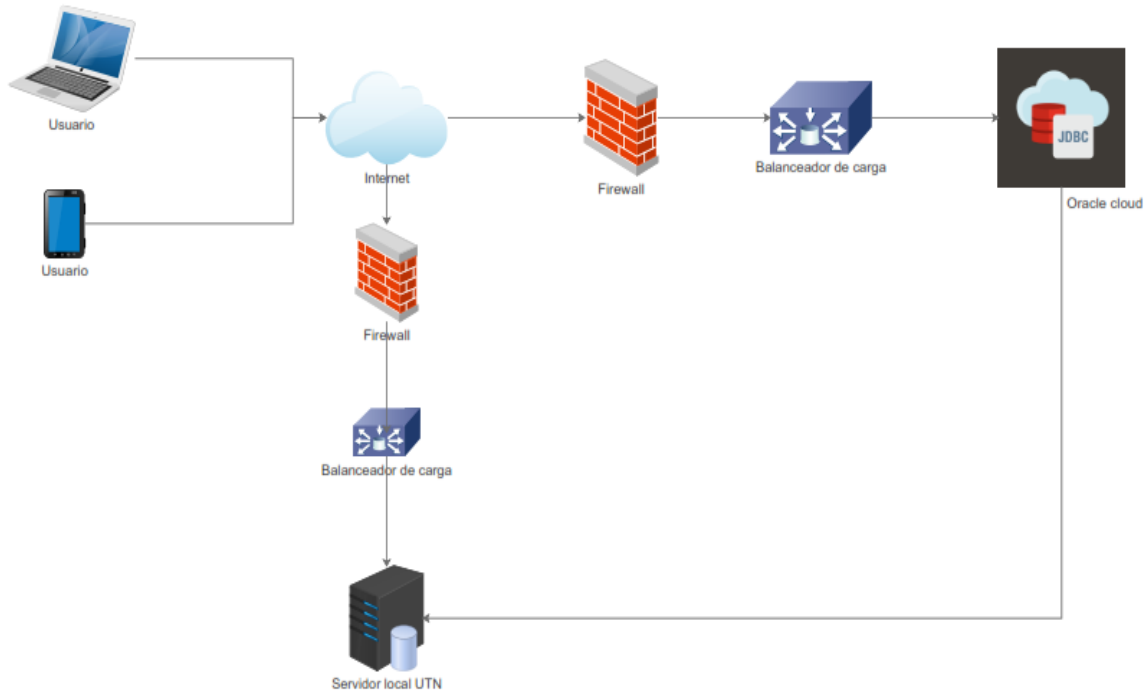


Fig. 27. Diagrama de hardware y red
Fuente: Propia

2.10.5.1. Descripción de componentes

Tabla 23. Descripción de componentes principales.

ID	COMPONENTE	DESCRIPCIÓN
C01	Usuario	Representación de equipo desde el cual se realiza una petición (PC o Notebook)
C02	Usuario 1	Representación del equipo móvil del cual se puede realizar la conexión o petición
C03	Internet	Elemento fundamental en el diagrama de red, ya que es el medio por el cual se puede realizar todas las peticiones
C04	Balancedador	Herramienta que permite direccionar a un cliente al servidor web que se encuentre con mayor disponibilidad entre los que cuentan con el mismo contenido
C05	Firewall	Dispositivo de seguridad de la red, que permite monitorear el tráfico de red entrante y saliente y este decide si permite la entrada o no.
C06	Oracle Cloud	Servidor en la nube, donde se encuentran todas las aplicaciones, portales y servicios de la UTN, también cuenta con sus propios backups y seguridad contra fallos y ataques, es menos propenso a estos ya que cuenta con varios estándares de seguridad, que hacen una aplicación segura
C06	Servidor local	Servidor físico, donde se almacenan todas las aplicaciones locales, backups, aplicaciones de prueba, entre otros datos y aplicaciones, funciona también como un respaldo si la conectividad con Oracle falla.

Fuente: Propia

2.10.6. Fundamentos y justificación del enfoque arquitectónico

En base a la arquitectura empresarial analizada y los procesos, se realiza la definición de los principales problemas y requerimientos que serán resueltos, en la arquitectura empresarial que será planteada.

2.10.6.1. Problemática del proceso

- No se cuenta con los recursos necesarios para el análisis de archivos generados por el portal web
- Para el análisis de status del portal web, se usa una única herramienta tecnológica denominada similar web.
- En el proceso de análisis del status del portal web de la UTN, no se utilizan los archivos .log.
- No existe un proceso de análisis de los archivos .log.
- No se utilizan herramientas de análisis de archivos .log

2.10.6.2. Principales requerimientos

En la siguiente **Tabla 24** se muestra los principales requerimientos por parte de los stakeholders:

Tabla 24. Requerimientos

ID	Requerimientos de la arquitectura a diseñar
R1	Diseñar una arquitectura tecnológica que permita fortalecer el proceso de analítica web de los archivos provenientes del portal web de la UTN.
R2	La arquitectura diseñada deberá permitir entender el proceso de analítica web de una manera más fácil y eficiente.
R3	La arquitectura diseñada deberá permitir conocer las herramientas necesarias que permitan manipular de una manera más sencilla el contenido de archivos .log.
R4	La arquitectura diseñada deberá permitir seleccionar herramientas de análisis de archivos log, para la obtención de datos o patrones que proporcionen información relevante del contenido de los archivos y lo que se puede descubrir con estos.
R5	La arquitectura diseñada deberá proporcionar la documentación o guía necesaria para poder determinar la mejor herramienta para el proceso de analítica de los .log
R6	La Herramienta seleccionada debe contar con su manual de instalación en caso de ser necesario y manual de uso de esta.
R7	Implementar la arquitectura

R8 La arquitectura diseñada deberá proporcionar un metamodelo del funcionamiento de esta.

Fuente: Propia

2.11. Arquitectura de destino

Para describir la arquitectura de destino, se toma en cuenta que hay factores anteriormente descritos que no necesitan ser modificados, y algunos parámetros simplemente se generan nuevos procesos que permitan el fortalecimiento del proceso de analítica web y el análisis de archivos .log.

2.11.1. Arquitectura de negocio

No se modifica la estructura organizacional del departamento informático de la UTN.

2.11.2. Mapa de procesos de negocio

El mapa de proceso del negocio tampoco se modifica ya que están todos sus procesos detallados.

Los roles tampoco se modifican ya que cuenta con la distribución adecuada detallada por la institución

2.11.3. Diagrama de actividades (solución)

En la **Fig. 28**, se muestra el diagrama de actividades propuesto para el proceso de analítica Web

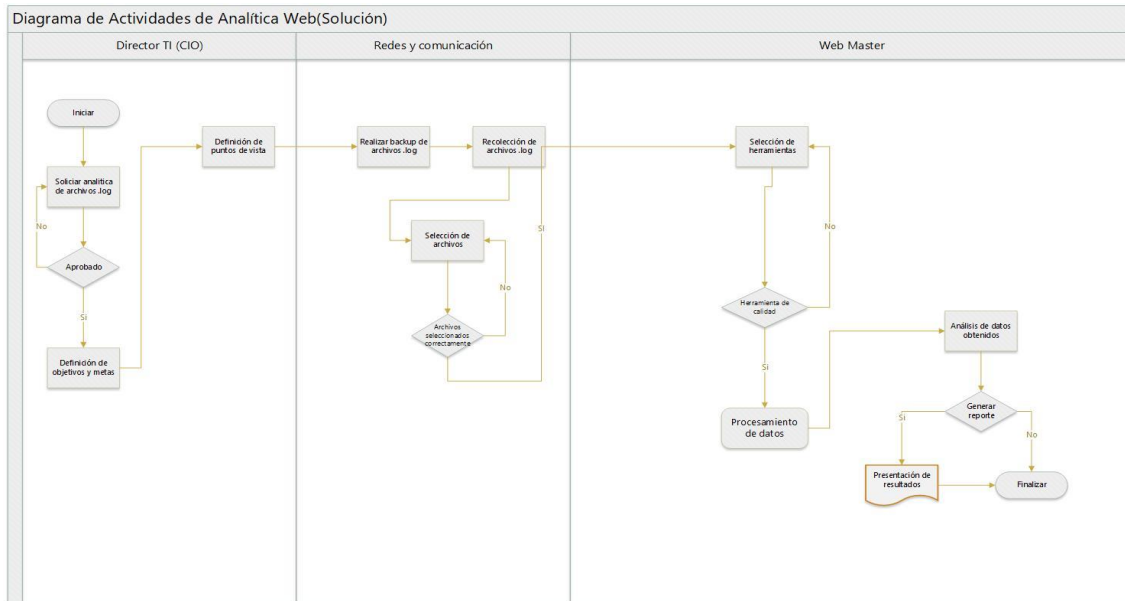


Fig. 28. Diagrama de Actividades
Fuente: Propia

2.11.4. Arquitectura de datos

2.11.4.1. Modelo de datos lógico

A continuación, se presenta un modelo de datos lógico propuesto el cual integra la línea base para el proceso de analítica Web, se elimina algún elemento relevante o redundante y se presenta un modelo de datos que soporte la solución propuesta.

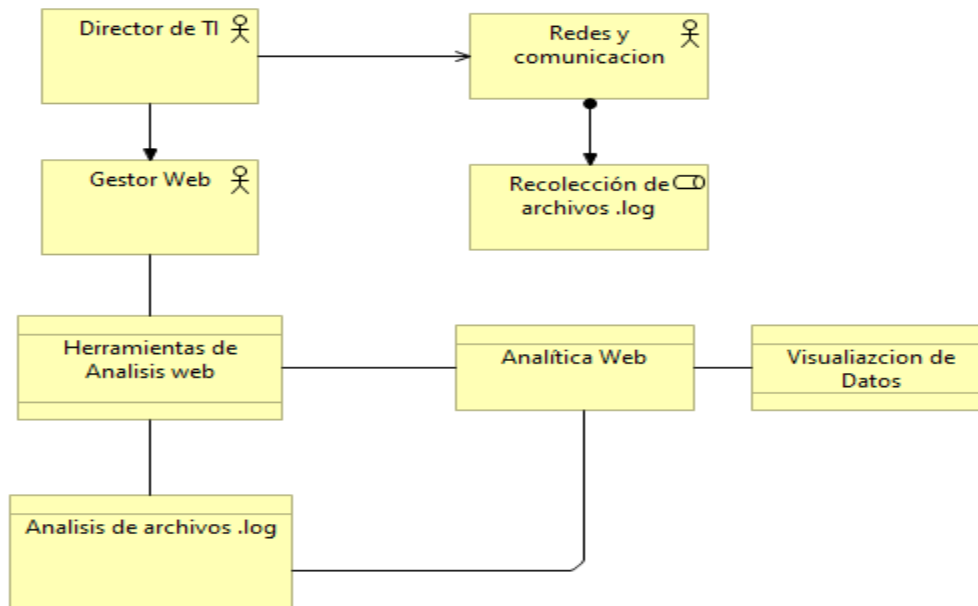


Fig. 29. Modelo Lógico (solución)
Fuente: Propia elaborado en Archi

En la **Tabla 25**, se muestra la descripción del modelo lógico de la solución.

Tabla 25. Descripción Modelo Lógico

ID	ENTIDAD	DESCRIPCIÓN
E01	Director TI	Director del departamento informático
E02	Redes y comunicación	Determina persona encargada de mantener en buen funcionamiento la conectividad interna
E03	Recolección de archivos .log	Determina una función asignada para el apoyo en la realización de backup de los archivos .log y su posterior selección de estos.
E04	Gestor web	Quien realiza el análisis
E05	Herramientas de análisis	Determina la herramienta utilizada para el proceso de analítica web
E06	Análisis de archivos .log	Sección donde mediante la herramienta seleccionada se realiza el análisis de los archivos .log
E07	Analítica web	Proceso de análisis del portal mediante los resultados obtenidos de los archivos .log
E08	Visualización de datos	Visualización de datos obtenidos mediante el análisis con la herramienta utilizada.

Fuente: Propia

2.11.5. Arquitectura de aplicaciones

Para determinar las herramientas que permita fortalecer el proceso de analítica Web del portal de la UTN, se propone incluir al proceso actual el uso de herramientas con que permitan la lectura, análisis y representación de archivos provenientes de las peticiones realizadas al portal, estos archivos son: error_log, acces_log,



Fig. 30. Diagrama de Arquitectura de Aplicaciones
Fuente: Propia elaborada con Archi

2.11.6. Arquitectura Tecnológica

A continuación, se presenta el modelo de arquitectura tecnológica propuesta, el cual consiste en mostrar las aplicaciones que se pueden utilizar para el proceso de Web Mining y análisis de archivos .log, con la finalidad de fortalecer el proceso de analítica Web.

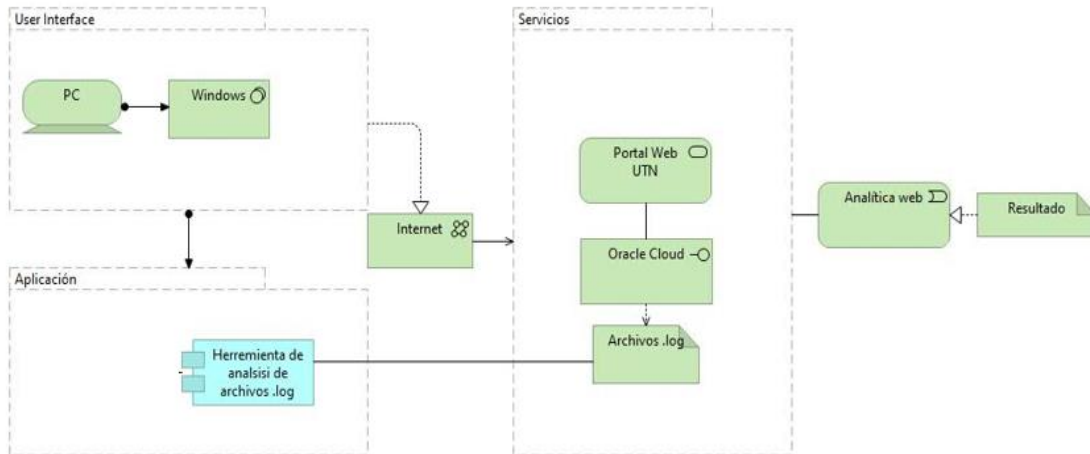


Fig. 31 Arquitectura Tecnológica
Fuente: Propia elaborada en archi

2.11.6.1. Descomposición de tecnología y su descripción

Para representar la arquitectura tecnológica se realiza un diagrama de componentes que integran en la analítica web.

2.11.6.2. Interfase de usuario

Dentro de esta sección se describe, el componente de hardware que se puede utilizar con su respectivo sistema operativo.

2.11.6.3. Aplicación

Para representar esta capa tomamos en cuenta las herramientas que se va a utilizar dentro del proceso de analítica web, además de las herramientas de análisis de archivos .log, las cuales son las siguientes:

- Similar web
- Weblog expert
- Screaming frog log analyzer
- RStudio
- Log File Analyzer/semrush

2.11.6.4. Servicios

Para la representación de la capa de data se toma en cuenta el contenido de la base de datos que se va a utilizar, en este caso los archivos .log, generados por el portal web, al final se tiene el proceso de analítica web que genera un reporte de los datos analizados.

2.11.7. Especificación de hardware y red

Nota: No se realiza ningún cambio al diagrama de hardware y red

2.12. Análisis de brechas

En esta sección se describen las brechas encontradas, por cada uno de los dominios de arquitectura, entre la arquitectura de línea base y la arquitectura objetivo.

2.12.1. Arquitectura de Negocio

En la **Tabla 26**, se muestra el análisis de brechas en el proceso de analítica web:

Tabla 26. Análisis de Brechas proceso de analítica web

	Arquitectura Destino										
Arquitectura línea base	Análisis de Rankin	Visitas totales en	Duración de visita	Paginas por visita	Número de	Palabras clave	Principales	Análisis del portal	Análisis de Archivos	Incidencia de errores	Generar reporte de análisis
Análisis de ranking											
Visitas totales en los últimos 6 meses											
Duración de visita											
Paginas por visita											
Número de Visitantes por país											
Palabras clave											
Principales											
Competidores											
Análisis del portal Web											

Backups de archivos .log	
Análisis de archivos .log	A
Incidencia de errores	A
Generar reporte de análisis	A

Fuente: Propia

Leyenda: (A= Actualizar; E=Eliminar; I=Implementar)

2.12.1.1. Brechas proceso de analítica web

Actualizar

- Análisis de archivos .log. – después de realizar el backup de los archivos .log, se puede realizar el análisis con las herramientas correspondientes.
- Generar reporte. – Se debe generar el reporte del análisis realizado con la herramienta seleccionada y presentar el reporte.

2.12.2. Arquitectura de aplicaciones

En la **Tabla 27**, se muestra la arquitectura de aplicaciones destino, con la implementación de la herramienta de análisis de archivos .log.

Tabla 27. Arquitectura de Aplicaciones

Arquitectura Destino	
Similar Web	Herramientas de análisis de archivos .log
Arquitectura línea base	
Similar Web	
Herramienta de análisis de archivos .log	I

Fuente: Propia

Leyenda: (A= Actualizar; E=Eliminar; I=Implementar)

2.12.2.1. Brechas de aplicaciones

Implementar

Herramientas de análisis de archivos .log. – La implementación de este tipo de herramientas, nos permite un análisis profundo y exhaustivo de los archivos generados por el portal web de la UTN, las cuales nos permite conocer varios aspectos primordiales, como, por ejemplo, la incidencia de errores, si la página ha sido víctima de ataque, la URL, el Bot mediante el cual se realiza la consulta entre otras especificaciones, como se muestra en la Figura 32:



Fig. 32. Estructura Archivos .log
Fuente: Propia

2.12.3. Arquitectura tecnológica

Análisis de brechas

Se determina que dentro de la arquitectura de tecnología no es necesario el cambio de ningún elemento de hardware, simplemente se propone la utilización o integración de una herramienta de analítica web de las anteriormente mencionadas.

2.13. Oportunidades y soluciones

2.13.1. Plan de implementación y migración

En la Fig. 33, se muestra la estructura y desglose del trabajo realizado



Fig. 33. Estructura y desglose del trabajo
Fuente: Propia

Para realizar el plan de implementación es necesario tomar en cuenta las acciones y pasos que se realizan para llegar a la aplicación correcta, para así lograr un cambio o innovación en el proceso de analítica web.

Para la implementación de la solución se toma en cuenta los siguientes pasos o procesos que a continuación se describen:

- a) Socializar a las partes interesadas el documento de descripción de arquitectura.
- b) Presentar la solución, que permita fortalecer el proceso de analítica web de la UTN.
- c) Socializar las herramientas de análisis de archivos .log
- d) Facilitar los instaladores y manuales de instalación de las herramientas a utilizar para las respectivas pruebas.
- e) Generar charlas informativas de la configuración y uso de las diferentes herramientas de análisis de archivos .log.
- f) Ejecutar el análisis de los archivos .log mediante una prueba de concepto, con las diferentes herramientas para así determinar cuál, de ellas, ayuda al fortalecimiento

de la analítica web, mediante el proceso de Web Usage Mining -análisis de archivos .log

g) Presentar resultados y discusiones.

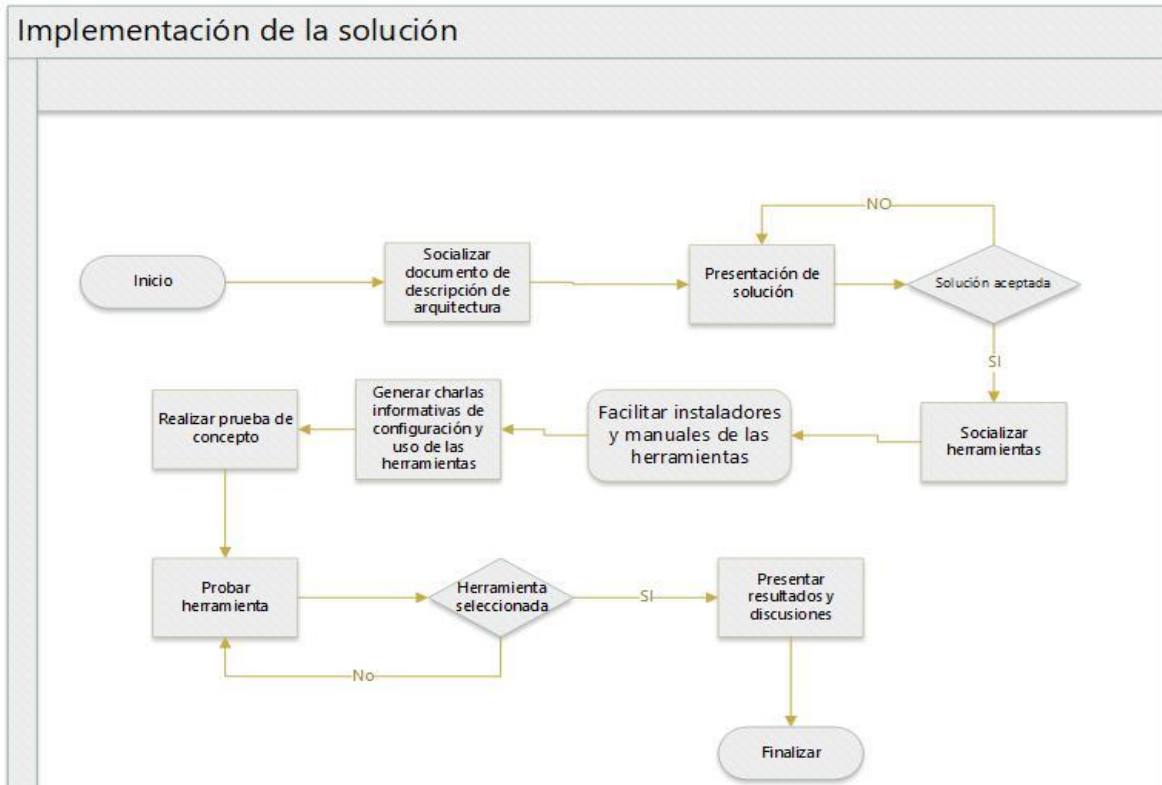


Fig. 34. Plan de implementación
Fuente: Propia

2.14. Solución (Metamodelo)

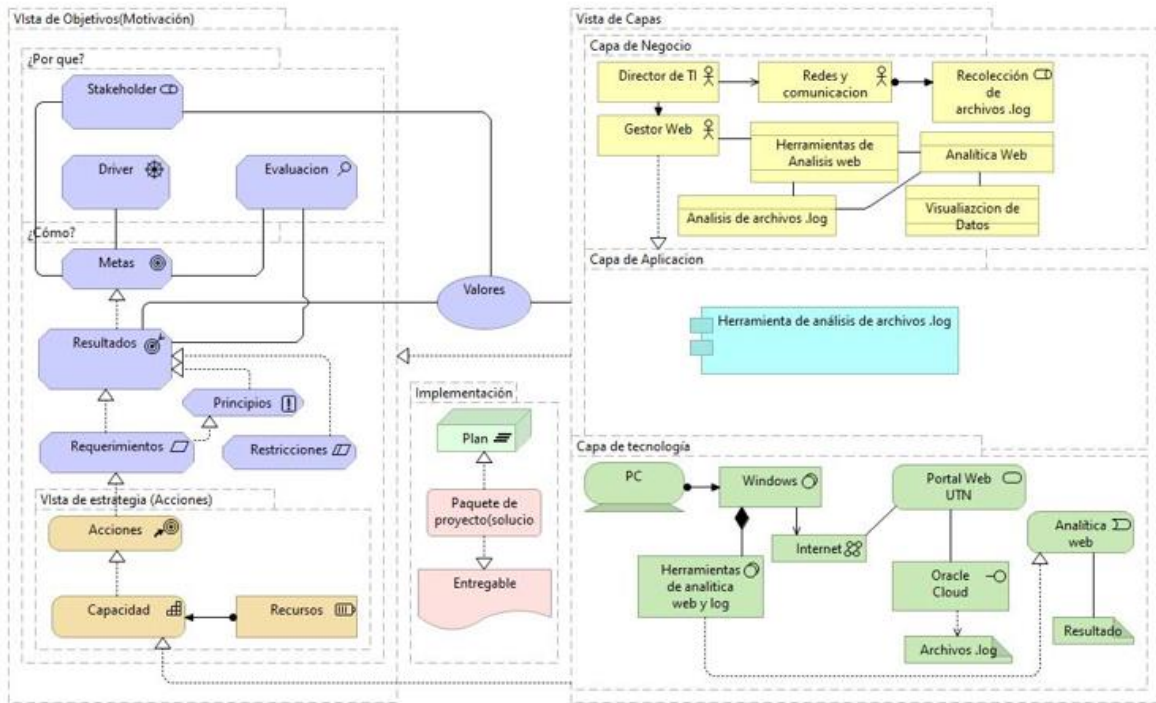


Fig. 35. Metamodelo analítico web
Fuente. Propia elaborada en Archi

2.15. Fase de evaluación de la arquitectura

Para esta fase se realiza un checklist que nos permite evaluar que principios del Estándar Internacional ISO/IEC/IEEE 42010, cumple la arquitectura sugerida, la lista de verificación se muestra en la **Tabla 28**.

Tabla 28 Checklist de evaluación de cumplimiento de principios del estándar ISO/IEC/IEEE 42010

Clausula	Principio	Especificación	Cumple	
			Si	No
Clausula N.º 5	Descripción de la arquitectura	5.2 Descripción de la arquitectura y descripción general.	✓	
		5.3 Identificación de las partes interesadas y preocupaciones.	✓	
		5.4 Puntos de vista de la arquitectura.	✓	
		5.5 Vistas de la arquitectura	✓	
		5.6 Modelos de la arquitectura.	✓	

		5.7.1 Consistencia dentro de una descripción de arquitectura.	✓	
		5.7.2 Correspondencias.		✓
		5.7.3 Reglas de correspondencia.	✓	
		5.8.1 Registro de justificación.	✓	
		5.8.2 Registro de decisión.	✓	
Clausula N.º 6	Marcos de arquitectura y lenguajes de descripción de arquitectura.	6.1 Marcos arquitectura.	✓	
		6.2 Adherencia de una descripción de arquitectura a un marco de arquitectura.		✓
		6.3 Lenguajes de descripción de la arquitectura.	✓	
Clausula N.º 7	Punto de vista de la arquitectura	a. Una o más preocupaciones enmarcadas por este punto de vista (según 5.3).	✓	
		b. Partes interesadas típicas para las preocupaciones enmarcadas por este punto de vista (según 5.3).	✓	
		c. Uno o más tipos de modelos utilizados en este punto de vista.		✓
		d. Para cada tipo de modelo identificado en c), los idiomas, notaciones, convenciones, técnicas de modelado, métodos analíticos y / u otras operaciones que se utilizarán en modelos de este tipo.	✓	
		e. Referencias a sus fuentes.		✓

Fuente: (Skriganov 2011)

Como se puede observar en la **Tabla 28** al realizar la evaluación sobre los principios del Estándar Internacional ISO/IEC/IEEE 42010, se determina que el cumplimiento al aplicar el estándar es de 89 % con 16 cumplidos y 2 no cumplidos.

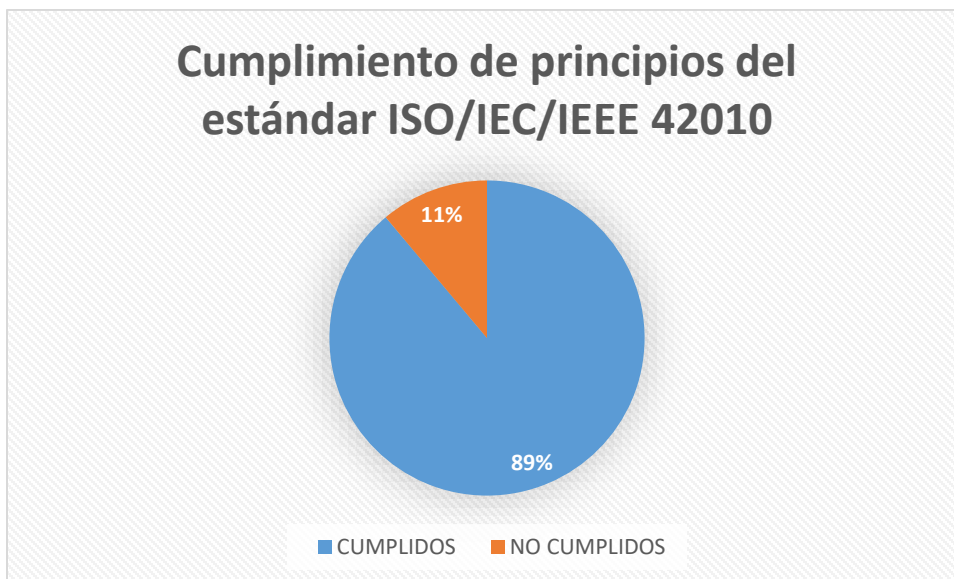


Fig. 36 Gráfico de cumplimiento de principios del estándar ISO/IEC/IEEE 42010
Fuente: Propia

2.16. Fase de implementación de la arquitectura

Siguiendo cada uno de los pasos descritos en el punto **2.15.1.** se procede a la implementación de la solución propuesta, que permita fortalecer el proceso de analítica de la UTN.

En un principio se realizó una revisión de varias fuentes, donde se presenta varias herramientas de analítica web, específicamente análisis de archivos de registro .log, después de varias pruebas realizadas se selecciona 4 herramientas con la mayor facilidad de uso, con las cuales se procede a realizar el análisis y comparativa de funcionalidad de estas, para ella se llevará a cabo la instalación de las distintas herramientas a utilizar.

2.17. Instalación y prueba de herramientas

Luego de haber realizado una búsqueda exhaustiva tanto bibliográfica como en la Web, de las posibles herramientas que permitan fortalecer el proceso de analítica Web mediante el análisis de los archivos .log, generados por el portal de la UTN.

Sé tomo como referencia 8 aplicaciones que realicen este tipo de análisis, de las cuales se selecciona 4 de ellas, que cumplen con las expectativas y generen los resultados de manera clara, de fácil entendimiento, con un orden específico, que facilita al profesional obtener los datos requeridos y la información adecuada.

En la **Fig. 37** se muestran las aplicaciones seleccionadas para realizar las respectivas pruebas.

R Studio	Web log Expert	Screaming Frog SEO Log File Analyser	Semrush/Log File Analyzer
			
RStudio es un entorno de desarrollo integrado para R, Incluye una consola, un editor de resaltado de sintaxis que admite la ejecución directa de código. Se lo utiliza en muchos campos de la informática como por ejemplo análisis de datos.	Web log Expert es una herramienta que permite analizar los archivos de registro de acceso de manera rápida, permite visualizar de información sobre los visitantes al sitio web, proporciona estadísticas de actividad, archivos accedidos, rutas a través del sitio, información sobre páginas de referencia, motores de búsqueda, sistemas operativos entre otros datos	Screaming Frog SEO Log File Analyser, es una potente herramienta que permite cargar sus archivos de registro (log), verificar bots de motores de búsqueda, identificar URL rastreadas y analizar los datos y el comportamiento de los bots de búsqueda para obtener información SEO invaluable.	Log File Analyzer te ayuda a conseguir datos sobre errores y detecta problemas técnicos, estructurales o navegacionales que obstaculizan a los bots de Google.
https://www.rstudio.com/	https://www.weblogexpert.com/	https://www.screamingfrog.co.uk/log-file-analyser/	https://www.semrush.com/log-file-analyzer/

Fig. 37 Herramientas de Análisis de archivos .log
Fuente: Propia

2.17.1. Descripción de las herramientas de análisis de archivos .log

En la **Tabla 29** se muestra las aplicaciones propuestas, que ayudaran en el fortalecimiento del proceso de analítica Web.

Tabla 29 Descripción herramientas análisis archivos .log

ID	Nombre	Descripción
APP01	RStudio	Basado en lenguaje R para análisis de datos
APP02	Screaming frog	Herramienta para análisis de archivos .log
APP03	Web log expert	Herramienta para análisis de archivos .log, genera reporte en PDF
APP04	LogFile analyzer/semrush	Aplicación en la web que permite la visualización y análisis de archivos .log

Fuente: Propia

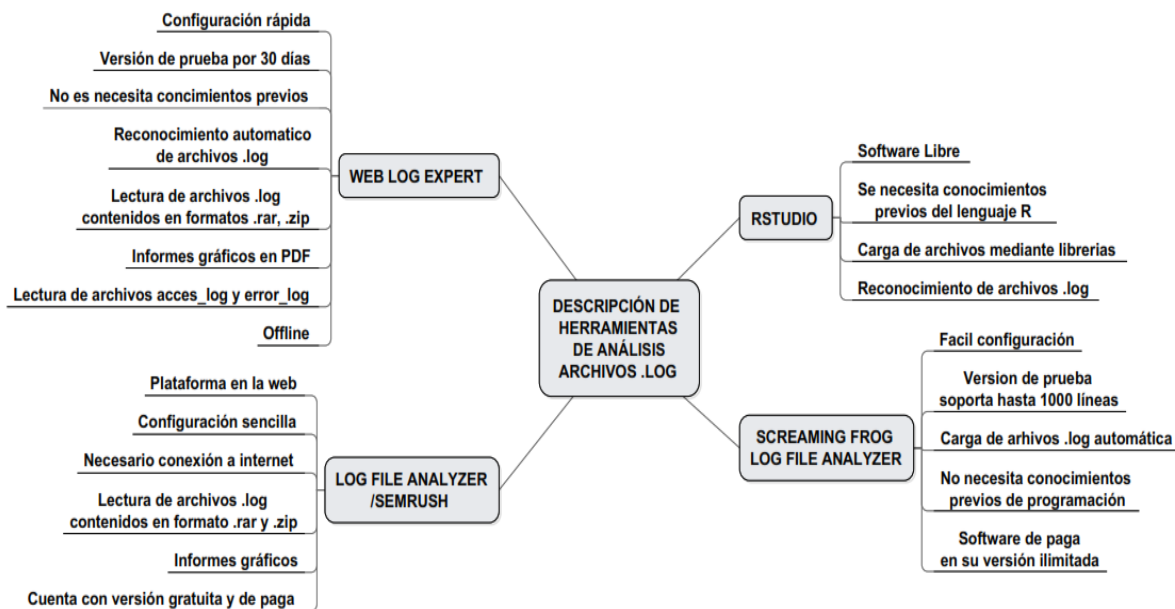


Fig. 38. Descripción herramientas de análisis de archivos .log
Fuente: Propia

2.18. Proceso de analítica Web

Para realizar la minería de datos Web es necesario utilizar el proceso de descubrimiento KDD, en esta sección se realiza el proceso de análisis de archivos .log con las diferentes herramientas anteriormente seleccionadas.

2.18.1. Herramienta RStudio

RStudio es un entorno de desarrollo integrado para R, Incluye una consola, un editor de resaltado de sintaxis que admite la ejecución directa de código(Anon n.d.).

El proceso de instalación es muy sencillo ya que es un entorno gráfico, en primera instancia se debe instalar el lenguaje R, que es un entorno de software libre y lenguaje de programación interpretado(UNIR 2019).

En la Tabla 30, se muestra los requisitos mínimos que el software necesita para su ejecución.

Tabla 30 Requisitos mínimos de instalación RStudio.

Características	Requerimientos mínimos
Sistema Operativo	Multiplataforma (Windows, MacOS, Unix, Linux) Windows 9x/ME/NT4.0/2000/XP/2003/Vista/7/8/2012 Server/8.1/10
Memoria RAM	2 GB
Procesador	2.00 ghz

Disco Duro	8GB mínimo, recomendable 40 GB para almacenar los scripts generados
Estado del Sistema	Se recomienda no tener instalado versiones anteriores

Fuente: Propia adaptada a (Anon n.d.)

- **Proceso de instalación RStudio**

Para realizar el proceso de instalación se deben seguir una serie de pasos secuenciales, los cuales están descritos en el **Anexo, MANUAL USO_RSTUDIO.docxs**.

2.18.1.1. Fase de integración y recopilación de Datos

- **Tipos de datos base**

Para el presente proyecto de titulación se utilizarán los archivos .log, generados en el portal web de la Universidad Técnica del Norte. Los tipos de datos que conforman estos archivos son de tipo, numérico, fecha, cadena de caracteres, además cabe recalcar que existe varios tipos de archivos log, como son: acces_log, error_log, sys_log, de los cuales se va a utilizar acces_log y error_log.

La estructura que nos presenta el archivo .log es la siguiente:

En la **Tabla 32** se detalla las características de cada uno de los componentes de un archivo .log.

Tabla 31. Estructura archivos .log

No.	Detalla	Valor	Significado
1	Dirección IP	139.99.122.199	Dirección IP del solicitante
2	Vacío	-	Por defecto, identidad RFC-1413 no identificada
3	Vacío(¿quién?)	-	Muestra al usuario en caso de que haya tenido lugar una autenticación HTTP, de lo contrario este espacio queda libre
4	Fecha/ Hora (¿Cuándo?)	[19/Mar/2013:14:49:23 - 0500]	Se indica fecha, hora y huso horario en el que se realizó la petición
5	¿Qué?	"GET / HTTP/1.1"	El evento que tuvo lugar, en este caso la solicitud

			de una imagen a través de HTTP
6	Estado de respuesta	404	Confirmación de la solicitud (Código de estado HTTP)
7	¿Cuánto?	209	Muestra la cantidad de datos obtenidos en bytes
8	¿Desde dónde?	"http://www.utn.edu.ec/"	Dirección web desde la que se solicitan los datos
		Google Chrome	Especificaciones técnicas del cliente: navegador,
9	¿Con que?	Firefox	sistema operativo, kernel,
		Safari	interfaz de usuario, idioma, versión

Fuente: Propia

2.18.1.2. Fase de Selección, limpieza y transformación

Una vez comprendido el contenido del archivo .log, se procede a seleccionar, limpiar y transformar los datos, para lo cual se realiza una revisión con las diferentes herramientas a utilizar, RStudio, weblog expert, screaming frog.

- **Selección**

En la etapa de selección y limpieza se inició determinando los atributos que no son necesarios para el estudio a realizar, por ejemplo, la columna X2 y X3 no presentan datos. Inicialmente se tiene 9 columnas nombradas X1 hasta X9 por defecto, una vez realizada la selección se tiene un total de 7 columnas.

Para realizar el proceso de selección y limpieza se procede a cargar los datos en RStudio ejecutando el siguiente comando: ***File<-file.choose()***, previamente a ello es necesario instalar el paquete “readr” de la siguiente manera ***install.packages('readr')***.

Una vez cargado los datos utilizamos la librería “readr”, esto nos genera un dataset como se muestra en la Fig 38, utilizando el siguiente comando `log<-read_log(File,skip=0,col_names = FALSE)`.

	X1	X2	X3	X4	X5	X6	X7	X8
1	110.77.135.152	NA	NA	30/Aug/2020:03:37:02 -0500	GET /fecyt/carreras/diseniopublicidad/?p=109111111111...	404	32854	NA
2	110.77.135.152	NA	NA	30/Aug/2020:03:37:03 -0500	GET /fecyt/carreras/diseniopublicidad/?p=109111111111...	301	NA	NA
3	110.77.135.152	NA	NA	30/Aug/2020:03:37:04 -0500	GET /fecyt/carreras/diseniopublicidad/?p=109111111111...	404	32856	NA
4	110.77.135.152	NA	NA	30/Aug/2020:03:37:06 -0500	GET /fecyt/carreras/diseniopublicidad/?p=109111111111...	301	NA	NA
5	110.77.135.152	NA	NA	30/Aug/2020:03:37:07 -0500	GET /fecyt/carreras/diseniopublicidad/?p=109111111111...	404	32853	NA
6	110.77.135.152	NA	NA	30/Aug/2020:03:37:09 -0500	GET /fecyt/carreras/diseniopublicidad/?p=109111111111...	301	NA	NA
7	110.77.135.152	NA	NA	30/Aug/2020:03:37:10 -0500	GET /fecyt/carreras/diseniopublicidad/?p=109111111111...	404	32854	NA
8	110.77.135.152	NA	NA	30/Aug/2020:03:37:12 -0500	GET /fecyt/carreras/diseniopublicidad/?p=109111111111...	301	NA	NA
9	110.77.135.152	NA	NA	30/Aug/2020:03:37:13 -0500	GET /fecyt/carreras/diseniopublicidad/?p=109111111111...	404	32853	NA
10	110.77.135.152	NA	NA	30/Aug/2020:03:37:14 -0500	GET /fecyt/carreras/diseniopublicidad/?p=109111111111...	301	NA	NA
11	139.99.122.199	NA	NA	30/Aug/2020:03:37:02 -0500	GET /fecyt/carreras/diseniopublicidad/?page_id=1040 H...	200	26278531	http
12	110.77.135.152	NA	NA	30/Aug/2020:03:37:15 -0500	GET /fecyt/carreras/diseniopublicidad/?p=109111111111...	404	32855	NA
13	110.77.135.152	NA	NA	30/Aug/2020:03:37:17 -0500	GET /fecyt/carreras/diseniopublicidad/?p=109111111111...	301	NA	NA
14	110.77.135.152	NA	NA	30/Aug/2020:03:37:18 -0500	GET /fecyt/carreras/diseniopublicidad/?p=109111111111...	404	32854	NA
15	110.77.135.152	NA	NA	30/Aug/2020:03:37:19 -0500	GET /fecyt/carreras/diseniopublicidad/?p=109111111111...	301	NA	NA
16	139.99.8.177	NA	NA	30/Aug/2020:03:37:05 -0500	GET /fecyt/carreras/diseniopublicidad/?page_id=1040 H...	200	26278531	http
17	110.77.135.152	NA	NA	30/Aug/2020:03:37:20 -0500	GET /fecyt/carreras/diseniopublicidad/?p=109111111111...	404	32856	NA
18	110.77.135.152	NA	NA	30/Aug/2020:03:37:22 -0500	GET /fecyt/carreras/diseniopublicidad/?p=109111111111...	301	NA	NA

Fig. 39 Selección de datos a analizar
Fuente: Propia

Una vez realizada la carga y lectura de los archivos log, se procede a renombrar las columnas utilizando el siguiente comando `log<-read_log(File,skip = 0, col_names = c("IP","2","3","Fecha","Petición","Status","Tamaño","Referer","Useragent"))`, en la Fig. 17, se puede observar la información que contiene el archivo acces.log, renombrada para mayor entendimiento.

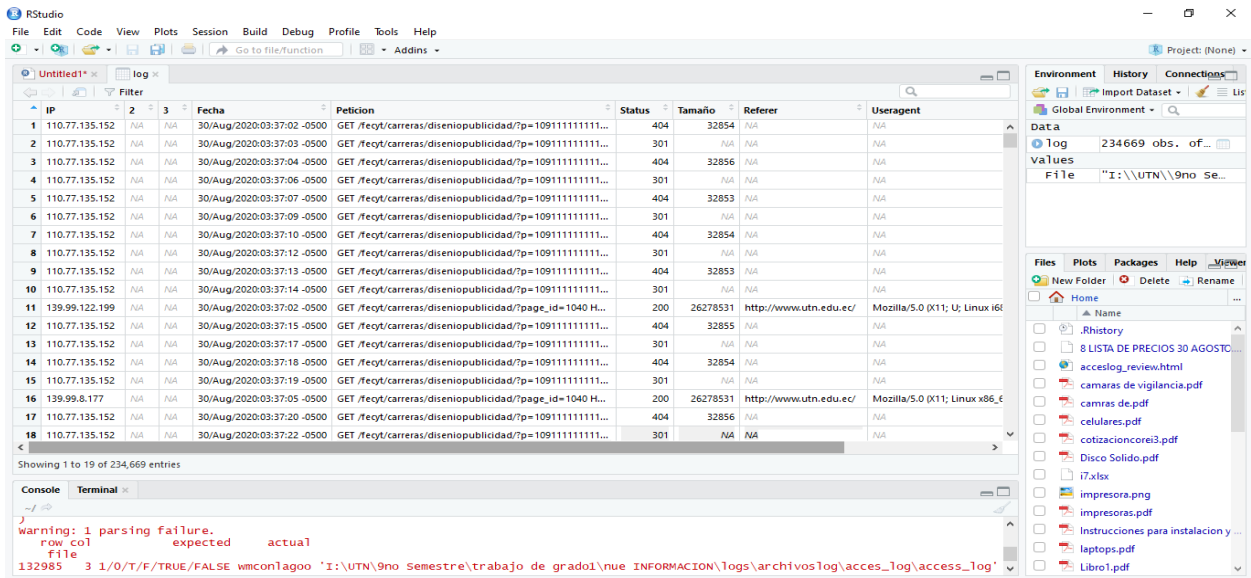


Fig. 40 Identificación de columnas a utilizar
Fuente: Propia

- Limpieza

Una vez seleccionado con los parámetros que se va a realizar el análisis, realizamos el proceso de limpieza, eliminando las columnas que se muestra con la denotación 2 y 3, ya que no contiene datos, para ello ejecutamos la siguiente línea de código: `log[,c("IP", "Fecha", "Peticion", "Status", "Referer", "Useragent")]`, nos queda como se muestra en la Fig. 41.

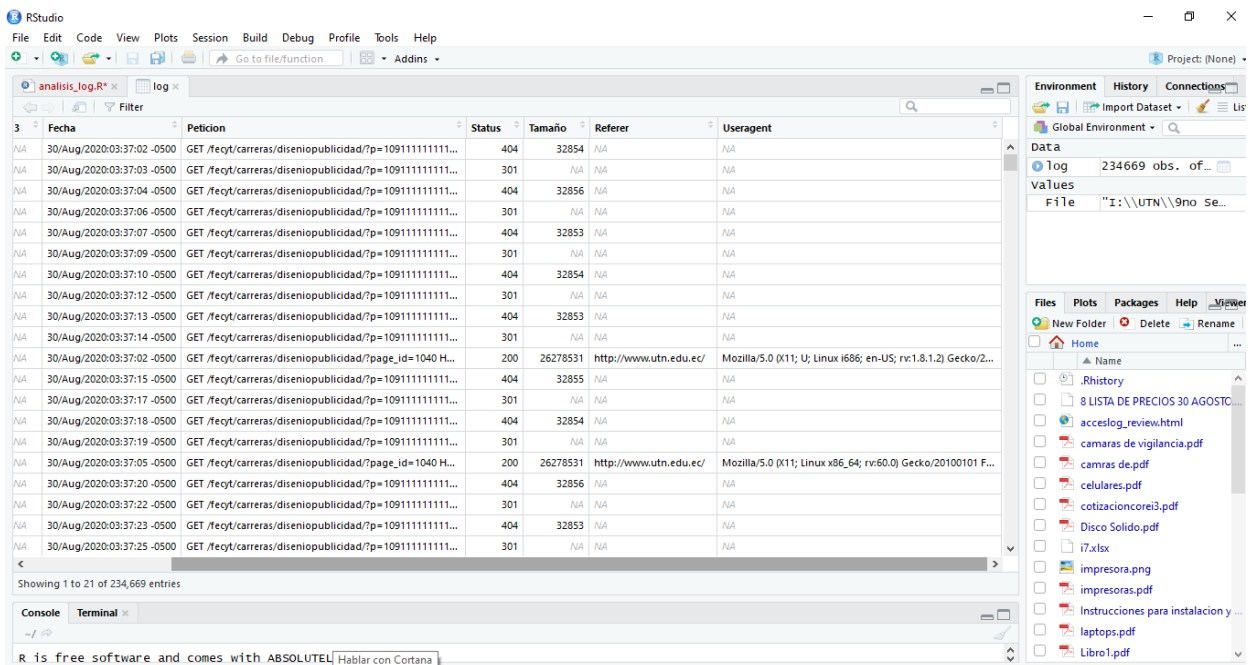


Fig. 41 Limpieza de columnas
Fuente: Propia

2.18.1.3. Fase minería de datos

En esta fase se procede a la realización y análisis de los datos generados, en el caso de RStudio solo nos quedaría filtrar los rastreos generados por Googlebot para conocer cuáles son las peticiones que hace Google al servidor, para ello solo tendremos que indicar que texto queremos filtrar y en qué columna de datos, ejecutando el siguiente con el código: `log.googlebot<-log%>%filter(grepl("googlebot",Useragent))`.

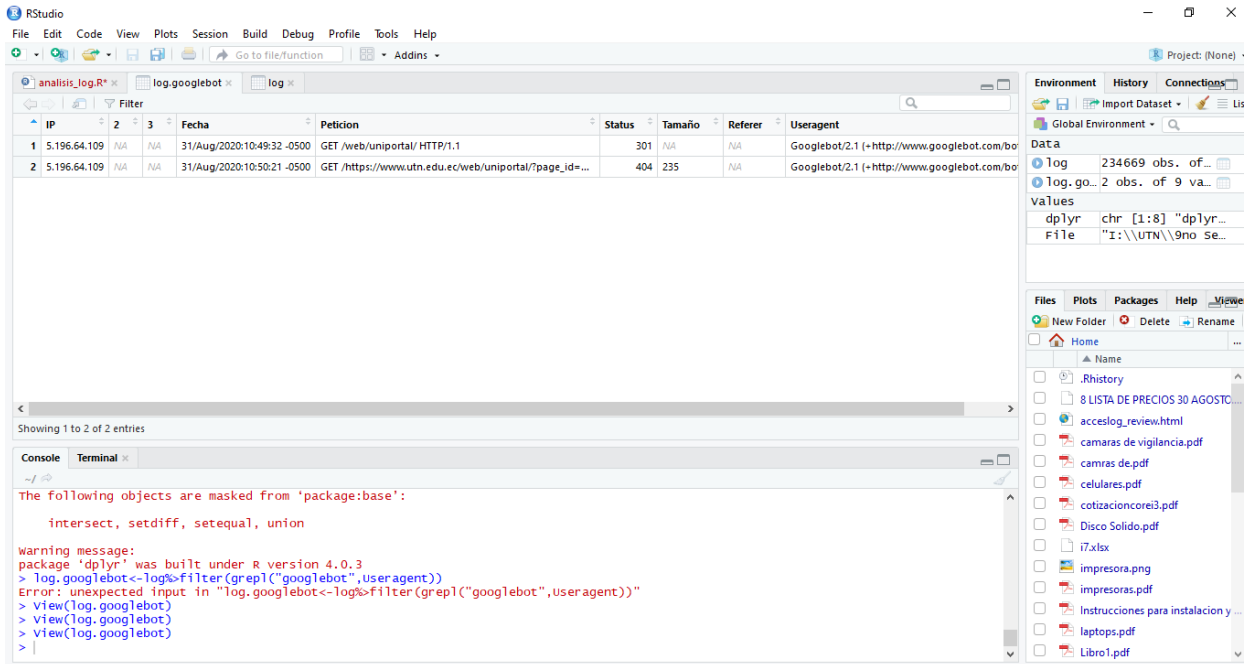


Fig. 42 Filtrado de atributos
Fuente: Propia

El lenguaje R, es un lenguaje muy eficaz para el manejo de archivos con distintos formatos, en la minería de datos, en los últimos años ha tenido una gran acogida ya que cuenta con una gran comunidad.

2.18.2. Herramienta WebLog Expert

Web log Expert es una herramienta que permite analizar los archivos de registro de acceso de manera rápida, además permite visualizar de información sobre los visitantes al sitio web, proporciona estadísticas de actividad, archivos accedidos, rutas a través del sitio, información sobre páginas de referencia, motores de búsqueda, sistemas operativos entre otros datos (Alentum Software 2019).

En la **Tabla 32** se muestra los requisitos mínimos que se necesita para la instalación del software Weblog Expert.

Tabla 32 Requisitos mínimos de instalación Weblog expert

Características	Requisitos mínimos
Sistema Operativo	Plataforma (Windows) Windows 9x/ME/NT4.0/2000/XP/2003/Vista/7/8/2012 Server/8.1/10
Memoria RAM	2 GB
Procesador	2.00 ghz
Disco Duro	50 GB
Estado del Sistema	Se recomienda no tener instalado versiones anteriores

Fuente: (Anon n.d.)

- **Proceso de instalación WebLog Expert**

Para realizar el proceso de instalación se deben seguir una serie de pasos secuenciales, los cuales están descritos en el **Anexo, MANUAL USO WEBLOGEXPERT**.

Una vez preparados los archivos que se van a analizar, se realiza la carga de archivos como se muestra a continuación:

- **Análisis de archivos .log mediante WebLog Expert**

Paso 1.- Clic en File > New Profile

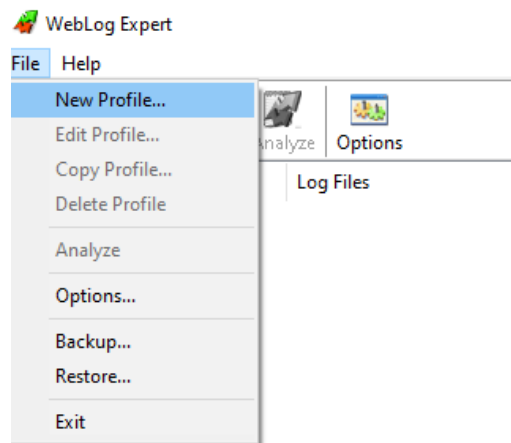


Fig. 43 Crear nuevo perfil
Fuente: Propia

Paso 2. - Asignar un nombre descriptivo al perfil, agregamos el dominio y el index.html, presionar en siguiente.

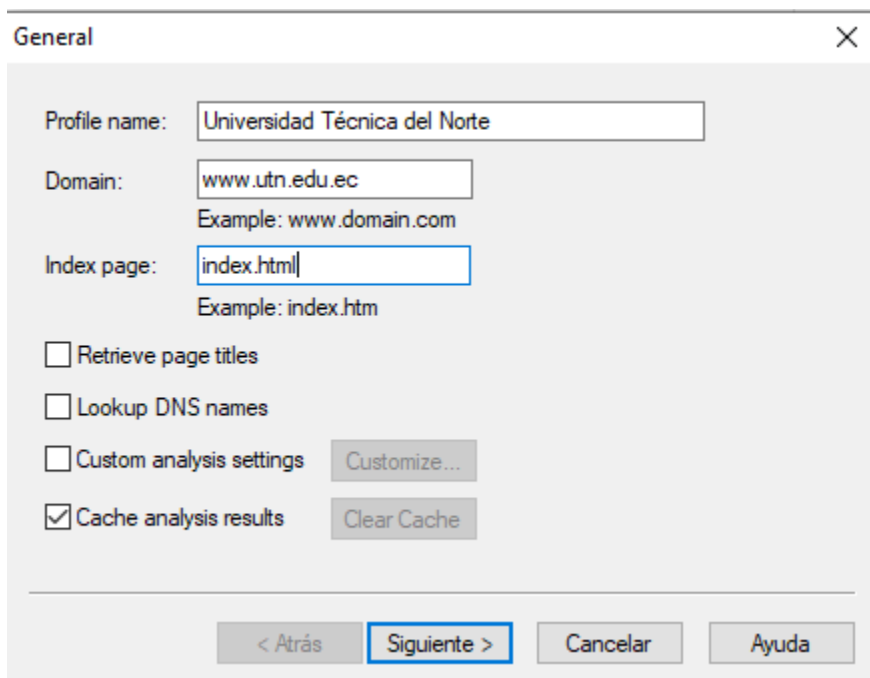


Fig. 44 Ingreso de perfil y dominio
Fuente: Propia

Paso 3. - Agregar el archivo .log(acces_log/error_log), dar clic en path-browse

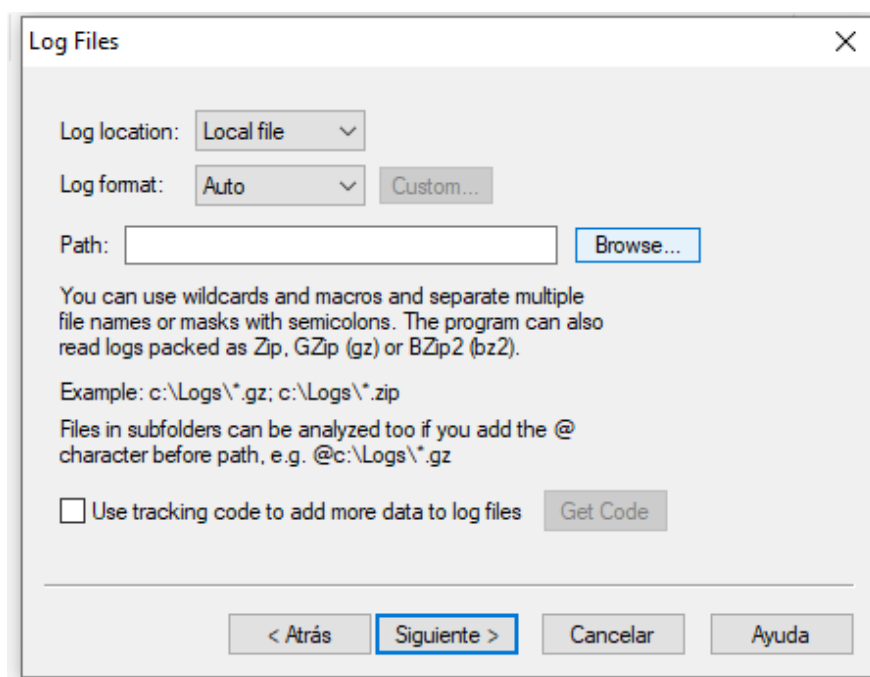


Fig. 45 Carga de archivos .log
Fuente: Propia

Paso 4. - Seleccionar archivo **acces_log**, dar clic en abrir, clic en siguiente

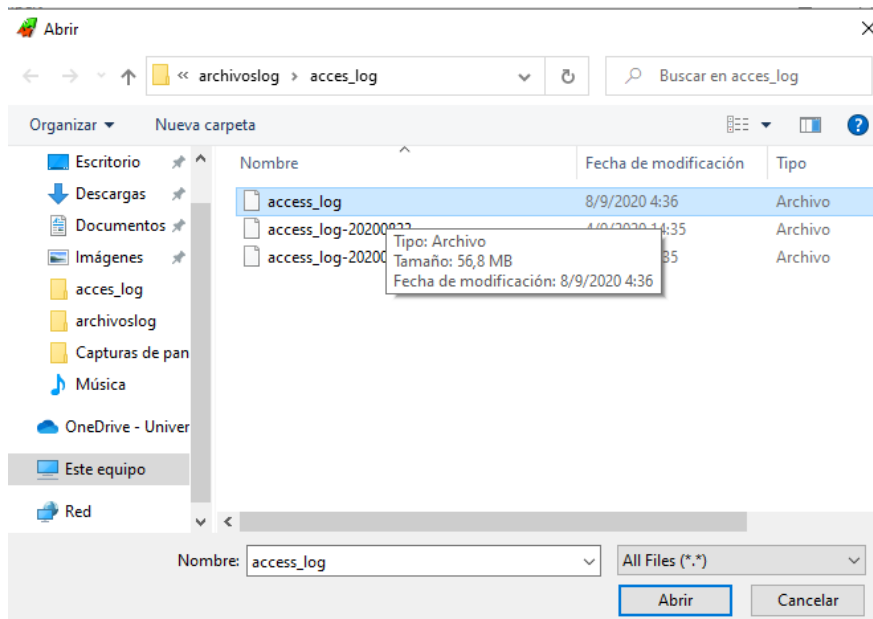


Fig. 46 Selección de archivos .log
Fuente: Propia

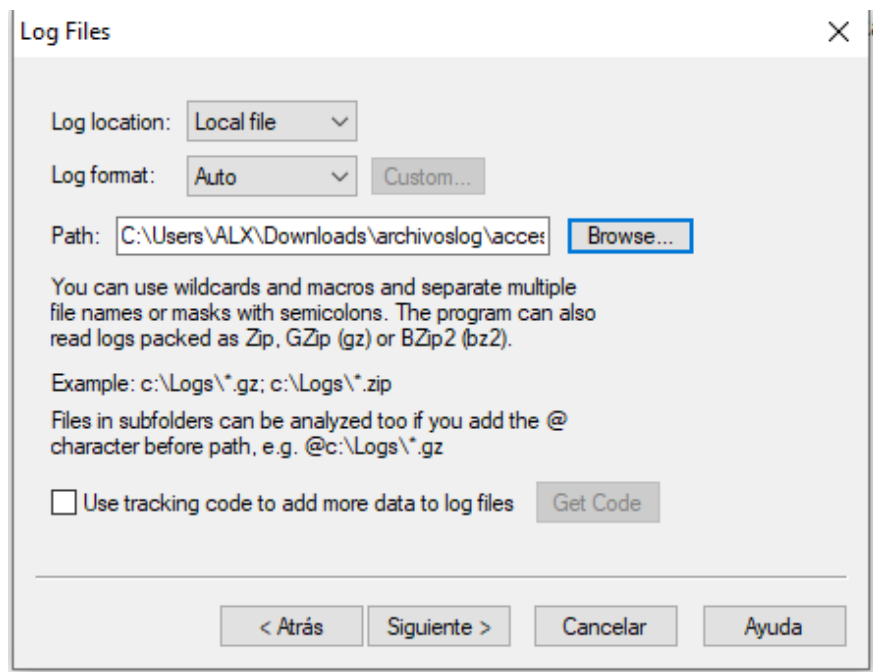


Fig. 47 Carga archivos .log
Fuente: Propia

Paso 5. - Seleccionar la actividad, en este caso, todas las actividades en un rango de tiempo especificado, dar clic en siguiente

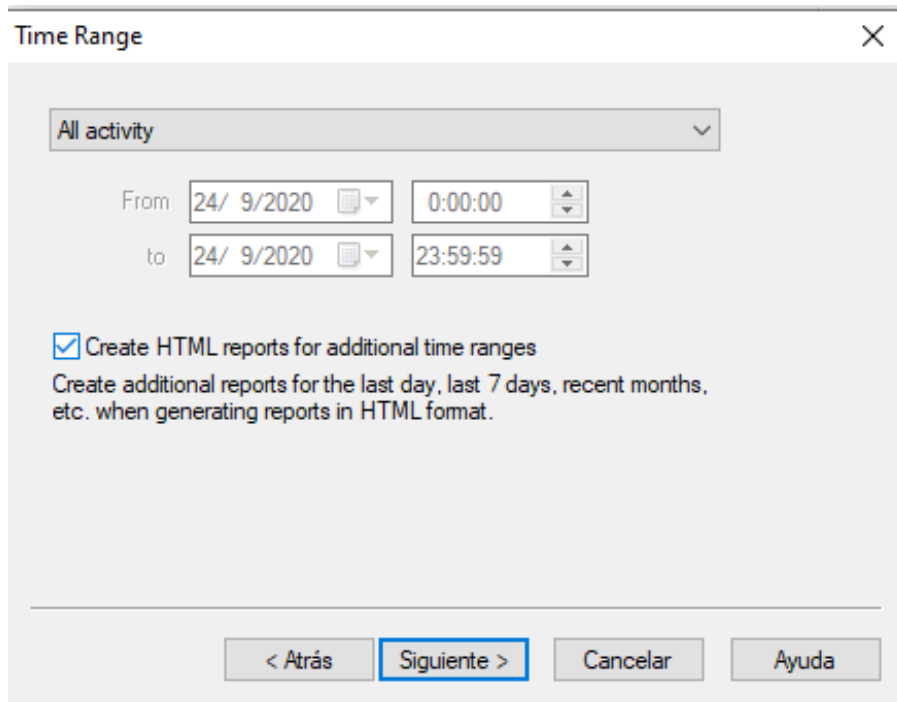


Fig. 48 Selección de rango de hora y fecha
Fuente: Propia

Nota: puede agregar filtros, en este caso no, se deja por defecto

Paso 6. - Generar reporte, seleccionar el tipo de archivo de salida, html, pdf, csv.

Paso 7. - Clic en analyze

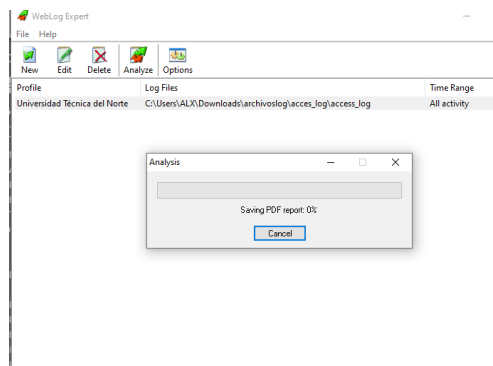


Fig. 49 Análisis de datos
Fuente: Propia

Generar reporte

El reporte completo generado ver el archivo **report for Universidad Técnica de Norte.**

Report for Universidad Técnica del Norte

Time range: 30/8/2020 03:37:02 - 1/9/2020 13:05:12

Generated on Thu Sep 24, 2020 - 12:27:22

General Statistics

Summary

Summary

Hits	
Total Hits	234,669
Visitor Hits	226,342
Spider Hits	8,327
Average Hits per Day	78,223
Average Hits per Visitor	9.20
Cached Requests	1,009
Failed Requests	13,012
Page Views	
Total Page Views	115,969
Average Page Views per Day	38,656
Average Page Views per Visitor	4.72
Visitors	
Total Visitors	24,592
Average Visitors per Day	8,197
Total Unique IPs	11,289
Bandwidth	
Total Bandwidth	176.68 GB
Visitor Bandwidth	174.67 GB
Spider Bandwidth	2.01 GB
Average Bandwidth per Day	58.89 GB
Average Bandwidth per Hit	789.46 KB
Average Bandwidth per Visitor	7.27 MB

Fig. 50 Sumario de análisis
Fuente: Propia

En un breve resumen de los resultados del análisis se puede visualizar datos de suma importancia como es la cantidad de visitas que el portal recibe, que se clasifican por total, por mes y por día incluso el total de páginas vistas, el total de Ancho de banda ocupado (bandwidth) esto se puede evidenciar en el **Anexo 3: Reporte**, los códigos de error y de respuesta que se generaron al momento de realizar el requerimiento.

En la Tabla 34, se enlista los códigos de respuesta http más comunes encontrados al realizar una solicitud específica.

Tabla 33 Códigos de estados de respuesta http encontrados

Código	Significado
200	(OK) La solicitud realizada ha tenido éxito
301	(Multiple choice) Esta solicitud tiene más de una posible respuesta. User-Agent o el usuario debe escoger uno de ellos. No hay forma estandarizada de seleccionar una de las respuestas.

302	(Found) Este código de respuesta significa que el recurso de la URI solicitada ha sido cambiado temporalmente. Nuevos cambios en la URI serán agregados en el futuro. Por lo tanto, la misma URI debe ser usada por el cliente en futuras solicitudes.
304	(Not Modified) Esta es usada para propósitos de "caché". Le indica al cliente que la respuesta no ha sido modificada. Entonces, el cliente puede continuar usando la misma versión almacenada en su caché.
403	(Forbidden) El cliente no posee los permisos necesarios para cierto contenido, por lo que el servidor está rechazando otorgar una respuesta apropiada.
404	(Not Found) El servidor no pudo encontrar el contenido solicitado.
405	(Method not Allowed) El código de estado de respuesta indica que el servidor conoce el método de solicitud, pero el recurso de destino no lo admite
408	(Request Timeout) El código de estado de respuesta significa que el servidor desea cerrar esta conexión no utilizada. Algunos servidores lo envían en una conexión inactiva, incluso sin ninguna solicitud previa por parte del cliente
500	(internal Server error) El servidor ha encontrado una situación que no sabe cómo manejarla.

Fuente: Propia

2.18.3. Herramienta Screaming Frog SEO Log File Analyser

Screaming Frog SEO Log File Analyser, es una potente herramienta que permite cargar sus archivos de registro (log), verificar bots de motores de búsqueda, identificar URL rastreadas y analizar los datos y el comportamiento de los bots de búsqueda para obtener información SEO invaluable (Anon n.d.).

En la Tabla 35 se muestra los requisitos mínimos requeridos para la instalación de la herramienta screaming frog.

Tabla 34 Requisitos de instalación Screaming Frog Log Analyzer

Características	Requisitos mínimos
Sistema Operativo	Multiplataforma (Windows, MacOS, Unix, Linux) Windows 9x/ME/NT4.0/2000/XP/2003/Vista/7/8/2012 Server/8.1/10
Memoria RAM	1 GB
Procesador	2.00 Ghz
Disco Duro	1 GB

Fuente: Propia adaptada a (Anon n.d.)

- **Proceso de instalación Screaming Frog Log Analyzer**

Para realizar el proceso de instalación se deben seguir una serie de pasos secuenciales, los cuales están descritos en el **Anexo MANUAL USOSCREAMINGFRGOG.docxs**.

- **Log Analyzer de Screaming Frog**

Una vez iniciado Screaming Frog Log Analyser se procede a realizar el respectivo análisis como se muestra a continuación:

Paso 1. - Clic en Project > Import log file

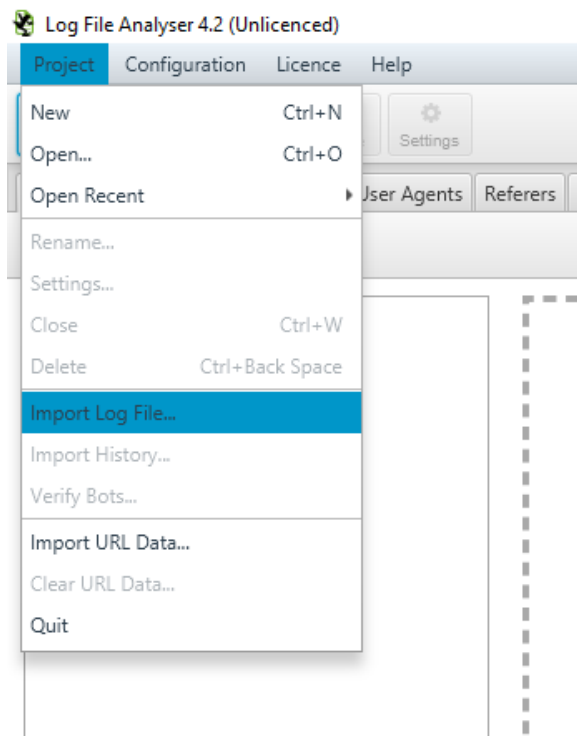


Fig. 51 Importar archivos log
Fuente: Propia

Paso 2. - Seleccionar archivo acces_log/ error_log

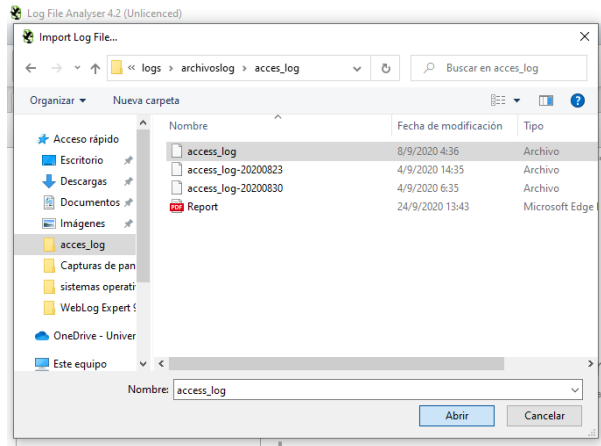


Fig. 52 Selección archivos log
Fuente: Propia

Paso 3. - Asignar nombre al proyecto / seleccionar zona horaria, clic en aceptar

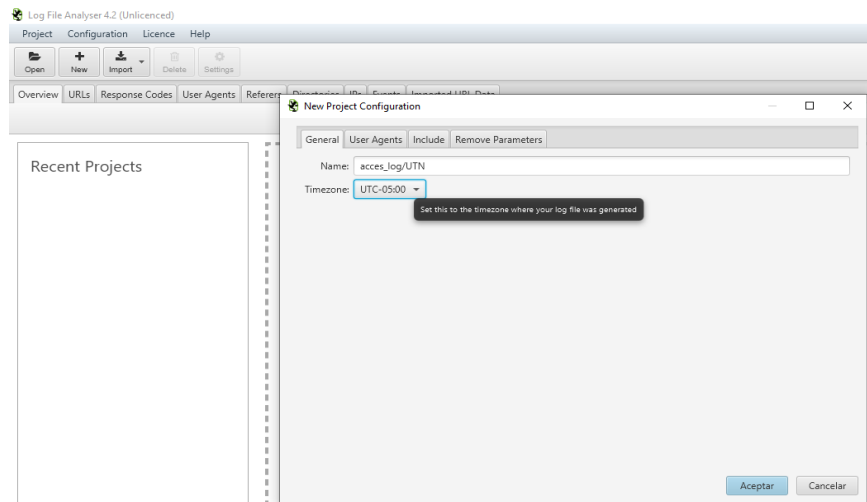


Fig. 53 Asignación de nombre y selección de uso horario
Fuente: Propia

Paso 4. - Escribir URL del sitio del que provienen los archivos .log, incluyendo protocolo http (<http://utn.edu.ec>), aceptar.

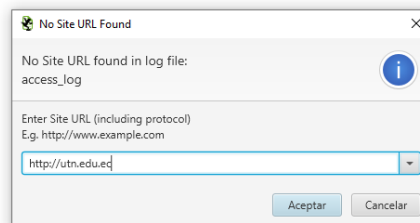


Fig. 54 URL -UTN
Fuente: Propia

Paso 5. - Una vez importado los archivos .log, automáticamente genera reportes correspondientes al contenido de estos.

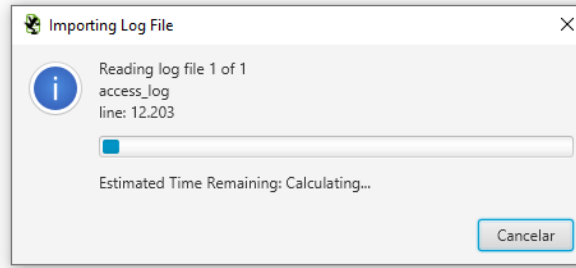


Fig. 55 Importar archivos log
Fuente: Propia

Paso 6. - Presentación de Resultados

- **Overview.** – Muestra una visión general de todos los datos que ofrece el registro de archivos logs.

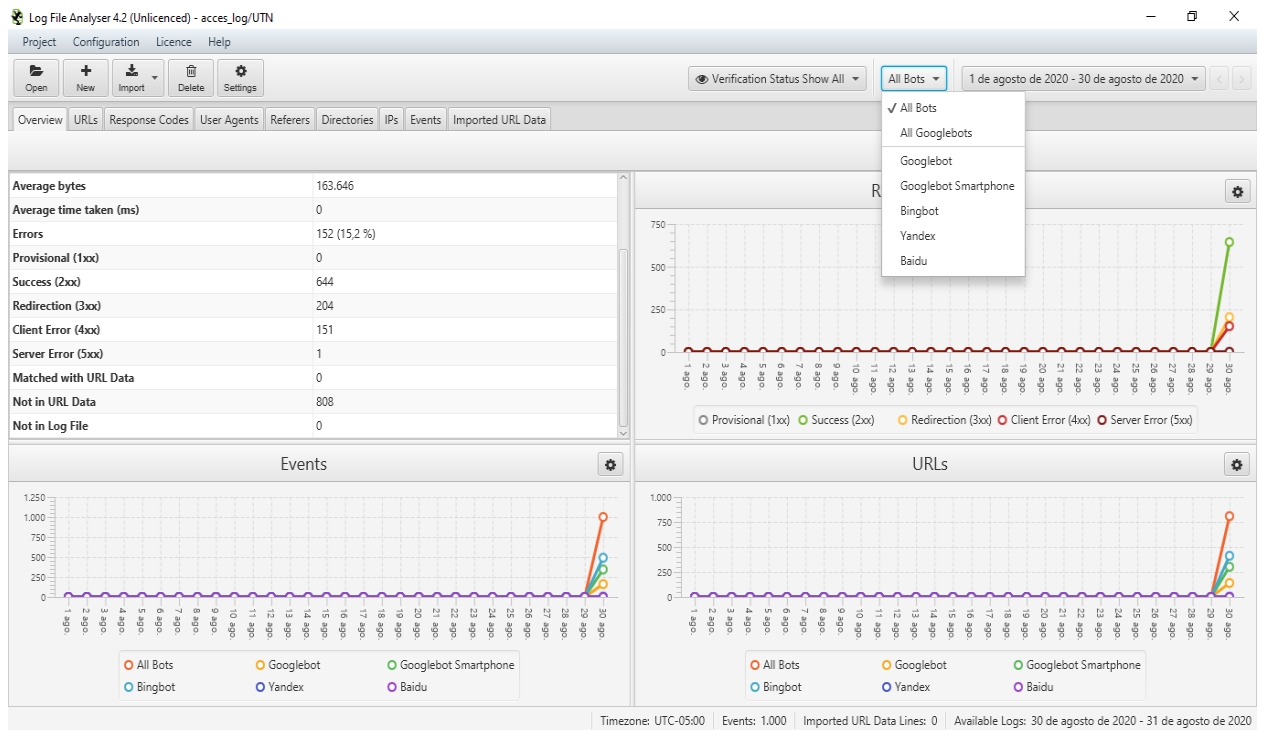


Fig. 56 Visión general
Fuente: Propia

- **URLs.** – Muestra las URLs con las cuales el usuario ha tenido interacción.

Row	URL	Last Response Code	Time Of Last Response	Content Type	Average Bytes	Average
1	http://utn.edu.ec/robots.txt	404	30 ago. 2020 19:13:40	text/plain	208	0
2	http://utn.edu.ec/web/uniportal	301	30 ago. 2020 19:18:12	text/html	244	0
3	http://utn.edu.ec/web/uniportal/	301	30 ago. 2020 19:18:36	text/html	0	0
4	http://utn.edu.ec/	302	30 ago. 2020 18:28:16	text/html	0	0
5	http://utn.edu.ec/bibliotecavirtual/	200	30 ago. 2020 19:51:22	text/html	81.248	0
6	http://utn.edu.ec/deportes/	200	30 ago. 2020 19:51:06	text/html	34.625	0
7	http://utn.edu.ec/ficaya/carreras/agroindustrias/?page_id=849	200	30 ago. 2020 19:33:06	text/html	27.675	0
8	http://utn.edu.ec/web/uniportal/?cat=3&paged=82	301	30 ago. 2020 14:29:44	text/html	0	0
9	http://utn.edu.ec/bibliotecavirtual/index.php/slider-page/slider-catalogo/	200	30 ago. 2020 10:29:37	text/html	55.565	0
10	http://utn.edu.ec/conduccion/index.php/organigrama/	200	30 ago. 2020 14:21:40	text/html	18.365	0
11	http://utn.edu.ec/cultura	301	30 ago. 2020 19:51:04	text/html	238	0
12	http://utn.edu.ec/cultura/	301	30 ago. 2020 19:51:05	text/html	0	0
13	http://utn.edu.ec/deportes	301	30 ago. 2020 19:51:04	text/html	239	0
14	http://utn.edu.ec/eduroam/	301	30 ago. 2020 19:51:21	text/html	0	0

Filter Total: 808

Export

Name	Value
No URL selected	

URL Info Events Referers

Timezone: UTC-05:00 | Events: 1.000 | Imported URL Data Lines: 0 | Available Logs: 30 de agosto de 2020 - 31 de agosto de 2020

Fig. 57 Información de URLs encontradas
Fuente: Propia

- **Response Codes.** – Muestra los códigos de respuesta que genera la URL con la cual el usuario ha tenido interacción.

Row	URL	Last Response Code	Time Of Last Response	Num Events	1xx	2xx	3xx
1	http://utn.edu.ec/robots.txt	404	30 ago. 2020 19:13:40	74	0	0	0
2	http://utn.edu.ec/web/uniportal	301	30 ago. 2020 19:18:12	8	0	0	8
3	http://utn.edu.ec/web/uniportal/	301	30 ago. 2020 19:18:36	7	0	0	7
4	http://utn.edu.ec/	302	30 ago. 2020 18:28:16	6	0	0	6
5	http://utn.edu.ec/bibliotecavirtual/	200	30 ago. 2020 19:51:22	4	0	4	0
6	http://utn.edu.ec/deportes/	200	30 ago. 2020 19:51:06	4	0	4	0
7	http://utn.edu.ec/ficaya/carreras/agroindustrias/?page_id=849	200	30 ago. 2020 19:33:06	4	0	4	0
8	http://utn.edu.ec/web/uniportal/?cat=3&paged=82	301	30 ago. 2020 14:29:44	4	0	0	4
9	http://utn.edu.ec/bibliotecavirtual/index.php/slider-page/slider-catalogo/	200	30 ago. 2020 10:29:37	3	0	3	0
10	http://utn.edu.ec/conduccion/index.php/organigrama/	200	30 ago. 2020 14:21:40	3	0	3	0

Filter Total: 808

Export Search UA & Referers

Row	Timestamp	Remote Host	Method	Response Code	Bytes	Time Taken(ms)	User Agent
No URL selected							

Total Events: 0

URL Info Events Referers

Timezone: UTC-05:00 | Events: 1.000 | Imported URL Data Lines: 0 | Available Logs: 30 de agosto de 2020 - 31 de agosto de 2020

Fig. 58 Códigos de respuesta
Fuente: Propia

- **User Agents.** – Muestra los agentes de usuario que utiliza el cliente para realizar la consulta requerida al portal web.

Row	User Agent	Unique URLs	Num Events	Average Bytes	Average Response Time (ms)	Errors
1	Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)	367	440	235.344	0	133
2	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/80.0.398...	268	309	32.539	0	12
3	Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)	130	151	304.999	0	5
4	Mozilla/5.0 (iPhone; CPU iPhone OS 7_0 like Mac OS X) AppleWebKit/537.51.1 (KHTML, like Gecko) Version/7.0 Mobile...	51	51	52.909	0	0
5	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/84.0.414...	26	26	29.817	0	0
6	Mozilla/5.0 AppleWebKit/537.36 (KHTML, like Gecko; compatible; Googlebot/2.1; +http://www.google.com/bot.html) ...	10	10	24.766	0	0
7	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/84.0.414...	8	8	17.084	0	0
8	Mozilla/5.0 (compatible; YandexBot/3.0; +http://yandex.com/bots)	3	4	28.545	0	2
9	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.227...	1	1	13.608	0	0

Fig. 59 Agentes de Usuario
Fuente: Propia

- **Referers.** – Esta vista muestra la página a la cual se realizó la consulta.

Row	Referer	Unique URLs	Num Events	Average Bytes	Average Response Time (ms)	Errors	1xx	2xx	3xx
1	-	748	927	174.369	0	152	0	576	199
2	http://www.utn.edu.ec/facae/carreras/mercadotecnia/?page_id=9	7	7	32.797	0	0	0	7	0
3	http://www.utn.edu.ec/fecyt/carreras/artesplasticas/	7	7	32.797	0	0	0	7	0
4	http://www.utn.edu.ec/fica/carreras/electronica/?paged=16	6	6	40.211	0	0	0	6	0
5	http://www.utn.edu.ec/fccss/carreras/enfermeria/?page_id=705	3	3	42.166	0	0	0	3	0
6	http://www.utn.edu.ec/colegio/?page_id=962	2	2	8.582	0	0	0	2	0
7	http://www.utn.edu.ec/cultura	1	2	0	0	0	0	0	2
8	http://www.utn.edu.ec/deportes	1	2	34.625	0	0	0	2	0
9	http://www.utn.edu.ec/fccss/carreras/enfermeria/?p=1220	2	2	40.015	0	0	0	2	0
10	http://www.utn.edu.ec/fecyt/carreras/educaciongeneralbasica/?page_id...	2	2	7.422	0	0	0	2	0
11	http://www.utn.edu.ec/ficaya/carreras/forestal/?page_id=827	2	2	5.567	0	0	0	2	0
12	http://www.utn.edu.ec/web/uniportal	1	2	0	0	0	0	0	2
13	http://www.utn.edu.ec/bibliotecavirtual	1	1	81.249	0	0	0	1	0
14	http://www.utn.edu.ec/bibliotecavirtual/wp-admin/	1	1	6.436	0	0	0	1	0
15	http://www.utn.edu.ec/colegio/?page_id=vaozlgjd%3Fpaged%3D5%3Fp...	1	1	6.131	0	0	0	1	0
16	http://www.utn.edu.ec/colegio/?page_id=vaozlgjd%3Fpaged%3D7%3Fp...	1	1	8.695	0	0	0	1	0

Fig. 60 Referencias
Fuente: Propia

- **Directories.** – Muestra los directorios hacia los cuales va dirigida la consulta o petición de información.

Path	Num Events	Average Bytes	Average Response Time (ms)	All Bots	Googlebot	Googlebot Smartphone	Bingbot	Yandex	Baidu
http/	1000	163647	0	1000	161	344	491	4	0
utn.edu.ec/	1000	163647	0	1000	161	344	491	4	0
colegio/	224	34957	0	224	5	162	57	0	0
fecyt/	143	52444	0	143	22	27	94	0	0
web/	100	119784	0	100	22	34	44	0	0
fica/	73	95201	0	73	16	22	34	1	0
bibliotecavirtual/	43	56822	0	43	39	1	3	0	0
ficaya/	40	56177	0	40	14	9	17	0	0
facea/	38	52741	0	38	10	12	16	0	0
ficayaemprende/	37	142287	0	37	1	12	24	0	0
reduca/	35	30032	0	35	0	11	24	0	0
fccss/	31	59788	0	31	6	12	13	0	0
oficinaestudiante/	16	28947	0	16	1	1	14	0	0
legislacion/	15	478690	0	15	0	2	13	0	0
internacional/	15	92653	0	15	1	2	11	1	0
postgrado/	10	222684	0	10	0	0	10	0	0
transparencia/	10	10286173	0	10	2	5	3	0	0
biblioteca/	10	243	0	10	1	5	4	0	0
cultura/	7	0	0	7	3	2	2	0	0
lecturasdificiles/	5	242	0	5	0	0	5	0	0

Fig. 61 Directorios de acceso
Fuente: Propia

- **IPs.** – Muestra las IPs de los Host remoto.

Row	Remote Host	Unique URLs	Num Events	Average Bytes	Average Response Time (ms)	Errors	1xx	2xx	3xx	4xx	5xx
1	66.249.70.30	151	168	264.465	0	6	0	114	48	6	0
2	66.249.70.2	119	125	25.971	0	3	0	102	20	3	0
3	35.240.117.238	98	103	41.800	0	0	0	69	34	0	0
4	157.55.39.100	34	101	15.886	0	71	0	24	6	71	0
5	66.249.70.4	85	92	55.968	0	1	0	68	23	1	0
6	157.55.39.235	46	49	137.313	0	10	0	33	6	9	1
7	207.46.13.119	44	47	65.001	0	3	0	38	6	3	0
8	207.46.13.232	46	47	86.268	0	2	0	41	4	2	0
9	40.77.167.171	46	46	48.780	0	9	0	33	4	9	0
10	207.46.13.108	31	32	171.614	0	1	0	23	8	1	0
11	157.55.39.135	23	23	82.977	0	6	0	15	2	6	0
12	40.77.167.58	18	18	3.687.713	0	0	0	13	5	0	0
13	157.55.39.250	16	16	197.528	0	4	0	7	5	4	0
14	35.187.23.223	10	11	13.658	0	1	0	4	6	1	0
15	40.77.167.75	11	11	11.487	0	2	0	7	2	2	0
16	207.46.13.117	10	10	20.715	0	1	0	8	1	1	0
17	40.77.167.212	10	10	510.579	0	1	0	5	4	1	0
18	40.77.167.123	8	8	47.250	0	1	0	5	2	1	0
19	40.77.167.186	7	7	29.860	0	3	0	4	0	3	0

Fig. 62 IPs Host remoto
Fuente: Propia

- **Events.-** Muestra los eventos generados en cada petición.

Row	URL	Timestamp	Remote Host	Method	Response Code	Bytes	Time
1	http://utn.edu.ec/bibliotecavirtual/	30 ago. 2020 19:51:22	66.249.70.30	GET	200	81.248	0
2	http://utn.edu.ec/eduroam/	30 ago. 2020 19:51:21	66.249.70.30	GET	301	0	0
3	http://utn.edu.ec/bibliotecavirtual	30 ago. 2020 19:51:19	66.249.70.30	GET	301	248	0
4	http://utn.edu.ec/oficinaestudiante/	30 ago. 2020 19:51:18	66.249.70.30	GET	200	38.845	0
5	http://utn.edu.ec/transparencia/	30 ago. 2020 19:51:17	66.249.70.30	GET	301	0	0
6	http://utn.edu.ec/internacional/	30 ago. 2020 19:51:16	66.249.70.30	GET	200	41.222	0
7	http://utn.edu.ec/oficinaestudiante	30 ago. 2020 19:51:15	66.249.70.30	GET	301	248	0
8	http://utn.edu.ec/transparencia	30 ago. 2020 19:51:14	66.249.70.30	GET	301	244	0
9	http://utn.edu.ec/internacional	30 ago. 2020 19:51:13	66.249.70.30	GET	301	244	0
10	http://utn.edu.ec/legislacion/	30 ago. 2020 19:51:12	66.249.70.30	GET	200	20.927	0
11	http://utn.edu.ec/legislacion	30 ago. 2020 19:51:10	66.249.70.30	GET	301	242	0
12	http://utn.edu.ec/deportes/	30 ago. 2020 19:51:06	66.249.70.30	GET	200	34.625	0
13	http://utn.edu.ec/cultura/	30 ago. 2020 19:51:05	66.249.70.4	GET	301	0	0
14	http://utn.edu.ec/deportes	30 ago. 2020 19:51:04	66.249.70.30	GET	301	239	0
15	http://utn.edu.ec/cultura	30 ago. 2020 19:51:04	66.249.70.4	GET	301	238	0
16	http://utn.edu.ec/bibliotecavirtual/index.php/tablero-informativo/recursos-fisicos/	30 ago. 2020 19:47:18	207.46.13.119	GET	200	54.749	0
17	http://utn.edu.ec/colegio/?cat=10&paged=3	30 ago. 2020 19:47:13	66.249.70.4	GET	200	41.886	0

Fig. 63 Eventos
Fuente: Propia

Method	Response Code	Bytes	Time Taken(ms)	User Agent	Referer	Verificatio...
GET	200	81.248	0	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit...	-	Verification ...
GET	301	0	0	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit...	-	Verification ...
GET	301	248	0	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit...	-	Verification ...
GET	200	38.845	0	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit...	-	Verification ...
GET	301	0	0	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit...	-	Verification ...
GET	200	41.222	0	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit...	-	Verification ...
GET	301	248	0	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit...	-	Verification ...
GET	301	244	0	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit...	-	Verification ...
GET	301	244	0	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit...	-	Verification ...
GET	200	20.927	0	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit...	-	Verification ...
GET	301	242	0	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit...	-	Verification ...
GET	200	34.625	0	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit...	-	Verification ...
GET	301	0	0	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit...	-	Verification ...
GET	301	239	0	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit...	-	Verification ...
GET	301	238	0	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit...	-	Verification ...
GET	200	54.749	0	Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)	-	Verification ...
GET	200	41.886	0	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit...	-	Verification ...

Fig. 64 Eventos 2
Fuente: Propia

- **Import URLs Data.** – En este punto, se pueden importar archivos CSV o Excel con cualquier dato que contenga un encabezado y una columna de

URL. Por ejemplo, se puede importar un sitemap del sitio o exportar “páginas principales” obtenidas mediante otros programas.

The screenshot shows the Log File Analyzer 4.2 (Unlicensed) interface. The main window displays a table of imported URL data. The table has columns for Row, URL, Level, Inlinks, and Outlinks. The data is as follows:

Row	URL	Level	Inlinks	Outlinks
1	http://example.com/	0	42	56
2	http://example.com/1xx.html	1	6	162
3	http://example.com/alwaysredirects.html	3	5	143
4	http://example.com/cute_cat.gif	1	6	157
5	http://example.com/download/press.pdf	1	5	91
6	http://example.com/image.jpg	1	6	41
7	http://example.com/inconsistent.html	1	10	140
8	http://example.com/intro.swf	2	11	139
9	http://example.com/js/bigfoot.js	2	1	1
10	http://example.com/jquery.min.js	1	7	143
11	http://example.com/nav_left.png	1	30	219
12	http://example.com/nav_right.png	1	7	25
13	http://example.com/not_matched1.html	1	16	158
14	http://example.com/not_matched2.html	2	17	159
15	http://example.com/not_matched3.html	3	18	160
16	http://example.com/returns404.html	1	16	158
17	http://example.com/robots.txt	1	15	70
18	http://example.com/server_500.html	1	6	146
19	http://example.com/server_503.html	1	1	1

At the bottom of the interface, there is a status bar showing: Timezone: UTC-05:00 | Events: 1.000 | Imported URL Data Lines: 21 | Available Logs: 30 de agosto de 2020 - 31 de agosto de 2020. A filter total of 21 is also indicated.

Fig. 65 Importar datos de URL
Fuente: Propia

2.18.4. Herramienta LogFile Analyzer/ Semrush

Log File Analyzer es una fuente de información de confianza para ayudarte a entender todos los matices del rastreo de tu web. Comprueba los códigos de estado Log File Analyzer te ayuda a conseguir datos sobre errores y detecta problemas técnicos, estructurales o navegacionales que obstaculizan a los bots de Google(Anon 2019).

A continuación, se muestra los requisitos mínimos que se requiere para la utilización de la herramienta LogFile Analyzer/semrush:

- Conexión a Internet
- Descargar los archivos .log
- Subir a la plataforma los archivos obtenidos.

En el caso de esta herramienta no se muestra las características mínimas para realizar el análisis respectivo ya que se realiza mediante la página oficial.

A continuación, se detalla cada uno de los pasos para utilizar la herramienta Log File Analyzer/Semrush:

Paso 1.- Desde cualquier navegador ingresar al URL <https://es.semrush.com/features/log-file-analyzer/>

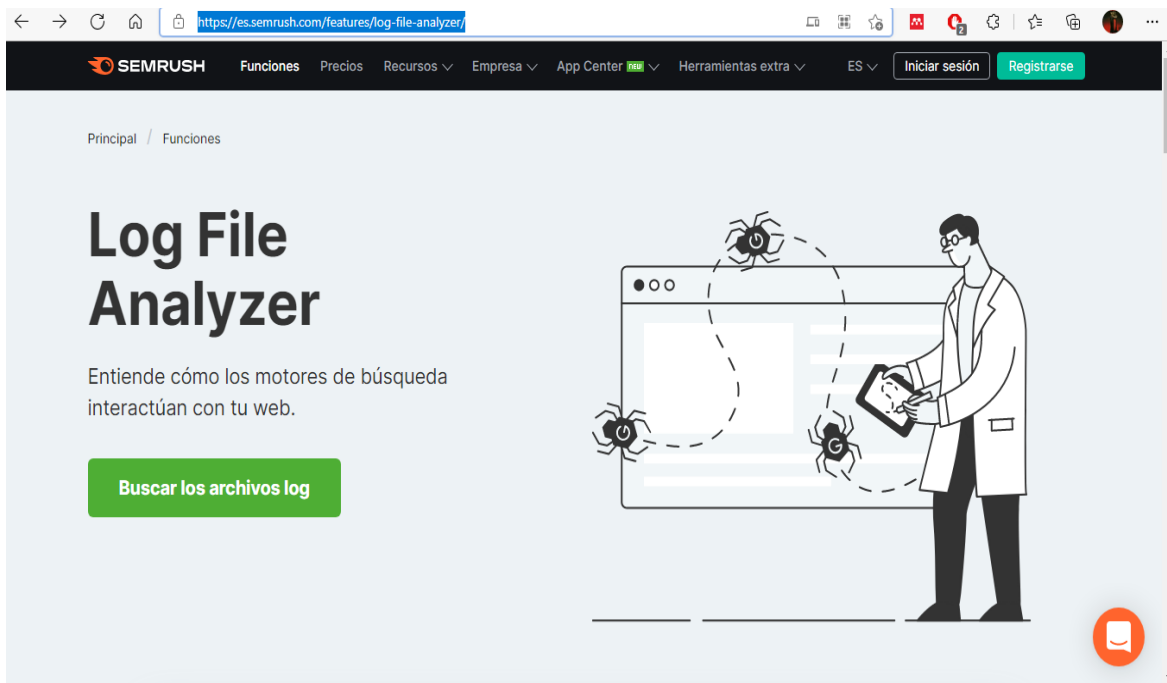


Fig. 66. Pantalla Principal Log File Analyzer/Semrush
Fuente: Propia adaptada a (Anon 2019)

Paso 2.- Dar clic en el botón Buscar los archivos log




Fig. 67. Botón Buscar archivos log
Fuente: Propia

Paso 3.- Crear una cuenta



Crear una cuenta

 Continuar con Google

or

Email

Contraseña

Al hacer clic en "Crear una cuenta", aceptas los [Términos de servicio](#) y la [Política de privacidad](#) de Semrush

Crear una cuenta

¿Ya tienes una cuenta? [Iniciar sesión](#)

Fig. 68. Crear cuenta
Fuente: Propia

Paso 4.- Seleccionar archivos .log

Log File Analyzer

We're offering you an exclusive opportunity to try our new tool that will give you an exact understanding of how search engines are interacting with your website. Be the first to try it!



Drag & drop your log files* here

or

Browse for log files

* Access log files, uncompressed, less than 1 GB in size.

We need information from your log files in order to provide you with analysis services in line with the Log File Analyzer. You don't need to provide any personal information to use our services.

Before uploading log files, make sure they do not contain personal data. If we detect personal data in your log files, we will delete it.

Fig. 69. Selección archivos .log
Fuente: Propia

Paso 5.- Seleccionar archivos .log, se recomienda utilizar los archivos contenidos en acces_log.

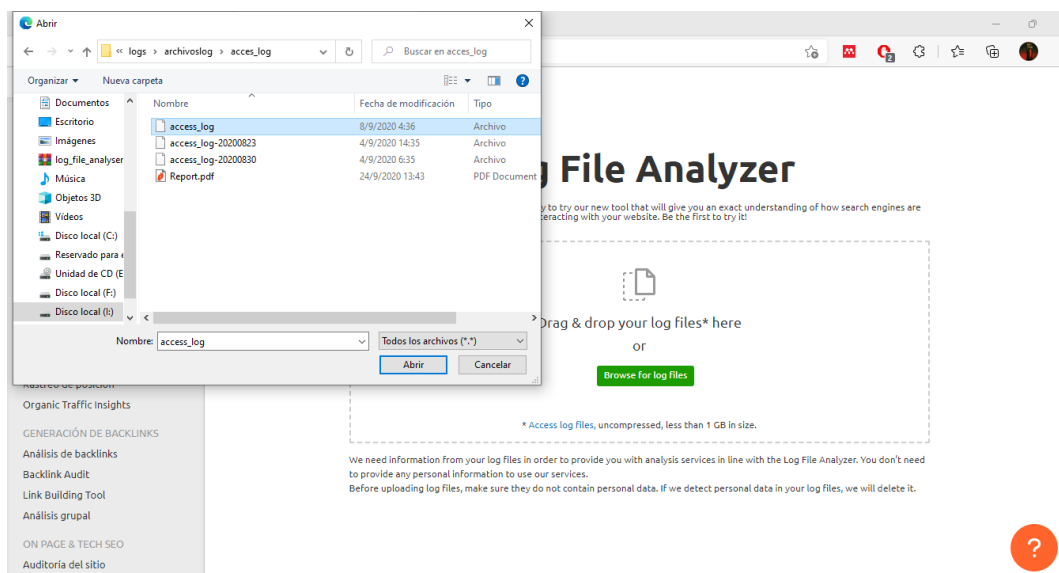
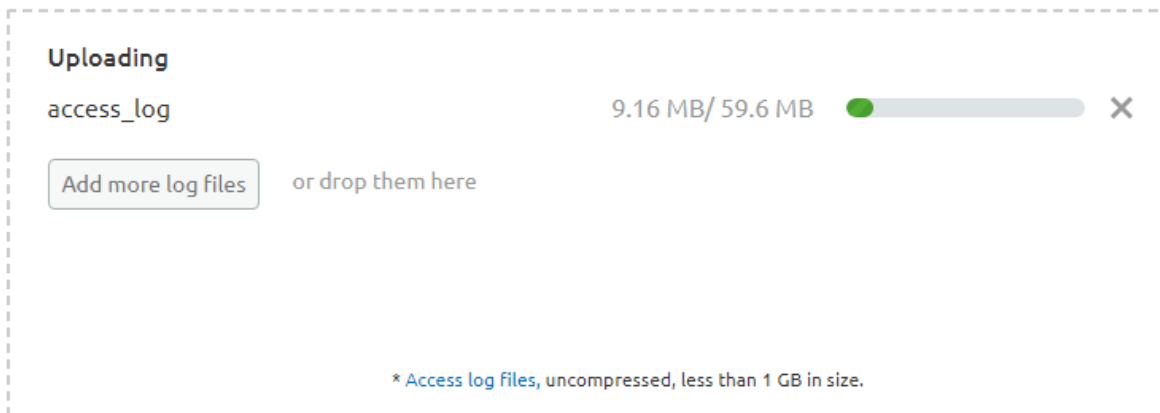


Fig. 70. Selección de archivos desde ruta especificada
Fuente: Propia

Log File Analyzer

We're offering you an exclusive opportunity to try our new tool that will give you an exact understanding of how search engines are interacting with your website. Be the first to try it!



We need information from your log files in order to provide you with analysis services in line with the Log File Analyzer. You don't need to provide any personal information to use our services.
Before uploading log files, make sure they do not contain personal data. If we detect personal data in your log files, we will delete it.

Fig. 71. Barra de proceso de subida de archivos log
fuente: Propia

Paso 6.- Dar clic en el botón Start Log File Analyzer

Log File Analyzer

We're offering you an exclusive opportunity to try our new tool that will give you an exact understanding of how search engines are interacting with your website. Be the first to try it!



Start Log File Analyzer

Fig. 72 Botón Start Log File Analyzer
Fuente: Propia

Log File Analyzer



We are processing your log files. This may take up to 15 minutes.

Processing files 0 / 1

Fig. 73. Proceso de análisis de archivos .log
Fuente: Propia

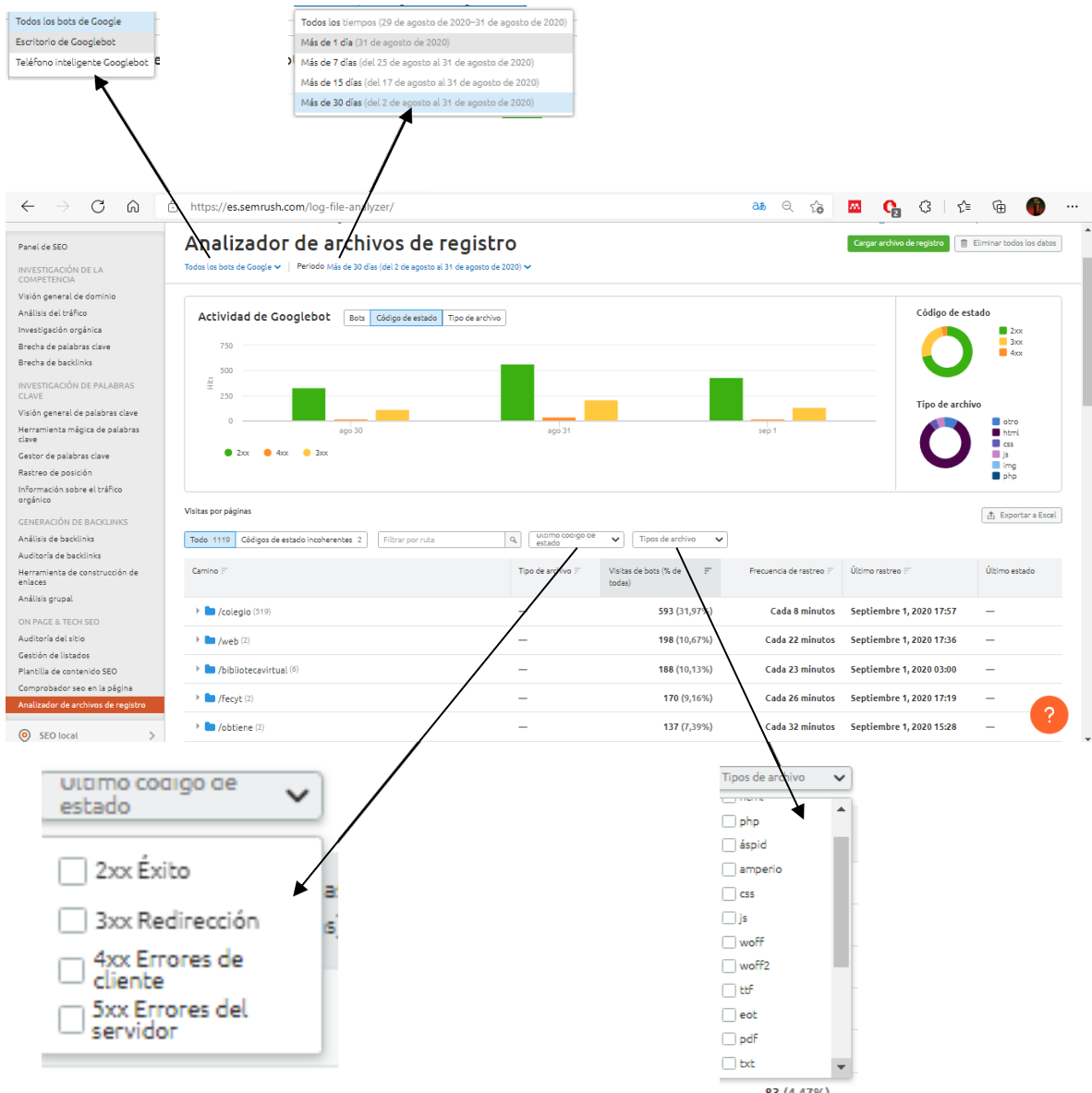


Fig. 74. Reporte de análisis de archivos .log
Fuente: Propia

Como se puede observar en la Fig. 73, Una vez realizado el análisis de archivos .log se puede evidenciar gráficamente el comportamiento de Googlebot en un periodo de tiempo determinado por la fecha que se haya realizado la copia de seguridad (backup) de los archivos .log.

Puede filtrar para mostrar solo la actividad de Bots de escritorio o móviles con el filtro "Todos los Bots de Google" que se encuentra en la parte superior de la tabla y ajustar

el período de tiempo. Los gráficos de la derecha le indican con cuántos de cada código de estado y el tipo de archivo con el cual interactuaron los Bots.

Se puede analizar todas las rutas que recibieron la mayor cantidad de visitas de un bot en el período de tiempo. Para profundizar en este informe, puede filtrar por código de estado, palabra clave en la ruta de acceso o tipo de archivo.

Con esta información obtenida se busca coherencia en los estados de respuesta para investigar cualquier problema de disponibilidad. También puede investigar las visitas de bot por tipo de contenido. Esto le ayuda a comprender si el gasto del presupuesto de rastreo ha cambiado con el tiempo.

Los filtros de tipo de archivo incluyen:

- HTML
- PHP (EN INGLÉS)
- amperio
- CSS
- Javascript
- JSON
- etc.

Visitas por páginas Exportar a Excel

Camino	Tipo de archivo	Visitas de bots (% de todas)	Frecuencia de rastreo	Último rastreo	Último estado
▸ /colegio (519)	—	593 (0,16%)	Cada 8 minutos	Septiembre 1, 2020 5:57	—
▸ /web (2)	—	198 (0,11%)	Cada 22 minutos	Septiembre 1, 2020 5:36	—
▸ /bibliotecavirtual (6)	—	188 (10,13%)	Cada 23 minutos	Septiembre 1, 2020 3:00	—
▸ /fecyt (2)	—	170 (9,16%)	Cada 26 minutos	Septiembre 1, 2020 5:19	—
▸ /obtiene (2)	—	137 (7,39%)	Cada 32 minutos	Septiembre 1, 2020 3:28	—
▸ /ficaya (1)	—	83 (4,47%)	Cada 53 minutos	Septiembre 1, 2020 5:03	—
▸ /ficayaemprende (38)	—	66 (3,56%)	Cada 2 horas	Septiembre 1, 2020 5:59	—
▸ /reducir (1)	—	59 (3,18%)	Cada 2 horas	Septiembre 1, 2020 5:35	—
▸ /facee (1)	—	57 (3,07%)	Cada 2 horas	Septiembre 1, 2020 2:38	—
▸ /fccss (1)	—	56 (3,02%)	Cada 2 horas	Septiembre 1, 2020 5:19	—
▸ /transparencia (2)	—	30 (1,62%)	Cada 3 horas	Septiembre 1, 2020 3:42	—
▸ /legislacion (3)	—	22 (1,19%)	Cada 4 horas	Septiembre 1, 2020 1:56	—
▸ /cultura (2)	—	20 (1,08%)	Cada 4 horas	Septiembre 1, 2020 11:36	—

Fig. 75. Vistas por pagina
Fuente: Propia

2.19. Evaluación de funcionalidad de herramientas

En la Tabla 35, se muestra la comparativa de las funciones presentadas por las diferentes herramientas utilizadas para el análisis de archivos .log.

Tabla 35 Comparativa de funcionalidad de las herramientas

Características	RStudio		WebLog Expert lite		Screaming Frog Analyzer		SEMRUSH Log file analyzer	
	Si	No	Si	No	Si	No	Si	No
	Open source	✓			✓		✓	✓
Multiplataforma	✓			✓	✓		✓	
Facilidad de uso		✓	✓		✓		✓	
Genera reportes en diferentes formatos (PDF, HTML, CSV)	✓		✓		✓			✓
Se necesita conocimientos previos de algún lenguaje de programación	✓			✓		✓		✓
Perfiles y sitios ilimitados	✓		✓			✓		✓
Detección automática de formato .log		✓	✓		✓		✓	
Reconoce archivos comprimidos (GZ, ZIP) que contienen archivos log		✓	✓		✓		✓	

Búsqueda de DNS multiproceso	✓	✓		✓	✓
Procesa un gran volumen de datos	✓	✓	✓		✓

Fuente: Propia

CAPÍTULO III

3. Validación de resultados

Para la validación de la arquitectura descrita en el Capítulo II, se planteó realizar las pruebas de uso de las diferentes herramientas de análisis de archivos log, esto permite evaluar cada una de ellas, juntamente con los expertos o partes interesadas del proyecto (Stakeholders), para así poder determinar cuál de ellas es la que más se ajusta a sus requerimientos, tomando en cuenta, facilidad de uso, seguridad de información, además, de determinar si los patrones o la información generada es importante para conocer el desempeño del portal Web de la UTN, determinar que URLs están siendo rastreadas y si el portal se encuentra seguro ante ataques externos.

Los datos obtenidos mediante la encuesta se dividen en diferentes dimensiones tales como: datos demográficos, calidad del Sistema, calidad de la información, calidad del servicio, uso - intensidad de uso, satisfacción del usuario y beneficios obtenidos.

El desarrollo del proyecto permitió fortalecer el proceso de analítica Web, alineada con un diseño de una arquitectura tecnológica estandarizada, ya que en la actualidad se habla mucho de aplicación de diferentes herramientas que permitan analizar los archivos .log, pero no se habla de que estén apoyadas por estándares de calidad.

Por otra parte, utilizar técnicas y herramientas que permitan realizar un análisis profundo de los archivos de registro .log, permitió a los expertos comprender el contenido de estos y a la vez la importancia de realizar una analítica web en el portal web de la UTN.

Una vez realizado las diferentes pruebas con las herramientas propuestas, se realizó un cuestionario, basado en las variables de éxito de Delone y Mclean, que permite evaluar productos de software para pequeños entornos y medir las distintas dimensiones que conforman un sistema (William and Ephraim 2003).

El cuestionario diseñado para la aplicación de la encuesta tiene como objetivo la obtención de datos confiables utilizando preguntas cerradas, las cuales se dividen en: datos informativos, calidad del Sistema, calidad de la información, calidad del servicio, uso - intensidad de uso, satisfacción del usuario y beneficios obtenidos.

La encuesta se aplicó a las partes involucradas en el proyecto, a quien se realizó la presentación y prueba de la herramienta seleccionada, de un total de 5 personas quienes se envió la invitación el 100 % uso prueba de la herramienta y realizaron la encuesta.

Este capítulo se centra en realizar la validación de la herramienta seleccionada para el análisis de archivos .log y determinar el grado de satisfacción de la herramienta de análisis de archivos .log, con base en un análisis e interpretación de los datos recopilados a partir de la encuesta aplicada a las partes involucradas.

3.1. Sección 1. Datos informativos

En esta sección se detalla los datos informativos de las partes involucradas relacionados con el proyecto (Stakeholders), comprendido únicamente, por la edad, género y relación con el proyecto

1. Edad

[Más detalles](#)

● Hasta 24 años	0
● Entre 25 y 34 años	2
● Entre 35 y 44 años	3
● Entre 45 y 54 años	1
● Mayor de 55 años	0



Fig. 76 Edad
Fuente: Propia

Análisis

En la **Fig. 76**, se puede evidenciar que las personas encuestadas la mayoría corresponden al rango de edad entre 35 y 44 años que corresponde al 50%, del rango de edad 25 y 34 años corresponde al 23% y 45 y 54 años corresponde al 17%, debido a que se realizó a los profesionales involucrados en el proyecto.

2. Género

[Más detalles](#)

● Masculino	4
● Femenino	2



Fig. 77. Género
Fuente: Propia

Análisis

En la **Fig. 77**, se puede evidenciar que la mayoría de las personas encuestadas corresponde al género masculino, dando un total de 67% de hombres y 33% de mujeres.

3. Relación con el proyecto

[Más detalles](#)

● Director de TI	1
● Docente Investigador	1
● Analista / Técnico de TI	3
● Tesista	1

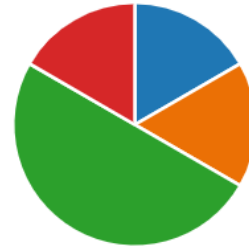


Fig. 78. Relación con el proyecto
Fuente: Propia

Análisis

En la **Fig. 78**, se puede observar los profesionales de la UTN involucrados en el proyecto, los cuales ayudarán a determinar el grado de satisfacción del software WebLog Expert, dando lugar el 50% de los encuestados corresponde a los analistas/ Técnicos de TI, y el resto de involucrados corresponde a director de TI, Docente Investigador (17%), y por último el tesista (17%).

3.2. Sección 2. Variables del modelo de éxito de los sistemas de información de Delone y Mclean

En esta sección se muestra la síntesis de respuestas relacionadas con las variables de éxito del modelo de DeLone y McLean: Calidad del sistema, Calidad de la información, Calidad del servicio, Uso e intención de uso, Satisfacción del usuario y beneficios obtenidos.

Se solicitó a los encuestados que indicaran si están de acuerdo o en desacuerdo con las afirmaciones presentadas en cada sección, para la obtención de las respuestas se usó la escala de Likert con una numeración de 1 a 5, donde 1 significa en total desacuerdo y 5 corresponde a una total aceptación. A continuación, se presenta un análisis más detallado de las diferentes variables.

3.2.1. Calidad del Software

La calidad del software es una característica que abarca el funcionamiento y la forma de procesar la información, se debe tomar en cuenta la accesibilidad, el tiempo de respuesta, además de la facilidad de uso.

4. Calidad de Software WebLogExpert

[Más detalles](#)

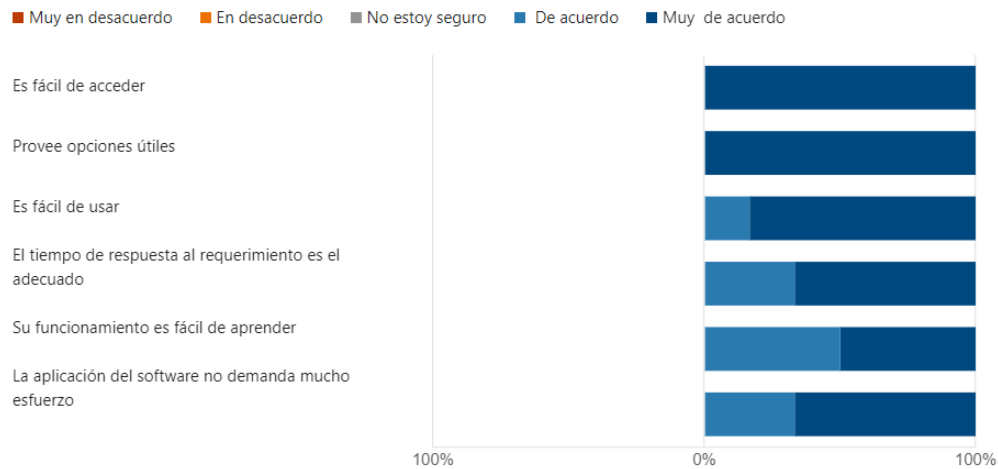


Fig. 79. Calidad del Software
Fuente: Propia

Análisis

En la **Fig. 79**, se evidencia las afirmaciones acerca de la calidad del software WebLog expert, donde se puede observar que los involucrados manifiestan de manera positiva dentro del rango de acuerdo y muy de acuerdo, lo que da a entender que el software cumple con las expectativas y es de buena calidad,

3.2.2. Calidad de la información

La calidad de la información se refiere a como se presenta la información o como se muestran los resultados del software con distintas características tales como: la confiabilidad, la precisión de los datos obtenidos, la calidad, la cantidad de información, la relevancia y si es adecuada.

5. Calidad de la Información del software WebLogExpert

[Más detalles](#)

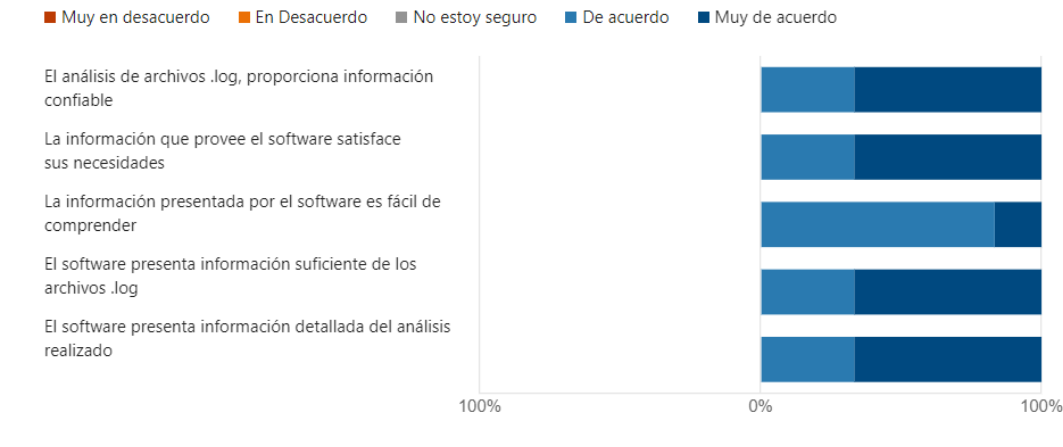


Fig. 80. Calidad de la información
Fuente: Propia

Análisis

En la **Fig. 80**, se puede evidenciar que del total de los encuestados en su mayoría están completamente satisfechos con la calidad de información que muestra el software con respecto al análisis de archivos .log.

El 100% de las respuestas obtenidas fueron de manera favorable con base en la comprensibilidad de la información, si es suficiente, al igual que en el resto de las afirmaciones fueron del 100% de aceptación, es decir estaban de acuerdo y muy de acuerdo respectivamente, lo que significa que los usuarios aprueban la calidad de la información y en general el sistema está organizado correctamente para facilitar su uso.

3.2.3. Calidad del Servicio

La calidad de servicio se relaciona con la efectividad del soporte brindado a los usuarios, como también la disponibilidad del sistema para tener acceso en cualquier momento, demás evidenciar si el sistema cumple o no con las funcionalidades para las que fue diseñado.

6. Calidad del Servicio del software WebLogExpert

[Más detalles](#)

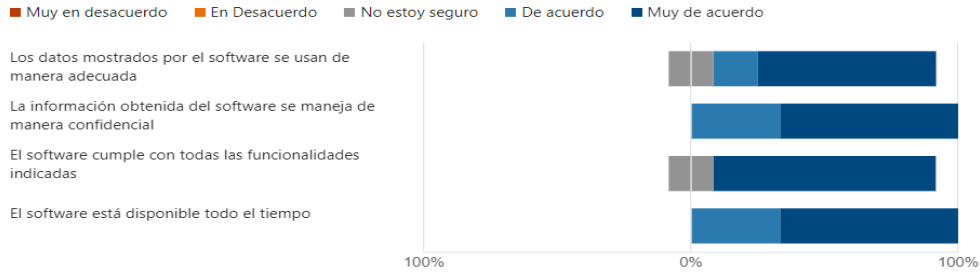


Fig. 81. Calidad del servicio
Fuente: Propia

Análisis

Para la validación de la variable de calidad de servicio en la encuesta se incluyeron las afirmaciones presentes en la **Fig. 81**. El 16,7% de los encuestados respondieron que no estaban seguros de que los datos mostrados por el software se usan de manera adecuada y el restante 16,77% están de acuerdo y el restante correspondiente al 66,7% están muy de acuerdo con esta afirmación, con respecto a la disponibilidad del sistema y que el sistema cumple con las funcionalidades indicada, el 16,7% no están seguros y el 83.3% están muy de acuerdo con la afirmación indicada, con respecto a la disponibilidad todos respondieron afirmativamente.

3.2.4. Uso – intención de Uso

En esta parte hace referencia a la intención de uso que tienen los usuarios, dependiendo de los beneficios que el sistema les brinde como es el incremento de la productividad, mejoras en el desempeño de su trabajo que en este caso es la realización de revisiones de literatura como se muestra en la Fig 82.

7. Intención de Uso del software WebLogExpert

[Más detalles](#)

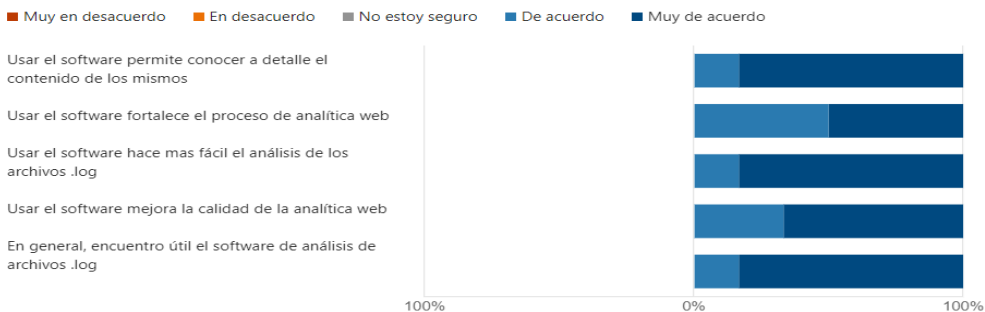


Fig. 82. Uso e intención de uso
Fuente: Propia

Análisis

En esta sección los encuestados en definitiva tienen la intención de usar el software, las respuestas son positivas en todas las afirmaciones, en donde el 100% de las respuestas está dividido entre de acuerdo y muy de acuerdo.

3.2.5. Satisfacción de usuario

Con esta variable se trata de identificar que tan satisfechos se sienten los usuarios con el software, si están de acuerdo que el sistema presenta la información necesaria, si cumple con lo que ellos esperaban, además de que si piensa en seguir usando el sistema.

8. Satisfacción del Usuario en relación con el software WebLogExpert

[Más detalles](#)

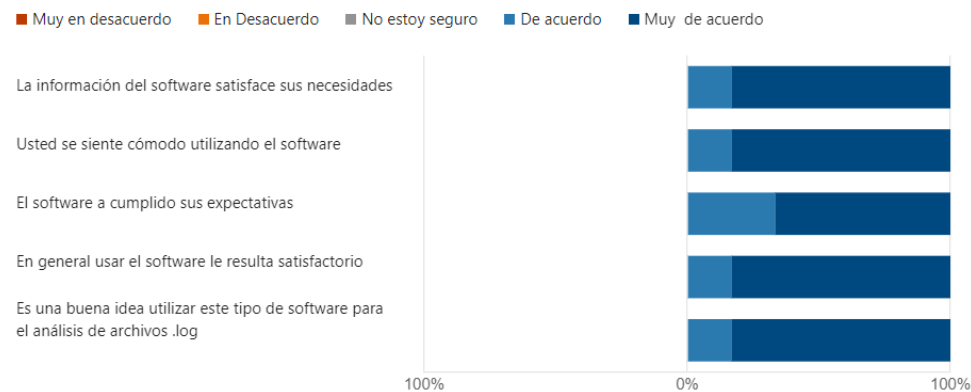


Fig. 83. Satisfacción del usuario

Fuente: Propia

Análisis

De igual manera que en la anterior sección los encuestados respondieron positivamente, lo que da a entender que el uso del sistema fue satisfactorio para ellos como se muestra en la **Fig. 83**. En la primera afirmación que se especifica si el sistema satisface sus necesidades dando lugar al 100% de las afirmaciones positivas, se sienten cómodos utilizando el software, la información del sistema satisface sus necesidades. Debido a la satisfacción de los usuarios, ellos respondieron de forma muy positiva a la afirmación; es buena idea utilizar el software de análisis de archivos .log.

3.2.6. Beneficios obtenidos

Esta variable se refiere a que beneficios consiguen los usuarios al utilizar el sistema, es decir la contribución que les brinda el software para fortalecer el proceso de analítica Web.

9. Beneficios Obtenidos del software WebLogExpert

[Más detalles](#)

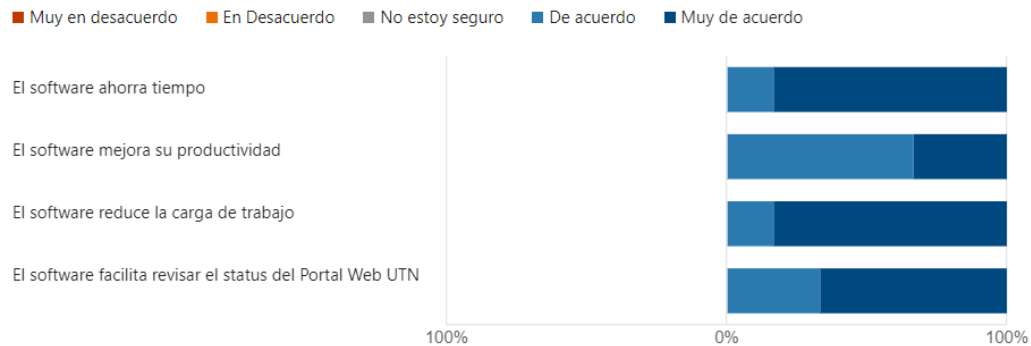


Fig. 84. Beneficios obtenidos
Fuente: Propia

Análisis

En las afirmaciones de esta sección que se muestran en la **Fig. 84**, la mayoría de los encuestados respondieron satisfactoriamente acerca de los beneficios que se obtienen al utilizar el software de análisis de archivos .log.

3.3. Calificación a la herramienta WebLog expert

Este apartado permite comprender la calificación que le da al software cada una de las partes involucradas.

10. ¿Cómo calificaría usted, de manera general a la herramienta WebLogExpert para el análisis web del Portal UTN?

[Más detalles](#)



Fig. 85. Calificación de la herramienta
Fuente: Propia

Análisis

En la **Fig. 85**, se puede evidenciar que el total de encuestados respondieron afirmativamente dando así que el 100 % de ellos le dan una calificación entre 9 y 10 respectivamente, es decir que la herramienta WebLog expert, tiene un alto grado de satisfacción del uso en general para el análisis de archivos .log.

El modelo planteado por Delone y Mclean para medir el éxito de los sistemas de información ayudo a evaluar la efectividad del sistema que se utilizó para el análisis de archivos .log, tras aplicar la encuesta a los usuarios, se obtuvo resultados positivos en las distintas categorías presentes en el modelo; En cuanto a la calidad del sistema, el 100% de los usuarios estuvieron muy de acuerdo que se puede acceder fácilmente al sistema, además de estar de acuerdo y muy de acuerdo que el sistema es fácil de usar y los tiempos de respuesta son los adecuados, los usuarios indicaron que la calidad de la información es muy aceptable, ya que la información que se presenta es bastante precisa, confiable y relevante para las investigaciones que se realizan, de igual forma en la calidad del servicio los encuestados estuvieron muy de acuerdo en que el sistema cumple con las funcionalidades especificadas y que presenta una alta disponibilidad, un 20% de los encuestados indico que no estaba seguro en cuanto al manejo de la información adecuada, ante ello se puede resaltar que el software trabaja con normas de seguridad, control de sesiones y encriptación de las contraseñas por lo cual la información está segura, como parte final los encuestados tiene la intención de seguir usando el sistema, ya que están muy satisfechos con los beneficios obtenidos, como son la mejora de productividad, la reducción de la carga de trabajo y en definitiva el ahorro de tiempo.

CONCLUSIONES

- Al realizar en análisis de los archivos .log, generados por el portal Web de la UTN, se logró obtener información oculta detallada, con información fiable y útil para determinar el estado de salud del portal institucional.
- De las herramientas utilizadas, la que mejor se ajusta a los requerimientos de las partes interesadas es la herramienta Weblog expert, ya que es una herramienta eficaz, de fácil uso, no se necesita tener conocimientos previos del manejo de esta.
- La aplicación del estándar ISO/IEC/IEEE 42010 como Norma Internacional para elaborar una arquitectura tecnológica permitió, comprender y establecer lineamientos base de implementación, garantizando el diseño de una arquitectura tecnológica estandarizada y robusta, enfocada a los requerimientos establecidos por las partes interesadas/stakeholders.
- La utilización de la herramienta RStudio tiene sus limitantes, ya que, para poder sacar provecho de todo el potencial de esta, se necesita un nivel de conocimiento alto acerca del lenguaje R.

RECOMENDACIONES

- Para poder elaborar una arquitectura tecnológica adecuada y robusta es de suma importancia estar permanentemente en contacto con las partes interesadas, para así poder cumplir con todos sus requerimientos, preocupaciones y puntos de vista.
- Se recomienda continuar realizando procesos de analítica Web, ya que permite conocer la frecuencia de visitas al sitio, manejo de errores, además de identificar si existe fuga de información, esto permite realizar una auditoria técnica profunda.
- Se recomienda realizar el proceso de Web Scraping, para así poder obtener datos de una manera más rápida y eficaz, datos que sirven para realizar una comparativa con otras instituciones con similares características de negocio.
- Elaborar una bitácora para el proceso de backups de los logs generados en la base de datos de la UTN, para así poder obtener datos de acceso actualizados.

BIBLIOGRAFÍA

- Alentum Software. 2019. "Index @ Www.Weblogexpert.Com." *Web Log Expert*. Retrieved (<https://www.weblogexpert.com/>).
- Anon. 2011. "ISO/IEC/IEEE Systems and Software Engineering -- Architecture Description." *ISO/IEC/IEEE 42010:2011(E) (Revision of ISO/IEC 42010:2007 and IEEE Std 1471-2000)* 1–46.
- Anon. 2019. "Índice de Log File Analyzer Semrush." Retrieved February 24, 2022 (<https://es.semrush.com/log-file-analyzer/>).
- Anon. n.d. "ESTRUCTURA ORGANIZACIONAL – Universidad Técnica Del Norte." Retrieved February 7, 2022a (<https://www.utn.edu.ec/estructura-organizacional/>).
- Anon. n.d. "ORGANIGRAMA APROBADO DEL DEPARTAMENTO DE INFORMÁTICA."
- Anon. n.d. "RStudio - RStudio." Retrieved December 7, 2020c (<https://rstudio.com/products/rstudio/>).
- Anon. n.d. "SEO Log File Analyser | Screaming Frog." Retrieved November 23, 2020d (<https://www.screamingfrog.co.uk/log-file-analyser/>).
- Anon. n.d. "Support @ Www.Weblogexpert.Com."
- Anon. n.d. "Website Traffic - Check and Analyze Any Website | Similarweb." Retrieved February 6, 2022f (<https://www.similarweb.com/>).
- Arcos, Doris Andrea. 2017. "INSTITUTO DE POSTGRADO Maestría En Ingeniería de Software Ingeniería de Software."
- Baeza-Yates, Ricardo. 2009. "Tendencias En Minería de Datos de La Web." *Profesional de La Informacion* 18(1):5–10. doi: 10.3145/epi.2009.ene.01.
- Cervantes Maceda, Humberto, Perla Velasco-Elizondo, and Castro Careaga Luis. 2016. *Arquitectura de Software*. 1st ed. Mexico.
- Frank, Eibe, Mark A. Hall, and Ian H. Witten. 2017. "The WEKA Workbench." *Data Mining* 553–71. doi: 10.1016/b978-0-12-804291-5.00024-6.
- Hasperue, Waldo. 2013. *Extracción Del Conocimiento En Grandes Bases De Datos Utilizando Estrategias Adaptativas*. Vol. 53.
- Hernandez Orallo, José, María José Ramírez Quintana, and Cesar Ferri Ramirez. 2004. *Introducción a La Minería de Datos*. 1 era. edited by Pearson. Madrid.
- Hilliard, Rich. 2011. "Architecture Viewpoint Template for ISO / IEC / IEEE 42010 v.2.2." 2.1b:1–11.
- Intelligent. 2015. "Minería Web: De Contenidos, de Estructuras y de Usos." *Minería de Datos Web* 1. Retrieved (<https://itelligent.es/es/mineria-web-de-contenidos-estructuras-usos/>).
- IONOS. 2016. "Análisis Web." 1. Retrieved December 7, 2020 (<https://www.ionos.es/digitalguide/online-marketing/analisis-web/el-log-el-archivo-de-registro-de-procesos-informaticos/>).

- Lara, Juan Alfonso. 2014. "MINERIA DE DATOS" edited by C. D. E. FINANCIEROS. 287.
- Leskovec, Jure, Anand Rajaraman, and Jeffrey David Ullman. 2014. "Mining of Massive Datasets." *Mining of Massive Datasets*. doi: 10.1017/cbo9781139924801.
- Liferay. n.d. "¿Qué Es Un Portal Web? | Liferay." Retrieved April 5, 2022 (<https://www.liferay.com/es/resources/l/web-portal>).
- Marinescu, Dan C. 2013. "Cloud Infrastructure." Pp. 67–98 in *Cloud Computing*.
- Martín, Jose. 2004. "Determinación de Tendencias En Un Portal Web Utilizando Técnicas No Supervisadas. Aplicación a Sistemas de Recomendaciones Basados En Filtrado Colaborativo."
- Mendoza, Marcelo. 2011. "Minería de Datos En La Web." (May):613–48.
- Ordoñez, Yoanni, and Darian Horacio Grass. 2015. "Integración Entre Python Y Weka Aplicado En La." *Research Gate* (October).
- Ordoñez, Yoanni, Ernesto Vázquez, and Darian grass boada. 2011. *INTEGRACIÓN ENTRE PYTHON Y WEKA APLICADO EN LA MINERIA DE DATOS*.
- Sierra, Basilio. 2006. *Aprendizaje Automático: Conceptos Básicos y Avanzados*.
- Skriganov, M. M. 2011. "Spectrum of Multidimensional Operators with Periodic Coefficients."
- The Open Group. 2018. "The TOGAF® Standard V9.2 Reference Cards." 1–16.
- Ulises Román, Luis Alarcón. 2005. "Minería de Uso de Web Para Predicción de Usuarios En La Universidad." 2(3):7–13.
- UNIR. 2019. "Lenguaje R." *INGENIERÍA Y TECNOLOGIA 1*. Retrieved (<https://www.unir.net/ingenieria/revista/lenguaje-r-big-data/>).
- Warf, Barney. 2018. "Web Mining." *The SAGE Encyclopedia of the Internet*. doi: 10.4135/9781473960367.n275.
- William, H. DeLone, and R. McLean Ephraim. 2003. "The DeLone and McLean Model of Information Systems Success: A Ten-Year Update." *Journal of Management Information Systems* 19(4):9–30.

ANEXOS

Anexo 1.- Encuesta, sondeo rápido: Grado de satisfacción de la herramienta de análisis de archivos .log

Sondeo rápido: Grado de satisfacción de la herramienta de análisis de archivos .log

El propósito de la presente encuesta es medir los beneficios obtenidos y el grado de satisfacción del uso de la herramienta WebLogExpert para el análisis de archivos .log que permitan fortalecer el proceso de analítica web del Portal Web UTN, el objetivo es medir el impacto de éxito de la propuesta tecnológica mediante un modelo de medida multidimensional DeLone & McLean: Calidad del Software, Calidad de la Información, Calidad del Servicio, Intención de Uso, Satisfacción del Usuario y Beneficios Obtenidos.

La encuesta tardará aproximadamente 9 minutos en completar y sus respuestas son anónimas.

* Obligatoria

Datos Informativos

1. Edad *

- Hasta 24 años
- Entre 25 y 34 años
- Entre 35 y 44 años
- Entre 45 y 54 años
- Mayor de 55 años

Fig. 86. Encuesta
Fuente: Propia

Anexo 2.- Manuales de uso de herramientas de análisis de los archivos .log

MANUAL DE USO WEBLOGEXPERT.pdf

MANUAL USO_SCREAMING FROG.pdf

MANUAL USO_LOGFILEANALYZER.pdf

MANUAL USO_RSTUDIO.pdf

Anexo 3.- Reporte generado por la herramienta WebLog expert

Reporte