

Tabla de Contenido

1.	Introducción.....	1
1.1.	Contexto.....	1
1.2.	Oportunidad de mejora.....	2
1.3.	Objetivos de la tesis.....	3
1.4.	Propuesta de solución.....	3
2.	Marco teórico.....	4
2.1.	Big Data.....	4
2.1.1.	Qué es Big Data.....	4
2.1.2.	Tecnologías Big Data.....	7
2.1.2.1.	Sistemas de almacenamiento distribuido (<i>Distributed File System</i>).....	8
2.1.2.2.	Bases de datos NoSQL.....	8
2.1.2.3.	Bases de datos NewSQL.....	9
2.1.2.4.	Procesamiento del tipo “ <i>Distributed Programming</i> ”.....	9
2.1.2.5.	Procesamiento del tipo “ <i>Stream Processing</i> ”.....	10
2.1.2.6.	Procesamiento del tipo “ <i>Graph Processing</i> ”.....	11
2.1.2.7.	Procesamiento del tipo “SQL Processing”.....	11
2.1.3.	Tipos de datos analizados en Big Data.....	12
2.2.	Hadoop.....	12
2.2.1.	Hadoop Distributed File System (HDFS).....	13
2.2.2.	Hadoop MapReduce.....	14
2.2.3.	Hadoop Common.....	15
3.	Esquema de Trabajo.....	16
4.	Etapa de Exploración.....	18
4.1.	Escenario Actual.....	18
4.2.	Oportunidad.....	20
5.	Etapa de Definición y Alcance.....	21
5.1.	Definición de Casos de Usos.....	21
5.1.1.	Trazabilidad de Usuarios.....	21
5.1.2.	Detección de Errores.....	21
5.2.	Definición de logs a utilizar.....	22
6.	Etapa de Solución Técnica.....	25
6.1.	Definir Tecnología de Big Data.....	25

6.2.	Arquitectura de la solución.....	27
6.2.1.	Fuentes de Conexión.....	28
6.2.2.	Capa de seguridad y balanceo de carga.....	28
6.2.3.	Granja de servidores web.....	28
6.2.4.	Repositorio central de logs web.....	28
6.2.5.	Cloudera Cluster (MR-HDFS).....	29
6.2.5.1.	Pre-procesamiento de datos.....	29
6.2.5.2.	Procesamiento mediante MapReduce.....	30
6.2.5.3.	Post-procesamiento de datos.....	31
6.2.6.	Data WareHouse Institucional (Base de datos).....	32
6.2.7.	Data WareHouse Institucional (Visualización).....	34
6.3.	Métodos Desarrollados.....	34
6.3.1.	MapReduce 01 “Job Trazabilidad”.....	35
6.3.2.	MapReduce 02 “Job Login”.....	38
6.3.3.	MapReduce 03 “Job Errores”.....	41
7.	Etapa de Evaluación.....	43
7.1.	Evaluación de Resultados.....	43
7.1.1.	Evaluación de resultados caso de uso “Trazabilidad de usuarios”.....	44
7.1.1.1.	PN-1 Trazabilidad de un usuario específico.....	44
7.1.1.2.	PN-2 Identificación de usuarios de un determinado navegador web....	45
7.1.1.3.	PN-3 Carga de usuarios en la granja de servidores web.....	46
7.1.1.4.	PN-4 Flujos del sitio web con dificultad.....	47
7.1.2.	Evaluación de resultados caso de uso “Detección de Errores”.....	50
7.1.2.1.	PN 1 Cantidad de errores más presente en el sitio web.....	50
7.1.2.2.	PN 2 Comportamiento de un error específico en la granja web.....	51
7.1.2.3.	PN-3 Detalle de personas con mayor número de errores.....	52
7.2.	Evaluación de escalabilidad.....	53
7.2.1.	Infraestructura de pruebas.....	54
7.2.2.	Evaluación de Escalabilidad Archivos bzip2 en 1, 2, 4 y 8 Nodos.....	55
7.2.2.1.	Estadísticas de escalabilidad MapReduce Job Trazabilidad.....	55
7.2.2.2.	Estadísticas de escalabilidad MapReduce Job Login.....	57
7.2.2.3.	Estadísticas de escalabilidad MapReduce Job Errores.....	58
7.2.3.	Evaluación de Compresión Archivos bzip2, gzip y texto en 8 nodos.....	60
8.	Conclusión.....	62
9.	Bibliografía.....	64

Índice de tablas

Tabla 1: Tabla de sistema de información en base byte.	4
Tabla 2: Características de máquina cluster Cloudera.	29
Tabla 3: Tabla de compresión Cloudera Hadoop.	29
Tabla 4: Tabla de preguntas de negocio para evaluación de resultados.	43
Tabla 5: Tabla de pruebas de escalabilidad.	53
Tabla 6: Tabla de resultados de escalabilidad del Job Trazabilidad.	55
Tabla 7: Tabla de resultados de escalabilidad del Job Login.	57
Tabla 8: Tabla de resultados de escalabilidad del Job Errores.	58
Tabla 9: Tabla de resumen de evaluación de compresión.	60
Tabla 10: Tabla de cumplimiento de objetivos específicos.	62

Índice de Figuras.

Ilustración 1: Gráfico de crecimiento del universo digital [1].	5
Ilustración 2: Gráfico de tipos de datos [1].	5
Ilustración 3: Las tres V (Volumen, Velocidad, Variedad) de Big Data.	6
Ilustración 4: Diagrama de la arquitectura de HDFS.	13
Ilustración 5: Diagrama de carga de archivo a Hadoop HDFS.	14
Ilustración 6: Diagrama del flujo de procesamiento de MapReduce.	15
Ilustración 7: Diagrama del esquema de trabajo.	16
Ilustración 8: Diagrama de la arquitectura actual de la organización.	18
Ilustración 9: Gráfico de líneas de log de ambiente web de la organización.	19
Ilustración 10: Gráfico de distribución entre los tipos de log de ambiente web.	20
Ilustración 11: Extracto de log wPortalLog.	22
Ilustración 12: Extracto de log wEmpresasLog.	23
Ilustración 13: Extracto de log wPersonasLog.	23
Ilustración 14: Extracto de log AccessLog.	23
Ilustración 15: Diagrama de ubicación de log dentro del sitio web.	24
Ilustración 16: Temporalidad de registro en archivos de log del sitio web.	24
Ilustración 17: Esquema de servicios de Cloudera CDH [25].	26
Ilustración 18: Servicios elegidos de Cloudera CDH [25].	26
Ilustración 19: Esquema de la arquitectura de la solución.	27
Ilustración 20: Esquema de pre-procesamiento de datos.	30
Ilustración 21: Esquema de procesamiento mediante MapReduce.	30
Ilustración 22: Esquema de post-procesamiento de datos.	32
Ilustración 23: Modelo de datos de trazabilidad de Infobright.	33
Ilustración 24: Modelo de datos de errores de Infobright.	33
Ilustración 25: Modelo de "Expresión Regular".	35
Ilustración 26: Diagrama de MapReduce de Trazabilidad.	35
Ilustración 27: Función "Map" de método MapReduce Trazabilidad.	36
Ilustración 28: Función "Reduce" de método MapReduce Trazabilidad.	37
Ilustración 29: Diagrama de MapReduce de Login.	38
Ilustración 30: Función primer "Map" de método MapReduce Login.	39
Ilustración 31: Función segundo "Map" de método MapReduce Login.	39
Ilustración 32: Función "Reduce" de método MapReduce Login.	40
Ilustración 33: Diagrama de MapReduce de Errores.	41
Ilustración 34: Función "Map" de método MapReduce Errores.	42
Ilustración 35: Resultado de trazabilidad de un usuario específico.	44
Ilustración 36: Consulta SQL de trazabilidad de un usuario específico.	45
Ilustración 37: Resultado de usuarios de un determinado navegador web.	45
Ilustración 38: Consulta SQL de usuarios de un determinado navegador web.	46
Ilustración 39: Gráfico de carga de usuarios concurrentes.	47
Ilustración 40: Consulta SQL de carga de usuarios concurrentes.	47
Ilustración 41: Flujo de navegación de recuperación de clave.	48
Ilustración 42: Registro de acciones del flujo de recuperación de clave.	48
Ilustración 43: Resultado de casos de flujo de navegación de recuperación de clave.	49
Ilustración 44: Consulta SQL de flujo de navegación de recuperación de clave.	49

Ilustración 45: Resultado de errores más presentes en sitio web.....	50
Ilustración 46: Consulta SQL de errores más presentes en sitio web.....	50
Ilustración 47: Selección de error específico dentro de los errores más comunes.	51
Ilustración 48: Gráfico de comportamiento de un error específico en la granja web.....	51
Ilustración 49: Consulta SQL de comportamiento de un error específico.	52
Ilustración 50: Resultado de las personas con mayor número de errores.	52
Ilustración 51: Consulta SQL de las personas con mayor número de errores.	53
Ilustración 52: Diagrama de arquitectura Cloud AWS de pruebas.....	54
Ilustración 53: Vista de la arquitectura dentro de AWS.	55
Ilustración 54: Gráfico de duración del Job Trazabilidad.	56
Ilustración 55: Gráfico de velocidad de procesamiento del Job Trazabilidad.....	56
Ilustración 56: Gráfico de duración del Job Login.....	57
Ilustración 57: Gráfico de velocidad de procesamiento del Job Login.	58
Ilustración 58: Gráfico de duración del Job Errores.....	59
Ilustración 59: Gráfico de velocidad de procesamiento del Job Errores.	59
Ilustración 60: Gráfico de volumen de procesamiento de la evaluación de compresión. ..	60
Ilustración 61: Gráfico de duración de procesamiento de la evaluación de compresión...	61
Ilustración 62: Gráfico de velocidad de procesamiento de la evaluación de compresión..	61