

TABLA DE CONTENIDOS

	página
Dedicatoria	I
Agradecimientos	II
Tabla de Contenidos	III
Índice de Figuras	v
Índice de Tablas	VI
Resumen	VII
1. Introducción	8
1.1. Definición del Problema	8
1.2. Propuesta de solución	8
1.3. Hipótesis	9
1.4. Objetivos	9
1.5. Resultados.	9
2. Contexto	10
2.1. Protein Data Bank y sus estructuras	10
2.1.1. Protein Data Bank	10
2.1.2. Proteínas y sus componentes	11
2.2. AFAL2	12
2.3. Apache Hadoop	15
2.4. Map Reduce	15
2.5. Apache Spark	16
2.6. Trabajo relacionado	18
3. Analizando Proteínas con Spark	21
3.1. Funcionamiento de AFAL2.	21
3.1.1. Frecuencia específica.	21
3.1.2. Frecuencia General.	22

3.1.3. Ocurrencia.	22
3.2. Base de datos de AFAL2.	22
3.2.1. Consultas SQL de AFAL2	25
3.3. Desnormalización de la Base de Datos	26
3.4. Pre-procesamiento de archivos.	27
3.4.1. Conversión de archivos para Spark.	27
3.5. Cálculo de Frecuencia General en Spark.	29
3.6. Cálculo de Frecuencia Específica en Spark.	30
3.7. Cálculo de Ocurrencia	31
4. Experimentos	33
4.1. Metodología.	33
4.1.1. Escenarios de prueba.	33
4.2. Resultados.	34
4.2.1. Comparación de tiempos en la misma base de datos	34
4.2.2. Comparación de tiempos en distintos conjuntos de datos	36
4.3. Conclusiones de los experimentos.	38
5. Conclusiones	39
5.1. Sobre los objetivos.	39
5.2. Comentarios adicionales.	40
Bibliografía	41
Anexos	
A: Códigos completos de las consultas en Spark	44
A.1. Código para la Frecuencia General	44
A.2. Código para la Frecuencia Específica	45
A.3. Código para la Ocurrencia	46
A.4. Ejecución de Spark por líneas de comandos.	48
A.5. Ejecución de Spark en un Cluster de Amazon EMR.	48

ÍNDICE DE FIGURAS

	página
2.1. Ejemplo de una proteína a nivel molecular, y la relación entre aminoácidos y ligandos.	12
2.2. Parametros de entrada de AFAL2.	13
2.3. Frecuencia Específica en AFAL2	14
2.4. Frecuencia General en AFAL2	14
2.5. Ocurrencia en AFAL2	14
2.6. Ejemplo de Map Reduce	16
2.7. Funcionamiento de WordCount en Spark	18
3.1. Esquema de base de datos de AFAL2	23
3.2. Tabla creada para desnormalizar la base de datos de AFAL2.	26
3.3. Archivos de entrada	28
3.4. Ejemplo del proceso de la Frecuencia General en Spark	30
3.5. Ejemplo del proceso de la Frecuencia Específica en Spark.	31
3.6. Ejemplo del proceso de la Ocurrencia en Spark.	32
4.1. Gráfico de los tiempos de respuesta al calcular la Frecuencia General	35
4.2. Gráfico de los tiempos de respuesta al calcular la Frecuencia Específica.	35
4.3. Gráfico de los tiempos de respuesta al calcular la Ocurrencia.	36
4.4. Gráfico del aumento de tiempo en PostgreSQL.	37
4.5. Gráfico del aumento de tiempo en Spark sobre un notebook.	37
4.6. Gráfico del aumento de tiempo en Spark sobre un Clúster.	38
A.1. Vista de un Cluster creado en Amazon EMR.	48
A.2. Parámetros para agregar un step en Amazon EMR.	48

ÍNDICE DE TABLAS

	página
4.1. Tiempos al calcular la Frecuencia General.	34
4.2. Tiempos de respuesta al calcular Frecuencia Específica.	35
4.3. Tiempos de respuesta al calcular la Ocurrencia.	36
4.4. Promedio de los tiempos de respuesta para cada conjunto de datos. .	37