Universidad Cenfotec

Maestría en Tecnologías de Bases de Datos

Documento final de Investigación Aplicada 2

Análisis y comparación de las capacidades de tecnologías OLAP con las de un clúster de motores de búsqueda para el análisis de datos en sistemas de ventas

Gamboa Ureña, Eduardo

Montoya Rodríguez, Malincy

Mayo, 2018

Declaratoria de Derechos de Autor

Resumen

Se estudian dos tecnologías para el análisis de datos: el cubo OLAP y el motor de búsqueda. El primero permite analizar las transacciones de un depósito de datos tomando en cuenta su contexto, mientras que el segundo es utilizado como filtro de documentos. Ambas tecnologías se comparan con el fin de determinar si pueden ser utilizadas para análisis de datos transaccionales.

Una de las ventajas que ofrecen las tecnologías actuales es el procesamiento en paralelo. La computación paralela es una arquitectura que permite que un grupo de procesadores trabaje de manera colaborativa, a fin de que se pueda resolver un problema determinado. Existen diferentes formas de implementación, sin embargo, una de las más usadas es la forma “nada compartido”, el cual trata de nodos independientes que trabajan en conjunto.

Con el objetivo de poder realizar una distribución de los datos de manera eficiente, se debe escoger un tipo de particionado para los mismos. Distintos algoritmos existen para esto, como el Round Robin, por medio de Hash y por rango de los datos.

Se han realizado diferentes estudios para ambas tecnologías, que van desde análisis de datos científicos hasta el de comportamiento de clientes para tiendas en línea. Cada una de dichas investigaciones muestra las capacidades de análisis de

las tecnologías en estudio, en las que se muestran las ventajas del uso de la misma en cada caso comparándolas con otras metodologías.

Entre ambas tecnologías, se denota una clara ventaja sobre las herramientas OLAP en los ámbitos de recursos y rendimiento. En una misma consulta, los casos de OLAP duran menos tiempo para la misma cantidad de recursos utilizados. Además, es fácil concluir que las aplicaciones de recuperación de información son más intensivas en memoria, mientras que las tecnologías OLAP son más orientadas al consumo del procesador.

Palabras Clave: almacén de datos, motor de búsqueda, comparación, clúster, OLAP, cubos.

Tabla de Contenidos

Índice de Tablas

Índice de Figuras

Capítulo 1 Introducción

1.1 Generalidades

Uno de los activos más importantes de una empresa son los datos que maneja. Ellos no solo reflejan la historia de la empresa, sino que permiten predecir el futuro de esta. Por ejemplo, las ventas de una empresa pueden reflejar tendencias el comportamiento de los clientes respecto a las compras que realizan, patrones de adquisición de productos por parte de los clientes según la época del año, o también indicar qué tipo de publicidad debe de ser utilizada para ciertos clientes.

Existen tecnologías bastante sólidas para poder visualizar datos históricos de manera fiel, una de ellas es el depósito de datos, o "*data warehouse*" en inglés. La implementación de uno de estos depósitos de datos se puede lograr por medio de la combinación de distintas tecnologías. Una de las más utilizadas es el modelo multidimensional, el cual permite realizar consultas a una base de datos de manera más rápida, sacrificando la velocidad y la consistencia a la hora de realizar escrituras de los datos.

Sin embargo, se hallan ciertas limitantes en el uso de un depósito de datos. Si bien es cierto, contienen gran cantidad de datos que son útiles para la organización, es posible ver que la cantidad crece rápidamente, llegando a acumularse una cantidad que no es posible procesar manualmente. A fin de realizar

estas integraciones, se desarrollan tecnologías tipo OLAP. Estas permiten tomar un depósito de datos, realizar operaciones de agregación sobre ellos, para que al final se puedan ver patrones o resultados totales sobre los mismos y no solo el detalle. Además de permitir este tipo de agregaciones, las tecnologías OLAP logran ir desde un nivel general a un nivel detallado de manera más rápida que realizar las consultas equivalentes a un depósito de base de datos directamente.

Con el auge de los datos no estructurados, se ha visto el surgimiento de modelos de bases de datos “no SQL”, es decir, arquitecturas que se alejan del modelo relacional. Además, se ha avanzado en la disciplina de recuperación de la información, lo que ha generado nuevas herramientas para el análisis de datos que permiten generar información más veraz, a partir de los datos de una empresa.

Para esta investigación, se propuso comparar una implementación de tecnologías OLAP contra un clúster de motores de búsqueda. Ambos especializados en el análisis de datos y capaces de hacer agregaciones sobre los mismos. Para comparar las tecnologías mencionadas anteriormente, se utilizó una base de datos de datos de ventas lo suficientemente extensa como para poder realizar pruebas de manera confiable.

1.2 Antecedentes del Problema

Después de una búsqueda preliminar, no parece existir una investigación que haya tratado de efectuar una comparación como la propuesta en este proyecto.

1.3 Definición y Descripción del Problema

Históricamente, se ha utilizado un almacén de datos para acumular datos históricos de una compañía, mezclando y estandarizando varias fuentes de datos, a fin de que se tenga una sola versión de la verdad y poder brindar ayuda a la toma de decisiones a nivel administrativo. Además de estas tecnologías, también se han utilizado motores de búsqueda, los cuales permiten obtener documentos a partir de consultas que realiza un usuario. Tal y como se ha mencionado, debido al avance de las herramientas de análisis de datos, los motores de búsqueda cuentan también con las capacidades de poder realizar diferentes añadiduras y operaciones sobre los datos, lo que genera la posibilidad de reemplazar el uso de tecnologías OLAP.

Se decidió efectuar una comparación de las capacidades de ambas tecnologías en distintos escenarios en los que su uso sea confrontable. En términos generales, se comparó la precisión de los resultados y el desempeño de una consulta completa. Con el objeto de realizar esta comparación, se analizaron diversos casos de uso típicos en negocios que utilizan almacenes de datos, a fin de poder efectuarlos en un clúster de motores de búsqueda. Con el fin de determinar que un caso es exitoso, se tomaron las siguientes métricas:

- **Resultados Correctos:** Ambas tecnologías deben mostrar, para el mismo caso de uso, resultados equivalentes.
- **Velocidad de respuesta:** Se realizaron medidas de cuánto se tarda en realizar una consulta. A fin de tener un punto de comparación, se midió el

tiempo promedio de respuesta de una implementación de cubo OLAP. Una vez calculado este tiempo, se comparó con la implementación del motor de búsqueda escogido. Se considera un fracaso que el motor de búsqueda tome significativamente más tiempo en responder que la implementación del cubo, o viceversa.

- **Uso de recursos:** Se efectuaron mediciones de uso de CPU y memoria RAM durante la ejecución de las pruebas. Ambas tecnologías deben mostrar, para un mismo escenario, una utilización equivalente de los recursos medidos.

1.4 Justificación

Esta investigación ayudará a las empresas, que deseen implementar una solución de inteligencia de negocios para procesar una base de datos relacional, a evaluar las capacidades de dos opciones, con el fin de elegir la opción que se ajuste mejor a sus necesidades. El objetivo principal es evitar invertir en la implementación de una tecnología que podría no ser la adecuada para la situación que enfrentan las empresas, ya sea por costo elevado, dificultad de uso o desempeño de la herramienta a implementar.

1.5 Viabilidad

1.5.1 Punto de Vista Técnico.

A fin de realizar esta investigación, se decidió conseguir servidores de alquiler en la nube, ya que existe amplia oferta con diversas capacidades y sistemas

operativos. En este caso, se usó el servicio Azure, principalmente debido a que se necesitaban servicios que se integraran con el sistema operativo Windows. Azure, al proveer ambientes basados en esta plataforma, encajó en las necesidades de esta investigación.

Además de lo anterior, se cuenta con información para la configuración correcta de los motores de base de datos, motores de búsqueda y sistemas operativos que permitan un buen desempeño en las pruebas a realizar.

1.5.2 Punto de Vista Económico.

Se tienen dos consultores que dedicaron cuatro horas semanales cada uno a la investigación. Para el desarrollo de este proyecto, se utilizaron los siguientes materiales:

- Una máquina virtual, provista por la plataforma Windows Azure. En la misma, se encuentra instalado el sistema SQL Server, en su versión Developer 2017, además de un sistema operativo Windows Server 2016. Esta máquina virtual se encarga de ser el almacén de datos y, asimismo, contener el sistema SQL Server Analysis Services para implementar el cubo OLAP
- Elastic Cloud: Servicio de Elastic que provee una implementación de clústeres del motor de búsqueda Elasticsearch. El servicio es ofrecido en

la nube. Se creó un clúster con 4GB de memoria y 96 GB de almacenamiento.

Los costos fueron asumidos por los investigadores; estos se reflejan en la Tabla 1.

Tabla 1
Costo teórico de investigación

Rubro	Valor
Horas de investigación	$ 1120
Costo de la máquina virtual de Windows Azure	$ 276,44
Costo de licenciamiento de Elastic Cloud	$ 260.62
Total	**$1657.06**

Nota. Valores en dólares americanos.

1.6 Objetivos

Se ha seleccionado la taxonomía de Bloom debido a que la misma posee una estructura idónea para la elaboración de objetivos estructurados de clase pedagógica y que provee un marco de referencia para poder revisar el cumplimiento de estos.

1.6.1 Objetivo General.

Comparar las capacidades de una implementación de tecnologías OLAP con las de un clúster de motores de búsqueda.

1.6.2 Objetivos Específicos.

1. Definir los conceptos de almacén de datos y motor de búsqueda.
2. Comprender las diferencias entre almacén de datos y motor de búsqueda.
3. Elaborar escenarios de uso comparables entre una herramienta OLAP y un clúster de motores de búsqueda.
4. Contrastar las capacidades de un clúster de motores de búsqueda y un cubo OLAP en los escenarios propuestos.
5. Sintetizar los beneficios en situaciones específicas de una tecnología sobre la otra.

1.7 Alcances y Limitaciones

1.7.1 Alcances.

Esta investigación tiene como entregable un documento escrito con la comparación entre las tecnologías planteadas, la cual puede ser usada como guía para la elección entre tecnología OLAP o un motor de búsqueda según los casos generales planteados.

1.7.2 Limitaciones.

Esta investigación no es una guía para crear almacenes de datos, motores de búsqueda ni clústeres. Lo anterior debido a que las configuraciones para casos

específicos de un negocio son dependientes el contexto y del tipo de datos por analizar.

1.8 Estado de la Cuestión

1.8.1 Planificación de la revisión

En esta sección se realiza una revisión sistemática de diversas fuentes de artículos referentes al tema de esta investigación. Además de ello, se procede a verificar dichos artículos, a fin de poder discernir cuáles son viables para la investigación y cuáles no son aplicables a la misma.

1.8.1.1 Formulación de la pregunta

En esta sección se define la pregunta de investigación, la cual demarca el área de trabajo y se anota qué artículos son los utilizados como estudios primarios.

1.8.1.1.1 Foco de la pregunta

En esta revisión sistemática se espera poder localizar y filtrar documentos que otorguen información acerca de depósitos de datos, y motores de búsqueda, a fin de poder contrastar ambas tecnologías en un caso de uso común.

1.8.1.1.2 Amplitud y calidad de la pregunta

Para poder definir la calidad de la pregunta, se usaron como base diversas secciones con el fin de analizar el problema por tratar, la pregunta en sí, los criterios

de búsqueda utilizados para obtener los estudios y, finalmente, cómo se mide la salida de dichas revisiones.

1.8.1.1.2.1 Problema

En toda organización, los datos son el insumo más importante a la hora de realizar una toma de decisiones. Lo anterior permite que el rumbo de una empresa tome una dirección específica, según la información suministrada a partir de los datos. Una de las soluciones por excelencia a este paradigma es un depósito de datos, el cual es una arquitectura que permite realizar un almacenamiento de los datos históricos, y visualizar los mismos. Existen otras herramientas para hacer esto; en particular, los motores de búsqueda, puesto que permiten el procesamiento de datos de manera rápida, además de implementar consultas en un lenguaje cercano al humano.

1.8.1.1.2.2 Pregunta de investigación

La pregunta de investigación, basada en el problema, y la cual denota el rumbo de la investigación es la siguiente:

¿Es posible que un clúster de motores de búsqueda sea capaz de reemplazar una implementación de tecnologías OLAP?

1.8.1.1.2.3 Palabras clave y sinónimos

Con el fin de poder realizar las consultas a las fuentes especificadas, se debe de tener una lista de sinónimos y conceptos que se relacionen con los términos en

cuestión. Lo anterior permite obtener la mayor cantidad de datos relevantes desde los buscadores de artículos. Los conceptos se enumeran en la Tabla 2.

Tabla 2
Conceptos por Investigar

Área	Palabras Clave	Conceptos Relacionados
Depósitos de datos	Data Warehouse, OLAP, OLAP Cube	Relacional, base de datos, SQL, multidimensional
Motores de búsqueda	Search Engines, Index, Aggregations	Information retrieval, recuperación de información
Clústeres	Clustering	Computación paralela, arquitectura nada compartido.

1.8.1.1.2.4 Intervención

En esta revisión fueron observadas las propuestas de investigación en los temas de depósitos de datos, con el objeto de conocer casos de usos particulares y ejemplos de los datos que se pueden analizar, además de estudios acerca de motores de búsqueda. Una vez examinados, se obtuvo un análisis de resultados que permitió determinar si los artículos encontrados son o no una fuente viable de información.

1.8.1.1.2.5 Control

Las únicas fuentes documentales que se incluyeron como trabajos primarios fueron aquellas analizadas por medio de los pasos especificados en esta revisión.

No fueron incluidos trabajos extra, o fuentes de terceros, que no sean arbitrados por este protocolo.

1.8.1.1.2.6 Resultado

El resultado esperado de esta revisión fue obtener fuentes documentales que permitan conocer las características tanto de los motores de búsqueda como de los depósitos de datos, con el objetivo de poder formar un marco conceptual para la investigación.

1.8.1.1.2.7 Medida de la salida

A fin de medir la salida, se agruparon las propuestas encontradas por cada uno de los conceptos que se deseaba; luego se aplicó revisión manual de cada documento, con el objeto de conocer el tema central del mismo.

1.8.1.1.2.8 Aplicación

Los beneficiarios de esta revisión se pueden dividir en dos grupos: en primer lugar, las empresas que deseen tener una forma de analizar datos históricos de transacciones, pues podrán discernir entre qué tecnologías usar; en segundo lugar, las personas que deseen realizar una investigación al respecto, ya que podrán conocer fuentes o documentos que permitan obtener información para sus escritos.

1.8.1.1.2.9 Diseño experimental

El meta-análisis de la revisión fue enfocado en dos objetivos principales: primero, se vieron las tendencias en almacenes de datos, a fin de conocer

implementaciones más recientes y las tecnologías utilizadas; segundo, se observaron las tendencias en motores de búsqueda, con el objetivo de conocer posibles formas de implementación, casos de uso y posibilidades de agregar de información.

1.8.1.2 Selección de fuentes

En la siguiente sección se plantean las fuentes de documentos usadas durante la revisión, con el fin de poder contrastar sus documentos contra los criterios de análisis que se desea utilizar para la filtración de los artículos científicos.

1.8.1.2.1 Definición de criterio de selección de fuentes

Para poder seleccionar las fuentes se utilizaron diversas métricas de medición, tanto objetivas como subjetivas, a fin de obtener una lista de repositorios considerados confiables para la investigación.

A fin de seleccionar las fuentes, se compararon los resultados por medio de plataformas de arbitraje de documentos, a fin de poder conocer la relevancia de dicho repositorio. En caso de no obtener una calificación, se procede a analizar su origen para analizar si es una fuente confiable o líder en el campo correspondiente.

1.8.1.2.2 Lenguaje de estudio

El lenguaje de los documentos, que se consultaron como fuentes principales, es el inglés, por medio de palabras clave en dicho idioma. Sin embargo, la presente revisión se realizó solamente en el idioma español.

1.8.1.3 Identificación de fuentes

En esta sección, se identifican las fuentes que se utilizan como repositorios principales de documentos, a fin de poder realizar una revisión de estos. Se define también la consulta de búsqueda que se utilizó para obtener los documentos más relevantes.

1.8.1.3.1 Método de selección de fuentes

Se utilizó un buscador de documentos de investigación gratuito para poder obtener una muestra inicial de documentos de diversas fuentes. Dichas fuentes fueron cotejadas, luego, contra los sitios de arbitraje y similares, con el objetivo de poder obtener las fuentes con mejores calificaciones. Si por el contrario no se encontró dicha calificación, se analizó el origen del artículo con el propósito de conocer la credibilidad de este.

1.8.1.3.2 Lista de fuentes

Las fuentes que se utilizan en esta revisión son las siguientes:

- Research Cell: An International Journal of Engineering Sciences,
- International Journal of Scientific & Engineering Research,

- Bulletin of the Marathwada Mathematical Society,
- Google Scholar,
- Research Gate,
- Documentación de los productos utilizados.

1.8.1.3.3 Cadenas de búsqueda

Se utiliza la siguiente cadena de búsqueda, cuando sea posible, a fin de poder obtener resultados relevantes en el área de almacenes de datos:

(Data AND Warehouse) OR OLAP OR (OLAP AND Cube)

Adicionalmente, se utiliza la siguiente cadena de búsqueda para obtener resultados en el área de recuperación de información:

Information AND Retrieval

Además, para obtener información acerca de motores de búsqueda, se utiliza la siguiente consulta:

Search AND engine

Finalmente, para obtener información acerca de clústeres y de computación paralela, se utilizó la siguiente cadena:

Cluster OR (Parallel AND computing)

1.8.1.4 Selección de fuentes después de la evaluación

Luego de ejecutar la evaluación y el refinamiento respectivo de los documentos que se utilizaron, se analizaron documentos que, aunque la relevancia no fue la mejor, es posible que aclaren conceptos útiles para los objetivos investigativos, o que permitan fortalecer o validar las ideas aquí descritas.

1.8.1.4.1 Selección de los estudios

Esta sección explica, de manera puntual, el proceso de selección de documentos seguido para esta revisión además de los criterios de inclusión y exclusión implementados.

1.8.1.4.1.1 Procedimiento para la selección de estudios

Con el fin de poder seleccionar los estudios que se utilizaron en esta investigación, se empleó un enfoque iterativo entre los documentos dentro de las diversas fuentes encontradas. Cada fuente es analizada de manera que se puedan encontrar los documentos relevantes a los objetivos planteados. Para esto, se examina cada documento para que cumpla con los criterios de inclusión y exclusión de documentos que son descritos en las próximas secciones de esta revisión.

Para todo documento que cumpla con los criterios para ser incluido, se procede a filtrar con las características de exclusión, de modo de que se puedan eliminar los documentos no relevantes.

1.8.1.4.1.2 Definición del criterio de inclusión y exclusión de estudios

Con el objetivo de poder incluir un documento dentro de la colección, se analiza el título del artículo, el resumen que contiene y las palabras clave que lo identifican. Esto permitió realizar el descarte de algunos documentos sin proceder a leer todo el contenido. En cuanto se tenga una idea general de cuáles aportes puede dar el documento, es posible filtrar con los criterios de exclusión.

A fin de poder excluir un documento, se procedió a la lectura completa del mismo, debido a que, aún si el documento es relevante a la investigación, el mismo podría no abordar el enfoque deseado.

1.8.1.4.1.3 Definición de tipos de estudio

Los tipos de estudio que se realizaron fueron los presentes en los artículos de las fuentes que se han seleccionado, luego de aplicar todos los criterios de inclusión y exclusión explicados anteriormente.

1.8.1.5 Ejecución de la revisión

1.8.1.5.1 Ejecución de la selección de la fuente Research Cell: An International Journal of Engineering Sciences

A continuación, se analiza la fuente de Research Cell: An International Journal of Engineering Sciences, una revista académica publicada por Vidya Publications. La misma contiene artículos de ingeniería, tecnologías de la información, entre otros temas.

1.8.1.5.1.1 Selección de estudios iniciales

Tal y como se mencionó anteriormente, la búsqueda de los artículos se realizó con un buscador de artículos gratuito, utilizando la siguiente consulta:

(Data AND Warehouse) OR OLAP OR (OLAP AND Cube)

La misma búsqueda se realiza por título, debido a que el contenido en sí del artículo no se encuentra indexado.

Una vez realizada la búsqueda, luego de verificar los procesos de inclusión y exclusión, se toman los siguientes documentos:

- Data Warehousing and Data Mining in Business Applications - Research Cell:
  An International Journal of Engineering Sciences - Volumen 13, páginas 133 a 137. Autor: Eesha Goel

1.8.1.5.1.2 Evaluación de la calidad de los estudios

Las publicaciones que se han encontrado dentro de esta fuente tienen como característica el hecho de ser revisadas por personal dentro de la misma revista, la cual es publicada dos veces al año.

1.8.1.5.1.3 Revisión de la selección

Los artículos elegidos han sido leídos de manera que se pueda notar el contenido real del artículo, además de buscar posibles referencias en ellos que puedan ser agregadas a la revisión, ya sea de los mismos autores, para

complementar el artículo en sí, o de otros autores que permitan aclarar conceptos de manera más precisa.

1.8.1.5.1.4 Extracción de la información

En esta sección se detalla el proceso de extracción de la información utilizado para conocer los detalles del contenido de los documentos especificados en esta fuente

1.8.1.5.1.4.1 Definición del criterio de inclusión y exclusión de información

Con el propósito de extraer la información que es pertinente para los objetivos planteados en esta investigación, los documentos se analizaron según los aportes y temas de interés para el trabajo. Dichos aportes deben de ser centralizados en el tema de los depósitos de datos, con aportaciones a minería de datos o incluyendo motores de búsqueda.

1.8.1.5.1.4.2 Formulario para la extracción de la información

Con el fin de documentar la extracción de la información, se utiliza un formulario que aclara diversas características del documento, como su título, autores y en qué área está enfocado. La información se almacena en formato tabular, siguiendo el formato de la Tabla 3.

Tabla 3
Formulario de extracción de información

Identificación	**Título** **Publicación** **Autores**
Descripción	Área Resumen
Características Generales	

1.8.1.5.1.4.3 Extracción de resultados objetivos y subjetivos

Tabla 4
Extracción de fuente "Research Cell: An international Journal of Engineering Sciences"

Identificación	**Título: Data Warehousing and Data Mining in Business Applications** **Publicación: Research Cell: An International Journal of Engineering Sciences** **Autores: Eesha Goel**
Descripción	Área: Data Warehouse y Data Mining Resumen: Artículo acerca de los depósitos de datos, minería de datos enfocada a la toma de decisiones de una compañía.
Características Generales	• Explica los conceptos de minería de datos, de manera clara a fin de poder realizar una introducción al tema. Incluye ejemplos reales del uso de ambas tecnologías. • Explicación de cómo implementar un depósito de datos, de manera general. • Provee casos de uso que pueden ser usados para la comparación entre depósitos de datos y clústeres de motores de búsqueda.

1.8.1.5.2 Ejecución de la selección de la fuente Bulletin of the Marathwada Mathematical Society

Ahora se analiza la fuente Bulletin of the Marathwada Mathematical Society, revista que cuenta con publicaciones acerca de almacenes de datos.

1.8.1.5.2.1 Selección de estudios iniciales

Se consultó el motor buscando información relacionada a data warehouse y analytics. Se elige:

- Data Warehousing and mining: an indispensable computational tool for real world problems. Autores: V. Sree Hari Rao y J.V.R Murthy

1.8.1.5.2.2 Evaluación de la calidad de los estudios

Se elige este artículo por varias razones. La primera, contiene términos que ayudan a aclarar conceptos de los depósitos de datos. Segundo, provee algunos casos de uso interesantes al paradigma. Tercero, los autores son de dos disciplinas diferentes, lo que hace interesante ver como un matemático puede observar un depósito de datos para sus utilidades.

1.8.1.5.2.3 Revisión de la selección

Ídem sección 1.8.1.5.1.3

1.8.1.5.2.4 Extracción de la información

En esta sección se detalla el proceso de extracción de la información utilizado para conocer los detalles del contenido de este documento.

1.8.1.5.2.4.1 Definición del criterio de inclusión y exclusión de información

Ídem sección 1.8.1.5.1.4.1

1.8.1.5.2.4.2 Formulario para la extracción de la información

La información se almacena en formato tabular, siguiendo el formato en la Tabla 3.

1.8.1.5.2.4.3 Extracción de resultados objetivos y subjetivos

Tabla 5
Extracción resultados de fuente "Data Warehousing and mining: an indispensable computational tool for real world problems"

Identificación	**Título: Data Warehousing and mining: an indispensable computational tool for real world problems** **Autores: V. Sree Hari Rao y J.V.R Murthy**
Descripción	Área: Data Warehouse y Data Mining Resumen: Artículo que demuestra que un depósito de bases de datos es capaz de ser usado para sistemas de soporte de decisiones.
Características Generales	• Analiza la evolución de las computadoras desde el punto de vista de los depósitos de datos. • Cuenta también con conceptos de OLTP, ERP, y sistemas de soporte de decisión

1.8.1.5.3 Ejecución de la selección de la fuente Google Scholar

Se analizó la fuente Google Scholar, sección meramente académica del motor de búsqueda Google que posee indexada una gran cantidad de material de investigación en todas las áreas.

1.8.1.5.3.1 Selección de estudios iniciales

Se consultó el motor buscando información relacionada a data warehouse y analytics. Se eligió:

- Data Warehousing and Analytics Infrastructure at Facebook - Facebook. Autores: Ashish Thusoo, Zheng Shao, Suresh Anthony, Dhruba Borthakur, Namit Jain, Joydeep Sen Sarma, Raghotham Murthy, Hao Liu.
- Online Analytical Processing (OLAP): A Fast and Effective Data Mining Tool for Gene Expression Databases. Autores: Nadim W. Alkharouf, D. Curtis Jamison y Benjamin F. Matthews1
- Log Analysis as an OLAP application. Autor: Leong Ying Siong Clement

También se consultó el motor en busca de información relacionada a recuperación de información. Se eligió:

- Introduction to Information Retrieval. Autores: Christopher Manning, Prabhakar Raghavan, Hinrich Schütze.
- Search Engines: A Study. Autores: K. Tarakeswar, D Kavitha
- Modern Information Retrieval, 2nd Edition. Capítulo 1: Enterprise Search. Autor: David Hawking
- Real-Time earthquake monitoring using a search engine method. Autores: Jie Zhang, Haijiang Zhang, Enhong Chen, Yi Zheng, Wenhuan Kuang y Xiong Zhang

- Information Retrieval using applied Supervised Learning for Personalized E-Commerce. Autor: Kjell Arne Hellum
- Machine literature searching X. Machine language; factors underlying its design and development. Autores: James W. Perry, Allen Kent, Madeline M. Berry

Finalmente, se buscaron también documentos acerca de computación paralela. Se analizaron los siguientes artículos:

- Parallel Computing. Autores: X. Cai, E. Acklam, H. P. Langtangen y A. Tveito
- Parallel Database Systems: The Future of High Performance Database Processing. Autores: David J. DeWitt, Jim Gray
- Cluster Computing: High Performance, High Availability and High-throughput Processing on a Network of Computers: Autores: Chee Shin Yeo, Rajkumar Buyya, Hossein Pourreza, Rasit Eskicioglu, Peter Graham, Frank Sommers
- The Case of Shared Nothing. Autor: Michael Stonebraker.

1.8.1.5.3.2 Evaluación de la calidad de los estudios

En el caso del artículo relacionado con almacenes de datos, al tratarse de una empresa de renombre en el área de tecnología, se toma como una fuente confiable.

En el caso de la publicación de recuperación de información, sus autores vienen de instituciones importantes. Adicionalmente, fue publicada por Cambridge University Press, parte de una universidad con gran renombre internacional.

1.8.1.5.3.3 Revisión de la selección

Ídem sección 1.8.1.5.1.3

1.8.1.5.3.4 Extracción de la información

En esta sección se detalla el proceso de extracción de la información utilizado para conocer los detalles del contenido de este documento.

1.8.1.5.3.4.1 Definición del criterio de inclusión y exclusión de información

Ídem sección 1.8.1.5.1.4.1

1.8.1.5.3.4.2 Formulario para la extracción de la información

La información se almacena en formato tabular, como en la Tabla 3.

1.8.1.5.3.4.3 Extracción de resultados objetivos y subjetivos

Tabla 6
Extracción resultados de fuente "Data Warehousing and Analytics Infrastructure at Facebook".

Identificación	**Título: Data Warehousing and Analytics Infrastructure at Facebook** **Autores: Ashish Thusoo, Zheng Shao, Suresh Anthony, Dhruba Borthakur, Namit Jain, Joydeep Sen Sarma, Raghotham Murthy, Hao Liu.**
Descripción	Área: Data Warehouse y Data Mining Resumen: Artículo acerca de la infraestructura que usa Facebook para su analítica de datos.
Características Generales	• Explica las generalidades de su almacenamiento y cómo manejan las dificultades que el manejo de tantos datos conlleva. • Expone el uso de Hive internamente y su papel en el descubrimiento de datos y su análisis.

Tabla 7
Extracción resultados de fuente "Online Analytical Processing (OLAP): A Fast and Effective Data Mining Tool for Gene Expression Databases"

Identificación	**Título: Online Analytical Processing (OLAP): A Fast and Effective Data Mining Tool for Gene Expression Databases** **Autores: Nadim W. Alkharouf, D. Curtis Jamison y Benjamin F. Matthews**
Descripción	Área: Data Warehouse y Data Mining Resumen: Caso de uso de tecnologías OLAP para el análisis de datos genéticos.
Características Generales	• Descripción del uso de estas tecnologías para el análisis de datos. • Comparativa del análisis OLAP y consultas SQL.

Tabla 8
Extracción resultados de fuente "Log Analysis as an OLAP Application"

Identificación	**Título: Log Analysis as an OLAP Application** **Autores: Leong Ying Siong Clement**
Descripción	Área: Data Warehouse y Data Mining

	Resumen: Caso de uso de tecnologías OLAP para el análisis de datos de registros de aplicación.
Características Generales	• Descripción del uso de estas tecnologías para el análisis de datos en el ámbito de la seguridad de las aplicaciones. • Comparativa del análisis OLAP y consultas SQL. • Uso de otras herramientas a las propuestas en esta investigación, que contienen los mismos conceptos.

Tabla 9
Extracción resultados de fuente "Introduction to Information Retrieval"

Identificación	**Título: Introduction to Information Retrieval** **Autores: Christopher Manning, Prabhakar Raghavan, Hinrich Schütze**
Descripción	Área: Recuperación de Información. Resumen: Libro de introducción a conceptos de la recuperación de información.
Características Generales	• Introduce los índices y el pre-procesamiento de documentos antes de su uso. • Explica estructuras de búsqueda para diccionarios. • Introduce los pesos para evitar evaluar un documento en estados booleanos.

Tabla 10
Extracción resultados de fuente "Search Engines: A Study"

Identificación	**Título: Search Engines: A study** **Autores: K. Tarakeswar, D Kavitha**
Descripción	Área: Recuperación de Información. Resumen: Artículo que habla acerca de motores de búsqueda en términos generales.
Características Generales	• Se explica la definición de un motor de búsqueda. • Se explica de manera clara los diferentes tipos de motores de búsqueda que existen.

Tabla 11
Extracción resultados de fuente "Modern Information Retrieval, 2nd Edition. Capítulo 1"

Identificación	**Título: Modern information Retrieval, 2nd Edition. Capítulo 1: Enterprise Search.** **Autores: David Hawking**
Descripción	Área: Recuperación de Información. Resumen: capítulo que explica el concepto de la búsqueda empresarial, además de brindar ejemplos de cómo funciona y algunas tareas que pueden realizarse por medio de búsquedas.
Características Generales	• Explica qué es un motor de búsqueda desde el punto de vista empresarial. • Ejemplifica algunas tareas que se pueden realizar con ayuda de motores de búsqueda. • Explica la arquitectura de la implementación básica de una solución de búsqueda.

Tabla 12
Extracción resultados de fuente "Real-time earthquake monitoring using a search engine method"

Identificación	**Título: Real-time earthquake monitoring using a search engine method.** **Autores: Jie Zhang, Haijiang Zhang, Enhong Chen, Yi Zheng, Wenhuan Kuang y Xiong Zhang**
Descripción	Área: Recuperación de Información. Resumen: Caso de uso de análisis de datos utilizando un motor de búsqueda, en este caso, datos históricos de sismos.
Características Generales	• Explica el origen de los datos. • Brinda resultados medibles de la implementación escogida.

Tabla 13
Extracción resultados de fuente " Information Retrieval using applied Supervised Learning for Personalized E-Commerce."

Identificación	**Título: Information Retrieval using applied Supervised Learning for Personalized E-Commerce.** **Autor: James W. Perry Allen Kent Madeline M. Berry**
Descripción	Área: Recuperación de Información. Resumen: Dicta algunas fórmulas usadas para medir el desempeño de un sistema de recuperación textual.
Características Generales	• Explica fundamentos de los sistemas de recuperación de información.

Tabla 14
Extracción resultados de fuente " Information Retrieval using applied Supervised Learning for Personalized E-Commerce"

Identificación	**Título: Machine literature searching X. Machine Language; factors underlying its design and development** **Autor: Kjell Arne Hellum**
Descripción	Área: Recuperación de Información. Resumen: Investigación acerca del uso de motores de búsqueda para analizar el historial de actividad de usuarios para mostrar resultados más relevantes.
Características Generales	• Utiliza datos reales de una tienda virtual. • Definen la facilidad de poder cambiar una consulta, según el historial de un usuario, para así poder mostrarle resultados más relevantes.

Tabla 15
Extracción resultados de fuente "Parallel Computing"

Identificación	**Título: Parallel Computing** **Autores: X. Cai, E. Acklam, H. P. Langtangen y A. Tveito**
Descripción	Área: Computación Paralela. Resumen: Explicación básica de los conceptos de computación paralela y modelos multicomputadora.
Características Generales	• Explica diferentes arquitecturas de hardware para computación paralela. • Introduce el concepto de multicomputadora. • Brinda un ejemplo de tareas que se pueden realizar en computación paralela.

Tabla 16
Extracción resultados de fuente "Parallel Database Systems: The Future of High Performance Database Processing"

Identificación	**Título: Parallel Database Systems: The Future of High Performance Database Processing** **Autores: David J. DeWitt, Jim Gray**
Descripción	Área: Computación Paralela. Resumen: Explicación de implementaciones de bases de datos SQL en arquitecturas paralelas.
Características Generales	• Explica conceptos como particionamiento de datos. • Explica el concepto de las arquitecturas nada compartido y disco compartido. • Explica conceptos y algoritmos para particionamiento de datos.

Tabla 17
Extracción resultados de fuente "Cluster Computing: High Performance, High Availability and High-throughput Processing on a Network of Computers"

Identificación	**Título: Cluster Computing: High Performance, High Availability and High-throughput Processing on a Network of Computers** **Autores: Chee Shin Yeo, Rajkumar Buyya, Hossein Pourreza, Frank Sommers**
Descripción	Área: Computación Paralela. Resumen: Explica la importancia de las redes de comunicación entre las computadoras en un clúster.
Características Generales	• Explicación de las diferentes consideraciones al configurar la comunicación entre diferentes nodos. • Introduce el concepto de imagen única en el clúster. • Explica diferentes modelos de programación para el clúster.

Tabla 18
Extracción resultados de fuente "The case of Shared Nothing"

Identificación	**Título: The Case of Shared Nothing.** **Autores: Michael Stonebraker.**
Descripción	Área: Computación Paralela. Resumen: Introducción y comparación de la arquitectura nada compartido con otras arquitecturas paralelas.
Características Generales	• Explicación de las diferentes arquitecturas, a nivel básico. • Comparación de las diferentes arquitecturas, según las diferentes características que poseen. • Explica de manera detallada las ventajas y desventajas de las arquitecturas nada compartido.

1.8.1.5.4 Ejecución de la selección de la fuente Research Gate

A continuación, se analiza la fuente Research Gate, un repositorio de datos online de artículos científicos de diversos temas, de origen académico.

1.8.1.5.4.1 Selección de estudios iniciales

Se consultó el motor buscando información relacionada a data warehouse y analytics. Se eligió el siguiente artículo:

- Mining, OLAPing, and Mining Biological Data: Towards a Data Warehousing Concept in Biology. Autores: Werner Dubitzky, Olga Krebs, Roland Eils

1.8.1.5.4.2 Evaluación de la calidad de los estudios

El artículo es de índole reciente, además de que su origen es de una empresa dedicada a la biotecnología y al análisis de datos biológicos, por medio de sistemas de bases de datos y minería de datos.

1.8.1.5.4.3 Revisión de la selección

Ídem sección 1.8.1.5.1.3

1.8.1.5.4.4 Extracción de la información

En esta sección se detalla el proceso de extracción de la información utilizado para conocer los detalles del contenido de este documento.

1.8.1.5.4.4.1 Definición del criterio de inclusión y exclusión de información

Ídem sección 1.8.1.5.1.4.1

1.8.1.5.4.4.2 Formulario para la extracción de la información

La información se almacena en formato tabular como es estipulado en la Tabla 3.

1.8.1.5.4.4.3 Extracción de resultados objetivos y subjetivos

Tabla 19
Extracción resultados de fuente "Mining, OLAPing and Mining Biological Data: Towards a Data Warehousing Concept in Biology"

Identificación	**Título: Mining, OLAPing, and Mining Biological Data: Towards a Data Warehousing Concept in Biology** **Autores: Werner Dubitzky, Olga Krebs, Roland Eils**
Descripción	Área: Data Warehouse y Data Mining. Resumen: Se diseña un sistema de almacenes de datos como origen de datos biológicos con el fin de poder realizar análisis al respecto. Se analiza un sistema de minería de datos contra un sistema OLAP para realizar consultas.
Características Generales	• Analiza los conceptos de OLAP y OLTP. • Registra como la minería de datos debe de tener cierta forma de normalización de datos. • Destaca el hecho de que un sistema de base de datos está diseñado para negocios y utilizarlo para otros casos de uso es una tarea compleja.

1.8.1.5.5 Ejecución de la selección de la fuente Elasticsearch

A continuación, se analiza la fuente Elasticsearch, documentación de una de las herramientas usadas en esta investigación.

1.8.1.5.5.1 Selección de estudios iniciales

Se eligieron los siguientes documentos a fin de poder diseñar las consultas necesarias:

- Terms Aggregation: Enlace:

  https://www.elastic.co/guide/en/elasticsearch/reference/current/search-aggregations-bucket-terms-aggregation.html

- Composite Aggregation. Enlace: https://www.elastic.co/guide/en/elasticsearch/reference/current/search-aggregations-bucket-composite-aggregation.html

1.8.1.5.5.2 Evaluación de la calidad de los estudios

La documentación coincide con la versión del producto ofrecida, la cual es además, la última que existe en el mercado.

1.8.1.5.5.3 Revisión de la selección

Ídem sección 1.8.1.5.1.3

1.8.1.5.5.4 Extracción de la información

En esta sección se detalla el proceso de extracción de la información utilizado para conocer los detalles del contenido de este documento.

1.8.1.5.5.4.1 Definición del criterio de inclusión y exclusión de información

Ídem sección 1.8.1.5.1.4.1

1.8.1.5.5.4.2 Formulario para la extracción de la información

La información se almacena en el formato tabular, estipulado en la Tabla 3.

1.8.1.5.5.4.3 Extracción de resultados objetivos y subjetivos

Tabla 20
Extracción resultados de fuente "Terms Aggregation"

Identificación	**Título: Terms Aggregation**
Descripción	Área: Documentación de herramientas. Resumen: Documentación de la agregación conocida como “terms”, la cual agrupa los documentos según los términos encontrados.
Características Generales	• Describe el formato de la agregación. • Brinda ejemplos prácticos de la misma.

Tabla 21
Extracción de fuente “Composite Aggregations”

Identificación	**Título: Composite Aggregations**
Descripción	Área: Documentación de herramientas. Resumen: Documentación de la agregación conocida como “composite”, la cual agrupa los documentos según los campos especificados.
Características Generales	• Describe el formato de la agregación. • Brinda ejemplos prácticos de la misma.

Capítulo 2 Marco Conceptual

A lo largo del tiempo han existido diferentes formas de almacenar los datos. Desde 1960, la tecnología de información y de base de datos ha evolucionado de manera constante: desde archivos sencillos hasta llegar a sistemas de base de datos complejos y sofisticados. A partir de 1970, se han tenido diferentes tipos de base de datos, desde jerárquicas, de red, hasta llegar al modelo relacional que se usa hoy en día, acompañado por técnicas de organización de los datos que se guardan dentro de ellas. Debido a esto, los sistemas gestores de bases de datos relacionales se han convertido en una herramienta eficiente para almacenar, obtener y manejar grandes cantidades de datos en una aplicación. (Hari Rao y Murthy, 2004)

Un concepto que se encuentra relacionado con los sistemas mencionados anteriormente es el de OLTP (*on-line transactional processing*, procesamiento transaccional en línea). Este tipo de sistemas se encuentran optimizados para realizar diversas transacciones sobre los datos, en un tiempo muy corto. Sin embargo, cada uno de estos datos representan interacciones con clientes, como una oportunidad de aprendizaje de los mismos. (Dubitzky, Krebs, & Roland, 2011)

Uno de los retos más grandes de una organización es poder predecir, con un alto grado de confianza, cómo se va a comportar el negocio en algún periodo en el

tiempo, por ejemplo, en el próximo trimestre. Para esto, es necesario tener datos históricos.

Una tecnología que ayuda en el estudio de datos es el almacén de datos. Según Goel (2014, p. 133) un depósito de datos es “un repositorio de bases de datos de negocio, que provee una vista de operaciones actuales e históricas del negocio". Es posible utilizar dicho repositorio para la toma de decisiones, debido a su valor histórico en los datos. Un ambiente de depósito de datos incluye, generalmente, la extracción de una base de datos relacional, una transformación y una carga de datos (ETL por sus siglas en inglés, Extraction, Transformation, Load).

Otros autores dan a los depósitos de datos cuatro características principales. Hari Rao y Murthy (2004) las definen como sigue:

- **Orientadas a un tema:** un almacén de datos está organizado alrededor de un concepto del negocio. Es decir, se engloban diferentes tipos de transacciones de manera que los que toman las decisiones puedan ver los datos más relevantes, eliminando aquellos que no sean significativos.
- **Integrada:** un depósito es construido tomando diferentes fuentes de datos heterogéneas. Los ETLs ayudan en estas operaciones.
- **Variante en el tiempo:** Cada registro almacenado guarda, ya sea implícita o explícitamente, un elemento de tiempo.

- **No volátil:** Solo existen dos operaciones dentro de un depósito de datos, la carga inicial de los registros, y la lectura de estos. En otras palabras, no existe una operación de actualización o borrado de los registros.

Los almacenes de datos tienen los datos requeridos, pero no pueden ser aprovechados sin las herramientas correctas. OLAP (*Online Analytical Processing*) es una tecnología que permite analizar datos de múltiples orígenes, teniendo no solo la transacción en sí, sino el contexto de esta. También, por medio de cálculos complejos, ayuda a los usuarios a visualizar tendencias o estadísticas acerca de los mismos.

Se define la recuperación de la información como “encontrar material (usualmente documentos) de naturaleza no estructurada (usualmente texto) que satisface una necesidad de información de dentro de grandes colecciones (usualmente almacenadas en computadoras)." (Manning, Raghavan, & Schütze, 2008, p. 1)

Bajo este concepto, se introducen diferentes términos: un documento es la unidad básica en la cual el sistema ha sido creado; esto varía de sistema en sistema, por ejemplo, pueden ser las facturas de ventas de una compañía o cada una de sus líneas. Una agrupación de documentos es conocida como una colección, o *corpus.* Un “dato no estructurado” se refiere a un tipo de dato que no sigue normas semánticas, o simplemente, que no es fácil de procesar para una computadora

(Manning, Raghavan, & Schütze, 2008). Adicionalmente, la recuperación de la información se usa para facilitar que los datos semiestructurados puedan ser buscados por los términos que contienen.

No obstante, la recuperación de información no es sólo obtener documentos. Este campo también cubre el procesamiento o exploración de los documentos que han sido considerados como resultados, por ejemplo, realizando agrupaciones por contenido. Finalmente, también es posible realizar una especificación por la escala en que la recuperación funciona. Manning, Raghavan, Schütze (2008) proponen tres diferentes escalas, las cuales son predominantes en la tecnología actual:

- **Búsquedas Web:** El sistema provee resultados a partir de millones de documentos, ubicados en millones de computadoras. Esto conlleva retos en cómo indexar los documentos, y cómo obtenerlos rápidamente, de forma que los usuarios confíen en el sistema.
- **Recuperación de la información personal:** Los últimos sistemas operativos han incluido recuperación de la información, lo que hace posible buscar los documentos en el disco duro. Otro ejemplo de esto son los filtros de correo, que permiten clasificar el correo importante del correo basura.
- **Búsqueda empresarial, institucional y de dominio específico:** Trata de buscar documentos internos, generalmente, almacenados de manera

centralizada, con la intención de que los colaboradores puedan obtener acceso a ellos de distintas maneras.

A su vez, los autores mencionados proponen diferentes maneras de poder atacar el problema de la búsqueda de texto. La más trivial, es realizar una lectura lineal por la colección para identificar cuáles documentos coinciden con lo que se busca. Si bien es cierto, las computadoras actuales tienen la capacidad de almacenar este tipo de información, no resulta enteramente viable si la colección es muy grande o las consultas son complejas. Para poder superar estos problemas, los documentos se *indexan* antes de que puedan ser buscados.

La técnica por la cual se indexan los documentos varía dependiendo del tipo de búsqueda por implementar. Manning, Raghavan, Schütze (2008) proponen el siguiente ejemplo, analizando las obras de Shakespeare una a una, extrayendo los términos usados y generando una matriz de incidencia similar a la Tabla 22, en donde un valor de 1 indica que el término está contenido en una obra.

Tabla 22
Matriz de incidencia para palabras en obras de Shakespeare en inglés

Término	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Anthony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
Mercy	1	0	1	1	1	1
Worser	1	0	1	1	1	0

Nota. Si (x, y) = 1, entonces el término *x* está contenido en la obra *y* (Manning, Raghavan, & Schütze, 2008)

La Tabla 22 permite responder de manera sencilla consultas que involucren saber si un término está o no dentro de un documento. Continuando con el ejemplo de los autores, si se busca algo como *Brutus AND Caesar AND NOT Calpurnia,* se obtendrían los siguientes vectores:

$$110100\,0\ AND\ 110111\ AND\ 101111 = \ 100100$$

Según el resultado de la ecuación anterior, se obtiene como respuesta que los términos están en las obras *Anthony and Cleopatra* y *Hamlet.* A este tipo de búsquedas se les conoce como Modelo de Recuperación Booleano. Si bien es cierto, este tipo de consultas pueden ser útiles, no es muy utilizado en casos reales pues limita mucho el cómo las consultas son respondidas. Además, si se tiene una colección muy grande, la matriz puede ser muy grande como para poder ser almacenada en memoria (Manning, Raghavan, & Schütze, 2008)

Para poder almacenar grandes cantidades de documentos se ha creado una estructura nueva conocida como índice invertido, el cual se compone de un diccionario de términos. Para cada término se guarda una lista de los documentos en los cuales el mismo aparece. Cada elemento de la lista puede también tener más información, como la posición del término en el documento y otros que pueden servir como datos a la hora de realizar búsquedas.

Si se repite la búsqueda hecha anteriormente, *Brutus AND Caesar AND NOT Calpurnia* en un índice invertido, simplemente se deben comparar las dos listas en las cuales los términos *Brutus* y *Caesar* aparecen. Finalmente, se deben sustraer del resultado final los documentos donde aparezca el termino *Calpurnia*.

Por último, una de las ventajas de tener un índice invertido es la capacidad de realizar ordenamientos basados en relevancia. Esto es por medio de dos factores fundamentales, primero, comparando el tamaño de la lista de documentos, lo que indica que tan común es el término. Segundo, cuántas veces un término aparece en cada uno de los documentos, dato que está almacenado en cada uno de los elementos de la lista.

Tarakeswar & Kavitha (2011, p. 29) definen un motor de búsqueda como “un programa que busca en una base de datos, obtiene y reporta la información que contiene los términos especificados o relacionados”. Dichos autores también definen cuatro tipos de motores de búsqueda, enfocados a búsquedas web:

- **Motores de búsqueda basados en rastreadores:** Este tipo de motores cuenta con tres partes: el rastreador, el cual lee los documentos, el índice, el cual registra los documentos, y el programa de búsqueda, el cual organiza los resultados de una consulta.
- **Directorios dependientes de humanos:** Estos motores de búsqueda dependen enteramente de humanos para actualizar sus registros. Las personas envían los contenidos, que luego son revisados por editores. Esto mejora la calidad de los documentos que son retornados como resultados pues existe un filtro manual del contenido. Sin embargo, los cambios no son actualizados hasta que la entrada sea enviada de nuevo.
- **Motores de búsqueda híbridos:** Mezclan los motores de búsquedas con rastreadores y los dependientes de humanos.
- **Metabuscadores:** Este tipo de motores retorna los resultados de otros buscadores.

Como se menciona anteriormente, otro tipo de búsqueda que puede ser implementada es la búsqueda empresarial. Definida como “la aplicación de la tecnología de la recuperación de información para encontrar información dentro de una compañía” (Hawking, 2010). Esto incluye tanto información estructurada como no estructurada, mezclando diferentes fuentes de datos y, opcionalmente, respetando los permisos de acceso que los usuarios tengan sobre la información

almacenada. En otras palabras, se necesita algo más que solo procesar texto e indexarlo para que los usuarios puedan buscarlo.

Varios fabricantes ofrecen motores de búsqueda empresarial, tanto de cobro como de código abierto, que permiten implementar este paradigma de búsqueda en empresas de distintos tamaños.

Hawking (2010) define ciertas tareas que pueden ser respaldadas por motores de búsqueda y su capacidad de centralizar la información. Se citan algunas a continuación:

- **Aprobar solicitudes de viaje de un empleado:** Una búsqueda empresarial puede dar como resultado toda la información que el empleado tiene, con la finalidad de que se pueda analizar si el viaje debe ser aprobado o no.
- **Respuesta de llamadas en un centro de llamadas:** Los operadores pueden buscar los documentos que necesitan al dar soporte, eliminando los que no son relevantes al caso que atienden.
- **Responder en una disputa:** Es posible buscar la información requerida de correos electrónicos u otra evidencia, antes de amonestar a un empleado por una disputa en acción.

- **Escribir una propuesta:** Es posible indexar propuestas anteriores para los trabajos de una compañía, con intención de reutilizar las propuestas anteriores en la creación de una nueva.
- **Obtener patentes:** En caso de que una empresa dependa enteramente de las patentes, es posible indexar una base de datos de patentes, a fin de que se pueda obtener de manera más rápida si existe un producto similar.

Al implementar un sistema de búsqueda, se debe pensar en que la información debe ser almacenada en un índice invertido. Esto generalmente tiene un flujo establecido. Hawking (2010) lo describe a continuación

1. **Obtención**: En esta parte del proceso, tal y como su nombre lo dice, se obtiene la información que se desea buscar. Esta fase presenta diferentes retos, tal como discernir cual versión de un documento usar en caso de que varias versiones existan o incluso, la identificación de los cambios, para evitar que todos los documentos sean procesados cada vez. Además, es necesario que todo programa que genere información sea capaz de proveer una interfaz para la extracción de esta. Agregado a esto, se debe de ser capaz de obtener los permisos de los documentos, si se requiere que los mismos sean respetados a la hora de realizar la búsqueda respectiva.

2. **Extracción**: Una vez que la información se obtiene, es necesario tener un filtrado sobre la misma. Una mala extracción de los datos desde las fuentes puede causar errores a la hora de mostrar los datos, bajando la precisión de estos, generando copias o resultados duplicados, títulos que no son correctos, entre otros. Existen varios problemas en esta etapa, principalmente causados por los tipos de archivos propietarios que no es posible leer, e incluso, archivos que son escaneados, los cuales deben de tener un tratamiento de reconocimiento de texto. Es común que esta etapa es la que tome más tiempo debido a los diferentes procesos que deben de ser ejecutados.

3. **Indexado**: Este proceso utiliza índices invertidos, manejados por el motor de búsqueda escogido. Sin embargo, un reto importante que se debe tomar es cómo manejar los cambios incrementales en una colección. En otras palabras, nuevos documentos deben de ser creados, mientras que los documentos que se borran deben de ser eliminados del índice, o marcados como borrados. Las actualizaciones son simplemente un borrado y una inserción. Otro enfoque para cambios es crear un índice base, e ir creando índices conforme sea necesario para los cambios y mezclar los resultados de estos.

El sistema de búsqueda empresarial también cuenta con mecanismos que permiten procesar las consultas hechas y clasificación de los documentos por

campos específicos, con el propósito de que los documentos que son considerados relevantes sean los primeros obtenidos por los usuarios. Otros tipos de ordenamiento son posibles dependiendo del tipo de documentos u otros factores que pueden ser tomados en cuenta.

Un elemento importante para un motor de búsqueda empresarial es la seguridad. Debido a la naturaleza del sistema es necesario que el motor tenga acceso a toda la información, haciendo que el mismo sea responsable de implementar la seguridad de los documentos que contiene. Dos enfoques son propuestos por Hawking (2010):

- Seguridad a nivel de colección, lograda dividiendo la información en colecciones, cada una con los mismos permisos. Esto hace que cada búsqueda sea realizada a un subconjunto de colecciones.
- Seguridad a nivel de documentos, que filtra todos los documentos con los permisos necesarios, eliminando de los resultados de las búsquedas los documentos que no son visibles.

El éxito de un sistema de recuperación de información se mide con dos valores sobre cada consulta, uno es la precisión y el otro el llamado *recall,* en inglés. Ambos conceptos fueron introducidos por Perry, Kent y Berry (1955).

La precisión se define con la siguiente fórmula

$$precision = \frac{documentosRelevantes \cap documentosRecuperados}{documentosRecuperados}$$

En otras palabras, es la fracción de documentos recuperados que son relevantes para la consulta.

El *recall* se define con esta fórmula:

$$recall = \frac{documentosRelevantes \cap documentosRecuperados}{documentosRelevantes}$$

Resumidamente, es el porcentaje de documentos relevantes que se recuperaron exitosamente de toda la colección de documentos relevantes para una consulta en particular.

Una de las áreas relacionadas con la recuperación de información es la computación paralela. Según Cai (2003) "la computación paralela o concurrente se refiere a un grupo independiente de procesadores trabajando colaborativamente para resolver un problema computacional grande". La computación paralela está motivada por la necesidad de ahorrar tiempo y aprovechar mejor los recursos en la resolución de problemas. Además, esto ayuda a almacenar y procesar gran cantidad de documentos, puesto que el trabajo es dividido de manera que pueda ser ejecutado de manera paralela.

Existen diversas arquitecturas para lograr esto y estas van en un amplio espectro. Cai (2003) las clasifica desde uno de los extremos, que llama

arquitecturas de memoria compartida, las cuales trabajan sobre un conjunto global de memoria de acceso aleatorio. En este espectro, cada aplicación asume que tiene acceso a toda la memoria del sistema, dando la ventaja de ocultar la comunicación entre los procesos y el método de utilizado para dividir el trabajo queda fuera de las responsabilidades de la aplicación. Este método es mejor utilizado en estructuras de datos estáticas, que permiten procesar ciclos en paralelo de manera más eficiente.

En otro extremo, se tienen arquitecturas de *memoria distribuida*, en la que cada procesador tiene memoria propia y no es compartida en los demás nodos de trabajo. Esto permite la implementación de un método de *paso de mensajes.* El sistema debe de ser capaz de dividir un problema grande en varios sub-problemas, cada uno de los cuales es asignado a un nodo para su procesamiento. No es posible acceder a la información de los demás procesos, dado que solo se tiene acceso a la memoria local, únicamente es posible acceder a la información del sub-proceso en sí. No obstante, existen maneras de comunicación por medio de mensajes, los cuales pueden ser usados para pasar datos necesarios o implementar sincronización de procesos.

Aparte de las arquitecturas de memoria, también es posible clasificar los diferentes medios de computación paralela por el tipo de sistemas usados para su implementación. Tales como: computadoras dedicadas a procesamiento paralelo, computadoras de bajo costo, entre otras.

Una forma de poder abstraer el concepto de computación paralela es utilizando el concepto teórico de *Multicomputadora.* Cai (2003) lo define así: “Una multicomputadora tiene P procesadores idénticos, y cada procesador tiene su propia memoria. Cada procesador tiene control de su computación local, que es efectuada de manera secuencial. Los procesadores están conectados por medio de una red de comunicación, así que dos o más procesadores pueden compartir datos enviando y recibiendo mensajes. Finalmente, la comunicación entre dos procesadores es de igual velocidad, sin importar la distancia física.”

Existen también otras formas de clasificación, en este caso, dependiendo de que tanto se comparte el equipo entre los nodos del grupo de procesadores. Esta clasificación es definida por Stonebraker (1986) de la siguiente manera:

- **Memoria compartida:** Múltiples procesadores tienen memoria central común, similar a la clasificación anterior.
- **Disco compartido:** Múltiples procesadores tienen memoria privada pero acceden a una colección de discos de almacenamiento común.
- **Nada compartido:** Los procesadores no comparten ni la memoria ni el almacenamiento.

Las distintas características de cada una de las arquitecturas se enumeran en la Tabla 23.

Tabla 23
Comparación de arquitecturas de computación paralela

Característica del sistema	**Nada compartido**	**Memoria compartida**	**Disco compartido**
Dificultad en control de concurrencia	2	2	3
Dificultad en recuperación de errores	2	1	3
Dificultad de diseño	3	2	2
Dificultad de balance de carga	3	1	2
Dificultad de alta disponibilidad	1	3	2
Número de mensajes	3	1	2
Ancho de banda requerido	1	3	2
Habilidad para escalar a gran número de máquinas	1	3	2
Habilidad de tener grandes distancias entre máquinas	1	3	2
Susceptibilidad a secciones críticas	1	3	2
Número de imágenes del sistema	3	1	3
Susceptibilidad a *hot spots*	3	3	3

Nota. Cada arquitectura se clasifica del 1 al 3, donde 1 significa que es la mejor en una característica. (Stonebreaker – 1986)

Se puede ver que, entre otras cosas, el modelo de memoria compartida no escala bien, es decir, es limitado en el número de máquinas que se le pueden agregar. Por una parte, lo anterior es fundamental, ya que vuelve más complicado tener un sistema capaz de procesar cantidades de datos cada vez más grandes. Por otra parte, el sistema de disco compartido no tiene ventajas respecto a los demás.

Debido a lo mencionado, la mayoría de las soluciones de código abierto son implementadas usando la arquitectura nada compartido. Es importante notar que esta arquitectura es una de las más baratas de implementar, pues cualquier equipo de cómputo puede ser usado para hacerla. Asimismo, de manera general, entre más nodos se agreguen a la red, la capacidad de procesamiento aumenta.

No obstante, también existen desperfectos en las arquitecturas nada compartido que vale la pena señalar. Uno de estos es la cantidad de mensajes, tanto de información como de sincronización, que se deben de enviar entre los nodos de las máquinas. Esto puede afectar el rendimiento de las transacciones si la red en la que están conectadas es lenta en la transmisión de mensajes. Otro defecto es el fenómeno denominado *hot spot*, el cual implica que todo el procesamiento sea realizado por un solo nodo. Esto puede suceder si toda la información relevante a una consulta o transacción no está completamente distribuida.

Para solucionar el problema de la distribución de los datos, existen diferentes maneras de particionarlos. Según DeWitt y Gray (1992), el uso de un particionamiento correcto puede incluso facilitar las operaciones de entrada y salida de los sistemas distribuidos, pues puede permitir que toda la información se escriba en distintos nodos en forma paralela. Existen diferentes maneras de realizar la distribución de los documentos, descritas brevemente a continuación:

- **Round Robin:** En este sistema, el cual es el más simple, se trata de asignar los datos a los nodos de manera secuencial. Si se desea revisar los datos de manera secuencial al realizar una consulta, este sistema puede funcionar. Sin embargo, búsquedas asociativas no se verían beneficiadas.
- **Particionamiento por hash:** Ideal para las aplicaciones que deseen tanto acceso secuencial como asociativo. Se aplica una función de hash a cada

uno de los datos, a fin de poder determinar donde serán ubicados. Con esto, incluso es posible redirigir las consultas a un nodo particular en lugar de consultar todos los nodos a la vez.

- **Particionamiento por rango:** En este sistema se toman valores específicos y se distribuyen por rangos en el grupo de nodos. Esto hace que datos con características similares estén en el mismo nodo. En otras palabras, se implementa un agrupamiento real de los datos. Empero, esto puede causar que todos los datos se guarden en una sola partición, lo que lleva a que las operaciones sean ejecutadas en ella. Esto es particularmente cierto si se utilizan llaves uniformes para los datos.

En las arquitecturas nada compartido es común hablar del término clúster, o grupo. Shin Yeo et al (2006) lo definen como “un tipo de sistema de computación paralela que consiste en varias computadoras interconectas que trabajan como un recurso integrado de computación único”. Otros componentes para un clúster son los sistemas operativos, una red de alto desempeño, ambientes de programación paralela y las aplicaciones en sí.

Una red rápida es necesaria en estos ambientes a fin de poder soportar la comunicación rápida entre los procesadores y así evitar cuellos de botella en el desempeño general del sistema. La decisión del tipo de red a usar debe de ser tomada con muchos factores, entre ellos la compatibilidad, precio y desempeño de la red.

Un factor importante de un clúster es la vista denominada *imagen única del sistema* (SSI por sus siglas en inglés, *Single System Image*) que representa todo el grupo de computadoras como una sola, ocultando a los usuarios la complejidad de implementación de este. Esto puede ser logrado en varios niveles de abstracción: en el equipo, el sistema operativo o las aplicaciones. Independientemente de cómo se realice, se debe tener las siguientes características (Shin Yeo et al, 2006, p 10):

- **Manejabilidad:** Se debe poder administrar los recursos.
- **Estabilidad:** El sistema se debe recuperar de fallos.
- **Desempeño:** Las operaciones deben de ser optimizadas y eficientes.
- **Extensible:** Se debe proveer integración con herramientas propias para extensión del clúster.
- **Escalable:** Se debe poder escalar sin afectar el desempeño.
- **Soporte:** Se debe tener soporte para usuarios y administradores del sistema.
- **Heterogeneidad:** Se debe ser capaz de portar el sistema del clúster a distintas arquitecturas, con el fin de poder soportar componentes heterogéneos.

Capítulo 3 Marco Metodológico

3.1 Tipo de Investigación

La investigación efectuada es de tipo evaluativa pues se comparan capacidades entre dos tecnologías existentes en escenarios específicos. A partir de los resultados obtenidos de los escenarios, se evalúa la tecnología respecto a los indicadores planteados.

3.2 Alcance Investigativo

Las tecnologías que se van a describir durante esta investigación han sido muy estudiadas por varios autores; sin embargo, no han sido comparadas a nivel de uso, por lo que no existe mucha documentación al respecto. Debido a lo anterior, se define el alcance de este estudio como una investigación exploratoria.

Tal y como se menciona en los objetivos, la idea es determinar cuál tecnología funciona mejor según el caso de uso que se necesite. Lo que hace útil realizar una exploración de ambas y documentar los resultados de su comparación.

3.3 Enfoque

Debido a la naturaleza de la investigación, así como la alineación de los objetivos de esta, se escoge un enfoque mixto para poder estudiar las variables asociadas a las tecnologías aplicadas. Lo anterior se debe a que la naturaleza de las comparaciones es meramente cualitativa, como eficiencia, facilidad de

implementación, entre otras. Finalmente, se realizaron mediciones sobre los datos capturados y se compararon los mismos de manera objetiva.

3.4 Diseño

Como se menciona en el apartado anterior, se utilizó un diseño exploratorio secuencial, tomando primeramente las variables cualitativas que interfieren en cada caso de uso y luego realizando mediciones en las mismas. Finalmente, las mediciones encontradas fueron comparadas para determinar cuál tecnología se adecúa mejor al caso de uso en cuestión. La razón principal de lo anterior es poder dar un veredicto que no sea guiado por experiencia anterior o que sea sesgado a la hora de realizar la evaluación respectiva.

3.5 Descripción del proceso de pruebas

Con el objetivo de poder obtener los datos necesarios para las comparaciones que se desean realizar, se comenzó por la creación de una base de datos relacional de pruebas, que contiene diferentes tipos de datos a transformar. Los mismos se almacenaron en un sistema OLAP y en un motor de búsqueda, haciendo que los datos sean equivalentes, pero diferentes en su presentación.

Una vez llevado a cabo este paso, se realizaron diferentes consultas y operaciones con el fin de obtener los datos que deseados. Los datos almacenados dentro de la base de datos raíz son aleatorios, con el fin de que no se puedan hacer optimizaciones a partir de los mismos al descubrir algún patrón específico.

3.6 Ejecución de las pruebas

Con el fin de recolectar los datos, se realizó un sistema de depósito de datos de pruebas, al igual que un clúster de motores de búsqueda de pruebas, el cual tiene la capacidad de realizar consultas comunes hacia los datos que se tienen almacenados. Una vez realizado este paso, se procedió a medir tiempos de respuesta y exactitud que permitirán realizar la comparación esperada. Lo anterior para los diferentes casos de uso observados durante la investigación.

3.7 Técnicas de Análisis de Información

La manera en la cual la información se analizó fue por medio de listas de cotejo. En cada caso de uso, se espera recolectar diferentes variables, las cuales se documentan. Ambos reportes de datos se contrastan, con el objeto de poder observar las fortalezas y debilidades encontradas. Esto incluso puede ser logrado por medio de software especializado que permita la comparación de dichos datos en manera tabular, como hojas de cálculo, para facilitar la visualización.

3.8 Estrategia de Desarrollo de la Propuesta

Para la implementación del clúster y el almacén de datos, se requirieron varias máquinas, todas con características similares para evitar que factores de hardware repercutieran en las mediciones que se efectuarían. Es por ello que se utilizaron servidores de alquiler, para tener a disposición todo el hardware necesario que cumpliera con los requisitos planteados.

Capítulo 4 Análisis del diagnóstico

4.1 Sistemas OLAP

Diversas investigaciones se han realizado empleando las tecnologías OLAP, con el objetivo de conocer las capacidades de resolver los diferentes casos de uso a los cuales son aplicables. A continuación, se describen algunas investigaciones y los resultados obtenidos en cada una de ellas.

4.1.1 Caso de estudio de investigación genética

La investigación, propuesta por Alkharouf, Curtis y Matthews (2005), trata de implementar tecnologías OLAP para realizar comparaciones genéticas. La implementación del sistema fue realizada con el programa Analysis Services, en la versión incluida dentro del gestor de base de datos SQL Server.

En su artículo, Alkharouf, Curtis y Matthews (2005) poseen datos capturados de genes de plantas de soya. El objetivo de la investigación era descubrir el cambio genético de ciertas plantas luego de ser expuestas a una peste particular, nombrada en el artículo como SCN. Con la información de los genes afectados, sería posible obtener plantas resistentes a la enfermedad, con lo cual se evitarían pérdidas en cultivos completos.

Otras técnicas usadas fueron las de agrupamiento de minería de datos, a fin de poder comparar los resultados obtenidos y determinar cuál era la mejor herramienta para ese caso de uso.

4.1.1.1 Implementación

Tal y como se menciona anteriormente, los investigadores utilizaron el producto Analysis Services, distribuido por Microsoft, para la implementación de una solución de tipo OLAP, a fin de encontrar genes candidatos a ser afectados por la enfermedad. Se contaban con 6000 mediciones de genes de distintas plantas de soya, todas en distintas etapas de SCN. Además de una implementación OLAP, también se utilizan algoritmos de agrupamiento y clasificación.

4.1.1.2 Resultados

En la investigación, la herramienta OLAP demostró ser eficiente en el tiempo de retorno de respuestas. Para la medición de tiempos se utilizó una computadora con un procesador Pentium 4 a 1.8GHz y 1 Gb de memoria RAM, equipo aceptable en el tiempo del que se realizó el estudio, es decir, en el año 2005.

Los tiempos obtenidos estaban entre 2 y 5 segundos, este tiempo es descrito como “una fracción del tiempo necesario para crear reportes similares desde una consulta SQL compleja” (Alkharouf, Curtis y Matthews, 2005, p. 184). Se menciona el ejemplo de un reporte tomó 25 segundos por medio de consultas SQL, pero solo tomó un segundo en ejecución a la hora de utilizar el modelo multidimensional OLAP diseñado, además de encontrar más información relevante respecto a los demás algoritmos utilizados.

Finalmente, los autores cierran la investigación destacando la facilidad de implementación del modelo, dadas las herramientas correctas, la gran compatibilidad de diferentes bases de datos, como Oracle, Sybase, MySQL, y la capacidad de usar programas populares, como Microsoft Excel, para poder realizar la visualización de los datos.

4.1.2 Análisis de registros de aplicación

Una de las herramientas más útiles a la hora de realizar el manejo de seguridad de una aplicación es el análisis de los registros del sistema. Sin embargo, la cantidad de registros puede ser muy voluminosa. Según Leong Ying (2003), el problema del manejo de altos volúmenes de información fue solucionado con sistemas OLAP.

En su artículo, Ying utiliza clientes y sistemas OLAP que permiten la visualización de la información referente a bitácoras de una información. Como demostración, utiliza el cliente Seagate Analysis con registros de bitácora de eventos de un sistema Windows NT.

Como acotación, Ying también aclara que una precondición importante es tener los registros en una base de datos para su manejo, aun cuando también acota que lenguajes SQL no son exactamente los más útiles para el manejo de dicha información, principalmente, porque no es muy intuitivo. Finalmente, menciona que es posible realizar una visualización grafica de los tipos de eventos registrados.

4.1.2.1 Implementación

Como se menciona anteriormente el autor utiliza el programa Seagate Analysis a fin de realizar el análisis OLAP requerido para la implementación del caso de estudio. El programa es utilizado con registros del sistema del sistema operativo de Windows NT, analizando ciertos campos determinados.

Un detalle importante acerca de esta implementación es que los tipos de registros no cumplen con una estructura de hechos y dimensiones. En este caso, ningún campo cumple con las características necesarias para ser considerado una medida. Como forma de evitar errores en los reportes, Ying recomienda escoger un campo que se pueda contar, por ejemplo, la fecha en la que el registro se generó. Esto hará que el sistema realice un conteo de registros por fecha, lo que causará la creación de una medición.

4.1.2.2 Resultados

En esta implementación, no se realizan comparaciones de desempeño contra otras tecnologías. Sin embargo, el enfoque del autor señala distintos aspectos de las herramientas.

Primeramente, las herramientas OLAP permiten ahorrar tiempo de análisis de los registros, pues el tiempo de realizar la misma consulta para definir la información necesaria se reduce. Además, diversas tareas pueden ser realizadas por estas tecnologías. Ying menciona las siguientes:

1. Monitoreo de cantidades de registros. Si se denotan ciertas repeticiones de registros, es posible generar una alerta.

2. Graficar medidas a través del tiempo.

3. Capacidad de analizar diversos parámetros al mismo tiempo, con ayuda de visualizaciones multidimensionales.

4. El avance de los conceptos OLAP hace que sea posible que usuarios sin conocimiento técnico utilicen este tipo de herramientas.

4.2 Sistemas de recuperación de información

Diversos esfuerzos se han realizado para analizar datos y realizar predicciones a partir de ellos utilizando herramientas alternativas. En esta sección, se definen algunos casos de estudio que fueron implementados con los conceptos de recuperación de información.

4.2.1 Caso de estudio de análisis de movimientos sísmicos.

Durante la investigación propuesta por Jie Zhang, Haijing Zhang, Enhong Chen et al (2014), se intenta utilizar un enfoque de recuperación de información alimentado por datos de sismos, con el objetivo de poder predecir en el menor tiempo posible, toda la información del evento e inclusive realizar alertas al público de posibles zonas afectadas.

En la actualidad, según Zhang et al, existen diferentes tipos de algoritmos y sistemas que son capaces de alertar eventos telúricos con 5 segundos de

antelación, basados en una red de monitoreo densa. Sin embargo, el problema de dichos sistemas es que se toma mucho tiempo en poder analizar toda la información necesaria. En otras palabras, el reto de esta investigación “está en una estimación rápida y automática de la fuente del temblor, en unos segundos después de recibir los datos de diferentes estaciones sísmicas” (Zhan et al, 2014).

4.2.1.1 Implementación.

Para la implementación del sistema propuesto, los autores utilizaron imágenes que representaban los datos de diferentes movimientos sísmicos registrados. Dichos datos contenían, entre otros detalles, la fuente del sismo y el medio por el cual se propagó. A fin de poder obtener los resultados, los autores implementaron un “Sistema de búsqueda de temblores basado en imágenes, similar a los motores de búsqueda web” (Zhan et al, 2014). Dichos datos son procesados, a fin de encontrar los mejores resultados para nuevos sismos, sin necesidad de configuración o intervención humana.

4.2.1.2 Resultados

En los datos de prueba, los cuales cubrían diferentes eventos en un área controlada, se encontró que el resultado que fue considerado más relevante predijo la fuente y la profundidad del evento, con una diferencia de 15 km en el plano horizontal, y 0.6 km de profundidad. Además, los primeros 200 resultados mostraron diferencias de 25 km en el plano horizontal y 5 km en la profundidad del sismo.

Respecto a los tiempos de respuesta, durante la investigación se demuestra que es posible realizar un reporte del sismo en menos de un segundo después de que el evento fue recibido. Además, no se retorna un solo resultado, sino que se retornan varios, lo que permite verificar la solución en tiempos menores a un segundo.

4.2.2 Análisis de registros para comercio electrónico personalizado.

En esta investigación, propuesta por Arne Hellum, se analizan diferentes herramientas de la disciplina de la recuperación de la información, analizando registros de compras y de navegación para predecir patrones de uso de los clientes, con el objetivo de poder mostrar productos relevantes en la tienda que implemente estas técnicas.

4.2.2.1 Implementación

Para esta investigación, Hellum (2017) utiliza una colección de 900 megabytes de texto, conteniendo registros de navegación, historia de transacciones de clientes, productos, y otros metadatos. Entre los datos analizados por el autor se encuentran:

- Productos: Datos como su nombre y el precio de cada producto
- Categorías de producto
- Registros de clic: cada interacción que el usuario haga con los productos, incluyendo el registro fuente

- Registro de vistas: cada vez que un producto fue visto por un cliente
- Registro de compras
- Registro de consultas: si un usuario realiza múltiples consultas, este registro refleja el orden en el que las mismas fueron realizadas.

Diferentes aspectos de cada uno de los registros son analizados de manera que puedan ser comparados, formando lo que el autor denomina como “indicador de relevancia” (Hellum, 2017).

Finalmente, luego de calcular y normalizar dichos indicadores, se procede a entrenar modelos, a fin de poder modificar la relevancia de los documentos retornados.

4.2.2.2 Resultados

El estudio muestra que diferentes formas de analizar los datos pueden cambiar drásticamente la forma en que los resultados son mostrados. Hellum concluye que los registros de clics tienen un mejor desempeño respecto a los demás registros, tomando en cuenta incluso que los registros de ventas, aun cuando dan buenos aspectos de relevancia, no aportan un cambio tan significativo.

Una de las maneras de utilizar esta información es integrarla dentro de los algoritmos de relevancia de un motor de búsqueda. Esto hace que cada persona tenga una experiencia propia.

Capítulo 5 Hallazgos Preliminares

Como parte de la investigación, se implementaron consultas y diseños tanto al motor de búsqueda como al depósito de datos, con el fin de encontrar resultados comparables entre ambas tecnologías. Los resultados finales serán descritos en el Capítulo 6; sin embargo, se procede a discutir las diferentes implementaciones realizadas como parte del proceso investigativo utilizado.

5.1 Exactitud

5.1.1 Motor de búsqueda

Durante la ejecución de las consultas, se encontraron diferencias respecto a la exactitud de estas, debido a que el motor de búsqueda retornaba cantidades menores que las retornadas en el cubo OLAP. Al analizar los datos, se encuentra que, en una agregación de tipo Terms “las cuentas de los documentos (y de cualquier sub-agregación) en la agregación Terms no siempre es preciso” (Terms Aggregation, n.d.). Para evitar cualquier diferencia entre los resultados, se elige utilizar otra forma de agregaciones más precisas. Siguiendo las recomendaciones de la misma documentación, se utilizan la agregación denominada “Composite”. Esta garantiza que todos los elementos son capturados de manera eficiente por el motor permitiendo una mejor exactitud de los resultados. (Terms Aggregation, n.d.).

5.1.2 Almacén de datos

Al diseñar el cubo OLAP se utilizó un miembro calculado para obtener el total de ventas a partir de la multiplicación de otras dos medidas: precio unitario y cantidad. Esto sin embargo presentó problemas en la exactitud de resultados.

La causa de la falta de exactitud en los resultados se debió a que el miembro se calcula en tiempo de ejecución, es decir, al momento de realizar la consulta. No obstante, esto causa que antes de efectuar la multiplicación de las dos medidas se agreguen ambas y el resultado de la operación sea mucho mayor que el real.

Lo anterior fue solucionado usando una medida derivada, la cual es computada cuando el cubo se procesa, esto es, en pre cálculo. En la Figura 1 se puede observar la diferencia entre resultados, la columna *sales* corresponde al miembro calculado y la columna *sales calc* es la medida derivada. Se puede observar que el miembro calculado toma las agregaciones de las columnas *price* y *quantity* y las multiplica. Al comparar ambas columnas de resultado de ventas se nota la diferencia.

Name	Sales Calc	sales	Price	Quantity
Abbey Kivelle	308433	3461974916	56263	61532
Abby Flowers	304995	3422763536	55586	61576
Abdul Bardsley	307730	3459063970	55886	61895
Abey Screen	304978	3403645104	55824	60971
Abraham Niece	306121	3412361876	55604	61369
Ada Proughten	306124	3383724400	55690	60760

Figura 1. Cálculos con miembro calculado vs medida derivada

Capítulo 6 Propuesta de Solución

6.1 Diseños de la implementación

A continuación, se describen los modelos que se utilizaron para hacer las pruebas. Se detallan en las secciones a continuación el diseño del depósito de bases de datos, el diseño del cubo OLAP y el diseño del esquema del motor de búsqueda seleccionado.

6.1.1 Depósito de datos

Para la implementación del depósito de datos se realiza una base de datos dimensional, con formato estrella, que detalla los diferentes hechos y dimensiones necesarias para el proyecto. Cabe destacar que se utilizaron alrededor de diez millones de registros a fin de poder realizar las pruebas necesarias. Dicho modelo consta de las siguientes tablas:

6.1.1.1 Dimensiones del modelo

6.1.1.1.1 Álbum

Contiene información de identificación de una película en particular. Es decir, solo tiene un nombre y un identificador único en el sistema. Los campos se muestran en la Figura 2.

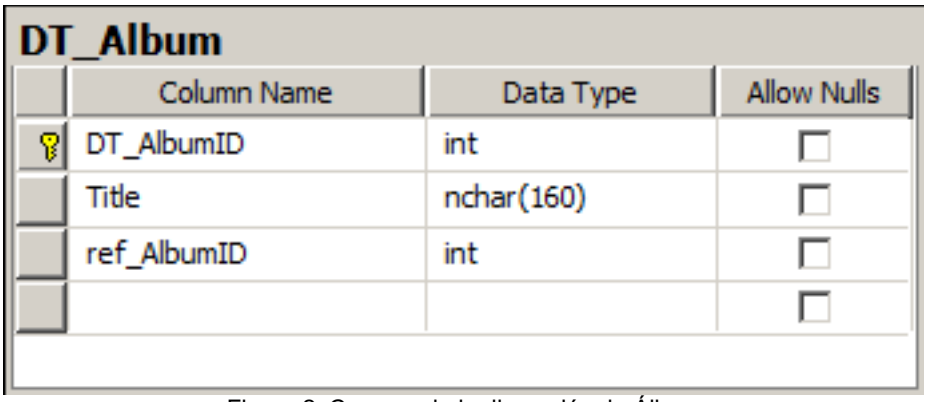


Figura 2. Campos de la dimensión de Álbum

6.1.1.1.2 Artista

Esta dimensión refleja los datos de un artista dentro del sistema, a fin de poder contextualizar las ventas de una película en particular. Los campos se muestran en la Figura 3.

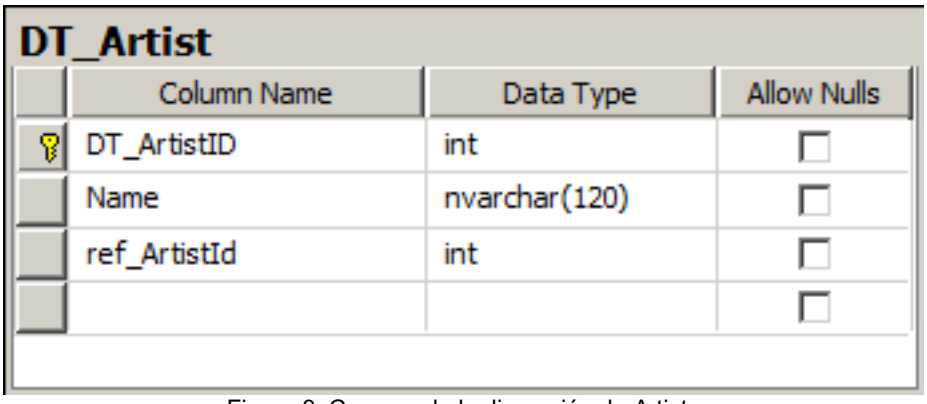
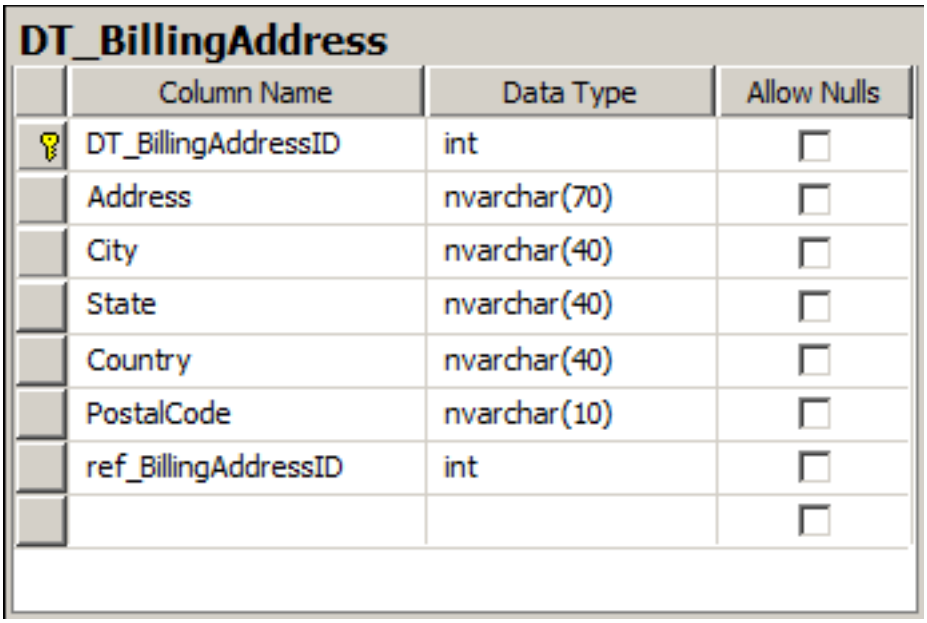


Figura 3. Campos de la dimensión de Artista

6.1.1.1.3 Dirección de cobro

Esta dimensión especifica la dirección en la que un cobro debe de ser enviado, según la factura que documenta una compra. Contiene todos los campos que se encuentran en la factura que son relacionados a la dirección. Los campos se muestran en la Figura 4.



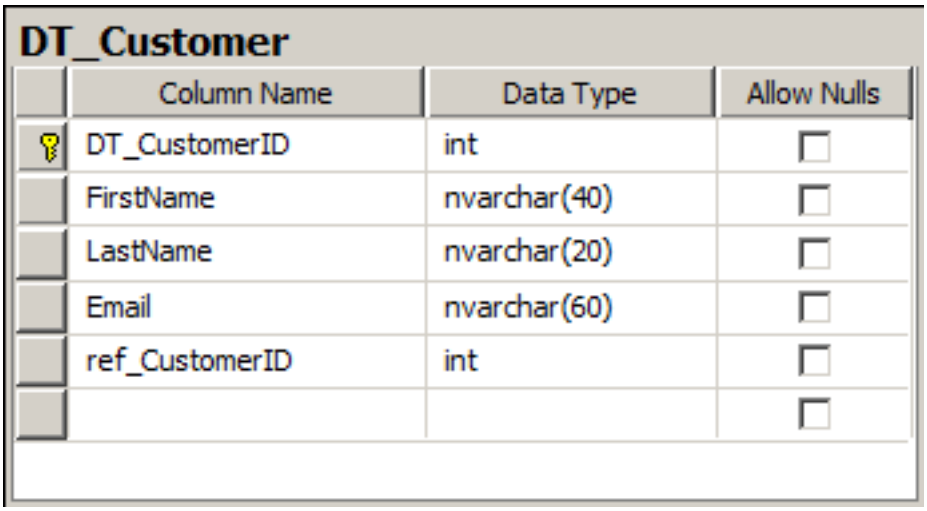
DT_BillingAddress

Column Name	Data Type	Allow Nulls
DT_BillingAddressID	int	☐
Address	nvarchar(70)	☐
City	nvarchar(40)	☐
State	nvarchar(40)	☐
Country	nvarchar(40)	☐
PostalCode	nvarchar(10)	☐
ref_BillingAddressID	int	☐
		☐

Figura 4. Campos de la dimensión de cobro

6.1.1.1.4 Cliente

Representa los datos que contextualizan un cliente. Campos de dirección no fueron incluidos, pues se encuentran en otra dimensión que se describirá en futuras secciones del documento. Los detalles se pueden observar en la Figura 5.



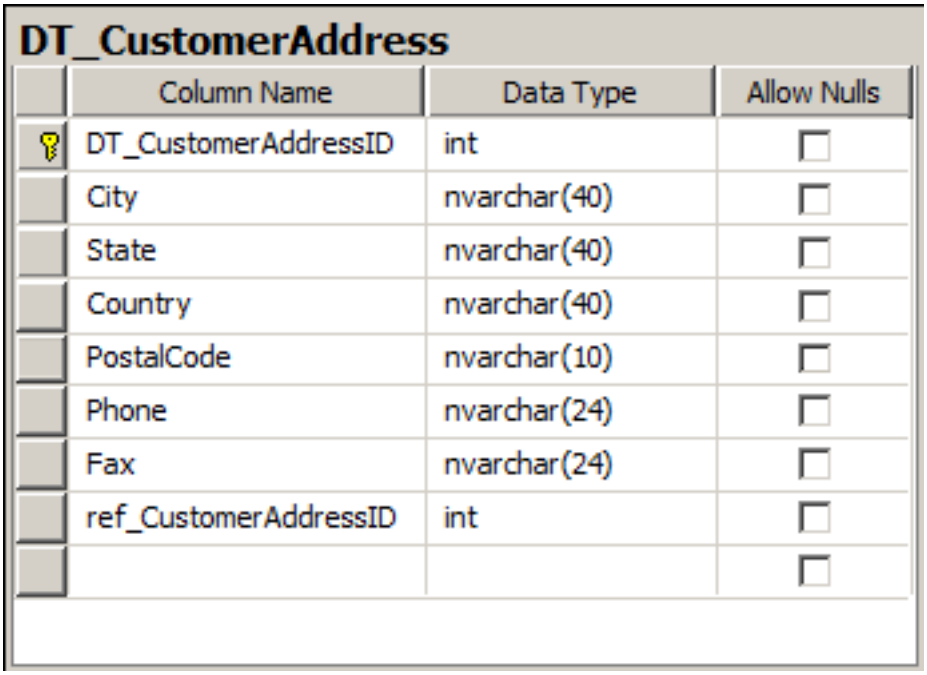
DT_Customer

Column Name	Data Type	Allow Nulls
DT_CustomerID	int	☐
FirstName	nvarchar(40)	☐
LastName	nvarchar(20)	☐
Email	nvarchar(60)	☐
ref_CustomerID	int	☐
		☐

Figura 5. Campos y tipos de la dimensión de Cliente

6.1.1.1.5 Dirección del cliente

Similar a una dirección de cobro, sin embargo, esta tiene datos que la atan a los clientes. Se decide separar por claridad del diseño, además de que se facilita el mantenimiento de la dimensión en caso de cambios. Al realizar una separación es posible cambiar la dirección de un cliente, sin que afecte su dirección de cobro. Los campos de la tabla se especifican en la Figura 6.



DT_CustomerAddress

Column Name	Data Type	Allow Nulls
DT_CustomerAddressID	int	☐
City	nvarchar(40)	☐
State	nvarchar(40)	☐
Country	nvarchar(40)	☐
PostalCode	nvarchar(10)	☐
Phone	nvarchar(24)	☐
Fax	nvarchar(24)	☐
ref_CustomerAddressID	int	☐
		☐

Figura 6. Campos y tipos de la dimensión dirección de clientes

6.1.1.1.6 Fecha

Esta dimensión representa una fecha, así como múltiples datos de esta. Sus campos se especifican en la Figura 7.

DT_Date

Column Name	Data Type	Allow Nulls
DT_DateID	int	☐
Date	datetime	☑
FullDateUK	char(10)	☑
FullDateUSA	char(10)	☑
DayOfMonth	varchar(2)	☑
DaySuffix	varchar(4)	☑
DayName	varchar(9)	☑
DayOfWeekUSA	char(1)	☑
DayOfWeekUK	char(1)	☑
DayOfWeekInMonth	varchar(2)	☑
DayOfWeekInYear	varchar(2)	☑
DayOfQuarter	varchar(3)	☑
DayOfYear	varchar(3)	☑
WeekOfMonth	varchar(1)	☑
WeekOfQuarter	varchar(2)	☑
WeekOfYear	varchar(2)	☑
Month	varchar(2)	☑
MonthName	varchar(9)	☑
MonthOfQuarter	varchar(2)	☑
Quarter	char(1)	☑
QuarterName	varchar(9)	☑
Year	char(4)	☑
YearName	char(7)	☑
MonthYear	char(10)	☑
MMYYYY	char(6)	☑
FirstDayOfMonth	date	☑
LastDayOfMonth	date	☑
FirstDayOfQuarter	date	☑
LastDayOfQuarter	date	☑
FirstDayOfYear	date	☑
LastDayOfYear	date	☑
IsWeekday	bit	☑
		☐

Figura 7. Campos y tipos de la dimensión de fechas

6.1.1.1.7 Género

Representa la información de un género de películas en particular, el cual va asociada a una película. Sus campos se encuentran en la Figura 8.

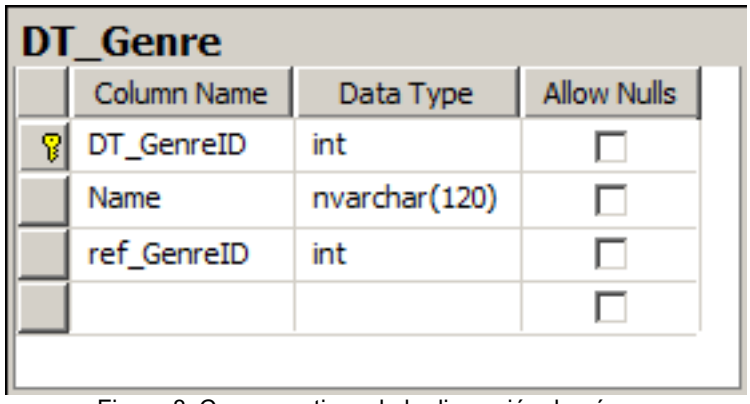


Figura 8. Campos y tipos de la dimensión de género

6.1.1.1.8 Tipo de media

Indica el tipo de archivo de una película, o el tipo de medio para ser reproducida. Los campos se presentan en la Figura 9.

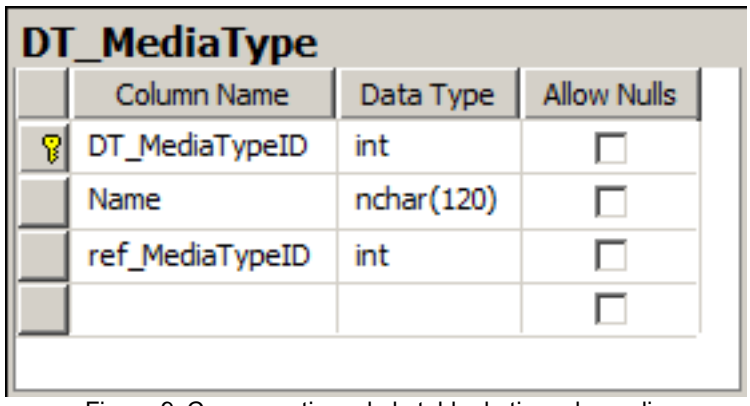


Figura 9. Campos y tipos de la tabla de tipos de media

6.1.1.1.9 Pista

Almacena la información de un idioma de una película en particular, además de datos de la película. Los tipos de datos y nombres de los campos se muestran en la Figura 10.

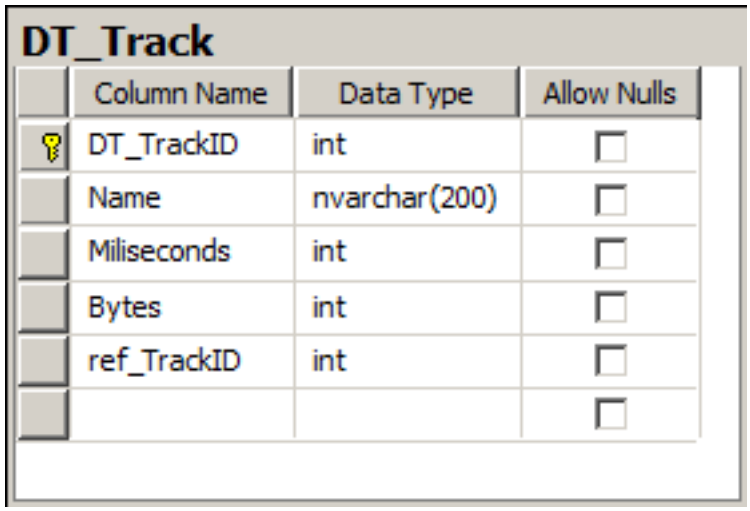
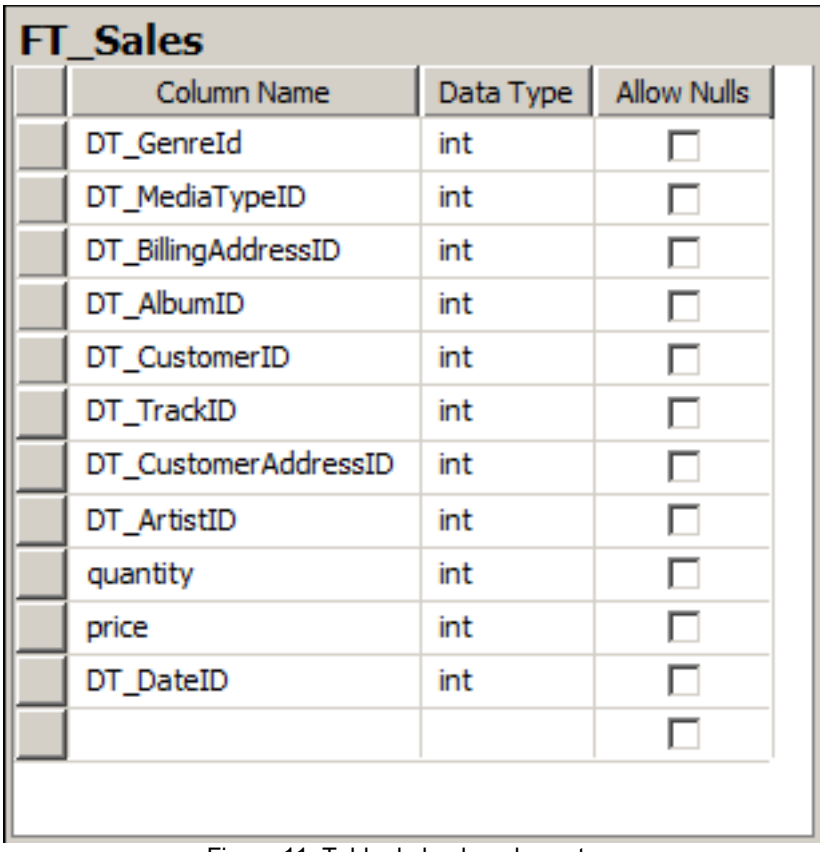


Figura 10. Campos de la dimensión de Pista

6.1.1.2 Tabla de Hechos

El depósito utilizado para este diseño tiene solamente una tabla de hechos, la cual une todas las dimensiones, consolidando una línea de factura en una compra. El diseño de la dimensión se muestra en la Figura 11.



FT_Sales

	Column Name	Data Type	Allow Nulls
	DT_GenreId	int	☐
	DT_MediaTypeID	int	☐
	DT_BillingAddressID	int	☐
	DT_AlbumID	int	☐
	DT_CustomerID	int	☐
	DT_TrackID	int	☐
	DT_CustomerAddressID	int	☐
	DT_ArtistID	int	☐
	quantity	int	☐
	price	int	☐
	DT_DateID	int	☐
			☐

Figura 11. Tabla de hechos de ventas.

6.1.2 Cubo OLAP

Usando como base el depósito de datos descrito en 6.1.1 se creó un cubo OLAP para que lleve a cabo las agregaciones y cálculos requeridos para los diferentes escenarios planteados. Es importante destacar que se implementó una medida derivada directamente en el cubo para ayudar en el cálculo de ventas, tal como fue descrito en 5.1.2

En la Figura 12 se puede observar el esquema completo del cubo final creado. La figura también muestra que cada dimensión creada para el modelo del depósito de datos es reflejada en el cubo, haciendo más fácil el análisis del mismo.

Finalmente, todos los atributos de las dimensiones son también atributos del cubo, lo que permite hacer consultas de manera más sencilla.

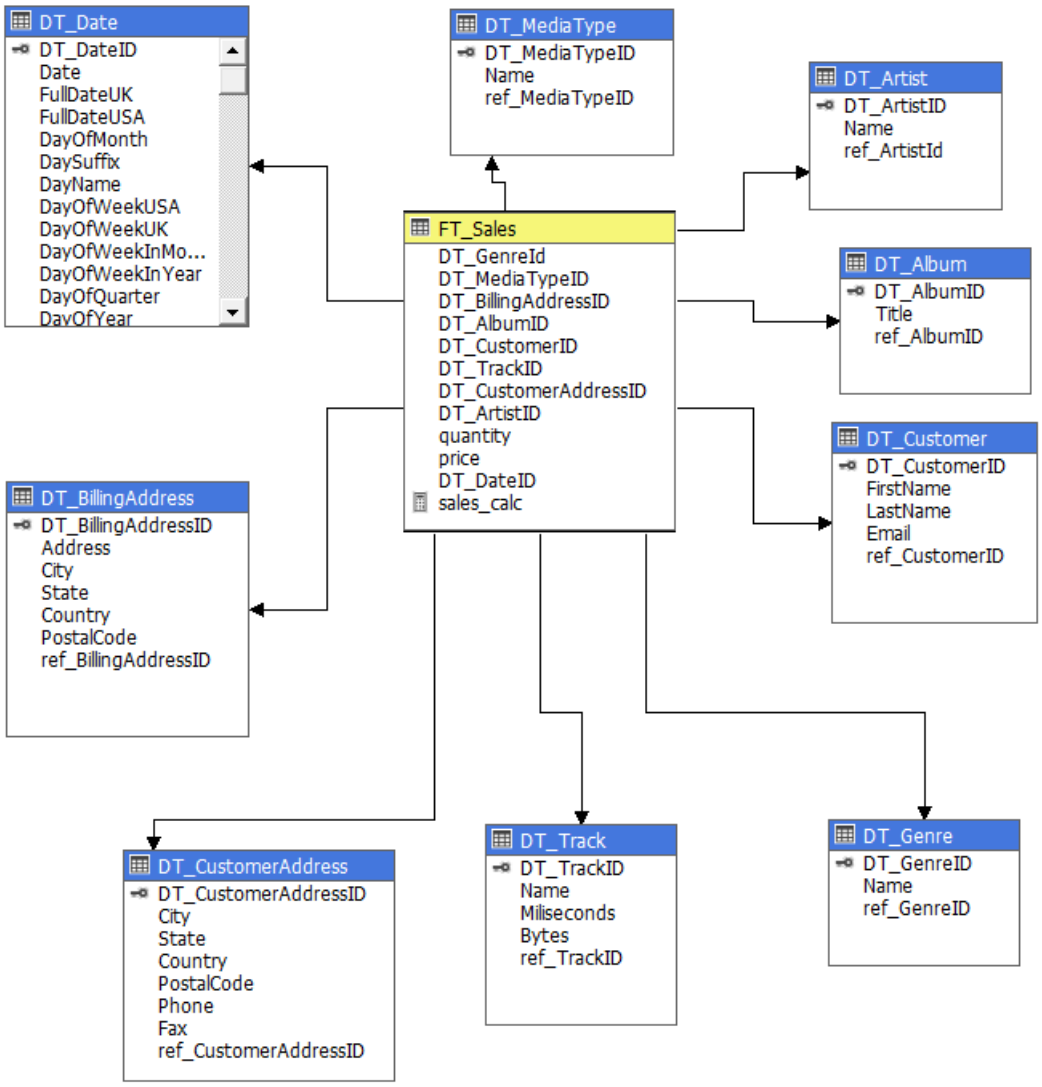


Figura 12. Esquema completo cubo OLAP.

En el caso de las fechas y las direcciones se crearon dos jerarquías, para que fuera posible realizar casos de uso con navegación de los datos. Las jerarquías creadas fueron las siguientes:

1. Fecha.
    a. Año.
    b. Mes.
    c. Día.
2. Región.
    a. País.
    b. Estado.
    c. Ciudad.
    d. Dirección.
    e. Código Postal.

6.1.3 Esquema del motor de búsqueda

En el caso de un motor de búsqueda, se realiza un modelo por documento, a fin de que los campos puedan ser indexados de la mejor manera. Para la investigación, se tomó como un documento cada tupla generada por la tabla de hechos. El diseño de cada uno de los campos que fueron indexados se muestra en la Tabla 24 Como nota general, cada campo de texto fue duplicado con el fin de que se puedan realizar conteos de manera más precisa.

Tabla 24
Especificación del esquema del motor de búsqueda.

Nombre del campo	Tipo
AlbumTitle	string
AlbumTitle.keyword	string
ArtistName	string
ArtistName.keyword	string
BillingAddress	string
BillingAddress.keyword	string
BillingCity	string
BillingCity.keyword	string
BillingCountry	string
BillingCountry.keyword	string
BillingPostalCode	string
BillingPostalCode.keyword	string
BillingState	string
BillingState.keyword	string
CustomerEmail	string
CustmerEmail.keyword	string
CustomerCity	string
CustomerCity.keyword	string
CustomerCountry	string
CustomerCountry.keyword	string
CustomerFax	string
CustomerFax.keyword	string
CustomerFirstName	string
CustomerFirstName.keyword	string
CustomerLastName	string
CustomerLastName.keyword	string
CustomerPhone	string
CustomerPhone.keyword	string
CustomerPostalCode	string
CustomerPostalCode.keyword	string
CustomerState	string
CustomerState.keyword	string
Date	date
DurationMiliseconds	number
Genre	string
Genre.keyword	string
IsWeekday	boolean
MediaType	string
MediaType.keyword	string

SizeInBytes	number
TrackName	string
TrackName.keyword	string
price	number
quantity	Number
sales_pre	Number

6.2 Configuración y Herramientas

En esta sección se describe el equipo utilizado para las pruebas. En este caso, se eligieron diferentes proveedores de servicios en la nube: para el almacén de datos se eligió Microsoft Azure, mientras que para el clúster fue escogido Elastic Cloud. En ambos casos siempre intentando utilizar especificaciones similares entre los proveedores seleccionados.

6.2.1 Depósito de datos

El almacén de datos fue alojado en una máquina virtual provista en el servicio en la nube de Microsoft Azure. La máquina virtual tenía instalado el sistema operativo Windows Server 2016. Además de esto, se instaló el motor de base de datos SQL Server 2017 Developer Edition. Respecto a las especificaciones de la máquina, el sistema contaba con el siguiente equipo:

- 2 CPUs virtuales.
- Disco de estado sólido de 16 gigabytes para el sistema operativo.
- Disco duro de estado sólido 1 terabyte para el almacenamiento de las bases de datos.

- 8GB de memoria RAM.

Para la recopilación de tiempos de los distintos escenarios, se crearon scripts de Powershell para cada escenario y los tiempos fueron capturados usando SQL Profiler. Esta herramienta permitió almacenar los resultados de cada iteración del script en una base de datos en SQL Server. Una vez que se tenían estos datos, los mismos fueron agregados para así poder comparar los resultados obtenidos.

6.2.2 Clúster de motores de búsqueda

Para la implementación del clúster del motor de búsqueda, se eligió el proveedor Elastic Cloud. Este brinda el servicio del motor de búsqueda Elasticsearch 6.2.2 y el servicio de monitoreo y visualización de datos Kibana. Las máquinas virtuales fueron provisionadas en la plataforma Amazon Web Services.

En total, se contaba con las siguientes configuraciones:

- 4GB de memoria RAM.
- 96GB de almacenamiento, utilizando discos duros de estado sólido.

A la hora de la realización de las pruebas, se configuró una ejecución automática de las diversas consultas generadas por las visualizaciones. Dado que, como parte de la respuesta, el motor de búsqueda devuelve el tiempo de ejecución, se captura esta duración y se normaliza para poder ser graficado.

6.3 Escenarios

En la presente sección, se detallan los diferentes casos de uso que se ejecutaron. El proceso para la generación de las ejecuciones automáticas fue el siguiente:

1. Se realiza una visualización con los clientes respectivos. Microsoft Excel para el cubo OLAP y el visualizador Kibana para el motor de búsqueda.

2. Se extrajo la consulta generada por cada una de las herramientas. En el caso del motor de búsqueda, debido a errores en las consultas generadas que afectaban la exactitud, se diseñaron las consultas por separado.

3. Se utilizaron herramientas para la ejecución automática, además de monitorear el uso de recursos durante la ejecución de las diferentes consultas.

Las consultas correspondientes a cada escenario fueron ejecutadas mil veces en cada una de las tecnologías planteadas en este trabajo. Antes de cada corrida individual, se limpiaron las cachés, con el objetivo de que la duración fuera más precisa. En las siguientes secciones se detallan los diferentes resultados a comparar, así como la exactitud de estos. En el ámbito de precisión, se limita la cantidad de elementos a retornar a diez, pues solo es necesaria una muestra para determinar si los resultados son correctos.

6.3.1 Ventas por Fecha

Este escenario muestra los resultados de calcular las ventas por una jerarquía de tiempo en el sistema. Las ventas no son divididas por ningún campo, salvo día, mes y año.

6.3.1.1 Motor de búsqueda

6.3.1.1.1 Exactitud

Con el fin de poder comparar los resultados de consulta, se procede a dividir la misma en cada una de sus partes.

6.3.1.1.1.1 Ventas por año

Los resultados obtenidos en venta por año se describen en la Tabla 25.

Tabla 25
Resultados de consulta ventas por fecha, elemento año

Fecha (timestamp)	Fecha (formato ISO)	Valor
1451606400000	Fri Jan 01 2016 00:00:00	153253033
1483228800000	Sun Jan 01 2017 00:00:00	152921228

6.3.1.1.1.2 Ventas por Mes

Los resultados obtenidos para esta consulta se muestran en la Tabla 26.

Tabla 26
Resultados de consulta ventas por fecha, dividida en meses

Fecha (timestamp)	Fecha (Formato ISO)	Valor
1451606400000	Fri Jan 01 2016 00:00:00	13009432
1454284800000	Mon Feb 01 2016 00:00:00	12160571
1456790400000	Tue Mar 01 2016 00:00:00	12981945
1459468800000	Fri Apr 01 2016 00:00:00	12567499
1462060800000	Sun May 01 2016 00:00:00	12954631
1464739200000	Wed Jun 01 2016 00:00:00	12532477
1467331200000	Fri Jul 01 2016 00:00:00	12967114
1470009600000	Mon Aug 01 2016 00:00:00	13004483
1472688000000	Thu Sep 01 2016 00:00:00	12572791
1475280000000	Sat Oct 01 2016 00:00:00	13001729

6.3.1.1.1.3 Ventas por día

Finalmente, una muestra de 10 resultados de las ventas por día se presenta en la Tabla 27.

Tabla 27
Resultados de consulta ventas por fecha, dividida por día

Fecha (timestamp)	Fecha (formato ISO)	Valor
1451606400000	Fri Jan 01 2016 00:00:00	416943
1451692800000	Sat Jan 02 2016 00:00:00	414790
1451779200000	Sun Jan 03 2016 00:00:00	425883
1451865600000	Mon Jan 04 2016 00:00:00	418447
1451952000000	Tue Jan 05 2016 00:00:00	422032
1452038400000	Wed Jan 06 2016 00:00:00	420359
1452124800000	Thu Jan 07 2016 00:00:00	419747
1452211200000	Fri Jan 08 2016 00:00:00	429440
1452297600000	Sat Jan 09 2016 00:00:00	421186
1452384000000	Sun Jan 10 2016 00:00:00	414025

6.3.1.1.2 Desempeño

6.3.1.1.2.1 Duración de la consulta

Durante las pruebas, se obtuvieron los siguientes datos:

- Duración máxima: 125841milisegundos
- Duración mínima: 4155 milisegundos
- Duración promedio: 15694.469 milisegundos

Los resultados se muestran en el Gráfico 1.

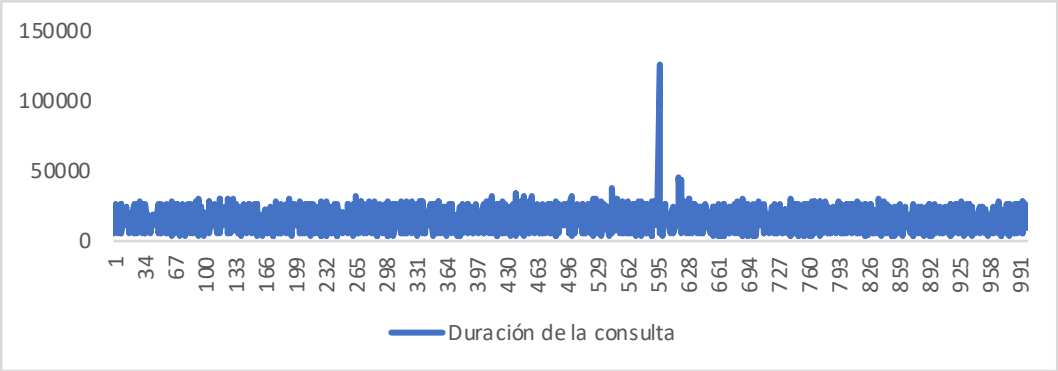


Gráfico 1. Duración de consultas en motor de búsqueda para caso ventas por año, mes y día.

6.3.1.1.2.2 Utilización de recursos

6.3.1.1.2.2.1Nodo 1

Para el tiempo de ejecución, en el primer nodo se nota una actividad baja, sin llegar a utilizar más de la mitad de sus recursos de CPU. Su punto máximo fue de 43.79%, mientras su punto mínimo fue de 2.89%. Se aprecian visualmente en el

Gráfico 2, además de que es posible ver que su uso se mantuvo entre los 20% y 30% durante la ejecución de la prueba.

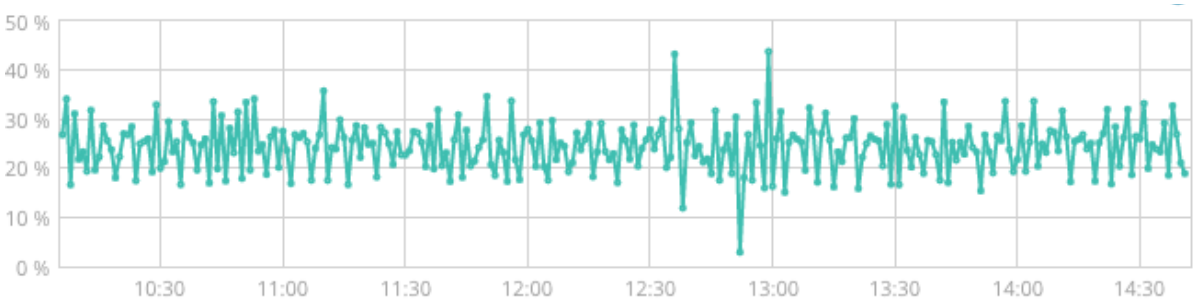


Gráfico 2. Utilización del CPU durante las pruebas caso ventas por año, mes y día

Respecto a la memoria utilizada, su punto más bajo fue un 24% y su punto más alto un 74%. La visualización de este se encuentra en el Gráfico 3. Como nota importante, es claro mencionar que el sistema detectó un uso grande de memoria, lo que hizo que se llamara una limpieza de esta.

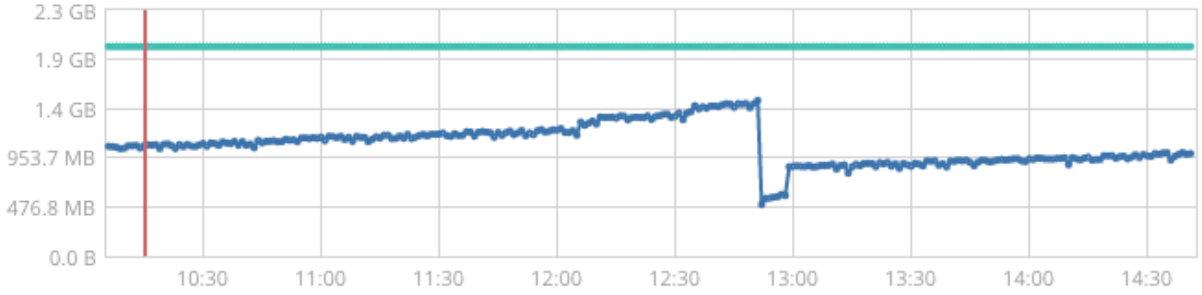


Gráfico 3. Utilización de memoria para caso de ventas por año, mes y día

6.3.1.1.2.2.2 Nodo 2

Para el nodo 2, se observó que el CPU utilizado llegó a un pico máximo de 100% y un mínimo de 94.05%. Es decir, este nodo fue el que realizó más trabajo a la hora de hacer la resolución de la consulta. Los resultados se muestran en el Gráfico 4.

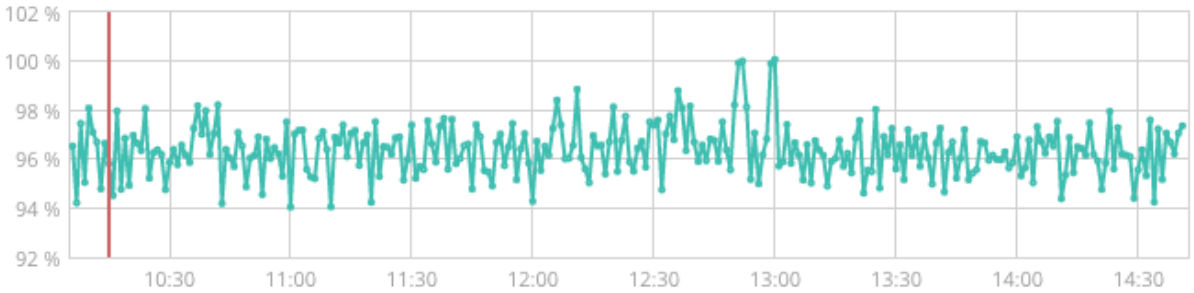


Gráfico 4. Utilización de CPU del nodo 2 en consulta ventas por año, mes y día

Para el caso de la memoria, su valor más alto fue de un 76%, mientras que su valor más bajo, de un 25%. Su uso se puede visualizar en el Gráfico 5. Similar al nodo anterior, se realizó una limpieza de memoria durante la ejecución de la consulta.

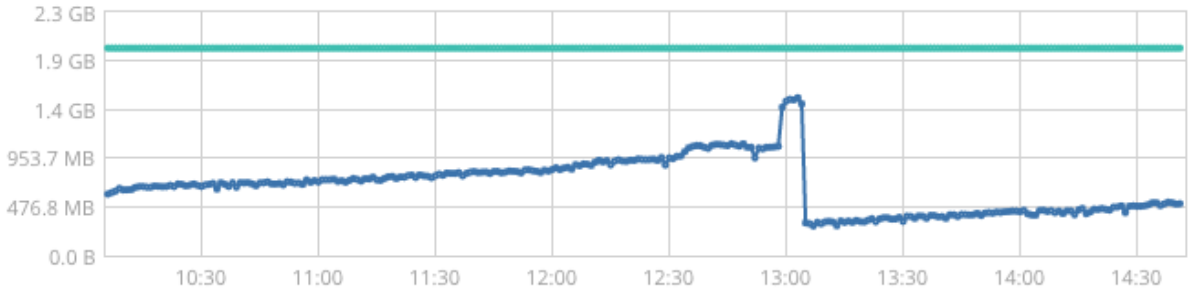


Gráfico 5. Utilización de memoria del nodo 2 durante consulta de ventas por año, mes y día

6.3.1.2 Almacén de Datos

6.3.1.2.1 Exactitud

Para la validación de resultados de este escenario, se dividió la jerarquía de tiempo en año, mes y día.

6.3.1.2.1.1 Ventas por año

En la Tabla 28 se pueden observar los resultados para el elemento año de la jerarquía de fecha.

Tabla 28
Resultados de ventas por fecha para elemento año

Año	Monto ventas
2016	153253033
2017	152921228

6.3.1.2.1.2 Ventas por mes

Para el elemento mes de la jerarquía, solo se seleccionaron los meses del año 2016. En la Tabla 29 se observan los resultados.

Tabla 29
Resultados consulta ventas por fecha en elemento mes

Mes	Monto ventas
Enero	13009432
Febrero	12160571
Marzo	12981945
Abril	12567499
Mayo	12954631
Junio	12532477
Julio	12967114
Agosto	13004483
Setiembre	12572791
Octubre	13001729

6.3.1.2.1.3 Ventas por día

En el caso del elemento día se seleccionaron los primeros diez días del mes de enero de 2016, tal y como se muestra en la Tabla 30.

Tabla 30
Resultados de ventas por fecha en elemento día

Día	Monto ventas
Enero 01 2016	416943
Enero 02 2016	414790
Enero 03 2016	425883
Enero 04 2016	418447
Enero 05 2016	422032
Enero 06 2016	420359
Enero 07 2016	419747
Enero 08 2016	429440
Enero 09 2016	421186
Enero 10 2016	414025

6.3.1.2.2 Desempeño

6.3.1.2.2.1 Duración de la consulta

En las pruebas se obtuvieron estos datos:

- Duración máxima: 1359 milisegundos
- Duración mínima: 875 milisegundos
- Duración promedio: 945 milisegundos

Los resultados completos se muestran en el Gráfico 6.

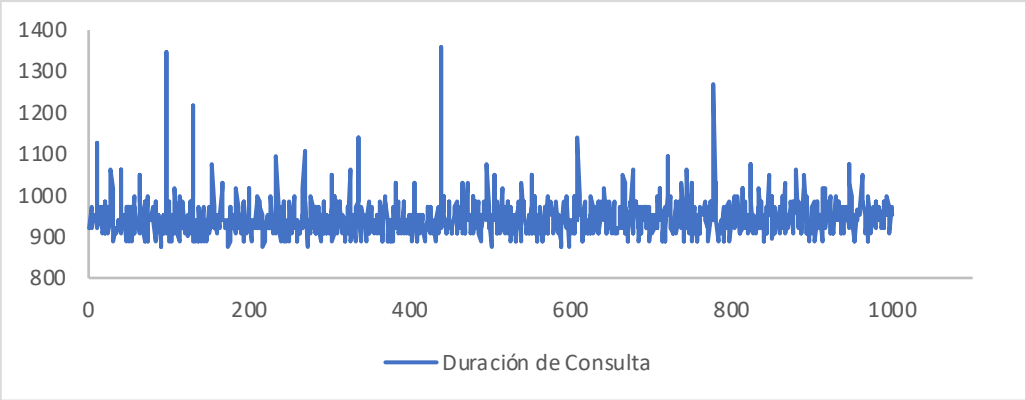


Gráfico 6. Duración consultas para escenario de ventas por año, mes y día

6.3.1.2.2.2 Utilización de recursos

Durante el tiempo de ejecución de este escenario, el consumo de CPU llegó en ocasiones al 100% de la máquina virtual y se mantuvo de forma sostenida por encima del 90%.

La memoria se mantuvo en uso alrededor de un 55% durante el mismo periodo de tiempo. Estos fenómenos se aprecian en el Gráfico 7.

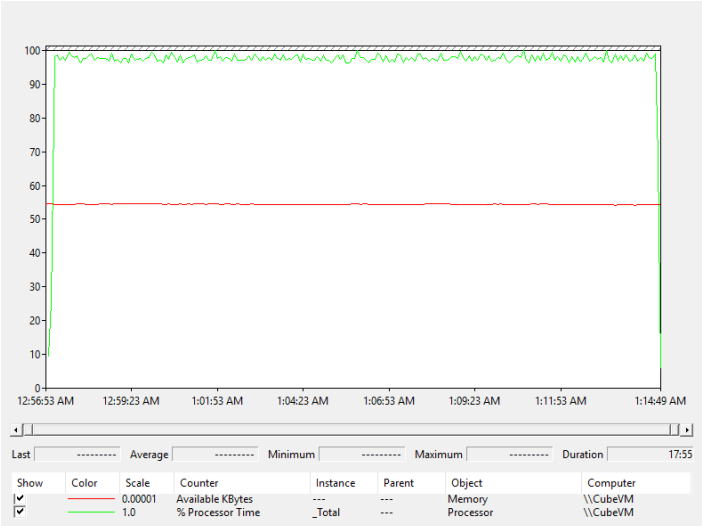


Gráfico 7. Utilización del CPU y memoria libre, pruebas ventas por año

6.3.2 Ventas por Región

Este escenario plantea la recuperación del total de ventas obtenidas agrupadas por país, estado, ciudad y dirección. A continuación, los resultados de la ejecución de la consulta.

6.3.2.1 Motor de Búsqueda

6.3.2.1.1 Exactitud

Similar a la jerarquía de fecha, las consultas se separan por los diferentes miembros de la jerarquía de la región. Los resultados se verán a continuación:

6.3.2.1.1.1 Resultados por país:

En la Tabla 31 se resume una muestra de los resultados por región, divididos por país.

Tabla 31
Resultados de ventas por región, divididas por país.

País	Valor
Afghanistan	1216806
Aland Islands	314559
Albania	298447
Angola	299644
Argentina	4886238
Armenia	307505
Australia	310748
Austria	312772
Azerbaijan	607229
Bahamas	308147

6.3.2.1.1.2 Resultados por estado

En la Tabla 32 se describe una muestra de los resultados de las ventas, divididas por país y estados. Se puede notar que no todos los países cuentan con información de estados, por lo que se muestra dicha celda vacía.

Tabla 32
Resultados consulta ventas por región, dividida por país y estado.

País	Estado	Valor
Afganistán		1216806
Aland Islands		314559
Albania		298447
Angola		299644
Argentina		4886238
Armenia		307505
Australia	Western Australia	310748
Austria	Steiermark	312772
Azerbaijan		607229
Bahamas		308147

6.3.2.1.1.3 Resultados por ciudad

Luego de tomar una muestra de resultados, se obtienen los valores reportados en la Tabla 33.

Tabla 33
Resultados ventas por región, divididas por ciudad.

País	Estado	Ciudad	Valor
Afghanistan		Larkird	302545
		Paghmān	307747
		Sang-e Chārak	301916
		‘Alāqahdārī Kirān wa Munjān	304598
Aland Islands		Brändö	314559
Albania		Kashar	298447
Angola		Lubango	299644
Argentina		Avellaneda	304684
		Berón de Astrada	303848
		Embarcación	305432

6.3.2.1.1.4 Resultados por dirección

La Tabla 34 contiene una muestra de los resultados de consulta de ventas por región, llegando a la categoría de la dirección.

Tabla 34
Resultados consulta de ventas por región, categorizada en direcciones

País	Estado	Ciudad	Dirección	Valor
Afghanistan		Larkird	4315 Mallory Crossing	302545
		Paghmān	48309 Pepper Wood Place	307747
		Sang-e Chārak	72 Spaight Street	301916
		ʻAlāqahdārī Kirān wa Munjān	718 Golf Course Parkway	304598
Aland Islands		Brändö	8680 Dahle Plaza	314559
Albania		Kashar	2 Farwell Street	298447
Angola		Lubango	7817 Northwestern Way	299644
Argentina		Avellaneda	8 Jackson Place	304684
		Berón de Astrada	6 Hollow Ridge Point	303848
		Embarcación	06819 Summer Ridge Junction	305432

6.3.2.1.2 Desempeño

6.3.2.1.2.1 Duración de la consulta

La duración de la consulta generó los siguientes indicadores:

- Duración máxima: 95565 milisegundos.
- Duración mínima: 12597 milisegundos.
- Duración promedio: 43685.701 milisegundos.

Las duraciones se encuentran en el Gráfico 8.

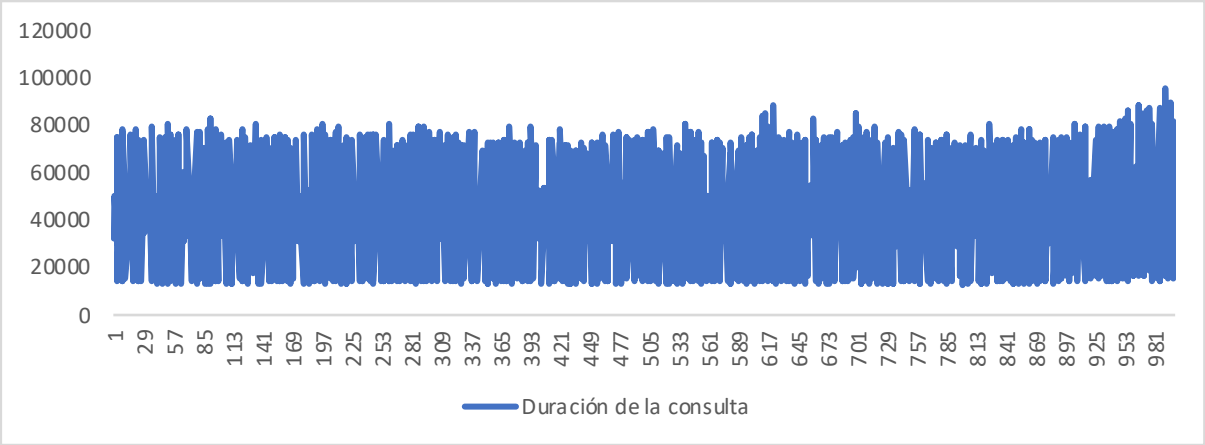


Gráfico 8. Duración de consultas de ventas por región.

6.3.2.1.2.2 Utilización de recursos

6.3.2.1.2.2.1 Nodo 1.

Durante tiempo de ejecución, en el primer nodo obtiene una utilización de procesador máxima fue de 35.99%, mientras su punto mínimo fue de 19.48%. Se aprecian visualmente en el Gráfico 9.

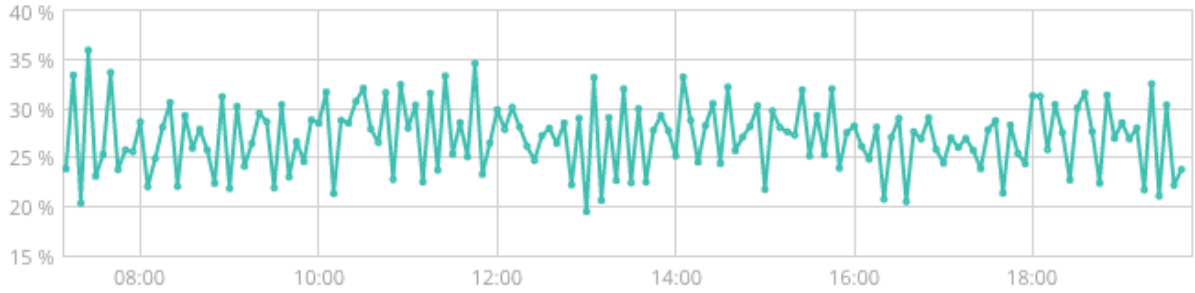


Gráfico 9. Utilización del CPU durante pruebas del caso ventas por región.

Respecto a la memoria utilizada, su punto más bajo de utilización fue un 75% y su punto más alto un 17%. La visualización de este se encuentra en el Gráfico 10.

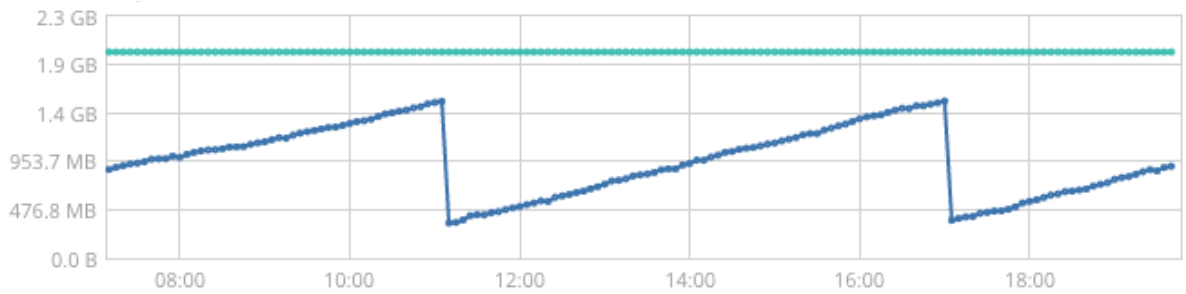


Gráfico 10. Utilización de memoria para caso ventas por región.

6.3.2.1.2.2.2 Nodo 2

Para el nodo 2, se observó que el CPU utilizado fue, en su valor máximo, de 98.54% y un valor más bajo de 94.99%. Los resultados se muestran en el Gráfico 11.

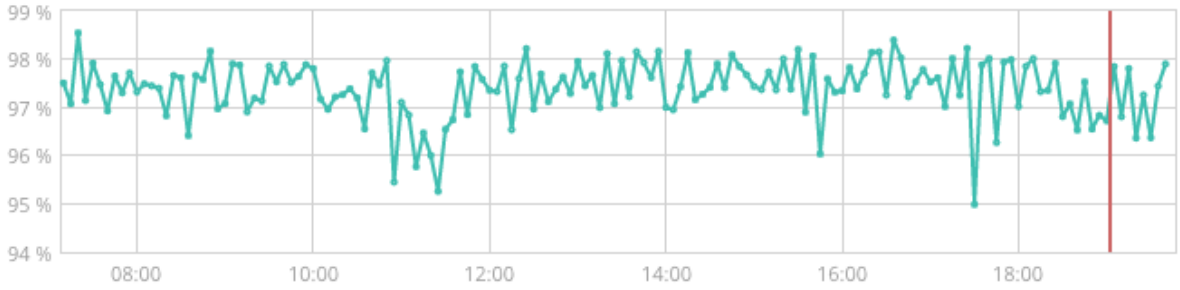


Gráfico 11. Utilización de CPU del nodo 2 en consulta ventas por región.

Para el caso de la memoria, el valor de uso fluctuó entre 17% y 76%, durante la ejecución de la prueba. Los datos se muestran en el Gráfico 12.

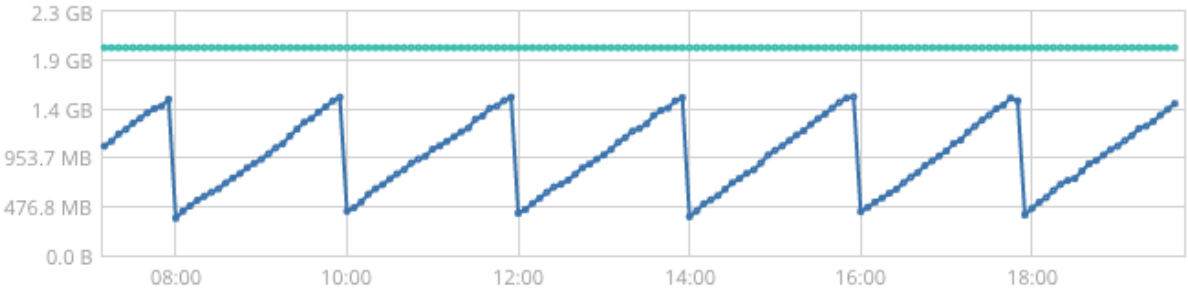


Gráfico 12. Utilización de memoria del nodo 2 durante consulta ventas por región.

6.3.2.2 Almacén de Datos

6.3.2.2.1 Exactitud

En este escenario también se dividió la jerarquía, en este caso región, por cada uno de sus elementos.

6.3.2.2.1.1 Resultados por país

En la Tabla 35 es posible observar algunos de los montos obtenidos para el elemento país de la jerarquía región.

Tabla 35
Resultados ventas por región, elemento país.

País	Ventas
Afghanistan	1216806
Aland Islands	314559
Albania	298447
Angola	299644
Argentina	4886238
Armenia	307505
Australia	310748
Austria	312772
Azerbaijan	607229
Bahamas	308147

6.3.2.2.1.2 Resultados por estado

En la jerarquía región, no todos los países cuentan con información de estado y por ello se encuentran con estado vacío. Además, los datos no necesariamente contienen ventas para múltiples estados de un país. Ver Tabla 36.

Tabla 36
Resultados ventas por región, partición por estado.

País	Estado	Ventas
Afghanistan		1216806
Aland Islands		314559
Albania		298447
Angola		299644
Argentina		4886238
Armenia		307505
Australia	Western Australia	310748
Austria	Steiermark	312772
Azerbaijan		607229
Bahamas		308147

6.3.2.2.1.3 Resultados por ciudad

En la Tabla 37 se muestran los resultados del elemento ciudad.

Tabla 37
Resultados ventas por región en categoría ciudad.

País	Estado	Ciudad	Ventas
Afghanistan		Larkird	302545
		Paghmān	307747
		Sang-e Chārak	301916
		‘Alāqahdārī Kirān wa Munjān	304598
Aland Islands		Brändö	314559
Albania		Kashar	298447
Angola		Lubango	299644
Argentina		Avellaneda	304684
		Berón de Astrada	303848
		Embarcación	305432

6.3.2.2.1.4 Resultados por dirección

La Tabla 38 muestra los resultados obtenidos divididas por dirección.

Tabla 38
Resultados por elemento dirección de ventas por región

País	Estado	Ciudad	Dirección	Ventas
Afghanistan		Larkird	4315 Mallory Crossing	302545
		Paghmān	48309 Pepper Wood Place	307747
		Sang-e Chārak	72 Spaight Street	301916
		ʻAlāqahdārī Kirān wa Munjān	718 Golf Course Parkway	304598
Aland Islands		Brändö	8680 Dahle Plaza	314559
Albania		Kashar	2 Farwell Street	298447
Angola		Lubango	7817 Northwestern Way	299644
Argentina		Avellaneda	8 Jackson Place	304684
		Berón de Astrada	6 Hollow Ridge Point	303848
		Embarcación	06819 Summer Ridge Junction	305432

6.3.2.2.2 Desempeño

6.3.2.2.2.1 Duración de la consulta

En la ejecución de las pruebas se obtuvieron los siguientes datos:

- Duración máxima: 1344 milisegundos.
- Duración mínima: 922 milisegundos.
- Duración promedio: 998 milisegundos.

El Gráfico 13 muestra todas las duraciones a continuación.

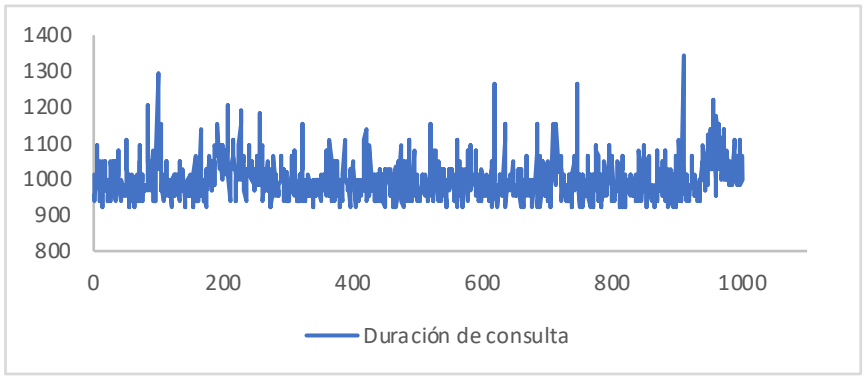


Gráfico 13. Duración de consultas escenario de ventas por región.

6.3.2.2.2.2 Utilización de recursos

Durante el tiempo de ejecución de las pruebas para este escenario, el CPU se mantuvo por encima del 95%, mientras que la memoria en uso estuvo consistentemente por debajo del 50%. Esto se muestra en el Gráfico 14.

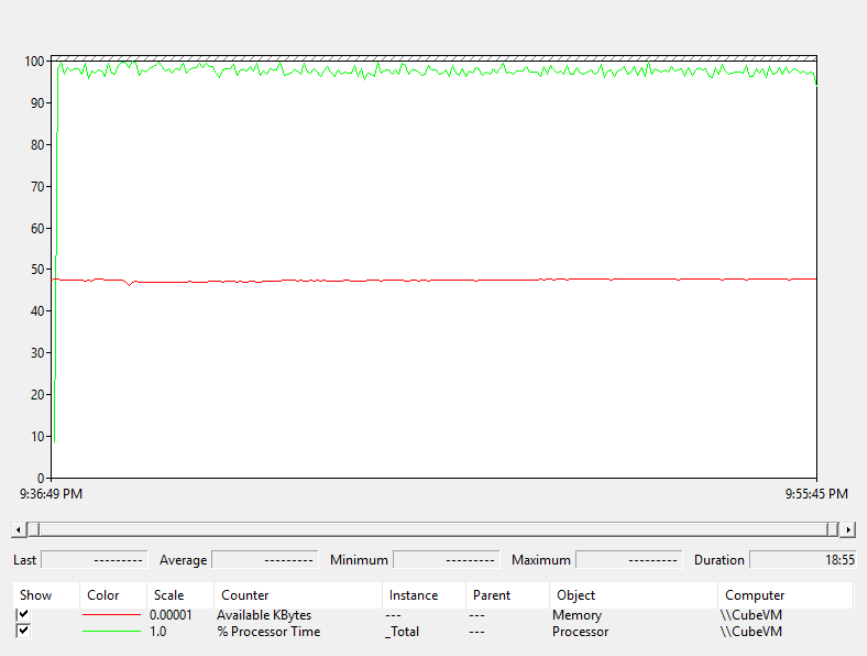


Gráfico 14. Utilización de CPU y memoria escenario de ventas por región.

6.3.3 Ventas por Artista

Este escenario captura las ventas de productos, agrupados esta vez por artistas. Los resultados de exactitud y desempeño se muestran a continuación:

6.3.3.1 Motor de búsqueda

6.3.3.1.1 Exactitud

Para esta medición, se toma un número fijo de artistas, ordenados alfabéticamente para facilitar la comparación. Los resultados se presentan en la Tabla 39.

Tabla 39
Resultados consulta de ventas por artista.

Artista	Valor
Abbey Kivelle	308433
Abby Flowers	304995
Abdul Bardsley	307730
Abey Screen	304978
Abraham Niece	306121
Ada Proughten	306124
Addy O'Clery	305230
Adelaida Semrad	301624
Adena Klemmt	302610
Adolf Commuzzo	306401

6.3.3.1.2 Desempeño

6.3.3.1.2.1 Duración de la consulta

Las duraciones de las consultas en el clúster generaron los siguientes resultados:

- Duración máxima: 5624 milisegundos.
- Duración mínima: 1488 milisegundos
- Duración promedio: 2748.226 milisegundos

La visualización de estos se puede apreciar en el Gráfico 15.

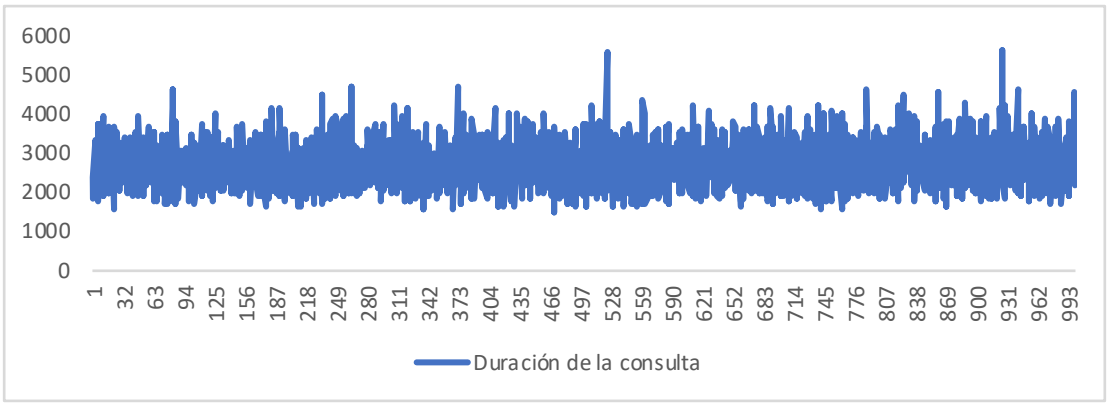


Gráfico 15. Duración de consultas de ventas por artista.

6.3.3.1.2.2 Utilización de recursos

6.3.3.1.2.2.1 Nodo 1

Para el nodo 1, se observó un uso del procesador bastante bajo por parte del motor de búsqueda, variando entre un 10.26% como valor máximo, y un 5.29% como uso mínimo. El Gráfico 16 muestra los datos obtenidos de rendimiento del procesador.

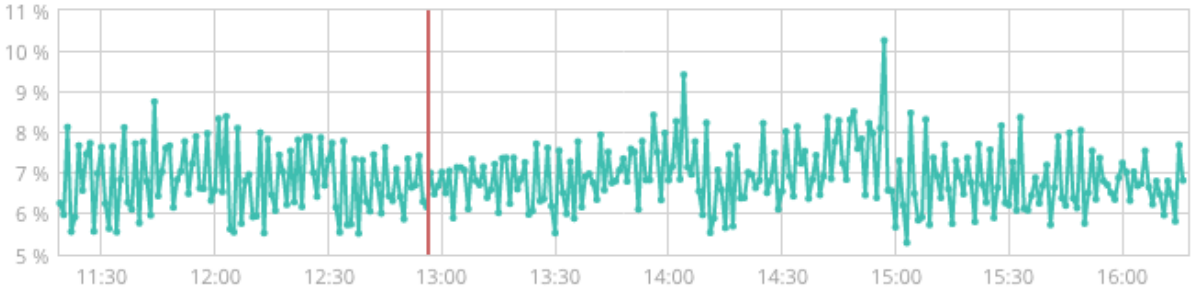


Gráfico 16. Uso del CPU durante consulta de ventas por artista en nodo 1.

En el caso de la memoria, sus valores extremos fueron de un 22% a un 75%. Su variación puede apreciarse en el Gráfico 17.

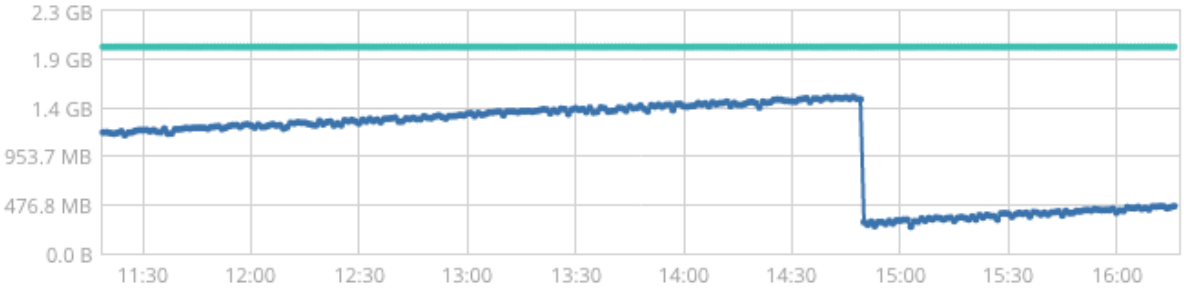


Gráfico 17. Uso de memoria durante consulta ventas por artista.

6.3.3.1.2.2.2Nodo 2

Durante la ejecución de la prueba, el uso promedio del CPU se calculó en un valor máximo fue de 38.3%, mientras que su valor mínimo se mostró como de 23.97%. Su variación durante la ejecución se muestra en el Gráfico 18.

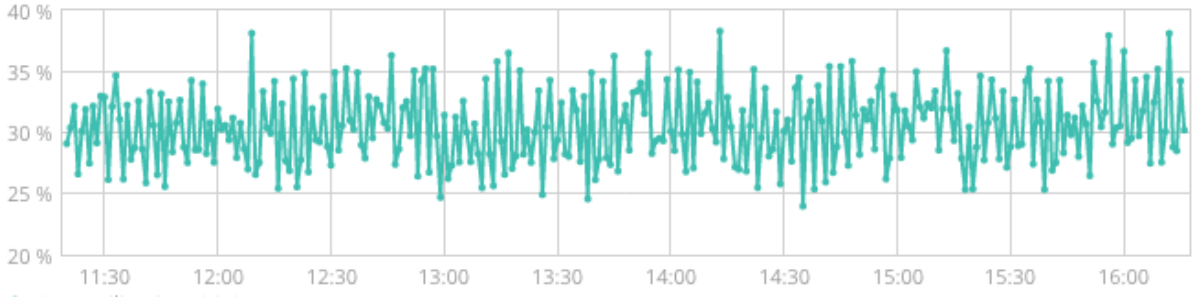


Gráfico 18. Uso del procesador del nodo 2 durante consulta ventas por artista.

Por parte de la memoria, su valor máximo fue de 75%, mientras que su mínimo fue de 14%. Su uso detallado a través del tiempo se muestra en el Gráfico 19.

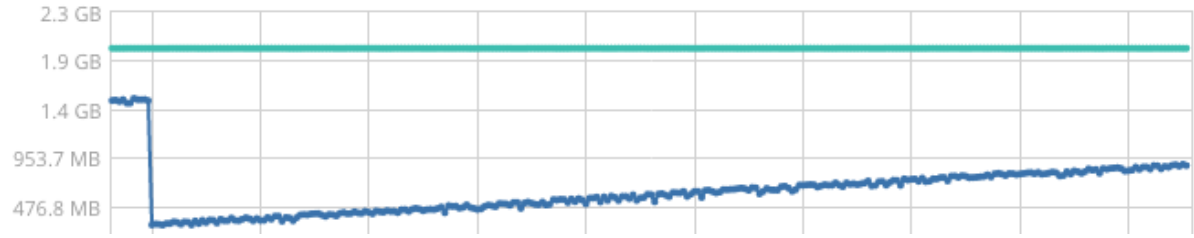


Gráfico 19. Uso de memoria asignada durante consulta ventas por artista.

6.3.3.2 Almacén de Datos

6.3.3.2.1 Exactitud

En este escenario se toman únicamente los primeros diez resultados para artistas, en orden alfabético. Los resultados se visualizan en Tabla 40.

Tabla 40
Resultados ventas por artista.

Artista	Ventas
Abbey Kivelle	308433
Abby Flowers	304995
Abdul Bardsley	307730
Abey Screen	304978
Abraham Niece	306121
Ada Proughten	306124
Addy O'Clery	305230
Adelaida Semrad	301624
Adena Klemmt	302610
Adolf Commuzzo	306401

6.3.3.2.2 Desempeño

6.3.3.2.2.1 Duración de la consulta

La ejecución de las consultas arrojó las siguientes estadísticas:

- Duración máxima: 1454 milisegundos.
- Duración mínima: 1063 milisegundos.
- Duración promedio: 1176 milisegundos.

Dicha información se puede visualizar en el Gráfico 20.

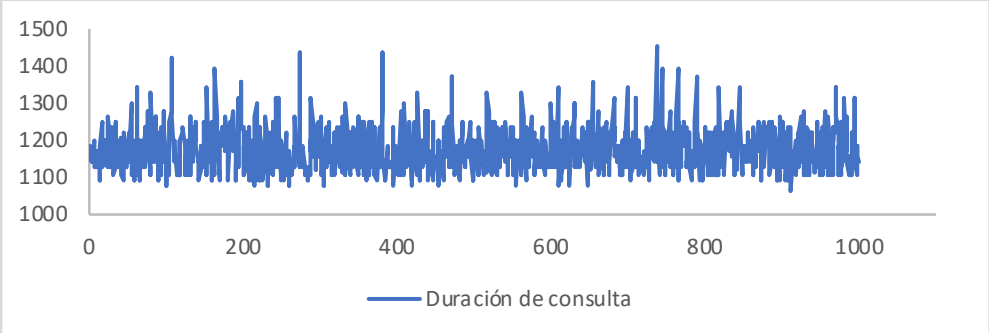


Gráfico 20. Duración consultas para ventas por artista.

6.3.3.2.2.2 Utilización de recursos

Al tomar muestra del uso del procesador durante estas pruebas, se evidencia que este estuvo al tope; en cambio, la memoria se mantuvo rondando el 55%. Esto se evidencia en el Gráfico 21.

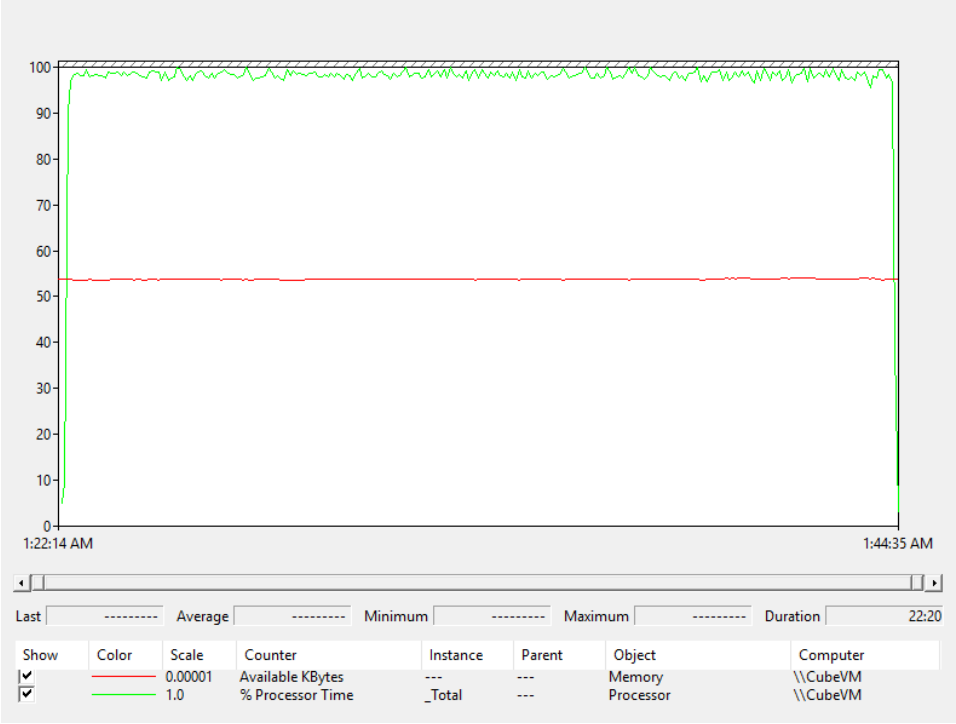


Gráfico 21. Utilización CPU y memoria en escenario ventas por artista.

6.3.4 Ventas por Género

Este escenario intenta responder a la consulta de cuántas películas de algún género específico fueron vendidas a lo largo de la historia de la compañía. La ejecución de las consultas dio los siguientes indicadores:

6.3.4.1 Motor de búsqueda

6.3.4.1.1 Exactitud

Luego de realizar una consulta de un número determinado de resultados, ordenados alfabéticamente, se obtuvieron los números que se ven en la Tabla 41.

Tabla 41
Resultados consulta de ventas por género.

Género	Valor
(no genres listed)	3070848
Action	1829212
Action|Adventure	614971
Action|Adventure|Animation	612433
Action|Adventure|Animation|Children|Comedy|Fantasy	311328
Action|Adventure|Animation|Fantasy	306082
Action|Adventure|Animation|Sci-Fi	609985
Action|Adventure|Children|Fantasy	307108
Action|Adventure|Comedy	307114
Action|Adventure|Comedy|Fantasy	314012

6.3.4.1.2 Desempeño

6.3.4.1.2.1 Duración de la consulta

El desempeño de la consulta presentó los siguientes valores:

- Duración máxima: 13204 milisegundos.

- Duración mínima: 1295 milisegundos.
- Duración promedio:5426.028 milisegundos.

La visualización de los datos se encuentra en el Gráfico 22.

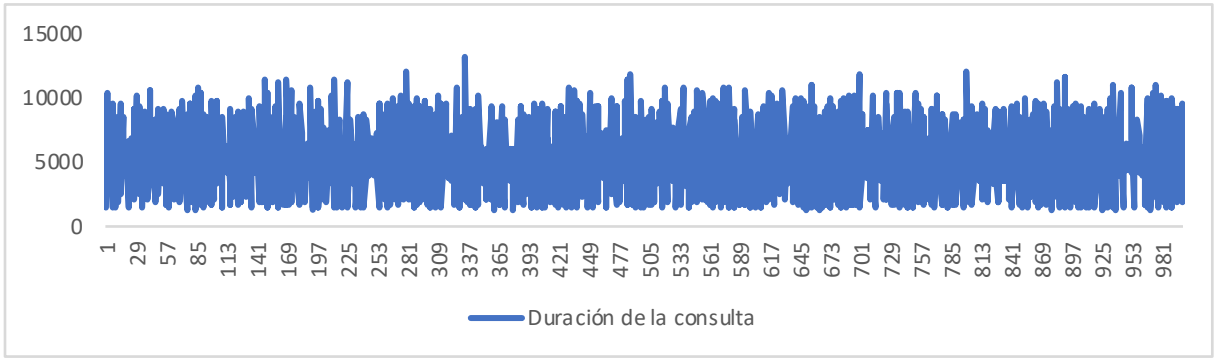


Gráfico 22. Duración de consultas ventas por género.

6.3.4.1.2.2 Utilización de los recursos

6.3.4.1.2.2.1Nodo 1

Para esta consulta, se ve poca actividad por parte del nodo 1, en cuanto al procesador se refiere. Con valores que van entre el 14.5% y el 23.95%, es posible ver ciertas variaciones frecuentes en los números del mismo. Se puede ver la información en el Gráfico 23.

Gráfico 23. Uso del procesador para consulta ventas por género en nodo 1

Además, el parámetro de la memoria fue utilizado también de manera leve. Su valor mínimo fue de un 34%, mientras que el máximo, de 43%. La tendencia de uso puede verse reflejada en el Gráfico 24.

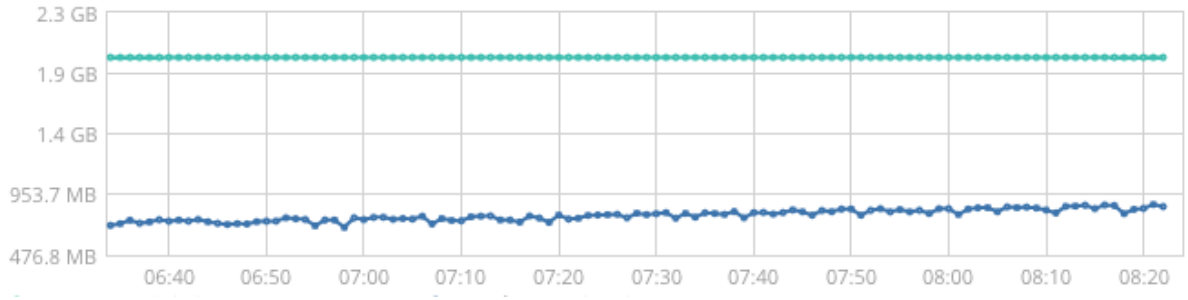


Gráfico 24. Uso de memoria para consulta ventas por género en nodo 1

6.3.4.1.2.2.2 Nodo 2

Es posible ver en este nodo una gran actividad en el procesador. En este caso, el uso fue desde un 84.71% a un 90.66%. Aun cuando hubo muchas veces en las que el procesador alcanzaba niveles cercanos al mínimo, es claro apreciar en el Gráfico 25 que en varios momentos el uso estuvo más cerca de sus valores altos.

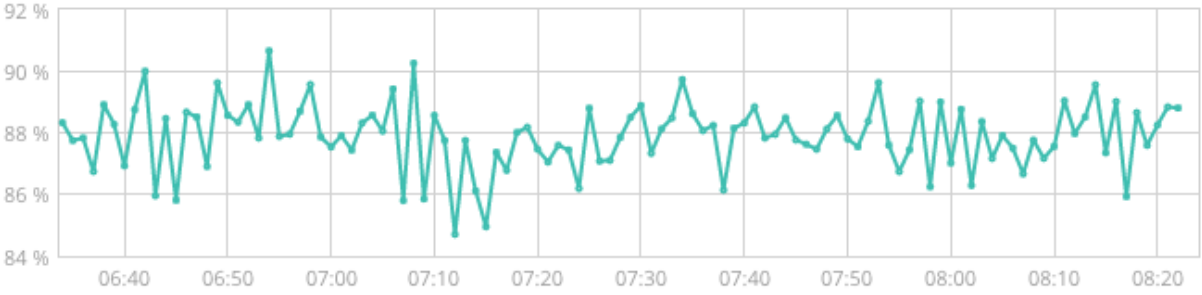


Gráfico 25. Uso del procesador del nodo 2 en consulta ventas por género

En el caso de la memoria usada, se aprecia que los valores fueron también altos, con tendencias entre el 63% y el 73%. Esto indica que este nodo se dedicó

más al procesamiento y almacenamiento de valores intermedios de la consulta en sí. El uso de la memoria se puede apreciar en el Gráfico 26.

Gráfico 26. Uso de memoria del nodo 2 para consulta ventas por género.

6.3.4.2 Almacén de Datos

6.3.4.2.1 Exactitud

En la Tabla 42 se observan los primeros resultados, en orden alfabético para ventas por género.

Tabla 42
Resultados consulta ventas por género

Género	Ventas
Sin género listado	3070848
Action	1829212
Action|Adventure	614971
Action|Adventure|Animation	612433
Action|Adventure|Animation|Children|Comedy|Fantasy	311328
Action|Adventure|Animation|Fantasy	306082
Action|Adventure|Animation|Sci-Fi	609985
Action|Adventure|Children|Fantasy	307108
Action|Adventure|Comedy	307114
Action|Adventure|Comedy|Fantasy	314012

6.3.4.2.2 Desempeño

6.3.4.2.2.1 Duración de la consulta

Las ejecuciones para este escenario generaron los siguientes indicadores:

- Duración máxima: 1359 milisegundos
- Duración mínima: 922 milisegundos
- Promedio de duración: 1023 milisegundos

El comportamiento general de las pruebas de este caso de uso puede ser observado en Gráfico 27.

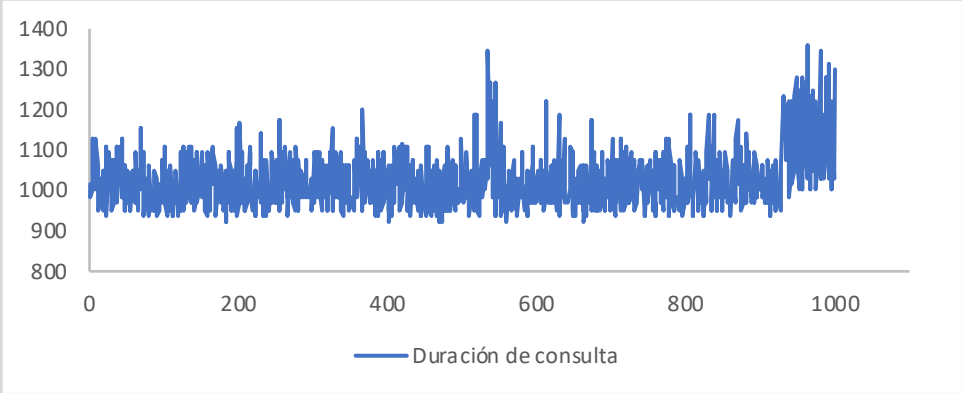


Gráfico 27. Duración consultas ventas por género

6.3.4.2.2.2 Utilización de Recursos

La memoria de la máquina virtual estuvo ligeramente sobre el 50% y el uso de procesador se mantuvo alto durante la ejecución de este escenario, tal como se observa en el Gráfico 28.

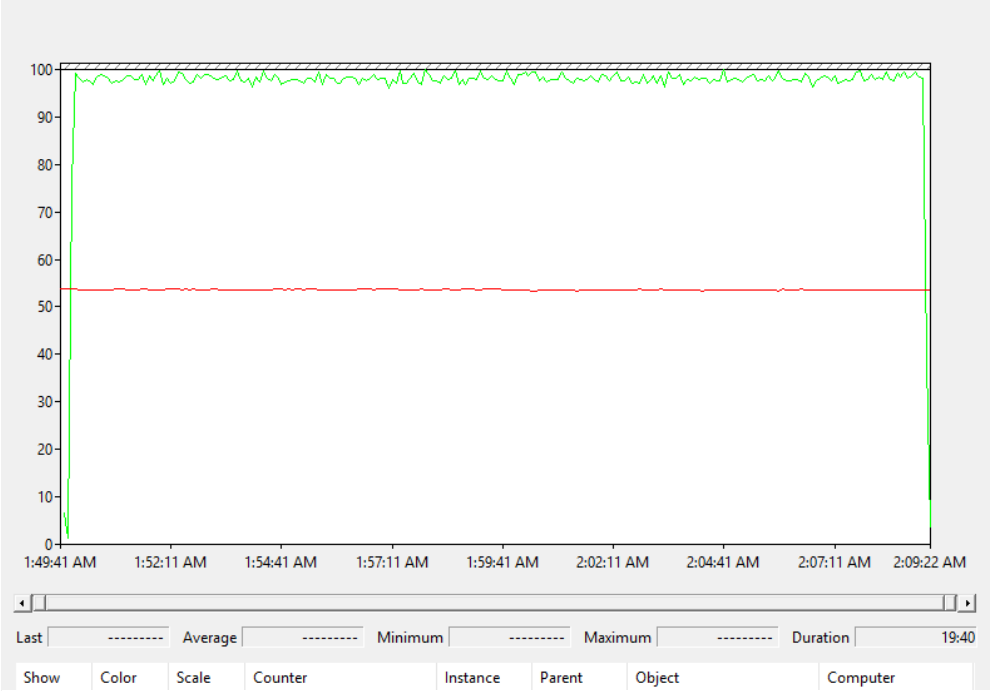


Gráfico 28. Utilización de CPU y memoria para ventas por género.

6.3.5 Ventas por Álbum

En esta consulta se desea obtener las ganancias de todas las ventas; esta vez, estas fueron agrupadas por los valores de la tabla Álbum.

6.3.5.1 Motor de búsqueda

6.3.5.1.1 Exactitud

Una muestra de resultados se aprecia en la Tabla 43.

Tabla 43
Resultados consulta ventas por álbum

Álbum	Valor
'Neath the Arizona Skies	307772
'burbs, The	301491
*batteries not included	303950
...And God Spoke	306284
10 Mountains 10 Years	309199
101 Reykjavik (101 Reykjavík)	305511
13 Sins	303981
18 Fingers of Death!	302322
1900 (Novecento)	302178
20 Dates	309902

6.3.5.1.2 Desempeño

6.3.5.1.2.1 Duración de la consulta

Al tabular los datos de las diferentes consultas, se obtienen los siguientes resultados:

- Duración máxima: 18172 milisegundos
- Duración mínima: 1960 milisegundos
- Duración promedio: 7600.15 milisegundos

Los resultados se aprecian visualmente en el Gráfico 29.

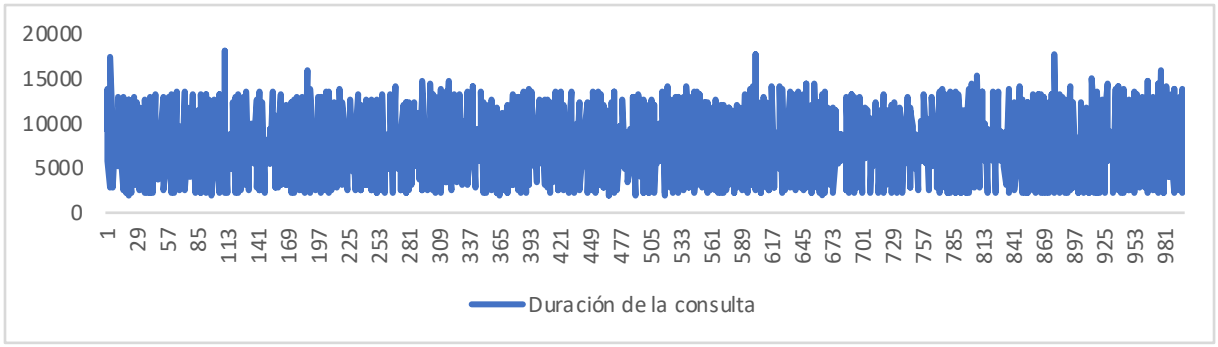


Gráfico 29. Duración de consultas ventas por álbum

6.3.5.1.2.2 Utilización de recursos

6.3.5.1.2.2.1Nodo 1

Al analizar el uso del procesador en el sistema mientas las pruebas fueron ejecutadas, se obtuvo un porcentaje de uso entre los valores 15.51% y 30.95%, observándose el comportamiento en el Gráfico 30.

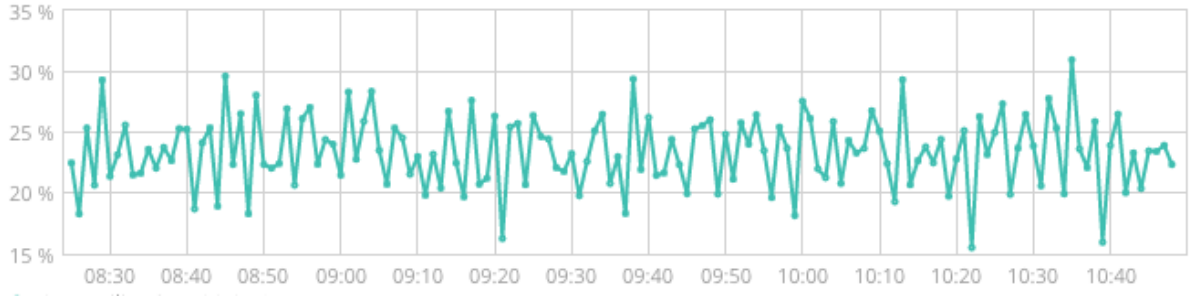


Gráfico 30. Uso del procesador en nodo 1 durante consulta ventas por álbum

Respecto al uso de memoria, se observa un comportamiento que tiende a crecer en este nodo, con valores registrados de entre un 41% y un 54%. Se aprecia el uso en el Gráfico 31.

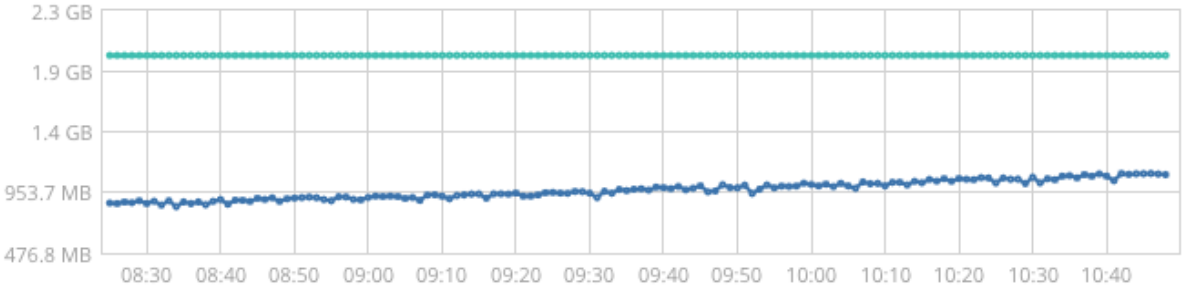


Gráfico 31. Uso de memoria en nodo 1 durante consulta ventas por álbum

6.3.5.1.2.2.2 Nodo 2

Es posible apreciar que este nodo fue el que realizó más trabajo al realizar las agregaciones en cuanto a CPU se refiere. Su valor máximo fue de 93.69%, mientras que el mínimo fue de un 87.62%. Se visualiza el uso en el Gráfico 32.

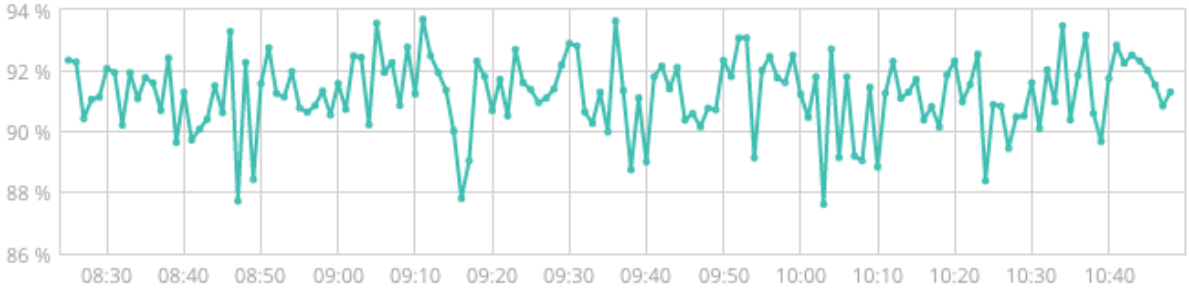


Gráfico 32. Uso del procesador en nodo 2 para consulta ventas por álbum

Respecto a la memoria utilizada, también este nodo se ve marcado por el gran uso de esta. Sus máximos valores fueron de 75% mientras que el mínimo fue de 15%. Sus valores se muestran en el Gráfico 33.

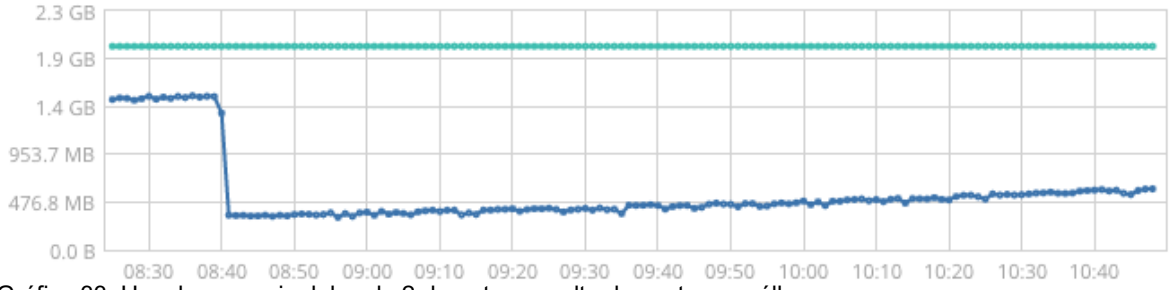


Gráfico 33. Uso de memoria del nodo 2 durante consulta de ventas por álbum.

6.3.5.2 Almacén de Datos

6.3.5.2.1 Exactitud

Para este escenario, se muestra un subconjunto de diez resultados en la Tabla 44.

Tabla 44
Resultados ventas por álbum.

Álbum	Ventas
'Neath the Arizona Skies	307772
'burbs, The	301491
*batteries not included	303950
...And God Spoke	306284
10 Mountains 10 Years	309199
101 Reykjavik (101 Reykjavík)	305511
13 Sins	303981
18 Fingers of Death!	302322
1900 (Novecento)	302178
20 Dates	309902

6.3.5.2.2 Desempeño

6.3.5.2.2.1 Duración de la consulta

Para las pruebas del escenario ventas por álbum se obtuvieron los datos a continuación:

- Duración máxima: 1609 milisegundos.
- Duración mínima: 1078 milisegundos.
- Promedio de duración: 1187 milisegundos.

El Gráfico 34 contiene los datos de todas las ejecuciones.

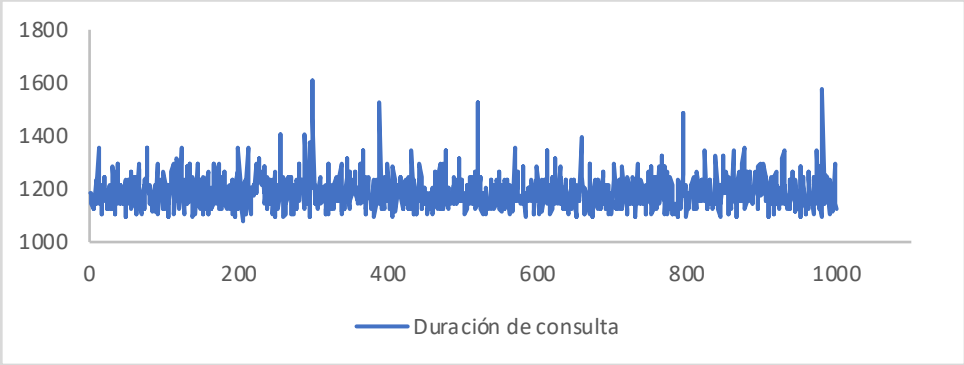


Gráfico 34. Duración consultas en ventas por álbum.

6.3.5.2.2.2 Utilización de Recursos

Al igual que en otros escenarios, el uso del CPU fue bastante elevado en este conjunto de pruebas. En cuanto a la memoria, se mantuvo cerca de 55% todo el tiempo. Esto es observable en el Gráfico 35.

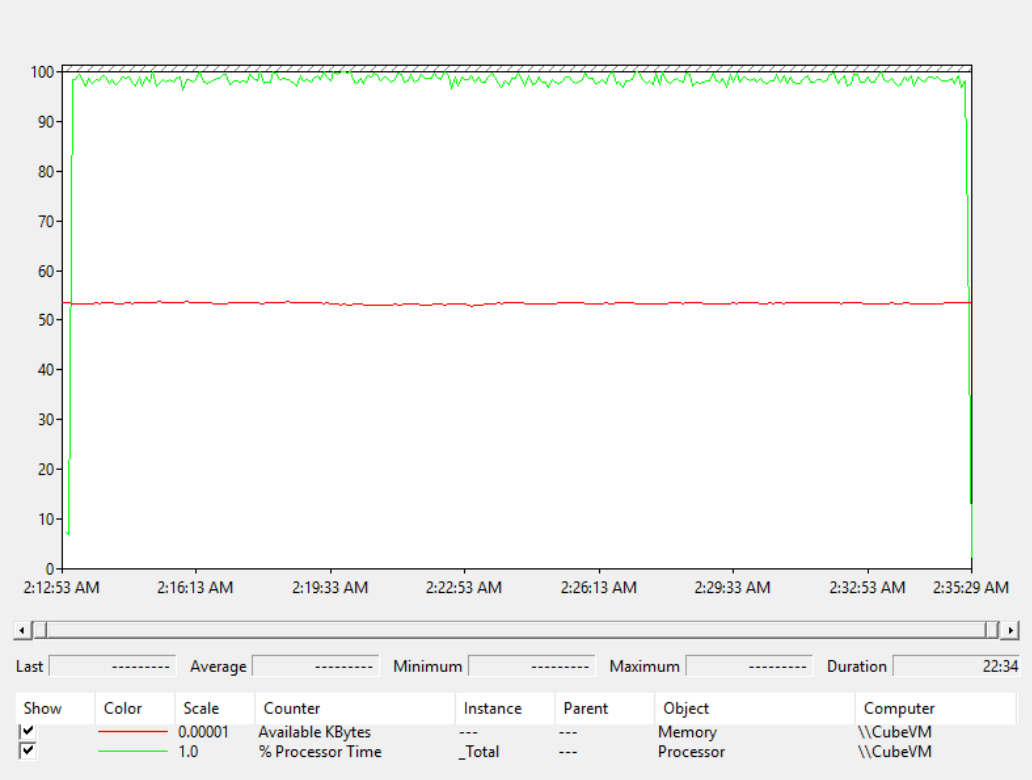


Gráfico 35: Utilización de CPU y memoria para ventas por álbum.

6.3.6 Cantidad de tipo de media vendida por fecha

Este escenario intenta responder a la pregunta de cuántos tipos de media fueron vendidos agrupados por año, mes y día. Se documentan en las siguientes secciones los resultados obtenidos:

6.3.6.1 Motor de búsqueda

6.3.6.1.1 Exactitud

Similar al caso descrito en la sección 0, se dividen los resultados por cada miembro de la jerarquía en las siguientes subsecciones.

6.3.6.1.1.1 Resultados por año

La Tabla 45 enumera una muestra de resultados de esta consulta.

Tabla 45
Resultados consulta cantidad de tipo de media vendida, separada por años.

Tipo de media	Fecha (timestamp)	Fecha (formato ISO)	Valor
application/x-troff-msvideo	1451606400000	Fri Jan 01 2016 00:00:00	618249
	1483228800000	Sun Jan 01 2017 00:00:00	617394
audio/mpeg3	1451606400000	Fri Jan 01 2016 00:00:00	601815
	1483228800000	Sun Jan 01 2017 00:00:00	601494
audio/x-mpeg-3	1451606400000	Fri Jan 01 2016 00:00:00	512383
	1483228800000	Sun Jan 01 2017 00:00:00	511676
video/avi	1451606400000	Fri Jan 01 2016 00:00:00	602361
	1483228800000	Sun Jan 01 2017 00:00:00	601631
video/mpeg	1451606400000	Fri Jan 01 2016 00:00:00	976254
	1483228800000	Sun Jan 01 2017 00:00:00	972288

6.3.6.1.1.2 Resultados por mes

Los resultados de esta consulta por mes se pueden visualizar en la Tabla 46.

Tabla 46
Resultados consulta cantidad de tipo de media, categorizada en meses.

Tipo de media	Fecha (timestamp)	Fecha (Formato ISO)	Valor
application/x-troff-msvideo	1451606400000	Fri Jan 01 2016 00:00:00	52509
	1454284800000	Mon Feb 01 2016 00:00:00	48690
	1456790400000	Tue Mar 01 2016 00:00:00	53089
	1459468800000	Fri Apr 01 2016 00:00:00	50364
	1462060800000	Sun May 01 2016 00:00:00	52351
	1464739200000	Wed Jun 01 2016 00:00:00	50623
	1467331200000	Fri Jul 01 2016 00:00:00	52166
	1470009600000	Mon Aug 01 2016 00:00:00	52616
	1472688000000	Thu Sep 01 2016 00:00:00	50671
	1475280000000	Sat Oct 01 2016 00:00:00	52286

6.3.6.1.1.3 Resultados por día

En la Tabla 47 se documentan algunos resultados de la consulta de ventas por tipo de media, esta vez categorizada por día.

Tabla 47
Resultados consulta ventas por tipo de media, categorizadas por día.

Tipo de media	Fecha (Timestamp)	Fecha (Formato ISO)	Valor
application/x-troff-msvideo	1451606400000	Fri Jan 01 2016 00:00:00	1702
	1451692800000	Sat Jan 02 2016 00:00:00	1657
	1451779200000	Sun Jan 03 2016 00:00:00	1696
	1451865600000	Mon Jan 04 2016 00:00:00	1767
	1451952000000	Tue Jan 05 2016 00:00:00	1658
	1452038400000	Wed Jan 06 2016 00:00:00	1730
	1452124800000	Thu Jan 07 2016 00:00:00	1667
	1452211200000	Fri Jan 08 2016 00:00:00	1717
	1452297600000	Sat Jan 09 2016 00:00:00	1665
	1452384000000	Sun Jan 10 2016 00:00:00	1705

6.3.6.1.2 Desempeño

6.3.6.1.2.1 Duración de la consulta

Los resultados de la ejecución de las consultas fueron los siguientes:

- Duración máxima: 65295 milisegundos.

- Duración mínima: 8078 milisegundos.
- Duración promedio: 31795.868 milisegundos.

Se pueden apreciar los tiempos en el Gráfico 36.

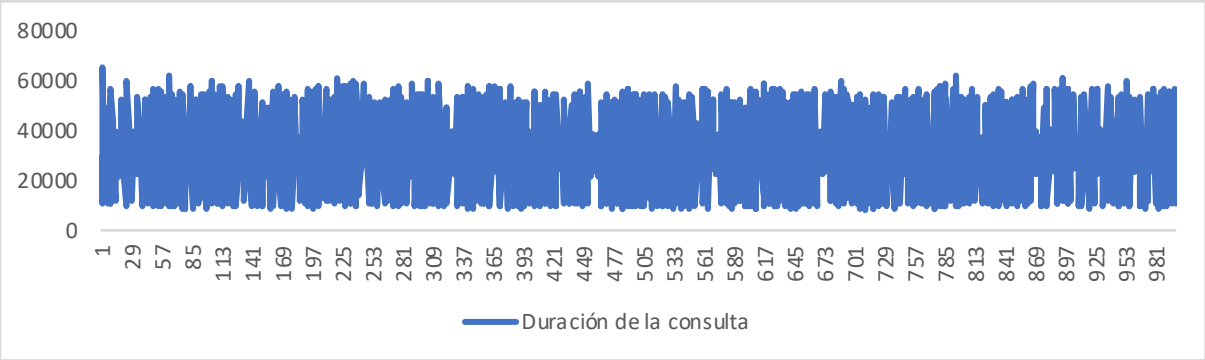


Gráfico 36. Duración de consultas cantidad de ventas por tipo de media separadas por país.

6.3.6.1.2.2 Utilización de recursos

6.3.6.1.2.2.1 Nodo 1

Durante las consultas realizadas en las pruebas, se mostró una tendencia en el uso del procesador que fluctuaba entre 7.94% y 28.37% del uso de este. Dicho uso es registrado en el Gráfico 37.

Gráfico 37. Uso del procesador en consulta cantidad de ventas por tipo de media.

Cabe destacar que para este nodo se vio un alto uso de la memoria, superando el 50% del uso de este recurso. Sus valores fueron entre 47% y 79%, en una tendencia creciente, como se aprecia en el Gráfico 38.

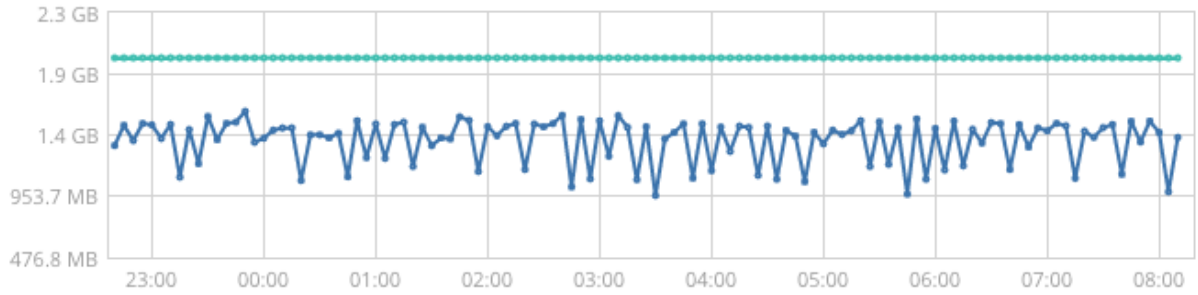


Gráfico 38. Uso del procesador en consulta de cantidad ventas por tipo de media.

6.3.6.1.2.2.2 Nodo 2

Para este nodo en particular, las consultas tuvieron un impacto significativo en el uso del procesador. Su uso porcentual estuvo entre el 51.34% y el 97.45%, manteniéndose en esas posiciones durante todo el tiempo de la ejecución de la prueba. Se aprecia el uso de este en el Gráfico 39.

Gráfico 39. Uso del procesador consulta ventas por tipo de media en nodo 2.

Para el mismo nodo, en el caso de la memoria, se nota una tendencia similar a nodo 1, con una tendencia a la subida del uso de esta. Su uso estuvo entre un 45% y un 80%, tal y como se muestra en el Gráfico 40.

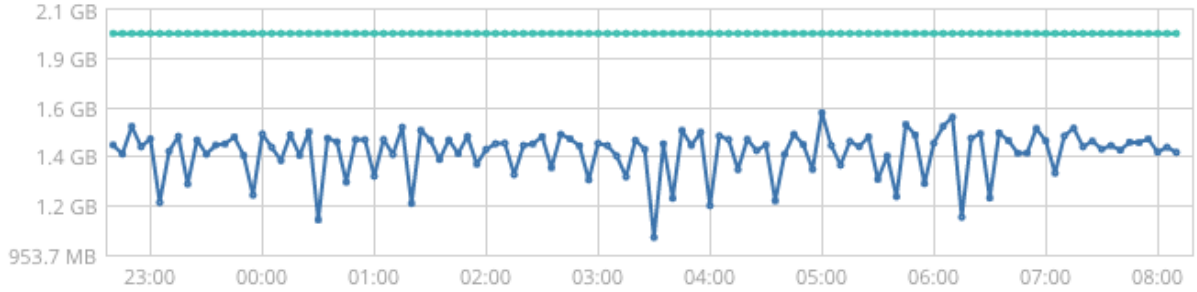


Gráfico 40. Uso de memoria en nodo 2 durante consulta ventas por tipo de media.

6.3.6.2 Almacén de Datos

6.3.6.2.1 Exactitud

Este escenario utiliza una jerarquía de fecha, la cual fue dividida para presentar una mayor claridad en los resultados.

6.3.6.2.1.1 Resultados por año

En la Tabla 48 se listan la cantidad de tipos de media por año.

Tabla 48
Resultados cantidad vendida agrupada por tipo de media y año.

Tipo Media	Año	Cantidad
application/x-troff-msvideo	2016	618249
	2017	617394
audio/mpeg3	2016	601815
	2017	601494
audio/x-mpeg-3	2016	512383
	2017	511676
video/avi	2016	602361
	2017	601631
video/mpeg	2016	976254
	2017	972288

6.3.6.2.1.2 Resultados por mes

Un subconjunto de los resultados para este escenario dividido por mes se puede ver en la Tabla 49.

Tabla 49
Resultados cantidad de media vendida agrupado por tipo y mes.

Tipo Media	Año	Mes	Cantidad
application/x-troff-msvideo	2016	Enero	52509
		Febrero	48690
		Marzo	53089
		Abril	50364
		Mayo	52351
		Junio	50623
		Julio	52166
		Agosto	52616
		Setiembre	50671
		Octubre	52286

6.3.6.2.1.3 Resultados por día

Algunos de los resultados obtenidos para el elemento día de la jerarquía se pueden observar en la Tabla 50.

Tabla 50
Resultados cantidad media vendida por día agrupados por tipo.

Tipo Media	Día	Cantidad
application/x-troff-msvideo	Enero 01 2016	1702
	Enero 02 2016	1657
	Enero 03 2016	1696
	Enero 04 2016	1767
	Enero 05 2016	1658
	Enero 06 2016	1730
	Enero 07 2016	1667
	Enero 08 2016	1717
	Enero 09 2016	1665
	Enero 10 2016	1705

6.3.6.2.2 Desempeño

6.3.6.2.2.1 Duración de la consulta

La ejecución de las pruebas arrojó los siguientes datos:

- Duración máxima: 1875 milisegundos
- Duración mínima: 1141 milisegundos
- Duración promedio: 1300 milisegundos

En el Gráfico 41 se pueden visualizar los resultados de la totalidad de ejecuciones.

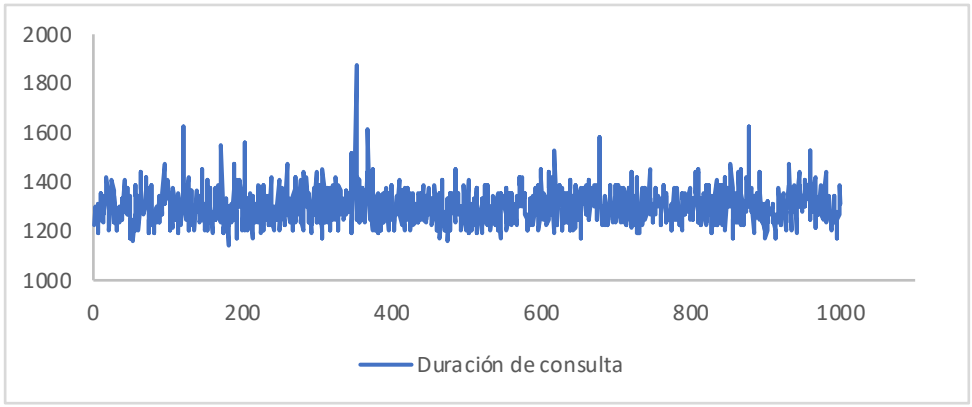


Gráfico 41. Duración de consultas cantidad de tipo de media vendida por año y mes.

6.3.6.2.2.2 Utilización de recursos

La utilización del procesador estuvo considerablemente alta, este se mantuvo por encima del 90% durante la ejecución de todo el escenario de prueba. La memoria, durante ese mismo periodo, rondó el 50%. A continuación, en el Gráfico 42 se observa el comportamiento descrito.

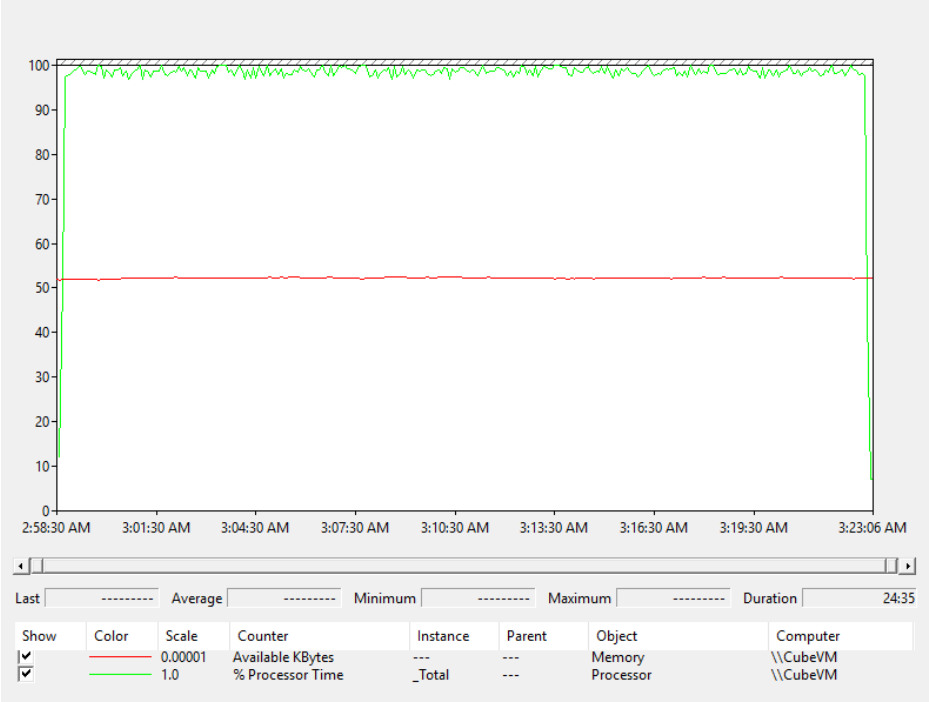


Gráfico 42. Utilización de CPU en cantidad de tipo de media en tiempo.

6.3.7 Ventas por Género por Región

Esta consulta responde a la cantidad de dinero obtenido en ventas, agrupada por género en cada una de las regiones. Los resultados de rendimiento se describen a continuación:

6.3.7.1 Motor de búsqueda

6.3.7.1.1 Exactitud

Se utilizará el mismo enfoque que el escenario descrito en la sección 6.3.2 dividiendo cada uno de los resultados en diferentes secciones, con el fin de realizar una comparación más sencilla. Los resultados se describen en las siguientes secciones.

6.3.7.1.1.1 Ventas por género por país

Los resultados separados por la categoría de país se muestran en la Tabla 51.

Tabla 51
Resultados consulta ventas por género por región, nivel país.

País	Género	Valor
Afghanistan	(no genres listed)	11335
	Action	7250
	Action|Adventure	2442
	Action|Adventure|Animation	2382
	Action|Adventure|Animation|Children|Comedy|Fantasy	1337
	Action|Adventure|Animation|Fantasy	1004
	Action|Adventure|Animation|Sci-Fi	2825
	Action|Adventure|Children|Fantasy	1481
	Action|Adventure|Comedy	815
	Action|Adventure|Comedy|Fantasy	970

6.3.7.1.1.2 Ventas por género por estado

Los resultados de las ventas por género por región, en la categoría por estado, se encuentran en la Tabla 52.

Tabla 52
Resultados ventas por género por región, separadas por estado.

País	Estado	Género	Valor
Afghanistan		(no genres listed)	11335
		Action	7250
		Action|Adventure	2442
		Action|Adventure|Animation	2382
		Action|Adventure|Animation|Children|Comedy|Fantasy	1337
		Action|Adventure|Animation|Fantasy	1004
		Action|Adventure|Animation|Sci-Fi	2825
		Action|Adventure|Children|Fantasy	1481
		Action|Adventure|Comedy	815
		Action|Adventure|Comedy|Fantasy	970

6.3.7.1.1.3 Ventas por género por ciudad

Los resultados de las ventas, esta vez por ciudad, se muestran en la Tabla 53.**Error! Reference source not found.**

Tabla 53
Resultados ventas por género por región, nivel ciudad.

País	Estado	Ciudad	Género	Valor
Afghanistan		Larkird	(no genres listed)	3281
			Action	1668
			Action|Adventure	316
			Action|Adventure|Animation	641
			Action|Adventure|Animation|Children|Comedy|Fantasy	53
			Action|Adventure|Animation|Fantasy	419
			Action|Adventure|Animation|Sci-Fi	715
			Action|Adventure|Children|Fantasy	250
			Action|Adventure|Comedy	225
			Action|Adventure|Comedy|Fantasy	294

6.3.7.1.1.4 Ventas por genero por dirección

Los resultados separados por direcciones se encuentran en la Tabla 54.

Tabla 54
Resultados consulta de ventas por género por región, separada por direcciones.

País	Estado	Ciudad	Dirección	Género	Valor
Afghanistan		Larkird	4315 Mallory Crossing	(no genres listed)	3281
				Action	1668
				Action|Adventure	316
				Action|Adventure|Animation	641
				Action|Adventure|Animation|Children|Comedy|Fantasy	53
				Action|Adventure|Animation|Fantasy	419
				Action|Adventure|Animation|Sci-Fi	715
				Action|Adventure|Children|Fantasy	250
				Action|Adventure|Comedy	225
				Action|Adventure|Comedy|Fantasy	294

6.3.7.1.2 Desempeño

6.3.7.1.2.1 Duración de la consulta

No fue posible capturar medidas para este caso de prueba en particular. Durante la ejecución de la consulta diseñada para obtener los datos el clúster no era capaz de responder debido a la utilización completa de la memoria asignada. Con base en lo anterior, cada uno de los nodos era reiniciado, causando un fallo de todas las consultas.

6.3.7.1.2.2 Utilización de los recursos

6.3.7.1.2.2.1 Nodo 1

Incluso con la ejecución fallida de la consulta, es posible monitorear el uso de recursos de este nodo. El Gráfico 43 muestra el uso del CPU, el cual se utilizó como mínimo en un 1.05%, y como máximo en un 74.44%.

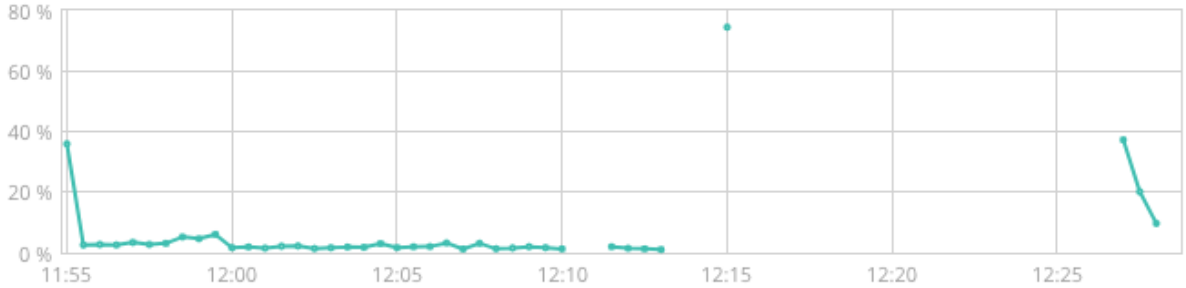


Gráfico 43. Uso del procesador del nodo 1 consulta ventas por género por región. Los espacios sin datos indican que el clúster fue reiniciado en ese espacio de tiempo.

La memoria obtuvo un uso creciente y alcanzó el uso máximo de la misma, dando como resultado el reinicio del nodo. Se puede visualizar el uso en el Gráfico 44.

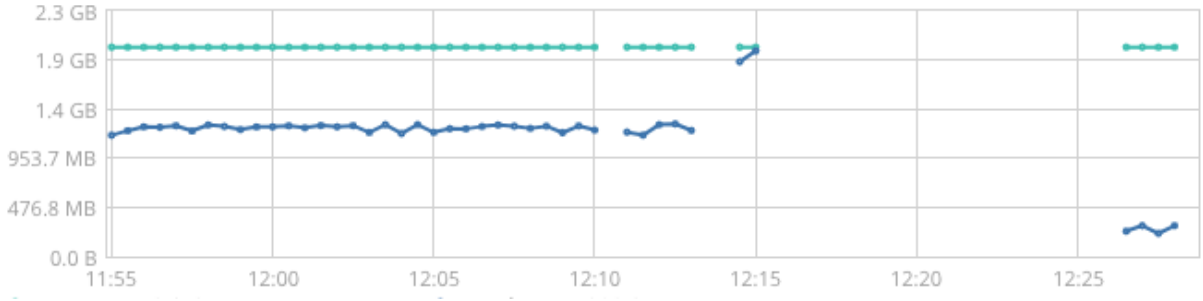


Gráfico 44. Uso de memoria del nodo 1 para consulta ventas según género, por región. Los espacios en blanco indican tiempo en el cual el clúster se reinició.

6.3.7.1.2.2.2 Nodo 2

En este caso, el procesador alcanzó niveles altos en su uso, llegando a utilizar hasta un 100% de su capacidad. Su valor mínimo fue de un 99.85%. Debido al uso extremo del procesador, la herramienta de visualización no fue capaz de recuperar datos suficientes para la generación del gráfico respectivo.

En este nodo, la memoria también fue utilizada de manera intensiva, llegando al 99% de la misma. Al verse utilizada en su totalidad, el nodo también fue reiniciado. El Gráfico 45 muestra los datos recopilados.

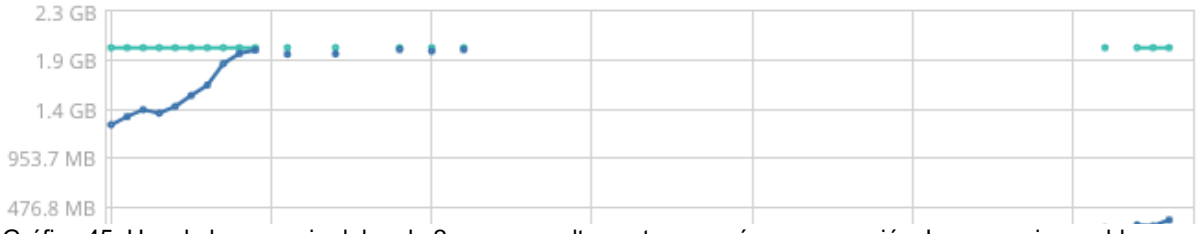


Gráfico 45. Uso de la memoria del nodo 2 para consulta ventas por género por región. Los espacios en blanco representan los momentos en los que el nodo debió ser reiniciado.

6.3.7.2 Almacén de Datos

6.3.7.2.1 Exactitud

Dado que la consulta para este caso tiene también una jerarquía, esta se dividirá en sus elementos.

6.3.7.2.1.1 Ventas por país

La Tabla 55 muestra las ventas agrupadas por algunos géneros para un país.

Tabla 55
Resultados consulta ventas por género en elemento país.

País	Género	Ventas
Afghanistan	(no genres listed)	11335
	Action	7250
	Action|Adventure	2442
	Action|Adventure|Animation	2382
	Action|Adventure|Animation|Children|Comedy|Fantasy	1337
	Action|Adventure|Animation|Fantasy	1004
	Action|Adventure|Animation|Sci-Fi	2825
	Action|Adventure|Children|Fantasy	1481
	Action|Adventure|Comedy	815
	Action|Adventure|Comedy|Fantasy	970

6.3.7.2.1.2 Ventas por estado

La Tabla 56 muestra algunos resultados para ventas por género por estado.

Tabla 56
Resultados consulta ventas por género por región, elemento estado.

País	Estado	Género	Ventas
Afghanistan		(no genres listed)	11335
		Action	7250
		Action|Adventure	2442
		Action|Adventure|Animation	2382
		Action|Adventure|Animation|Children|Comedy|Fantasy	1337
		Action|Adventure|Animation|Fantasy	1004
		Action|Adventure|Animation|Sci-Fi	2825
		Action|Adventure|Children|Fantasy	1481
		Action|Adventure|Comedy	815
		Action|Adventure|Comedy|Fantasy	970

6.3.7.2.1.3 Ventas por ciudad

A continuación, en la Tabla 57 determinados resultados para el escenario planteado y el elemento ciudad.

Tabla 57
Resultados ventas por género por región, categoría ciudad.

País	Estado	Ciudad	Género	Ventas
Afghanistan		Larkird	(no genres listed)	3281
			Action	1668
			Action|Adventure	316
			Action|Adventure|Animation	641
			Action|Adventure|Animation|Children|Comedy|Fantasy	53
			Action|Adventure|Animation|Fantasy	419
			Action|Adventure|Animation|Sci-Fi	715
			Action|Adventure|Children|Fantasy	250
			Action|Adventure|Comedy	225
			Action|Adventure|Comedy|Fantasy	294

6.3.7.2.1.4 Ventas por dirección

Se observan en la Tabla 58 las ventas por género, según la región, para la categoría dirección de la jerarquía.

Tabla 58
Resultados ventas por género, según región, elemento dirección.

País	Estado	Ciudad	Dirección	Género	Ventas
Afghanistan		Larkird	4315 Mallory Crossing	(no genres listed)	3281
				Action	1668
				Action|Adventure	316
				Action|Adventure|Animation	641
				Action|Adventure|Animation|Children|Comedy|Fantasy	53
				Action|Adventure|Animation|Fantasy	419
				Action|Adventure|Animation|Sci-Fi	715
				Action|Adventure|Children|Fantasy	250
				Action|Adventure|Comedy	225
				Action|Adventure|Comedy|Fantasy	294

6.3.7.2.2 Desempeño

6.3.7.2.2.1 Duración de la consulta

Las distintas ejecuciones muestran los datos a continuación:

- Duración máxima: 2438 milisegundos.
- Duración mínima: 1625 milisegundos.
- Duración promedio: 1814 milisegundos.

La visualización del Gráfico 47 tiene los tiempos de cada una de las ejecuciones para este escenario.

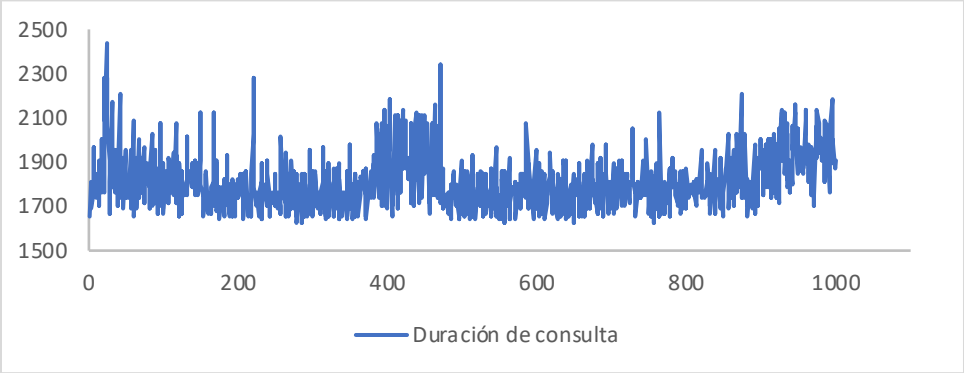


Gráfico 46. Duración de consultas en escenario ventas por género por región.

6.3.7.2.2.2 Utilización de los recursos

Para este conjunto de pruebas el procesador se mantuvo con porcentaje mayor a 90% en su uso, en cambio la memoria de la máquina virtual se conservó ligeramente por debajo del 55%. El Gráfico 48 muestra la información referente a esta consulta.

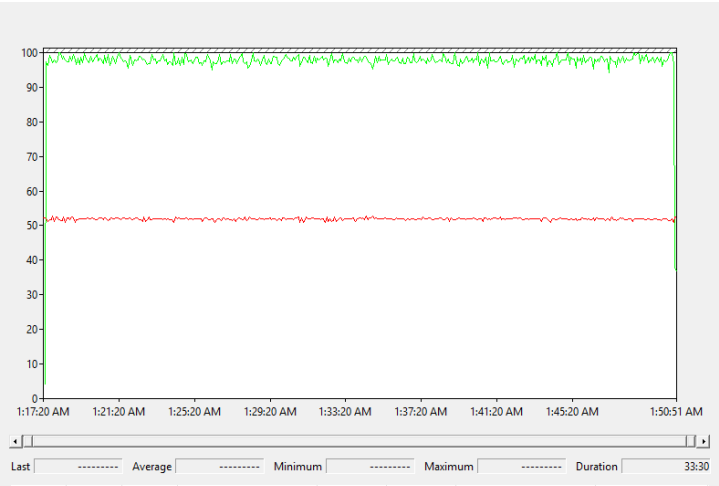


Gráfico 47. Utilización de CPU y memoria en ventas por género y región.

6.3.8 Ventas por tipo de media por región

Esta consulta mezcla el monto de ventas por tipo de media en conjunto a las regiones almacenadas en el sistema. Es decir, ventas por tipos de media que se dieron en una ciudad, estado, o país especifico.

6.3.8.1 Motor de Búsqueda

6.3.8.1.1 Exactitud

Se describen a continuación los diferentes resultados, divididos por jerarquía, para las ventas de tipo de media por región.

6.3.8.1.1.1 Ventas por país

Los resultados categorizados por país se muestran en la Tabla 59.

Tabla 59
Resultados consulta ventas por tipo de media, por país

País	Tipo de media	Ventas
Afghanistan	application/x-troff-msvideo	128855
	audio/mpeg3	133809
	audio/x-mpeg-3	113681
	video/avi	126841
	video/mpeg	216050
	video/msvideo	127882
	video/quicktime	117073
	video/x-mpeg	126824
	video/x-msvideo	125791
Aland Islands	application/x-troff-msvideo	35800

6.3.8.1.1.2 Ventas por estado

De igual manera, se presentan los resultados de las ventas por estado en la Tabla 60.

Tabla 60
Resultados de consulta de ventas por tipo de media por estado

País	Estado	Tipo de media	Ventas
Afghanistan		application/x-troff-msvideo	128855
		audio/mpeg3	133809
		audio/x-mpeg-3	113681
		video/avi	126841
		video/mpeg	216050
		video/msvideo	127882
		video/quicktime	117073
		video/x-mpeg	126824
		video/x-msvideo	125791
Aland Islands		application/x-troff-msvideo	35800

6.3.8.1.1.3 Ventas por ciudad

En la Tabla 61 se muestran los datos de ventas, categorizadas por ciudad

Tabla 61
Resultados consulta ventas por tipo de media, por ciudad

País	Estado	Ciudad	Tipo de media	Ventas
Afghanistan		Larkird	application/x-troff-msvideo	31658
			audio/mpeg3	32371
			audio/x-mpeg-3	29820
			video/avi	32503
			video/mpeg	53942
			video/msvideo	30811
			video/quicktime	28674
			video/x-mpeg	31266
			video/x-msvideo	31500
		Paghmān	application/x-troff-msvideo	32489

6.3.8.1.1.4 Ventas por dirección

En la Tabla 62 se muestran los resultados de consulta de ventas por tipo de media, categorizadas por dirección.

Tabla 62
Resultados consulta ventas por región, categorizada por dirección

País	Estado	Ciudad	Dirección	Tipo de media	Ventas
Afghanistan		Larkird	4315 Mallory Crossing	application/x-troff-msvideo	31658
				audio/mpeg3	32371
				audio/x-mpeg-3	29820
				video/avi	32503
				video/mpeg	53942
				video/msvideo	30811
				video/quicktime	28674
				video/x-mpeg	31266
				video/x-msvideo	31500
		Paghmān	48309 Pepper Wood Place	application/x-troff-msvideo	32489

6.3.8.1.2 Desempeño

6.3.8.1.2.1 Duración de las consultas:

Durante la ejecución de las consultas para establecer la duración de este escenario, se mostraron los siguientes resultados:

- Duración máxima: 173953 milisegundos
- Duración mínima: 29342 milisegundos
- Duración promedio: 77337.439 milisegundos

Como comparación visual, se incluye el Gráfico 49 con los datos obtenidos durante el tiempo de ejecución.

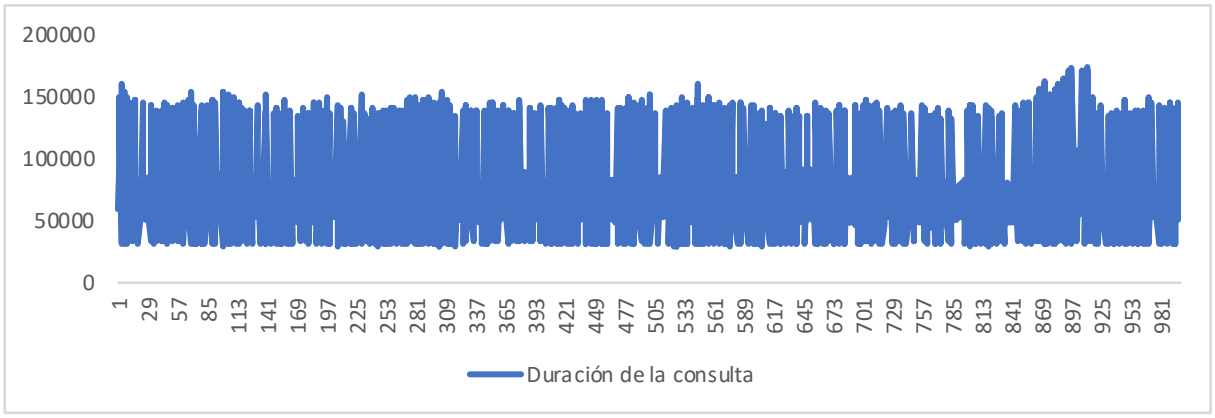


Gráfico 48. Duración de consultas ventas de tipo de media por país

6.3.8.1.2.2 Utilización de los recursos

6.3.8.1.2.2.1Nodo 1

En este nodo se vio un uso normal del procesador, pues no se vieron valores extremos en el mismo. Durante la prueba, se reportó un uso de entre un 16.11% y un 30.24%, lo cual puede ser visto en el Gráfico 50.

Gráfico 49. Utilización del procesador consulta ventas por tipo de media por región

Del lado de la memoria, si se contó con un uso grande de este recurso. Su utilización se encontró entre el 63% y el 92%, es decir, gran cantidad de memoria fue invertida para la resolución de las consultas. Esto se puede visualizar en el Gráfico 51.

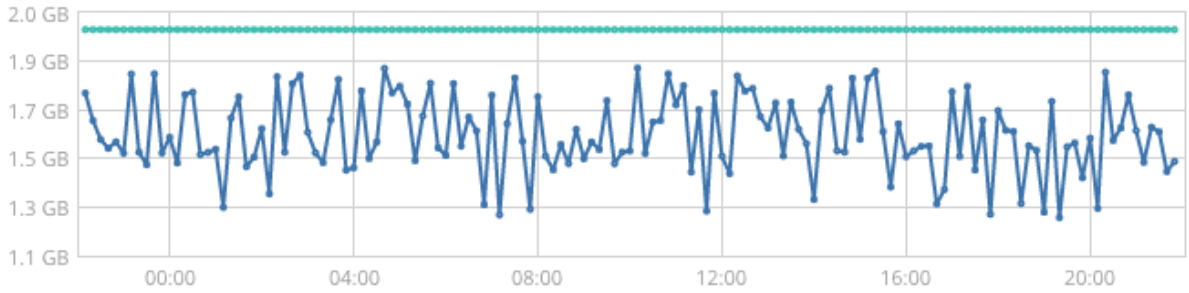


Gráfico 50. Utilización de memoria durante ejecución de consulta ventas por tipo de media por país

6.3.8.1.2.2.2Nodo 2

Para el segundo nodo, se denota un uso alto del procesador, alcanzando un 92.92% del mismo. Su valor mínimo fue de un 63.77%. Su uso se describe visualmente en el Gráfico 52.

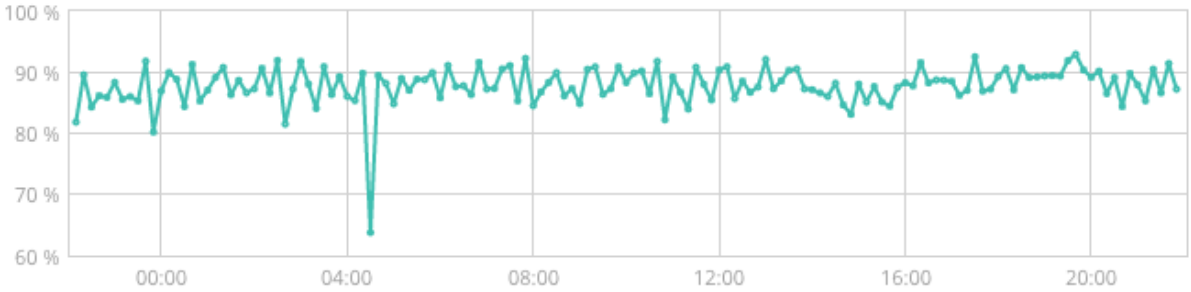


Gráfico 51. Uso del procesador del nodo 2 en consulta ventas por tipo de media por región

La memoria también fue utilizada de manera alta en esta consulta para este nodo. Con un uso mínimo de un 70% y un valor máximo de 97%, es claro ver que la memoria si se encuentra en un uso máximo. Se visualiza en el Gráfico 53.

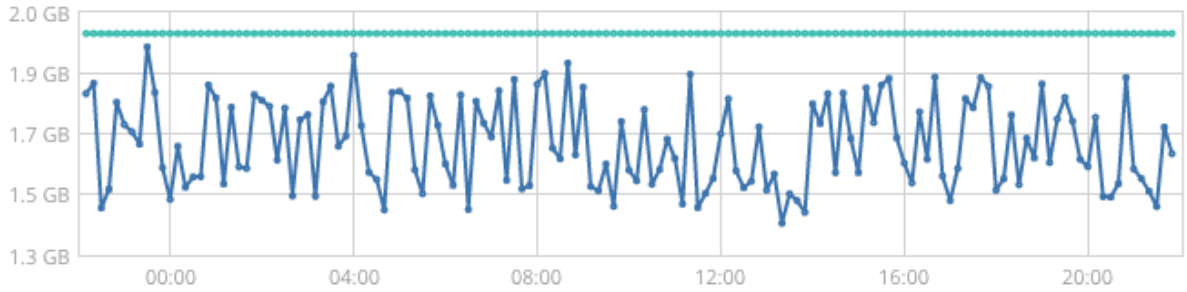


Gráfico 52. Uso de memoria nodo 2 durante consulta ventas de tipo de media por región

6.3.8.2 Almacén de Datos

6.3.8.2.1 Exactitud

Al igual que en otros escenarios con jerarquías, en este se divide por sus elementos.

6.3.8.2.1.1 Ventas por país

Para el elemento país se presentan los resultados de la Tabla 63.

Tabla 63
Resultados ventas por tipo de media por región, elemento país

País	Tipo Media	Ventas
Afghanistan	application/x-troff-msvideo	128855
	audio/mpeg3	133809
	audio/x-mpeg-3	113681
	video/avi	126841
	video/mpeg	216050
	video/msvideo	127882
	video/quicktime	117073
	video/x-mpeg	126824
	video/x-msvideo	125791
Alan Islands	application/x-troff-msvideo	35800

6.3.8.2.1.2 Ventas por estado

La Tabla 64 muestra algunos resultados para el escenario con el elemento estado.

Tabla 64
Resultados ventas por tipo de media por región, categoría estado

País	Estado	Tipo Media	Ventas
Afghanistan		application/x-troff-msvideo	128855
		audio/mpeg3	133809
		audio/x-mpeg-3	113681
		video/avi	126841
		video/mpeg	216050
		video/msvideo	127882
		video/quicktime	117073
		video/x-mpeg	126824
		video/x-msvideo	125791
Alan Islands		application/x-troff-msvideo	35800

6.3.8.2.1.3 Ventas por ciudad

A continuación, en la Tabla 65, se visualizan algunos resultados del escenario para el elemento ciudad de la jerarquía.

Tabla 65
Resultados ventas por tipo de media por región, categoría ciudad

País	Estado	Ciudad	Tipo Media	Ventas
Afghanistan		Larkird	application/x-troff-msvideo	31658
			audio/mpeg3	32371
			audio/x-mpeg-3	29820
			video/avi	32503
			video/mpeg	53942
			video/msvideo	30811
			video/quicktime	28674
			video/x-mpeg	31266
			video/x-msvideo	31500
		Paghmān	application/x-troff-msvideo	32489

6.3.8.2.1.4 Ventas por dirección

La Tabla 66 muestra algunos resultados para este escenario dividido por dirección.

Tabla 66
Resultados ventas por tipo de media por región, elemento dirección

País	Estado	Ciudad	Dirección	Tipo Media	Ventas
Afghanistan		Larkird	4315 Mallory Crossing	application/x-troff-msvideo	31658
				audio/mpeg3	32371
				audio/x-mpeg-3	29820
				video/avi	32503
				video/mpeg	53942
				video/msvideo	30811
				video/quicktime	28674
				video/x-mpeg	31266
				video/x-msvideo	31500
		Paghmān	48309 Pepper Wood Place	application/x-troff-msvideo	32489

6.3.8.2.2 Desempeño

6.3.8.2.2.1 Duración de la consulta

Las pruebas para el escenario de ventas por tipo de media dieron los siguientes tiempos:

- Duración máxima: 1937 milisegundos.
- Duración mínima: 1234 milisegundos.
- Duración promedio: 1375 milisegundos.

A continuación, en el Gráfico 54, se pueden observar los tiempos de cada corrida del escenario:

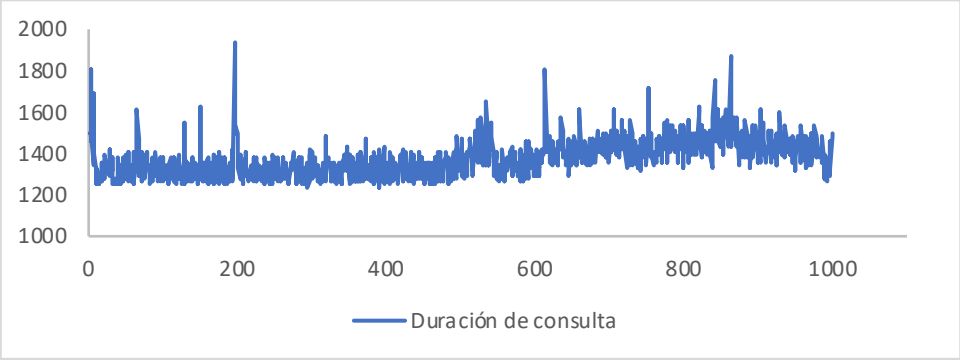


Gráfico 53. Duración de consultas para ventas por tipo de media por región.

6.3.8.2.2.2 Utilización de Recursos

El uso del procesador en este escenario fue consistentemente elevado y la memoria levemente por encima de la mitad de memoria disponible, esto se aprecia en el Gráfico 55.

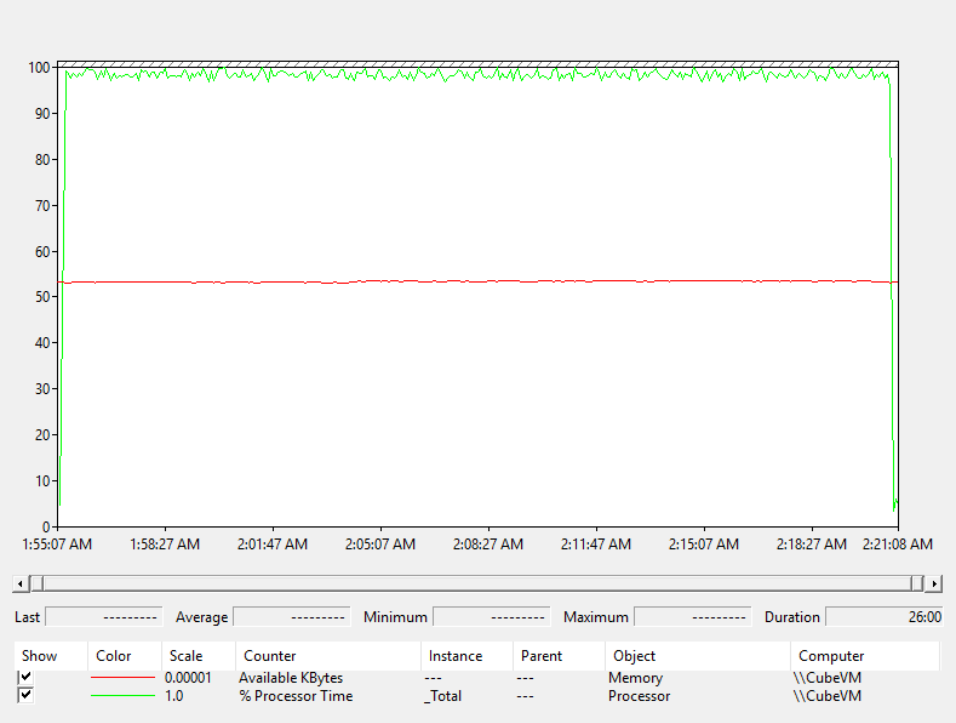


Gráfico 54. Utilización de CPU y memoria escenario ventas por tipo de media por región.

6.4 Análisis de resultados

En esta sección se discutirán brevemente los resultados obtenidos luego de ejecutar las pruebas descritas. Se discutirá por separado cada elemento.

6.4.1 Exactitud

En cada escenario, se tomó la misma muestra tanto para el cubo OLAP como para el clúster de motores de búsqueda, con el fin de comparar adecuadamente en ambas tecnologías.

Dado que fueron utilizadas herramientas distintas para la recuperación de resultados, se tiene diferencia en el formato en el cual son mostrados. Un caso particular de esto es la fecha, pues el motor de búsqueda Elasticsearch retorna las fechas en formato numérico, mientras que el cubo OLAP muestra la fechas en formato textual. Sin embargo, una vez las fechas se convierten a formatos comparables, estas son totalmente equivalentes.

Al cotejar uno a uno los escenarios presentados, se puede observar que las agregaciones en cada una de las granularidades (año, mes, región, etc.) son iguales. Por ejemplo, para el caso específico de ventas por artista, tanto el cubo como el clúster presentan para el artista Adolf Commuzzo un total de 306401. Esto se repite en las muestras tomadas para cada escenario; los valores finales de ventas o cantidades son idénticos para ambas tecnologías, lo cual implica capacidades similares para efectuar cálculos exactos.

Incluso en casos especiales, como la consulta de ventas por género, por región, la cual tuvo problemas de desempeño al ejecutarse las pruebas, es capaz de retornar resultados equivalentes a los del cubo OLAP. Siempre y cuando la consulta del motor de búsqueda esté limitada a un número bajo de registros finales. Esto indica, igualmente que el análisis descrito anteriormente, que la información almacenada en el clúster de Elasticsearch es capaz de ser agregada de manera exacta, y que el motor funciona de manera más eficiente con consultas limitadas a un rango específico o a una cantidad pequeña de resultados a retornar.

Un aspecto el cual debe de ser tomado en cuenta es que en el motor de búsqueda no existe el concepto de jerarquía de datos, tal y como las tecnologías OLAP las implementan. Con el objetivo de emular este comportamiento, fue necesario ejecutar una sola consulta con todos los datos necesarios para todas las partes de la jerarquía. Por ejemplo, en vez de solicitar solamente los resultados por año en el motor de búsqueda, fueron solicitados todos los datos de los años, meses, y días. Similarmente con la jerarquía de región. Sin embargo, es claro que esto no afecta en la precisión de los datos.

En resumen, ambas tecnologías examinadas fueron capaces de realizar cálculos exactos y correctos utilizando los mismos datos.

6.4.2 Uso de recursos

Ambas tecnologías se encontraban alojadas en arquitecturas distintas, esto por la naturaleza de ambas. Un clúster requiere múltiples nodos para ejecutarse mientras un cubo OLAP puede ser ejecutado desde una única máquina.

Para esta comparación, se tomaron en cuenta las métricas del recurso en cada máquina individual pero también el total disponible del mismo en toda la arquitectura.

El clúster de motor de búsqueda tiende a incrementar el uso de memoria RAM. Esto se concluye al observar los gráficos de monitoreo para el clúster, los cuales muestran que durante el tiempo de ejecución de las pruebas hubo un uso creciente de la memoria. Esto no sucede en la máquina que alberga el cubo OLAP, pues su memoria tiene un uso constante y cercano al 50%.

A pesar de que el uso de memoria es estable en el cubo, el porcentaje utilizado representa alrededor de 4GB de memoria RAM, el cual es la misma cantidad total disponible para ambos nodos del clúster. Esto significa que ambas tecnologías emplean, aproximadamente, la misma cantidad de memoria. No obstante, se debe recordar que los nodos tuvieron que ser sometidos a limpiezas de memoria cuando esta alcanzó un porcentaje de utilización elevado e incluso esto llegó a no ser suficiente para finalizar una consulta, como en el caso de ventas por

género por región, en que la memoria se consumió completamente y el servicio quedó sin respuesta.

En el área de procesamiento, ambas tecnologías usan bastante los procesadores. Sin embargo, el cubo OLAP muestra un uso intensivo y constante, llegando al 100% en todas las consultas ejecutadas, mientras que el clúster mantiene generalmente porcentajes menores.

En otras palabras, es posible notar que el uso de un motor de búsqueda es muy intensivo en memoria, forzando a la limpieza de la misma en reiteradas ocasiones, mientras que las tecnologías OLAP son más intensivas en el uso del procesador.

6.4.3 Desempeño de las consultas

Al finalizar cada una de las corridas de un escenario de pruebas se realizó una limpieza de caché, de modo que esta no agilizara la siguiente ejecución y se obligara a recuperar todos los datos nuevamente.

Al llevar a cabo un cotejo, en cada una de las pruebas efectuadas para las dos tecnologías, se puede notar una clara ventaja de las tecnologías OLAP. En el caso específico del escenario ventas por álbum, el clúster tiene una duración promedio de 7.6 segundos, en cambio OLAP cuenta con una duración promedio de 1.1 segundos. En términos generales se ve que, en algunas ocurrencias, para retornar todos los datos solicitados, el cubo dura menos de un segundo. A diferencia

del motor de búsqueda que en ciertas consultas llega a durar hasta más de un minuto para obtener una respuesta equivalente a su contraparte.

Es importante resaltar que existieron casos en que la consulta no pudo finalizar. Nuevamente el ejemplo de la consulta de ventas por género por región no alcanza una respuesta aun cuando la consulta es ejecutada en diversos nodos. Este caso, no obstante, aparenta estar ligado al tamaño de las respuestas y a la cantidad de valores por obtener y no tanto a los cálculos requeridos. Es necesario investigar este caso más a detalle a fin de conocer la causa de este comportamiento en el clúster creado.

En síntesis, para la métrica de desempeño, la tecnología OLAP se mostró superior. Ya que retornó los resultados esperados en tiempos notablemente menores a su homólogo.

Capítulo 7 Conclusiones

Durante esta investigación, fue posible definir los conceptos y contrastar las diferencias de dos tecnologías que intentan cumplir el mismo objetivo: el almacenamiento y procesamiento de datos de una organización específica. En el caso de un almacén de datos, se puede deducir que el objetivo es principalmente ser un depósito histórico de datos, que no permite la actualización o el cambio de estos. Esto ayuda a obtener datos históricos de manera más sencilla a los gerentes de una compañía. Además, el motor de búsqueda está diseñado para facilitar el acceso a los documentos de texto, sirviendo como una especie de filtro entre un usuario determinado y una colección completa de documentos, esto para evitar que un usuario deba leer todos los documentos para poder encontrar lo que necesita.

Si bien es cierto ambas tecnologías trabajan con datos, existen diferencias en cómo son manejadas. Conceptualmente hablando, un depósito de datos divide la información entre hechos y dimensiones, mientras que un motor de búsqueda usa el concepto de documentos. Además, el almacenamiento usado es implementado de manera diferente. Mientras el motor de búsqueda utiliza un índice con los valores necesarios para poder realizar su función, el depósito de datos utiliza otras tecnologías, como bases de datos relacionales para el almacenamiento. De igual manera, el depósito no tiene de manera definida una forma de procesamiento dentro de sí misma, lo que hace necesario la utilización de tecnologías OLAP para realizar

las agregaciones de los datos. Un motor de búsqueda cuenta con capacidades de agregación básicas.

Otra diferencia notable entre las tecnologías comparadas es el uso de los recursos. Durante la investigación, se observa que las tecnologías OLAP son muy orientadas al uso del procesador, haciendo que se utilice este recurso cerca de un 100% en las pruebas ejecutadas. Sin embargo, los motores de búsqueda tienden a utilizar más memoria principal para trabajar. En otras palabras, si se decide utilizar una tecnología sobre la otra, se debe de pensar también en que el equipo donde se va a implementar necesita estar orientado hacia uno de los dos recursos en cuestión.

Cabe acotar que ambas tecnologías utilizan métodos de comunicación diferentes hacia los clientes. Los clústeres generalmente reciben la información por medio de servicios basados en el protocolo HTTP, el cual permite que sean creados clientes de manera sencilla y rápida. Para esta investigación, se utilizaron herramientas de medición de resultados que, aunque no fueron diseñadas específicamente para el motor de búsqueda, fueron capaces de poder extraer resultados y medir el tiempo tomado en cada consulta. Sin embargo, los cubos OLAP utilizan protocolos propietarios que, aunque están documentados y se describen de manera detallada, restringen la creación de un cliente en particular. Como ejemplo, para estas pruebas se debieron utilizar clientes especializados, provistos por Microsoft para extraer una consulta y poder repetirla múltiples veces.

Otra diferencia notable entre las tecnologías es el costo de implementación en términos monetarios. En el caso de la creación de la infraestructura en los proveedores seleccionados, Azure para el sistema operativo Windows y Elastic Cloud para el clúster de motores de búsqueda, se es capaz de ver que una implementación de un clúster es más barata que la implementación de un cubo OLAP. Esto es debido a que una implementación OLAP con las tecnologías escogidas en esta investigación no solo requiere el alquiler del equipo necesario, sino que es necesario adquirir licencias extra para los servidores de procesamiento multidimensional optimizados para producción. Este caso no se cumple con el motor de búsqueda seleccionado, que requiere solamente costo de licencia por el equipo, sin contar sistema operativo, lo cual abarata el costo.

Aún con lo anterior, un cubo OLAP necesita tener un almacén de datos funcional, a fin de que pueda tomar las tablas de hechos y de dimensiones de manera automática. El motor de búsqueda puede ir directo a la capa de inteligencia de negocios, creando una especie de depósito de datos que es capaz de ser analizada y visualizada de manera casi inmediata.

Debido a las diferencias de implementación es difícil comparar un almacén y un motor de manera directa, por lo que es mejor compararlos en términos de usabilidad, razón por la cual se elaboraron escenarios para esto. Los casos propuestos en esta investigación intentan representar consultas comunes que una organización necesita conocer de forma rápida y precisa, con el propósito de

obtener información relevante para el negocio. Para esto se utilizaron agregaciones como sumatorias de ventas, categorizaciones por jerarquías, entre otras.

A la hora de ejecutar los casos propuestos, es notoria una gran ventaja en los tiempos de respuesta que tiene un cubo OLAP respecto a un motor de búsqueda, al tratar de emular dichas funcionalidades. Lo anterior se da debido a que, a la hora de procesar un cubo, existen múltiples rutinas realizadas automáticamente para optimizar las consultas. Estos procesos no se realizan en un índice invertido, el cual solamente analiza el texto para poder encontrar términos de búsqueda, no para realizar cálculos previos de ningún tipo, lo que conlleva a tener que procesarlos en tiempo de consulta, aumentando el tiempo y recursos a utilizar para su resolución. Fue posible ver que los resultados que el usuario recibiría serían equivalentes, cambiando solamente como los mismos son presentados, fallando solamente en la métrica del tiempo de respuesta.

Similarmente, se puede observar que el motor de búsqueda no cuenta con capacidades de crear jerarquías de datos, haciendo necesario que las mismas deban de ser implementadas o emuladas de alguna forma. Dependiendo de la implementación de algunos clientes para motores de búsqueda, esto podría implicar más consultas hacia el índice donde estén ubicados los datos, lo cual generaría más tráfico en la red, o una consulta con los datos de la jerarquía calculados para ser luego mostrados sin necesidad de realizar una nueva conexión, lo cual aumenta el consumo de recursos. No obstante, el hecho de que las tecnologías OLAP permitan

realizar este tipo de manejo de datos de forma nativa hace que sea más transparente para los usuarios que consultan la información de este tipo de sistemas.

Para esta investigación, el resultado es evidente, la mejor forma de consultar grandes cantidades de datos con un mejor desempeño en el tiempo de duración de las consultas, es el uso de las tecnologías OLAP. Sin embargo, también se aprecia que es posible realizar los mismos análisis con una tecnología de motor de búsqueda.

En resumen, es capaz de apreciarse una gran ventaja en la utilización de cubos OLAP para el análisis de grandes datos, siempre y cuando se desee trabajar con la totalidad de los mismos. A fin de poder emular esta capacidad en un motor de búsqueda, es necesario tener más recursos que su homólogo OLAP, o realizar optimizaciones a nivel de los datos. Por otro lado, también es fácil incrementar dichos recursos en un motor con solo añadir nuevas instancias de datos, haciendo que el sistema incremente sus capacidades de procesamiento.

Capítulo 8 Trabajos Futuros

Como continuación de este trabajo y al igual que en cualquier otro proyecto de investigación, quedan abiertas diversas líneas de investigación en las que es posible continuar trabajando. Algunas de estas líneas surgieron directamente del desarrollo de este trabajo, mientras que otras obedecen a cuestiones más generales.

Seguidamente se enumeran trabajos posibles para el futuro:

- Utilizar un clúster de motores de búsqueda con más cantidad de nodos y/o mayor memoria en cada uno de los nodos.
- Ninguna de las dos implementaciones recibió optimización alguna en sus métodos de almacenamiento. Es posible abrir una línea de investigación que explore la posibilidad de mejorar el rendimiento de las consultas aquí propuestas, a fin de mejorar su consumo de recursos o el tiempo tomado para la resolución de estas.
- La comparación de esta investigación se hizo en datos completamente estructurados, representados por tablas en una base de datos relacional. La inclusión de datos no estructurados podría brindar otros resultados de interés.
- Realizar pruebas de análisis en tiempo real. Es decir, como se comportarían ambas tecnologías en caso de que sea necesario que los

datos se encuentren listos para ser agregados al momento de ser ingresados.

- Comparar más productos en ambas tecnologías, utilizando esta investigación como base, a fin de poder determinar si otras ofertas del mercado, tanto de motores de búsqueda como de implementaciones OLAP, ofrecen mejor desempeño que las aquí descritas.

Capítulo 9 Referencias

Alkharouf, N. W., Jamison, C. D., & Matthews, B. F. (2005). Online Analytical Processing (OLAP): A Fast and Effective Data Mining Tool for Gene Expression Databases. *Journal of Biomedicine and Biotechnology*, 181-188.

Cai, X., Acklam, E., Langtangem, H. P., & Tveito, A. (2003). Parallel Computing. En *Advanced Topics in Computational Partial Differential Equations: Numerical Methods and Diffpack Programming* (págs. 1-55).

Clement, L. Y. (2003). Log Analysis as an OLAP Application.

Composite Aggregations. (s.f.). Obtenido de Elastic: https://www.elastic.co/guide/en/elasticsearch/reference/current/search-aggregations-bucket-composite-aggregation.html#search-aggregations-bucket-composite-aggregation

DeWitt, D., & Gray, J. (1992). Parallel database systems: the future of high performance database systems. *Communications of the ACM*, 85-98.

Dubitzky, W., Krebs, O., & Roland, E. (2011). Minding, OLAPing, and Mining Biological Data: Towards a Data Warehousing Concept. .

Goel, E. (2014). Data Warehousing and Data Mining in Business Applications. *Research Cell*, 133-137.

Hawking, D. (2010). Enterprise Search. En *Modern Information Retrieval, 2nd Ed.* (págs. 641-684). Pearson Educational.

Hellum, K. A. (2017). *Information Retrieval using applied Supervised Learning for Personalized E-Commerce.*

Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press.

Perry, J. W., Kent, A., & Berruy, M. M. (1955). Machine literature searching X. Machine language; factors underlying its design and development. *Journal of the Association for Information Science and Technology*, 242-254.

Sree HAri Rao, V., & Murthy, J. (2004). Data Warehousing and Mining An Indispensable Computational Tool for Real World Problems. *Bulletin of the Marathwada Mathematical Society*, 38-52.

Stonebreaker, M. (1986). The Case for Shared Nothing. *Database Engineering*, 4-9.

Tarakeswar, D., & Kavitha, D. (2011). Search Engines: A Study. *Journal of Computer Applications*.

Terms Aggregation. (s.f.). Obtenido de Elastic: https://www.elastic.co/guide/en/elasticsearch/reference/current/search-aggregations-bucket-terms-aggregation.html

Thusoo, A., Shao, Z., Anthony, S., Borthakur, D., Jain, N., Sen Sarma, J., . . . Liu, H. (s.f.). *Data Warehousing and Analytics Infrastucture at Facebook.* Facebook.

Yeo, C., Buyya, R., Pourreza, H., Eskicioglu, R., Graham, P., & Sommers, F. (2006). Cluster Computing: High-Performance, High-Availability, and High-Throughput Processing on a Network of Computers. *Handbook of nature-inspired and innovative computing. Integrating classical models with emerging technologies*, 521-551.

Zhang, J., Zhang, H., Chen, E., Zheng, Y., Kuang, W., & Zhang, X. (2014). Real-time earthquake monitoring using a search engine method. *Nature Communications*.