

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO - CHILE



**“COMPARACIÓN DE RENDIMIENTO ENTRE UNA BASE DE
DATOS RELACIONAL Y UNA COLUMNAR, PARA EL DISEÑO
DE UN *DATA WAREHOUSE* PARA CUMPLO CHILE S.A.”**

ORLANDO PABLO CONTRERAS REBOLLEDO

**MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN INFORMÁTICA**

Profesor Guía: José Luis Martí Lara
Profesor Correferente: Federico Meza Montoya

Marzo - 2020

AGRADECIMIENTOS

Agradezco a todas aquellas personas que me brindaron su apoyo a lo largo de esta etapa, todo hubiese sido mucho más difícil sin ustedes. A mi familia, padre y madre, quienes sacrificaron gran parte de su energía y tiempo en darme la oportunidad de llegar hasta acá, seré un eterno agradecido, más aún, por el amor y la confianza. A mi hermana Loreto, que es para mí un ejemplo de superación ante la adversidad. A mi amor, Vania, que recibiste mucho de lo bueno y de lo malo de pasar por esta etapa juntos, gracias por acompañarme y ser así de fuerte ante la presión que significó a ratos y gracias también a tu familia que siempre tuvo palabras de motivación. A quienes no están, pero que formaron parte importante de este proceso, mi abuela y de forma muy especial a Marcelo Santander, que, sin tu disposición y buena onda para enseñarme esos primeros días, los más difíciles, difícilmente hubiese llegado hasta acá, donde sea que estés amigo, gracias. A mis amigos de universidad, desde plan común hasta quienes me acompañaron en los últimos ramos. Un especial agradecimiento a los 'XLP', quienes fueron y serán mi compañía en largas noches de risas y rabias, como dice un amigo, el cielo es el límite. Finalmente, quiero dedicar este trabajo de título a mi sobrino, Vicente, que todo lo que te depare el mundo sea acogido con amor antes que cualquier otra cosa. Puedes ir o no a la universidad, lo que hagas se que lo harás bien si lo haces con amor y escojas lo que escojas, voy a estar para acompañarte.

RESUMEN

A lo largo del tiempo, arquitecturas enfocadas en el análisis de los datos de las organizaciones, como los *data warehouses*, han sido ampliamente diseñadas en base a modelos estrellas, copos de nieve u otras variantes. Sin embargo, la llegada de las bases de datos NoSQL proponen una mejora de rendimiento sustancial en cuanto a los tiempos de ejecución de las consultas sobre *data warehouses*.

En este trabajo se presenta un estudio comparativo de los tiempos de ejecución de consultas analíticas realizadas sobre una base de datos columnar versus una relacional, con el objetivo de saber qué tipo de base de datos se comporta de mejor forma de acuerdo con los requerimientos de Cumplo Chile S.A. en el diseño de un *data warehouse*. Para la implementación de las alternativas evaluadas se usó MySQL como base de datos relacional y *Cassandra* como alternativa columnar, además de herramientas de visualización como Tableau, *Pentaho Data Integration* para procesos ETL y Hyper como base de datos en memoria primaria.

INDICE DE CONTENIDOS

AGRADECIMIENTOS	2
RESUMEN	3
CAPÍTULO 1: DEFINICIÓN DEL PROBLEMA	6
1.1. Contexto.....	6
1.2. Interesados y actores involucrados	9
1.3. Dificultades.....	9
1.4. Objetivos del trabajo	10
1.5. Alcance	11
CAPÍTULO 2: MARCO CONCEPTUAL.....	12
2.1. Inteligencia de negocios	12
2.2. Data Warehouse.....	13
2.3. Bases de datos relacionales.	13
2.4. Modelo Estrella	15
2.5. Bases de datos NoSQL	15
2.6. Bases de datos columnares y Cassandra	17
2.7. Técnicas de modelado de <i>data warehouses</i> en arquitecturas columnares	19
2.8. Bases de datos en memoria primaria y en memoria secundaria.....	19
2.9. Estado del arte	20
CAPÍTULO 3: DISEÑO DE LA SOLUCIÓN	23
3.1. Modelos de negocio	23
3.2. Diseño de la visualización de datos	25
3.3. Diseño del Modelo Estrella	29
3.4. Diseño del modelo columnar	30

3.5.	Vistas materializadas	31
3.8.	Arquitectura del sistema a construir	34
3.9.	Integración de bases de datos con Tableau Desktop	35
3.10.	Pruebas de rendimiento	35
CAPÍTULO 4: IMPLEMENTACION Y VALIDACIÓN DE LA SOLUCIÓN		37
4.1.	Evidencias.....	37
4.2.	Estructura de las pruebas de rendimiento	39
4.3.	Resultados de las pruebas de rendimiento	42
4.4.	Análisis de pruebas de rendimiento	46
4.5.	Selección de propuesta de solución	47
CONCLUSIONES.....		48
REFERENCIAS BIBLIOGRÁFICAS.....		51
ANEXO A: Procesos de negocio		54
ANEXO B: Códigos SQL		64
ANEXO C: Tablas de resultados		66

CAPÍTULO 1: DEFINICIÓN DEL PROBLEMA

En este capítulo se da a conocer la problemática actual de Cumpló Chile S.A., la empresa de financiamiento colaborativo más grande de Latinoamérica. Además, se enumeran los objetivos y el alcance de este trabajo.

1.1. Contexto

1.1.1. Empresa

Cumpló Chile S.A. es una empresa dedicada al financiamiento colaborativo o *Crowdfunding*; esta organización es catalogada como la empresa más grande de América Latina en el rubro del *Crowdfunding* financiero, basándose en la cantidad de capital transado desde su fundación [1]. Pero ¿Qué es el *Crowdfunding*? Este término se refiere a un mecanismo colaborativo de financiación de proyectos, el cual consiste en poner en contacto promotores de proyectos que demandan fondos mediante la emisión de valores y participaciones sociales o mediante la solicitud de préstamos, con inversores u ofertantes de fondos que buscan en la inversión un rendimiento [2]. En otros términos, Cumpló Chile S.A. provee una plataforma tecnológica, mediante la cual se conecta a pequeñas y medianas empresas que necesitan financiamiento, con inversionistas que ven en cada empresa una oportunidad de inversión con tasas y plazos definidos.

Existen distintos tipos de *Crowdfunding*, los cuales nacen de las diferencias que presenta cada empresa al implementar sus propuestas en base a cada modelo de negocio, manteniendo los lineamientos centrales del financiamiento colaborativo. Algunos tipos de *Crowdfunding* son [3]:

- 1) *Crowdfunding* de donación: consiste en que un grupo de personas aporta capital a una organización; esto se realiza de forma desinteresada y sin esperar nada a cambio. Por lo general los proyectos que son financiados son de carácter social o catástrofes. Gran parte de las veces este tipo de proyecto tiene una duración determinada.
- 2) *Crowdfunding* de recompensa: a diferencia del *crowdfunding* de donación, en este caso el grupo de personas que aporta con capital al proyecto sí espera algo a cambio; por lo general las personas que participan en el financiamiento del proyecto o propuesta tienen garantizado el acceso anticipado al producto o servicio desarrollado o bien, a un menor precio.
- 3) *Crowdfunding* de inversión: en este tipo de *crowdfunding* las personas que aportan en el proyecto obtienen participación en el capital social de la organización. Es por esto que resulta importante que las organizaciones que solicitan financiamiento mediante este método tengan la capacidad de otorgar o repartir beneficios al grupo de inversionistas en forma de sociedad.

Cumplo Chile S.A. se desempeña en el área del *Crowdfunding* de préstamo, también conocido como *Crowdlending*, el cual consiste en un sistema de financiamiento por medio del cual los inversionistas de fondos obtienen rentabilidad mediante intereses y la devolución de la inversión inicial, mientras los prestatarios o empresas solicitantes de crédito obtienen una rentabilidad más atractiva que la que podría obtener mediante un método de financiamiento tradicional.

Cumplo Chile S.A. se fundó el año 2013, cumpliendo hoy en día 6 años en el mercado del financiamiento colaborativo en Chile. El éxito de la compañía a lo largo del tiempo ha sido sustancial, contando con más de 250.000 millones de pesos financiados, en los cuales han participado más de 4.800 inversionistas y, aproximadamente, 1.500 solicitantes de crédito. Además, el año 2018 Cumplo inició un proceso de transformación global, situándose en el mercado mexicano, con lo que busca convertirse, al igual que en Chile, en una alternativa interesante para las Pymes al momento de necesitar financiamiento.

El acelerado crecimiento de la compañía ha propuesto diversos desafíos a lo largo del tiempo, para cada una de las áreas del negocio, demandando grandes transformaciones en periodos cortos de tiempo. Es así como el área tecnológica no ha sido ajena a estas demandas; más aún, dado que el eje central de Cumplo Chile S.A. es una plataforma tecnológica de financiamiento colaborativo, las transformaciones y el crecimiento acelerado proponen desafíos interesantes en los cuales la tecnología es un actor principal.

A lo largo de los años que Cumplo Chile S.A. ha estado en funcionamiento, los cambios en el modelo de negocios de la empresa han sido sustanciales; es por esto que los requerimientos tecnológicos se han vuelto obsoletos en cortos periodos de tiempo, originando nuevas necesidades para ser cubiertas por herramientas tecnológicas de distinta índole. El tipo de situaciones mencionadas ha provocado que los esfuerzos del equipo de tecnología estén dedicados principalmente al desarrollo de software, con lo que se han desarrollado herramientas que funcionan estrictamente dentro del ambiente operacional de la organización, sin tener dominio del ambiente analítico de la empresa.

Conforme Cumplo Chile S.A. ha crecido como organización, comenzando desde un *Start-Up* hasta consolidarse e internacionalizarse, la calidad de software y herramientas tecnológicas también ha sido perfeccionada, en cuanto a la calidad de las aplicaciones y la externalización de servicios.

El crecimiento de Cumplo Chile S.A. puede ser visto en diversos aspectos, por ejemplo, en el ambiente financiero, donde los montos facturados desde sus inicios hasta hoy se han duplicado cada año. Otro aspecto en el cual resulta fácil notar un notable crecimiento, es en la cantidad de empresas e inversionistas que participan de la dinámica del *Crowdlending* para llevar a cabo sus proyectos. Por último, otro aspecto importante que proviene del rápido crecimiento de la organización es la información.

La cantidad de información que maneja la organización a lo largo del ciclo de cada proyecto publicado en la plataforma es muy importante. Detallando, cada proyecto tiene

un conjunto de condiciones de pago distintas, fechas y plazos diferentes para cada conjunto de cuotas, como también tasas y costos especiales de acuerdo a cada producto. Sumado a esto, desde el punto de vista de los inversionistas, también existen asociaciones entre cada inversionista y cada cuota de un producto, generando información relacionada al cruce entre costos, plazos, inversiones y cuotas de forma característica para un conjunto de productos determinado. A estos factores es importante agregar la cantidad de operaciones que se realizan día a día.

Dado lo anterior, es que surge la necesidad de analizar los datos de forma inteligente, para orientar la toma de decisiones con un respaldo basado en la información. Hasta cierto punto resulta posible diagnosticar y orientar el negocio basándose en la experiencia y en algunas métricas que definen, de forma general, cómo se encuentra la empresa; sin embargo, dado el rápido crecimiento de la cantidad de información y el aumento en la complejidad de la operación organizacional, es que tareas como diagnosticar y tomar decisiones se vuelven cada vez más complicadas. Es por lo mencionado, que actualmente se fundan áreas en la organización en el ámbito de la ciencia de datos, ya sea para el análisis del riesgo crediticio como también para la inteligencia de negocios.

Actualmente el área de la inteligencia de negocios es un área profundamente validada y estudiada. Uno de los elementos característicos correspondientes a ésta son las aplicaciones y soluciones tecnológicas que se proponen como área científica para abordar los diversos problemas de cada organización. En Cumpló Chile S.A. hoy en día no se implementa ninguna herramienta con características que permita realizar un trabajo analítico sobre los datos provenientes de la operación, por lo que las decisiones con respecto a la ruta que debe tomar el negocio se realizan basadas en la experiencia y el conocimiento de los líderes de la organización, quienes constantemente realizan evaluaciones y diagnósticos para tomar decisiones ejecutivas con respecto a la empresa. Por otra parte, debido al gran volumen de datos que se maneja en la operación de cada día, resulta necesario obtener información relevante al estado del negocio de forma oportuna, por lo que es importante implementar técnicas específicas de diseño para que los datos sean procesados en el momento justo en cuanto la información es requerida.

1.1.2. Técnica

Como se indica en la sección anterior, la operación de las organizaciones a nivel de información y datos ha sido respaldada por bases de datos durante una gran cantidad de años, siendo las bases de datos relacionales las primeras preferencias al momento de resguardar la información y diseñar la base tecnológica de las operaciones organizacionales. Sin embargo, en los últimos años se ha descubierto que no necesariamente las bases de datos relacionales son las mejores para todo el flujo de información a través de la organización. Es por esto por lo que se usan formas de guardar la información, lejos de los esquemas clásicos relacionales de almacenamiento de datos, para dar paso a los esquemas *NoSQL* y *NewSQL*.

Algunos estudios [17] sugieren que las bases de datos basadas en columnas tienen un mejor rendimiento al momento de sostener plataformas de análisis de datos.

1.2. Interesados y actores involucrados

El tema a tratar en este estudio tiene dominio específicamente en el área de la inteligencia de negocios y está orientado de forma puntual al procesamiento de información útil para la toma de decisiones a nivel organizacional. Algunos de los interesados, actores y usuarios de la solución a proponer son:

- 1) Área de ciencia de datos: la fundación de un área destinada a la ciencia de datos es de carácter reciente en la organización, dado que el esfuerzo a lo largo del tiempo que la organización lleva operando ha sido orientado al área de desarrollo. Dado esto, el estudio realizado propone uno de los primeros avances en términos de diseño de un *data warehouse* para la empresa.
- 2) Gerencia: organizacionalmente la empresa se divide en áreas operativas, lideradas cada una por un gerente en particular. Además, cada gerencia está supeditada a una gerencia general, la cual transmite requerimientos basados en su visión del negocio conforme a una futura ejecución por parte de las áreas respectivas. Es por esto que el estudio realizado propone grandes bases para la toma de estas decisiones, fundamentándolas en datos.
- 3) Tecnología: una de las características del área de inteligencia de negocios, es el respaldo tecnológico en base a plataformas y *suites*, que se ofrecen para el procesamiento de datos en las distintas etapas del flujo diseñado. Resulta fácil enumerar diversas herramientas para la inteligencia de negocios, como muchas otras para la realización de *ETL*¹. Es por esto que el área tecnológica es protagonista de la solución propuesta, dado que es responsabilidad de esta área entregar soporte constante al funcionamiento de cada herramienta.
- 4) Directorio: son el directorio de la empresa y socios los principales interesados de los resultados que obtiene la empresa. De forma reiterada requieren de información clara y precisa con respecto al estado de la organización y su crecimiento. Es por esto que parte del resultado que se entrega mediante las aplicaciones de inteligencia de negocios junto a la solución propuesta, es de especial interés de este sector.

1.3. Dificultades

Actualmente en la organización, muchas de las decisiones ejecutivas que se toman son fundamentadas y respaldadas por la experiencia, gran capacidad de liderazgo y extenso conocimiento de la industria por parte de los directivos de la organización. No obstante, conforme la organización y el contexto en el cual la empresa se desempeña crece, resulta

¹ Sigla en ingles para Extract Transfor and Load

cada vez más complejo tener noción del panorama completo, dado que son cada vez más variables las que conforman el espectro de toma de decisiones. Por otra parte, existen decisiones que se toman en el transcurso del día, con la finalidad de orientar la operación de la organización a mejores estados contribuyendo al crecimiento constante. Es por esto que de no existir la solución propuesta se produce un riesgo para la organización en diversos aspectos, algunos de los cuales se enumeran a continuación.

- 1) Ventaja competitiva: el mercado de las empresas financieras tecnológicas (*Fintech*) es considerablemente más grande cada año. En algunos países, entre ellos México, existe actualmente una regulación a nivel gubernamental para las empresas *Fintech*, lo que hace notar cuánto crece este negocio en América latina. Hoy en día, en Chile, existen numerosas empresas que promueven distintos tipos de *Crowdfunding* como alternativa a las formas clásicas de financiamiento, con lo que se generan buenos niveles de competencia en el mercado, haciendo importante el uso de tener datos para tomar decisiones y orientar a la organización hacia el crecimiento continuo. Es por esto que, de no existir una solución a proponer, se pierde una ventaja competitiva con respecto al resto de las organizaciones.
- 2) Sesgo en el diagnóstico: en las actividades operativas que realiza la organización todos los días, surge la necesidad, en diversas ocasiones, de tomar decisiones con respecto a la orientación de la organización. Muchas de estas decisiones son tomadas en reuniones con las personas con conocimiento y experiencia en el área específica. Sin embargo, al no tener una base que se respalde en la información para analizar el estado de la organización y desarrollar un diagnóstico de la problemática, se pueden tomar decisiones erradas que lleven a la empresa a destinos alejados de lo definido en sus objetivos estratégicos de esta.

1.4. Objetivos del trabajo

Se propone como solución el diseño de una plataforma destinada al análisis y la toma de decisiones para un caso real, correspondiente a Cumplo Chile S.A., para lo cual se realizará una comparación de rendimiento entre dos tipos de base de datos, un modelo columnar versus un modelo relacional, usando como motores de base de datos Cassandra y Mysql respectivamente.

Luego, el objetivo general de este trabajo es disminuir los tiempos de análisis de las áreas del negocio modeladas, mediante un mejor rendimiento en las consultas realizadas por la plataforma de inteligencia de negocio.

A su vez los objetivos específicos asociados son:

- 1) Comparar el rendimiento de consultas entre modelos lógicos columnares y relacionales, para comprobar una mejora en los tiempos de ejecución de los modelos columnares.

- 2) Diseñar un *data warehouse* que tenga como base los modelos previstos y se integre a las tecnologías mencionadas, para entregar una plataforma que apoye la toma de decisiones con base en los datos.
- 3) Analizar y comparar los resultados obtenidos con el fin de contribuir con la implementación de estas tecnologías en un caso real, aportando con esto a la literatura.

1.5. Alcance

La solución por desarrollar limita principalmente con las áreas de negocio pertenecientes a la organización que serán modeladas para realizar la comparación; por esto se escoge sólo un área de negocio en particular. Además, se escoge una plataforma de visualización de datos en vez de desarrollar visualizaciones personalizadas, dado que el objetivo de esta memoria es realizar una comparación entre los modelos mencionados y no sobre su rendimiento a nivel de visualización. Finalmente se utiliza MySQL, debido a que es el motor de base de datos relacional que se usa actualmente en la organización, así por la perspectiva columnar se opta por el motor de base de datos Cassandra debido a ser uno de los motores que cuenta con mejores evaluaciones, dejando la posibilidad de comparar estos aspectos con otras alternativas en estudios futuros.

CAPÍTULO 2: MARCO CONCEPTUAL

En este capítulo se dan a conocer los principales aspectos técnicos en los cuales se basa la solución propuesta, con el objetivo de proporcionar una base de conocimiento técnico para la comprensión de la implementación de la propuesta de solución. El tema principal sobre el que se define esta memoria es la Inteligencia de negocios.

2.1. Inteligencia de negocios

La inteligencia de negocios es un área de la informática, la cual está compuesta por un conjunto de procesos, aplicaciones y tecnologías, que permite acceder a información relevante para la toma de decisiones y al análisis organizacional de forma rápida y sencilla, contribuyendo con un ambiente bajo el cual se puedan desempeñar labores alineadas a la toma de decisiones en una organización en particular. Es importante mencionar que las arquitecturas sobre las cuales se realizan procesos de inteligencia de negocios son especialmente diseñadas para estos efectos, de forma que se diferencian de forma significativa de las arquitecturas que soportan la operación cotidiana de la organización; es más, usualmente es necesario realizar procesos de extracción, transformación y carga de datos (ETL) [4], desde las arquitecturas operacionales y/o distintas fuentes de datos hacia arquitecturas diseñadas para el análisis, como se muestra en la figura 1.

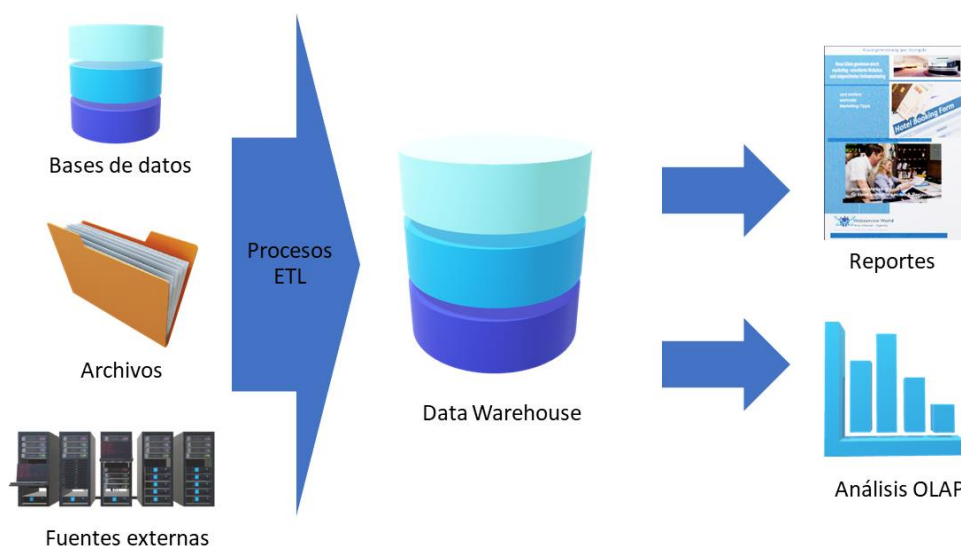


Figura 1: Diagrama de proceso de inteligencia de negocio y sus etapas, desde fuentes de datos a visualizaciones, de autoría propia.

Estas arquitecturas de almacenamiento de datos, con orientación al análisis de datos son denominadas *Data Warehouses*.

2.2. Data Warehouse

Un *data warehouse* es un almacén de datos, el cual corresponde a uno de los elementos centrales de la inteligencia de negocios [5]. Esta arquitectura está especialmente diseñada para soportar el análisis sobre la organización. Estas bases de datos se diferencian de esquemas de datos relacionales dado que no usan la información sobre la cual opera la empresa diariamente. Los *data warehouses* son colecciones de datos usualmente enriquecidos por procesos de carga denominados ETL, los cuales extraen información desde las bases de datos transaccionales y operativas, la cual es opcionalmente enriquecida por fuentes de datos externas.

A lo largo del tiempo se han desarrollado diversas técnicas para el diseño de los *data warehouses*; algunas de las más conocidas corresponden a modelos multidimensionales, dentro de los cuales se pueden nombrar a los de tipo estrella, copo de nieve o constelaciones.

Para el diseño de un *data warehouse* que soporte el desarrollo de las plataformas analíticas de la organización existen dos grandes corrientes. Una de estas corrientes es la descrita por Bill Inmon [5], quien propone un enfoque *top-down* de diseño, aludiendo a que en una primera instancia se debe diseñar el *data warehouse* completo de la organización, de forma que se tomen en cuenta todos los datos corporativos para así, finalmente, dar paso a la creación de los *data marts* que se enfoquen en áreas específicas del negocio. Por otra parte, Ralph Kimball [5], estipula que una de las mejores formas de diseñar un *data warehouse* es con un enfoque *bottom-up*, donde en una primera instancia se desarrollan los *data marts* específicos de áreas atómicas del negocio y mediante la unión o integración de los distintos *data marts* se da origen al *data warehouse* corporativo. De acuerdo con esto, y asociado a la naturaleza de la organización en estudio, la cual se puede describir como una cadena o flujo de valor sumamente definido en áreas con responsabilidades atómicas sobre cada proceso, considerando además que la evolución y constante cambio del negocio demandan grandes esfuerzos en cuanto a mantenibilidad de las herramientas, es que, para este trabajo, a modo de ejemplo, una opción interesante sería una modalidad de desarrollo *bottom-up*.

2.3. Bases de datos relacionales.

Una relación representa una tabla que no es más que un conjunto de filas, donde cada fila es un conjunto de campos y cada campo representa un valor [6]. El modelo de datos relacional fundamenta su idea en el uso de relaciones, considerándose la base de datos con un conjunto definido de relaciones. La información almacenada en este tipo de bases de datos es recuperada usualmente mediante consultas.

Algunos de los elementos las bases de datos relacionales son:

- 1) Relaciones: cada dato en una base de datos relacional es accedido mediante sus relaciones. Existen diversos tipos de relaciones; las relaciones base y las

relaciones derivadas. Las relaciones base corresponden a las relaciones que almacenan datos y su implementación son las tablas. Las relaciones derivadas no almacenan datos, pero se calculan mediante la aplicación de operaciones a través de vistas o consultas [6].

2) Claves:

- Clave primaria: es un identificador único de un elemento en una tabla; esta puede ser conformada por uno o varios elementos de la tabla y distingue a este elemento de cualquier elemento dentro de la misma tabla [6].
- Clave foránea: hace referencia a la clave de un elemento de otra tabla, generando así la relación entre cada una de las tablas. Estas no necesitan ser únicas, como las claves primarias [6].

3) Procedimientos almacenados: hace referencia a una sección de código que está almacenada en la base de datos, usualmente son usados para realizar acciones comunes como la inserción de registros en alguna tabla o la recopilación de información estadística [6].

Por ejemplo, en un sistema de almacenamiento de datos relacional los datos son almacenados en estructuras bidimensionales denominadas tablas, compuestas por filas y columnas donde cada fila es guardada contiguamente. Una tabla en un modelo de datos relacional puede tener la siguiente estructura:

Id	Nombre	Edad	País	Sueldo
1	Vania	23	Chile	100000
2	Ana	43	Argentina	200000
3	Marcelo	45	Alemania	300000

Tabla 1: Representación conceptual de estructura de base de datos relacional.

En el sistema de base de datos relacional esta estructura sería almacenada por filas, donde cada fila de la tabla hace referencia a una cadena, a la cual se le atribuye un identificador interno de cada fila, el cual en este caso se denominará “RowId”, teniendo la siguiente estructura:

RowId: Id, Nombre, Edad, Pais, Sueldo;

De forma que, aplicado a los datos del ejemplo, se guardarían múltiples filas con la estructura indicada:

001: 1, Vania, 23, Chile, 100000;
002: 2, Ana, 43, Argentina, 200000;
003: 3, Marcelo, 45, Alemania, 300000;

Esta forma de almacenar los datos tiene grandes ventajas cuando el desafío a abordar involucra conocer toda o gran parte de la información de un registro o fila.

Si bien las bases de datos relacionales se ocupan, en gran medida, para almacenar la información de la operación de la organización, con una fuerte correlación entre la definición del modelo lógico y las características del negocio. Las bases de datos relacionales también se pueden diseñar de forma que entreguen grandes facultades para el análisis de datos. Desde esta perspectiva analítica nacen los modelos de datos tipo estrella.

2.4. Modelo Estrella

Un modelo estrella corresponde a una técnica de diseño de bases de datos relacionales, se usan de forma recurrente en el diseño de bases de datos con fines analíticos [7] y son frecuentemente poblados por procesos ETL. Algunos de los elementos centrales de este tipo de modelo son la tabla de hechos y las dimensiones:

- 1) Tabla de hechos: diseñada para almacenar datos generalmente numéricos, es en esta tabla en donde se registran las métricas o medidas de un evento correspondiente al negocio en específico. Por ejemplo “Se giraron US\$30000 dólares a un cliente sobre un crédito de tipo hipotecario compuesto de 20 cuotas con una tasa de interés de 3% mensual en la ciudad de Jalisco, México, el día 20 de marzo de 2017 a las 19:00 horas” [7].

La granularidad que contiene la tabla de hechos corresponde al detalle con el cual se almacenan los datos y depende de la granularidad de las dimensiones.

- 2) Dimensiones: las dimensiones de un modelo estrella son múltiples tablas, las cuales tienen un bajo número de registros en comparación con la tabla de hechos y se encargan de almacenar atributos que describen la tabla de hechos.

Sin embargo, se han desarrollado nuevas técnicas de diseño de bases de datos, bajo las cuales, de acuerdo a las necesidades del negocio, se pueden obtener mejores resultados. Esto ocurre principalmente con el uso de las bases de datos NoSQL.

2.5. Bases de datos NoSQL

Las bases de datos NoSQL (por sus siglas en inglés ‘*Not only SQL*’), hacen alusión al desarrollo de una nueva forma de almacenar datos que toma distancia en base a diferencias sustanciales con el modelo relacional de almacenamiento. Este tipo de base de datos encapsula diversos tipos de técnicas y tecnologías, las cuales nacen desde la necesidad impuesta por las aplicaciones modernas en cuanto a procesamiento de información. Es decir, las nuevas aplicaciones y el nacimiento de conceptos como el *big data*, entre otros, propusieron desafíos que como respuesta requieren nuevas tecnologías

y estrategias para solucionar de manera óptima los problemas asociados, particularmente en la forma en que se almacenan los datos. Debido a esto, emergieron motores de bases de datos acordes a ciertos desafíos en común, por ejemplo, para algunos casos no era necesario mantener un nivel de consistencia estricto en cuando a los datos. No obstante, otros negocios requerían tiempos de respuesta mínimos al momento de realizar lecturas, mientras otra parte otras áreas del negocio necesitaban realizar una gran cantidad de inserciones o edición de registros en el menor tiempo posible. Una clasificación básica de estos tipos de base de datos, como también algunos ejemplos de motores de base de datos, es la siguiente [8]:

- 1) Clave-Valor: las bases de datos categorizadas como clave-valor son similares en gran medida a una estructura de datos de tipo diccionario, donde los datos son encapsulados y accedidos mediante una clave única. Estos datos difieren en la forma en la cual se almacenan del modelo relacional, en la medida en que no resulta necesario determinar una estructura ni tipo para los campos que se almacenarán, sino que éstos se dejan variables. Es decir, los valores accedidos por una clave pueden ser de tipo numérico o cadenas de caracteres, mientras que otra clave puede tener asociada valores solo de tipo numérico, que no tengan relación entre claves. Uno de los motores de datos más reconocidos de esta categoría es *“Amazon Dynamo DB”* [9].

Clave	Valor
K1	AAA, 123, BBB
K2	AAA, DDD

Tabla 2: Representación conceptual de estructura de base de datos clave-valor.

- 2) Documental: las bases de datos documentales guardan datos en forma de documentos. Algunos de los formatos más usados son JSON y XML, puesto que ambos proveen estructuras de orden o jerarquías mediante etiquetas. Una de las principales ventajas que definen a este paradigma es la eventualidad en la presencia de los campos, puesto que es totalmente flexible en este aspecto. Un documento puede tener o no un campo que existe en otro documento, a diferencia del modelo relacional, donde los campos son establecidos en la definición del modelo. Además, los objetos son guardados completos en la base de datos, con lo que se vuelve prescindible la necesidad de un mapeado objeto relacional. Uno de los motores mas reconocidos de esta categoría es *“Mongo DB”* [10].
- 3) Columnar: las bases de datos columnares a basan su estructura en columnas. En estas bases de datos, cada llave es asociada con una columna lo que permite un gran desempeño en la ejecución de agregaciones. Uno de los motores más reconocidos de esta categoría es *“Apache Cassandra DB”* [11].

- 4) Grafos: Las bases de datos de grafos se estructuran de nodos, los cuales representan cada clase que se desee almacenar, como también de vertices que representan las relaciones entre cada nodo o clase. Estas bases de datos son ampliamente usadas en escenarios donde la relación entre los datos es de gran importancia. Uno de los motores más reconocidos de esta categoría es “Neo4j” [12].

Para esta memoria y por las características que su diseño aporta a el diseño de bases de datos con orientaciones analíticas, parece pertinente ahondar en detalle sobre las bases de datos columnares.

2.6. Bases de datos columnares y Cassandra

Las bases de datos columnares tienen sus primeras raíces en técnicas implementadas a mediados de 1970, con los primeros estudios de técnicas de archivos transpuestos y particionamiento vertical como indicios de lo que serían en un futuro las técnicas de *clustering* en bases de datos [13].

La orientación de bases de datos a modelos columnares propone grandes beneficios de acuerdo con la naturaleza del desafío a abordar, dado que esta forma de almacenamiento de datos tiene grandes diferencias estructurales con respecto al modelo relacional. Por ejemplo, un modelo relacional resulta poco eficiente cuando se desea obtener información sobre un grupo pequeño de columnas, dado que esto significaría realizar una extracción y filtro de grandes volúmenes de datos, obteniendo mas información de la que es requerida. Esta es la principal razón por la que se usan las bases de datos orientadas a columnas; por ejemplo, la estructura señalada en la tabla 1 puede ser almacenada en registros contiguos según la columna que el usuario desee o que el problema a resolver requiera; de esta forma una estructura tentativa de almacenamiento sería:

001: Vania, 002: Ana, 003: Marcelo;

001:100000, 002:200000, 003:300000;

Se pueden percibir ventajas sustanciales de un modelo sobre otro de acuerdo con el tipo de consulta que se requiera realizar. Por ejemplo, si se quiere analizar los sueldos promedio de la totalidad de los registros, en el modelo relacional se debe acceder a cada uno de los registros de donde sea posible extraer el sueldo de cada usuario, para de esta forma calcular el campo requerido. Sin embargo, en el modelo columnar es suficiente con realizar una agregación sobre la columna que contiene el campo respectivo, sin necesidad de acceder al resto de los campos.

Uno de los motores de bases de datos mas conocidos de este paradigma de diseño de bases de datos es Cassandra, según lo indica el ranking ‘*db-engines.com*’ [14], el cual se dedica exclusivamente a medir la popularidad de los distintos motores de bases de datos, considerando relevantes aspectos como el número de menciones en páginas web mediante el número de resultados arrojados en motores de búsqueda como Google, Bing

y Yandex, el interés mediante la frecuencia de búsquedas obtenidas por Google Trend, así como también la frecuencia de discusiones técnicas sobre el motor extraídas desde Stack; Overflow y DBA Stack Exchange. Se consideran, además, el número de ofertas de trabajo, perfiles en redes profesionales donde el motor es mencionado y su relevancia en redes sociales como Twitter. Según la ponderación de todas estas características Cassandra se posiciona ampliamente en el primer lugar, manteniendo grandes diferencias con otros sistemas de bases de datos de la misma categoría, como se muestra en la tabla 3.

Posición	Sistema	Tipo	Puntaje
1	Cassandra	Columnar	120,66
2	HBase	Columnar	53,34
3	Microsoft Azure Cosmos DB	Multi-Modelo	31,51

Tabla 3: Ranking de bases de datos multi modelo y columnar, enero 2020 [11].

Algunas de las principales características de Cassandra tienen que ver con la capacidad de distribuir información en distintos nodos; esto lo convierte, también, en una base de datos distribuida, con lo cual se tienen ventajas como:

- 1) Replicación: Cassandra soporta replicación, con lo cual se pueden tener múltiples *data center* alojados en distintas posiciones geográficas.
- 2) Tolerancia a fallos: por el hecho de ser una base de datos distribuida no tiene un punto único de fallo, con lo cual en caso de existir un fallo no se compromete la totalidad de los datos almacenados, sino que parte o ninguno de ellos, según la estrategia utilizada en cuanto a la distribución o replicación.
- 3) Lenguaje de consultas: Cassandra incorpora un nuevo lenguaje de consultas denominado CQL (*Cassandra Query Language*), el cual proporciona una interfaz de fácil acceso a los datos de forma alternativa, pero similar en muchos aspectos al tradicional SQL (*Structured Query Language*).
- 4) Consistencia eventual: Cassandra es un motor de base de datos eventualmente consistente, lo que ayuda a alcanzar una alta disponibilidad de sus nodos.

Las características técnicas mencionadas sobre el motor de base de datos Cassandra han sido ampliamente validas por diversas compañías en el ámbito de la tecnología; es así como empresas como *Apple* utiliza 100.000 nodos de Cassandra, la aplicación de comunicación en videojuegos *Discord* migró a Cassandra billones de mensajes el 2015, Netflix usa este motor como *Backend* para su servicio de *streaming*, además de otras compañías como Uber, Reddit, Cisco, entre otros.

Dado a que la forma en la cual se guardan los datos en este tipo de bases de datos difiere de forma sustancial con respecto a las maneras mencionadas y consideradas como tradicionales, aludiendo a modelos estrella contruidos sobre bases de datos relacionales, resulta necesario proponer técnicas de modelado asociadas al paradigma columnar.

2.7. Técnicas de modelado de *data warehouses* en arquitecturas columnares

Dados los beneficios de las arquitecturas basadas en columnas con respecto al desempeño de las consultas según su tipo, resulta posible inferir buenos comportamientos también en plataformas asociadas a dicho comportamiento. La principal diferencia de estas arquitecturas con respecto a las arquitecturas tradicionales relacionales radica en cuales son los objetivos que se quiere alcanzar, ya sea conocer información variada aludiendo a varios campos de un conjunto delimitado de registros, o consultar información general o proveniente de agregaciones de ciertos campos en específico, con gran volumen de registros. Este último caso es recurrentemente observado en las plataformas analíticas, donde generalmente las organizaciones requieren del análisis y cálculo sobre grandes volúmenes de datos, y ciertos campos que permiten resumir el comportamiento de ciertas áreas en específico de cada negocio. A lo largo del tiempo, se ha encontrado la forma de tener buen desempeño en estas plataformas de análisis, mediante el uso de técnicas de diseño basadas en modelos relacionales, tales como modelos estrella, copos de nieve o constelaciones; sin embargo, con la llegada de las arquitecturas basadas en columnas se ha propuesto una serie de técnicas que mejoran el desempeño de las consultas analíticas al margen de los modelos relacionales. Una de ellas propone migrar desde un modelo estrella a una gran tabla de hechos que contenga toda la información de las dimensiones desnormalizada de forma columnar [17]; otra propuesta es imitar el comportamiento de un modelo estrella para una arquitectura relacional, dejando así cada dimensión del modelo estrella en una familia de columnas teniendo una familia de columnas aislada para la tabla de hechos [17]. Estas técnicas se encuentran ampliamente estudiadas y los resultados, según indican los estudios realizados [16] [17], concuerdan con lo esperado al momento de realizar consultas de carácter analítico sobre las columnas, entregando mejores resultados en termino de desempeño en cuanto al tiempo de ejecución de las consultas en modelos relacionales. Sin embargo, las mejoras en el desempeño de la ejecución de las consultas también pueden variar dependiendo de otras variables, como el nivel de memoria sobre el cual se encuentran las bases de datos.

2.8. Bases de datos en memoria primaria y en memoria secundaria

Además de tener distintos tipos de modelos y arquitecturas de bases de datos, los avances tecnológicos en cuanto al almacenamiento físico de los datos han permitido considerar que estos tipos de arquitecturas no solo existan en memorias secundarias, sino que también apoyen la posibilidad de diseñar y llevar a cabo la construcción de bases de datos en memoria primaria. Uno de los principales desafíos para este tipo de arquitectura radica en como alcanzar la persistencia de datos considerando que la memoria primaria es volátil. A modo de ejemplo, el proyecto *Hyper* [18] alcanza este objetivo mediante la captura de lo que se denomina un *snapshot*, que es una fotografía del estado de la base de datos en un momento dado; además, las transacciones que llevan a la base de datos a

estados posteriores, ya sea por inserciones o ediciones de sus registros son guardadas en *logs*; de esta forma, cada vez que se quiere construir esta base de datos, se ejecutan las transacciones almacenadas en el *log* de transacciones, sobre el *snapshot* de la base de datos. Esta estrategia de almacenamiento propone grandes beneficios sobre el rendimiento de las consultas, dada la rapidez de la memoria primaria en comparación con la memoria secundaria.

2.9. Estado del arte

En la actualidad, gran parte de las organizaciones cuenta con sistemas que soportan, si no de forma completa, una parte de la operación. Es por esto que las organizaciones generan y guardan información la cual contiene la historia de la organización y se puede explotar a favor de la toma de decisiones. El poder competitivo de una organización se basa en la cantidad de información que sea capaz de usar en la toma de decisiones [20]. Para sacar provecho de la información relativa a la operación de la organización se hace uso de metodologías y herramientas correspondientes a la Inteligencia de negocios, en la cual se contempla el uso de arquitecturas de almacenamiento diseñadas de acuerdo a las necesidades de la empresa, como son los *data warehouses*, los cuales entregan la información necesaria para ser desplegada en *dashboards* que facilitan la interpretación de los distintos datos para respaldar la toma de decisiones.

- 1) Inteligencia de negocios: es un conjunto de técnicas y herramientas que se ha utilizado a lo largo del tiempo con el fin de hacer un uso apropiado con respecto a la interpretación de los datos de la organización con el fin de sacar provecho de estos, transformándolos en un respaldo para la toma de decisiones. La implementación de herramientas de inteligencia de negocios produce un efecto positivo en cuanto a una ventaja competitiva, tanto en la gestión interna de la organización como ante competidores externos [20].

Actualmente se proponen distintas técnicas, la cual enumera una serie de lineamientos a seguir para alcanzar la obtención de valor para la organización mediante la implementación de inteligencia de negocios. Estos lineamientos son: Adaptabilidad continua, Trabajo continuo, Jerarquías con respecto al desarrollo y manejo de las herramientas propuestas por la inteligencia de negocios. Una de estas metodologías es denominada *Agile BI Governance* flexibles y Personas antes que procesos [21].

- 2) *Data warehouse*: para la correcta implementación de inteligencia de negocios en las distintas organizaciones, es necesario el uso de una base de datos que almacene la información bajo un diseño exclusivo orientado en el análisis de estos. Bajo esta premisa, existen diversas formas de diseñar estas bases de datos; por un lado, se encuentran perspectivas consideradas clásicas, puesto que se han usado durante un periodo largo de tiempo hasta el día de hoy, como también, existen técnicas de diseño basadas en nuevas tecnologías. Es en este punto donde se encuentran modelos de base de datos columnares,

provenientes del auge de las tecnologías NoSQL, como buenas alternativas al momento de implementar las bases de datos [22].

Independientemente de las técnicas de diseño que se decidan utilizar para el desarrollo de las distintas bases de datos, existen procesos que son indispensables al momento de construir un *data warehouse*. Es aquí donde resulta sumamente necesario realizar un estudio acabado con respecto a los requerimientos de la plataforma a construir, su alcance y el valor que se puede obtener de la información disponible. De esta forma toma suma importancia las entrevistas que se puedan realizar con los respectivos interesados, conforme se desarrollan modelos entidad relación que describan de la forma mas completa posible el área a modelar, con el fin de obtener un panorama general detallado de las dimensiones a abordar.

- 3) *Dashboard*: Finalmente, una parte importante de la implementación de inteligencia de negocios en una organización es el desarrollo de la visualización que permita el análisis de los datos modelados y almacenados. Hoy en día existen diversas plataformas, que permiten la conexión con distintas fuentes de datos y proponen formas interesantes de visualizar la información. Algunas de estas promueven grados de libertad en la persona encargada de analizar, permitiendo hacer uso de filtros dinámicos o de la selección de las variables a observar, mientras otros tipos de visualizaciones resultan ser mas rígidos y estandarizados.

Las áreas de la organización involucradas en el análisis mediante inteligencia de negocios suelen abarcar diversos números de variables, con lo cual se genera una gran variedad de gráficos que ayudan a facilitar la visualización de los datos y centran el esfuerzo en un correcto análisis para la toma de decisiones. De esta forma, los distintos gráficos que se desarrollan son incluidos en un cuadro de mando denominado *Dashboard*.

Dado lo mencionado en los puntos 1, 2 y 3, es posible obtener una perspectiva general de los componentes necesarios para el desarrollo de una herramienta de inteligencia de negocios, tanto desde su base de datos hasta las visualizaciones consultadas por el usuario final; sin embargo, la capa correspondiente al diseño del *data warehouse* es la que motiva la comparación entre modelos de bases de datos relacionales y columnares de este estudio. Uno de los estudios previos que realiza una comparación llamado "*Using the column oriented NoSQL model for implementing big data warehouses*" [17], propone distintos métodos con el fin de mapear los elementos relacionales correspondientes a 3 distintas capas de desarrollo, en modelos columnares. Estas capas corresponden a los modelos conceptual, lógico y físico. Junto con esto, el estudio propone 3 formas distintas de realizar este mapeo de relaciones a columnas. En primera instancia un acercamiento lógico normalizado, en segunda instancia un modelo lógico desnormalizado y, finalmente, un modelo desnormalizado con familias de columnas.

Por otra parte, el estudio “*Columnar NoSQL Star Schema Benchmark*” [19], propone estrategias para la migración de datos desde un modelo estrella ya implementado hacia un modelo columnar. Este acercamiento es desnormalizado, considerando un elemento denominado *Big Fact Table*, el cual termina siendo una tabla de hecho la cual contiene en si los datos correspondientes a todas las dimensiones, atribuyendo a cada dimensión una familia de columna.

CAPÍTULO 3: DISEÑO DE LA SOLUCIÓN

En este capítulo se propone una serie de pasos para desarrollar una solución al problema planteado. Se describen los pasos para realizar una comparación de rendimiento entre dos modelos de bases de datos y técnicas para el desarrollo de bases de datos que soporten estructuras y plataformas analíticas, considerando en primera instancia, las necesidades del negocio y las principales áreas donde resulta necesario hacer un análisis, esto apoyado de técnicas de modelado de procesos de negocios, para posteriormente desarrollar modelos para arquitecturas columnares y relacionales, y finalizando con la integración de plataformas de visualización de datos que permite estudiar el rendimiento de cada arquitectura para el área de negocio modelada.

3.1. Modelos de negocio

El análisis de las distintas áreas del negocio y sus cadenas productivas tiene como objetivo la elección de las entidades que resultan relevantes o urgentes de analizar con respecto a las otras. Este análisis entrega una base en cuanto a cuáles son las variables asociadas a cada proceso y resulta útil como documentación de los procesos de la cadena de valor de la organización, y de los atributos que tendrán las distintas dimensiones de los modelos analíticos. La cadena productiva de la organización a analizar está compuesta, a grandes rasgos, de 3 procesos: Activación, Crédito y Cobranza, los cuales se encuentran ordenados secuencialmente de la forma en la cual se describen a continuación:

- 1) Activación: proceso mediante el cual una empresa o *pyme* solicita un monto a la organización, para que así esta publique la solicitud en una plataforma mediante la cual diversos inversionistas puedan realizar sus inversiones sobre la solicitud de la empresa solicitante. De esta forma, y una vez financiada completamente la solicitud, se transfieren los montos invertidos al solicitante.
- 2) Crédito: proceso donde el solicitante cuenta con los montos obtenidos desde los inversionistas, los cuales generan, de acuerdo con las condiciones estipuladas en los respectivos contratos, distintos montos de intereses asociados a las fechas estipuladas de pago de cuotas.
- 3) Cobranza: proceso donde se procede a recuperar los montos invertidos dado el no cumplimiento de las condiciones contractuales o incumplimiento de plazos de pago.

Dados estos tres procesos de la cadena de valor, los cuales se encuentran compuestos de variados subprocesos, tareas y flujos de mayor complejidad, resulta imperante para Cumplir la realización de una plataforma que permita el análisis sobre el primer proceso identificado, Activación. Este proceso fue modelado bajo la notación BPNM [23], la cual está especialmente orientada a la visualización y el correcto entendimiento y documentación de los procesos de negocio de las organizaciones. Dado el análisis realizado del proceso de activación, se desarrollaron tres diagramas asociados a los subprocesos que describen el proceso en cuestión, estos son:

- 1) Subproceso de solicitud de crédito: donde los principales actores involucrados son el solicitante de crédito, el representante comercial de dicho solicitante y el área que evalúa el riesgo asociado al solicitante de crédito. Las tareas de este subproceso hacen referencia a las etapas por la que pasa una solicitud de crédito con respecto al cumplimiento de los estándares definidos por Cumplimiento con respecto a las garantías y el riesgo asociado al solicitante, para así, terminar con la publicación de la solicitud en la plataforma pública de inversionistas.
- 2) Subproceso de inversión: considera las etapas mediante las cuales un inversionista externo formaliza una inversión en una solicitud de crédito publicada en la plataforma pública. Sus principales actores son el inversionista, la plataforma pública y el solicitante del crédito.

Entidad	Atributo
Solicitud de crédito	Identificador de solicitud
	Tipo de crédito
	Estado de solicitud
	Representante comercial
	Fecha de activación
Inversionista	Identificador de inversionista
	Nombre de inversionista
	Sexo
	Nivel educacional
	Profesión
Empresa solicitante	Identificador de solicitante
	Tipo de empresa
Respaldo	Identificador de respaldo
	Empresa pagadora
	Monto
Activación	Monto de activación
	Cobro por uso de plataforma
	Tasa interna de retorno
	Tasa de crédito

Tabla 4: Entidades y atributos clave propuestos por el análisis de los subprocesos de negocio.

- 3) Subproceso de activación: en este subproceso se verifica que la solicitud publicada cumpla con todos los criterios definidos por Cumplimiento y se garanticen todas las condiciones contractuales que permiten un flujo limpio de dinero entre inversionistas y solicitantes. Los principales actores de este proceso de negocio son los representantes comerciales, solicitantes de crédito y área de operaciones.

Dado el análisis realizado de cada uno de los subprocessos pertenecientes al proceso de activación, resulta posible encontrar las principales entidades que describen a este proceso, además de sus atributos. De esta forma resulta posible tener un mejor manejo de la selección de variables a analizar en una futura visualización. En la tabla 4 se muestran las principales entidades y atributos propuestos en esta etapa.

El detalle y los subprocessos de negocio detallados en forma de procesos modelados en notación BPMN, puede ser consultado en el anexo A.

3.2. Diseño de la visualización de datos

Posterior al modelado de los procesos de negocio identificados, resulta útil según la propuesta, desarrollar esquemas que representen los resultados finales requeridos por Cumplio. En este caso se diseñan múltiples interfaces gráficas que hacen uso de las variables identificadas en el diseño de los procesos de negocio, para luego ser incorporadas en un tablero de control que otorgue una perspectiva general del estado del proceso de Activación según las variables escogidas. Los diseños no se basan en una metodología estándar, puesto que el objetivo de esta memoria es realizar una comparación a nivel de rendimiento y no un estudio de las características gráficas de las interfaces. Sin embargo, al ser un producto real para una empresa, se trabaja con el equipo de diseñadores en la evaluación de las propuestas a implementar, dando un mayor énfasis a explotar la variedad de estilos de gráficos por sobre la utilidad práctica de estos, con el fin de presentar en una primera instancia a los actores interesados el abanico de propuestas visuales que la plataforma de visualización propone. Entre las gráficas que componen el tablero de control a proponer se encuentran las siguientes.

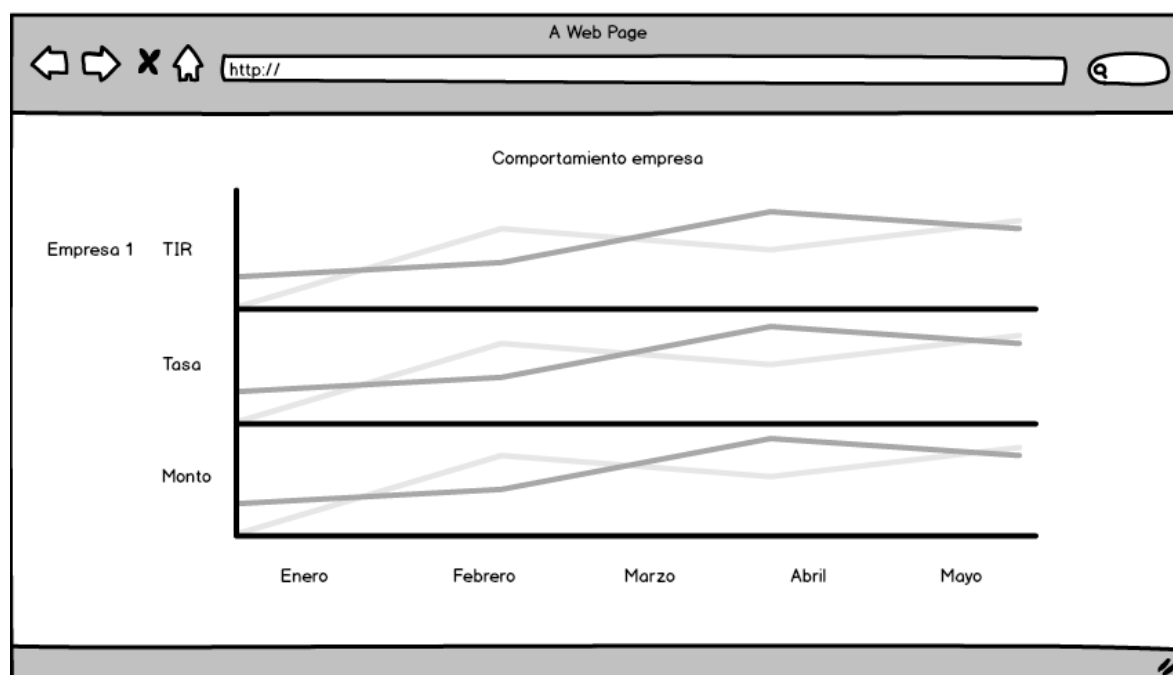


Figura 2: Diseño propuesto para visualización de comportamiento de empresas en el tiempo.

- 1) Comportamiento de empresas en el tiempo: corresponde a una visualización compuesta de un gráfico de líneas múltiples, el cual contiene las principales variables seleccionadas a visualizar para cada empresa, tales como tasa interna de retorno, tasa del crédito y monto total de los créditos solicitados; cada una mostrada a lo largo del periodo, desde el inicio de actividades en Cumplimiento con la opción de filtrar por cada empresa solicitante. El diseño alcanzado mediante las especificaciones previas es representado en la figura 2.
- 2) Comportamiento de representantes comerciales: corresponde a una visualización con amplias similitudes con respecto a la gráfica de descrita previamente, salvo que el área de dominio de esta se encuentra en los representantes comerciales asociados a cada empresa solicitante, de forma que al escoger una empresa sobre la cual analizar su evolución, también es posible analizar el comportamiento del representante comercial a lo largo del tiempo, desplegando información mensual con respecto al número de créditos que el solicitante ha activado, la tasa promedio de los créditos activados y el monto total de los créditos activados. Esta visualización es representada mediante el diseño presente en la figura 3.

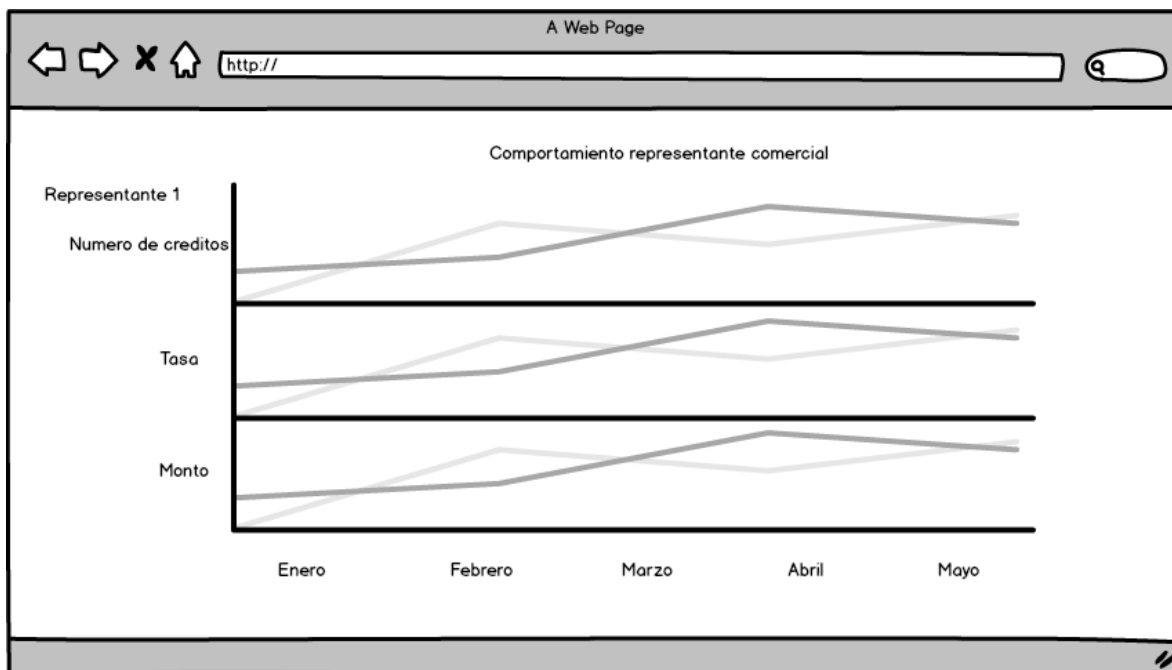


Figura 3: Diseño propuesto para visualización de comportamiento de representantes comerciales.

- 3) Montos activados por empresa: corresponde a una visualización en la cual se representa mediante el tamaño de diferentes burbujas, el monto activado por cada empresa solicitante asociada a un representante en particular, considerando que un representante comercial puede ser responsable de la gestión de múltiples empresas solicitantes. La visualización mencionada presenta un diseño presentado en la figura 4.

Montos por empresa del representante

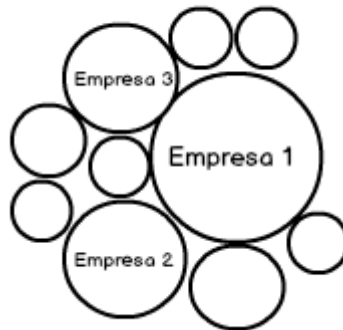


Figura 4: Diseño propuesto para visualización de montos activados por empresa del representante comercial.

- 4) Cantidad de créditos activados: corresponde a una visualización en la cual se representa mediante el tamaño de diferentes burbujas, la cantidad de créditos activados por cada empresa solicitante asociada a un representante en particular, considerando que un representante comercial puede ser responsable de la gestión de múltiples empresas solicitantes. El diseño de la visualización se presenta en la figura 5.

Número de créditos por empresa del representante

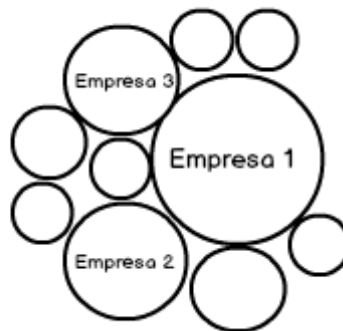


Figura 5: Diseño propuesto para visualización de número de créditos activados por empresas del representante comercial.

- 5) Inversiones según profesión: corresponde a una visualización que incorpora las variables asociadas a la inversión de cada crédito. En este gráfico es posible observar la agrupación de inversionistas para cada crédito activado según su profesión, haciendo posible el análisis en base a la situación profesional de los inversionistas y de sus preferencias de acuerdo con cada solicitante. El diseño de la visualización se presenta en la figura 6.

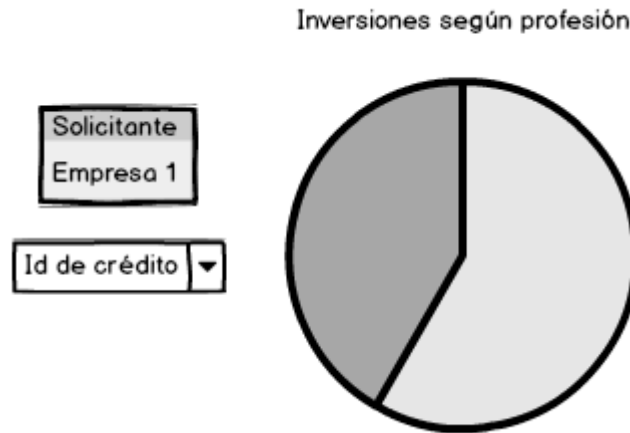


Figura 6: Diseño propuesto para visualización de inversiones según profesión del inversionista.

- 6) Pagadores según monto: esta visualización incorpora la variable relacionada con los respaldos de cada pagador asociado al crédito activado. Es así como se representan en un mapa de árbol los distintos montos respaldados por las distintas empresas, considerando que un crédito activado puede estar respaldado por múltiples empresas, las cuales proporcionan montos diferentes de la totalidad del monto activado. El diseño de la visualización se presenta en la figura 7.



Figura 7: Diseño propuesto para visualización de montos de facturas por créditos de empresa solicitante.

Finalmente, el tablero de control involucra la combinación de todos los diseños y visualizaciones presentados; es así como se propone tener una noción general del proceso de Activación que involucre las principales variables obtenidas desde el modelado del proceso de negocios. Posteriormente, en base a lo desarrollado se propone realizar un

modelo lógico y trasladar los diseños definidos a niveles de modelos de base de datos y arquitectura, para finalmente realizar pruebas de rendimiento sobre las opciones propuestas.

3.3. Diseño del Modelo Estrella

Tras el trabajo anterior, resulta necesario diseñar el modelo relacional sobre el cual se construye el esquema estrella; para esto se identifican las dimensiones y hechos a incorporar a cada una de las tablas, resultando el modelo de la figura 8.

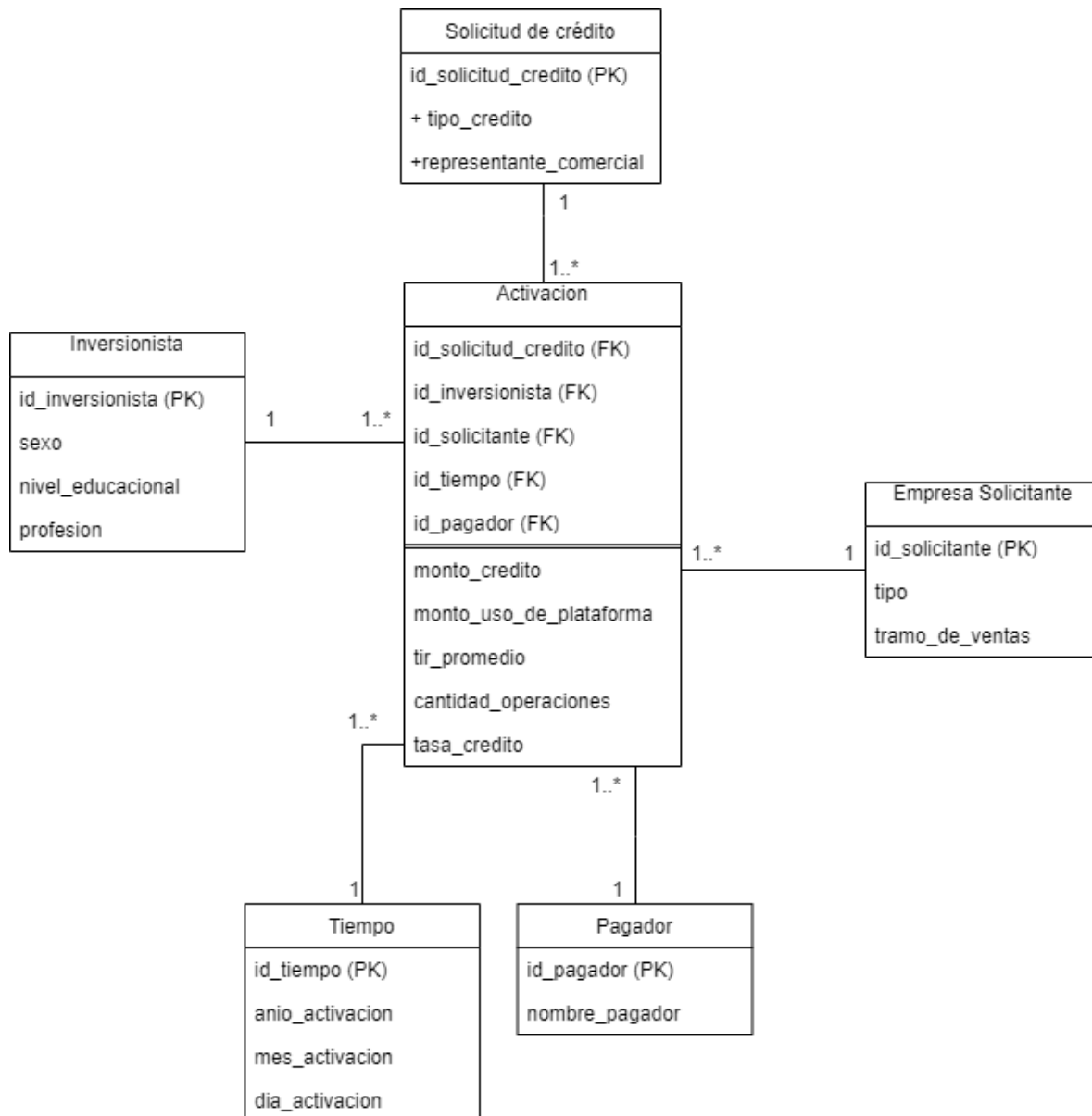


Figura 8: Modelo de datos estrella para el proceso de Activación.

La realización de este modelo permite el análisis del proceso de Activación, considerando variables importantes asociadas a las inversiones, pagadores y representantes

comerciales, además de incorporar información relevante al tiempo generando un análisis histórico del comportamiento de representantes comerciales con respecto a las empresas solicitantes.

Con la finalidad de poder poblar la base de datos correspondiente al *data warehouse*, basado en el modelo relacional anterior, se propone el uso de la herramienta *Pentaho Data Integration* la cual cuenta con funciones o pasos preestablecidos que son un aporte relevante al desarrollo de este tipo de arquitecturas analíticas, ya que proporciona una interfaz de alto nivel que evita transitar por el diseño de consultas mediante código, usando íconos con estilo de uso *Drag and Drop*. Así, el diseño del esquema ETL usado para poblar el *data warehouse* se muestra en la figura 9, donde la primera tarea consiste en la extracción de datos, realizando una consulta sobre estos datos por cada una de las dimensiones que se definirán. Las consultas correspondientes a cada dimensión del modelo contienen los campos que se definieron en el modelo de datos, para finalmente poblar la tabla de hechos. Ejecutada esta transformación, se tiene en una base de datos relacional el modelo estrella definido previamente.



Figura 9: Diseño del *job* para el modelo estrella propuesto, desarrollado en plataforma *Pentaho Data Integration*.

3.4. Diseño del modelo columnar

Como estrategia para el diseño del modelo de datos columnar se escogió la técnica definida como “Acercamiento Lógico Desnormalizado” (“DLA” por las siglas en inglés para “*Denormalized Logical Approach*”), que fue presentada el año 2015[17], la cual consiste en construir una tabla asociada al modelo desnormalizado, es decir, una tabla denominada “*Big Fact Table*”, la cual contiene todas las dimensiones y los hechos correspondientes. De esta forma la tabla “*BigFactTable*” estaría conformada como muestra la tabla 5, donde un registro tendría todos los campos que componen la tabla. Resulta importante mencionar que cada uno de los campos presentados en la tabla tiene relación con una columna y los datos referentes a esta pueden ser extraídos de forma independiente con respecto al resto de las columnas, sin la necesidad de extraer el registro completo en caso de necesitar realizar agregaciones grandes que involucren gran cantidad de datos.

Atributo	Tipo
id_inversionista	INT
sexo	VARCHAR
nivel_educacional	VARCHAR
profesion	VARCHAR
id_solicitud_de_credito	INT
tipo_credito	VARCHAR
representante_comercial	VARCHAR
id_solicitante	INT
tipo_solicitante	VARCHAR
tramo_de_ventas	DECIMAL
id_pagador	INT
nombre_pagador	VARCHAR
monto_credito	INT
uso_de_plataforma	DECIMAL
tir_promedio	DECIMAL
cantidad_de_operaciones	INT
tasa_credito	DECIMAL

Tabla 5: Atributos del modelo columnar seleccionados para la creación de las columnas en base de datos *Cassandra*.

3.5. Vistas materializadas

Dada la desnormalización del modelo presentado, para su utilización en forma columnar es posible notar que se generan considerablemente mas registros que con las tablas normalizadas; dado esto, los beneficios esperados en tiempo de ejecución de las consultas pueden verse opacados. En consecuencia, se propone realizar vistas materializadas para el modelo columnar, las cuales usan la información proporcionada por los diseños realizados y extraen para cada visualización las columnas que solo son requeridas por cada una de ellas. Por otra parte, de acuerdo a las visualizaciones desarrolladas también resulta en un beneficio realizar vistas materializadas para el modelo relacional, puesto que con esta práctica se pretende acelerar su rendimiento y realizar una comparación parcial entre ambos modelos:

3.6. Vistas materializadas para el modelo relacional

Se propone el desarrollo de tres vistas materializadas, cada una diferenciada según el contexto de cada subprocesso del proceso de Activación:

- 1) Vista materializada de empresas y representantes comerciales: contiene la información respectiva a las empresas en conjunto con sus representantes comerciales, la cual es capaz de proporcionar la información necesaria por las visualizaciones de la figura 6, 7, 8 y 9, generando la tabla 6.

Atributo	Tipo
id_funding	INT(11)
admin_name	VARCHAR(255)
credit_type_name	VARCHAR(255)
requester_business_name	VARCHAR(255)
amount	INT(11)
fee	INT(11)
tir	DOUBLE
nominal_rate	DECIMAL(30, 20)
activated_date	DATETIME

Tabla 6: Atributos seleccionados para la vista materializada correspondiente a la empresa solicitante y los representantes comerciales a crear en *MySQL*.

La creación en la base de datos asociada a la tabla 6 fue realizada mediante el código disponible en el anexo B.1.

- 2) Vista materializada de inversionistas: contiene la información necesaria para presentar en la visualización de la figura 10, la cual contiene información respectiva a los inversionistas, generando la tabla 7.

Atributo	Tipo
id_funding	INT(11)
invested_amount	INT(11)
requester_business_name	VARCHAR(255)
investor_id	INT(11)
investor_sex	VARCHAR(255)
profession_name	VARCHAR(255)

Tabla 7: Atributos seleccionados para la vista materializada correspondiente a inversionistas, a crear en *MySQL*.

La creación en la base de datos asociada a la tabla 7 fue realizada mediante el código disponible en el anexo B.2.

- 3) Vista materializada de pagadores: contiene la información necesaria para presentar en la visualización de la figura 11, la cual contiene información respectiva a los pagadores, generando la tabla 8.

Atributo	Tipo
id_funding	INT(11)
invoice_amount	INT(11)
requester_business_name	VARCHAR(255)
invoice_payer_id	INT(11)
invoice_payer_name	VARCHAR(255)

Tabla 8: Atributos seleccionados para la vista materializada correspondiente a pagadores, a crear en *MySQL*.

La creación en la base de datos de la tabla presentada en la tabla 8 fue realizada mediante el código disponible en el anexo B.3.

3.7. Vistas materializadas para el modelo columnar

El desarrollo de vistas materializadas en el modelo columnar se realiza mediante la plataforma *Pentaho Data Integration*, debido a que presenta un amplio soporte con respecto a arquitecturas columnares, contando con controladores especializados y una comunidad de desarrolladores activa. Para este proceso se definen tres vistas materializadas asociadas a cada uno de los subprocesos comprendidos en el proceso de Activación; junto con esto, se realizarán las agrupaciones pertinentes para así disminuir la cantidad de registros resultante de la desnormalización del modelo lógico relacional y contar con una vista materializada en forma columnar. Las vistas propuestas son:

- 1) Vista materializada de empresas y representantes comerciales: para este desarrollo se propone efectuar una selección de los valores que se considerarán en la visualización, para luego realizar una agregación por el identificador del crédito activado, y finalmente insertar los datos agrupados en las respectivas columnas del modelo columnar. Cabe destacar que las columnas en la base de datos pueden ser definidas mediante el software de forma automática mediante los controladores incorporados. El diseño del *job* se grafica de la forma expuesta en la figura 10.



Figura 10: Diseño del *job* para la creación de la vista materializada de empresas solicitantes y representantes comerciales en *Cassandra DB*.

En el paso de agrupación se define un criterio; dado que en este caso muchos de los registros están duplicados, debido a que una solicitud activada tiene múltiples inversiones y que las inversiones no son consideradas en esta vista, se decide agrupar por el identificador de la solicitud, almacenando como monto invertido el total de las inversiones realizadas.

Realizado este trabajo se tiene almacenada en el espacio definido como “*materialized_views*”, una vista con los datos necesarios para formar la visualización mencionada en Cassandra.

- 2) Vista materializada de inversionistas y pagadores: para el desarrollo de la vista materializada de inversionistas resulta necesario realizar el mismo trabajo de extracción que en el subproceso de empresas y representantes comerciales, con

la diferencia que en la etapa de selección de datos y agrupación se eligen parámetros diferentes, los cuales corresponden a los mencionados en la tabla 7 para inversionistas y la tabla 8 para pagadores. De esta forma se tiene una vista con los datos que serán requeridos por las interfaces gráficas correspondientes, sin disponer de datos extra que no sean visualizados.

Con el desarrollo de las vistas materializadas para el modelo columnar y el modelo estrella almacenados ambos en memoria secundaria, se tienen como bases de datos propuestas para el análisis a dos estructuras con arquitecturas basadas en conceptos distintos. Dado esto, el paso siguiente tiene relación con la integración de estas arquitecturas a una plataforma de análisis que permita la realización de pruebas de ejecución de consultas con datos y visualizaciones reales pertenecientes a Cumpro.

3.8. Arquitectura del sistema a construir

El sistema a construir, con el fin de soportar las distintas herramientas necesarias para realizar la comparación de rendimientos entre base datos, contempló distintos niveles de memoria para la ejecución de sus pruebas. Por una parte, la primera propuesta contempló el uso de memoria secundaria con bases de datos MySQL para el modelo estrella, y por otra parte para la propuesta columnar Cassandra. Además, se desarrollaron escenarios de prueba con bases de datos en memoria primaria con una extracción de datos en archivos Hyper, correspondientes a los *snapshots* de la base de datos, tanto para el modelo columnar como el relacional.

El diseño de las distintas arquitecturas para las bases de datos se ejemplifica según la figura 11, donde se muestran los mismos tipos de bases de datos en diferentes tipos de memoria.

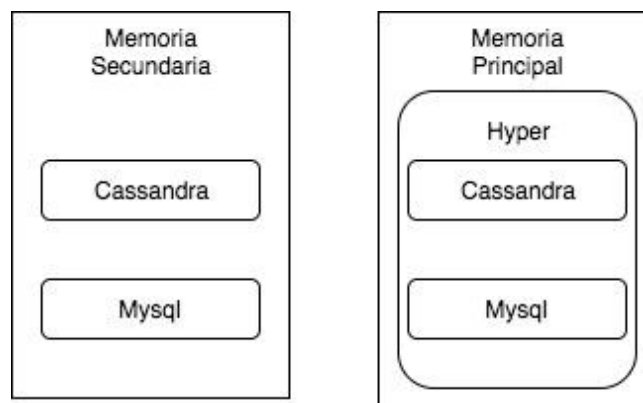


Figura 11: Diagrama de división de arquitectura de pruebas de rendimiento.

Una vez realizada la integración, la misma herramienta ofrece como opción realizar una extracción de los datos, para así mejorar considerablemente el rendimiento de las consultas realizadas para el renderizado de cada visualización; de esta forma, se extraen los datos desde el disco a la memoria primaria, creando una instancia de base de datos en memoria primaria para cada base de datos. La plataforma crea un archivo con extensión *“.hyper”*, el cual conforma la base de datos en memoria primaria, almacenando de esta

forma los distintos estados de tipos de base de datos y las transacciones en *logs* que son ejecutados para pasar de un estado de la base de datos a otro. Los beneficios de esta forma de almacenamiento radican principalmente en la velocidad de lectura y escritura, que es considerablemente mayor en memoria primaria que en disco. Además, cabe destacar que este tipo de almacenamiento cuenta con sofisticados algoritmos de compresión con los cuales tablas como la de hechos del modelo presentado, que en su formato relacional tiene un peso de 74MB, mientras que, en un archivo de extracción de datos de Hyper en Tableau pesa tan solo 6MB con los datos usados en este trabajo.

De la forma propuesta se realizan extracciones de datos para cada una de las bases de datos, teniendo de esta forma cuatro bases de datos, agrupadas en arquitecturas en memoria primaria y memoria secundaria, cada una con una base de datos de tipo columnar y otra relacional, de forma que se puedan comparar rendimientos entre cada uno de los elementos internos de cada categoría.

3.9. Integración de bases de datos con Tableau Desktop

Con ambas bases de datos ya pobladas resulta posible la integración de estos modelos con diversas plataformas de inteligencia de negocios. Estas plataformas ofrecen una diversidad de servicios que, usualmente, convergen en la apropiada visualización de los datos modelados.

La evaluación de una plataforma de visualización de datos, que permita una presentación apropiada de la información, puede contener diversos factores a considerar, desde el presupuesto disponible para la contratación de estos servicios hasta la disponibilidad de opciones de visualización que se adapten a las necesidades de la organización. En este caso, el factor más importante a evaluar fue la compatibilidad; la decisión con respecto a la plataforma en la cual visualizar los datos se orientó a la que mejor soporte la arquitectura columnar desarrollada. Es así como evaluando distintas herramientas se llega a la opción de usar Tableau Desktop, puesto que esta herramienta cuenta con los controladores necesarios para realizar conexiones directas a nodos pertenecientes a la arquitectura columnar.

La conexión con las bases de datos relacionales se realiza de forma directa, puesto que el soporte para este tipo de bases de datos es más natural en comparación con otros tipos de arquitectura, por lo que se utiliza un *driver* común de bases de datos, MySQL en este caso. Por otra parte, para el caso columnar, es necesario obtener *drivers* especializados que proporcionan servicios externos y compatibilizan la plataforma con el modelo; en este caso se utiliza el *driver* para conexiones de bases de datos Cassandra de CData.

3.10. Pruebas de rendimiento

Con las cuatro bases de datos pobladas e integradas con la plataforma de visualización seleccionada, la propuesta de solución se encuentra en el escenario en donde resulta necesario estudiar y discernir sobre qué opción de diseño resulta mejor implementar de acuerdo con las necesidades de la empresa. Cada una de estas propuestas se validó en su mismo ambiente con su par, midiendo el tiempo que cada base de datos tomó en ejecutar

las consultas solicitadas por las visualizaciones, con el fin de comparar los tiempos de ejecución obtenidos por un modelo versus los del otro.

Dado que la forma de validar las propuestas fue mediante el tiempo de ejecución de las distintas consultas que se generan según las necesidades de la empresa, por lo que se generaron escenarios de prueba asociados a cada visualización y se ejecutaron las consultas que serán realizadas por los usuarios finales de la plataforma. De esta forma, mediante la comparación de tiempos de ejecución, se concluirá sobre las propuestas presentadas.

Para los tiempos de ejecución sobre cada tipo de base de datos, se generan tres escenarios de prueba, los cuales corresponden a consultas reales requeridas por los usuarios finales:

- a) Escenario 1: se solicita la información asociada a los montos transados por un representante comercial, para luego consultar en detalle los montos asociados a empresas particulares de dicho representante. En este escenario se evalúa la estrategia de *Drill-Down* desde representantes comerciales a empresas solicitantes.
- b) Escenario 2: se consulta sobre los montos transados por los distintos representantes a lo largo del tiempo, proponiendo la opción de variar las agrupaciones con respecto a los meses y años de activación de los créditos correspondientes. En este escenario se evalúa el comportamiento de los distintos modelos con respecto a una agrupación de fechas.
- c) Escenario 3: se solicita información de diversas dimensiones para ser expuesta en la plataforma de visualización. En este escenario se busca medir el rendimiento de cada modelo con respecto a una consulta en la cual se desarrollan operaciones *JOIN* sobre distintas dimensiones.

CAPÍTULO 4: IMPLEMENTACION Y VALIDACIÓN DE LA SOLUCIÓN

En el este capítulo se desarrolla una estructura bajo la cual se llevan a cabo las pruebas comparativas entre cada uno de los modelos propuestos previamente. Dados los resultados de las pruebas ejecutadas, se realiza un análisis del rendimiento y el contexto de cada arquitectura, y cómo se asocia cada uno de estos resultados con las necesidades actuales y proyecciones futuras del área de inteligencia de negocios de Cumplio. Finalmente, y posterior a este análisis, se realiza una selección propuesta de la arquitectura o tecnología a utilizar para la implementación de futuros *data marts* con dirección a la construcción de un *data warehouse* para Cumplio.

4.1. Evidencias

De acuerdo con el diseño planteado como solución, se poblaron los modelos de base de datos columnar y estrella, junto con sus respectivas vistas materializadas, obteniendo en cada uno de estos, fuentes de datos disponibles para ser consumidas por la plataforma de visualización. Con lo que se obtuvo las siguientes vistas a utilizar por el usuario final.

- a) Comportamiento de empresa: corresponde a la visualización que permite analizar los indicadores principales de las empresas solicitantes de acuerdo al tiempo de operación en Cumplio, y puede ser vista en la figura 12.

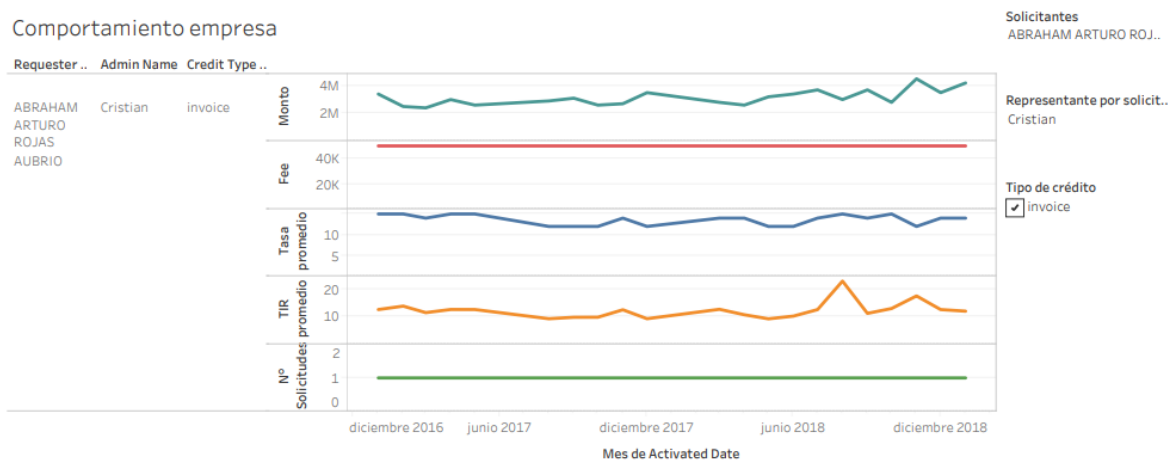


Figura 12: Visualización desarrollada para el comportamiento de empresas solicitantes según el diseño propuesto para el modelo columnar y relacional.

- b) Comportamiento de representante: corresponde a la visualización que permite analizar los indicadores principales de los representantes comerciales que llevan activando créditos en la organización, y puede ser vista en la figura 13.

Indices representante

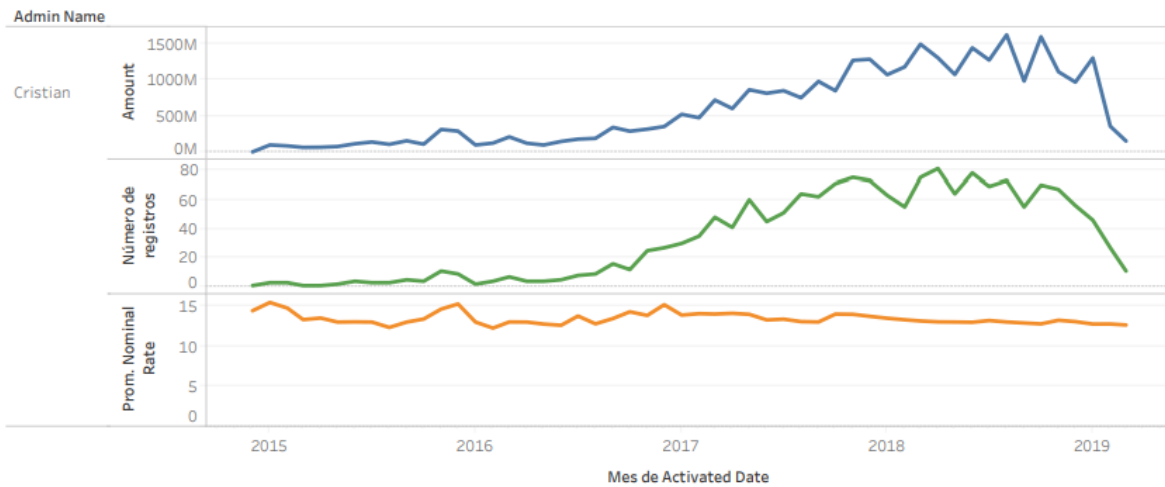


Figura 13: Visualización desarrollada para el comportamiento de representantes comerciales según el diseño propuesto para el modelo columnar y relacional.

- c) Montos por empresa y número de Créditos: corresponde a dos gráficos de burbujas que grafican qué tan grandes son los montos solicitados por algunas empresas en relación a otras que operan con el mismo representante comercial. De la misma forma, en otra visualización se da a conocer la cantidad de créditos activados por empresa de un representante comercial en particular. Estas visualizaciones pueden ser vistas en la figura 14.

Montos activados por empresa por representante



Número de creditos activados



Figura 14: Visualización desarrollada para montos activados y cantidades de créditos de empresas según el diseño propuesto para el modelo columnar y relacional.

- d) Inversiones y Pagadores: corresponde a una visualización en la cual se aprecian las distintas inversiones asociadas a un crédito activado, proporcionando en un gráfico de torta las cantidades invertidas. Por otra parte, en un mapa de árbol se subdividen los distintos montos respaldados por los pagadores asociados al crédito. Los gráficos resultantes pueden ser vistos en la figura 15.

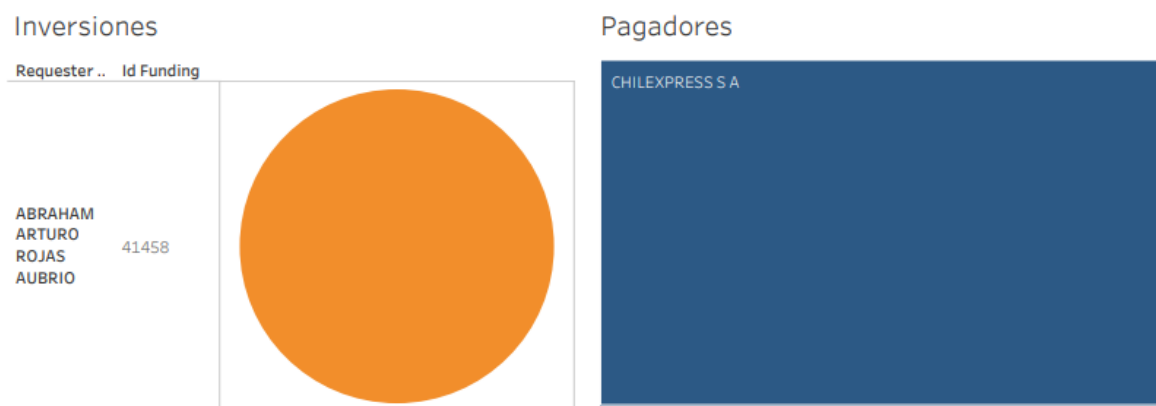


Figura 15: Dashboard desarrollado para la profesión de inversionistas y los montos de las facturas según el diseño propuesto para el modelo columnar y relacional.

4.2. Estructura de las pruebas de rendimiento

Para realizar pruebas de rendimiento se seleccionaron 3 escenarios diferentes, los cuales representan 3 tipos de consultas que el usuario final realiza al momento de interactuar con la plataforma. Estos escenarios fueron escogidos de acuerdo con el diseño del tablero de control y del análisis de los requisitos del negocio sobre la plataforma; además, los resultados fueron obtenidos mediante la herramienta de registro de rendimiento de Tableau Desktop, la cual respalda y entrega una visualización detallada del tiempo de ejecución de la renderización de las visualizaciones como del tiempo de ejecución de las consultas a las bases de datos, que son parte de la información que resulta interesante considerar.

a) Escenario 1: *DrillDown*

En este escenario se generó una visualización para generar las pruebas respectivas, la cual consiste en el despliegue de indicadores obtenidos a partir de la tabla de hechos, siendo filtrados por representante comercial en una primera instancia, para luego acceder a datos más detallados, considerando de esta forma una empresa solicitante en particular del representante seleccionado y entregando los mismos indicadores clave, como se aprecia en la figura 16.

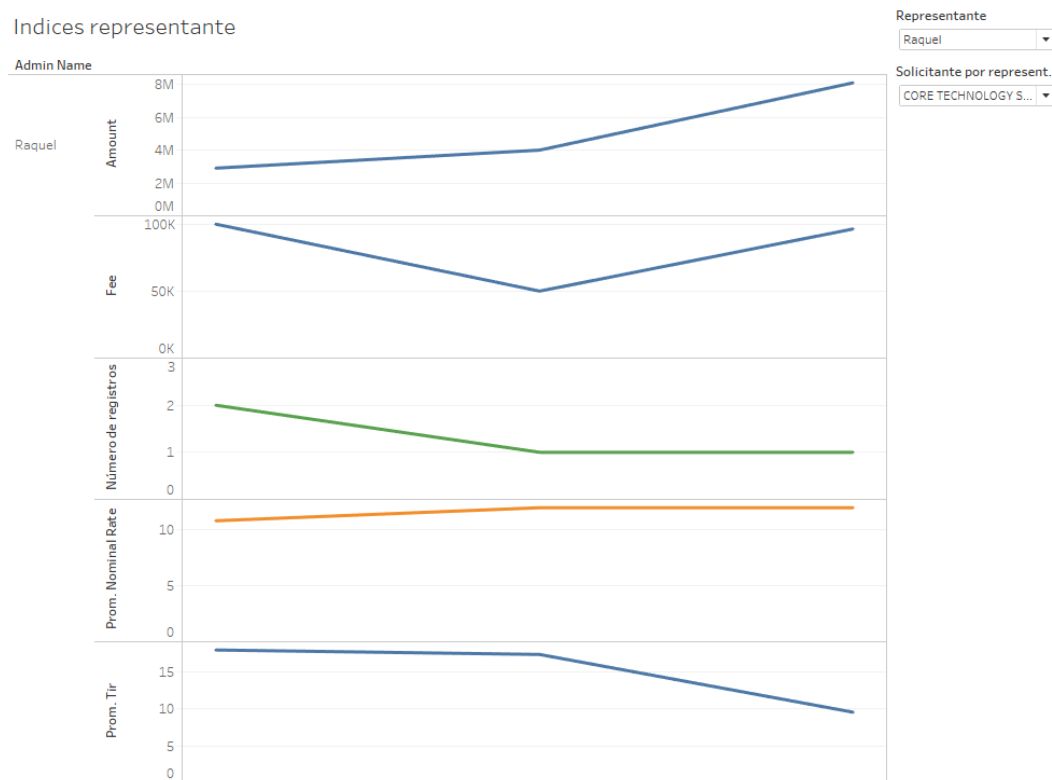


Figura 16: Visualización creada para la prueba de rendimiento para ambos ambientes y modelos de datos según el primer escenario.

b) Escenario 2: Agrupación por fechas

En este escenario se generó una visualización en la cual se accede a los datos considerados como indicadores clave según representante comercial, otorgando la posibilidad de acceder información cada vez mas detallada con respecto a las fechas de activación, dando la posibilidad de seleccionar, en una primera instancia, un año de activación para posteriormente visualizar información de un mes en particular, como se puede ver en la figura 17.

c) Escenario 3: Intersección de dimensiones

En este escenario se desarrolló una visualización en la cual resulta necesario la combinación de 3 dimensiones distintas para la generación de los gráficos. Se muestran dos gráficos de torta, correspondientes a la distribución porcentual según sexo de los inversionistas en los créditos activados de una empresa seleccionada, además de un segundo gráfico de torta, el cual presenta la distribución porcentual de la profesión de los inversionistas. Junto con estos gráficos provenientes de la dimensión de las inversiones, se muestra un mapa de árbol con los montos proporcionados por los pagadores de las facturas de los créditos activados, donde la

clave sobre la cual se realizan las intersecciones es la empresa solicitante, como se puede ver en la figura 18.

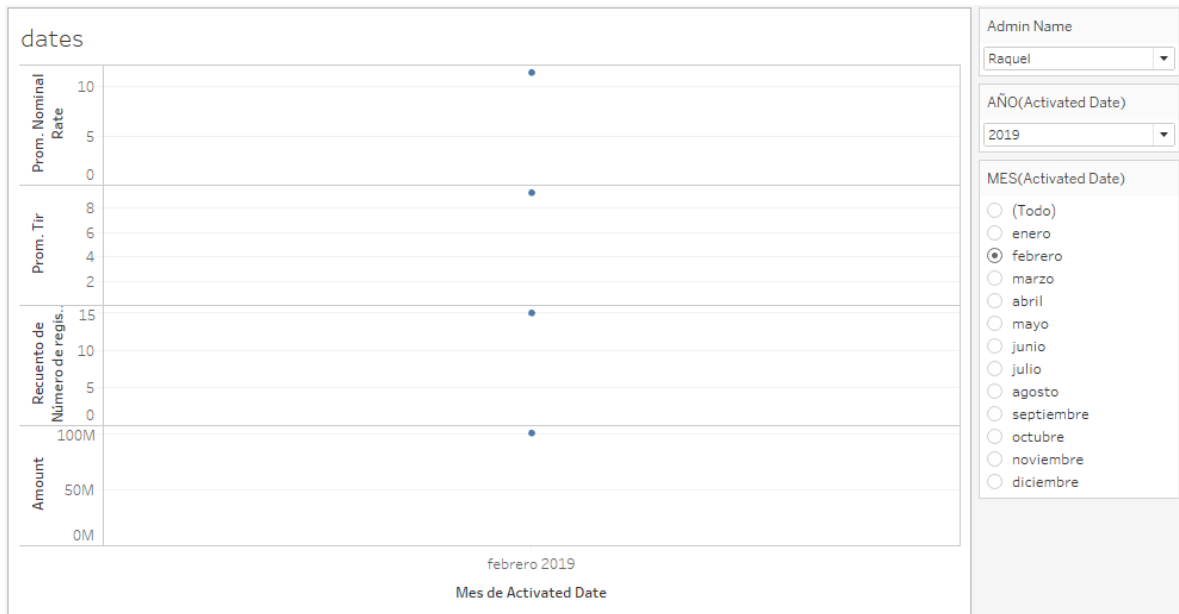


Figura 17: Visualización creada para la prueba de rendimiento para ambos ambientes y modelos de datos según el segundo escenario.

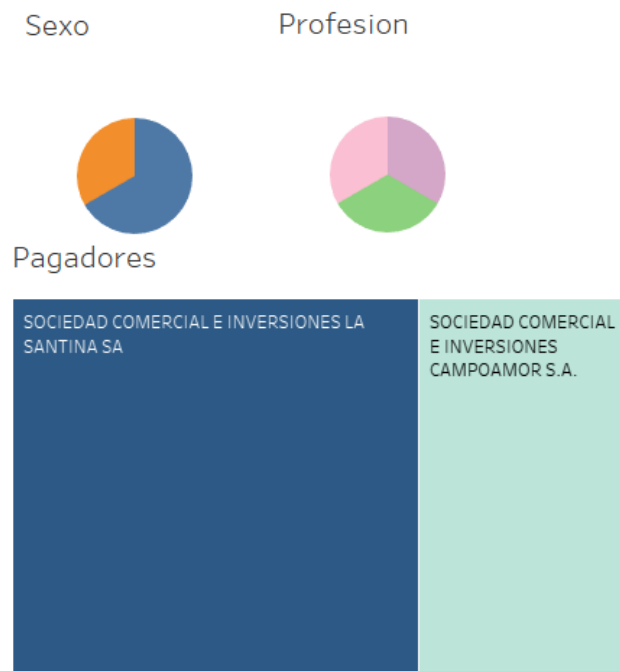


Figura 18: Dashboard creado para la prueba de rendimiento para ambos ambientes y modelos de datos según el tercer escenario.

Con cada uno de estos escenarios se desarrollaron pruebas de rendimiento, que consistieron en ejecutar las consultas respectivas, comparando en cada escenario el modelo columnar con el modelo de bases de datos relacional tanto en memoria primaria como en memoria secundaria. Además, cabe destacar que las pruebas realizadas en memoria secundaria se desarrollaron bajo un disco duro en estado sólido.

4.3. Resultados de las pruebas de rendimiento

Para las pruebas de rendimiento, en una primera instancia se probaron los 3 escenarios, comparando el modelo columnar con el relacional en memoria secundaria. Para el primer escenario se seleccionó un conjunto de representantes donde para cada uno se escogió una empresa solicitante asociada a él, de forma que se ejecutara la consulta respectiva a cada base de datos desplegando la información resultante en la visualización. Se registraron los datos para 12 representantes y para 12 empresas solicitantes; promediando los valores del tiempo de ejecución de las consultas se obtuvieron los resultados del gráfico 1.

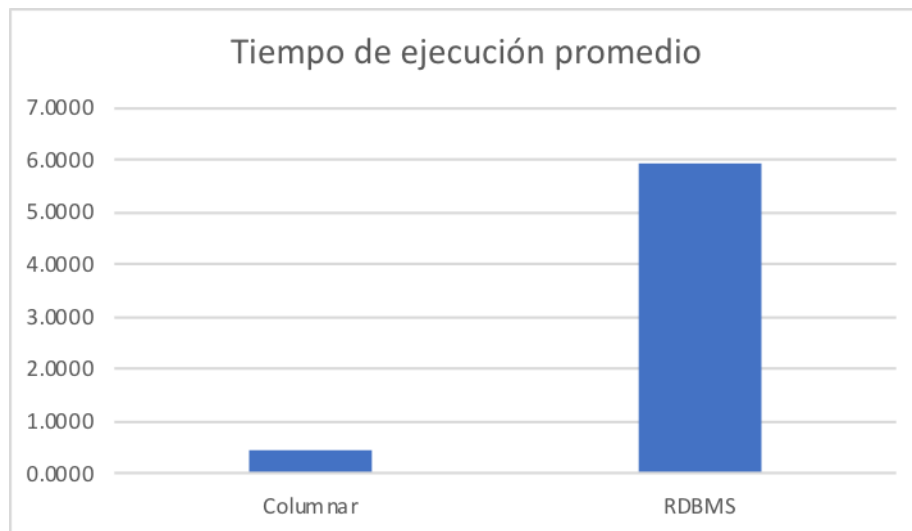


Gráfico 1: Tiempo promedio de ejecución de consultas para el primer escenario en memoria secundaria.

En el gráfico 1 es posible notar como el modelo columnar emplea una cantidad considerablemente menor de tiempo de ejecución al realizar *DrillDown* desde un representante a su empresa representada en memoria secundaria.

Para el segundo escenario de pruebas en memoria secundaria se seleccionó un grupo de 12 representantes comerciales, para los cuales se filtró la información requerida por la visualización para un año en particular, para luego filtrar por un mes en particular. Se obtuvieron los promedios del tiempo de ejecución de las consultas para cada modelo de datos, obteniendo los resultados del gráfico 2.

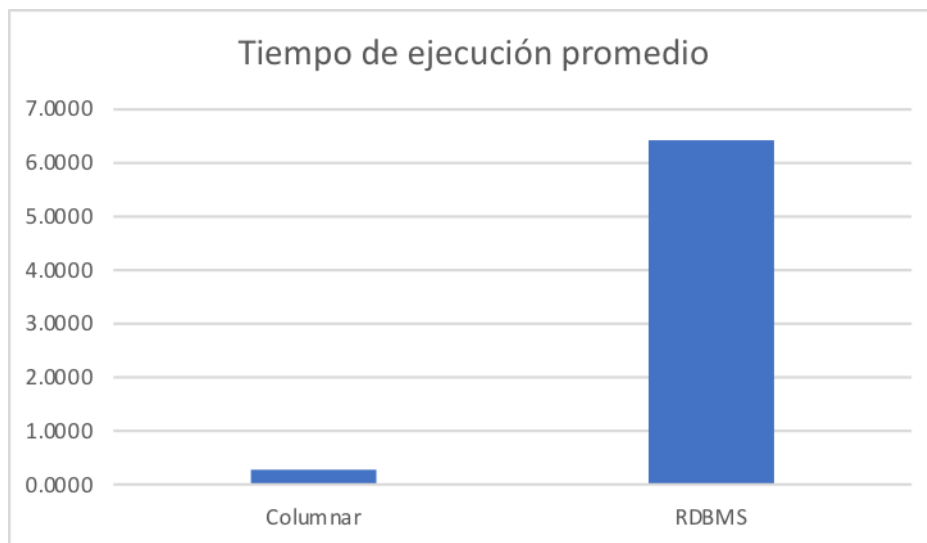


Gráfico 2: Tiempo promedio de ejecución de consultas para el segundo escenario en memoria secundaria.

En el gráfico 2 se nota como el modelo columnar emplea una menor cantidad de tiempo en la obtención de valores clave y su filtro por fechas en memoria secundaria.

Para el tercer escenario de pruebas en memoria secundaria se seleccionó un grupo de 14 empresas solicitantes, para las cuales se solicitó, mediante consultas a cada modelo, la información respectiva al sexo de los inversionistas, su profesión y los montos asociados a los pagadores de facturas de los créditos solicitados por la empresa, conforme fueron necesarios para el despliegue de la visualización. Se obtuvieron los promedios de cada modelo, obteniendo los resultados del gráfico 3.

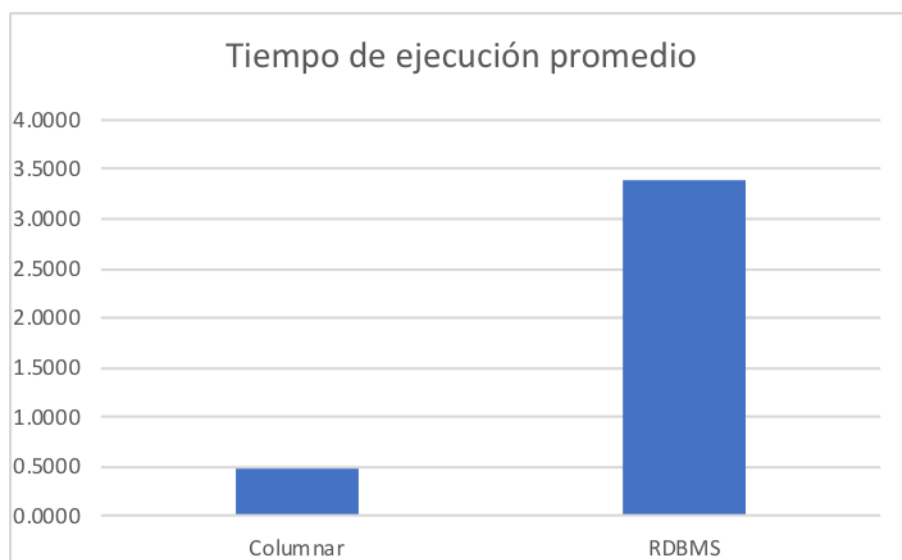


Gráfico 3: Tiempo promedio de ejecución de consultas para el tercer escenario en memoria secundaria.

En el gráfico 3 es posible notar como el modelo columnar emplea menor tiempo de ejecución al momento de consultar multiples dimensiones en memoria secundaria.

Posterior a las pruebas de rendimiento en memoria secundaria, se procedió a realizar pruebas de rendimiento en memoria primaria. Los escenarios bajo los cuales se realizaron las pruebas fueron los mismos que en memoria secundaria, y la variable a registrar fue el tiempo de ejecución de las consultas. Para el primer escenario de prueba se registro el tiempo de ejecución de las consultas ejecutadas, seleccionando 12 representantes comerciales y 12 empresas sobre las cuales se desplegó la infomación requerida por la visualización, obteniendo los resultados del gráfico 4.

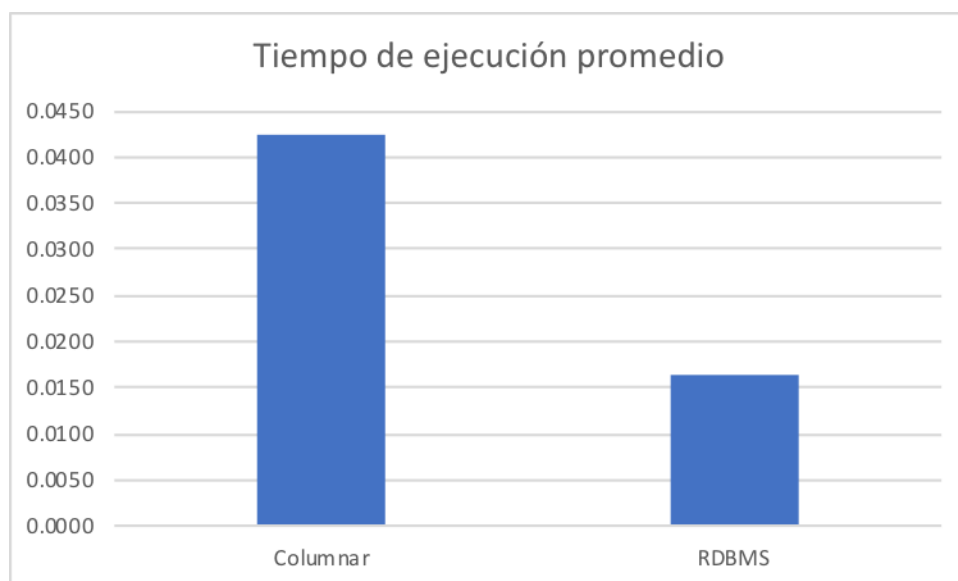


Gráfico 4: Tiempo promedio de ejecución de consultas para el primer escenario en memoria primaria.

En el gráfico 4 es posible notar como el modelo columnar empleó mayor tiempo de ejecución que el modelo relacional al momento de hacer consultas del tipo *DrillDown* en memoria primaria.

Para el segundo escenario de prueba en memoria primaria, se seleccionó 12 representantes comerciales, para los cuales se filtró la información para un año y un mes en específico, resgistrando el tiempo de resupuesta de cada modelo, de forma que se obtuvieron los resultados del gráfico 5.

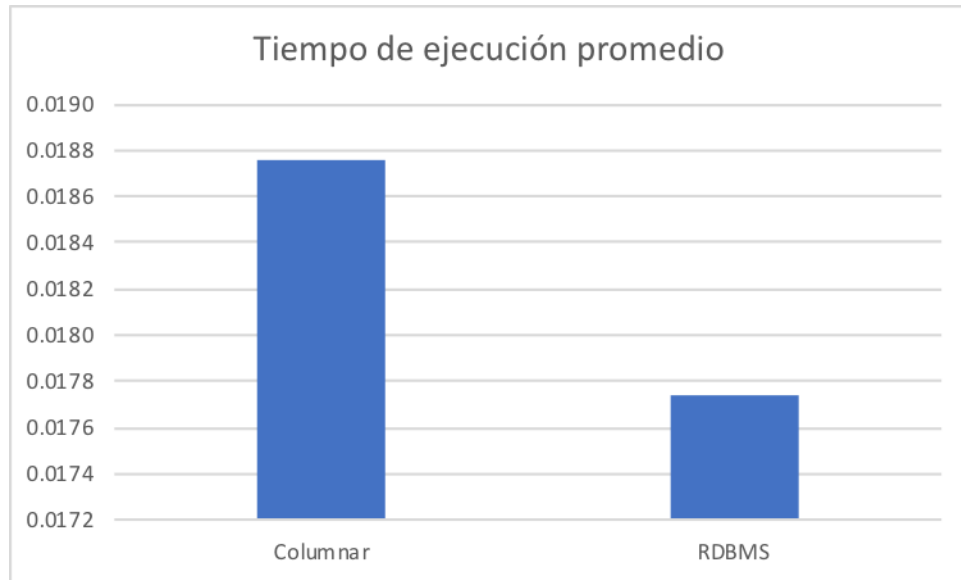


Gráfico 5: Tiempo promedio de ejecución de consultas para el segundo escenario en memoria primaria.

En el gráfico 5 se aprecia como el modelo columnar empleó un mayor tiempo de ejecución que el modelo relacional al realizar filtros por fecha en memoria primaria.

Finalmente, para el tercer escenario de pruebas se seleccionaron 14 empresas solicitantes, de las cuales se desplegó la información correspondiente al sexo y profesión del total de los inversionistas de sus créditos activados, como también la información respectiva al total de los montos de las facturas de los pagadores asociados a estos créditos. Se registró el tiempo de ejecución de las consultas, promediando los resultados obtenidos para cada modelos, obteniendo los resultados del gráfico 6.

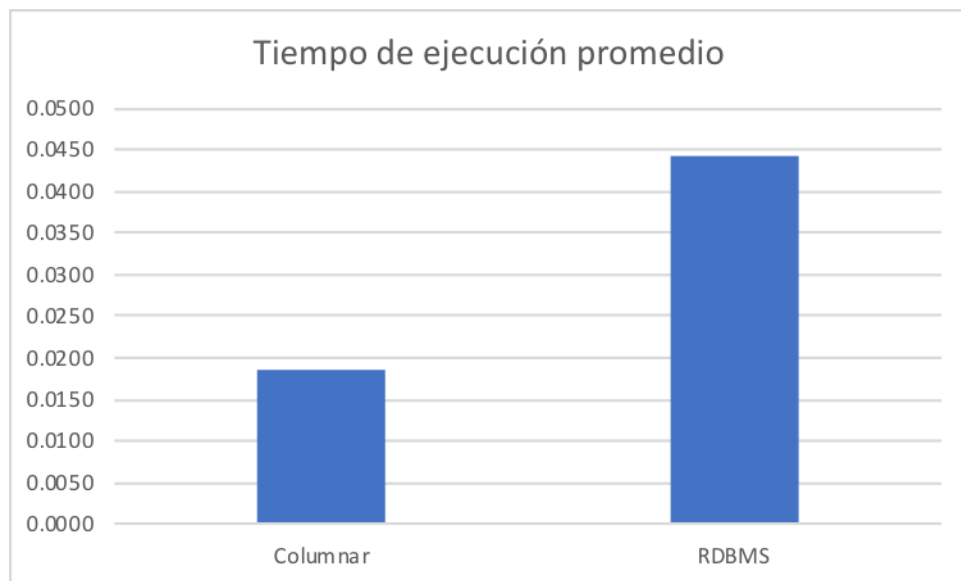


Gráfico 6: Tiempo promedio de ejecución de consultas para el tercer escenario en memoria primaria.

En el gráfico 6 es posible notar como el modelo relacional tomó un mayor tiempo de ejecución que el modelo columnar al realizar consultas que involucraran múltiples dimensiones.

4.4. Análisis de pruebas de rendimiento

En esta sección se realiza un análisis detallado de los gráficos que reflejan los tiempos de ejecución de cada escenario de prueba.

Con respecto a las pruebas realizadas en memoria secundaria, fue posible observar que para los 3 escenarios de prueba los resultados fueron similares; además, estos resultados concuerdan con la información presentada por la literatura actual [17], la que plantea que las arquitecturas columnares tienen mejores desempeños al momento de realizar consultas que involucren agregaciones de gran porcentaje de datos de las tablas respectivas. Para los tres escenarios el tiempo de ejecución de las consultas fue inferior a un segundo, mientras que para el modelo relacional se mantienen entre los 3 y 6 segundos. Gran parte del buen desempeño que presenta el modelo columnar sobre el relacional en este tipo de agregaciones radica en el desarrollo de las vistas materializadas, dado que estas estructuras de datos no contienen dimensiones, sino que usan solamente los datos necesarios para la visualización permitiendo que no se produzcan operaciones *JOIN* para poder entregar la información consultada.

En cuanto a las arquitecturas en memoria primaria, los tiempos de ejecución de las consultas son significativamente menores que los tiempos registrados por las pruebas en memoria secundaria. Las pruebas en memoria primaria siguen presentando grandes diferencias con respecto a aquellas en memoria secundaria. Los tiempos de ejecución de las consultas sobre las bases de datos construidas en memoria primaria son inferiores a las décimas de segundos; es más, las consultas que poseen menor complejidad presentan resultados en el orden de las centésimas. Sin embargo, en este sentido se aprecia una diferencia con respecto a las pruebas realizadas en memoria secundaria y es que en los escenarios que involucran solicitar más detalles en la granularidad de los datos, los modelos estrella presentan mejor desempeño que los modelos columnares, no así en el escenario 3, que involucra mayor número de operaciones *JOIN*. Sin embargo, la diferencia en los tiempos de ejecución entre cada modelo, en el ambiente de memoria primaria, resultan despreciables ya que son del orden de las centésimas de segundo.

De acuerdo con los resultados obtenidos, es posible notar una gran diferencia en los tiempos de ejecución de consultas entre los modelos propuestos en memoria secundaria versus los modelos propuestos en memoria primaria. Por esta razón es que se decidió no comparar ambas categorías entre sí, sino que comparar los modelos agrupados dentro de cada una de esas categorías. Estas diferencias quieren decir que la propuesta de realizar el diseño de un *data mart* sobre una base de datos en memoria primaria sea siempre la adecuada o la más conveniente, sino que esta decisión tiene que ver con el contexto en el cual el *data mart* es utilizado

4.5. Selección de propuesta de solución

Para seleccionar una propuesta de solución a implementar, resulta necesario analizar los resultados obtenidos por todas las propuestas, en base a los requerimientos de Cumpló, basados en las visualizaciones que se necesiten desplegar en el tablero de control. De acuerdo con la naturaleza de la organización y la información que hoy en día se analiza en ella, no se tienen grandes volúmenes de datos bajo los cuales se propongan soluciones del tipo *Big Data*. Por ejemplo, hoy en día se trabaja con información relacionada a la operación diaria de la empresa, donde la mayor granularidad en cuanto a la información trabajada es del nivel de las cuotas pertenecientes a un crédito, además, la cantidad de créditos activados diariamente es cercana a los 50, lo cual es un problema diferente a, por ejemplo, analizar el comportamiento mediante los *clicks* de los usuarios en la plataforma pública, nutriendo esta información con datos externos a ser analizados. Dado esto, el problema actual que se requiere solucionar no significa un desafío importante para la capacidad de cómputo de los sistemas actuales que se pueden implementar al momento de requerir analizar esta información, lo que descarta de cierta forma la opción de diseñar una estructura que aloje bases de datos en memoria primaria. Sin embargo, pese a que no se aprovechó a cabalidad el potencial que ofrecen este tipo de técnicas, no se descarta que este tipo de técnicas pueda ser útil en un futuro cercano; además, en plataformas como Tableau Desktop, realizar este tipo de extracciones a memoria primaria contempla gran cantidad de soporte por parte de la plataforma, puesto que no significa más complejidad que una sencilla configuración, la cual consta de la selección de una opción dentro de la visualización, para hacer uso de esta funcionalidad.

Por otra parte, al momento de comparar modelos columnares versus modelos relacionales en el diseño de un *data mart* sobre ambientes en memoria secundaria, se puede apreciar diferencias sustanciales en cuanto a los tiempos de ejecución. Bajo este estudio, se ratifica mediante los resultados expuestos, que las agregaciones y consultas necesarias para proporcionar información a las visualizaciones diseñadas y resolver el problema del análisis de los datos de Cumpló, contemplan un mejor tiempo de ejecución en arquitecturas columnares sobre las arquitecturas relacionales, además estas arquitecturas permiten modularizar los aspectos a analizar. Si bien este tipo de arquitecturas supone un desafío mayor al momento de realizar los diseños, puesto que se debe tener claro previamente con mayor precisión cuales serán las variables y tablas, junto con el modelo a consultas para visualizar, este estudio también propone un marco de trabajo sobre el cual se definen a cabalidad todas las variables necesarias, previo al diseño de los modelos columnares.

Bajo las premisas planteadas, la propuesta a mantener y sobre la cual se propone trabajar, es la de modelos columnares con vistas materializadas, por sus ventajas en cuanto al tiempo de respuesta, dejando disponible la alternativa de extraer los datos a bases de datos en memoria primaria.

CONCLUSIONES

De acuerdo con el trabajo realizado, tanto en la parte investigativa como en las partes técnicas involucradas en la implementación de las soluciones propuestas, fue posible percibir la importancia que tiene una disciplina perteneciente a la tecnología, como lo es la inteligencia de negocios. Esta disciplina atribuye ventajas competitivas importantes a las organizaciones que implementan sus herramientas, puesto que proporciona una visión completa en distintos niveles de el estado del negocio, dando la posibilidad de proyectarse hacia el futuro con las decisiones organizacionales, colaborando así con una gestión de mayor calidad, y analizar el crecimiento de la organización o su comportamiento pasado.

Fue posible, además, percibir como la implementación de soluciones y herramientas del área de la inteligencia de negocios requiere de arquitecturas especiales con respecto a su diseño, y con el fin de proporcionar un funcionamiento adecuado y congruente con las velocidades y estándares de calidad que demandan las organizaciones modernas. Es así como se desarrollan áreas completas de especialización, tanto en arquitectura como en herramientas de visualización, que hacen necesario que el análisis de las organizaciones se desprenda de su ambiente operacional, resultando necesario ambientes propios y especializados de análisis de datos.

Una empresa como Cumpló Chile S.A., que se define desde sus orígenes como una empresa tecnológica, es una empresa en constante y acelerado crecimiento, la cual demanda soluciones tecnológicas diariamente, de tal forma que su crecimiento como organización es mas acelerado que su crecimiento tecnológico. Dado esto, fue posible observar que muchos de los requisitos tecnológicos que la empresa ha demandado a lo largo del tiempo, se encuentran netamente ligados al contexto operacional de la organización, el cual se encarga de organizar el funcionamiento diario de la empresa, proporcionando herramientas desarrolladas a la medida de sus procesos de negocio, sin tener aún procesos de negocio definidos de manera formal. En este sentido, se desprenden dos variables de gran valor que fueron proporcionadas por este trabajo. Por una parte, se concluye que se logró documentar y definir de manera detallada gran parte de los procesos de negocio involucrados en la cadena de valor de la organización, en una notación estándar y reconocida en el mundo de los procesos; y por otra, se logró identificar la necesidad de implementar una plataforma de análisis de datos que entregue un estado general (en una primera instancia) de la información respectiva a los procesos de la organización, involucrando en sí, parte importante de la cadena de valor respectiva al proceso de Activación, proporcionando como valor agregado el diseño de una plataforma analítica construida sobre arquitecturas con tecnologías modernas escogidas desde procesos comparativos con tecnologías clásicas, con el fin de fundamentar su uso.

Es posible concluir que Cumpló Chile S.A. se encuentra en un proceso de expansión tal, que se hace necesario una capa específica y de calidad para el análisis de datos de la organización, además, las tecnologías propuestas en los últimos años, que son alternativas

que nacen desde la minería de datos o del *big data*, ofrecen opciones interesantes en términos de desempeño de las consultas analíticas. Por esto, se concluye que la organización califica como un caso de estudio interesante para una comparativa entre dos de las técnicas más conocidas en cuanto al análisis de información: las bases de datos relacionales versus las bases de datos de tipo columnar, para el desarrollo de *data marts*.

Se concluye, además, que de acuerdo a la naturaleza de la organización, la cual presenta procesos de negocios sumamente modulares, con áreas de negocio establecidas de manera precisa y limitada, la mejor forma de desarrollar un *data warehouse* para la organización, es mediante técnicas *bottom-up*, implementando en primera instancia *data marts* que se encarguen de guardar la información necesaria por áreas atómicas específicas del negocio, en vez del diseño de un *data warehouse top-down*, donde gran parte del *data warehouse* se desarrolla en una primera instancia, opción que pone en riesgo la flexibilidad y escalabilidad de este.

Se concluye que, en una arquitectura basada en memoria secundaria, los modelos desarrollados de tipo columnar presentan un mejor comportamiento con respecto al tiempo de ejecución de las consultas en comparación con el modelo estrella, mientras que, en memoria primaria, las consultas que involucran menor cantidad de operaciones *JOIN*, presentan un mejor resultado en un modelo relacional, no así cuando la cantidad de estas operaciones es mayor. Sin embargo, los tiempos en memoria primaria son reducidos de forma drástica en comparación con los tiempos registrados en memoria secundaria, tanto así que las diferencias entre los modelos columnares y estrella en esta arquitectura son imperceptibles.

De acuerdo con los resultados obtenidos en la comparación de cada uno de los modelos, se concluye que la opción a mantener implementada y disponible para trabajos futuros es la de un modelo columnar en memoria secundaria, proporcionando así la opción de realizar extracciones a memoria primaria mediante la herramienta de visualización propuesta, en caso de que se necesita inmediatez en cuanto al despliegue de la información en las visualizaciones desarrolladas.

Se concluye que las bases de datos desarrolladas y las herramientas de visualización y análisis de datos implementadas son de gran valor para los procesos de gestión de los procesos de negocio de la organización, puesto que otorgan una visión general del proceso de Activación, visión que previamente se obtenía mediante la consulta directa de datos en las bases de datos operacionales, sin la opción de visualizar y resultando en procesos engorrosos y demandantes.

Además, se concluye que este trabajo forma parte de un marco de trabajo propuesto para el diseño e implementación de futuras visualizaciones, con el fin de conformar una primera versión de un estándar de desarrollo de *data marts* y visualización de datos para Cumplio Chile S.A.

Finalmente, se propone como trabajo futuro, el diseño e implementación de nuevos *data marts* y visualizaciones, que involucren distintas áreas del negocio interesantes a analizar.

Algunas de estas áreas de negocio involucran datos con respecto al riesgo asociado a los solicitantes como también puede involucrar información interesante para los cargos directivos de la organización. Además, la arquitectura desarrollada y las plataformas involucradas permiten la integración con distintos tipos de fuentes de datos, proporcionando así la posibilidad de integrar mas modelos y arquitecturas además de bases de datos columnares o relacionales.

Para la realización de este estudio fue de suma importancia las bases que se fundaron sobre los cursos de Bases de Datos, además de los cursos de profundización como Bases de Datos Avanzadas y el contenido obtenido de Inteligencia de Negocios. Además, parte del trabajo realizado involucró información adquirida desde la asignatura de Modelado de Procesos de Negocios. Como herramientas que hubiesen sido deseables al momento de desarrollar el estudio, vale la pena mencionar Bases de datos en memoria primaria.

REFERENCIAS BIBLIOGRÁFICAS

1. El Mostrador Mercados. (2015, 12 noviembre). Cumplo.cl, el polémico emprendimiento financiero que desafió al sistema y ganó. Recuperado 2 octubre, 2019, de <https://m.elmostrador.cl/mercados/2015/11/12/cumplo-cl-el-polemico-emprendimiento-financiero-que-desafio-al-sistema-y-gano/>
2. Colaboradores de Wikipedia. (2019, 26 septiembre). mecanismo colaborativo de financiación de proyectos. Recuperado 2 octubre, 2019, de <https://es.wikipedia.org/wiki/Micromecenazgo>
3. López, C. (2017, 9 febrero). ¿Qué tipos de crowdfunding existen? Recuperado 2 octubre, 2019, de <https://www.rankia.com/blog/crowdfunding-crowdlending/3468321-que-tipos-crowdfunding-existen>
4. Colaboradores de Wikipedia. (2020, 11 febrero). Inteligencia empresarial - Wikipedia, la enciclopedia libre. Recuperado 20 febrero, 2020, de https://es.wikipedia.org/wiki/Inteligencia_empresa
5. Wikipedia contributors. (2019, 13 diciembre). system used for reporting and data analysis. Recuperado 19 diciembre, 2019, de https://en.wikipedia.org/wiki/Data_warehouse
6. Colaboradores de Wikipedia. (s.f.). Relational database - Wikipedia, la enciclopedia libre. Recuperado 19 diciembre, 2019, de https://es.wikipedia.org/wiki/Relational_database
7. Colaboradores de Wikipedia. (2019, 15 diciembre). Esquema en estrella - Wikipedia, la enciclopedia libre. Recuperado 19 diciembre, 2019, de https://es.wikipedia.org/wiki/Esquema_en_estrella
8. ¿Qué es una Base de Datos NoSQL? (2016, 19 abril). Recuperado 19 diciembre, 2019, de <https://blogs.oracle.com/spain/qu-es-una-base-de-datos-nosql>
9. Amazon. (s.f.). AWS | Servicio de base de datos gestionada NoSQL (DynamoDB). Recuperado 2 octubre, 2019, de <https://aws.amazon.com/es/dynamodb/>
10. Mongo. (s.f.). The most popular database for modern apps. Recuperado 2 octubre, 2019, de <https://www.mongodb.com/>
11. Apache Cassandra. (s.f.). Apache Cassandra. Recuperado 2 octubre, 2019, de <http://cassandra.apache.org/>

12. Neo4j. (2019, 1 octubre). Neo4j Graph Platform – The Leader in Graph Databases. Recuperado 2 octubre, 2019, de <https://neo4j.com/>
13. Wikipedia contributors. (2019b, 25 noviembre). Column-oriented DBMS - Wikipedia. Recuperado 19 diciembre, 2019, de https://en.wikipedia.org/wiki/Column-oriented_DBMS
14. DB-Engines - Knowledge Base of Relational and NoSQL Database Management Systems. (s.f.). Recuperado 19 diciembre, 2019, de <https://db-engines.com/en/>
15. Nayak, A., Poriya, A., & Poojary, D. (2013, 1 marzo). Type of NOSQL Databases and its Comparison with Relational Databases. Recuperado 2 octubre, 2019, de https://www.researchgate.net/profile/Dikshay_Poojary/publication/302557703_Article_Type_of_nosql_databases_and_its_comparison_with_relational_databases/links/5aeaa2b50f7e9b837d3c40e7/Article-Type-of-nosql-databases-and-its-comparison-with-relational-databases.pdf
16. Solid IT gmbh. (s.f.). DB-Engines Ranking. Recuperado 2 octubre, 2019, de <https://db-engines.com/en/ranking/wide+column+store>
17. Dehdouh, K., Bentayeb, F., Boussaid, O., & Kabachi, N. (2015). Using the column-oriented NoSQL model for implementing big data warehouses. Recuperado 2 octubre, 2019, de <https://pdfs.semanticscholar.org/d0ad/8a604492d9aea8a737e081f2404e4b39391c.pdf>
18. Hyper. (s.f.). Hyper BD Distributions. Recuperado 20 febrero, 2020, de <https://hyper-bd.com/>
19. Dehdouh K., Boussaid O., Bentayeb F. (2014). Columnar NoSQL star schema benchmark. Model and Data Engineering, LNCS 8748, Springer, 2014.
20. Rosado, A., & Rico, D. (2010, abril). BUSINESS INTELLIGENCE: STATE OF THE ART. Recuperado 10 diciembre, 2019, de https://www.researchgate.net/publication/277231717_Inteligencia_de_negocios_Estado_del_arte
21. Fernández González, J. (2008, 21 enero). Los 4 Valores del Agile BI Governance.. Recuperado 10 diciembre, 2019, de <http://sistemasdecisionales.blogspot.com/2008/01/los-4-valores-del-agile-bi-governance.html>

22. Antiñanco, M. (2013, diciembre). Bases de Datos NoSQL: Escalabilidad y alta disponibilidad a través de patrones de diseño. Recuperado 10 diciembre, 2019.
23. Hitpass, B. (2017, 20 septiembre). BPM Fundamentos y Conceptos de Implementación. Recuperado 29 febrero, 2020.

ANEXO A: PROCESOS DE NEGOCIO

Los procesos de negocios son elementos de una organización compuestos por una secuencia de actividades que, relacionadas entre sí, producen un servicio o un producto. Estos pueden ser diagramados mediante una notación especial denominada BPMN (Business process management notation). Esta notación está compuesta de una serie de elementos que permiten describir el proceso de negocio de principio a fin. Algunos de los elementos más importantes de esta notación son:

- 1) Eventos: un evento es representado en la notación BPMN como un círculo, estos representan situaciones que ocurren en el proceso de negocio y no actividades que son realizadas, estos pueden ser eventos de inicio, termino o bien eventos intermedios y son representados según la figura A.1.



Figura A.1: Representación gráfica de ejemplos de eventos en notación de modelado de procesos de negocio BPMN.

- 2) Actividades: las actividades son representadas mediante un rectángulo y representan las tareas que serán realizadas. Son representados según la figura A.2.



Figura A.2: Representación gráfica de ejemplos de actividades en notación de modelado de procesos de negocio BPMN.

- 3) Compuertas: son representadas mediante un rombo y se encargan de bifurcar o unir el flujo de acuerdo con la lógica que se desee en la definición del negocio. Se representan según la figura A.3.



Figura A.3: Representación gráfica de ejemplos de actividades en notación de modelado de procesos de negocio BPMN.

- 4) Conexiones: las conexiones son representadas por líneas y pueden ser de distintos tipos, su función principal es atribuir de dirección y secuencialidad a las distintas tareas del proceso de negocio y son representadas según la figura A.4.

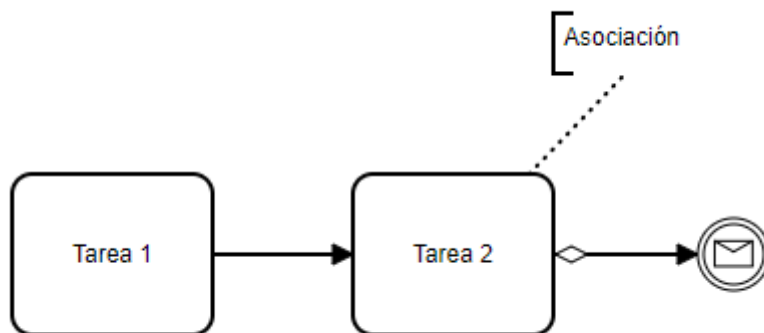


Figura A.4: Representación gráfica de ejemplos de conexiones en notación de modelado de procesos de negocio BPMN.

A continuación, se presenta el modelo del proceso de Solicitud de crédito en notación BPMN.

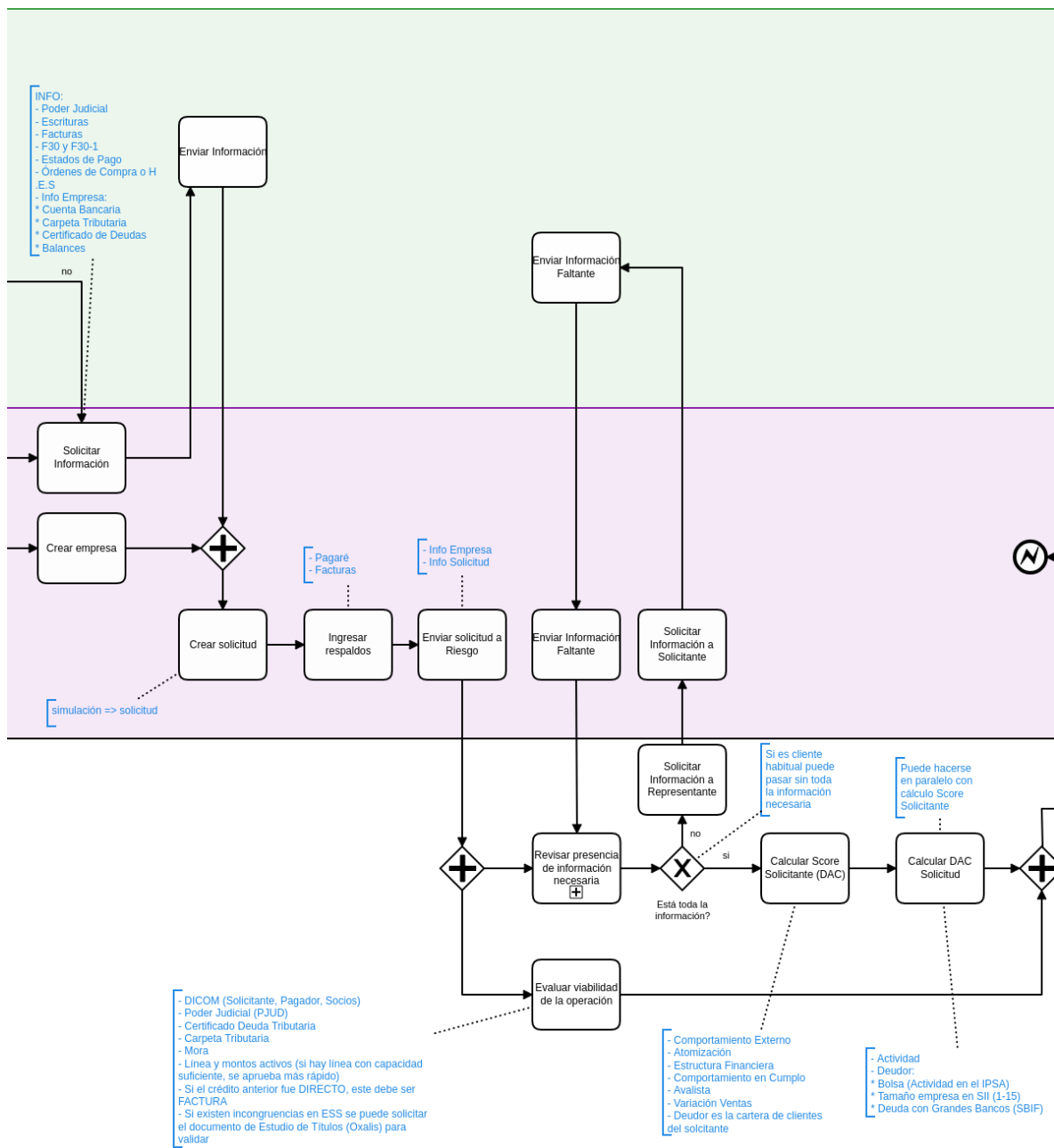


Figura A.5: Proceso de negocio detallado de solicitud de crédito, en notación BPMN, parte 2.

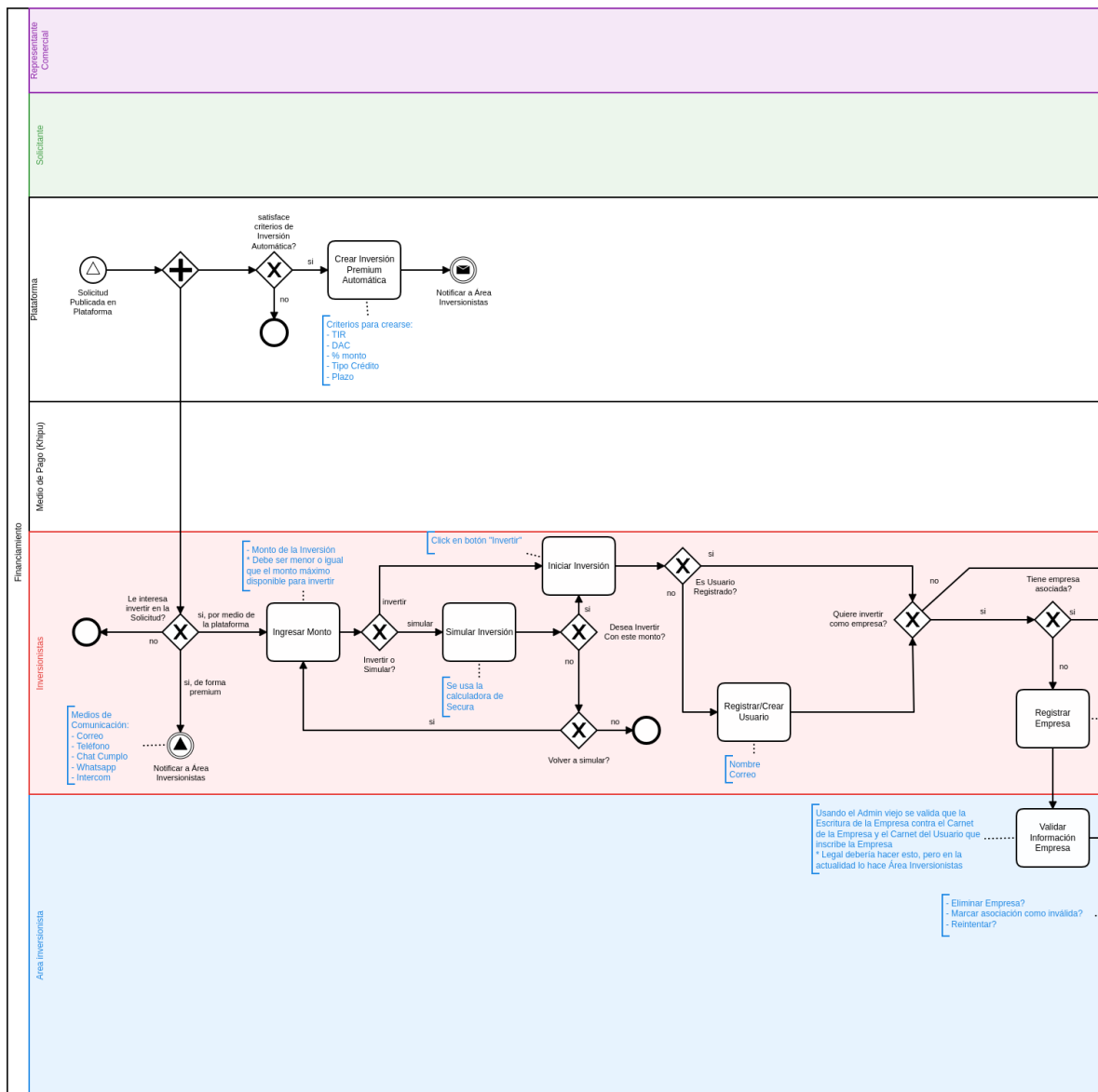


Figura A.6: Proceso de negocio detallado de inversión, en notación BPMN, parte 1.

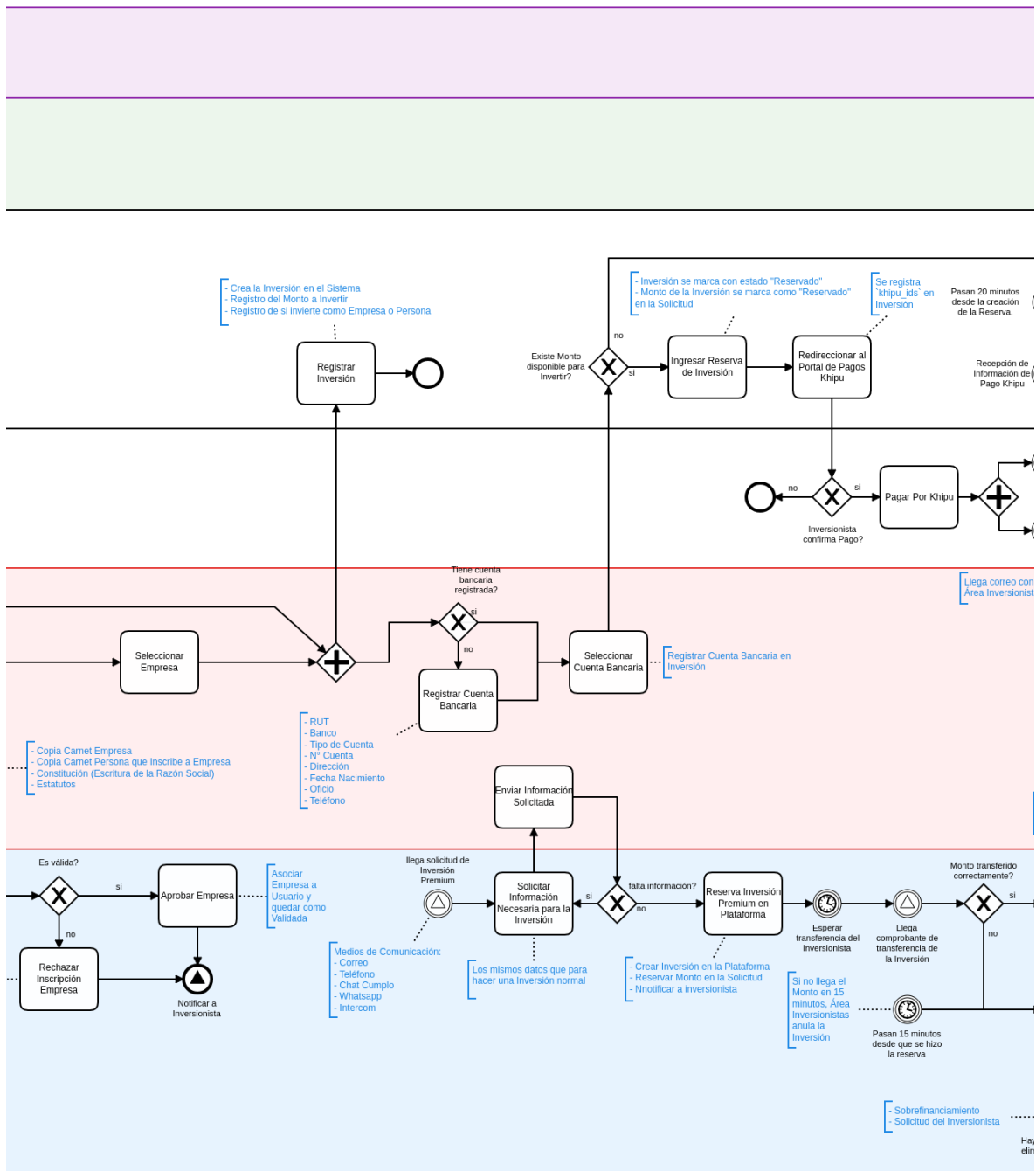


Figura A.6: Proceso de negocio detallado de inversión, en notación BPMN, parte 2.

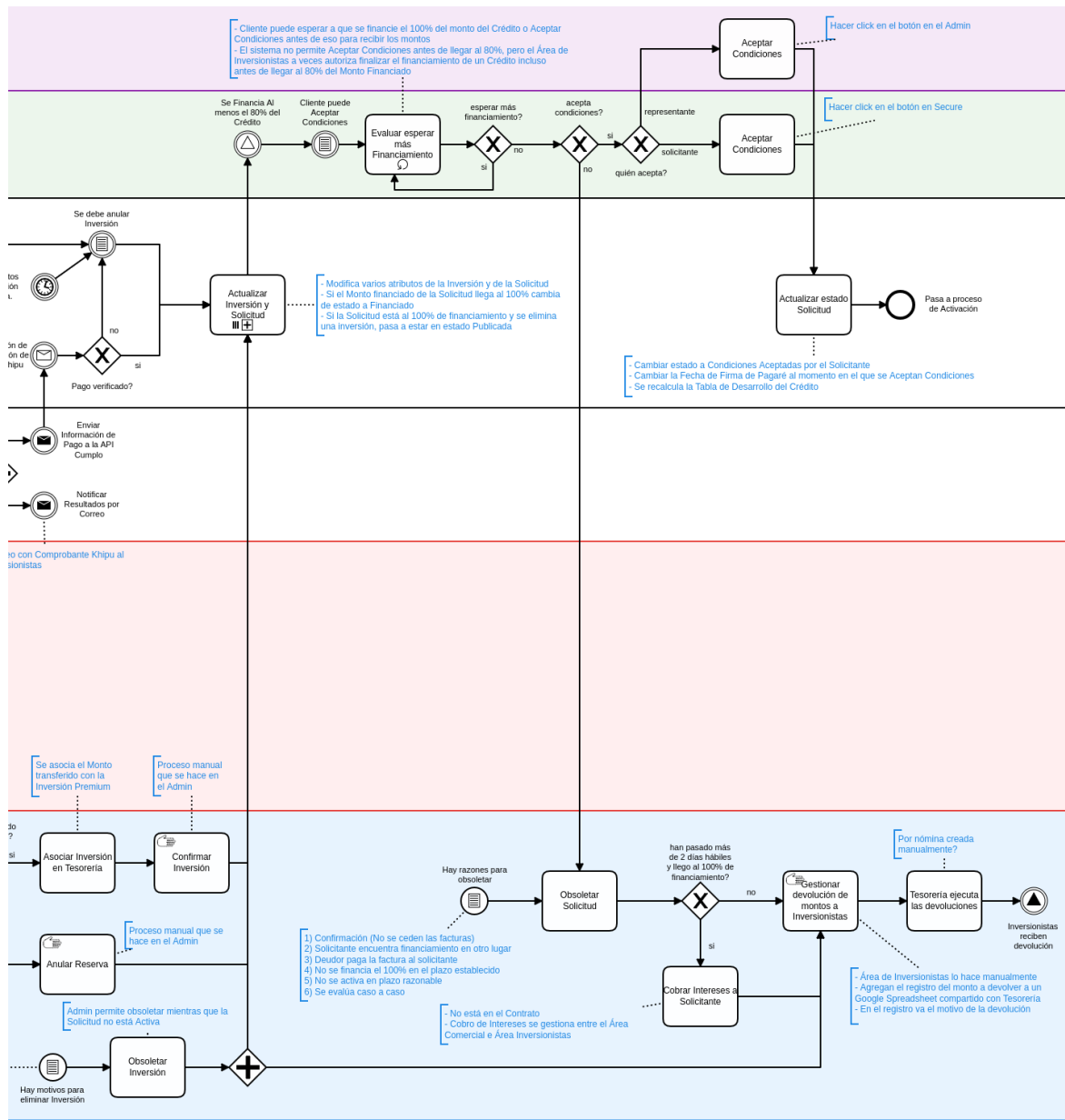


Figura A.6: Proceso de negocio detallado de inversión, en notación BPMN, parte 3.

A continuación, se muestra el modelo correspondiente al proceso de Activación en notación BPMN.

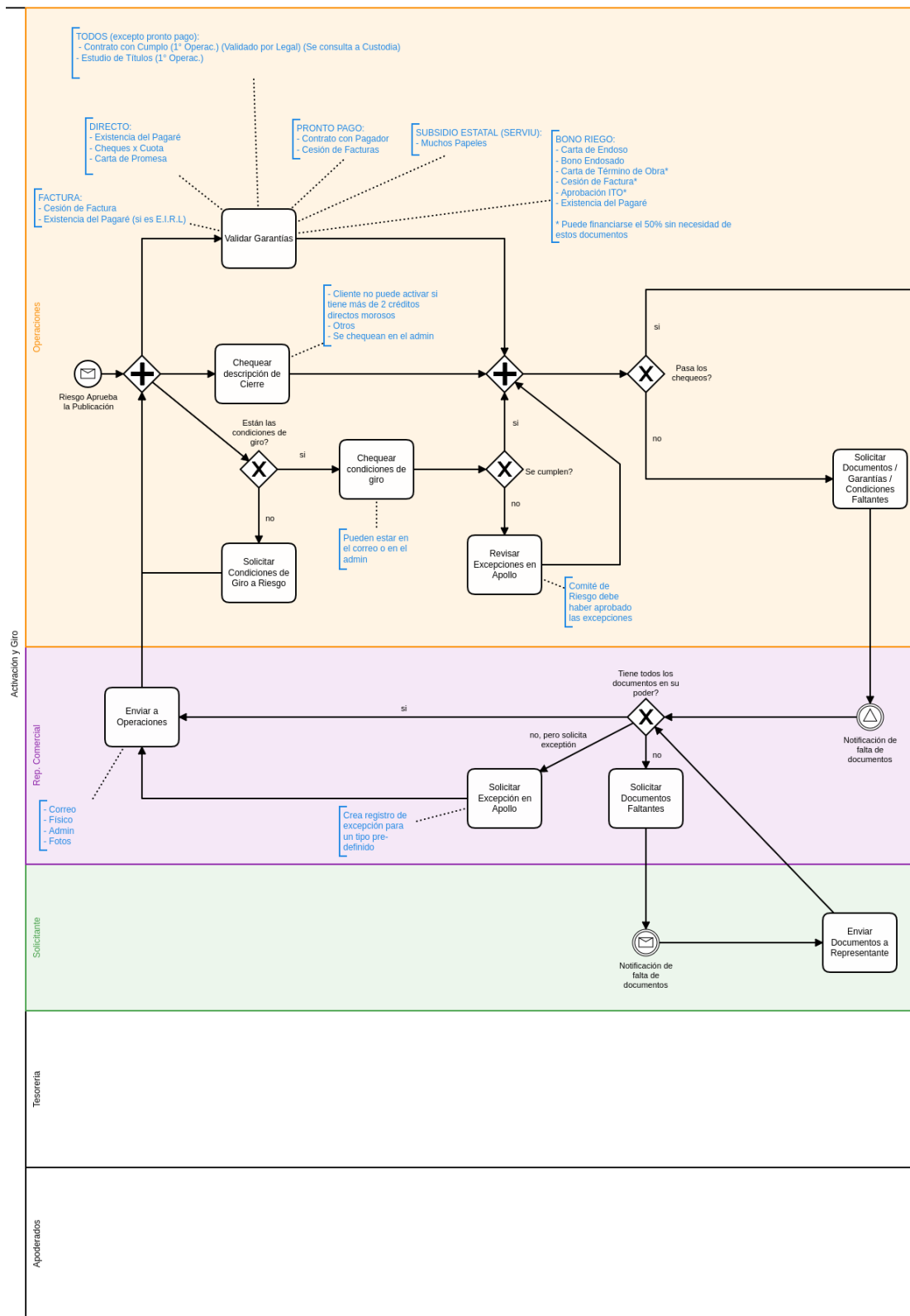


Figura A.7: Proceso de negocio detallado de activación, en notación BPMN, parte 1.

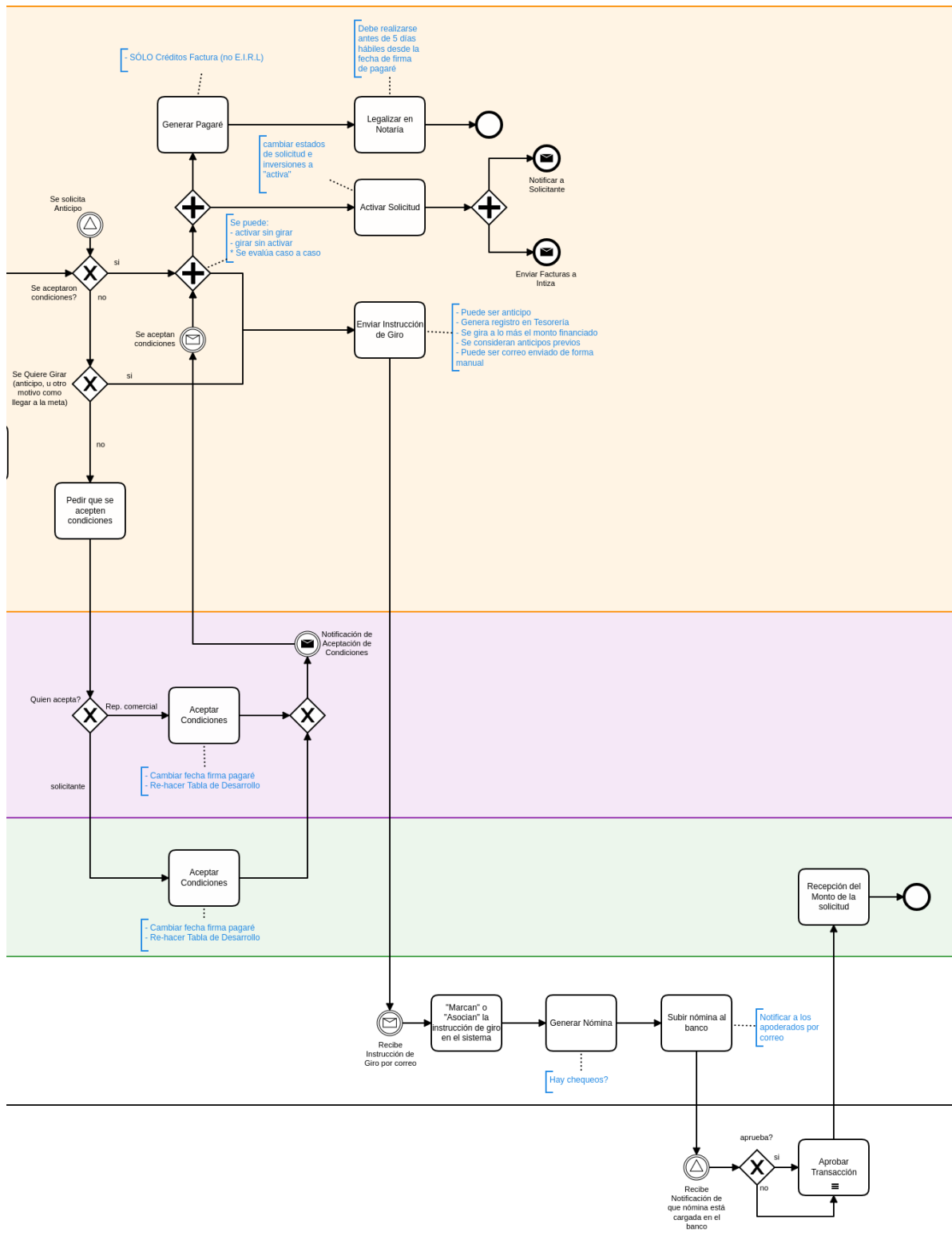


Figura A.7: Proceso de negocio detallado de activación, en notación BPMN, parte 2.

ANEXO B: CÓDIGOS SQL

```
CREATE

    ALGORITHM = UNDEFINED
    DEFINER = `root`@`localhost`
    SQL SECURITY DEFINER
VIEW `cumplodb_dwh`.`enterprise_view` AS
    SELECT
        `dfr`.`id_funding` AS `id_funding`,
        `dfr`.`admin_name` AS `admin_name`,
        `dfr`.`credit_type_name` AS `credit_type_name`,
        `dr`.`id_dim_requester` AS `id_dim_requester`,
        `dr`.`requester_business_name` AS `requester_business_name`,
        `fa`.`amount` AS `amount`,
        `fa`.`fee` AS `fee`,
        `fa`.`tir` AS `tir`,
        `fa`.`nominal_rate` AS `nominal_rate`,
        `fa`.`activated_date` AS `activated_date`
    FROM
        (((`cumplodb_dwh`.`fact_activation` `fa`
        JOIN `cumplodb_dwh`.`dim_requester` `dr` ON
        ((`fa`.`id_dim_requester` = `dr`.`id_dim_requester`)))
        JOIN `cumplodb_dwh`.`dim_funding_request` `dfr` ON
        ((`dfr`.`id_dim_funding` = `fa`.`id_dim_funding`)))
        GROUP BY `dfr`.`id_funding`
```

Código B.1: Código en lenguaje SQL necesario para la creación de la vista materializada de representantes comerciales y empresa solicitante en MySQL

```
CREATE

    ALGORITHM = UNDEFINED
    DEFINER = `root`@`localhost`
    SQL SECURITY DEFINER
VIEW `cumplodb_dwh`.`investor_view` AS
    SELECT
        `dfr`.`id_funding` AS `id_funding`,
        `fa`.`invested_amount` AS `invested_amount`,
        `dr`.`id_dim_requester` AS `id_dim_requester`,
        `dr`.`requester_business_name` AS `requester_business_name`,
        `din`.`investor_id` AS `investor_id`,
        `din`.`investor_sex` AS `investor_sex`,
        `din`.`profession_name` AS `profession_name`
    FROM
        (((`cumplodb_dwh`.`fact_activation` `fa`
        JOIN `cumplodb_dwh`.`dim_requester` `dr` ON
        ((`fa`.`id_dim_requester` = `dr`.`id_dim_requester`)))
        JOIN `cumplodb_dwh`.`dim_funding_request` `dfr` ON
        ((`dfr`.`id_dim_funding` = `fa`.`id_dim_funding`)))
        JOIN `cumplodb_dwh`.`dim_investor` `din` ON
        ((`din`.`id_dim_investor` = `fa`.`id_dim_investor`)))
        GROUP BY `din`.`investor_id`
```

Código B.2: Código en lenguaje SQL necesario para la creación de la vista materializada inversionistas en MySQL

```
CREATE
  ALGORITHM = UNDEFINED
  DEFINER = `root`@`localhost`
  SQL SECURITY DEFINER
VIEW `cumpl_o_db_dwh`.`invoice_view` AS
  SELECT
    `dfr`.`id_funding` AS `id_funding`,
    `fa`.`invoice_amount` AS `invoice_amount`,
    `dr`.`id_dim_requester` AS `id_dim_requester`,
    `dr`.`requester_business_name` AS `requester_business_name`,
    `dpy`.`invoice_payer_id` AS `invoice_payer_id`,
    `dpy`.`invoice_payer_name` AS `invoice_payer_name`
  FROM
    (((`cumpl_o_db_dwh`.`fact_activation` `fa`
      JOIN `cumpl_o_db_dwh`.`dim_requester` `dr` ON
        ((`fa`.`id_dim_requester` = `dr`.`id_dim_requester`)))
      JOIN `cumpl_o_db_dwh`.`dim_funding_request` `dfr` ON
        ((`dfr`.`id_dim_funding` = `fa`.`id_dim_funding`)))
      JOIN `cumpl_o_db_dwh`.`dim_payer` `dpy` ON ((`dpy`.`id_dim_payer`
        = `fa`.`id_dim_payer`)))
```

Código B.3: Código en lenguaje SQL necesario para la creación de la vista materializada pagadores en MySQL

ANEXO C: TABLAS DE RESULTADOS

Numero	Nombre	q1	q2	Total
1	Alejandro	0,2932	0,2325	0,5257
2	Aracelli	0,2447	0,2390	0,4837
3	Catalina	0,2168	0,2159	0,4327
4	Christian	0,2132	0,2111	0,4243
5	Daniela	0,2088	0,2009	0,4097
6	Francisca	0,2305	0,1973	0,4278
7	Ignacio	0,2199	0,2254	0,4453
8	Joseph	0,2146	0,2368	0,4514
9	Luis	0,2140	0,2080	0,4220
10	Paolo	0,2210	0,2020	0,4230
11	Pedro	0,2087	0,2078	0,4165
12	Raquel	0,2241	0,1989	0,4229
	Promedio	0,2258	0,2146	0,4404
	Desvest	0,0235	0,0151	0,0333

Tabla C.1: Detalle de resultados de pruebas de rendimiento en memoria secundaria para el escenario 1, en modelo columnar.

Numero	Nombre	q1	q2	total
1	Alejandro	2,5905	2,6566	5,2471
2	Aracelly	2,6498	2,6583	5,3081
3	Catalina	2,6697	4,1029	6,7726
4	Christian	2,6826	2,6252	5,3077
5	Daniela	5,7116	2,6359	8,3475
6	Francisca	2,6487	2,6701	5,3188
7	Ignacio	5,4098	2,7724	8,1822
8	Joseph	2,6525	2,6760	5,3284
9	Luis	2,6509	2,6598	5,3106
10	Paolo	2,6235	2,6454	5,2689
11	Pedro	2,6387	2,6814	5,3201
12	Raquel	2,6490	2,6847	5,3337
	Promedio	3,1314	2,7890	5,9205
	Desvest	1,1368	0,4154	1,1736

Tabla C.2: Detalle de resultados de pruebas de rendimiento en memoria secundaria para el escenario 1, en modelo relacional.

Numero	Nombre	q1	q2	q3	Total
1	Alejandro	0,1186	0,1122	0,1011	0,3319
2	Antonia	0,0887	0,0846	0,0990	0,2724
3	Aracelli	0,0933	0,0944	0,1035	0,2913
4	Cristian	0,1251	0,0992	0,1071	0,3314
5	Daniela	0,0939	0,1360	0,1067	0,3365
6	Fabian	0,1195	0,1019	0,0986	0,3201
7	Ignacio	0,0938	0,0968	0,0876	0,2782
8	Joseph	0,0948	0,0960	0,0818	0,2726
9	Luis	0,0777	0,0788	0,0958	0,2522
10	Nicolas	0,0873	0,0934	0,0914	0,2720
11	Oliver	0,1306	0,0982	0,1051	0,3339
12	Raquel	0,0812	0,0944	0,0822	0,2579
	Promedio	0,1004	0,0988	0,0966	0,2959
	Desvest	0,0180	0,0143	0,0090	0,0325

Tabla C.3: Detalle de resultados de pruebas de rendimiento en memoria secundaria para el escenario 2, en modelo columnar.

Numero	Nombre	q1	q2	q3	Total
1	Alejandro	2,2549	2,0858	1,8519	6,1926
2	Antonia	2,2362	1,7302	2,0181	5,9845
3	Aracelli	2,1724	2,2236	2,0473	6,4433
4	Cristian	2,1385	1,9433	2,2786	6,3604
5	Daniela	2,2555	2,1555	2,2737	6,6846
6	Fabian	2,1812	2,1887	1,8506	6,2205
7	Ignacio	2,2225	2,0266	2,0274	6,2764
8	Joseph	2,2670	2,7297	2,0656	7,0622
9	Luis	2,2118	1,9280	2,1179	6,2577
10	Nicolas	2,0330	1,8342	2,2434	6,1106
11	Oliver	2,2772	2,2472	2,3422	6,8666
12	Raquel	2,1850	2,2802	1,6812	6,1463
	Promedio	2,2029	2,1144	2,0665	6,3838
	Desvest	0,0684	0,2596	0,2007	0,3260

Tabla C.4: Detalle de resultados de pruebas de rendimiento en memoria secundaria para el escenario 2, en modelo relacional.

Numero	Nombre	q1	q2	q3	Total
1		0,1757	0,1788	0,1635	0,5180
2		0,1556	0,1682	0,1528	0,4766
3		0,1660	0,1893	0,1778	0,5331
4		0,1708	0,1798	0,1475	0,4981
5		0,1503	0,1601	0,1369	0,4473
6		0,1712	0,1620	0,1681	0,5013
7		0,1560	0,1252	0,1608	0,4420
8		0,1553	0,1287	0,1646	0,4486
9		0,1661	0,1371	0,1511	0,4543
10		0,1553	0,1677	0,1735	0,4965
11		0,1294	0,1596	0,1604	0,4494
12		0,1720	0,1554	0,1361	0,4635
13		0,1379	0,1652	0,1518	0,4549
14		0,1498	0,1409	0,1661	0,4568
	Promedio	0,1580	0,1584	0,1579	0,4774
	Desvest	0,0134	0,0193	0,0125	0,0311

Tabla C.5: Detalle de resultados de pruebas de rendimiento en memoria secundaria para el escenario 3, en modelo columnar.

Numero	Nombre	q1	q2	q3	Total
1		1,0953	1,1615	1,1608	3,4176
2		0,5289	1,1370	0,5512	2,2170
3		1,1010	1,1894	1,1885	3,4789
4		0,5274	1,1462	0,5452	2,2189
5		1,0891	1,1461	1,1459	3,3811
6		0,5312	1,1240	0,5261	2,1813
7		1,0669	5,0925	5,0925	11,2519
8		0,5534	1,2121	0,5395	2,3050
9		1,0956	1,2142	1,2140	3,5237
10		0,5255	1,1235	0,5554	2,2045
11		1,0656	1,1556	1,1507	3,3719
12		0,5261	1,1483	0,5312	2,2057
13		1,0742	1,1698	1,1678	3,4118
14		0,5243	1,1244	0,5316	2,1803
	Promedio	0,8075	1,4389	1,1357	3,3821

	Desvest	0,2872	1,0520	1,1816	2,5208
--	---------	--------	--------	--------	--------

Tabla C.6: Detalle de resultados de pruebas de rendimiento en memoria secundaria para el escenario 3, en modelo relacional.

Numero	Nombre	q1	q2	total
1	Alejandro	0,0481	0,0228	0,0710
2	Aracelly	0,0201	0,0225	0,0426
3	Catalina	0,0181	0,0214	0,0395
4	Christian	0,0198	0,0195	0,0393
5	Daniela	0,0181	0,0213	0,0393
6	Francisca	0,0190	0,0198	0,0388
7	Ignacio	0,0222	0,0193	0,0415
8	Joseph	0,0194	0,0194	0,0387
9	Luis	0,0209	0,0191	0,0400
10	Paolo	0,0193	0,0191	0,0383
11	Pedro	0,0193	0,0216	0,0409
12	Raquel	0,0205	0,0195	0,0400
	Promedio	0,0221	0,0204	0,0425
	Desvest	0,0083	0,0014	0,0090

Tabla C.7: Detalle de resultados de pruebas de rendimiento en memoria primaria para el escenario 1, en modelo columnar.

Numero	Nombre	q1	q2	total
1	Alejandro	0,0040	0,0058	0,0099
2	Aracelly	0,0051	0,0054	0,0105
3	Catalina	0,0048	0,0056	0,0103
4	Christian	0,0049	0,0054	0,0104
5	Daniela	0,0060	0,0782	0,0842
6	Francisca	0,0050	0,0054	0,0105
7	Ignacio	0,0051	0,0051	0,0102
8	Joseph	0,0054	0,0050	0,0105
9	Luis	0,0052	0,0050	0,0101
10	Paolo	0,0049	0,0063	0,0112
11	Pedro	0,0051	0,0045	0,0097
12	Raquel	0,0049	0,0056	0,0105
	Promedio	0,0050	0,0115	0,0165
	Desvest	0,0005	0,0210	0,0213

Tabla C.8: Detalle de resultados de pruebas de rendimiento en memoria primaria para el escenario 1, en modelo relacional.

Numero	Nombre	q1	q2	q3	Total
1	Alejandro	0,0057	0,0082	0,0069	0,0209
2	Antonia	0,0048	0,0058	0,0056	0,0161
3	Aracelli	0,0076	0,0092	0,0065	0,0234
4	Cristian	0,0058	0,0065	0,0061	0,0184
5	Daniela	0,0056	0,0054	0,0060	0,0170
6	Fabian	0,0063	0,0063	0,0053	0,0179
7	Ignacio	0,0072	0,0051	0,0056	0,0179
8	Joseph	0,0058	0,0077	0,0064	0,0199
9	Luis	0,0069	0,0056	0,0063	0,0189
10	Nicolas	0,0066	0,0056	0,0075	0,0197
11	Oliver	0,0057	0,0057	0,0063	0,0177
12	Raquel	0,0060	0,0056	0,0058	0,0173
	Promedio	0,0062	0,0064	0,0062	0,0188
	Desvest	0,0008	0,0013	0,0006	0,0020

Tabla C.9: Detalle de resultados de pruebas de rendimiento en memoria primaria para el escenario 2, en modelo columnar.

Numero	Nombre	q1	q2	q3	Total
1	Alejandro	0,0052	0,0050	0,0055	0,0157
2	Antonia	0,0047	0,0055	0,0049	0,0152
3	Aracelli	0,0052	0,0056	0,0084	0,0192
4	Cristian	0,0053	0,0055	0,0079	0,0188
5	Daniela	0,0053	0,0061	0,0058	0,0172
6	Fabian	0,0059	0,0068	0,0052	0,0179
7	Ignacio	0,0059	0,0051	0,0055	0,0165
8	Joseph	0,0052	0,0071	0,0059	0,0183
9	Luis	0,0051	0,0061	0,0058	0,0170
10	Nicolas	0,0063	0,0060	0,0059	0,0181
11	Oliver	0,0054	0,0062	0,0071	0,0187
12	Raquel	0,0065	0,0065	0,0072	0,0202
	Promedio	0,0055	0,0060	0,0063	0,0177
	Desvest	0,0005	0,0006	0,0011	0,0015

Tabla C.10: Detalle de resultados de pruebas de rendimiento en memoria primaria para el escenario 2, en modelo relacional.

Numero	Nombre	q1	q2	q3	Total
1		0,0072	0,0055	0,0055	0,0182
2		0,0066	0,0056	0,0054	0,0176
3		0,0070	0,0053	0,0059	0,0182
4		0,0054	0,0060	0,0061	0,0175
5		0,0072	0,0061	0,0059	0,0192
6		0,0065	0,0059	0,0053	0,0178
7		0,0052	0,0060	0,0063	0,0174
8		0,0070	0,0061	0,0061	0,0191
9		0,0053	0,0062	0,0059	0,0174
10		0,0058	0,0054	0,0061	0,0173
11		0,0061	0,0101	0,0063	0,0225
12		0,0059	0,0060	0,0059	0,0177
13		0,0088	0,0062	0,0080	0,0230
14		0,0058	0,0063	0,0060	0,0181
	Promedio	0,0064	0,0062	0,0060	0,0186
	Desvest	0,0010	0,0012	0,0006	0,0028

Tabla C.11: Detalle de resultados de pruebas de rendimiento en memoria primaria para el escenario 3, en modelo columnar.

Numero	Nombre	q1	q2	q3	Total
1		0,0046	0,0052	0,0063	0,0161
2		0,0439	0,0062	0,0287	0,0788
3		0,0049	0,0048	0,0069	0,0166
4		0,0391	0,0055	0,0257	0,0703
5		0,0042	0,0070	0,0070	0,0183
6		0,0423	0,0056	0,0271	0,0750
7		0,0034	0,0059	0,0060	0,0153
8		0,0354	0,0058	0,0303	0,0714
9		0,0048	0,0058	0,0058	0,0163
10		0,0370	0,0044	0,0253	0,0667
11		0,0049	0,0063	0,0058	0,0169
12		0,0395	0,0065	0,0255	0,0714
13		0,0042	0,0064	0,0055	0,0160
14		0,0397	0,0038	0,0257	0,0692
	Promedio	0,0220	0,0057	0,0165	0,0442

	Desvest	0,0183	0,0009	0,0108	0,0300
--	---------	--------	--------	--------	--------

Tabla C.12: Detalle de resultados de pruebas de rendimiento en memoria primaria para el escenario 3, en modelo relacional.