Tesis de Maestría

BIG DATA APLICADO EN EL SECTOR BANCARIO TRADICIONAL PARA LOGRAR UNA MAYOR VENTAJA COMPETITIVA FRENTE A LAS FINTECH

Angelo Martínez C.

FACULTAD DE INGENIERÍA

Maestría en Tecnología de la Información

Director de Tesis

Prof. Mg. Daniel Tokman

Buenos Aires, Argentina

Abril 2018

DEDICATORIA

A Dios por haberme permitido llegar hasta este punto y haberme dado fuerzas y salud para culminar con éxito este largo viaje.

A mi madre por haberme apoyado incondicionalmente, su amor, consejos y valores me han permitido ser una persona de bien.

A mi novia Mercedes y mi hija Jenny por su constante motivación y apoyo, por recordarme cada día que todo es posible cuando le ponemos ganas y empeño suficiente.

RECONOCIMIENTOS

Me gustaría que estas líneas sirvieran para expresar mi más profundo y sincero agradecimiento a todas aquellas personas que con su ayuda han colaborado en la realización de esta tesis.

Mi más profundo agradecimiento a mi director de tesis, Daniel Tokman, por tener tan claro el camino de la investigación y compartirlo conmigo, por ser mi guía en el proceso de aprendizaje y por estar siempre dispuesto a compartir conmigo su tiempo y conocimientos.

Mi mayor gratitud a mis profesores de la Maestría en Tecnología de la Información de la Universidad de Palermo por haberme abierto las puertas del mundo de la ingeniería.

Agradezco a mis compañeros de la maestría, porque cada trabajo en equipo fue una experiencia enriquecedora.

RESUMEN DE LA TESIS

Es común en muchas organizaciones la toma de decisiones basadas en la experiencia de sus directores y no apoyada en los datos. En otros casos, éstas se ven sumidas en un mar de datos sin saber qué hacer con ellos. Disponer de información exacta, clara y oportuna puede considerarse como la base del éxito de una organización.

En este trabajo, se establecen lineamientos para el aprovechamiento de grandes volúmenes de datos que dispone el sector bancario tradicional, que por su gran magnitud y complejidad se ha visto limitado en aprovechar estas tecnologías disruptivas. Esto, ha dado origen a nuevos competidores, las fintech, que haciendo uso de la tecnología más innovadora están ganando terreno en poco tiempo.

Mediante la creación de un modelo de trabajo ágil para administrar tecnologías Big Data en el sector financiero tradicional se pretende lograr una ventaja competitiva, enfocarse en el cliente ofreciéndole productos innovadores y personalizados que cumplan con sus necesidades y extraer conocimiento de los datos para aprovechar nuevas oportunidades de negocios mediante una toma de decisiones más inteligente sustentada en los datos.

Palabras claves: Big data, Cloudera, Hadoop, Fintech, Bancos Tradicionales, Modelo de trabajo.

TABLA DE CONTENIDOS

1. INTRODUCCIÓN 1
1.1. Antecedentes 1
1.2. Problemática 3
1.3. Justificación 5
1.4. Objetivos 6
1.4.1. Objetivo General 6
1.4.2. Objetivos Específicos 6
1.5. Alcance 7
1.6. Estructura del documento 8
2. MARCO TEÓRICO 9
2.1. Sector Bancario Tradicional 9
2.1.1. Banco Central de la República Argentina 11
2.1.2. Ámbito Regulatorio 11
2.2. Fintech 13
2.2.1. Características de las fintech 15
2.2.2. Tipos de fintech 16
2.2.3. Ámbitos de actividad de las fintech 16
2.2.4. Fintech en la Argentina 18
2.2.5. Marco Regulatorio de las fintech en la Argentina 19
2.3. Almacenamiento de datos 20
2.3.1. Tipos de datos: estructurados, semiestructurados y no estructurados 20
2.3.2. Base de datos relacional 21
2.3.3. Data Warehouse 22
2.3.3.1. Ventajas de usar Data Warehouse 24
2.3.3.2. Desventajas de usar Data Warehouse 24
2.3.3.3. Características del Data Warehouse 25
2.3.3.4. Proceso ETL 26
2.3.3.5. OLTP y OLAP 26
2.3.3.6. Esquemas: Estrella y Copo de Nieve 27
2.3.4. Data Mart 29
2.3.4.1. Ventajas de usar Data Mart 30

2.3.5. Diferencias entre Data Warehouse y Data Mart 31
2.3.6. Bases de datos NoSQL 32
2.3.6.1. Teorema de CAP 33
2.3.6.2. Tipos de bases de datos NoSQL 36
2.3.7. Data Lake 38
2.3.7.1. Características de un Data Lake 39
2.3.7.2. Ventajas de un Data Lake 40
2.3.7.3. Desventajas de un Data Lake 40
2.3.8. Diferencias entre un Data Lake y un Data Warehouse 41
2.4. Big Data 42
2.4.1. Dimensiones de Big Data 42
2.4.2. ¿De dónde proviene toda esa información? 43
2.4.3. Big Data en el Sector Bancario 44
2.4.4. Hadoop 45
2.4.5. Ecosistema Hadoop 47
2.4.6. Distribuciones de Hadoop 50
2.4.6.1. Cloudera 50
2.4.6.2. Hortonworks 53
2.4.6.3. MapR 56
3. MODELO PROPUESTO 58
3.1. Motivos de fracaso de un proyecto de Big Data 58
3.2. Aspectos a tomar en cuenta al iniciar un proyecto de Big Data 59
3.3. Modelo para la administración de Big Data en el sector bancario 61
3.3.1. Fase 1: Definición 62
3.3.2. Fase 2: Identificar fuentes de datos 64
3.3.3. Fase 3: Diseño de la aplicación 66
3.3.4. Fase 4: Captura y almacenamiento de datos 68
3.3.5. Fase 5: Modelado y limpieza de datos 69
3.3.6. Fase 6: Análisis 70
3.3.7. Fase 7: Evaluación y monitoreo 71
4. IMPLEMENTACIÓN DEL MODELO 72
4.1. Implementación 72
4.2. Caso de estudio 73

4.3. Prototipo 73
4.3.1. Aplicación de la fase de Definición 73
4.3.2. Aplicación de la fase de Identificación de fuentes de datos 74
4.3.3. Aplicación de la fase de Diseño de la Aplicación 74
4.3.4. Aplicación de la fase de Captura y almacenamiento de datos 75
4.3.5. Aplicación de la fase de Modelado y limpieza 77
4.3.6. Aplicación de la fase de Análisis 78
4.3.7. Aplicación de la fase de Evaluación y Monitoreo 79
4.4. Presentación de resultados 80
4.4. Estudio comparativo entre las entidades bancarias tradicionales y las fintech 88
5. CONCLUSIONES 90
6. FUTURAS LÍNEAS DE INVESTIGACIÓN 91
BIBLIOGRAFÍA 92

LISTA DE TABLAS

Tabla 1. Diferencias entre Data Warehouse y Data Mart 31
Tabla 2. Comparación entre las características de NoSQL y base de datos relacionales . 35
Tabla 3. Comparación entre las características de NoSQL y base de datos relacionales . 41
Tabla 4. Dimensiones de Big Data 43
Tabla 5. Características de Hadoop 47

LISTA DE FIGURAS

Figura 1. Mapa Fintech de Argentina 18
Figura 2. Esquema Estrella 27
Figura 3. Esquema de Copo de Nieve..... 28
Figura 4. Teorema de CAP .....**Error! Bookmark not defined.**
Figura 5. Ecosistema de Hadoop 49
Figura 6. Ecosistema de Cloudera 52
Figura 7. Ecosistema de Hortonworks..... 54
Figura 8. Plataforma de datos convergentes MapR 57
Figura 9. Fases del modelo propuesto..... 62
Figura 10. Estilo de Arquitectura para Big Data..... 67
Figura 11. Diseño de la Aplicación..... 75
Figura 12. Importar datos a Apache Hive..... 76
Figura 13. Consultar los datos con HUE 77
Figura 14. Conector Cloudera Hadoop 78
Figura 15. Consultar los datos con Cloudera Manager..... 79
Figura 16. Cantidad de transacciones por canal..... 80
Figura 17. Cantidad de transacciones por meses 80
Figura 18. Promedio de edad de los clientes por canal..... 81
Figura 19. Promedio de edad de clientes por mes 81
Figura 20. Canal más usado de un cliente..... 82
Figura 21. Pagos que realiza en Internet..... 82
Figura 22. Cantidad de transacciones por rango de edad..... 83
Figura 23. Cantidad de transacciones por rango de edad..... 83
Figura 24. Tipos de consumo..... 84
Figura 25. Promedio de interacción de un cliente en los canales 85
Figura 26. Monto de ingresos y gastos por meses (buena salud)..... 86
Figura 27. Situación financiera (buena salud) 86
Figura 28. Monto de ingresos y gastos por meses (mala salud) 87
Figura 29. Situación financiera (mala salud)..... 87

LISTA DE SIGLAS

SIGLA	SIGNIFICADO
AWS	Amazon Web Services
BCRA	Banco Central de la República Argentina
CDH	Cloudera Distribution for Hadoop
CRM	Customer Relationship Management
DW	Data Warehouse
ERP	Enterprise Resource Planning
ETL	Extract, Transform and Load
GAFA	Google, Amazon, Facebook y Apple
HDFS	Hadoop Distributed File System
HDF	Hortonworks DataFlow
HDP	Hortonworks Hadoop distribution
HTML	HyperText Markup Language
JSON	JavaScript Object Notation
OLAP	On-Line Analytical Processing
OLTP	On-line Transaction Processing
P2P	Peer to Peer
P2B	Peer to Business
XML	Extensible Markup Language
YARM	Yet Another Resource Negotiator

1. INTRODUCCIÓN

1.1. Antecedentes

En la actualidad las organizaciones se ven sumidas por el ambiente global en el que se desempeñan, siendo de gran importancia la toma de decisiones que les permita actuar de forma estratégica anticipándose a tendencias de mercado y cubriendo las necesidades de sus clientes, lo cual les ayudará a diferenciarse del resto de manera exitosa y rentable. Para lograr una eficiente gestión es imprescindible tomar en cuenta un factor clave que es la toma de decisiones. Disponer de información exacta, clara y oportuna puede considerarse como la base del éxito.

Es evidente el incremento de información que se ha generado durante los últimos años. Prácticamente todo lo que hacemos genera datos. El uso masivo de internet y la creación de grandes comunidades en línea a las que comúnmente llamamos redes sociales significó un cambio profundo en nuestros hábitos. A diario recurrimos a Facebook para saber de nuestros amigos, a Google para buscar información, a Twitter para estar al tanto de lo sucedido minuto a minuto. Asimismo, Internet de las cosas está registrando un crecimiento significativo y se espera que siga evolucionando con el despliegue de las redes 5G. Estamos atravesando una época de revolución de datos. Esto es impulsado no solo por la abundancia de datos actual, sino por las tecnologías fundamentales que cambia la forma en que reunimos, almacenamos, analizamos y transformamos la información.

De hecho, el 90 por ciento de los datos actuales se crearon sólo en los últimos años (IBM, 2015), y ahora estamos duplicando la velocidad con que se producen los datos. Un estudio de Seagate, líder global en administración y almacenamiento de datos, pronostica un incremento en 10 veces del volumen de datos a nivel mundial para el año 2025 (IDC, 2017). Por otra parte, la constante evolución y reducción de costos de la tecnología ha representado un gran crecimiento en las capacidades de generar y colectar datos, debido al gran poder de procesamiento de los computadores como a su bajo costo de almacenamiento.

Esto se presenta para el sector bancario tradicional como un gran desafío para aprovechar nuevas oportunidades provenientes de los datos con el fin de poder ofrecer productos y servicios financieros personalizados que satisfagan las necesidades de los clientes. Las entidades financieras cuentan con datos extraordinariamente ricos y variados de sus clientes como el manejo de sus cuentas personales, productos y servicios financieros contratados y operaciones en los cajeros automáticos y demás canales de atención. Usando los datos correctamente podrían hacer un marketing dirigido, conocer qué compran, en qué cantidad y cada cuánto tiempo y mejorar la reputación de la marca (KPMG, 2017).

Las nuevas tecnologías e internet han hecho posible que en el sector financiero aparezcan nuevos competidores que están revolucionando los procesos habituales de la banca, las Fintech. Obligando a las entidades bancarias a replantearse el modelo de negocio para un cliente más exigente y cada vez más ligado a la tecnología (Igual, 2016).

1.2. Problemática

El nacimiento de las nuevas tecnologías a través del internet y el incremento desmesurado de los datos ha obligado a muchas organizaciones a reinventar su modelo de negocio con el fin de conocer a sus clientes para ofrecerles los productos y servicios a medida en el momento que lo necesiten. El conocimiento de lo que van a demandar los clientes en el futuro es la principal razón por la que apuestan las organizaciones.

El sector bancario ha estado tradicionalmente centrado en el producto y no en el cliente. Su naturaleza, magnitud y complejidad han ralentizado su adaptación a las nuevas tecnologías y necesidades del mercado. La actual revolución de los hábitos de los clientes expuestos en las redes sociales y otros medios ha conseguido que la banca se plantee seriamente la necesidad de centrarse en el cliente. Si bien es cierto, que cuentan con información valiosa sobre los consumidores que cualquier otra empresa. Por ejemplo, los bancos conocen el nivel de ingreso y gasto aproximado de sus clientes, en qué empresas gastan el dinero, lo que les proporciona también una gran comprensión de qué tipos de bienes y servicios el cliente está comprando. Conocen también exactamente cuándo los hacen la mayor parte de estas compras y dónde estaban cuando se hizo la compra. No han sabido usar favorablemente esta información que además puede ser más completa añadiendo fuentes de datos externas que aportarían valor adicional sobre sus clientes para una toma de decisiones basada en los clientes.

Esto ha creado una brecha que han sabido aprovechar nuevos competidores que usan las nuevas tecnologías y el poder que ofrece el Big Data para ofrecer servicios financieros que resuelvan las necesidades reales los clientes de manera sencilla y de forma digital, su nombre son fintech. El gran secreto ha sido saber aprovechar la información disponible en Internet, para optimizar procesos y servicios. En particular, se centran en un solo segmento o producto, dando gran importancia a la relación con el cliente. Si bien su tamaño en la actualidad es muy pequeño y no preocupa ahora a la banca, está previsto que este sector crezca de forma exponencial y que para los próximos años las cifras de millones que muevan las empresas fintech sean de miles de millones.

Las fintech son empresas que intermedian en todos los ámbitos del mundo de las finanzas actuando como brokers, como mediadores de pago, como emisores y receptores de transferencias o como asesores financieros. Éstas, se han dado cuenta de la gran oportunidad que existe y han decidido a incursionar en el monopolio de las finanzas manejado por unos pocos bancos que se han dormido en los laureles, ofreciéndole al usuario una alternativa mucho más rápida, económica y transparente para gestionar sus finanzas. Su fortaleza consiste en ofrecer soluciones más rápidas, con menos recursos y empleando menos dinero. Las nuevas generaciones ya se están acostumbrando y adaptando a trabajar de esta forma y es solo cuestión de tiempo que esa nueva forma de hacer y de actuar se traslade al mundo de las finanzas tradicionales.

1.3. Justificación

Los bancos deben innovar constantemente para seguir siendo competitivos en un entorno de gran disrupción donde nuevos competidores se suman cada día. La transformación digital ha dejado de ser una opción para las empresas y es cada vez más una necesidad real. El mercado de hoy exige que las entidades bancarias generen decisiones inteligentes permitiéndoles ser más eficaces y eficientes. Poder contar con información relevante y en tiempo real, permite tomar acciones correctivas sobre la marcha de modo más oportuno que viendo resultados históricos.

Aunque las entidades bancarias están realizando inversiones en este sentido, todavía hay mucho recorrido para conseguir una digitalización que, más allá de vender los productos de siempre por nuevos canales, se enfoque en una nueva forma de hacer negocio. Mientras tanto nuevas organizaciones como las fintech se dirigen a los clientes de forma ágil para cubrir estas necesidades a una generación dinámica que se adapta fácilmente a los constantes cambios tecnológicos.

A esto hay que sumar la situación que experimenta el sector bancario con una regulación y supervisión que supone un sobre esfuerzo en costes y flexibilidad, limitando en algunos casos la competitividad con nuevos competidores potenciales.

Esta amenaza podría convertirse en una gran oportunidad para las entidades bancarias pudiendo experimentar y aprender de las fintech para mejorar sus procesos e integrar el nuevo modelo de negocio centrado en el cliente.

1.4. Objetivos

Los objetivos de esta tesis surgen a partir de la problemática ya planteada. A continuación, se presenta el objetivo general y los objetivos específicos.

1.4.1. Objetivo General

Construir un modelo de trabajo para la implementación de Big Data aplicado en el sector bancario tradicional para ofrecer productos y servicios innovadores que satisfagan las necesidades de los clientes permitiendo así lograr una ventaja competitiva frente a las fintech.

1.4.2. Objetivos Específicos

- Realizar un estudio comparativo de las entidades bancarias tradicionales y las empresas fintech.
- Analizar los gustos y hábitos de los clientes con información de sus compras realizadas.
- Realizar una segmentación avanzada de clientes para enviar ofertas personalizadas.
- Anticiparse a la detección de abandonos de clientes para realizar acciones de retención.
- Crear un prototipo del modelo propuesto.
- Realizar un análisis comparativo de los resultados obtenidos en base al modelo propuesto.

1.5. Alcance

En el desarrollo de esta tesis se considera los siguientes aspectos como delimitantes del alcance de la misma:

- Para la implementación del modelo propuesto se van a usar algunas tecnologías asociadas a Big Data. La más conocida, Hadoop como un repositorio centralizado de datos con el fin de abatir el costo de almacenamiento e incrementar la capacidad de procesamiento de grandes volúmenes de datos.
- Se va a utilizar información de una entidad bancaria tradicional para tener una visión completa del cliente. Logrando primero conocer sus gustos, cómo interactúa, dónde interactúa, en qué gasta, etc. y así crear marketing personalizado que sirva para casos específicos.
- El resultado del modelo propuesto a partir del caso de estudio permite que se ajuste a las necesidades y objetivos perseguidos de cualquier institución bancaria tradicional, sirviendo como una guía que debe ser adaptada.
- Se busca lograr un entendimiento de la situación actual teniendo como principal actor a la banca tradicional y su nuevo competidor la fintech.
- Se va a realizar un estudio de las distribuciones que ofrece Hadoop y se va a elegir uno para elaborar el caso de estudio.

1.6. Estructura del documento

El trabajo está organizado en cinco secciones: Introducción, Marco Teórico, Modelo Propuesto, Implementación del Modelo, Conclusiones y Futuras Líneas de Investigación.

Los temas que hemos tratado, corresponden a la primera sección que se centró en los antecedentes, problemática, objetivo general y los objetivos específicos que se pretenden alcanzar mediante la presente investigación.

En el Marco Teórico, se mencionan una serie de conceptos y teorías usadas para formular y desarrollar el presente trabajo. Por ello se hacen necesario estudiar las diferentes tecnologías relacionadas a Big Data y relacionarlas al sector bancario.

Mediante el uso de las tecnologías antes estudiadas, se pretende demostrar un Modelo Propuesto que me permita identificar y comprender las oportunidades que nos ofrece el Big Data como apoyo al sector bancario tradicional. Partiendo del Modelo Propuesto, la cuarta sección hace referencia a la Implementación del Modelo Propuesto.

Como resultado de la presente investigación, se presentan las conclusiones obtenidas y se mencionan las futuras líneas de investigación.

2. MARCO TEÓRICO

2.1. Sector Bancario Tradicional

El sector financiero ha estado tradicionalmente centrado en el producto. Su naturaleza y complejidad han ralentizado su adaptación a las nuevas realidades sociales. La actual revolución de los hábitos de vida ha conseguido que la banca en la transformación digital se plantee seriamente la necesidad de centrarse en el cliente. Las entidades financieras tienen un conocimiento superficial de sus clientes y una cartera de productos muy amplia y compleja. Apostar por conseguir tener un profundo conocimiento del perfil y sus hábitos de consumo de sus clientes y un portafolio de productos muy sencillo marcará el éxito futuro de las entidades financieras (Macario, 2018).

En todo el mundo, la industria de los servicios financieros está atravesando por una etapa de transformación digital que amenaza no solo su volumen y margen de utilidades, sino también, y en algunos casos, la existencia misma de aquellos actores tradicionales que no logren adaptarse a los avances tecnológicos y a las exigencias de sus clientes. Aunque las entidades financieras están realizando inversiones en este sentido, todavía hay mucho recorrido para conseguir una digitalización que, más allá de vender los productos de siempre por nuevos canales, se enfoque en una nueva forma de hacer negocio. Las entidades se enfrentan ahora a un nuevo desafío: explotar y obtener rentabilidad de la gran cantidad de datos que manejan de sus clientes lo que les va a permitir tener ventaja competitiva con el uso de la información (KPMG, 2017).

La regulación, la atracción de nuevos clientes y la necesidad de recuperar la reputación perdida se convierte en las principales preocupaciones para el sector bancario tradicional. El cliente cada vez es más exigente y preciso en la demanda de productos y servicios bancarios que cumplan sus expectativas como usuario. El cambio de mentalidad del usuario y la tecnología están revolucionando su relación con los servicios y productos ofrecidos por las entidades financieras (Groenfeldt, 2013).

Cambios que deberán realizar los bancos:

- Centrar el modelo de negocio en el cliente
- Optimizar la distribución
- Simplificar los modelos operativos
- La información como ventaja competitiva
- Innovación
- Gestión proactiva de riesgos, capital y regulación

El sector de la banca tradicional se enfrenta a una serie de retos: regulaciones, competir con los nuevos actores, darle valor a la información que posee, promover la innovación y agilidad empresarial, entre otros. En definitiva, la búsqueda de un modelo flexible y eficiente que sea capaz de generar valor en los usuarios. En definitiva, el consumidor tiene más poder que nunca, por eso, la relación con el cliente y el cuidado de su experiencia serán vitales para garantizar su éxito.

2.1.1. Banco Central de la República Argentina

El BCRA (Banco Central de la República Argentina) es el organismo rector de todo el sistema financiero nacional, regulando a 120 entidades financieras, a través de normas y procedimientos estrictamente auditados.

Entre sus principales objetivos estratégicos se encuentran (Banco Central de la República Argentina, 2017):

- Estabilidad monetaria: baja sostenible y sistemática de la tasa de inflación.
- Estabilidad financiera: desarrollo y profundización del sistema financiero doméstico, en conjunto con una supervisión y regulación macroprudencial del sector de acuerdo a los mejores estándares internacionales.
- Bancarización, inclusión financiera y medios de pago electrónicos: reducción del uso de efectivo, agilización y mayor seguridad en las transacciones, mayor educación financiera.

2.1.2. Ámbito Regulatorio

Se mencionan a continuación algunas de las regulaciones más importantes dictadas por BCRA (Polo, 2017) (Banco Central de la República Argentina, 2018) vinculadas con la apertura tecnológica:

- Alias CBU: ya no es necesario recordar los 22 números del CBU para realizar una transferencia. Hoy se puede realizar a través de un alias, sin importar qué banco o tipo de cuenta hay detrás.
- Aplicaciones Mobile: las entidades financieras deben proveer a sus clientes, sin costo alguno, la posibilidad de operar a través de aplicaciones mobile.
- Depósito remoto de cheques: ya no es necesario acercarse a una sucursal a depositar un cheque. Las entidades financieras deben proveer una solución de digitalización y depósito remoto de cheques a través de dispositivos mobile.
- Cajas de ahorro y tarjeta de débito gratis: las entidades financieras deben proveer una caja de ahorro sin costo a cualquier habitante argentino.
- Menores bancarizados: menores de entre 12 y 17 años pueden ser clientes de entidades financieras, con restricciones puestas por sus padres.
- Baja y creación de servicios: a través de canales digitales se puede crear o dar de baja una cuenta de cualquier tipo y utilizarla inmediatamente.
- Billetera Virtual: no hace falta disponer de DNI y tarjetas físicas para realizar un pago. Las entidades financieras deben brindar sistemas de pago electrónico a sus clientes; es decir, brindar la posibilidad de adherir a diferentes medios de pago de diferentes entidades (tarjeta de débito, crédito o alias), brindando un medio de pago seguro a los clientes. La normativa relacionada a este ítem nombra sistemas como POS Móvil, Botón de Pago y la Billetera Móvil PEI.
- Débito inmediato (DEBIN): este medio habilito a las entidades financieras y a nuevos actores de la industria de medios de pago a debitar fondos de las cuentas bancarias de sus clientes, previa autorización de los mismos, para cursar pagos.

2.2. Fintech

Las fintech (finance + technology) son empresas innovadoras que están emergiendo en estos últimos años y que ofrecen productos y servicios financieros innovadores utilizando la tecnología. Las primeras fintech tienen su origen en el año 2008, aunque su impulso y desarrollo no se produce hasta el año 2010, principalmente en Estados Unidos y Reino Unido. Estas, ofrecen soluciones tanto para personas físicas como para empresas, y son auténticas especialistas en áreas concretas de los servicios financieros. Una de las principales vías de éxito de las fintech, es precisamente el desarrollo desde cero de una única propuesta de negocio centrada en un área específica, a diferencia de la banca tradicional que ofrece muchos productos caracterizados por ser complejos, difíciles de encontrar, con problemas de transparencia y nada ágiles (BBVA, 2018).

Las fintech han encontrado la forma de hacer que los servicios financieros sean mucho más simples a través del uso de la tecnología y la innovación, con estructuras muy flexibles y absolutamente orientadas a satisfacer las necesidades y mejorar la experiencia del cliente. Esta facilidad de desarrollo les permite ir mucho más rápido que las instituciones financieras tradicionales (Schultze, 2018). Las empresas fintech se han dado cuenta de la brecha en el sector financiero, creada por los bancos en su inmovilismo y en sus ansias de aumentar sus cuentas de resultados, con su monopolio de las finanzas, se han dormido en los laureles y ahora las fintech están aprovechando esa oportunidad ofreciéndole al usuario una alternativa mucho más sencilla y económica para sus finanzas (Igual, 2016)

Las empresas fintech están creciendo rápidamente como una alternativa respecto a la banca tradicional, sus argumentos principales son presentarse como una nueva opción, más transparente y eficiente que los productos ofrecidos por la banca que no gozan de una buena reputación. Al mismo tiempo, a medida que se aprueban nuevas regulaciones legales y el inversor ve que son empresas de confianza, poco a poco se ha ido perdiendo ese miedo y ahora la intermediación en las finanzas a través de empresas fintech se está multiplicando. (Gasalla, 2016). Asimismo, el nacimiento de las nuevas tecnologías a través de internet están permitiendo al consumidor un mejor control de su dinero y de sus inversiones, transformándolos en clientes mucho más exigentes e informados de lo que eran antes de la masificación del internet, ya que, conocen en tiempo real lo que está sucediendo en el resto del mundo, qué hay nuevo, qué opinan los que ya han consumido un producto o servicio, cuáles alternativas ofrece la competencia y cuáles son los precios en cualquier lugar del mundo (Hoder, Wagner, Sguerra, & Bertol, 2016).

Si bien su tamaño en la actualidad es muy pequeño en términos de negocio y no preocupa ahora a la banca, está previsto que este sector crezca de forma exponencial en los próximos años. Las fintech son un desafío real y están cambiando el sector de las finanzas caracterizados por la tecnología, innovación, practicidad, bajos costos de operación y flexibilidad en el mercado. El principal reto para la banca tradicional, es que estas fintech son capaces de ofrecer un servicio más eficiente, digital y pensado desde cero para satisfacer las necesidades de los usuarios mejor que ellos (Vives, 2017).

El desarrollo futuro de las fintech dependerá en buena parte de la regulación. El reto para el regulador es mantener un campo de juego equilibrado entre la banca tradicional y las fintech de manera que se promueva la innovación y siempre enfocándose al cliente para que así las nuevas tecnologías proporcionen un mayor nivel de bienestar y satisfacción, todo esto con el fin de preservar la estabilidad financiera (Martínez, 2018).

2.2.1. Características de las fintech

- Su propuesta está centrada en algún aspecto concreto de las finanzas, por ejemplo, préstamos, pagos, asesoramiento financiero, transferencias, análisis de datos, inversiones, entre otros).
- Mediante el uso de nuevas tecnologías ofrecen soluciones a problemas financieros de los clientes o a necesidades mal atendidas por la banca y con una gran especialización. Es decir, añaden valor sobre los servicios financieros actuales haciéndolos más intuitivos, ágiles, confiables y transparentes.
- Uno de los puntos más fuertes es su cultura de la innovación con una filosofía de romper los anteriores esquemas utilizando tecnología actual para dar mejores soluciones a los servicios financieros.
- Da la posibilidad de que uno mismo gestione cualquier tipo de información y que tenga el control absoluto, sin intermediarios, es decir productos que puedan ofrecerse de persona a persona.

2.2.2. Tipos de fintech

Dependiendo de su naturaleza se pueden clasificar en dos grupos: startups y compañías de tamaño pequeño y GAFAs.

- Startups y compañías de tamaño pequeño: las fintech pueden considerarse como startups orientadas a las finanzas. Una startup, es una empresa pequeña o mediana, de creación reciente vinculada al mundo tecnológico. Tratan de explotar nichos de mercados específicos que están de moda o transformaciones al mundo digital de alguna necesidad que se tenga.
- GAFAs: es el acrónimo de Google, Amazon, Facebook y Apple. Son conocidas también como bigtechs y cuentan con una capacidad para introducir en cualquier ámbito de negocio. A diferencia de las startups, estas producen un gran poder intimidatorio para la banca tradicional por el potencial económico y tecnológico que tienen para penetrar en el negocio financiero y desplazar a los bancos.

2.2.3. Ámbitos de actividad de las fintech

De acuerdo con los sectores de actividad a los que están orientadas, se pueden establecer los siguientes ámbitos (Igual, 2016):

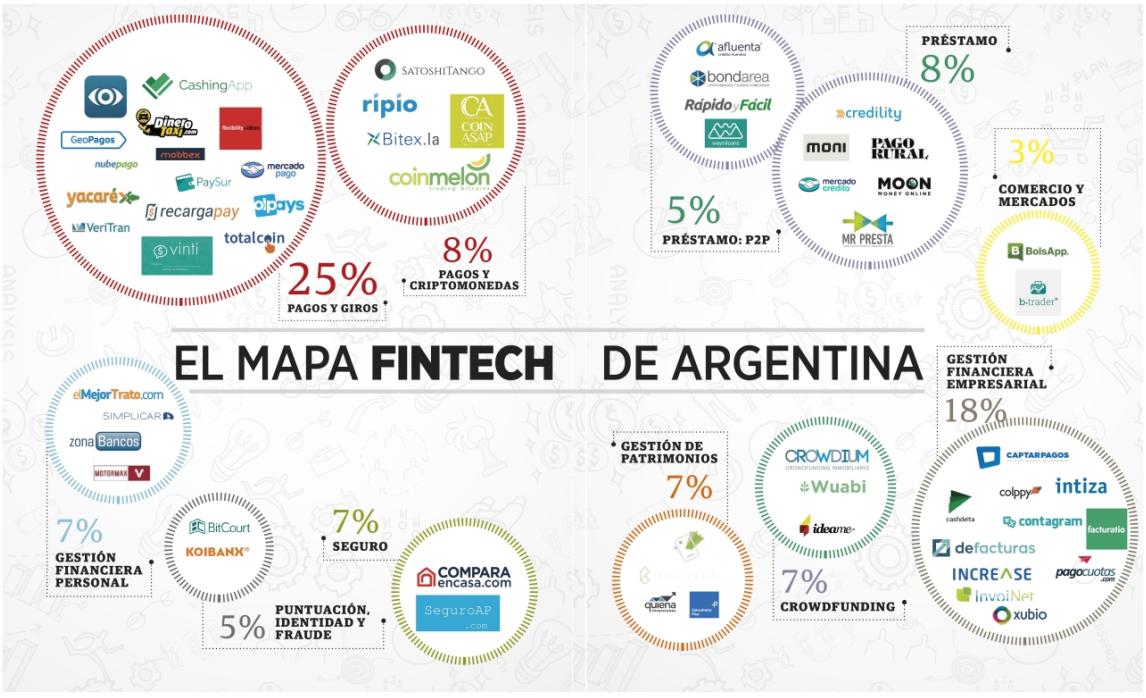
- Pagos y transferencias, en este grupo se encuentran fintech que ofrecen soluciones sobre medios de pagos electrónicos y transferencias de manera independiente de los bancos. Frecuentemente, los costes son inferiores a la banca.

- Financiación de particulares y empresas, las fintech ofrecen mediante sus plataformas financiación y préstamos. Son conocidas como crowdlending pueden ser entre particulares (P2P) o particulares y empresas (P2B).
- Participación en proyectos de inversión, se conoce como crowdequity y permite invertir en empresas con potencial de crecimiento a través de una plataforma tecnológica.
- Seguridad y control de fraude, son fintech que han desarrollado innovaciones en seguridad como la identificación y la gestión de la identidad digital, control de fraude aprovechando tecnologías que aporta el Big Data.
- Asesoramiento financiero y en inversiones, son plataformas que permiten a los usuarios en una sola vista a través de sus dispositivos móviles todas sus cuentas y operaciones financieras.
- Análisis de datos, fintech dedicadas a predecir el comportamiento de los usuarios, analizar patrones logrando una comercialización más eficaz de los productos bancarios.
- Criptomonedas, se caracterizan por ser un sistema de pago seguro que solo funciona en internet y al ser descentralizado no tiene propietarios siendo los propios usuarios de la moneda los que controlan su uso y su valor.
- Compliance, son fintech dedicadas al cumplimiento regulatorio de las actividades financieras, así como, la implementación de procedimientos para cumplir con las normas del regulador.

2.2.4. Fintech en la Argentina

En Argentina operan unas 72 startups relacionadas con esta actividad. El principal sector en el que incursionaron las fintech es pagos (25%) y enterprise financial management (18%) (BID y Finnovista, 2017). Otro dato interesante es el proporcionado por la consultora PwC Argentina, la mitad de las organizaciones del sector financiero ya están asociadas con un desarrollo tecnológico fintech (Guarino, 2018) (PwC Argentina, 2018). La infografía realizada por Forbes Argentina permite entender cuáles son los actores principales del ecosistema Fintech nacional (Valleboni, 2017):

Figura 1. Mapa Fintech de Argentina



Fuente: (Valleboni, 2017)

2.2.5. Marco Regulatorio de las fintech en la Argentina

El presidente del Banco Central de la República Argentina, Federico Sturzenegger, indicó "No hay ninguna regulación en marcha, ni la habrá. Para nosotros es una fuerte competencia que le hará bien al sistema financiero" (Burgueño, 2017). La decisión es definitiva aun cuando los bancos presionaban para que se les aplique algún tipo de regulación similar a la que tienen las entidades financieras para operar, permitiendo facilitar la competencia lo cual llevaría a mejores condiciones para el cliente. Por otra parte, los bancos cuentan con el permiso para que puedan abrir sus propias fintech o asociarse con las que ya existen en el mercado. Mientras los bancos y financieras realizan actividades de intermediación, las fintech ejecutarían operaciones de préstamos comerciales en contratos de uno que vende y otro que compra un servicio.

Las fintech podrán aplicar así sus propios criterios de scoring personal sobre los tomadores de los créditos que ofrecen. Un ejemplo es Mercado Crédito, la fintech de Mercado Libre, tiene su propio sistema, a partir de la fidelidad y comportamiento de los clientes que operan desde hace tiempo a través de la plataforma de la empresa. Si la persona es habitual operadora y tiene un nivel de calificación alto, la tasa será menor que otro que recién comienza a operar. El BCRA aclara, además, que no pondrá trabas a que los propios bancos quieran competir de igual a igual con las fintech en ese mismo mercado, ya que habilitó a las entidades financieras a abrir de cero una compañía de este tipo, o a comprar o asociarse con alguna que ya está operativa (Burgueño, 2017).

2.3. Almacenamiento de datos

2.3.1. Tipos de datos: estructurados, semiestructurados y no estructurados

Los datos estructurados tienen perfectamente definido la longitud, el formato y el tamaño de sus datos. Se almacenan en formato tabla hojas de cálculo o en bases de datos relacionales, generalmente conocidas como bases de datos SQL. Este tipo de datos son fáciles de introducir, analizar y almacenar.

Los datos no estructurados se caracterizan por no tener un formato específico, y se podría decir que no tienen una estructura rígida. Normalmente su estructura no es uniforme y se tiene habitualmente poco y nulo control sobre ellos. La información no está representada por datos elementales y su interpretación y manipulación es mucho más compleja. Ejemplos: audios, vídeos, fotografías, documentos, mensajes de correo electrónico, Twitter, etc.

Los datos semiestructurados son una mezcla de los dos anteriores no presenta una estructura perfectamente definida como los datos estructurados, pero si presentan una organización definida en sus metadatos donde describen los objetos y sus relaciones, suelen tener un formato que no es fácil su comprensión por el usuario y requiere habitualmente el uso de reglas complejas que ayuden a determinar cómo leer cada pieza de la información., como por ejemplo los formatos HTML, XML o JSON.

2.3.2. Base de datos relacional

Una base de datos relacional es una base de datos que se trata como un conjunto de tablas con columnas y filas relacionadas entre sí y que se manipulan de acuerdo con relaciones predefinidas entre ellas. Contiene un conjunto de objetos que se utilizan para almacenar y gestionar los datos, así como para acceder a los mismos. Las tablas, vistas, índices, funciones, activadores y paquetes son ejemplos de estos objetos (IBM, 2018).

Una entidad tiene columnas. Cada columna se identifica mediante un nombre y un tipo. Una entidad tiene registros o filas. Cada fila representa un conjunto exclusivo de información que normalmente representa los datos permanentes de un objeto. Cada entidad tiene una o varias claves principales. Las claves principales identifican de forma exclusiva a cada registro (Cabello, 2010).

En las bases de datos relacionales se tratan de evitar los datos redundantes mediante la normalización de las entidades. También se garantiza la disponibilidad, es decir que siempre esté preparada para su utilización. Otro aspecto relevante es la integridad de la base de datos, se refiere al hecho de que los datos sean correctos (Gómez, 2013). Una base de datos bien diseñada le brinda un completo acceso a la información deseada. Con un buen diseño dedicará menos tiempo a crear la base de datos y obtendrá resultados más exactos en menos tiempo.

2.3.3. Data Warehouse

Surgió en la década de los 90 ante la necesidad de desarrollar un sistema de almacenamiento de datos que proporcione una visión global de la organización garantizando la fluidez, el orden y el fácil manejo de la información y que, a la vez, supusiera un ahorro en tiempo y presupuesto para las empresas frente a los sistemas utilizados hasta el momento. Un Data Warehouse es un repositorio unificado en el que se almacenan los datos procedentes de fuentes internas y externas que puedan existir en una organización, quedando éstos integrados, depurados y ordenados en una única base de datos centralizada con las siguientes propiedades: estable, coherente, fiable y con información histórica. Esto quiere decir que la información que se puede obtener de un Data Warehouse es información fidedigna y confiable para la toma de decisión (Inmon, 2005).

Los Data Warehouse almacenan los datos durante el período de tiempo requerido para cumplir con las necesidades de consulta de una organización. Con este sistema, las compañías consiguen tener integrados en un único contenedor todos los datos de sus diferentes procesos de negocio, incluyendo datos históricos listos para soportar la constante necesidad de consultas estructuradas, ad hoc, reportes analíticos y soporte de decisiones. La construcción de los Data Warehouse están ganando cada vez mayor popularidad en las organizaciones, al considerar las ventajas que involucra el análisis de los datos históricos de forma multidimensional para apoyar el proceso de toma de decisiones (Conesa & Curto, 2015).

La información ingresada al Data Warehouse debe ser integrada y limpia, debiendo pasar por un proceso ETL (Extraer, Transformar y Cargar) el cual se hace referencia más adelante. A diferencia de las estructuras relacionales de una base de datos operacional donde la información es normalizada, la idea principal, de un Data Warehouse, es que la información sea presentada desnormalizada para optimizar las consultas, para ello, existen estructuras en las que se almacena la información: modelo estrella y modelo copo de nieve. Por otra parte, las bases de datos relacionales son las más comúnmente usadas para almacenar las estructuras de estos datos y sus grandes volúmenes, pero no es la única opción factible, también es posible considerar las bases de datos orientadas a columnas o incluso basadas en lógica asociativa (Caralt & Díaz, 2011).

Una vez creado el Data Warehouse es necesario realizar continuamente una limpieza, transformación e integración de los datos. Es un trabajo constante que garantiza el éxito de los datos en el diagnóstico y las soluciones de inteligencia de negocios que serán implementadas en la compañía (Ponniah, 2001). Por último, la integridad de los datos es un problema importante en la mayoría de las organizaciones, y el desarrollo de un Data Warehouse se utiliza con frecuencia como un vehículo para mejorar la calidad de los datos de manera significativa. De hecho, es mucho más sencillo controlar la calidad de los datos en un Data Warehouse centralizado que hacerlo en múltiples repositorios independientes (Humphries, Hawkins, & Dy, 1999).

2.3.3.1. Ventajas de usar Data Warehouse

- Se emplea para hacer el trabajo analítico, dejando las bases de datos transaccionales libres para centrarse en las transacciones.
- Al integrar datos de múltiples sistemas de origen, permite una visión central y completa de toda la organización.
- Almacena datos históricos permitiendo consultas en cualquier línea de tiempo.
- Mejora la calidad de los datos, al documentar y estructurar la información.
- Proporciona información clave para la toma de decisiones.
- Presenta la información de la organización de forma coherente.
- Reestructura los datos de manera que tienen sentido para los usuarios de negocios.
- Excelente rendimiento para consultas analíticas complejas.
- Permite conocer qué está pasando en el negocio

2.3.3.2. Desventajas de usar Data Warehouse

- Su elevado costo de implementación y mantención.
- Aplicable para grandes empresas ya que el esfuerzo de crear un Data Warehouse sería muy superior a los beneficios obtenidos.
- No es muy útil para la toma de decisiones en tiempo real debido al largo tiempo de procesamiento que puede requerir.
- Datos que no se almacenaron en el Data Warehouse que podría aportar con el análisis.
- Mayores demandas del usuario final.

2.3.3.3. Características del Data Warehouse

- Orientado a temas concretos: los datos se organizan alrededor de temas específicos que son de interés para el negocio en lugar de hacerlo alrededor de aplicaciones tales como la gestión de inventario o el procesamiento de pedidos. Por ejemplo, un Data Warehouse se puede utilizar para analizar las ventas.
- No volátil: una vez que los datos están en almacenados, ya no van a cambiar y no se actualizan en tiempo real. Hay cargas incrementales de información desde bases de datos transaccionales y otras fuentes para refrescarlos con datos nuevos, pero sin variar los antiguos.
- Integrado: un Data Warehouse integra datos recolectados de diferentes sistemas operacionales de la organización y también de fuentes externas. Por ejemplo, la fuente A y fuente B pueden tener diferentes maneras de identificar un producto, pero en un almacén de datos, sólo habrá un único modo de identificar un producto. Tiene que ver con la forma en que los datos se extraen y transforman de las diversas fuentes de datos.
- Variante en el tiempo: en un Data Warehouse se mantienen los datos históricos ordenados por diferentes periodos de tiempo. De esta forma, se pueden recuperar datos de 3 meses, 6 meses, 12 meses, o incluso datos más antiguos. Esto contrasta con un sistema basado en transacciones, donde a menudo sólo se mantienen los datos más recientes. Por ejemplo, de un sistema transaccional se puede obtener la dirección más reciente de un cliente, mientras que en un Data Warehouse podemos encontrar todas las direcciones asociadas históricamente a un cliente.

2.3.3.4. Proceso ETL

Es un proceso responsable de la extracción de datos de los sistemas de origen y de colocarlo en un Data Warehouse. Este proceso implica las siguientes tareas (Kimball & Caserta, 2004):

- Extracción: Se trata de obtener los datos deseados a partir de las distintas fuentes de origen, tanto internas como externas.
- Transformación: Es el filtrado, limpieza, depuración, homogeneización y agrupación de los datos de las diferentes fuentes.
- Carga: Es el proceso de almacenar los datos en el Date Warehouse.

2.3.3.5. OLTP y OLAP

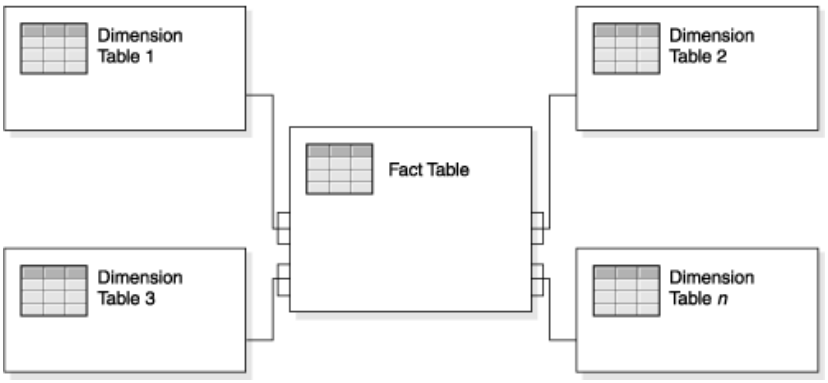
Los sistemas OLTP son bases de datos orientadas al procesamiento de transacciones. Se caracterizan por un gran número de transacciones cortas en línea (INSERT, UPDATE, DELETE). El énfasis principal de los sistemas OLTP se basa en el procesamiento de consultas muy rápido, manteniendo la integridad de los datos en entornos de acceso múltiple y una efectividad medida por el número de transacciones por segundo. En la base de datos OLTP hay datos detallados y actuales, y el esquema utilizado para almacenar bases de datos transaccionales es el modelo de entidad, generalmente 3NF (Caralt & Díaz, 2011).

Los sistemas OLAP son bases de datos orientadas al procesamiento analítico. Se caracterizan por un volumen relativamente bajo de transacciones. Las consultas suelen ser muy complejas e implican agregaciones. Para los sistemas OLAP, un tiempo de respuesta es una medida de efectividad. Las aplicaciones OLAP son utilizadas por las técnicas de Data Mining. En la base de datos OLAP hay datos históricos agregados, almacenados en esquemas multidimensionales (esquema en estrella) (Caralt & Díaz, 2011).

2.3.3.6. Esquemas: Estrella y Copo de Nieve

Un esquema en estrella es un tipo de esquema de base de datos relacional que se compone de una única tabla de hechos central que está rodeada por tablas de dimensiones. Un esquema en estrella puede tener cualquier cantidad de tablas de dimensiones. Las ramas al final de los enlaces que conectan las tablas indican una relación de varios a uno entre la tabla de hechos y cada tabla de dimensiones (IBM, 2018).

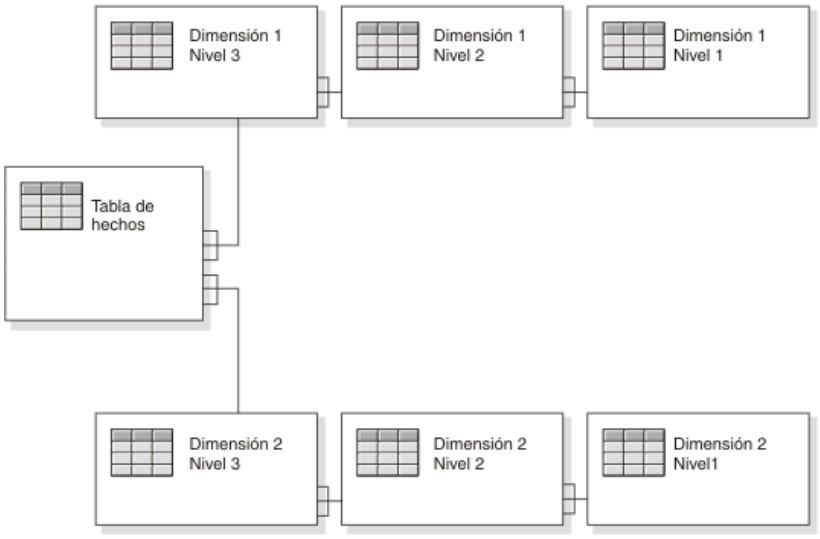
Figura 2. Esquema Estrella



Fuente: (IBM, 2018)

El esquema de copo de nieve consta de una tabla de hechos que está conectada a muchas tablas de dimensiones, que pueden estar conectadas a otras tablas de dimensiones a través de una relación de muchos a uno. Las tablas de un esquema de copo de nieve generalmente se normalizan en el tercer formulario de normalización. Cada tabla de dimensiones representa exactamente un nivel en una jerarquía (IBM, 2018).

Figura 3. Esquema de Copo de Nieve



Fuente: (IBM, 2018).

El esquema de estrella es el más sencillo además de ser quizás el más utilizado ya que su estructura es simple, sin embargo, para su uso mucha información debe estar contenida en cada una de las tablas de dimensión generando mayor uso de almacenamiento. Por otro lado, al utilizar el modelo copo de nieve existen más relaciones en el modelo y este se volvería poco eficiente para buscar la información además de volverse complejo de mantener.

2.3.4. Data Mart

El trabajo de construir un Data Warehouse corporativo puede generar inflexibilidades, o ser complejo, costoso y requerir plazos de tiempo que las organizaciones no están dispuestos a aceptar. En parte, estas razones originaron la aparición de los Data Mart. Un Data Mart es una base de datos departamental, especializada en el análisis, almacenamiento e integración de los datos y está destinado a satisfacer las necesidades de un segmento de negocio en particular o por un grupo de trabajo multidisciplinar con objetivos comunes. Al igual que en un Data Warehouse, los datos están estructurados en modelos de estrella o copo de nieve, y un Data Mart puede ser dependiente o independiente de un Data Warehouse. Por ejemplo, un posible uso de un Data Mart sería para la minería de datos o para la información de marketing (Caralt & Díaz, 2011).

Generalmente, los Data Mart son más pequeños que los Data Warehouse. Funcionan como una aplicación del Data Warehouse o una alternativa para empresas medianas que no pueden afrontar los costos de implementar un sistema tan amplio de almacenamiento de data. Esta estrategia es particularmente apropiada cuando el Data Warehouse central crece muy rápidamente y los distintos departamentos requieren sólo una pequeña porción de los datos contenidos en él. La creación de estos Data Mart requiere algo más que una simple réplica de los datos: se necesitarán tanto la segmentación como algunos métodos adicionales de consolidación. Tienen menos cantidad de información, menos modelos de negocio y son utilizados por un número inferior de usuarios. Los Data Mart pueden ser independientes o dependientes. Los primeros son alimentados directamente de los orígenes

de información, mientras que los segundos se alimentan desde el Data Warehouse corporativo. Con un Data Mart es posible acceder a información clave más rápidamente y ayuda a evitar que los departamentos dentro de la organización interfieran con los datos de los demás (Cano, 2007).

Los Data Mart se crean solo una vez, al comienzo del proceso analítico, pero se actualizan de forma cíclica y automática para contener toda la información relevante relacionada, por ejemplo, con los clientes, productos, transacciones, etc. en un período de tiempo determinado.

2.3.4.1. Ventajas de usar Data Mart

- Podemos agilizar las consultas evitando recorrer un gran volumen de datos que la verdad no es necesario recorrer.
- Menor riesgo de errores en el análisis, lo que significa que los resultados son más creíbles.
- Disponibilidad de la información analítica más actualizada, gracias a la actualización cíclica de Data Mart.
- Los costos en la construcción de un Data Mart son menores en comparación a los de un Data Warehouse, que puede tardar muchos meses en construirse.

2.3.5. Diferencias entre Data Warehouse y Data Mart

La principal diferencia entre un Data Warehouse y un Data Mart es su alcance y área de uso. Así, mientras un Data Warehouse contiene todos los datos de una organización, un Data Mart solamente recoge un subconjunto de éstos, centrándose en un área específica dentro del negocio. A continuación, se mencionan las diferencias:

Tabla 1. Diferencias entre Data Warehouse y Data Mart

Data Warehouse	**Data Mart**
Construido para satisfacer las necesidades de información de toda la organización.	Construido para satisfacer las necesidades de información a un área específica de la organización.
Tiene múltiples áreas temáticas.	A menudo tiene solo un área temática específica.
Integra y administra todas las fuentes de datos.	Integra datos de una determinada temática o conjunto de sistemas fuente.
Guarda información muy detallada (desagregada).	Contiene información resumida o totalizada (agregada).
Los datos están levemente desnormalizados.	Los datos están altamente desnormalizados.
Pertenece a toda la organización.	Pertenece al área de negocios al cual está orientado.
Esquema constelación de hechos.	Esquema estrella y copo de nieve.
Tiempo de implementación pueden ser meses a años.	Tiempo de implementación pueden ser meses.
Difícil de construir.	Fácil de construir.

Fuente: Elaboración propia.

2.3.6. Bases de datos NoSQL

NoSQL es un término que describe a las bases de datos no relacionales de alto desempeño. También conocidas como “No solo SQL” (Not Only SQL), son famosas por la facilidad de desarrollo, ser no relacionales, la escalabilidad horizontal, alta disponibilidad y la tolerancia a fallos (Amazon Web Services, 2017). En su mayoría se originaron a partir de la aparición de la web 2.0 en donde los usuarios se transformaron de ser simples consumidores a productores de contenido. Hasta este momento existen unos 225 sistemas de gestión bases de datos NoSQL disponibles en el mercado. Una diferencia clave entre las bases de datos de NoSQL y las bases de datos relacionales tradicionales, es el hecho de que NoSQL es una forma de almacenamiento no estructurado. Esto significa que NoSQL no tiene una estructura de tabla fija como las que se encuentran en las bases de datos relacionales. Además, no usan SQL (Structured Query Language) como lenguaje principal de consulta (NoSQL, 2018).

Aunque son conocidas desde la década de los 60 del pasado siglo, las principales compañías de internet como Amazon, Google, Twitter y Facebook son las que han propiciado el uso de las bases de datos NoSQL. En un principio, para solucionar estos problemas de accesibilidad, las compañías optaron por utilizar un mayor número de máquinas, pero pronto se dieron cuenta de que esto no solucionaba el problema, además de ser una solución muy cara (Acens, 2014). Motivados por el rápido crecimiento de la web, donde se requería dar respuesta a la necesidad de proporcionar información procesada a partir de grandes volúmenes de datos con unas estructuras horizontales, más o menos,

similares y con aplicaciones web que debían dar respuesta a las peticiones de un número elevado e indeterminado de usuarios en el menor tiempo posible. Estas compañías tuvieron que priorizar el rendimiento y sus necesidades de tiempo real sobre la consistencia de los datos, este último aspecto en donde las bases de datos tradicionales dedicaban una gran cantidad de tiempo de proceso (Oracle, 2016).

2.3.6.1. Teorema de CAP

Las bases de datos NoSQL, no soportan totalmente ACID, esto lo explica el teorema de CAP, también llamado formalmente Teorema de Brewer. Este fue referenciado por Eric Brewer en el año 2000 en un simposio sobre los principios de la computación distribuida, dice que en sistemas de datos distribuidos es imposible garantizar estas tres características simultáneamente: consistencia, disponibilidad y tolerancia a particiones (Estada & Ruiz, 2016). A continuación, vamos ver qué son estas características:

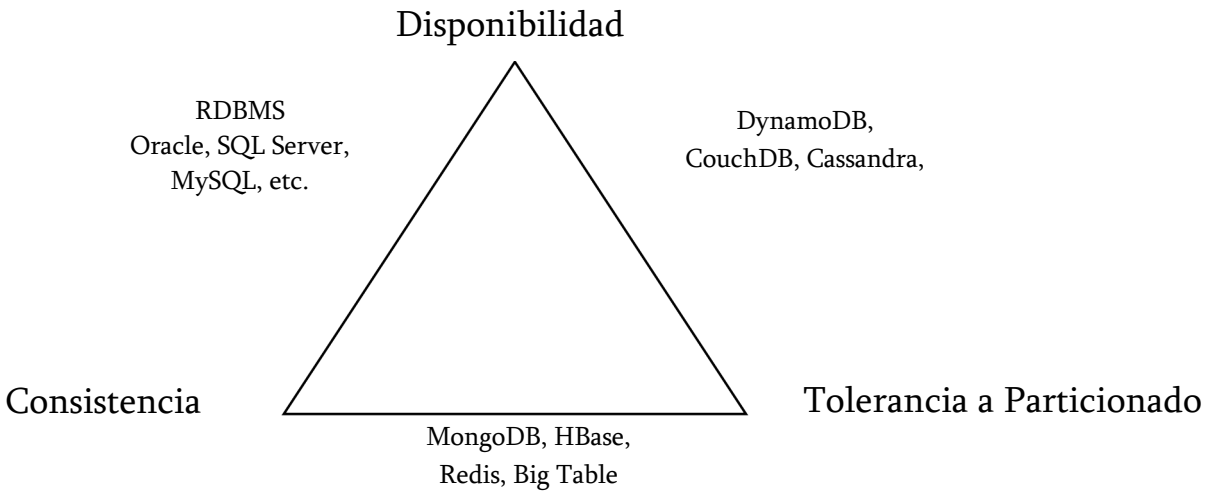
- Consistencia (Consistency): Implica que la información permanece coherente y consistente después de cualquier operación sobre los datos, de modo que cualquier usuario que acceda a los mismos verá la misma información. Si insertamos datos (todos los nodos deben insertar los mismos datos), si actualizamos datos (todos los nodos deben aplicar la misma actualización a todos los datos) y si consultamos datos (todos los nodos deben devolver los mismos datos).

- Disponibilidad (Availability): Independientemente si uno de los nodos se ha caído o a dejado de emitir respuestas, el sistema debe seguir en funcionamiento y aceptar peticiones tanto de escritura como de lectura. Si se pierde comunicación con un nodo, el sistema automáticamente debe tener la capacidad de seguir operando mientras éste se restablece y una vez que lo hace, se debe sincronizar con los demás.
- Tolerancia a particionado (Partition Tolerance): El sistema debe estar disponible así existan problemas de comunicación entre los nodos, cortes de red que dificulten su comunicación o cualquier otro aspecto que genere su particionamiento.

El teorema sólo nos puede garantizar las siguientes combinaciones:

- CP (Consistencia y Particionado): El sistema aplicará los cambios de forma consistente, aunque existan problemas de conexión entre los nodos (particiones de red) y no se asegura que haya disponibilidad.
- AP (Disponibilidad y Particionado): El sistema siempre estará disponible a las peticiones, aunque existan problemas de conexión entre los nodos (particiones de red), y en consecuencia por la pérdida de comunicación existirá inconsistencia en los datos temporalmente porque no todos los nodos serán iguales.
- CA (Consistente y Disponibilidad): El sistema siempre estará disponible respondiendo peticiones con información consistente. Por lo cual el sistema no soporta pérdida de comunicación entre los nodos (Particiones de red).

Figura 4. Teorema de CAP



Fuente: Elaboración propia.

Esta tabla ofrece una comparativa entre las funcionalidades más representativas de las bases de datos NoSQL y las bases de datos relacionales:

Tabla 2. Comparación entre las características de NoSQL y base de datos relacionales

Características	**Bases de datos NoSQL**	**Bases de datos relacionales**
Rendimiento	Alto	Bajo
Confiabilidad	Pobre	Buena
Disponibilidad	Buena	Buena
Consistencia	Pobre	Buena
Almacenamiento	Optimizado para gran cantidad de datos	De mediano a grandes cantidades de datos
Escalabilidad	Alto	Alto (pero más caro)

Fuente: Elaboración propia.

2.3.6.2. Tipos de bases de datos NoSQL

En el mundo de las bases de datos NoSQL nos encontramos con distintos modelos o tipos, que se desempeñan mejor en algunos ambientes específicos. Existen 4 tipos de bases de datos NoSQL que describiremos a continuación:

1. Bases de datos en columnas: Una base de datos columnar está optimizada para leer y escribir columnas de datos en lugar de filas. El objetivo de una base de datos columnar es escribir y leer datos de manera eficiente, desde y hacia el almacenamiento en disco duro, para acelerar el tiempo que se tarda en devolver el resultado de una consulta. Uno de los principales beneficios de una base de datos columnar es que los datos pueden ser altamente comprimidos. La compresión permite que las operaciones de agregación se realicen muy rápidamente (Amazon Web Services, 2017).
2. Bases de datos de documentos: Una base de datos documental está diseñada para almacenar datos semiestructurados como documentos, en algún formato estándar como puede ser JSON, XML o BSON y donde se utiliza una clave única para cada registro. Permite, además de realizar búsquedas por clave–valor, realizar consultas más avanzadas sobre el contenido del documento. Son las bases de datos NoSQL más versátiles. Se pueden utilizar en gran cantidad de proyectos, incluyendo muchos que tradicionalmente funcionarían sobre bases de datos relacionales (Amazon Web Services, 2017).

3. Bases de datos de grafos: Este tipo de bases de datos utiliza la topología de un grafo con nodos como vértices y relaciones como aristas y propiedades, utilizada para almacenar y representar datos conectados sin necesidad de utilizar un índice (que es el método tradicional de simular una relación en una base de datos relacional) (Robinson, Webber, & Eifrem, 2015). Son muy útiles para guardar información en modelos con muchas relaciones, como redes y conexiones sociales (Neo4j, 2018). Este tipo de bases de datos ofrece una navegación más eficiente entre relaciones que en un modelo relacional (Acens, 2014).
4. Bases de datos de clave-valor: Son el modelo de base de datos NoSQL más popular, además de ser la más sencilla en cuanto a funcionalidad. Optimizada para grandes cargas de trabajo de aplicaciones de lectura (como redes sociales, juegos, uso compartido de archivos multimedia y portales de preguntas y respuestas) o para cargas de trabajo informáticas intensivas (como un motor de recomendaciones) (Amazon Web Services, 2017). En este tipo de sistema, cada elemento está identificado por una clave única, lo que permite la recuperación de la información de forma muy rápida, información que suele almacenarse como un objeto binario. Para conseguir baja latencia, almacena los datos críticos en memoria siendo muy eficiente tanto para lecturas como para las escrituras. Son útiles también para almacenar información estadística y en definitiva para cualquier tipo de problemática que se puede solventar a través del concepto de diccionario orientado a almacenar grandes volúmenes de información. (Oracle, 2016).

2.3.7. Data Lake

Las herramientas analíticas y de almacenamiento de datos tradicionales ya no pueden brindar la agilidad y flexibilidad requeridas para brindar información empresarial relevante. En la era digital y tecnológica que vivimos no solo implica el crecimiento desproporcionado de la información, sino más bien qué hacemos con ella y cómo la gestionamos y organizamos para no desperdiciarla. Datos que hoy pueden carecer de utilidad para tu empresa o estrategia de negocio pueden tenerlos en el futuro, por lo que si no se usa o se gestiona bien este conocimiento estás perdiendo valor. Es por eso que muchas organizaciones están cambiando a una arquitectura de lago de datos (Amazon Web Services, 2017).

Un lago de datos (Data Lake) es un repositorio en el que se almacenan grandes cantidades de datos en su formato original y sin ser sometidos a ninguna transformación previa, incluyendo datos estructurados, semiestructurados y no estructurados, para ser analizados posteriormente. En vista que los datos se pueden almacenar como están, no es necesario convertirlos a un esquema predefinido. De hecho, las organizaciones vierten los datos en una ubicación central y los recuperan cuando estos se necesitan. Únicamente en ese momento se procede a categorizarlos, procesarlos, ordenarlos y a diseñar una estructura que haga más fácil su posterior análisis por diversos grupos dentro de una organización (Tomcy & Pankaj, 2017).

El término Data Lake generalmente se asocia con el almacenamiento de objetos orientado a Hadoop en el que los datos de una organización se cargan en la plataforma Hadoop y luego se aplican las herramientas analíticas y minería de datos a los datos donde reside el clúster Hadoop. Sin embargo, los Data Lake también se pueden usar de manera efectiva sin incorporar Hadoop en función de las necesidades y los objetivos de la organización (IBM, 2018). A cada elemento del Data Lake se le asigna un identificador único y se etiqueta con un conjunto de etiquetas de metadatos extendidas. Cuando surge una cuestión comercial, resulta posible efectuar una consulta al lago de datos en busca de datos relevantes y, al mismo tiempo, cabe la posibilidad de analizar dicho conjunto de datos más pequeño para ayudar a responder a la consulta (Pasupuleti & Purra, 2015).

2.3.7.1. Características de un Data Lake

- No desecha ningún dato. Se almacena y conserva todos los datos e información que le llegan sin tener en cuenta su estructura y su fuente original. Se conservan todos los datos, aunque no sean utilizados a corto plazo porque quizás sean utilizados algún día.
- Flexible y rápido en los cambios. Permite adaptarse a cualquier análisis, en cualquier momento y con más detalles. El usuario puede usar los datos almacenados como mejor le parezca, cuando quiera y las veces que necesite.
- Para todo tipo de perfiles. Es útil para todo tipo de usuarios, tanto para los científicos de datos para un análisis más profundo y para otros usuarios que usan vistas más estructuradas.

2.3.7.2. Ventajas de un Data Lake

- El lago de datos permite que los usuarios comerciales tengan acceso inmediato a toda la información existente.
- Los datos situados en el lago no se limitan a los datos relacionales o transaccionales.
- Con un lago de datos, el usuario nunca necesita desplazar los datos.
- El lago de datos otorga facultades a los usuarios comerciales y los libera de las ataduras que supone la dominación de TI.
- El lago de datos acelera la entrega permitiendo que las unidades de negocio alimenten las aplicaciones rápidamente.

2.3.7.3. Desventajas de un Data Lake

- Conocimientos técnicos avanzados en procesamiento de datos.
- Gobernanza de datos
- Gestionar el caos.
- Problemas de privacidad.
- Complejidad de los datos heredados.
- Gestión del ciclo de vida de los metadatos.
- Islas de datos aislados.
- Problemas de integración.

2.3.8. Diferencias entre un Data Lake y un Data Warehouse

Tabla 3. Comparación entre las características de NoSQL y base de datos relacionales

Data Lake	**Data Warehouse**
Almacena datos en su formato original y sin ser sometidos a ninguna transformación.	Almacena datos que han sido sometidos a procesos ETL.
Cualquier tipo de datos: estructurado, semiestructurado y no estructurado.	Solamente datos estructurados.
Almacenan información de cualquier fuente de datos.	Almacenan información extraída de los sistemas transaccionales.
Se basa en tecnologías que permiten almacenar datos sin procesar.	Se basa en la tecnología de base de datos relacional.
Costo de almacenamiento es relativamente bajo, se puede usar hardware de bajo costo y software de código abierto como Hadoop.	Costo de almacenamiento superior.
Carece de estructura, ofreciendo la posibilidad de configurar y reconfigurar los modelos, consultas y aplicaciones con facilidad.	Es altamente estructurado, cambiar la estructura puede implicar modificar gran parte de los procesos de negocios vinculados.
Más idóneo para los perfiles más técnicos que buscan explotar al máximo las capacidades analíticas.	Orientado a facilitar la interacción del usuario de negocio.

Fuente: Elaboración propia.

Toda información es importante para las organizaciones, aunque esta no sea usada en el corto plazo. Un Data Lake no sustituye a un Data Warehouse. Ambos están optimizados para diferentes propósitos, y el objetivo es utilizar cada uno para lo que fueron diseñados no malgastando recursos en algo que no tendrá uso en la organización.

2.4. Big Data

Es la gestión y análisis de enormes volúmenes de datos que no pueden ser tratados de manera convencional, ya que superan los límites y capacidades de las herramientas de software habitualmente utilizadas para la captura, gestión y procesamiento de datos. Este concepto engloba infraestructuras, tecnologías y servicios que han sido creados para dar solución al procesamiento de enormes conjuntos de datos estructurados, no estructurados o semiestructurados. Cuando hablamos de Big Data estamos hablando de grandes cantidades de datos, por ejemplo, petabytes (PB), exabytes (EB), zettabytes (ZB), yottabytes (YB).

El aumento de las capacidades de procesamiento y ancho de banda junto con el abaratamiento del almacenamiento de la información han hecho posible de guardar y procesar las grandes cantidades de datos generadas en la actualidad y que años atrás era casi imposible por su alto costo y complejidad. Es precisamente al almacenamiento y procesamiento de los datos para obtener información que aporte valor para la toma de decisiones a lo que se refiere el concepto de Big Data.

2.4.1. Dimensiones de Big Data

A continuación, se define la tecnología Big Data con sus 4 dimensiones conocidas como las 4V:

Tabla 4. Dimensiones de Big Data

Dimensiones	Descripción
Velocidad	Se refiere al ritmo en que los datos fluyen tanto para el almacenamiento en tiempo real como la capacidad de análisis de dichos datos reduciendo los largos tiempos de procesamiento de las herramientas tradicionales.
Volumen	Ser capaz de gestionar un gran volumen de datos que se generan de diversas fuentes y que no caben en un disco duro normal. Debiendo formarse un clúster de ordenadores.
Variedad	Tener la capacidad de combinar una gran variedad de información digital en los diferentes formatos en las que se puedan presentar ya sean en formato video, audio, texto, etc.
Veracidad	Ser capaz de tratar y analizar inteligentemente este vasto volumen de datos con la finalidad de obtener una información verídica y útil que nos permita mejorar nuestra toma de decisiones.

Fuente: Elaboración propia.

2.4.2. ¿De dónde proviene toda esa información?

No hace mucho tiempo, la recopilación de datos implicaba realizar observaciones escritas a mano para estudiar cómo evoluciona el problema que se estaba investigando. Ahora, los datos son generados constantemente y cada vez más en cantidades astronómicas. Si bien mantener esta cantidad de datos fue alguna vez costoso y difícil, las capacidades de almacenamiento crecieron y los costos cayeron. Sumado con la tecnología ha desempeñado un papel importante en la adopción de Big Data, haciendo posible el almacenamiento de grandes cantidades de información.

Todo esto, mediante un marco de trabajo denominado Hadoop, el cual se hace referencia más adelante, donde es posible recolectar datos de toda índole, por ejemplo, mensajes en redes sociales, internet, apps, teléfono móvil, archivos de audio, sensores, imágenes digitales, datos de formularios, tarjetas de crédito, emails, datos de encuestas, logs, micrófonos, cámaras, escáneres médicos, bases de datos relacionales, internet de las cosas, archivos XML, ERP, CRM, etc. A esto habría que añadirle las grandes cantidades de datos transaccionales que almacenan las compañías y que contienen información de sus clientes, proveedores, operaciones, etc. Si bien los datos se encuentran en todas partes, éstos tienen un potencial enorme que debe ser aprovechado con herramientas de Big Data (BSA, 2015).

2.4.3. Big Data en el Sector Bancario

La revolución del Big Data no ha llegado en su totalidad al sector financiero, pero está en un proceso de transformación. En su mayoría se han dado cuenta del inmenso valor que representa Big Data para entender mejor al cliente permitiéndoles tomar mejores decisiones. Por su parte, el Big Data permite analizar grandes volúmenes de datos de múltiples fuentes. La banca maneja información interna como transacciones mediante pago por tarjeta, su propio sistema de información y analítica web, y otra externa como tipologías de familias, ingresos, gastos, actividad inmobiliaria, redes sociales, matriculaciones de vehículos, etc., que se pueden vincular geográficamente a oficinas, sucursales y puntos de venta. Permitiéndoles a los Bancos comprender a los clientes de una mejor manera y ofrecerles asistencia diferenciada e identificada para satisfacer sus necesidades (Groenfeldt, 2013).

2.4.4. Hadoop

Distribuido por la Apache Software Foundation. Hadoop es una plataforma de software de código abierto para el almacenamiento y procesamiento distribuido de conjuntos de datos muy grandes en clústeres de computadoras creados a partir de hardware básico (White, 2012). Hadoop se inspiró en los documentos de Google para MapReduce y Google File System (GFS). El origen de Hadoop fue en 2004, dos ingenieros escribieron un artículo titulado "MapReduce: Simplified Data Processing on Large Clusters". El artículo hace referencia a un nuevo modelo de programación, MapReduce, que permite simplificar el procesamiento de grandes volúmenes de datos. Este nuevo modelo nació de la necesidad que tenía Google para procesar grandes volúmenes de datos que manejaban (documentos, referencias web, páginas, etc.) además de abaratar los costos de almacenamiento y lograr una mejor escalabilidad y rendimiento (Jain, 2017).

La idea de este nuevo modelo es montar un esquema en paralelo de computación que permita distribuir el trabajo (procesamiento de datos) entre diferentes máquinas (nodos dentro de una red) para que se pueda reducir el tiempo total de procesamiento. Básicamente, la idea de "divide y vencerás", al dividir el trabajo en tareas más pequeñas y ejecutarlas de manera distribuida, lo que por su contraparte sería un único y gran trabajo. Los artículos de Google le sirvieron de inspiración a Cutting para crear Apache Hadoop. En 2006 se marchó a Yahoo! y allí completó el desarrollo de la plataforma que fue lanzada en 2008. El propio buscador utilizaría la tecnología para su negocio, así como otras grandes compañías de Internet, como Facebook, Twitter o eBay (Rayón, 2016).

Hadoop se compone de los siguientes módulos donde cada uno lleva a cabo una tarea esencial para un sistema informático diseñado para grandes análisis de datos (Apache Software Foundation, 2018):

- HDFS: es un sistema de ficheros distribuido, altamente escalable y tolerante a fallos escrito en Java y diseñado para utilizarse en hardware básico. Proporciona un alto rendimiento en el acceso a datos y trabaja en estrecha colaboración con una amplia variedad de aplicaciones concurrentes de acceso a datos, coordinadas por YARN. Los archivos en HDFS se almacenan en bloques de tamaño de bloque (128 MB). Cada bloque se replica en diferentes nodos (Hortonworks, 2018).
- YARN: es el centro de la arquitectura de Hadoop, administra los recursos del clúster permitiendo procesar datos de múltiples formas al mismo tiempo (datos Bach y tiempo real).
- MapReduce: el procesamiento está separado en dos operaciones: Map (mapea) transforma un conjunto de datos de partida en pares (clave, valor) a otro conjunto de datos intermedios también en pares (clave, valor). Un formato, que hará más eficiente su procesamiento y sobre todo, más fácil su "reconstrucción" futura y Reduce (reduce) recibe los valores intermedios procesados en formato de pares (clave, valor) para agruparlos y producir el resultado final (Esteso, 2018).
- Common: proporciona el acceso a los sistemas de archivos soportados por Hadoop y contiene el código necesario para poder ejecutar el framework. Además, se pone a disposición del usuario el código fuente y la documentación necesaria para aprender a utilizar la herramienta.

Entre sus características más importantes se encuentran las siguientes:

Tabla 5. Características de Hadoop

Características	Descripción
Distribuido	Se ejecuta en un conjunto de ordenadores (clúster) conectados entre sí mediante una red interconexión que funciona de forma coordinada.
Escalable	Puede crecer simplemente añadiendo nuevos nodos y no es necesario hacer ajustes que modifiquen la estructura inicial.
Tolerante a fallos	No hay pérdida de datos gracias a la replicación. Se asignan tareas a otro nodo para continuar operando.
Open Source	Su código fuente está disponible de forma abierta. Es compatible en todas las plataformas ya que está basado en Java.
Bajo costo	Diseñado para usarse en commodity hardware, lo cual supone una gran flexibilidad y un importante ahorro.

Fuente: Elaboración propia.

2.4.5. Ecosistema Hadoop

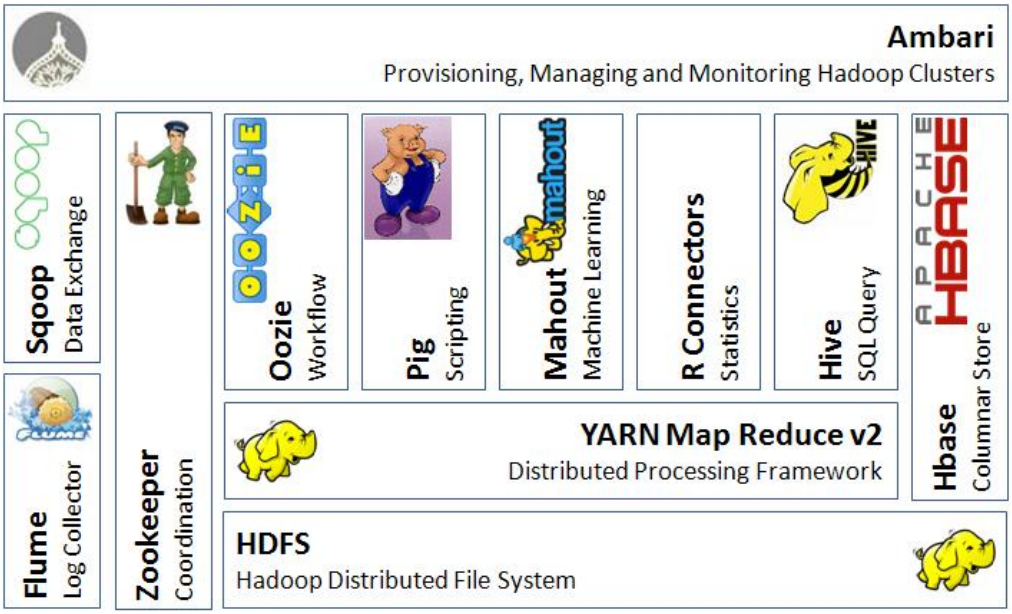
Hadoop está suplementado por un ecosistema de proyectos Apache que incrementan el valor de Hadoop y mejora sus posibilidades siendo adaptable a necesidades particulares. A continuación, se mencionan los más relevantes (Apache Software Foundation, 2018):

- Ambari: proporciona una interfaz de usuario web intuitiva para el aprovisionamiento, administración, monitoreo y seguridad de los clústeres de Apache Hadoop. Simplifica la administración de Hadoop al proporcionar una plataforma segura y consistente para el control operativo.

- Avro: es un sistema de serialización de datos. La serialización se usa para procesarlos y almacenar estos datos, de forma que el rendimiento en tiempo sea efectivo, además de facilitar el intercambio de datos entre diferentes programas.
- Cassandra: Es de las denominadas bases de datos NoSQL, orientada a almacenamiento en columnas. Su objetivo principal es la escalabilidad lineal y la disponibilidad.
- Flume: recopila, agrega y mueve grandes cantidades de datos log de diferentes orígenes a un repositorio central.
- HBase: es una base de datos orientada a columnas, proporciona acceso de lectura / escritura en tiempo real a grandes conjuntos de datos. No admite SQL.
- Hive: es considerado el Data Warehouse para Hadoop, facilita el uso de la agregación de los datos, ad-hoc queries, y el análisis de grandes datasets almacenados. Usa un lenguaje parecido al SQL, llamado HiveQL.
- Kafka: usado para el análisis en tiempo real y la reproducción de datos de transmisión. Funciona como un servicio de mensajería (topics), donde productores publican mensajes en las listas y los consumidores se pueden subscribir.
- Mahout: es una biblioteca escalable de algoritmos de aprendizaje automático y minería de datos gratuitos.
- Pig: permite a los usuarios escribir scripts usando un lenguaje de escritura simple llamado Pig Latin, en vez de escribir una aplicación de MapReduce.
- Spark: es un motor de análisis unificado y veloz para grandes volúmenes de datos y aprendizaje automático. Tiene API fáciles de usar para operar en grandes conjuntos de datos y puede ser mucho más rápido que MapReduce.

- Sqoop: permite conectarnos a una base de datos relacional para transferir de forma eficiente los datos a Hadoop y viceversa. Permite también importar los datos SQL a Hive.
- ZooKeeper: es un servicio centralizado de coordinación para diversas tareas como mantenimiento de configuración, registro de nombres y sincronización para sistemas distribuidos.

Figura 5. Ecosistema de Hadoop



Fuente: (Taie, 2015)

2.4.6. Distribuciones de Hadoop

Existen diferentes herramientas o distribuciones que nos permiten administrar nuestro clúster de manera sencilla. Nos vamos a centrar en las tres principales: Cloudera, Hortonworks y MapR.

2.4.6.1. Cloudera

Cloudera fue fundada en 2008 por tres ingenieros de Google (Christophe Bisciglia), Yahoo! (Amr Awadallah) y Facebook (Jeff Hammerbacher) más un antiguo ejecutivo de Oracle (Mike Olson). Cloudera tiene más de 1,600 empleados y oficinas en 24 países de todo el mundo, con sede en Palo Alto, California. Su software está basado en Apache Hadoop y ofrecen soporte, servicios y formación para grandes clientes. Cloudera ha sido la primera en lanzar componentes propietarios de valor añadido sobre Hadoop, como Impala, que junto con otras herramientas como Cloudera Manager y Cloudera Navigator son usadas por un gran número de clientes. (Cloudera, 2018).

Cloudera proporciona una plataforma escalable, flexible e integrada que facilita la administración de volúmenes y variedades de datos en rápido crecimiento. Los productos y soluciones de Cloudera permiten implementar y administrar Apache Hadoop y proyectos relacionados con el ecosistema para manipular y analizar datos de forma segura. Su naturaleza es open source aunque también tiene software propietario y cobra por las licencias y el soporte técnico.

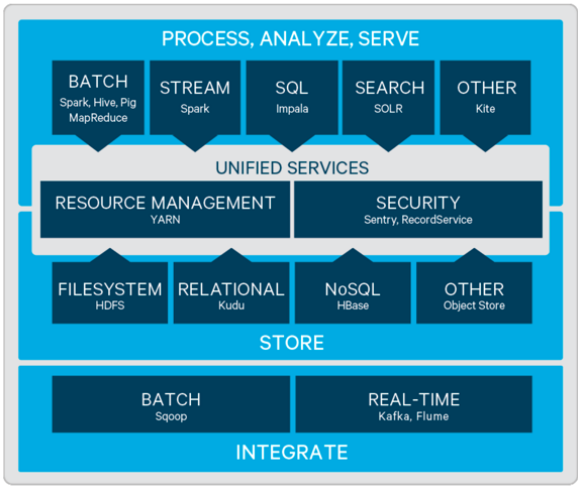
Cloudera ofrece un plan de estudios dinámico que se actualiza para mantener el ritmo de la innovación donde hay diversas formas de aprender para mejorar las habilidades en tecnología de big data a través de Cloudera University. Permitiendo convertirse en un profesional certificado en Big Data a través del programa Cloudera Certified Professional, el cual ofrece la credencial Big Data más rigurosa y reconocida. Además, existen diversas rutas de aprendizaje de acuerdo a lo que deseemos especializarnos por ejemplo, para los científicos de datos con el aprendizaje automático y análisis avanzado, para los desarrolladores que deseen crear aplicaciones de big data, para los administradores que dominen el monitoreo, gobernabilidad, seguridad y solucionen problemas en el clúster y para los analistas de datos que deseen analizar grandes cantidades de datos de manera sencilla y rápida (Cloudera, 2018).

Los principales productos y herramientas son:

- CDH (Cloudera Distribution for Hadoop): es la distribución líder y más conocida la cual incluye Hadoop y otros proyectos relacionados de código abierto, incluidos Apache Impala y Cloudera Search.
- Apache Impala: es un motor SQL de procesamiento paralelo masivo para análisis interactivo e inteligencia de negocios. Permite ejecutar consultas de estilo BI tradicionales con combinaciones, agregaciones y subconsultas.
- Cloudera Search: proporciona acceso fácil y natural a los datos almacenados o ingeridos en Hadoop, HBase o en la nube.

- Cloudera Manager: proporciona una Consola de administración utilizada para implementar, administrar, monitorear y diagnosticar problemas en CDH con una interfaz de usuario basada en la web que la hace simple y directa.
- Cloudera Navigator: es una solución completa para la gestión de datos, la auditoría y las tareas de gestión de datos permitiendo a los grupos de cumplimiento, administradores de datos, administradores y otros trabajen de manera efectiva con los datos a escala.

Figura 6. Ecosistema de Cloudera



Fuente: (Cloudera, 2018)

Por último, Cloudera facilita la tarea práctica de CDH mediante las máquinas virtuales Cloudera QuickStart usadas con fines de prueba, demostración y autoaprendizaje donde se incluye Cloudera Manager para administrar el clúster con un tutorial, datos de muestra y scripts para comenzar.

2.4.6.2. Hortonworks

Fue fundada en 2011 por 24 ingenieros del equipo original de Hadoop en Yahoo! que hicieron un giro para formar Hortonworks. Es 100% open source. Solo cobra por el soporte y no tiene ningún tipo de licencia siendo es el único proveedor comercial que distribuye exclusivamente Hadoop de código abierto completo sin software propietario adicional. Con sede central en Santa Clara, California, cuenta con más de 2300 socios colaboradores en ingeniería, distribución estratégica, tecnología e integración de sistemas. Hortonworks proporciona una plataforma de fuente abierta basada en Apache Hadoop para analizar, almacenar y administrar big data (Hortonworks, 2018). Los productos principales son Hortonworks Hadoop distribution (HDP) y Hortonworks DataFlow (HDF).

HDP es una distribución de código abierto que permite implementar, integrar y trabajar con volúmenes sin precedentes de datos estructurados y no estructurados y puede ser descargada e integrada fácilmente. HDP está centrada en mejorar la usabilidad del ecosistema Hadoop y está basada en una arquitectura centralizada (YARN) la cual coordina los servicios en todo el clúster para operaciones, gobernanza de datos y seguridad Su instalación es completamente gratuita, existiendo una versión de pago con la única diferencia de ofrecer soporte técnico (Hortonworks, 2018).

HDF proporciona una plataforma segura de extremo a extremo desde la fuente hasta el destino que recolecta, analiza y actúa sobre datos en tiempo real, permitiendo también

integrar nuevas fuentes y controlar en tiempo real el rendimiento de las operaciones. Además, cuenta con una interfaz visual sencilla de arrastrar y soltar. HDF es una solución integrada con Apache Nifi/MiNifi, Apache Kafka, Apache Storm y Druid (Hortonworks, 2018).

Figura 7. Ecosistema de Hortonworks

Fuente: (Hortonworks, 2018)

Hortonworks ofrece una formación en directo mediante Hortonworks University, la cual cuenta con una amplia oferta de cursos en directo que proporcionan una formación basada en casos reales. La mayor parte de los cursos en directo son impartidas por instructores certificados de Hortonworks University y están disponibles tanto en clases presenciales como de manera virtual. Permitiendo así, convertirse en un profesional certificado y para ello existen diversos tipos de certificaciones, por ejemplo, para administradores de HDP permitiendo implementar y administrar clústeres de Hadoop, para desarrolladores de

Hadoop que usan frameworks como Pig, Hive, Sqoop y Flume, para desarrolladores que diseñan soluciones basadas en Apache Hadoop escritas en Java, para desarrolladores responsables del desarrollo de aplicaciones Spark Core y Spark SQL en Scala o Python, entre otras (Hortonworks, 2018).

HDP y HDF se pueden usar de forma independiente, pero también se combinan para convertirse en una plataforma cohesiva para administrar y analizar la transmisión de datos históricos y en tiempo real. Por último, Hortonworks Sandbox (HDP o HDF) es un entorno de escritorio personal rápido y fácil para empezar a aprender, desarrollar, evaluar y probar nuevas características y cuenta con numerosos tutoriales online. El Sandbox de HDP, un VM de nodo único, permite que sea fácil comenzar con Apache Hadoop, Apache Spark, Apache Hive, Apache HBase y muchos más proyectos de datos de Apache. El HDF Sandbox hace que sea fácil comenzar con Apache NiFi, Apache Kafka, Apache Storm, Druid y el Gestor de Análisis de Retransmisión (Hortonworks, 2018).

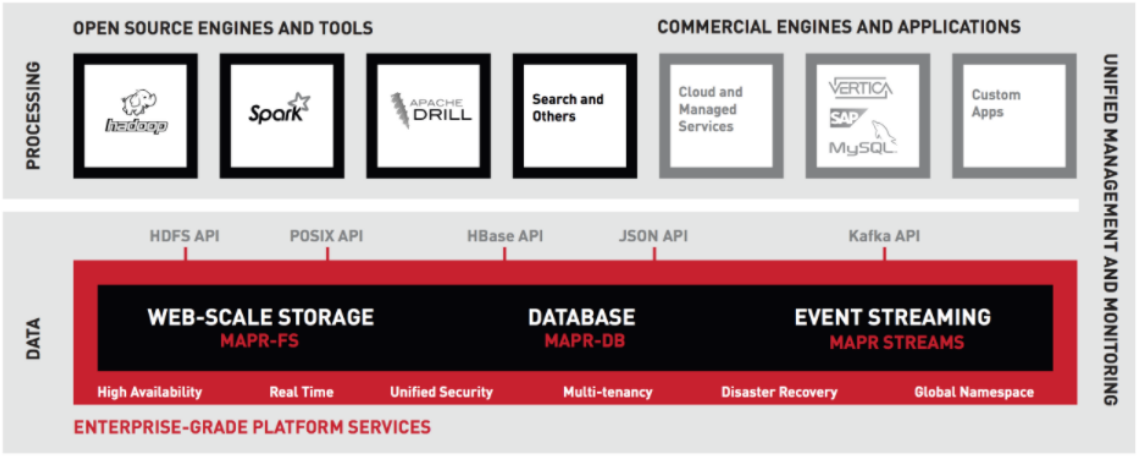
En poco tiempo, Hortonworks se ha convertido en uno de los principales proveedores de Hadoop, alcanzando rápidamente Cloudera que por ser la primera compañía de distribución de Hadoop es más grande que Hortonworks, pero ambas están creciendo simultáneamente. Ambas compañías están innovando en el mundo de Hadoop y ambas están revolucionando el espacio de Big Data.

2.4.6.3. MapR

MapR es una empresa que proporciona una plataforma de datos distribuidos de nivel empresarial para almacenar y procesar Big Data, con sede en Santa Clara, California. Su tecnología funciona tanto en hardware básico como en la nube. Su principal producto es la plataforma de datos convergentes MapR. Está diseñado con muchas optimizaciones de rendimiento para aprovechar al máximo su hardware siendo capaz de realizar análisis con velocidad y confiabilidad. Dentro de una única plataforma en una única base de código, converge las tecnologías que lo componen, que incluye un sistema de archivos distribuido, una base de datos NoSQL multimodelo (MapR-DB), un motor de transmisión de eventos de publicación y suscripción (MapR Streams), ANSI SQL (Apache DrillTM) y un conjunto amplio de tecnologías de análisis y administración de datos de código abierto (MapR, 2018).

Es el distribuidor de Hadoop que mayor esfuerzo ha hecho en hacer fiables y eficientes las mayores implementaciones en Hadoop. Para ello ha desarrollado su propio sistema de ficheros nativo en Unix. MapR reemplaza al componente HDFS y en su lugar usa su propio sistema de archivos patentado, llamado MapR-FS. Map-RFS ayuda a incorporar funciones de nivel empresarial en Hadoop, lo que permite una administración de datos más eficiente, confiable y fácil de usar. Proporciona funciones de alta disponibilidad (HA) de misión crítica, recuperación ante desastres (DR) y recuperación de datos para maximizar el tiempo de actividad y reducir el riesgo de pérdida de datos (Bloomberg, 2018).

Figura 8. Plataforma de datos convergentes MapR



Fuente: (MapR, 2018)

MapR ofrece entrenamiento técnico mediante clases virtuales y e-learning bajo demanda permitiendo obtener diferentes certificaciones que incluyen administración de clústeres, desarrollador HBase, desarrollador Spark o analista de datos. También ofrece servicios profesionales de diseño e implementación de clústeres, migración de datos, servicio de actualización de clúster, optimización de ETL, diseño convergente de aplicaciones analíticas, diseño de aplicación operacional convergente, entre otros (MapR, 2018).

Por tanto, mientras Cloudera y Hortonworks distribuyen los metadatos y procesamiento en los DataNodes y NameNodes propios de HDFS, MapR tiene una arquitectura distinta, que vale la pena tener en cuenta al momento de implementar una solución, ya que posee un enfoque más distribuido que se traduce en mejores rendimientos y también diferenciándose de sus competidores por sus características de alta disponibilidad.

3. MODELO PROPUESTO

En este capítulo se describe el modelo genérico propuesto para la administración de una solución de Big Data que permita su aplicación en el sector bancario tradicional. En primera instancia, se va a señalar algunos motivos que pueden llevar al fracaso de un proyecto de esta índole y de esas evidencias, poner énfasis en cinco aspectos antes de iniciar un proyecto de Big Data. Luego, se presentará el modelo y se describirá las particularidades de cada fase.

3.1. Motivos de fracaso de un proyecto de Big Data

A pesar del potencial y oportunidad de negocio que brinda el uso de la tecnología de Big Data, el 60% de estos proyectos no están teniendo éxito y no de Big Data precisamente por la tecnología (Viaña, 2016). Disponer de centenares de miles de millones de datos sin saber qué hacer con ellos o equivocarse en la estrategia de recogida y análisis, es realmente una pérdida de tiempo y dinero. Por otra parte, hay empresas que han formado la estructura de Big Data, implementado toda la tecnología obteniendo los datos en tiempo real, pero las decisiones las adoptan comités de dirección que se reúnen una vez al mes, lo cual no tendría sentido al disponer de información inmediata.

El error es seguir pensando como una empresa tradicional cuando la tecnología ya nos permite adoptar decisiones en el momento, tal como lo exige el mercado de hoy en día. Reinventarse es la clave de muchas empresas que ahora son más innovadoras y ágiles

gracias a que se embarcaron en la ola de esta revolución llamada Big Data. Sin embargo, contar con un modelo mal definido que hago foco solamente en la tecnología nos situará en ese grupo de proyectos que no tienen éxito.

Entre los motivos de fracaso de un proyecto de Big Data podemos mencionar los siguientes:

- No tener objetivos claramente definidos.
- Baja calidad de datos.
- Falta de patrocinio ejecutivo, presupuesto y prioridad.
- Falta de personal con las habilidades adecuadas.
- Elección de la arquitectura incorrecta.
- Proyecto demasiado ambicioso enfocándose en varios frentes.
- Creer que es un proyecto que pertenece a tecnología y no se necesita del negocio.
- Adaptarse a los procesos tradicionales.
- Expectativas sobredimensionadas.

3.2. Aspectos a tomar en cuenta al iniciar un proyecto de Big Data

Al iniciar un proyecto de Big Data es fundamental poner énfasis en los siguientes cinco aspectos:

1. Definir una metodología: Por lo general, los proyectos de Big Data comienzan con un caso de uso específico y un conjunto de datos. En el transcurso de las implementaciones, las necesidades evolucionan a medida que se comprenden e integran los datos y es entonces cuando se comienza a aprovechar su real valor. Además, de necesitar nuevos roles (arquitectos de datos, ingenieros, científicos de datos y expertos en visualización) perfectamente coordinados. Por ello se hace necesario utilizar una metodología ágil de implementación y un enfoque iterativo que permita un profundo análisis de los datos disponibles y la modelización de los mismos produciendo resultados alineados con el negocio a corto plazo.

2. Planificación del proyecto de Big Data: Una correcta definición de los objetivos en términos de negocios es el mejor punto de partida. Es recomendable no comenzar por algo demasiado grande. Lo ideal es empezar con un proyecto conciso, acerca de un tema específico que no abarque demasiados ámbitos, y centrarse en él. Además, mantener motivado al equipo para lograr el objetivo y alinear las expectativas de la alta dirección, serán claves.

3. Obtener la aprobación de la alta dirección: Es fundamental y es uno de los aspectos que garantiza el éxito del proyecto. Tener la aprobación de la dirección da confianza y garantiza al equipo el acceso a todos los datos de negocio relevantes y también el apoyo de otras áreas, y así poder encontrar los patrones y las relaciones que responderán a las preguntas de negocio.

4. Ejecución del proyecto de Big Data: Al igual que la planificación, la buena gestión del proyecto es importante. La ejecución del proyecto requiere la atención continua del equipo de trabajo mediante un monitoreo constante del proyecto, así como una comunicación efectiva y ágil entre el equipo y demás interesados.

5. Contar con equipo capacitado: Se debe tener un equipo multidisciplinario que esté enfocado en el objetivo con una buena comprensión tanto de la organización como de sus datos y entender la relación con el negocio para ayudar a la organización. El equipo deberá estar formado por analistas de negocios, analista de datos, arquitecto de datos, gestores de visualización, entre otros). Es relevante que al menos un miembro del equipo tenga experiencia en estas tecnologías, aunque no es indispensable.

3.3. Modelo para la administración de Big Data en el sector bancario

El modelo propuesto para la administración de Big data aplicado al sector bancario consta de 7 fases: definición, fuentes de datos, diseño de aplicación, capturar y almacenar, modelado y limpieza, análisis y evaluación y monitoreo. A continuación, se describe cada fase tomando en consideración que el entregable de una fase sirve de entrada para la siguiente fase.

Figura 9. Fases del modelo propuesto

Fuente: Elaboración propia.

3.3.1. Fase 1: Definición

En esta fase se identifican las necesidades del negocio. El equipo de desarrollo del proyecto debe trabajar junto al área de negocios, para elaborar y refinar constantemente sus necesidades de información. Los objetivos del proyecto deben definirse con máxima claridad evitando ambigüedades, ya que muchas dudas se irán despejando en la medida que se avance, y podrán apreciarse mejor los beneficios (Bellé, 2016).

Debido al fuerte componente tecnológico se tiende a definir los resultados únicamente en función de la tecnología. Llevando al equipo de desarrollo a concentrarse en conseguir los datos y luego no se sabe qué hacer con éstos. Esta forma de pensar provoca que se diluya la responsabilidad sobre los resultados finales, que tiene que ser también del negocio. En síntesis, los objetivos no pueden apuntar a la tecnología, tienen que extenderse abarcando el uso final de los datos y su conversión en valor para el negocio.

Por ello, es necesario identificar a los usuarios finales y poder saber cuáles son las preguntas que requieren responder para tomar mejores decisiones. Además de las preguntas del negocio, debemos contestar una serie de preguntas del tipo:

- ¿Dónde se originan los datos?
- ¿Cuál es el tamaño de estos datos y qué tanto crecen?
- ¿Con qué frecuencia se originan y con qué frecuencia se van a necesitar?
- ¿Qué tipos de datos se necesitan?
- ¿Cómo garantizamos su fiabilidad y veracidad?
- ¿Cómo se van a almacenar?
- ¿Es necesario analizarlos en tiempo real?
- ¿Podemos combinar los datos internos con otros datos externos que ayuden a buscar correlaciones valiosas?

Al tratarse de una implementación nueva, es recomendable empezar con un tema específico que no abarque muchos ámbitos de un conjunto de datos que pueda ser evaluado fácilmente, con el fin de monitorear el proyecto y corregir errores sobre la marcha. Por su contraparte, al intentar abarcar un alcance muy amplio y ambicioso será más difícil de evaluar los resultados y cuando se lo quiera ajustar puede resultar más complejo.

El desarrollo de aplicaciones de negocios Big Data es un proceso iterativo, que requiere paciencia. Hay que asegurarse de que todas las personas involucradas estén de acuerdo con el alcance y los objetivos del proyecto. En esta fase también se determina la aceptación del presupuesto del proyecto y se identifican a los interesados del proyecto para gestionar sus expectativas e incertidumbre.

3.3.2. Fase 2: Identificar fuentes de datos

Una vez identificadas las necesidades del negocio, se tiene que ubicar las fuentes de información requeridas para responder a las preguntas del negocio. Se trata de identificarlas primero para analizar y consumir sólo aquellos datos que sean relevantes a los objetivos antes planteados. Por lo general, la mayor cantidad de fuentes de información de una organización serán internas, es decir, se encontrarán en un Data Warehouse, bases de datos relacionales, bases de datos NoSQL, CRM, ERP, archivos de texto plano, libros de Excel, entre otros.

Una de las ventajas del uso de tecnologías de Big Data es que permite enriquecer la información interna disponible en una organización con información de fuentes externas proporcionadas de la web, redes sociales, por otras empresas, open data, entre otras. Sin embargo, no toda la información externa es útil o tiene la calidad suficiente para satisfacer nuestros objetivos de análisis y permitir la extracción de conocimiento fiable. Si el conocimiento extraído no es fiable puede llevar a la toma de decisiones errónea respecto al proceso de negocio que pretendemos mejorar dando lugar a pérdidas económicas y al fracaso del proyecto. En efecto, es importante verificar la calidad de las fuentes de datos internas y externas usadas, así como la correlación entre éstas.

Por otra parte, es necesario identificar los tipos de datos que se necesitarán, por ejemplo, la mayor parte de datos estructurados la encontramos en fuentes de datos internas, aunque pueden existir también datos semiestructurados y no estructurados en correos, documentos de texto, etc. En el caso de las fuentes externas la mayor parte de los datos serán semiestructurados y no estructurados.

Una vez identificadas cuáles son las fuentes de datos necesarias para nuestra estrategia, podremos consumirlas y almacenarlas para luego aplicar técnicas de analítica y convertir los datos en información que aporte valor de negocio, algo que sólo será posible sabiendo gestionar tanto la información que nos ofrecen los datos estructurados, como no estructurados con la ayuda de las herramientas de analítica adecuada.

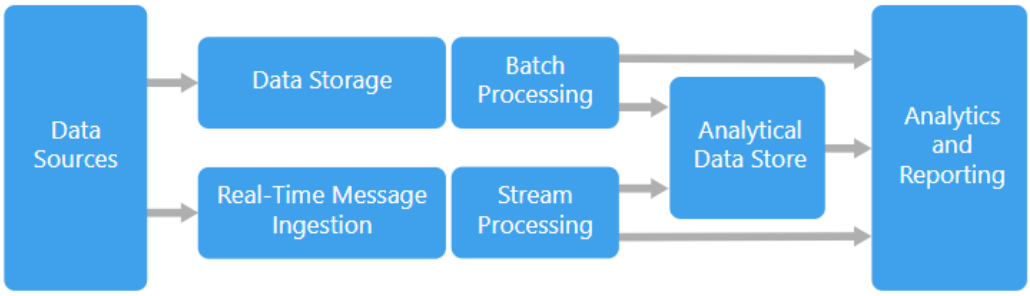
3.3.3. Fase 3: Diseño de la aplicación

El análisis realizado en la fase anterior nos va a permitir elegir una arquitectura que cumpla las necesidades del negocio. El tener los objetivos claros y los tipos de fuentes de datos identificados ayudará a saber qué se debe potenciar en la arquitectura. Aun cuando, al inicio, es recomendable centrarnos en un problema de negocio específico para lograr mejores resultados, hay que tener en cuenta que la arquitectura crecerá y deberá ser capaz de soportar futuros casos de uso (Rivera, 2017). Es por eso que la estructura deberá tener al menos escalabilidad, flexibilidad y ser tolerante a fallos.

En la ingesta de la información, no todas las herramientas sirven para cualquier fuente de datos, en algunos casos lo mejor sería combinar varias herramientas que cubran todos los casos. Para el procesamiento se debe evaluar si el sistema tiene que ser streaming o batch. Lo ideal sería aprovechar el procesamiento en streaming que ofrece Big Data. Asimismo, es recomendable utilizar herramientas para controlar, monitorizar y gestionar la arquitectura, esto facilitará y centralizará todo tipo de tareas.

La mayoría de las arquitecturas de Big Data incluyen algunos o todos los siguientes componentes (Wasson, 2017):

Figura 10. Estilo de Arquitectura para Big Data



Fuente: (Wasson, 2017)

- Fuentes de datos: todas las fuentes de las cuales se ingesta datos.
- Almacenamiento de datos: esta capa recibe datos de las fuentes internas y externas.
- Procesamiento por lotes: al ser los conjuntos de datos tan grandes, a menudo una solución de Big Data debe procesar archivos de datos mediante trabajos por lotes de larga ejecución para filtrar, agregar y preparar los datos para el análisis.
- Ingestión de mensajes en tiempo real: si la solución incluye fuentes en tiempo real, la arquitectura debe incluir una forma de capturar y almacenar mensajes en tiempo real para el procesamiento de flujo.
- Procesamiento de flujo: después de capturar mensajes en tiempo real, la solución debe procesarlos mediante el filtrado, la agregación y la preparación de los datos para el análisis.
- Almacén de datos analíticos: se preparan los datos para el análisis.

- Análisis e informes: finalmente los usuarios podrán elaborar informes y análisis mediante herramientas de autoservicio de BI, además sirve también de forma de exploración interactiva de datos por parte de los científicos de datos o analistas de datos.

3.3.4. Fase 4: Captura y almacenamiento de datos

Una vez que se ha diseñado el despliegue de un clúster el siguiente punto a considerar es como se van a cargar los datos. En esta fase se consumen los datos estructurados o no estructurados sin sufrir adaptaciones y se almacenan para ser analizados. Los datos pueden ser almacenados en infraestructura propia o mediante la contratación de un proveedor que ofrezca una plataforma Big Data en la nube.

Una arquitectura basada en el procesamiento y almacenamiento distribuido son una buena solución para almacenar y procesar el flujo continuo de datos, frente a lo poco que pueden hacer las bases de datos relacionales. Hadoop es capaz de almacenar y tratar los datos de un modo eficiente, pudiendo procesar con rapidez grandes cantidades de información.

Para el caso de los datos estructurados almacenados en bases de datos relacionales o en un Data Warehouse, se podría considerar a Hadoop como un complemento óptimo y el uso de este último no supone la renuncia a una estrategia de Data Warehouse. De igual manera, existen herramientas como Sqoop que permiten importar datos de una base de datos

relacional a HDFS, Hive o HBase. Además, se pueden exportar ficheros de HDFS a bases de datos relacionales.

En el caso de los datos semiestructurados y no estructurados existen diversas herramientas NoSQL que nos permitirán almacenarlos y agregan valor con características propias que no dispone Hadoop. Ambas son soluciones de Big Data para el almacenamiento de los grandes datos, complementarias y compatibles entre sí y también con respecto a las tradicionales bases de datos relacionales. La integración entre sistemas NoSQL y Hadoop es casi nativa y, asimismo, cada base NoSQL tiene su propia interfaz.

3.3.5. Fase 5: Modelado y limpieza de datos

En esta fase se preparan los datos para el análisis y luego son procesados en un formato estructurado que se puede consultar utilizando herramientas analíticas. El almacén de datos analíticos utilizado para atender estas consultas puede ser un almacén de datos relacionales de estilo Kimball, como se ve en la mayoría de las soluciones de inteligencia de negocios (BI) tradicionales.

Los datos de baja calidad conducirán a un conocimiento de baja calidad. Por lo tanto, el preprocesamiento de datos es una esencial en esta fase cuyo principal objetivo es obtener conjuntos de datos finales que puedan considerarse fiables y útiles para el análisis de los datos en la siguiente fase. El preprocesamiento en Big Data es una tarea desafiante por el

tamaño del conjunto de datos. Las mayores cantidades de datos recopilados requieren mecanismos más sofisticados para analizarlo. El preprocesamiento incluye técnicas como la preparación de datos que abarca la transformación de datos, integración, limpieza y normalización y técnicas de reducción de datos que apunta a reducir la complejidad de los datos por selección de características, selección de instancias o por discretización. Después de la aplicación de una etapa de preprocesamiento de datos, el conjunto final de datos obtenido puede considerarse como una fuente confiable y adecuada para cualquier algoritmo aplicado posteriormente (García, Ramírez-Gallego, Luengo, Benítez, & Herrera, 2016).

3.3.6. Fase 6: Análisis

Teniendo toda la información consolidada en un repositorio común, se puede empezar a realizar el análisis de grandes volúmenes de información utilizando técnicas avanzadas de análisis predictivo y minería de datos, para poder encontrar patrones de comportamiento comunes a diferentes segmentos de clientes, así como identificar las tendencias que nos servirán para predecir los escenarios más factibles a futuro para lograr los objetivos planteados.

En esta fase entran en acción los científicos de datos, analistas de datos, expertos en Data Mining, entre otros. Apoyándose en ideas, herramientas matemáticas, estadísticas e informáticas con las que trabajan para hacer análisis inteligentes de los grandes datos. Y

alineados a los objetivos de la organización con el uso de la tecnología para buscar soluciones, realizar pronósticos, proporcionar información en tiempo real accesible a través de distintos canales, mediante una fácil e intuitiva visualización de los resultados obtenidos para luego ser validados y consumidos por los usuarios del negocio. En conclusión, la fase analítica permite interactuar con los datos preprocesados y almacenados en una especie de almacén de datos analíticos para extraer inteligencia empresarial.

3.3.7. Fase 7: Evaluación y monitoreo

Finalmente, esta fase se encarga se evaluar los resultados obtenidos y los objetivos planteados al inicio del proyecto para ver si estos fueron alcanzados. Los datos se ponen a disposición de analistas de negocio. Estos analistas podrán generar informes personalizados y realizar análisis ad hoc para responder a preguntas de negocio y así poder afianzar la toma de decisiones.

En esta fase también se realiza el monitoreo de toda la aplicación desde la ingesta de datos hasta el preprocesamiento, existen algunas herramientas disponibles para la provisión, administración y monitoreo del clúster para hacer seguimiento al progreso completo en pos de la consecución de sus objetivos, y para guiar las decisiones de gestión de manera ágil. Por último, se realiza también en esta fase la mejora continua la cual me permitirá aplicar procesos de mejora durante cada fase del modelo propuesto con el fin de obtener mejores resultados que mejoran la toma de decisiones.

4. IMPLEMENTACIÓN DEL MODELO

En este capítulo se presenta la aplicación del modelo propuesto para la administración y análisis de Big Data aplicado al sector bancario, con el fin de validar y evaluar la viabilidad de su implementación. Para lograrlo, se plantea la implementación que se va a usar y además se propone un caso de estudio que nos servirá para entender y evaluar el modelo propuesto.

4.1. Implementación

Para validar el modelo propuesto se construyó una solución que contempla las fases de este modelo y se adapta al caso de estudio que se presenta a continuación. Se tomaron varias fuentes de datos estructurados y no estructurados. Para el diseño de la aplicación (Figura 11) se consideró una solución proporcionada por Cloudera, QuickStarts para CDH. QuickStarts es una máquina virtual que brinda un clúster de un solo nodo con fines de prueba y autoaprendizaje. También incluye Cloudera Manager que facilita las tareas de administración del clúster. En la fase de captura y almacenamiento de datos se va a usar Sqoop para transferir los datos a Hive. Tanto Sqoop como Hive vienen integrados en Cloudera QuickStarts. En el modelado y limpieza de datos se usará Python porque es ideal para trabajar con grandes volúmenes de datos y para el análisis se considera Tableau Software así también para la visualización. Finalmente, Cloudera Manager nos va a permitir monitorear toda la aplicación.

4.2. Caso de estudio

Para validar la arquitectura propuesta se tomó como caso de estudio los datos de un banco real, que por cuestiones de confidencialidad no es posible brindar mayor información. Tras el éxito de la primera transición a la banca online, el banco enfrenta un nuevo desafío: explotar y obtener rentabilidad de los datos de sus clientes. Una información que comprende movimientos en su cuenta cada mes y operaciones que realiza en los diferentes canales de atención como son ventanillas, cajeros automáticos, teléfono, internet y celular. En efecto, el banco persigue ofrecer productos y servicios financieros con la mejor calidad que atiendan las necesidades de los clientes. Todo dentro de un esquema de eficiencia y rentabilidad.

4.3. Prototipo

La creación del prototipo se concentra en el caso particular de los datos del banco y para ello, se toman en cuenta las 7 fases del modelo propuesto. A continuación, se detalla lo realizado en cada una:

4.3.1. Aplicación de la fase de Definición

Las preguntas de negocio que se identificaron y se pretenden responder para este caso de estudio son las siguientes:

1. ¿Cuál es el canal de atención por el que más se realizan transacciones?
2. ¿Cuál es el promedio de edad de los clientes por cada canal?

3. ¿Cuál es el canal más usado para cada cliente?
4. ¿En qué rango de edad se encuentran los clientes que más realizan transacciones?
5. ¿En qué rubro gastan más los clientes?
6. ¿Cómo me anticipo a la detección de abandono de los clientes para realizar acciones
7. ¿Cómo es la salud financiera de los clientes?

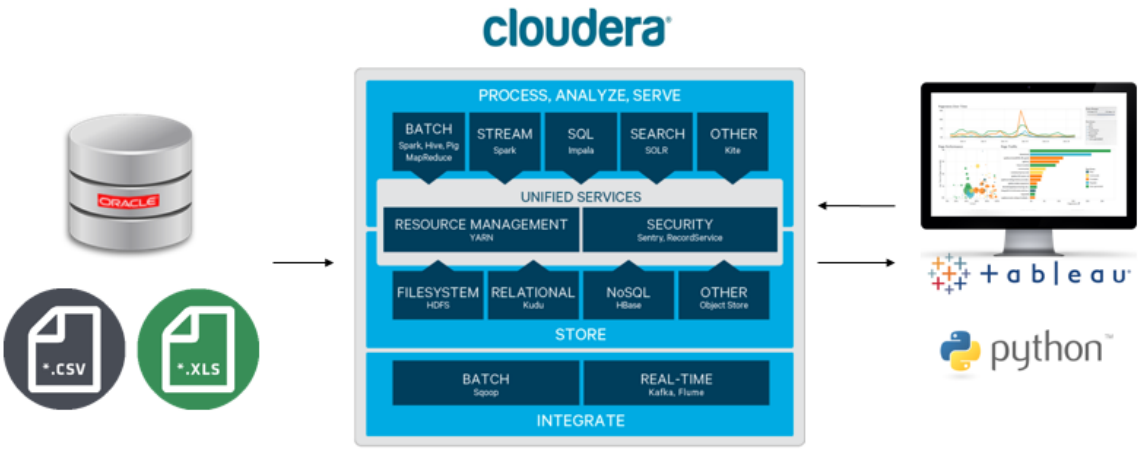
4.3.2. Aplicación de la fase de Identificación de fuentes de datos

La fuente de datos principal a usarse es Oracle, con más de 45 millones de registros correspondientes al año 2015. Estas 6 tablas disponen de información de los clientes y las transacciones que realizaron. Las transacciones proceden de los pagos y de algún producto o servicio que compraron en los canales de atención que dispone el banco. Asimismo, se van a usar libros de Excel y archivos de texto.

4.3.3. Aplicación de la fase de Diseño de la Aplicación

Tomando como base la solicitud del negocio y las fuentes de datos identificadas, se logró diseñar la siguiente arquitectura. Esta consta de las fuentes de datos, identificadas anteriormente, y que serán consumidas y procesadas por Cloudera para luego esos datos ser usados por científicos de datos con ayuda Python y también para analistas de negocios y analistas de datos con la herramienta líder en análisis e inteligencia de negocios y autoservicio Tableau Software.

Figura 11. Diseño de la Aplicación

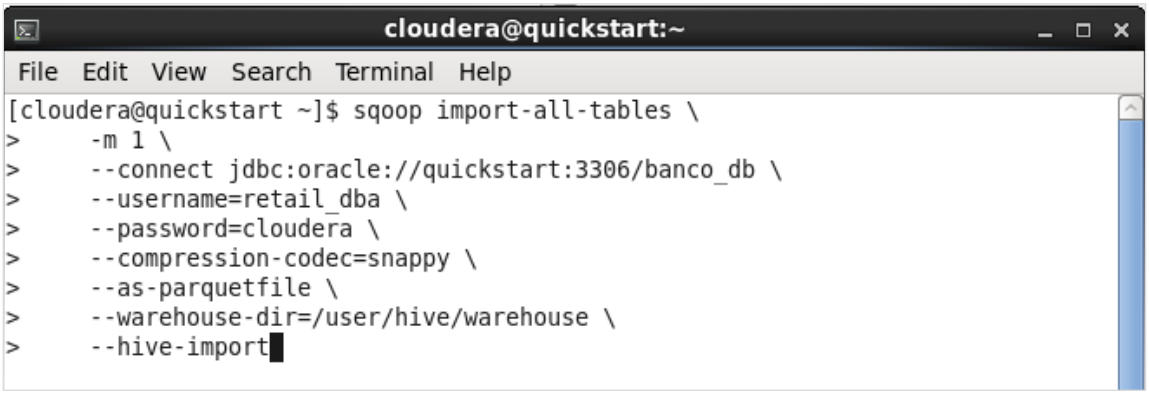


Fuente: Elaboración propia.

4.3.4. Aplicación de la fase de Captura y almacenamiento de datos

Se van a importar los datos provenientes de Oracle mediante Apache Sqoop. Sqoop es una herramienta que usa MapReduce para transferir datos entre clústeres de Hadoop y bases de datos relacionales de manera eficiente. Funciona al generar tareas en múltiples nodos para descargar los datos en paralelo. Al terminar, cada dato se replicará para garantizar la confiabilidad y se distribuirá en todo el clúster para el procesamiento en paralelo. Lo bueno de Sqoop es que podemos cargar automáticamente los datos relacionales de Oracle a HDFS conservando la estructura. Luego, estos datos son enviados a Apache Hive, también llamado el Data Warehouse de Hadoop (Cloudera, 2018). En la siguiente imagen se observan los comandos usados en el terminal para lanzar Sqoop.

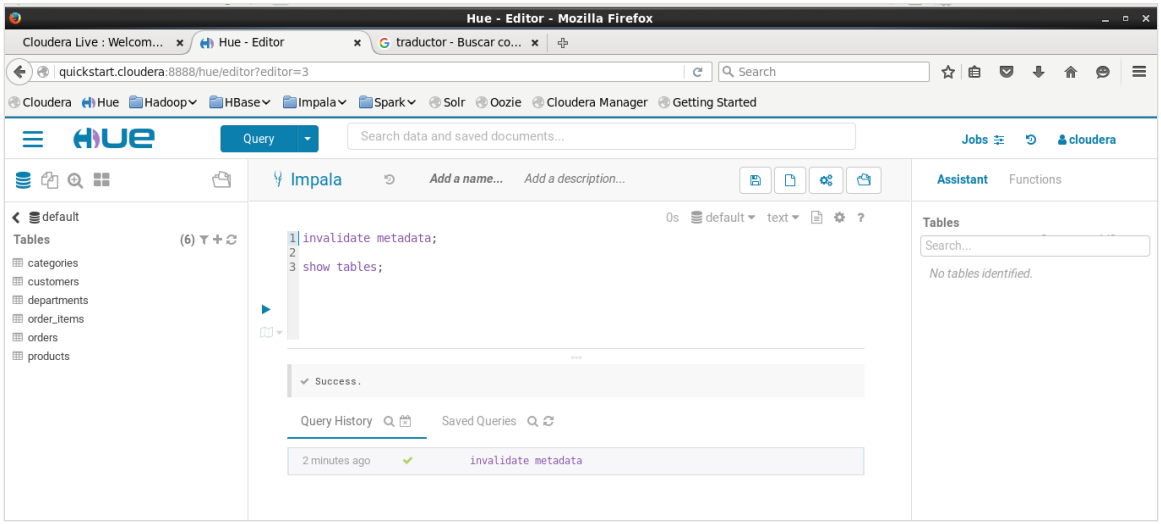
Figura 12. Importar datos a Apache Hive



Fuente: Elaboración propia.

Ahora usaremos Cloudera Impala para consultar los datos. Impala es un motor de consultas SQL open source de Cloudera para el procesamiento masivo en paralelo (MPP) de los datos almacenados en un clúster de computadoras corriendo en Apache Hadoop. Además, Impala puede compartir los archivos de datos y los metadatos de las tablas con Hive (Cloudera, 2018). En la siguiente imagen se verifica que todo esté correcto desde Cloudera HUE usando las consultas de Impala. HUE (Hadoop User Experience) es una interfaz de usuario web para la gestión de Hadoop creada como proyecto open source por Cloudera. Desde HUE podemos ejecutar querys como si estuviéramos en un motor de base de datos relacional.

Figura 13. Consultar los datos con HUE



Fuente: Elaboración propia.

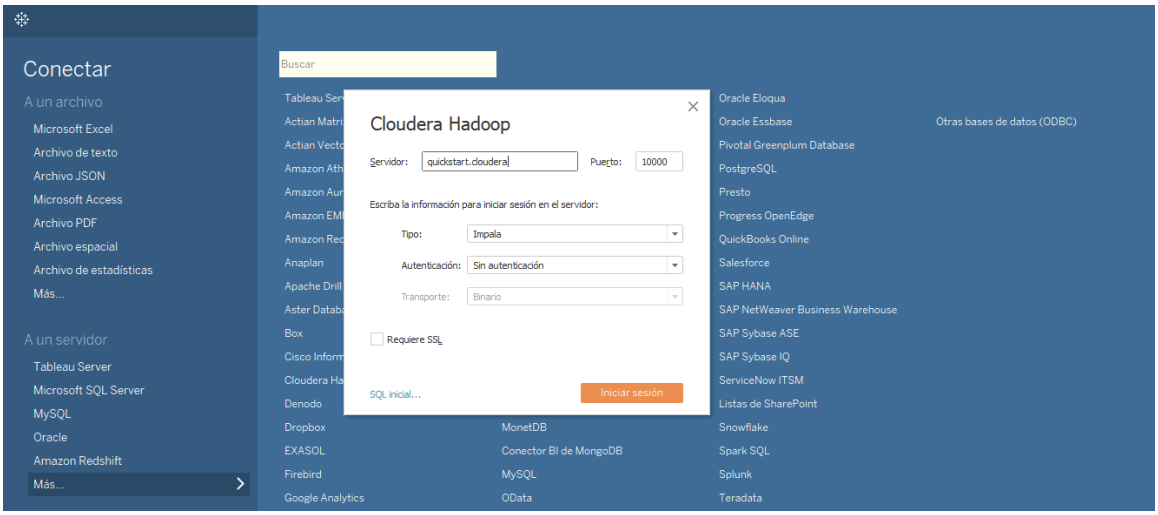
4.3.5. Aplicación de la fase de Modelado y limpieza

Para la aplicación de la fase de Modelado y limpieza de datos se va a usar Python. Python es un lenguaje de programación muy sencillo de aprender, fácil de leer y cuenta con una amplia variedad de librerías con múltiples posibilidades. La curva de aprendizaje es muy corta debido a su sintaxis bastante intuitiva y posee un tipado dinámico. Tiene una licencia de código abierto, denominada Python Software Foundation License. Python tiene distintos y variados paquetes en el campo del procesamiento de datos que además se encuentran vinculados con herramientas GIS, matemáticas, estadística, etc. Esta versatilidad de Python la convierte en una herramienta multifuncional para usar en sinergia con Big Data (Python, 2018).

4.3.6. Aplicación de la fase de Análisis

Para la aplicación de la fase de Análisis usaremos Tableau Software. Tableau cuenta con un conector de Cloudera para Hadoop y es posible conectarse con Impala o Hive. Antes de realizar la conexión se debe descargar el controlador ODBC para Impala o Hive desde el sitio web de Cloudera. Una vez instalado el conector procederemos a configurar la conexión y luego se cargan las tablas de Hadoop en Tableau y así poder realizar nuestro análisis. Al ser un set de datos tan grande es recomendable filtrar para reducir la cantidad de datos en Tableau y además realizar un extracto que a partir de la versión 10.5 de Tableau usa el nuevo motor hyper. El extracto nos permitirá mejorar el rendimiento de análisis y consulta más rápido para conjuntos de datos de mayor tamaño.

Figura 14. Conector Cloudera Hadoop

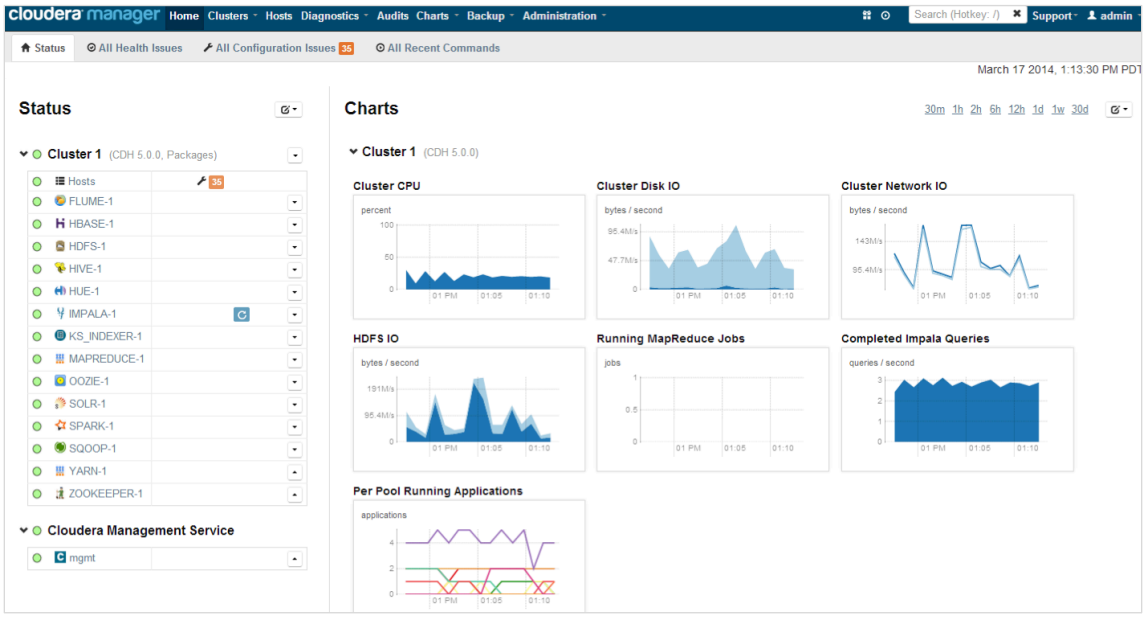


Fuente: Elaboración propia.

4.3.7. Aplicación de la fase de Evaluación y Monitoreo

En esta fase usaremos Cloudera Manager, aunque solo está disponible para la versión de pago de Cloudera Enterprise. Proporciona una interfaz centralizada, siendo posible ajustar fácilmente configuraciones y recursos, gestionar una amplia gama de funciones de usuario para el acceso de autoservicio entre departamentos, administrar múltiples clústeres e incorpora una gama completa de herramientas de informes y diagnóstico para ayudarlo a optimizar el rendimiento y la utilización (Cloudera, 2018).

Figura 15. Consultar los datos con Cloudera Manager



Fuente: Elaboración propia.

4.4. Presentación de resultados

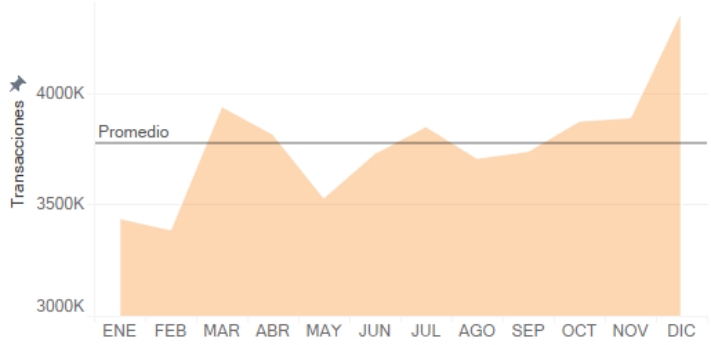
Para la primera pregunta, podemos determinar que, de los 45 millones de transacciones efectuadas en un periodo completo, el 95% corresponden a pagos realizados por Ventanilla, seguido por Internet con un 4,5% y Cajeros Automáticos con 1% (Figura 16). Permitiendo determinar que el canal ventanilla siguie siendo el más usado por los clientes, a pesar de existir los mismos servicios en el resto de canales.

Figura 16. Cantidad de transacciones por canal

Canal Descripción	Transacciones	% de total
Ventanilla	42.662.747	94,23%
Intenet	2.070.195	4,57%
Cajero Automático	412.237	0,91%
Celular	108.256	0,24%
Teléfono	21.230	0,05%
Total general	45.274.665	100,00%

Fuente: Elaboración propia.

Figura 17. Cantidad de transacciones por meses



Fuente: Elaboración propia.

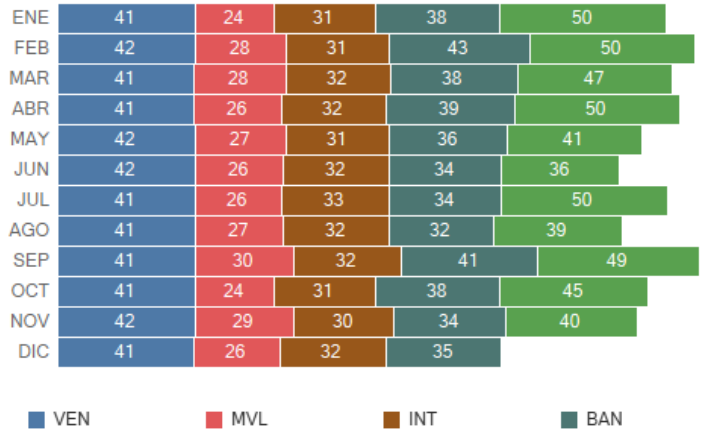
En cuanto a la segunda pregunta, el promedio de edad de uso de los canales son 45 años para Teléfono, 42 años para Ventanilla, 37 años para Cajeros Automáticos y como era de esperarse para Internet y Celular los promedios de edad son de 31 y 27 respectivamente (Figura 18).

Figura 18. Promedio de edad de los clientes por canal

Canal Descripción	
Teléfono	44,9
Ventanilla	41,5
Cajero Automático	36,7
Intenet	31,4
Celular	26,8

Fuente: Elaboración propia.

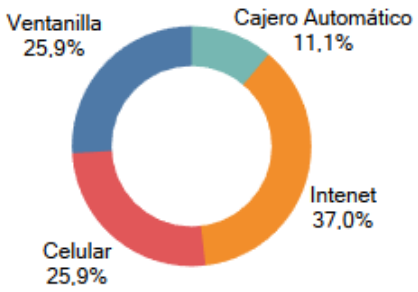
Figura 19. Promedio de edad de clientes por mes



Fuente: Elaboración propia.

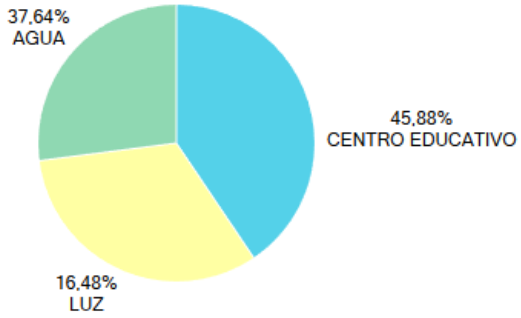
Para la tercera pregunta, analizamos un cliente en particular que tuvo 37 transacciones a lo largo del periodo, observamos que el canal más usado es Internet (Figura 20) y esto se correlaciona con su edad de 28 años. En este caso podríamos realizar marketing personalizado enfocado a sus preferencias. Para el caso analizado, observamos que el 46% de sus pagos es a Centros Educativos (Figura 21). Otra forma de llegar al cliente es mediante el envío de mensajes o publicidad que lo haga conocer y lo motive a que también puede pagar algún servicio de Ventanilla en Internet.

Figura 20. Canal más usado de un cliente



Fuente: Elaboración propia.

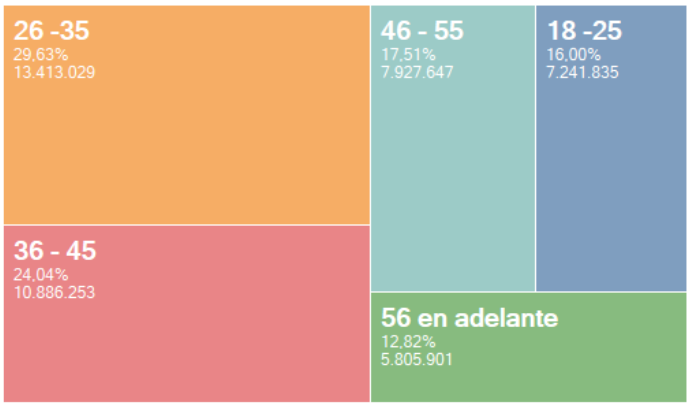
Figura 21. Pagos que realiza en Internet



Fuente: Elaboración propia.

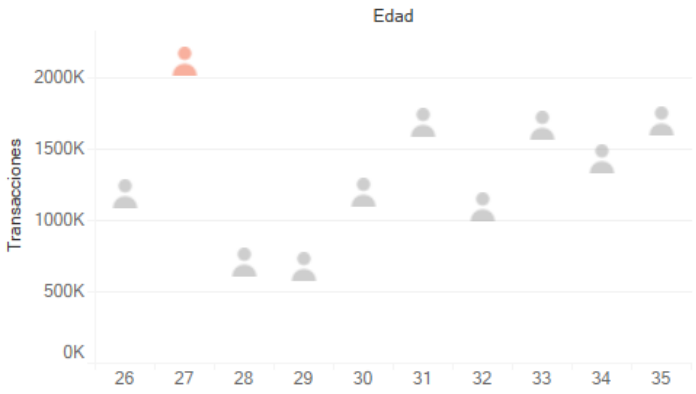
En la cuarta pregunta, se agruparon los clientes por rangos de edades dando como resultado que los clientes que más realizan transacciones tienen entre 26 y 35 años, seguido por los clientes entre los 36 y 45 años. Siendo los clientes de 56 años en adelante los que menos interactúan con los canales (Figura 22). En la Figura 23 observamos que los clientes de 27 años son los que más interactúan en los canales.

Figura 22. Cantidad de transacciones por rango de edad



Fuente: Elaboración propia.

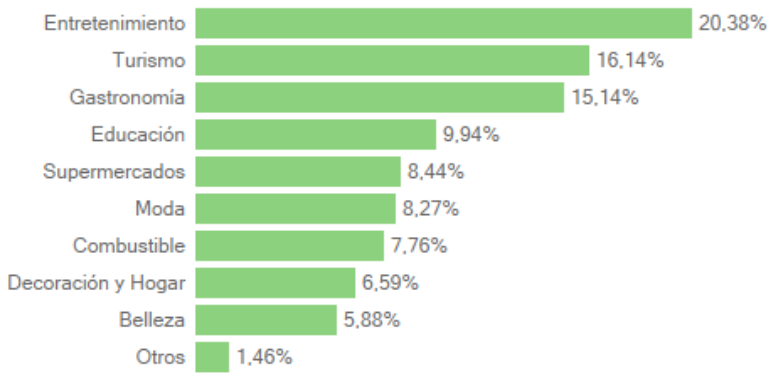
Figura 23. Cantidad de transacciones por rango de edad



Fuente: Elaboración propia.

Para la quinta pregunta, se realizó una agrupación de los tipos de consumo de los clientes para categorizarlos en 10 rubros y así poder identificar el rubro en que más gastan. Como resultado dio al entretenimiento el primer lugar con 20%, seguido por turismo con 16% y gastronomía con 15% (Figura 24). Con esta información podríamos ir más al detalle identificando qué gastos realizó el cliente, dónde compró, en qué fecha o mes realizó más compras, los horarios que realizó ese consumo, edad, estado civil, etc. y así poder conocer sus gustos para ejercer una comunicación y acciones comerciales mucho más a medida con publicidad que realmente le llame la atención y lo más importante en tiempo real gracias al procesamiento en streaming que ofrece la tecnología de Big Data.

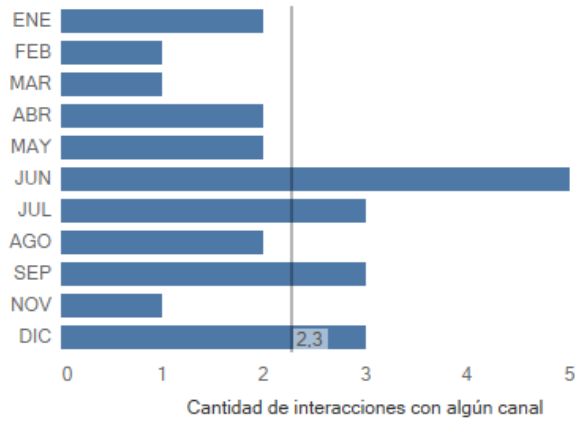
Figura 24. Tipos de consumo



Fuente: Elaboración propia.

La sexta pregunta corresponde al anticipo a la detección de abandonos de clientes para realizar acciones de retención. Para lograrlo podemos tomar en cuenta las veces que interactuó un cliente con algún canal y obtener un promedio en un periodo determinado y el caso que ese promedio baje menos de la mitad podremos crear una alerta que nos indique que ese cliente es propenso al abandono y así realizar acciones de retención y fidelización con dichos clientes. En la Figura 25 se muestra un cliente que tiene interacciones en todos los canales con un promedio de 2,3. Si a esto le aplicamos el criterio antes mencionado más otras variables como el retiro total de sus fondos o alguna queja sin resolver, podremos darle una calificación en base a esos criterios y atender los casos con mayor calificación que son propensos al abandono.

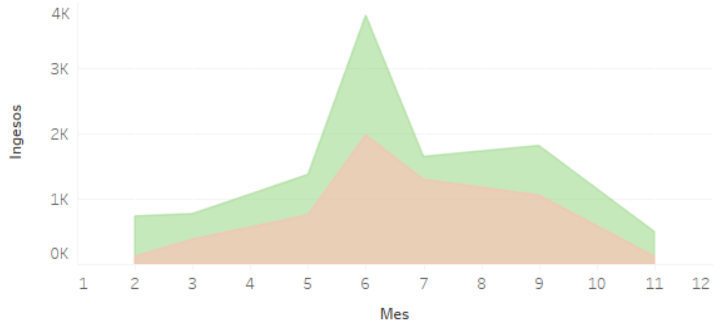
Figura 25. Promedio de interacción de un cliente en los canales



Fuente: Elaboración propia.

La séptima pregunta se enfoca en la salud financiera del cliente. Para eso vamos a analizar los ingresos y gastos de un cliente y obtener la diferencia. En el caso, de ser un valor positivo me indica que tiene una buena salud financiera, caso contrario, una mala salud financiera. Luego de este análisis se podría ofrecer al cliente algún tipo de inversión que se adapte al valor que tiene "ahorrado". Esta publicidad o recomendación se la haremos llegar por el canal que el cliente más interactúe.

Figura 26. Monto de ingresos y gastos por meses (buena salud)



Fuente: Elaboración propia.

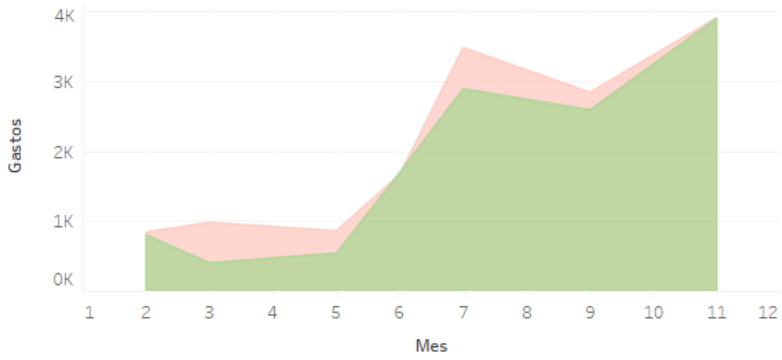
Figura 27. Situación financiera (buena salud)

Mes	Ingesos	Gastos	Estado Financiero
FEB	$ 736	$ 119	$ 617
MAR	$ 772	$ 380	$ 392
ABR	$ 1.833	$ 732	$ 1.101
MAY	$ 1.374	$ 754	$ 620
JUN	$ 3.817	$ 1.982	$ 1.835
JUL	$ 1.650	$ 1.298	$ 352
AGO	$ 1.405	$ 1.047	$ 358
SEP	$ 1.818	$ 1.058	$ 760
NOV	$ 498	$ 119	$ 379
Total general	$ 13.903	$ 7.489	$ 6.414

Fuente: Elaboración propia.

Por otro lado, para los clientes que no cuentan una buena salud financiera se podría ofrecer asesoría para el manejo correcto de sus finanzas, así como recomendaciones y complementarlo con alguna aplicación que permita tener el control de los ingresos y gastos.

Figura 28. Monto de ingresos y gastos por meses (mala salud)



Fuente: Elaboración propia.

Figura 29. Situación financiera (mala salud)

Mes	Ingesos	Gastos	Estado Financiero
FEB	$ 804	$ 843	-$ 39
MAR	$ 402	$ 981	-$ 579
ABR	$ 910	$ 642	$ 268
MAY	$ 542	$ 861	-$ 319
JUN	$ 1.697	$ 1.670	$ 27
JUL	$ 2.896	$ 3.484	-$ 588
AGO	$ 2.271	$ 3.115	-$ 844
SEP	$ 2.592	$ 2.844	-$ 252
NOV	$ 3.905	$ 3.912	-$ 7
Total general	$ 16.019	$ 18.352	-$ 2.333

Fuente: Elaboración propia.

4.4. Estudio comparativo entre las entidades bancarias tradicionales y las fintech

El sector financiero se enfrenta a grandes desafíos. Uno de ellos es adaptarse a la revolución digital donde en poco tiempo las fintech se han posicionado como las principales propulsoras del cambio, obligando a los actores más tradicionales a evolucionar al mismo ritmo.

Las fintech hacen uso de la tecnología más innovadora disponible para construir interfaces para clientes de uso intuitivo. Además, tienen una mayor cercanía y entendimiento de las necesidades de los mercados y una mayor eficiencia en la gestión de los proyectos en comparación con las instituciones financieras tradicionales. Dado que no poseen una gran infraestructura, su misión es hacer que la experiencia de usuario sea mejor que la disponible a través de los bancos y esto lo consiguen enfocándose en un solo producto o servicio financiero. Además, las fintech han demostrado ser mejores en la extracción de conocimiento muy valioso del cliente para ofrecer un mejor servicio y tomar decisiones rápidas. Están mostrando al sector tradicional nuevas oportunidades y modelos de negocio, a menor costo y con mayor velocidad.

Por otro lado, las entidades bancarias tradicionales todavía no están explotando los beneficios que ofrecen las nuevas tecnologías para innovar en la oferta de productos y servicios financieros. Las regulaciones y el rápido ritmo de la tecnología hacen que sea difícil para ellos mantenerse a la vanguardia de la innovación. Sin embargo, los bancos

cuentan con relaciones duraderas de sus clientes, están regulados y controlados constantemente lo cual incrementa la confianza de sus clientes, tienen una visión profunda de la infraestructura financiera y cuentan con una gran billetera.

La fortaleza de las fintech es la debilidad de la banca, y viceversa, entonces para subsistir entre ambas industrias tendrán que aprovechar la ventaja competitiva para preservar su valor dentro de la cadena. Por ello, muchas entidades bancarias están optando por la colaboración de las fintech mediante la compra de éstas o mediante algún tipo de asociación o inversión. La disrupción de las fintech ha contribuido en la oferta de mejores servicios financieros enfocados al cliente y esto es lo que los bancos tradicionales tienen que apuntar.

5. CONCLUSIONES

El desarrollo de esta tesis se orienta a buscar alternativas a las limitaciones encontradas respecto a la implementación y administración de grandes volúmenes de datos para las entidades bancarias tradicionales. Se logró definir un modelo que consta de varias fases para la gestión y tratamiento de grandes volúmenes de datos. La aplicación de este modelo de trabajo es una guía que debe ser adaptada a la situación y necesidad particular de cada organización.

En este trabajo se trata también de demostrar que las nuevas tecnologías asociadas a Big Data juegan un papel importante en el éxito de una organización. Asimismo, la innovación y agilidad permiten crear nuevas oportunidades de negocios que marquen la diferencia de la competencia.

Las conclusiones extraídas del análisis de los datos beneficiarán tanto a las empresas como a los usuarios. Los bancos tendrán un mayor conocimiento de sus clientes, podrán personalizar su oferta, tomar decisiones basadas en hechos y mejorar sus sistemas de atención lo que permitirá un incremento en sus ventas. Mientras que, los consumidores finales, obtendrán productos y servicios más adecuados y ajustados a sus necesidades, además de una experiencia comercial personalizada. Es necesario dar el paso, adaptarse a los nuevos tiempos y a las tecnologías actuales que permiten enfocarse en satisfacer las necesidades.

6. FUTURAS LÍNEAS DE INVESTIGACIÓN

Como futuras líneas de investigación se plantean diferentes aspectos enmarcados dentro de las diferentes fases del modelo propuesto.

Para la fase de fuentes de datos, se espera poder incluir nuevas fuentes de datos y poder contar con mayor información y nuevas variables que permitan tener una visión más completa y así entregar otro tipo de análisis más avanzado y predictivo. En la fase de diseño de la aplicación se plantea la posibilidad de implementar múltiples nodos, para aprovechar todo el potencial de Hadoop con el procesamiento y almacenamiento distribuido. Para la fase de almacenamiento se plantea trabajar en la aplicación de una arquitectura que comprenda la integración de herramientas para datos estructurados y no estructurados al mismo tiempo. En la fase de análisis, se pueden usar nuevas herramientas que permitan realizar análisis predictivo, análisis de sentimiento y también aprendizaje automático con ayuda de herramientas como, por ejemplo, Apache Mahout y lenguajes de programación estadístico y de datos como R y Python. Finalmente, para la fase de monitoreo se plantea usar cualquiera de las dos versiones de Cloudera Manager, la Express que requiere de mayores recursos de hardware y la Enterprise que tiene costo.

BIBLIOGRAFÍA

Acens. (2014). Bases de datos NoSQL. Qué son y tipos que nos podemos encontrar. *acenswhitepaper*, 2.

Adell, F., & Guersenzvaig, A. (2013). Big Data y los nuevos métodos de visualización de de la información. En F. Adell, & A. Guersenzvaig.

Amazon Web Services. (2017). *¿Qué es NoSQL?* Obtenido de Amazon Web Services: https://aws.amazon.com/es/nosql/

Amazon Web Services. (2017). *¿Qué es un Lago de datos?* Obtenido de Amazon Web Services: https://aws.amazon.com/es/big-data/data-lake-on-aws/?nc1=h_ls

Amazon Web Services. (2018). *Amazon Elastic Compute Cloud*. Obtenido de Amazon Web Services: https://docs.aws.amazon.com/es_es/AWSEC2/latest/UserGuide/concepts.html

Amazon Web Services. (2018). *Amazon Virtual Private Cloud*. Obtenido de Amazon Web Services: https://aws.amazon.com/es/vpc/

Amazon Web Services. (2018). *Cloudera EDH en AWS*. Obtenido de Amazon Web Services: https://aws.amazon.com/es/quickstart/architecture/cloudera/?nc1=h_ls

Apache Software Foundation. (2018). *Introducción*. Obtenido de Apache Ambari: https://ambari.apache.org/

Apache Software Foundation. (2018). *What is Apache Hadoop?* Obtenido de Apache Hadoop: http://hadoop.apache.org/

Banco Central de la República Argentina. (2017). *Medidas Adoptadas*. Obtenido de Banco Central de la República Argentina: http://www.bcra.gob.ar/Institucional/Medidas_adoptadas.asp

Banco Central de la República Argentina. (2018). *Banco Central de la República Argentina*. Obtenido de Política de Pagos: http://www.bcra.gob.ar/MediosPago/Politica_Pagos.asp

BBVA. (06 de Abril de 2018). *¿Qué es el fintech? Innovación en servicios financieros*. Obtenido de BBVA: https://www.bbva.com/es/que-es-el-fintech/

Bellé, A. (15 de Diciembre de 2016). *Por qué los proyectos de Big Data fallan: el éxito comienza en la definición de los objetivos*. Obtenido de itUser: http://www.ituser.es/opinion/2016/12/por-que-los-proyectos-de-big-data-fallan-el-exito-comienza-en-la-definicion-de-los-objetivos

BID y Finnovista. (2017). Innovaciones que no sabías que eran de América Latina y el Caribe. *Fintech*.

Bloomberg. (06 de Abril de 2018). *Company Overview of MapR Technologies, Inc.* Obtenido de Bloomberg: https://www.bloomberg.com/research/stocks/private/snapshot.asp?privcapId=139916695

BSA. (2015). ¿Por qué son tan importantes los datos? . *BSA The Software Alliance*.

Burgueño, C. (20 de Octubre de 2017). *Decisión definitiva: no habrá regulación del BCRA para las fintech*. Obtenido de ambito.com: http://www.ambito.com/900920-decision-definitiva-no-habra-regulacion-del-bcra-para-las-fintech

Cabello, V. N. (2010). *Introducción a las Bases de Datos Relacionales.* Madrid: Vision Libros.

Cano, J. L. (2007). *Business Intelligence: Competir con Información.*

Caralt, J. C., & Díaz, J. C. (2011). *Introducción al Business Intelligence.* Barcelona: Editorial UOC.

Cloudera. (05 de Abril de 2018). *About us*. Obtenido de Cloudera: https://www.cloudera.com/more/about.html

Cloudera. (05 de Abril de 2018). *Cloudera University*. Obtenido de Cloudera: https://www.cloudera.com/more/training.html

Cloudera. (2018). *Simple administration for Apache Hadoop*. Obtenido de Cloudera: https://www.cloudera.com/products/product-components/cloudera-manager.html

Conesa, J., & Curto, J. (2015). *¿Cómo crear un data warehouse?* Barcelona: Editorial UOC.

Estada, R., & Ruiz, I. (2016). *Big Data SMACK. A guide to Apache Spark, Mesos, Akka, Cassandra anda Kafka.* México City: Apress.

Esteso, M. P. (2018). *Fundamentos de Apache Hadoop y MapReduce*. Obtenido de Geeky Theory: https://geekytheory.com/fundamentos-de-apache-hadoop-y-mapreduce

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. *Big Data Analytics*.

Gartner. (2015). Big Data.

Gasalla, P. (5 de Junio de 2016). ¿Estamos caminando hacia un futuro sin bancos? *Diario El País*.

Gómez, J. M. (2013). *Bases de datos relacionales y modelado de datos.* España: Ediciones Parainfo.

Groenfeldt, T. (2013). Los bancos apuestan en grande por Big Data e información de los clientes en tiempo real. *Bloomberg Businessweek*.

Guarino, J. (06 de Abril de 2018). *La mitad de los bancos argentinos ya se asoció con una "fintech"*. Obtenido de Fintech Argentina: http://www.fintechargentina.com.ar/bancos/la-mitad-de-los-bancos-argentinos-ya-se-asocio-con-una-fintech/

Hoder, F., Wagner, M., Sguerra, J., & Bertol, G. (2016). La Revolución Fintech. *Oliver Wyman*.

Hortonworks. (2018). *Apache Hadoop HDFS*. Obtenido de Hortonworks: https://es.hortonworks.com/apache/hdfs/

Hortonworks. (06 de Abril de 2018). *Data Platform (HDP)*. Obtenido de Hortonworks : https://es.hortonworks.com/products/data-platforms/hdp/

Hortonworks. (06 de Abril de 2018). *DataFlow (HDF)*. Obtenido de Hortonworks: https://es.hortonworks.com/products/data-platforms/hdf/

Hortonworks. (05 de Abril de 2018). *Hortonworks Apache Hadoop y Certificaciones de Macrodatos*. Obtenido de Hortonworks: https://es.hortonworks.com/services/training/certification/

Hortonworks. (06 de Abril de 2018). *Sobre nosotros*. Obtenido de Hortonworks: https://es.hortonworks.com/about-us/quick-facts/

Humphries, M., Hawkins, M., & Dy, M. (1999). *Datawarehousing Architecture and Implementation.* New Jersey: Prentice Hall PTR.

IBM. (2015). *IBM Analytics*. Obtenido de IBM: http://www.ibm.com/software/data/bigdata/what-is-big-data.html

IBM. (2018). *¿Qué es Hadoop?* Obtenido de IBM: https://www-01.ibm.com/software/cl/data/infosphere/hadoop/que-es.html

IBM. (2018). *Apache Avro*. Obtenido de IBM: https://www.ibm.com/analytics/hadoop/avro

IBM. (03 de Marzo de 2018). *Bases de datos relacionales*. Obtenido de IBM Knowledge Center: https://www.ibm.com/support/knowledgecenter/es/SSEPGG_8.2.0/com.ibm.db2.udb.doc/admin/c0004099.htm

IBM. (2018). *Data Lake: Descubrimiento de hechos, patrones en datos e informes ad hoc*. Obtenido de IBM Analytics: https://www.ibm.com/analytics/data-management/data-lake

IBM. (18 de Marzo de 2018). *Esquema de Estrellas*. Obtenido de IBM Knowledge Center: https://www.ibm.com/support/knowledgecenter/SS9UM9_9.1.1/com.ibm.datatools.dimensional.ui.doc/topics/c_dm_star_schemas.html

IBM. (18 de Marzo de 2018). *Esquemas de copo de nieve*. Obtenido de IBM Knowledge Center: https://www.ibm.com/support/knowledgecenter/es/SS9UM9_9.1.1/com.ibm.datatools.dimensional.ui.doc/topics/c_dm_snowflake_schemas.html

IDC. (2017). The evolution of data trough 2025. *Data Age 2025*.

Igual, D. (2016). *Fintech: Lo que la tecnología hace por las finanzas.* Barcelona: Profit Editorial.

Inmon, W. H. (2005). *Building the Data Warehouse.* Indiana: Wiley Publishing.

Jain, V. K. (2017). Big Data & Hadoop. Khanna Publishing.

John, T., & Misra, P. (2017). *Data Lakes for Enterprises.* Birmingham: Packt Publishing.

Kimball, R., & Caserta, J. (2004). *The Data Warehouse ETL Toolkit.* Wiley Publishing.

KPMG. (2017). El nivel de madurez digital. *KPMG*.

Macario, A. (06 de Abril de 2018). *El blog de Andrés Macario*. Obtenido de El blog de Andrés Macario: https://andresmacario.com/hay-sitio-para-la-banca-en-la-transformacion-digital/

MapR. (06 de Abril de 2018). *A platform engineered for next-generation applications*. Obtenido de MapR: https://mapr.com/datasheets/mapr-converged-data-platform/

MapR. (06 de Abril de 2018). *MapR Services*. Obtenido de MapR: https://mapr.com/services/

MapR. (06 de Abril de 2018). *Productos MapR*. Obtenido de MapR: https://mapr.com/products/

Martínez, R. (06 de Abril de 2018). *¿Qué son y cuáles son las características de emprender a través de las empresas fintech en Latinoamérica?* Obtenido de Cámara Internacional de Emprendedores: https://www.cainem.com/que-son-y-cuales-son-las-caracteristicas-de-emprender-a-traves-de-las-empresas-fintech-en-latinoamerica/

Neo4j. (13 de Febrero de 2018). *Diez razones principales para elegir Neo4j*. Obtenido de Neo4j: https://neo4j.com

NoSQL. (10 de Febrero de 2018). *¡Su última guía para el universo no relacional!* Obtenido de NoSQL: http://nosql-database.org/

Oracle. (19 de April de 2016). *Un café con Oracle*. Obtenido de Oracle Blog: https://blogs.oracle.com/spain/qu-es-una-base-de-datos-nosql

Oracle. (06 de Abril de 2018). *La base de la innovación de datos*. Obtenido de Oracle: https://www.oracle.com/lad/big-data/index.html

Pasupuleti, P., & Purra, B. S. (2015). *Data Lake Development with Big Data.* Birmingham: Packt Publishing.

Polo, F. (28 de Julio de 2017). *Cuando el contexto impulsa al negocio*. Obtenido de Flux: https://medium.com/flux-it-thoughts/cuando-el-contexto-impulsa-al-negocio-9e95c0d36e5d

Ponniah, P. (2001). *Datawarehousing Fundamentals.* John Wiley & Sons.

PwC Argentina. (2018). *La influencia de las FinTech renueva la industria financiera en Argentina*. Obtenido de PwC Argentina: https://www.pwc.com.ar/es/prensa/la-influencia-fintech-renueva-industria-financiera-en-argentina.html

Python. (2018). *About*. Obtenido de Python: https://www.python.org/

Rayón, Á. (19 de Agosto de 2016). *Cuándo empieza esta era del Big Data: MapReduce*. Obtenido de Deuston Data: https://blogs.deusto.es/bigdata/cuando-empieza-esta-era-del-big-data-mapreduce/#comment-2026

Rivera, T. A. (05 de Septiembre de 2017). *¿Cómo diseñar una arquitectura Big Data y no morir en el intento?* Obtenido de Future Bites: https://bites.futurespace.es/2017/09/05/como-disenar-una-arquitectura-big-data-y-no-morir-en-el-intento/

Robinson, I., Webber, J., & Eifrem, E. (2015). *Graph Databases.* O'Reilly Media, Inc.

Schultze, J. F. (23 de Enero de 2018). Haciendo historia en el mundo Fintech hace 15 años. *La República*.

Taie, M. Z. (2015). *Hadoop Ecosystem: an Integrated Environment for Big Data*. Obtenido de Agroknow: http://blog.agroknow.com/?p=3810

Taylor, C. (8 de junio de 2017). *Big Data Architecture*. Obtenido de Datamation: https://www.datamation.com/big-data/big-data-architecture.html

Tomcy, J., & Pankaj, M. (2017). *Data Lake for Enterprises.* Birmingham: Packt Publishing.

Valleboni, C. (16 de Julio de 2017). *El crecimiento de las Fintech argentinas*. Obtenido de Forbes Argentina: http://www.forbesargentina.com/crecimiento-las-fintech-argentinas/

Viaña, E. (20 de Mayo de 2016). *El vértigo de usar el 'big data'*. Obtenido de Expansión: http://www.expansion.com/directivos/2015/05/20/555cd277268e3e30148b4581.html

Vives, X. (02 de Octubre de 2017). Fintech. *La Vanguardia*.

Wasson, M. (28 de Noviembre de 2017). *Big data architecture style*. Obtenido de Microsoft Azure: https://docs.microsoft.com/en-us/azure/architecture/guide/architecture-styles/big-data

White, T. (2012). Hadoop: The Definitive Guide. O'Reilly Media.