Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Facultad de Ingeniería de Sistemas e Informática

Escuela Profesional de Ingeniería de Sistemas

Integración de entorno big data en la gestión financiera de un sistema bancario para optimizar productos, servicios internos y toma de decisiones

TRABAJO DE SUFICIENCIA PROFESIONAL

Para optar el Título Profesional de Ingeniero de Sistemas

AUTOR

José Alberto HUAPAYA VÁSQUEZ

ASESOR

Norberto Ulises ROMÁN CONCHA

Lima, Perú

2021

Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

https://creativecommons.org/licenses/by-nc-sa/4.0/

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

Referencia bibliográfica

Huapaya, J. (2021). *Integración de entorno big data en la gestión financiera de un sistema bancario para optimizar productos, servicios internos y toma de decisiones*. [Trabajo de suficiencia profesional de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática, Escuela Profesional de Ingeniería de Sistemas]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	JOSÉ ALBERTO HUAPAYA VÁSQUEZ
DNI	70423018
URL de ORCID	https://orcid.org/0000-0001-8390-511X
Datos de asesor	
Nombres y apellidos	Ulises Roman Concha
DNI	08510560
URL de ORCID	https://orcid.org/0000-0002-3302-7539
Datos de investigación	
Línea de investigación	Data Science
Grupo de investigación	ITDATA
Agencia de financiamiento	Financiamiento Propio
Ubicación geográfica de lainvestigación	Perú, Lima, Lima, Santiago de Surco, Calle W 115, Urb. Santa Rosa de Surco II Latitud: -12.156277 Longitud: 76.970968
Año o rango de años en que serealizó la investigación	2021
URL de disciplinas OCDE	Ingeniería de sistemas y comunicaciones https://purl.org/pe-repo/ocde/ford#2.02.04

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
Escuela Profesional de Ingeniería de Sistemas

Acta Virtual de Sustentación
del Trabajo de Suficiencia Profesional

Siendo las 20:00 horas del día 30 de junio del año 2021, se reunieron virtualmente los docentes designados como Miembros de Jurado del Trabajo de Suficiencia Profesional, presidido por la Dra. Pró Concepción Luzmila Elisa (Presidente), Ing. Yañez Duran Carlos Enrique (Miembro) y el Lic. Román Concha Norberto Ulises (Miembro Asesor), usando la plataforma Meet (https://meet.google.com/xek-fqfj-kau), para la sustentación virtual del Trabajo de Suficiencia Profesional intitulado: ***"INTEGRACIÓN DE ENTORNO BIG DATA EN LA GESTIÓN FINANCIERA DE UN SISTEMA BANCARIO PARA OPTIMIZAR PRODUCTOS, SERVICIOS INTERNOS Y TOMA DE DECISIONES",*** *por el Bachiller* ***Huapaya Vásquez José Alberto****; para obtener el Título Profesional de Ingeniero de Sistemas.*

Acto seguido de la exposición del Trabajo de Suficiencia Profesional, el Presidente invitó al Bachiller a dar las respuestas a las preguntas establecidas por los miembros del Jurado.

El Bachiller en el curso de sus intervenciones demostró pleno dominio del tema, al responder con acierto y fluidez a las observaciones y preguntas formuladas por los señores miembros del Jurado.

Finalmente habiéndose efectuado la calificación correspondiente por los miembros del Jurado, el Bachiller obtuvo la nota de 18 (en letras) DIECIOCHO

A continuación el Presidente de Jurados la Dra. Pró Concepción Luzmila Elisa, declara al Bachiller ***Ingeniero de Sistemas.***

Siendo las 21:00...horas, se levantó la sesión.

Presidente
Dra. Pró Concepción Luzmila Elisa

Miembro
Ing. Yañez Duran Carlos Enrique

Miembro Asesor
Lic. Román Concha Norberto Ulises

FICHA CATALOGRÁFICA

INTEGRACIÓN DE ENTORNO BIG DATA EN GESTIÓN FINANCIERA PARA OPTIMIZAR PRODUCTOS / SERVICIOS Y TOMA DE DECISIONES

AUTOR: JOSÉ ALBERTO HUAPAYA VÁSQUEZ

ASESOR: ULISES ROMAN CONCHA

LIMA-PERÚ, 2021

Título profesional: Ingeniero de Sistemas

Línea de Investigación: Tecnologías de ciencia de datos.

Pregrado: Escuela Profesional de Ingeniería de Sistemas - Facultad de Ingeniería de Sistemas e Informática, UNMSM

Formato 28 x 20 cm Página x, 68

DEDICATORIA

Este trabajo se lo dedico a mi hermano Daniel, quien me motivó a culminar mis estudios profesionales y que continuó mis pasos estudiando Ingeniería de Software en nuestra UNMSM.

A mi hermano menor Fernando, que me enseña cada día que los sueños dependen de uno.

Y a mi novia Paola, que hoy en día es mi fortaleza para seguir adelante.

AGRADECIMIENTO

Agradezco principalmente a mis padres, que impulsaron desde siempre el logro de mis objetivos frente a todas las adversidades.

Agradezco también a la comisión de titulación que nos da esta oportunidad de demostrar nuestros conocimientos con el objetivo de culminar nuestros estudios profesionales.

De igual manera a Alonso Tejada por la oportunidad y confianza que me dio para integrar el equipo de Arquitectura de Datos y llevar esta transformación de la entidad financiera.

Finalmente quiero expresar mi mayor agradecimiento al profesor Ulises Román que con sus consejos y compromiso me apoyó en la realización del presente trabajo.

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

INTEGRACIÓN DE ENTORNO BIG DATA EN LA GESTIÓN FINANCIERA DE UN SISTEMA BANCARIO PARA OPTIMIZAR PRODUCTOS, SERVICIOS INTERNOS Y TOMA DE DECISIONES

Autor:	Huapaya Vásquez, José Alberto
Asesor:	Román Concha Ulises
Título:	Trabajo de Suficiencia Profesional
Fecha:	Julio 2021

RESUMEN

El presente informe de experiencia profesional describe la integración de un entorno Big Data en la gestión de productos y servicios de una entidad bancaria optimizando productos financieros y tomas de decisiones. Nos encontramos con una entidad bancaria donde sus distintas áreas de negocio tienen sistemas y bases de datos aisladas, provocando mayor consumo de recursos informáticos, mayor mantenibilidad y en muchos casos retraso de procesos. Este problema es común en entidades financieras y se vuelve crítico cuando nuestra empresa es transnacional, multiplicando estas bases de datos aisladas, como es nuestro caso.

Para todo ello, el área de Arquitectura de Datos planteó lineamientos como centralizar la información en un entorno big data asegurando la accesibilidad progresiva de los usuarios de negocio para nuevas iniciativas de analítica financiera y con ello ir apagando las bases de datos aisladas. El entorno big data comprende la capa de ingesta de datos en un Data Lakemediante el procesamiento distribuido de spark; y otra capa de consumo de información a través de sandboxes, donde los usuarios podrán realizar analítica avanzada y reportería con herramientas de inteligencia de negocios.

Como resultados de la integración, se tienen varios proyectos trabajando sobre este entorno Big Data, que están productivos con data gobernada, flujos definidos y con mejores tiempos de respuesta en procesos que anteriormente se desarrollaban en base de datos tradicionales. Con ello, las áreas de negocio tienen productos y servicios más eficientes.

Palabras claves: Big Data, Data Lake, Gestión Financiera, Toma de Decisiones, Integración, Optimización.

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

INTEGRATION OF BIG DATA ENVIRONMENT IN THE FINANCIAL MANAGEMENT OF A BANKING SYSTEM TO OPTIMIZE PRODUCTS, INTERNAL SERVICES AND DECISION-MAKING

Author:	Huapaya Vásquez, José Alberto
Adviser:	Román Concha Ulises
Title:	Trabajo de Suficiencia Profesional
Date:	Julio 2021

ABSTRACT

Thisprofessionalexperiencereport describes theintegrationof a Big Data environment in themanagementofproducts and servicesof a bankingentity, optimizingfinancialproducts and decisionmaking. Wefind a bankingentitywhereitsdifferentareashaveisolatedsystems and databases, causinggreaterconsumptionofcomputerresources, greatermaintainability and in many cases delayofprocesses. Thisproblemiscommon in financialinstitutions and becomescriticalwhenourcompanyistransnational, multiplyingtheseisolateddatabases, as in this case.

Forallthis, the Data Architectureareaproposedguidelinessuch as centralizinginformation in a big data environment, ensuringtheprogressiveaccessibilityofbusinessusersfor new financialanalyticsinitiatives and therebyshuttingdowntheisolateddatabases.The big data environmentcomprisesthelayerof data ingestion in a Data Lakethroughthedistributedprocessingof Spark; and anotherlayerofinformationconsumptionthroughsandboxes, whereusers can performadvancedanalytics and reportingwithbusinessintelligencetools.

As a resultoftheintegration, there are someprojectsworkingonthis Big Data environment, which are productivewithgoverned data, definedflows and withbetter response times in processesthatwerepreviouslydeveloped in traditionaldatabases. Withthis, thebusinessareashave more efficientproducts and services.

Keywords: Big Data, Data Lake, Financial Management, Decisionmaking, Integration, Optimization.

ÍNDICE

ÍNDICE FIGURAS x
ÍNDICE TABLAS xi
INTRODUCCIÓN 1
CAPÍTULO I: TRAYECTORIA PROFESIONAL 3
CAPÍTULO II: CONTEXTO EN EL QUE SE DESARROLLÓ LA EXPERIENCIA 8
2.1. EMPRESA - ACTIVIDAD QUE REALIZA 8
2.2. VISIÓN 9
2.3. MISIÓN 9
2.4. ORGANIZACIÓN DE LA EMPRESA 9
2.5. ÁREA, CARGO Y FUNCIONES DESEMPEÑADAS 10
2.6. EXPERIENCIA PROFESIONAL REALIZADA EN LA ORGANIZACIÓN 12
CAPÍTULO III: ACTIVIDADES DESARROLLADAS 14
3.1.1 DEFINICIÓN DEL PROBLEMA 14
3.2 SOLUCIÓN 15
3.2.1 OBJETIVOS 15
3.2.2 ALCANCE 16
3.2.3 ETAPAS Y METODOLOGÍA 18
3.2.4 FUNDAMENTOS UTILIZADOS 23
3.2.4.1 Entorno Big data 23
3.2.4.2 Gestión Financiera en el sistema bancario 30
3.2.4.3 Analítica de Datos Financieros y Toma de decisiones 32
3.2.5 IMPLEMENTACIÓN DE LAS ÁREAS, PROCESOS, SISTEMAS Y BUENAS PRÁCTICAS 32
3.2.5.1 Diagnóstico Geográfico 33
3.2.5.2 Formación de la masa crítica 39
3.2.5.3 Validación de Plataforma 43
3.2.5.4 Diseño y Desarrollo del Piloto 50
3.2.5.5 Validación del piloto 53
3.2.5.6 Despliegue 54
3.2.5.7 Reliability and Research 55

3.3 EVALUACIÓN ECONOMICA 58
3.3.1 EVALUACIÓN COSTO 58
3.3.2 BENEFICIO PARA LA ORGANIZACIÓN 60
CAPÍTULO IV: REFLEXIÓN CRÍTICA DE LA EXPERIENCIA 61
CONCLUSIONES Y RECOMENDACIONES 62
CONCLUSIONES 62
RECOMENDACIONES 63
FUENTES DE INFORMACIÓN 64
GLOSARIO 66
ANEXO 69

ÍNDICE FIGURAS

Figura 1: Organigrama de la empresa.....9
Figura 2: Integrantes del Organigrama de la Empresa.....10
Figura 3:Estructura Organizacional del Área de Ingeniería.....16
Figura 4:Presencia Geográfica de la Institución.....17
Figura 5:As Is del proceso de campañas de tarjeta de crédito.....33
Figura 6:As Is del proceso de campañas en grupos funcionales.....33
Figura 7:To Be del proceso de campañas de tarjeta de crédito.....34
Figura 8:Mejoras en el Modelo Proactivo de Personas Naturales.....34
Figura 9:Plan de transformación de la Entidad Financiera.....35
Figura 10:Esquema de Transcendence en el Plan de transformación.....35
Figura 11:Ejemplo de configuraciones para ingesta de Datos.....38
Figura 12:Herramientas de Gestión de la configuración.....38
Figura 13:Esquema Transcendence en el Plan de transformación.....39
Figura 14:Ceremonias Ágiles.....41
Figura 15:Distribución de la infraestructura.....41
Figura 16:Distribución de la infraestructura del Data Lake.....42
Figura 17:Repositorio de artefactos Nexus.....43
Figura 18:Diagrama general del entorno de procesamiento de datos 44
Figura 19:Jupyter y analítica avanzada.....45
Figura 20:Diseño de la distribución de Sandboxes.....45
Figura 21:Diseño de Componentes.....46
Figura 22:Buenas Prácticas de Gobierno.....47
Figura 23:Flujos de Ingestas.....48
Figura 24:Buenas Prácticas en Sandboxes.....48
Figura 25:Modelo de Propensión de TC.....49
Figura 26:Ejecución de Propensión de TC.....49
Figura 27:Variables cualitativas de validación.....51
Figura 28:Ejemplo de Seguimiento del entorno.....53
Figura 29:Herramientas de monitoreo del entorno.....54

ÍNDICE TABLAS

Tabla 1: Listado de experiencia profesional……..3
Tabla 2: Formación Académica Profesional……6
Tabla 3: Cursos y Eventos Académicos..……....6
Tabla 4: Otras Capacidades…...……..7
Tabla 5: Etapas de la Integración.……..…18
Tabla 6: Comparación de formatos de archivos de datos…….…27
Tabla 7: Evaluación de Diagnóstico Geográfico…..……..…31
Tabla 8: Formación de usuarios de negocio…...…..……..…37
Tabla 9: Formación de desarrolladores…...…..…….……....39
Tabla 10: Formación de masa crítica..…...…..…….……..…40
Tabla 11: Costo de RRHH del Piloto..…...…..…….……..…55
Tabla 12: Costo Mensual del proceso de campañas TC..…….……..…56

INTRODUCCIÓN

El presente trabajo de suficiencia profesional expone uno de los principales problemas que aquejan a muchas empresas en el país, que luego de solucionar sus problemas de negocio se encuentran con distintos sistemas y bases de datos no homologadas, poniendo en duda la veracidad y calidad de los datos.

En este trabajo el auto ha tenido la oportunidad de aplicar conceptos de gestión, experiencia de desarrollo de software, modelado de datos, marcos de trabajo agile, aplicando su experiencia en el uso de arquitectura de software y analizar requerimientos para soluciones de negocio; todo ello para participar en la integración de un entorno Big Data que soluciona los problemas anteriormente resumidos.

Las limitaciones que encontré fueron que las piezas se despliegan en servidores de México, con lo cual no tenemos información de primera mano y hay dependencia de operadores en dicho país. Adicional a ello los usuarios tienen mucha operatividad arraigada que los hace resistentes al cambio, lo cual es nuestra labor como informáticos darles todas ventajas y facilidades para que se involucren y se sientan satisfechos de las mejoras.

La integración de este entorno Big Data se aplicó usando el marco de trabajo agile SCRUM, con un equipo multidisciplinario agrupando a los propios usuarios de negocio como Data Scientist de la mano con los developers y arquitectos. Las soluciones que se montan sobre este entorno BIG DATA se basan en las ventajas de SPARK como motor de procesamiento y HDFS (Hadoop Distributed File System) como herramienta de almacenamiento.

El trabajo está estructurado de la siguiente manera: En la Sección I: Se describe la trayectoria profesional del autor del presente trabajo en las instituciones que he trabajado; en la Sección II: La experiencia se desarrolla en el contexto de una entidad financiera; en la Sección III: Se muestran las actividades desarrolladas del presente trabajo; en la Sección

IV: Se describe la reflexión crítica en base a la experiencia del autor y sus aportes en la integración de la solución; y finalmente presentamos las conclusiones y recomendaciones así como los trabajos futuros.

CAPÍTULO I:
TRAYECTORIA PROFESIONAL

El autor actualmente posee el grado de Bachiller en Ingeniería de Sistemas e Informática de la UNMSM; así mismo tiene habilidades en desarrollo de software, marcos agile (Scrum y Kanban), gestión y explotación de bases de datos (Data Scientist), conocimiento de Inteligencia de Negocios (Data Mart) y herramientas de procesamiento distribuido (Spark y HDFS).

De la misma manera cuenta con habilidades blandas para el trabajo colaborativo y resolución de problemas denotando la importancia de la calidad de servicio hacia el cliente.

La trayectoria profesional del autor se detalla a continuación:

Tabla 1. Listado de experiencia profesional

EXPERIENCIA PROFESIONAL				
Entidad	Área	Cargo	Fecha	Tiempo
BBVA	Engineering	Arquitecto de Datos	Setiembre 2017 - Actualidad	45 meses
	Funciones: ● Gestión de Arquitectura de Datos para Transformación digital con herramientas Big Data. ● Implementación de aprovisionamiento de data para Campañas de TC de en Arquitectura Big Data.			
EXPERIS	BI	Líder Técnico	Octubre 2016 - Septiembre 2017	12 meses
	Funciones: ● Análisis e Implementación de ETL para DataLake de Rímac Seguros (Siniestros, Productos SCTR) usando			

<table>
<tr><td></td><td colspan="4">InfoSphere DataStage 9.1, PL-SQL en Oracle 11G.</td></tr>
<tr><td rowspan="2">TECNOCOM</td><td>BI</td><td>Analista Técnico</td><td>Septiembre 2014 - Septiembre 2016</td><td>25 meses</td></tr>
<tr><td colspan="4">Funciones:
● Análisis e Implementación de ETLs para poblar el Data Lake de Rímac Seguros (Productos como Rímac Salud, Pólizas Vehiculares) usando InfoSphere DataStage 9.1 y PL-SQL en Oracle 11G.
● Modelado de datos para el Data Lake de Rímac Seguros (Pólizas y Siniestros) usando InfoSphere Data Architect (IBM).</td></tr>
<tr><td rowspan="2">GMD</td><td>Proyectos Evolutivos</td><td>Analista Programador</td><td>Abril 2013 - Septiembre 2014</td><td>18 meses</td></tr>
<tr><td colspan="4">Funciones:
● Optimización del sistema SACP (Sistema de Archivo Central de Planillas) dentro del Proyecto RIA SUB 3 para la ONP, bajo plataforma Java Web con arquitectura MVC, utilizando JSP, servelts, Oracle 11G y Servicios Web bajo metodología RUP.
● Mantenimiento del sistema SCA (Sistema de Cálculo Actuarial) para la estimación anual de pensiones y aportes a la ONP, bajo plataforma Java Web en una arquitectura MVC usando JSP, servlets, ORACLE 11G y con metodología RUP.
● Desarrollos Evolutivos del sistema FONAHPU WEB - módulo de “Calificaciones” y “Reportes” bajo arquitectura MVC en Java Web utilizando Strut, iReport, EJB, Oracle 11G y bajo la metodología RUP.
Logros:</td></tr>
</table>

<table>
<tr><td></td><td colspan="4">● Nominación al Círculo de Excelencia 2014 en GMD-Graña y Montero.</td></tr>
<tr><td rowspan="2">QUIPUCAMAYOC
UNMSM</td><td>Mantenimiento</td><td>Analista Programador</td><td>Junio 2011 - Marzo 2013</td><td>22 meses</td></tr>
<tr><td colspan="4">Funciones:
● Implementación de Módulo de Información Personal de la UNMSM
● Implementación de Módulo de “Planilla CAS” del sistema integral de RRHH de la UNMSM
● Implementación del Sistema Integral de Planificación (SIP) de la Oficina General de Planificación de la UNMSM, automatizando procesos de planificación y presupuesto de distintas unidades bajo el enfoque de RESULTADOS implantado por el Ministerio de Economía y Finanzas.
● Implementación del Cuadro Nominativo de Personal de la Oficina General de Recursos Humanos de la UNMSM, automatizando procesos de asignación de servidores públicos para las distintas plazas presupuestadas de la universidad apoyando a la generación de PAP (Presupuesto Analítico de Personal) que requiere el Ministerio de Economía y Finanzas.
● Implementación del Módulo de Cuadro de Asignación de Personal (CAP) de la Oficina General de Racionalización de la UNMSM.

(Todos con lenguaje de programación Java Web y arquitectura MVC, usando frameworks como Jsf2, Primefaces 3.4, Bootstrap, HTML 5, MyBatis 3.1/ iBatis 2.3, Spring 3.1, Spring Security 3.1 y iReport 5.0; bajo</td></tr>
</table>

	Scrum como marco de trabajo agile. Logros: ● Ascenso de practicante, pasando por desarrollador, hasta llegar a CAS como Analista Programador.

Tabla 2. Formación Académica Profesional

FORMACIÓN ACADÉMICA PROFESIONAL			
Formación Recibida	Institución que Acredita	Documento de acreditación	Fecha
Ingeniería de Sistemas - Facultad de Ingeniería de Sistemas e Informática	UNMSM	Diploma de Grado Académico de Bachiller	2007 - 2012

Tabla 3. Cursos y Eventos académicos

CURSOS Y EVENTOS ACADÉMICOS		
Curso	Institución	Año
Introducción a Machine Learning	Oxford	2021
Data Specialist Fundamentals	BBVA	2019
Big Data Analytics Summit	DMC	2018
Java 8.0 Advanced Developer	CIBERTEC	2016
Aplicaciones Java Web - Nivel Intermedio	GMD	2014
Power Builder - Nivel Intermedio	GMD	2014

Fundamentos SOA	GMD	2013
Frameworks de Desarrollo Java	GMD	2013
Desarrollos en PLSQL - Oracle 11g	GMD	2013
Desarrollo en SQL Server y Oracle 11g	GMD	2013

Tabla 4. Otras Capacidades

OTRAS CAPACIDADES	
Lenguajes de Programación	C++, PHP, JAVA, Powerbuilder.
Motores de Bases de Datos	Oracle, SQL Server, MySql, Postgres
Gestores de Bases de Datos	SQL Developer, Toad for Oracle, Toad for MySql, MySqlQueryBrowser, PL SQL Developer
Metodologías	RUP, Scrum, Kanban
Herramientas de Modelado	InfoSphere Data Architect (IBM), PowerDesigner, Erwin, Rational Rose
Herramientas de Desarrollo	Spring tool suite, Eclipse, Netbeans, JDeveloper, iReport, BIRT, SoapUI
Arquitecturas y Servidores	MVC, Cliente / Servidor, Tomcat, Oracle Application Server, PHPMyAdmin
Herramientas BI	InfoSphere DataStage 9.1
Sistemas Operativos	Windows, Linux, MacOs

CAPÍTULO II:
CONTEXTO EN EL QUE SE DESARROLLÓ LA EXPERIENCIA

2.1. EMPRESA - ACTIVIDAD QUE REALIZA

BBVA es una institución financiera subsidiaria bancaria con sede matriz en España, teniendo como actividades principales gestión de activos, depósitos, cartera de crédito, cajeros automáticos y sucursales de atención al cliente.

En Perú la entidad financiera se distribuye internamente en 9 áreas que se desarrolla a continuación:

- **C&IB Perú:** Se encarga de las inversiones corporativas que realiza la entidad financiera.
- **Desarrollo de Negocio**: Es el área encargada de centralizar y desarrollar los procesos y productos bancarios que la entidad financiera disponibiliza a todos los clientes.
- **Distribución Red**: Involucra a todas las redes de oficinas desplegadas en todo el país que son el primer frente hacia los clientes.
- **Medios**: Son los responsables de la comunicación y marketing hacia el cliente de acuerdo a los productos que despliegue el área de Desarrollo de Negocio.
- **Desarrollo Corporativo y Transformación**: Es el área que se encarga de mantener la infraestructura tecnológica de la entidad financiera, así como los responsables de la transformación digital.
- **Recursos Humanos**: Se encarga del bienestar y cumplimiento interno de los colaboradores en las distintas sedes de la institución financiera
- **Finanzas**: Son los responsables de mantener los lineamientos financieros de los productos.
- **Riesgos**: Esta área se encarga de evaluar financieramente a los clientes y posibles nuevos clientes.
- **Servicios Jurídicos y Cumplimiento**: Se encargan del cumplimiento legal de todos los procedimientos internos y externos de la entidad financiera.

Los datos de la entidad financiera con sede en Perú se detallan a continuación:

- **Razón Social**: BANCO BBVA PERÚ S.A.
- **Domicilio Legal**: AV. REP DE PANAMA NRO. 3055 URB. EL PALOMAR LIMA - LIMA - SAN ISIDRO
- **Teléfono**: 5(01)595-0000
- **RUC**: 20100130204

2.2. VISIÓN

"Ser el banco número uno del Perú, reconocido por las soluciones financieras y servicio al cliente para que tome mejores decisiones a través de una experiencia cómoda y fácil, así como dedicar esfuerzos al desarrollo de la responsabilidad Social, fomentar el conocimiento, la cultura e innovación"

2.3. MISIÓN

"Poner al alcance de todos los peruanos las oportunidades de esta nueva era, contribuyendo a la realización de sus sueños y proyectos"

2.4. ORGANIZACIÓN DE LA EMPRESA

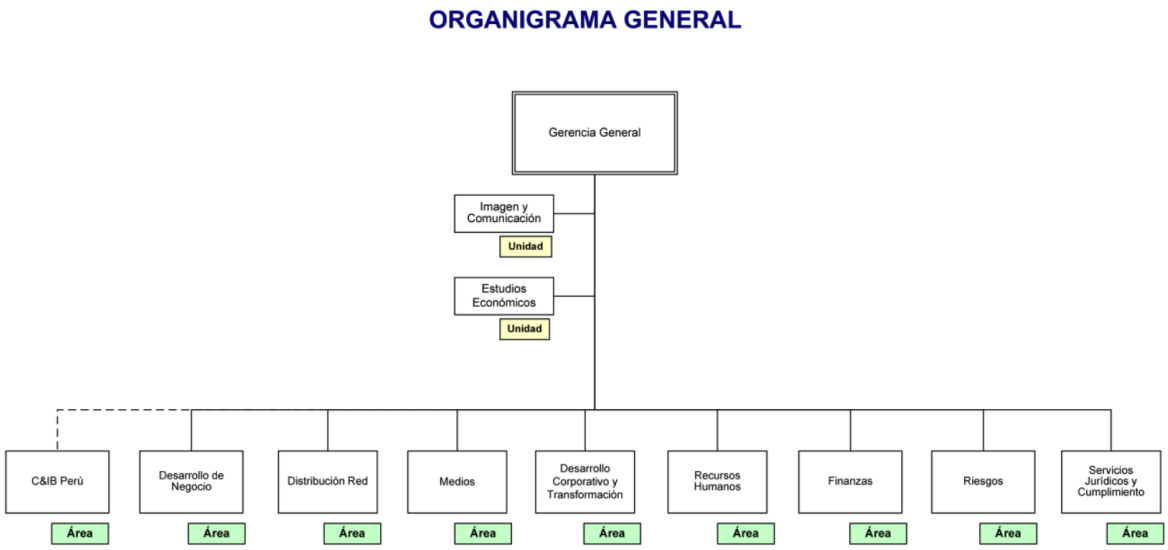


Figura 1. Organigrama general

Fernando Eguiluz Lozano Gerente General	**Gustavo Delgado-Aparicio Labarthe** Gerente de Banca Empresa y Corporativa	**Frank Erick Babarczy Rodríguez** Gerente de Corporate & Investment Banking
Luis Morales Espinosa Gerente de Client Solutions	**Marco Antonio Galdo Marín** Gerente de Banca Minorista	**Ignacio Férnandez-Palomero Morales** Gerente de Finanzas
Vicente Puig Payá Gerente de Riesgos	**María Guadalupe Pérez Suárez** Gerente de Engineering	**Ruth Anabelí González Velapatiño** Gerente de Servicios Jurídicos
Sandra Elba Bianco Roa Gerente de Talento y Cultura	**Enrique Medina García** Gerente de Transformation and Data	**Walter Ary Borra Nuñez** Auditor General (Participa en el comité de dirección como invitado)

Figura 2. Integrantes del organigrama de la empresa

fuente: www.bbva.pe

2.5. ÁREA, CARGO Y FUNCIONES DESEMPEÑADAS

El autor del presente informe de Experiencia Profesional se desempeñó como Arquitecto de Datos en el área de Arquitectura bajo la Gerencia de Engineering; esta área es la responsable de las soluciones tecnológicas para las necesidades de negocio de las otras áreas y/o gerencias.

Como Arquitecto de Datos somos responsables del entorno Big Data desplegado en Perú para el banco, incluyendo coordinaciones con arquitectos Globales (En México o España) en búsqueda de nuevas piezas tecnológicas que se puedan añadir funcionalidades o resolver incidencias en las piezas ya desplegadas, así como aterrizar flujos para el correcto desempeño de la arquitectura.

Funciones Realizadas en la presente entidad financiera:

- Gobernar los Datos dentro del Data Lake con los lineamientos y estándares corporativos.
- Montar pilotos en coordinación con el equipo de gestión de proyectos para el despliegue de nuevas piezas en el entorno big data.
- Analizar soluciones técnicas que involucren impacto o uso en datos.
- Dar apoyo al Reliability de la plataforma en la parte de arquitectura y piezas del entorno Big Data.
- Coordinar con otras disciplinas como seguridad, redes, backend, con el objetivo de resolver soluciones integrales.
- Coordinar con el holding de arquitectura de la entidad financiera por actualizaciones del roadmap de la plataforma, de esta manera apoyamos al área de Gestión de Proyectos en la priorización de iniciativas por recursos técnicos.
- Definir y monitorear KPIs estratégicos para Arquitectura de Datos, orientados al uso eficiente del entorno big data y al servicio hacia los usuarios finales.
- Disponibilizar flujos para el correcto uso y desempeño de los usuarios en el entorno big data.

En la organización interactúo principalmente con el área de Data Portfolio Manager en la coordinación de proyectos que usarán el entorno de Big Data, así como áreas de negocio que tengan iniciativas de procesamiento analítica que también podrían aprovechar la plataforma.

Otra área con quien Arquitectura de Datos interactúa es Data Engineering, quienes son los responsables de ejecutar los desarrollos en el entorno Big Data, aquí se encuentran los desarrolladores, analistas, modeladores, y el equipo de Reliability; estos últimos cumplen un rol vital en la estabilidad del entorno Big Data productivo, por ello como Arquitectos de Datos brindamos los lineamientos respectivos y acompañamiento en este proceso de transformación digital.

2.6. EXPERIENCIA PROFESIONAL REALIZADA EN LA ORGANIZACIÓN

La primera participación, del auto, en la entidad financiera fue ser parte del piloto de la integración del entorno Big Data en el cual tuve distintos roles que me ayudaron a reforzar los conceptos y actividades de un proyecto implementado en un entorno Big Data, estos roles los detallo a continuación:

- Analista: Identificar el origen de los datos que son parte del requerimiento de analítica, ya que estos deben provenir de un TDS (Trust Data Source).
- Desarrollador: Implementar transmisiones desde los TDS a la capa Staging del entorno Big Data.
- Desarrollador: Procesar la información transmitida para ingestarla al Data Lake de manera gobernada, esto implica que las estructuras están en un diccionario de datos y tanto los objetos como las estructuras tengan nomenclaturas estandarizadas del corporativo.
- QA: Validar las reglas establecidas por el usuario en la data ingestada en el Data Lake.
- Desarrollador: Desarrollo de procesamientos basados en la lógica de los usuarios de negocios para cumplir sus modelos analíticos.
- Gestor: Enmallar los procesamientos anteriores para que queden como procesos en batch.

Además, la experiencia del autor en la aseguradora Rímac Seguros contribuyó en la gestión de datos en esta nueva plataforma. En Rímac, mediante una empresa consultora de servicios TI, participé en otro proyecto de transformación digital en colaboración con IBM para construir un Data Lake que soporte todos los productos de Rímac Seguros y a su vez que se comunique con el Data Warehouse de la compañía.

Así como en los proyectos de desarrollo de software (previos a Rímac Seguros) también me ayudaron a comprender y adaptarme con practicidad en los ciclos de vida de los distintos componentes y automatizaciones del nuevo entorno de Big Data de la entidad financiera donde actualmente trabajo. Trabajé con herramientas de gestión de versionado,

lenguajes de programación, buenas prácticas de desarrollo de software, ETLs para Inteligencia de Negocios; todos estos conocimientos se aplican en el aprovisionamiento de cualquier Data Lake incluido el de un entorno Big Data.

CAPÍTULO III:
ACTIVIDADES DESARROLLADAS

3.1 SITUACIÓN PROBLEMÁTICA

Durante los últimos 10 años la entidad bancaria ha trabajado con una variedad de bases de datos tradicionales como Teradata, Oracle, Sql Server; con las cuales soportaba las necesidades de negocio de manera individual, y conforme aumentan los datos en la empresa aparecen oportunidades para explotar la información a través de modelos analíticos avanzados.

Estos modelos analíticos se continuaron trabajando bajo las bases de datos existentes, aquí fue donde los data scientist evidenciaron que los motores de procesamiento tradicionales eran muy lentos provocando que procesos críticos del banco tomen semanas o mes para su despliegue, por ejemplo: Procesos de Campañas de Tarjeta de Crédito, Evaluación de Riesgos Minoristas y Mayoristas.

A pesar de que algunas áreas trabajaban con modelos analíticos acorde con las nuevas tendencias, la competencia externa del banco entregaba productos y/o resultados más rápido, logrando fidelizar a sus clientes y en algunos casos aumentar su cartera con clientes de nuestra entidad financiera.

Adicional a este problema de rendimiento, aparece otro problema con la integridad del dato, que nos lleva a dudar de la veracidad de la información almacenada en las bases de datos dispersas, esto adquiere relevancia para la generación de KPIs estratégicos a partir del consumo de estos datos no centralizados y no confiables de las distintas áreas de negocio, implicando demoras en las tomas de decisiones.

3.1.1 DEFINICIÓN DEL PROBLEMA

No existe una integración adecuada de datos en la gestión financiera de la entidad bancaria, porque están dispersos, redundantes, soporte a la volumetría de datos y no cuentan con la veracidad correspondiente para una toma de decisiones óptima en la empresa. Es decir, no se tiene un gobierno del dato ni las herramientas para explotarlo de una manera óptima.

CAUSAS:

- La mayoría de áreas de negocio tiene procesos operativos muy arraigados a las herramientas tradicionales, lo que complica posibles migraciones.
- A partir de que la unidad de Ingeniería demoraba mucho en automatizar y/o desarrollar aplicativos, las propias unidades de negocio adquirían base de datos para soportar sus operaciones, lo que lleva a crear bases de datos locales o satélites.
- Con el tiempo el mantenimiento de la infraestructura de estas bases de datos satélites cae en responsabilidad de la unidad de ingeniería, lo que provoca gastos excesivos por la demanda de las distintas unidades de negocio.
- Los servicios internos, por parte del área de Ingeniería, son muy lentos y engorrosos, por lo que cada área de negocio tiene su propia sub área de TI, generando más sistemas y bases de datos aisladas, muy aparte de subsanar el problema de procedimientos lentos no se soluciona el problema de autonomía inexistente del usuario en el consumo de datos productivos, veraces y centralizados.

3.2 SOLUCIÓN

Integración de un entorno Big Data (basado en la tecnología de procesamiento Spark y también en la tecnología de almacenamiento HDFS) en una entidad bancaria para optimizar la generación de productos financieros, servicios internos y toma de decisiones.

3.2.1 OBJETIVOS

Objetivo General

- Integrar un entorno Big Data en la gestión financiera de un sistema bancario para optimizar productos, servicios internos y toma de decisiones.

Objetivos Específicos

- Validar y gobernar cada una de las piezas del entorno big data (tanto para analítica avanzada como para herramientas BI).
- Velar por los lineamientos del Gobierno del Dato para la centralización de información financiera en el Data Lake que es parte del entorno Big Data integrado.
- Diseñar soluciones de negocio usando las piezas integradas en el entorno big data e implantar pilotos de las piezas anteriormente recepcionadas para un afianzamiento del uso por parte de los usuarios y demostrar resultados confiables al negocio.
- Determinar los KPIs para la gestión del entorno Big Data.
- Diseñar flujos de atención de incidencias y requerimientos, con niveles de escalamiento respectivo.

3.2.2 ALCANCE

- Alcance Funcional

El alcance del proyecto de integración cubre inicialmente necesidades de analítica avanzada. Pero como uno de los objetivos es centralizar la información de la entidad bancaria en un Data Lake puede cubrir operativa no crítica del negocio (como reportería local con Dashboards gobernados) Ya que la información almacenada estará con una temporalidad D-1 imposibilitando procesos operativos críticos.

- Alcance Organizacional

La analítica avanzada es mayormente usada en las principales áreas de negocio como Riesgos, Finanzas, Comercial; siendo el área de Arquitectura (de la unidad de Engineering) el responsable técnico del entorno Big Data a integrar.

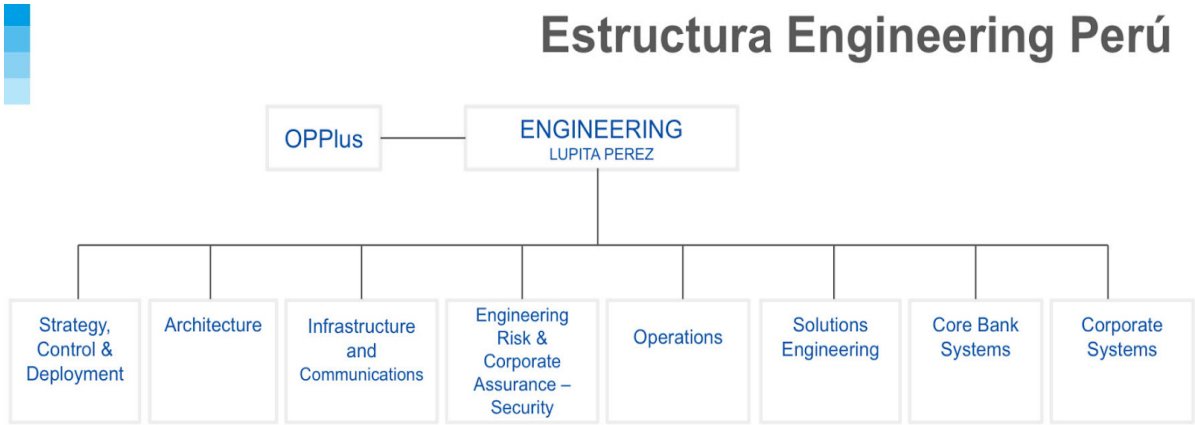


Figura 3. Estructura Organizacional del Área de Ingeniería

- ❑ Strategy Control & Deployment. - Encargados del Staff por roles asignado a los proyectos.
- ❑ Architecture. - Responsables de las piezas y soluciones Next-Gen.
- ❑ Infrastructure and Communications. - Responsables de las infraestructuras Next-Gen y Legacy.
- ❑ Engineering Risk & Corporate Assurance. - Security. - Responsable de soluciones asociadas a riesgos financieros.
- ❑ Operations. - Encargados de las transacciones financieras.
- ❑ Solutions Engineering. - Responsables de soluciones Legacy.
- ❑ Core Bank Systems. - Se encargan de implementar y mantener los sistemas internos del banco.
- ❑ Corporate Systems. - Se encarga de implementar y mantener los procedimientos tecnológicos.

- Alcance Geográfico

Al ser una entidad financiera transnacional (32 países), esta integración se realiza en 7 los países con las particularidades propias de cada país, pero manteniendo la comunicación bidireccional para feedback que podamos adaptar a nuestras realidades.

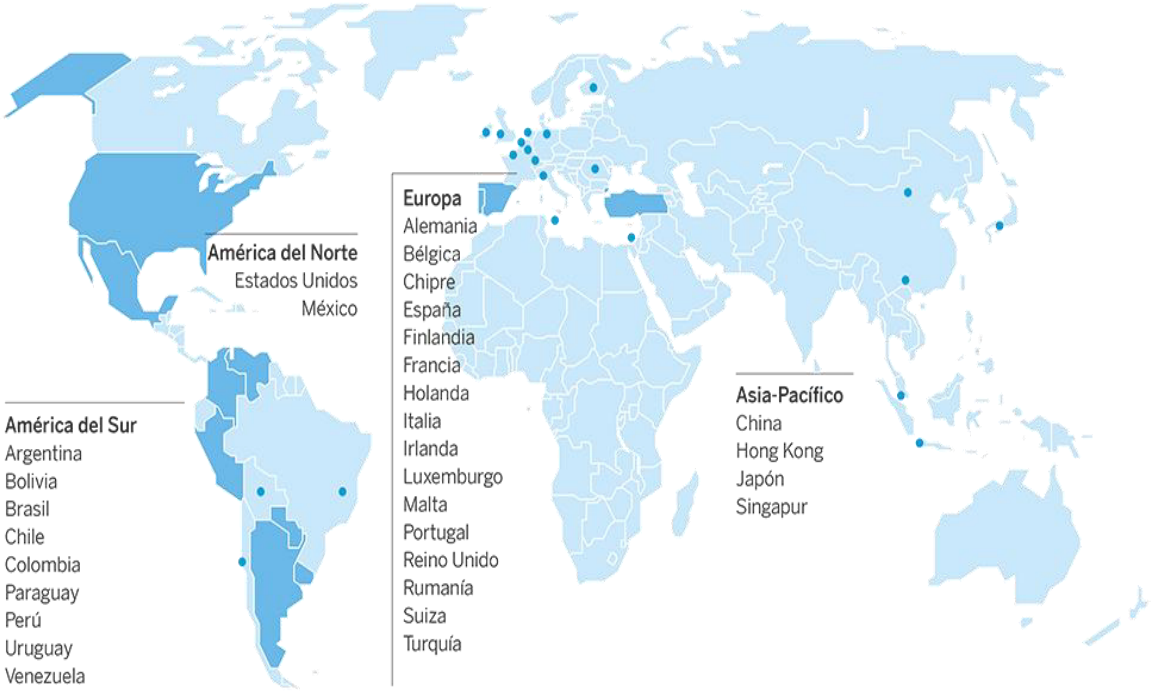


Figura 4. Presencia geográfica de la institución

Fuente: https://accionistaseinversores.bbva.com/microsites/cuentasanuales2018/informe-de-gestion/acerca-de-bbva

Los países con este entorno Big Data son:

España, USA, México, Colombia, Perú, Argentina, Uruguay. Todos ellos realizan analítica avanzada de la gestión financiera de la entidad bancaria y con ello optimizan productos.

3.2.3 ETAPAS Y METODOLOGÍA

La integración del entorno Big Data siguió las etapas detalladas en la Tabla 5 bajo el marco metodológico Agile Scrum, asegurando la **velocidad** del acceso a la información con un almacenamiento distribuido (HDFS), con formatos de archivos con sistema de compresión en el Data Lake asegurando los grandes volúmenes de datos (formatos AVRO y PARQUET), asegurando la variedad de la información con las transmisiones de distintos orígenes de datos a una zona staging del Data Lake, dándole valor a los datos ingestados mediante interfaces para procesamientos analíticos, y por último dándole veracidad a la información resultante con el gobierno del Data Lake (a través de estándares globales en la

entidad financiera)

Tabla 5. Etapas de la Integración

Fase	Tareas	Entregables	Técnica empleada
Diagnóstico Geográfico	Estudio y Evaluación del contexto de analítica avanzada de las principales áreas de negocio.	Propuesta de Proyectos Pilotos	-Sesiones con Gobierno de Proyectos y el negocio para que los pilotos estén alineados a la estrategia del banco. -Al ser parte del plan de transformación tecnológica de la entidad financiera, se involucró con ceremonias agile al negocio.
Formación de la masa crítica	Afianzar conocimiento de lenguajes de programación y herramientas Big Data. Proveedores para el desarrollo del piloto.	Lista de usuarios, desarrolladores y arquitectos a participar del piloto.	-Se buscó proveedores de desarrolladores con experiencia en Datos. -Se coordinó con el Holding de la transnacional para apoyo de España donde tenían experiencia de este tipo de integraciones.
Validación de plataforma	Validar las piezas de la plataforma a usar en el piloto.	V.B. por parte de Arquitectura en la usabilidad de las	-Se coordinó con el Holding para las liberaciones de las piezas en la infraestructura

		piezas.	asignada al país. -Se hicieron pruebas de concepto por pieza para usarlas posteriormente en el piloto.
Diseño y Desarrollo del Piloto	-Acondicionar el Data Lake. -Poblar el Data Lake -Diseñar el piloto (proceso analítico) por parte de los usuarios. -Implementar los notebook en el entorno	-Data lake con inputs necesarios para el piloto. -Gobierno del Data Lake. -Notebooks con el diseño del piloto. -Jobs productivos del piloto	-Coordinación constante con el Holding para la instalación y acondicionamiento del Data Lake. -Equipo de Arquitectura inmerso en el desarrollo del piloto de manera estratégica. -Acompañamiento constante con Soporte Holding para las puestas en producción.
Validación del Piloto	Consumo del Data Lake por parte del negocio.	Tiempos de respuesta antes y después del piloto.	-Los procesos analíticos son usados para el piloto de campañas de tarjetas de crédito, por ello los flujos del proceso de tarjeta de crédito tuvieron que cambiar. Se involucró al equipo de procesos y se tuvo acompañamiento técnico por parte de arquitectura.
Despliegue	-Masificar el conocimiento del	Flujos de uso.	En alianza con el área de gestión de proyectos se masificó

	nuevo entorno Big Data. -Involucrar a otras áreas usuarias de analítica avanzada.	Documentación.	el correcto uso del entorno y las ventajas de usar marcos ágiles con el usuario inmerso.
Reliability and Research	-Tener un soporte local. -Atender el gobierno del dato -Recepcionar nuevas piezas o mejoras en el entorno Big Data.	-Actualizaciones de flujos y documentaciones. -Herramientas locales de soporte. -Roadmap del entorno.	-Usar ceremonias ágiles para sincronizar estrategias de arquitectura y soporte. -Comunidades globales de arquitectura para herramientas de soporte y roadmap

En la fase de Diseño y Desarrollo del piloto se usaron como referencia conceptos de Pre Procesamiento de datos, Data Cleaning, Integración de datos y Reducción de Datos (Han y otros, 2012). Donde describe el tratamiento de datos que se pueden aplicar en proyectos de analítica avanzada como es el caso del piloto para la presente integración de un entorno Big Data, a continuación, describimos dichas fases:

- "Pre Procesamiento de datos. - Procesamiento previo con el fin de facilitar el trabajo de las siguientes etapas, solucionando problemas como registros de datos desactualizados, problemas en la puntualidad en la entrega de datos, información falsa provista, problemas en la interpretación de los datos, entre otros casos..."
- "Data Cleaning. - Busca "limpiar" los datos por medio de diversos procesos, entre estos están la eliminación de ruidos, el completado de valores faltantes y la corrección de inconsistencias..."
- "Integración de datos. - Luego de pre procesar los datos debemos incluir los datos a la estructura que vayamos a usar...

…Aparte de analizar la redundancia, debemos ver si hay duplicaciones de tuplas. Dentro del mundo de las bases de datos esto puede ser causado por tener tablas no normalizadas correctamente, o también por el ingreso incorrecto de los datos...”

- “Reducción de Datos. -Conjunto de técnicas que permiten el entendimiento y aumento de eficiencia en la representación de los datos cuando estos resultan altamente complejos en tamaño y significado, manteniendo en sí la estructura inicial...”

Adicionalmente, las fases de Validación de plataforma, Despliegue y Reliability and Research, descritas en la Tabla 5, pertenecen a las cualidades que se debe conseguir en una arquitectura Big Data, según Quiroz y otros (2019), las cuales son contar con un procesamiento distribuido de datos, ser escalable tanto en software como en hardware, estar siempre disponibles y que distribuya un gran volumen de datos. En este artículo también propone 3 etapas para la implementación de la Arquitectura Big Data:

- “Obtención de datos de diferentes fuentes”: Comprende la identificación de datos relevantes, estructurados y no estructurados; así como procesos de extracción, transformación y carga.
- “Procesamiento de datos en tiempo real”: Implica tener hardware para un adecuado procesamiento y visualización de datos; así como Spark y Hadoop como herramientas de procesamiento y almacenamiento respectivamente.
- “Análisis, visualización y toma de decisiones”: Lenguajes de programación asociados a Spark; así como herramientas para Inteligencia de Negocios y aprendizaje automático (Machine Learning)

Estas etapas se alinean a las etapas propuestas en la tabla 5 del presente informe, con la diferencia de que nuestra entidad financiera concretó la integración del entorno big data con un piloto funcional del área de negocio de Clients Solutions, un área con productos significativos como el perfilamiento de clientes para las campañas de tarjetas de crédito.

3.2.4 FUNDAMENTOS UTILIZADOS

3.2.4.1 Entorno Big data

- **Características de Big Data**

  Según Camargo-Vega y otros (2014), en la práctica la implementación del Big Data debe cumplir tres características o dimensiones básicas:

  - **Volumen**: "cada día, las empresas registran un aumento significativo de sus datos (terabytes, petabytes y exabytes), creados por personas y máquinas. En el año 2000 se generaron 800.000 petabytes (PB), de datos almacenados y se espera que esta cifra alcance los 35 zettabytes (ZB) en el 2020. Las redes sociales también generan datos, es el caso de Twitter, que por sí sola genera más de 7 terabytes (TB) diariamente, y de Facebook, 10 TB de datos cada día. Algunas empresas generan terabytes de datos cada hora de cada día del año, es decir, las empresas están inundadas de datos…"

  - **Variedad**: "Se puede mencionar que va muy de la mano con el volumen, pues de acuerdo con éste y con el desarrollo de la tecnología, existen muchas formas de representar los datos; es el caso de datos estructurados y no estructurados; estos últimos son los que se generan desde páginas web, archivos de búsquedas, redes sociales, foros, correos electrónicos o producto de sensores en diferentes actividades de las personas; un ejemplo es el convertir 350 mil millones de lecturas de los medidores por año para predecir el consumo de energía…"

  - **Velocidad**: "Se refiere a la velocidad con que se crean los datos, que es la medida en que aumentan los productos de desarrollos de software (páginas web, archivos de búsquedas, redes sociales, foros, correos electrónicos, entre otros) ..."

  Según Puyol (2014) indica que desde que salieron esas 3 V's propuestas por Doug Laney de la consultora Gartner en 2002, muchos han tratado de describir hasta 11 V's para el Big Data, pero en su artículo "Una Aproximación a Big Data" describe 2

V's muy relevantes adicionales a las 3 descritas en el párrafo anterior haciendo un total de 5 características que debe cumplir un proyecto de Big Data:

- **Valor**: "en el contexto de Big Data, valor hacer referencia a los beneficios que se desprenden del uso de Big Data (reducción de costes, eficiencia operativa, mejoras de negocio) ..."

- **Veracidad**: "la incertidumbre de los datos. La veracidad hace referencia al nivel de fiabilidad asociado a ciertos tipos de datos. Esforzarse por conseguir unos datos de alta calidad es un requisito importante y un reto fundamental de big data, pero incluso los mejores métodos de limpieza de datos no pueden eliminar la imprevisibilidad inherente de algunos datos, como el tiempo, la economía o las futuras decisiones de compra de un cliente. La necesidad de reconocer y planificar la incertidumbre es una dimensión de Big Data que surge a medida que los directivos intentan comprender mejor el mundo incierto que les rodea…"

También BBVA (2020), en portal de Api Market BBVA considera como características del Big Data a 7 Vs: Volumen, Velocidad, Variedad, Valor, Veracidad

- **La variabilidad**. - "hace referencia a la variabilidad en el significado, en el léxico. Esto es relevante a la hora de llevar a cabo análisis de percepciones. Los algoritmos deben ser capaces de comprender el contexto y descifrar el significado exacto de cada palabra en su respectivo entorno. Este análisis semántico resulta mucho más complejo…"

- **La visualización**. - "es lograr que toda la cantidad de datos recolectados y analizados sean comprensibles y sencillos de leer. Sin una visualización adecuada, no se puede sacar el máximo rendimiento y aprovechamiento de los datos en bruto…"

Así mismo otras empresas de tecnología consideran características importantes en el proceso de Big Data:

IBM: 4vs (Volumen Velocidad, variedad, Veracidad)

ORACLE: 5vs (Volumen Velocidad, variedad, Veracidad, Valor)

- **Arquitectura de Componentes Big Data**

  - **Sandbox**: Según Wright y otros (2006) para desarrollar analítica define: "Es un entorno de trabajo de concepto flexible que admite tareas analíticas tanto para una entidad formal como para tareas puntuales. Este componente recopila pruebas para elaborar capacidades innovadoras de la interacción de la información humana. También soporta escalabilidad para grandes tareas de análisis…"

    En nuestro caso, orientamos el Sandbox a un ambiente de trabajo con almacenamiento distribuido y capacidad de procesamientos en lenguajes de programación para analítica avanzada, donde el Data Scientist se desenvuelva en sus modelos analíticos.

  - **Data Lake**: Como comenta Miloslavskaya y Tolstoy (2016), para extraer datos, procesarlos y posteriormente almacenarlos, es necesario tener en cuenta que un Data Lake o Data Hub "hace referencia a repositorios de almacenamiento masivo y escalable que albergan data cruda o nativa necesaria para procesamientos de datos. En este repositorio se puede ingestar cualquier tipo de información sin comprometer la estructura de los datos, a diferencia de los data warehouses que depende de data estructurada. Los Data Lake también se caracterizan por ser construidos para manipular grandes volúmenes de información y de manera rápida. Los Data Lake a menudo incluyen una base de datos semántica, un modelo conceptual que aprovecha los mismos estándares y tecnologías utilizados para crear hipervínculos de Internet y agregar una capa de contexto sobre los datos que define el significado de los mismos y sus interrelaciones con otros datos. Las estrategias del Data Lake pueden combinar enfoques de bases de datos SQL y NoSQL y procesamiento de análisis en línea (OLAP) y Capacidades de procesamiento de transacciones en línea (OLTP)..."

    En nuestro caso, hay lineamientos corporativos para los estándares de los objetos y atributos en el Data Lake, asociados a los conceptos funcional de la

entidad financiera, lo que viene a ser el modelo conceptual para Miloslavskaya y Tolstoy (2016).

- **Ingesta de Datos**: Según el artículo de Erraissi y otros (2018), La ingesta de datos es un componente de carga de información relevante y limpia en la capa de almacenamiento distribuida de Hadoop. "Debe tener la capacidad de validar, limpiar, transformar, reducir, e integrar datos en el sistema de Big Data para su posterior procesamiento"

- **Tecnologías de procesamiento y almacenamiento de Big Data**

  - **HADOOP**. - La biblioteca Hadoop utiliza modelos de programación simples para el almacenamiento y procesamiento distribuido de grandes conjuntos de datos en clusters, dando redundancia para no perder nada y, al mismo tiempo, aprovechando muchos procesos a la vez.

    Dispone de un sistema de archivos distribuido en cada nodo del cluster: el HDFS (Hadoop Distributed File System), y se basa en el proceso de MapReduce de dos fases.

    En el presente trabajo se expone el uso exclusivo de HDFS para el entorno Big Data.

  - **APACHE SPARK**. - Considerado el primer software open source que hace la programación distribuida (muy en esencia, consiste en distribuir el trabajo entre un grupo de ordenadores, "cluster", que trabajan como uno) realmente accesible a los científicos de datos.

    Se pueden programar aplicaciones usando diferentes lenguajes como Java, Scala, Python o R. pudiendo ser, según el programa, hasta 100 veces más rápido en memoria o 10 veces más en disco que Hadoop MapReduce.

  - **PYTHON**. - Es una herramienta para Big Data muy eficiente, en parte debido a la gran comunidad existente, por lo que Python dispone de muchas librerías ya hechas por otros usuarios.

Sin embargo, tiene en su contra que no es un lenguaje muy rápido en su ejecución, por lo que suele ser empleado para tareas de integración o tareas donde no haya cálculos pesados.

- **SCALA**. - Lenguaje de programación que combina la programación funcional con la programación orientada a objetos. Por su practicidad de implementación se considera un lenguaje conciso de alto nivel. Permite construir aplicaciones de alto rendimiento por su ecosistema de librerías y su comunicación con Java Virtual Machine y JavaScript, muy usados en desarrollos de software.

  Para los efectos del presente trabajo, este lenguaje de programación tiene un excelente performance trabajando sobre Spark, lo cual llevó a ser usado en el entorno Big Data para el desarrollo de motores de procesamientos que alimentan a los modelos analíticos.

- **AVRO**: Como indican Plase y otros (2017) en su comparativo de formatos de archivos, y en línea de lo que manifiesta la empresa Apache (responsable de Spark), lo resume como: "...es un formato de almacenamiento basado en filas, también descrito como un sistema de serialización de datos similar a la serialización de Java. Avro proporciona estructuras de datos enriquecidas, un formato de datos binarios, compactos y de rápido acceso. No requiere la generación de código para leer o escribir archivos de datos. Además, Avro puede proporcionar un rendimiento de tiempo de ejecución más optimizado..."

- **PARQUET**: Plase y otros (2017) también nos comentan sobre PARQUET, en un análisis a lo mencionado por la empresa Apache lo siguiente: "...Es un formato de almacenamiento basado en columnas, optimizado para trabajar con conjunto de datos de varias columnas. Los casos de uso de Parquet generalmente implican trabajar con un subconjunto de esas columnas en lugar de registros completos. Una de las ventajas más citadas del ordenamiento columnar es que mejora el rendimiento de las consultas analíticas. Los algoritmos de compresión de datos funcionan mejor en datos con baja

entropía de información (Datos con gran valor local). Por lo tanto, el sistema de archivos logra los beneficios en lectura y escritura por rendimiento en compresión sin aumentar la carga de la CPU durante la descompresión. El diseño de los archivos de datos de Parquet está optimizado para consultas que procesan grandes volúmenes de datos."

Tabla 6. Comparación de formatos de archivos de datos

File format	Schema integration	Compression support
Text/CSV (Shafranovich 2005)	–	–
JSON (Bray 2014)	+	–
Avro (Apache 2009a)	+	+
SequenceFile (Apache 2009b)	–	+
RCFile (He *et al*. 2011)	–	+
ORC file (Apache 2017)	–	+
Parquet (Apache 2013)	+	+

Fuente: (Place, 2017)

- **Gestión de Big Data**

  - **Reliability**. - Nachiappan y otros (2017) describen al Reliability o Confiabilidad de la siguiente manera: "...las fallas son inevitables en los sistemas de almacenamiento en la nube, ya que están compuestos por componentes de hardware a gran escala. Mejorar la tolerancia a fallas en los sistemas de almacenamiento en la nube para aplicaciones de Big Data es un desafío importante. La codificación de replicación y borrado son las técnicas de confiabilidad de datos más importantes empleadas en los sistemas de almacenamiento en la nube. Ambas técnicas tienen su propia compensación en varios parámetros como durabilidad, disponibilidad, sobrecarga de almacenamiento, ancho de banda y tráfico de la red, consumo de energía y rendimiento de recuperación…"

  - **Data Scientist**. - Wang y otros (2021) resumen el concepto del científico de datos como: "...usamos el término 'Data Scientist' para referirnos a las personas que codifican un modelo de compilación/analítico, y usamos 'practicantes de DS / ML' para referirnos colectivamente a todas las personas que participan en un proyecto, incluyendo estos roles: Data Scientists, Engineers, Designers y Domain Experts... "

Estos roles que usan Wang y otros (2021) en IBM se aplican en la mayoría de entornos analíticos y el presente trabajo no es la excepción. Para ello la experiencia en dichos roles son la clave en el desarrollo de los proyectos sobre el entorno Big Data.

- **Infraestructura de Big Data**
    - **Cluster**. - Según Galeano y Peña (2019) en la aplicación de Big Data en un ámbito financiero, define Cluster con lo siguiente: "...Esta forma de trabajos muestra su potencia cuando se conectan varios ordenadores a la web, en general de forma remota sin proximidad física, formando un clúster, es decir un grupo de ordenadores conectados entre sí mediante un sistema de red de alta velocidad. Además, debe existir un programa que controle la distribución de la carga de trabajo entre los equipos. Por lo general, este tipo de sistemas cuentan con un centro de almacenamiento de datos único..."

3.2.4.2 Gestión Financiera en el sistema bancario

- **Detección de Fraude**. - Galeano y Peña (2019) indica que el aumento de operaciones digitales ha generado que los delincuentes encuentren nuevas formas de cometer fraudes, para ello detalla que: "...Los dos tipos de fraude financiero más importantes en la actualidad son el fraude a través de compras con tarjeta de crédito y mediante la utilización fraudulenta de la cuenta bancaria. La utilización de las tarjetas de crédito ha crecido mucho en los últimos tiempos y también el fraude, aunque el uso creciente de técnicas de prevención y detección parecen haber detenido ese avance: El Banco Central Europeo en su último informe de 2018 (https:// www.ecb.europa.eu/pub/cardfraud/html/ecb.cardfraudreport201809.en.html), estima que este fraude ha sido en 2016 de 1,8 billones de euros en el Single Euro PaymentsArea. El fraude estuvo aumentado fuertemente desde 2012 a 2015, pero se ha estabilizado en los últimos años, representando un crecimiento del 14,8 por 100 comparado con el de 2011. Respecto a la composición del fraude, el más frecuente (73 por 100) corresponde a pagos no presenciales realizados por Internet o teléfono, 19 por 100 a pagos con la tarjeta en la venta y 8 por 100 a transacciones al sacar dinero con la tarjeta..."

- **Red de Cliente**. - Como comenta Galeano y Peña (2019) la red de clientes en un banco es "...una fuente de información muy importante en varios aspectos. En primer lugar, es una pieza fundamental para entender la importancia de sus usuarios analizando su posición en dicha red. Tradicionalmente, se ha tendido a considerar que la importancia de un cliente se mide a partir de sus recursos económicos. Esta medida ignora su importancia estratégica dentro de la institución y su posición y relaciones en el conjunto de la red de clientes. La importancia de un cliente depende, claramente, de las consecuencias para el banco de que deje de serlo…"

- **Predicción del abandono de clientes. -** Para Galeano y Peña (2019) la predicción de abandono es una variable muy significativa que debe tener toda entidad financiera y lo sustenta de esta manera: "...Debido al actual alto grado de competencia en todos los sectores, tan importante o más que intentar ganar clientes es retener los actuales. Su pérdida implica la ganancia de otros, y nuestra disminución de cuota de mercado. Además, un cliente enfadado difunde su descontento en redes sociales con mayor probabilidad que uno satisfecho, con efectos muy negativos sobre la imagen de la empresa. El término en inglés*churnrate*se refiere a la tasa de pérdida de clientes de una compañía o institución…"

- **Predicciones personalizadas y geolocalización. -**Según Galeano y Peña (2019) la personalización cobra cada día mayor relevancia , más aún cuando tenemos mucha más información de nuestros clientes y esto lo describe así : "...El nuevo paradigma del Big Data está llevando a muchas compañías e instituciones a utilizar técnicas para personalizar sus predicciones. Por ejemplo, cada vez es más frecuente que un banco decida promocionar un producto a un determinado espectro de la población que tenga unas característicasespecíficas, como una hipoteca con unas condiciones ventajosas a todos sus clientes de entre 25 y 45 años, ya que en ese rango de edades es cuando la gran mayoría de personas decide comprarse una vivienda. Sin embargo, si solo se tiene en cuenta la edad es posible que la oferta publicitaria recibida por el cliente no sea de su interés, perdiendo la oportunidad de ofrecerle un producto más adecuado..."

3.2.4.3 Analítica de Datos Financieros y Toma de decisiones

- **Modelo Estadístico:** Hoy en día la competencia obliga a que conceptos como predicción, propensión, recomendación, etc., sean el día a día, y de ello no se eximen las entidades financieras. SegúnHan y otros (2012) los conceptos anteriormente mencionados se engloban en otro llamado 'Modelos Estadísticos' que son "...un conjunto de funciones matemáticas que describen el comportamiento de objetos en una clase objetivo en términos de variables aleatorias y sus distribuciones de probabilidad asociada. Los modelos estadísticos son ampliamente usados para modelos y clases de datos...".

- **Toma decisiones basadas en datos**: o DecisionMaking, es la aplicación práctica del conocimiento obtenido a partir de reportes o informes de resultados del análisis big data. Este concepto se aplica luego de haber realizado la extracción de valor y el análisis de los datos. Por ejemplo, en un contexto de cambios en los patrones de compra de los consumidores, la toma de decisiones consistirá en actuar para adaptarse a dichos cambios y anticiparse a la materialización de las tendencias previstas.

- **Inteligencia de negocios**: o Business Intelligence, es el conjunto de métodos para mejorar la toma de decisiones de negocio mediante el uso de sistemas y/o herramientas de apoyo basados en hechos. Su implementación se apoya sobre modelos dimensionales que se montan en Data Warehouse o Data Marts.

3.2.5 IMPLEMENTACIÓN DE LAS ÁREAS, PROCESOS, SISTEMAS Y BUENAS PRÁCTICAS

La implementación de la integración del entorno de BIG DATA se inició con un piloto en el área de Desarrollo de Negocio (área comercial de productos del banco); como se detalla en las siguientes etapas, primero determinando el alcance del piloto, pasando por la validación de herramientas hasta del despliegue de los productos que validan la integración del entorno y lo vuelve disponible para otros proyectos analíticos.

3.2.5.1 Diagnóstico Geográfico

El equipo conformado por 16 personas hicimos el estudio y evaluación del contexto de analítica avanzada de las principales áreas de negocio que se muestran a continuación:

Tabla 7. Evaluación de diagnóstico Geográfico

Evaluación de diagnóstico Geográfico		
Área	Resultados de la evaluación	Consideraciones
Finanzas	No requerían analítica avanzada, ya que contienen cargar operativas para reportería controlada hasta el momento. Esta área provee información al área de "Desarrollo Comercial" para evaluar nuevos productos.	Tienen SQL Server, MainFrame - Host
Riesgos	Actualmente tienen una base de datos Oracle dedicada para sus procesos analíticosu operativos. Algunos integrantes de este equipo participan en el proceso de Campañas de Tarjetas de Crédito de la otra área "Desarrollo Comercial" en la evaluación de riesgos crediticios.	Tienen Oracle, SQL Server, SAS.
Talento y Cultura	Básicamente las necesidades son de Inteligencia de Negocio y reportería básica. Así que cuando tengamos componentes BI, se priorizará.	Usan Data Studio, Office, Power BI
Engineering	No hay prioridad de analítica avanzada.	Soporte a los usuarios.

C&IB	Por el momento están enfocados en la reportería, su área de analítica está en desarrollo.	Usan Data Studio, Office
Servicios Jurídicos	Priorizar la madurez del entorno para implementar sus procedimientos.	No requiere
Desarrollo de Negocio	Tienen en cartera muchos modelos analíticos que podrían usar este entorno Big Data, algunos de ellos requieren inputs de otra área de negocio incrementando la manualidad en dichos procesos y con las herramientas actuales el performance no es el indicado para la respuesta al cliente.	Tienen Oracle, SQL Server, SAS.

El área de implementación adoptada es Desarrollo de Negocio (o área comercial), por tener varios modelos analíticos para generar productos financieros y por interactuar con otras áreas de negocio, gracias a estas dependencias con las áreas de Finanzas y Riesgos se conformó un equipo multidisciplinario con distintos roles como Data Scientist, Data Developers, Arquitectos de Datos, Gestores de Proyectos y Agile Coaches.

El piloto elegido consistió en migrar todo el proceso mensual de campañas de tarjeta de crédito para personas naturales. El cual hasta ese momento tomaba alrededor de 35 días desde su concepción hasta la llegada de la oferta de tarjeta de crédito a cada cliente.

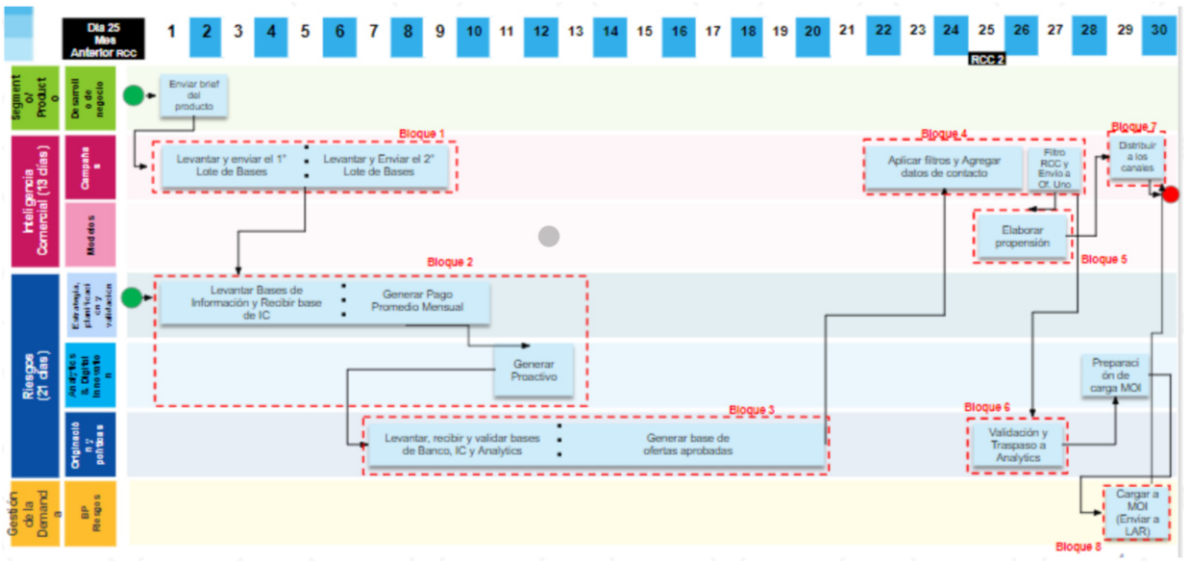


Figura 5. As Is del proceso de campañas de tarjeta de crédito

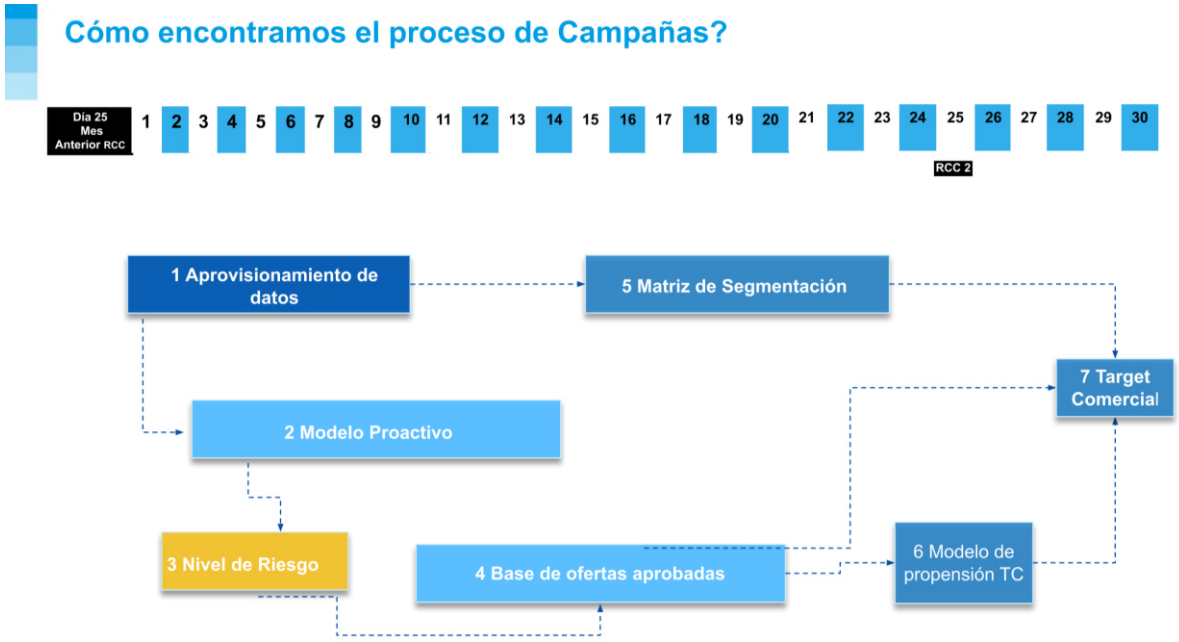


Figura 6. As Is del proceso de campañas en grupos funcionales.

Aprovechando el procesamiento y almacenamiento distribuido de una plataforma Big Data se pone el reto de reducir dicho proceso a 5 días.

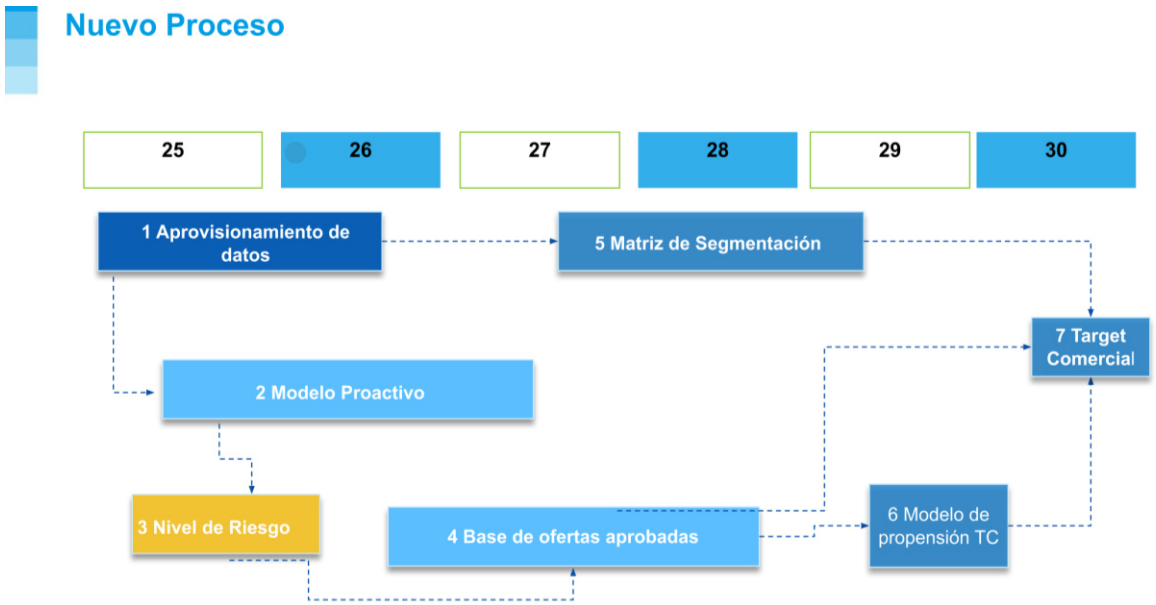


Figura 7. To Be del proceso de campañas de tarjeta de crédito.

Adicional a ello, se plantea enriquecer los modelos analíticos con fuentes de información más confiables, un ejemplo de ello es el modelo Proactivo de Personas Naturales que realizaron los Data Scientist del área de Riesgos. En este proceso se evalúan riesgos en clientes, dan puntuación y estiman cuotas de crédito.

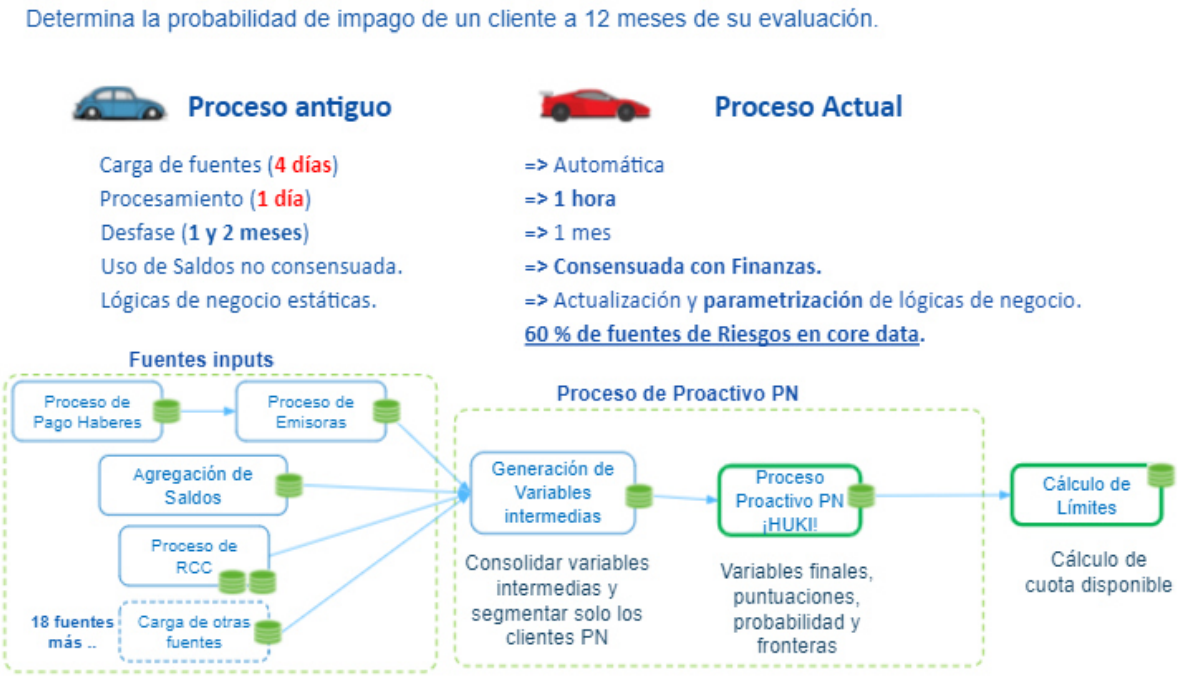


Figura 8. Mejoras en el Modelo Proactivo de Personas Naturales.

Y, por último, para alinearnos al plan de transformación de datos en la empresa, tanto el piloto como el Modelo de Gobierno, son el inicio del plan de transformación. Esta iniciativa es promovida por las áreas de Gestión de Proyectos, Engineering y Arquitectura.

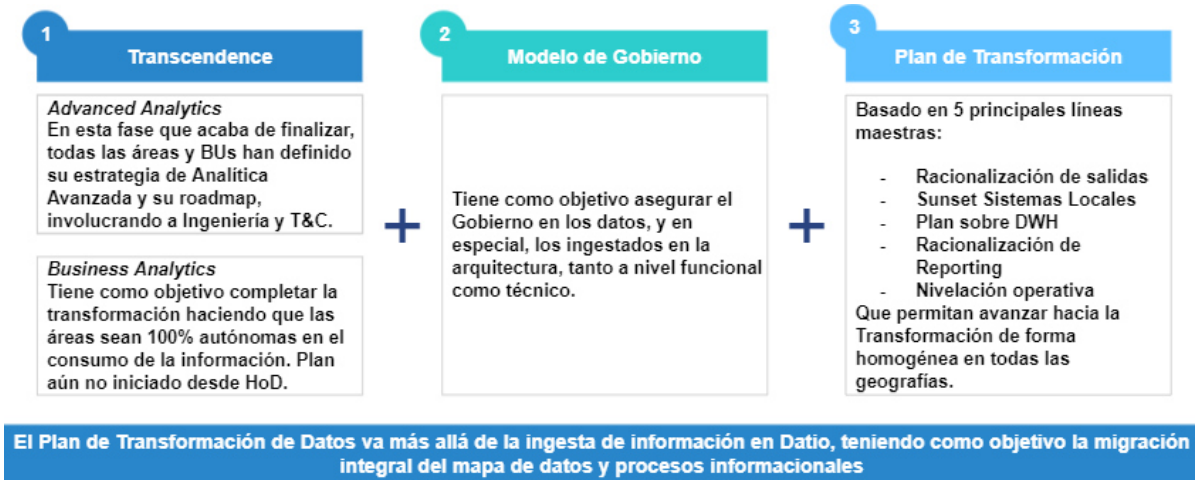


Figura 9. Plan de transformación de la Entidad Financiera

En este gran plan de transformación, se enmarcan 3 etapas:

- **El Proyecto Transcendence**: La integración del entorno Big Data del presente informe se ubica en esta etapa del plan de transformación, con la identificación de la situación actual, los planes a seguir y la puesta en marcha de un piloto funcional, considerando los aspectos de analítica avanzada y analítica de negocios.

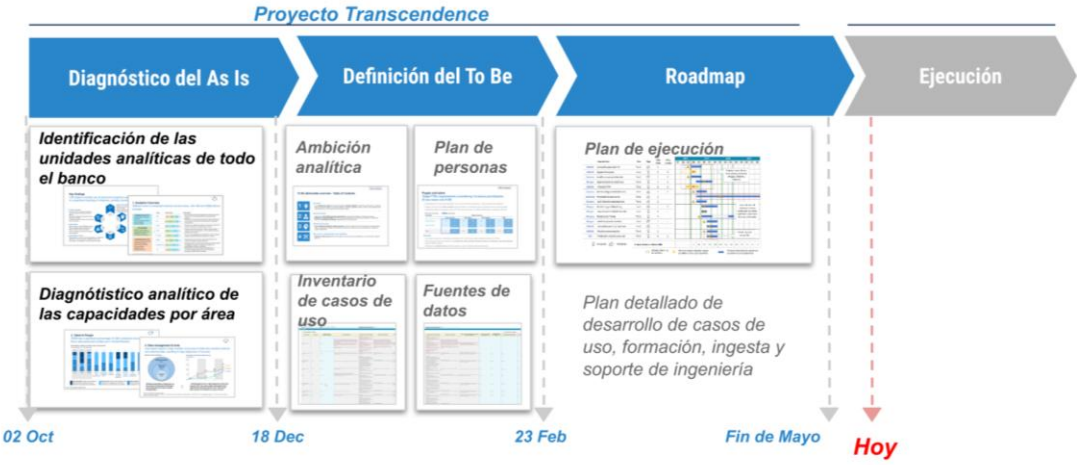


Figura 10. Esquema Transcendence en el Plan de transformación.

- **Diagnóstico del As Is**.- Evaluación actual no solo del proceso de campañas, sino de todas las unidades analíticas de la empresa y las fuentes de información. También de las capacidades técnicas de las distintas áreas de negocio involucradas.

- **Definición del To Be**. - Implica identificar Casos de Uso y estimar recursos para todas esas iniciativas, así como las fuentes de datos que se requerirán y las bases de datos locales que se irían apagando.

- **Roadmap**. - Diseñar el cronograma de ejecución de los casos de uso y formaciones en todas las áreas de negocio evaluadas.

- **El Modelo de Gobierno**: Implica que toda ingesta en el Data Lake cumpla con los lineamientos Globales de estructuras de datos y nomenclaturas, ya que estas fuentes quedarán en un diccionario de datos accesible para cualquier data scientist que requiera inputs para sus modelos analíticos. Queda el área de Engineering como responsable de la ingesta gobernada en el Data Lake, y Arquitectura como ente validador.

- **Plan de Transformación**:

  - Racionalización de Salidas. - El objetivo es reducir las implementaciones en el mainframe host a los distintos aplicativos, el entorno big data es una alternativa a centralizar información gobernada y protegida.
  - Sunset de sistemas locales. - Muchos desarrollos locales pueden migrar al entorno Big Data, aprovechando el procesamiento distribuido.
  - Plan sobre el Data WareHouse. - El Data WareHouse actual debe interactuar con el Data Lake para homologar fuentes y siga trabajando para todas las necesidades BI.
  - Racionalización de Reporting. -Los desarrollos de reportes departamentales pueden salir del mismo Data Lake, evitando descargas de información y mayor costo en licencias de reporteadoreslegacy o sistemas hechos a medida.
  - Nivelación Operativa. - Los datos centralizados y confiables deben homologar los resultados de la operativa de las distintas áreas de negocio.

3.2.5.2 Formación de la masa crítica

Conformado el nuevo equipo, se inicia la fase de capacitación, tanto de herramientas técnicas como de la herramienta metodológica Agile, Scrum.

- **Formación de Usuarios de Negocio**. - Iniciamos las capacitaciones con lenguajes de programación Scala y Spark para los procesamientos que migrarán los Data Scientist y con el manejo de HDFS.

  Así como trabajar en un Sandbox con Notebooks donde se pueda desarrollar en dichos lenguajes sobre el entorno Big Data.

Tabla 8. Formación usuarios de negocio

Formación de usuarios de Negocio			
Personal Formado	Áreas de conocimiento	Tecnología	Nivel requerido
8	Lenguajes de Programación	Python, Spark	Intermedio

- **Formación de los Desarrolladores**. - Este rol adquirió conocimientos de HDFS y de ingesta de datos con librerías proporcionadas por los holdings de arquitectura. Esta librería está basada en Scala y Spark y tiene como objetivo darle estructura a los inputs (que normalmente son archivos planos o csv), dichas estructuras sirven para el gobierno del Data lake.

  Adicional a la estructura, estas librerías de ingesta brindan transformaciones de datos básicas y configurables, que por debajo no son más que DMLs (Lenguaje de Manipulación de Datos) en Spark SQL que serían los conocidos Join, GroupBy, Union, etc.

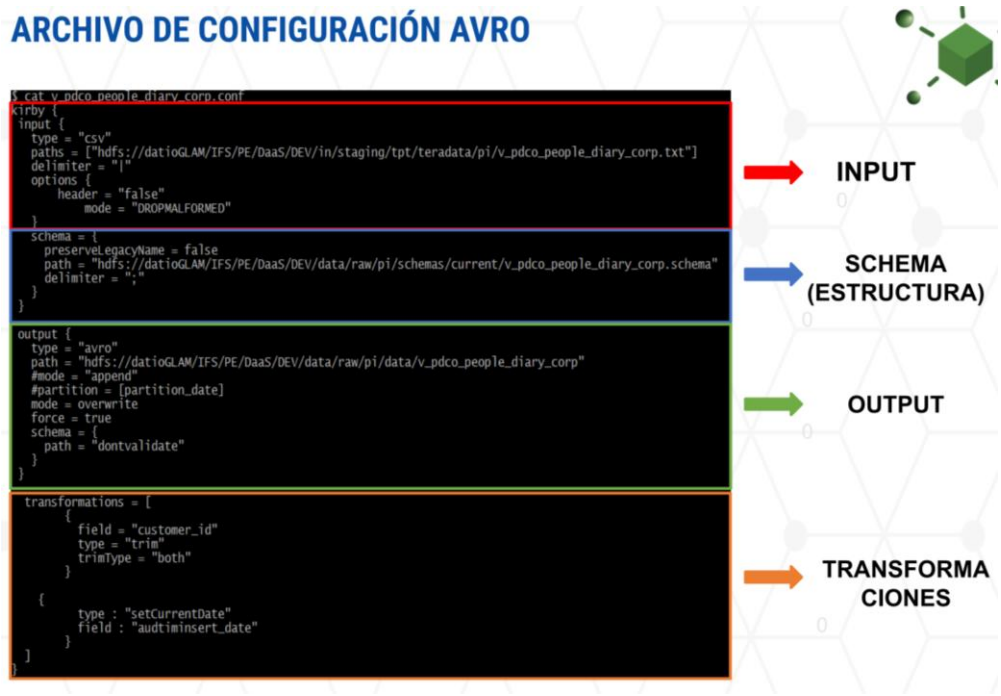


Figura 11. Ejemplo de configuraciones para ingesta de Datos

Hoy en día tenemos herramientas con Integración Continua para el despliegue de estas ingestas, usando componentes como Jenkins (para orquestar el pipeline del despliegue) y Bitbucket (como repositorio de las configuraciones y jobs).

Figura 12. Herramientas de Gestión de la configuración

Las capacitaciones incluyeron lineamientos de elaboración del diccionario del Data Lake para el gobierno de las estructuras de datos anteriormente descritas.

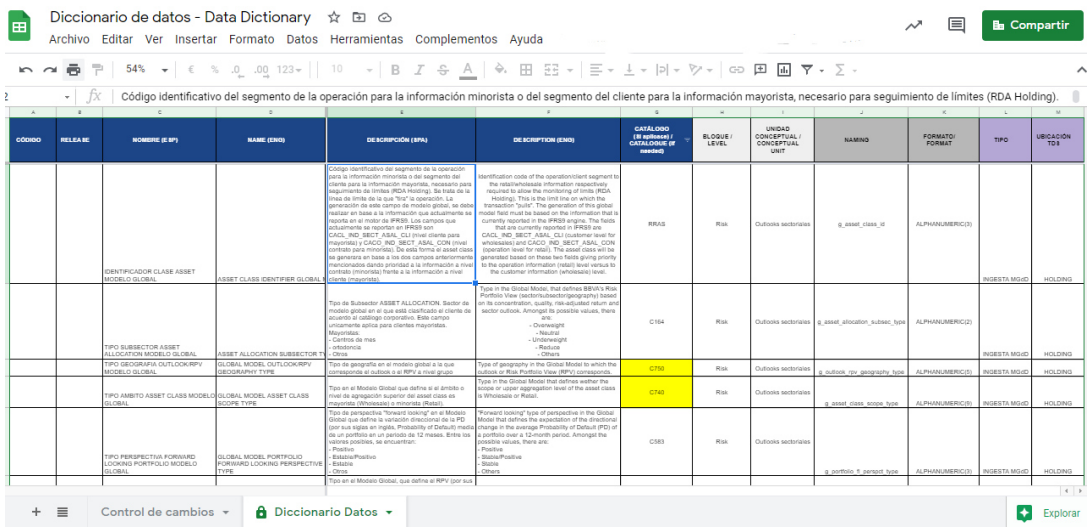


Figura 13. Diccionario de Datos.

En el piloto usamos las Hojas de Cálculo de Google, pero hoy en día ya tenemos herramientas para agilizar dicha elaboración de diccionario de datos.

Tabla 9. Formación de desarrolladores

Evaluación de diagnóstico Geográfico			
Personal Formados	Áreas de conocimiento	Tecnología	Nivel requerido
8	Lenguajes de Programación	Scala, Spark, Bitbucket	Intermedio
8	Estructura de archivos	Json, Avro, Parquet	Básico

- **Formaciones generales**. - Por último, los agile coaches nos capacitaron en el marco de trabajo ágil. Al tener un equipo multidisciplinario tomamos las ceremonias de Scrum como:
    - Planning
    - Dailys
    - Reviews
    - Retrospective

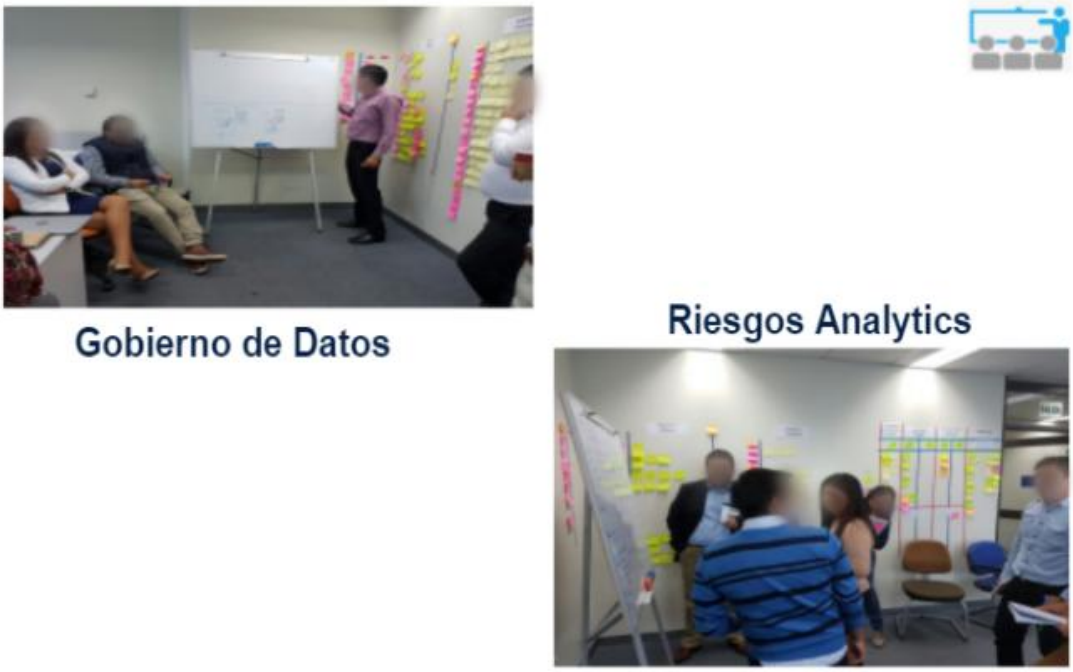


Figura 14. Ceremonias Agile

Y de la metodología Extreme Programming recogimos el uso de las historias de usuario.

El equipo necesitaba conocer el marco metodológico para trabajar en conjunto y de manera incremental, ya que las distintas gerencias involucradas (Riesgos, Finanzas, Desarrollo de Producto e Ingeniería) esperaban entregables quincenales para seguir aportando recursos. Scrum nos ayudó en desarrollar entregables iterativos como parte del producto final, ello mediante Sprints y la colaboración de los Stakeholders

Tabla 10. Formación masa crítica

Evaluación de diagnóstico Geográfico			
Personal Formados	Áreas de conocimiento	Tecnología	Nivel requerido
16	Marco Metodológico Agile	Scrum, XP	Intermedio

3.2.5.3 Validación de Plataforma

Cada una de las piezas son desplegadas tanto en Servidores localizados en España como en México (CCR) y estos últimos habilitan a toda América, el equipo de Arquitectura de Datos Global es el responsable de esa infraestructura tecnológica.

Figura 15. Distribución de la infraestructura

El equipo de Arquitectura de Datos Local, donde pertenezco, es el responsable de las piezas instaladas a nivel de su geografía, estas pasan por un proceso de recepción mediante POC (Pruebas de Concepto) donde se ejecutan casos básicos probando individualmente cada pieza con el objetivo de conocer las funcionalidades y el alcance antes de desplegarlo a los usuarios que requerirán capacitaciones. No se requieren pruebas de estrés o de rendimiento ya ello lo realiza el equipo de QualityAssurance al momento de instalar la pieza. De todas maneras, somos el segundo frente de validación funcional para el entorno Big Data ya que estamos en continua comunicación con Arquitectos Globales / Holding.

Estas piezas las podemos agrupar en dos frentes, Data Lake y Sandbox:

- **Data Lake**. - Todas las piezas que se involucran en el aprovisionamiento o ingesta de datos gobernados al Data Lake.

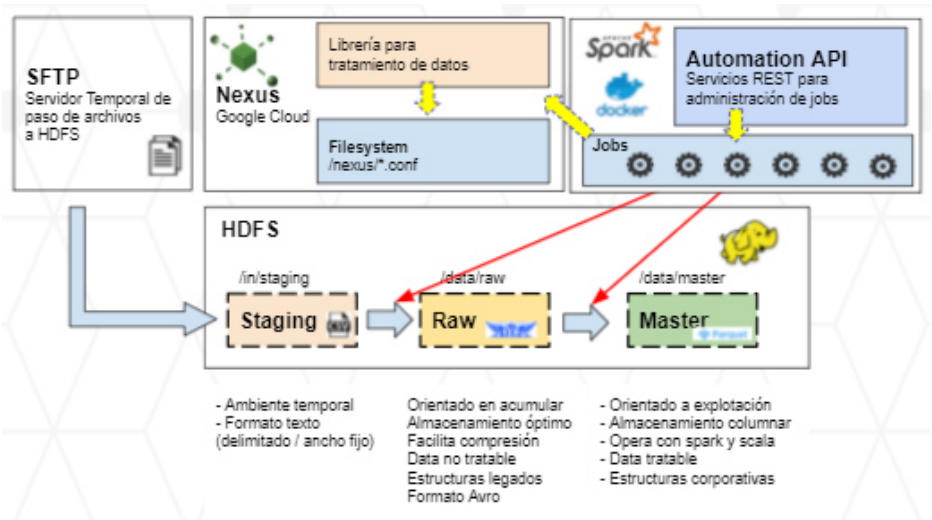


Figura 16. Distribución de la infraestructura del Data Lake

- **SFTP**.- Aquí se observa que la transmisión de datos de entrada es por el protocolo de transferencia SFTP, esta transmisión se hace por medio de Control M que es una herramienta de programación Batch, esta herramienta sirve en todo el ciclo de vida de aprovisionamiento del Data Lake.

- **HDFS**. - En este cuadro tenemos 3 capas de aprovisionamiento:
  - **Staging**. - Capa temporal de almacenamiento donde residen los archivos tal cual están en el origen, para efectos del piloto se homologó a archivos csv. La temporalidad de estos archivos debe ser la mínima posible, ya que la información puede contener data sensible (en las siguientes capas está la posibilidad de encriptar) y también debe ser temporal porque los archivos planos ocupan mucho espacio que debe ser aprovechado por otros formatos consumibles por los usuarios en Big Data.

  - **Raw**. - Es la capa de datos crudos en el entorno Big Data, esta capa contiene archivos con sistema de compresión a nivel de filas (formato AVRO) para posteriores transformaciones, pero no para consultas y analítica avanzada. Esta capa también nos sirve de almacenamiento histórico por su ahorro en disco y también sirve de respaldo ante posibles reprocesos de la siguiente capa.

- **Master**. - Capa final del Data Lake para la cual está compresa a nivel columnar con el objetivo de soportar análisis exploratorios y procesamiento de modelos analíticos, este formato de archivos es el PARQUET. Esta capa es la que tienen acceso los Data Scientist en sus respectivos Sandboxes.

- **Nexus Repository**. - Es el repositorio de artefactos donde almacenamos todos los archivos de configuración que usamos en las ingestas de archivos a las distintas capas del Data Lake, así como las librerías que se consumen en procesamientos complejos.

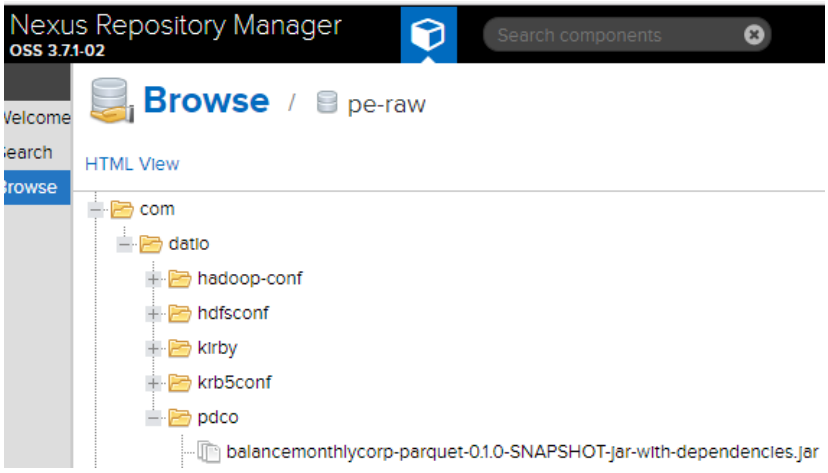


Figura 17. Repositorio de artefactos Nexus.

Actualmente reemplazamos este componente por JFrog Artifactory, una herramienta Open Source con la misma funcionalidad de Nexus pero que cumple de manera más eficiente en el ecosistema instalado.

- **Automation Api**. - Es el motor de procesamiento basado en Spark y en clusterización. Está distribuido por una serie de contenedores (Docker) y agentes (Nodos) que distribuyen los cores de manera eficiente para ejecutar los distintos jobs que aprovisionan archivos en el Data Lake. Como se observa en la figura 16, en esta capa se alojan los jobs (que servirán para procesamientos de información, como para ingesta al Data Lake), los cuales se crean y ejecutan mediante peticiones rest (comunicación entre servidores vía HTTP)

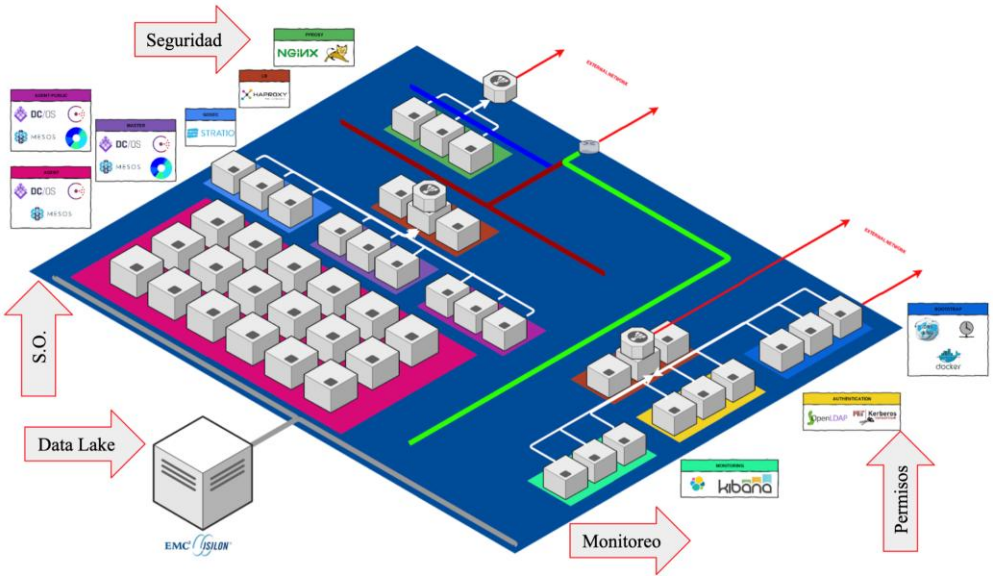


Figura 18. Diagrama general del entorno de procesamiento de datos.

En conclusión, este es uno de los componentes más importante de la plataforma ya que triangula con otros componentes de seguridad como se observa en la figura 18.

La infraestructura que soporta el Data Lake, y todo el entorno Big Data está respaldada por piezas de seguridad como Ldap para la autentificación de usuarios nominales y de servicio a aplicaciones, y con las distintas capas de proxy para evitar cualquier filtración al repositorio de datos crítico de la entidad financiera.

- **Sandboxes**. - Los componentes analíticos que consumen la información del Data Lake, descrito en los puntos anteriores, están distribuidos a nivel de sandboxes los cuales se montaron sobre Google Cloud. En la actualizar ya no usamos los servicios de Google y las herramientas Jupyter Lab están montadas en los mismos servidores de la entidad financiera.

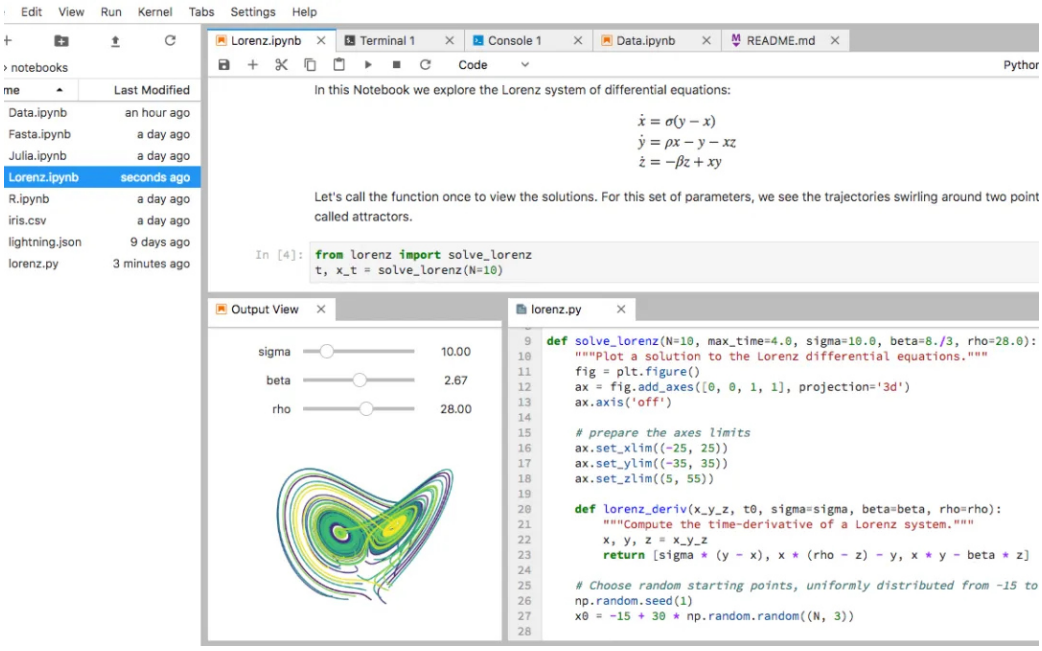


Figura 19. Jupyter y analítica avanzada.

Los sandboxes contienen permisos determinados a los archivos del Data Lake, según requieran los responsables del sandbox, como se puede ver en la figura 20, distintos Data Scientist pueden acceder a distintos sandboxes que a su vez pueden acceder a distintos archivos gobernados del Data Lake.

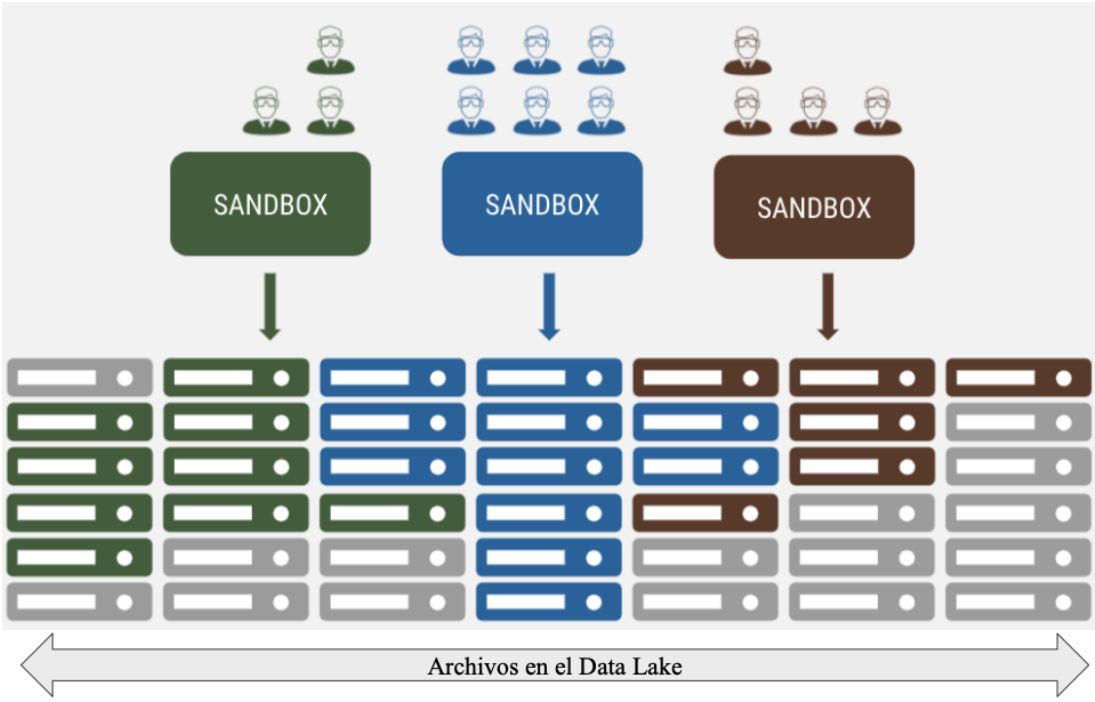


Figura 20. Diseño de la distribución de Sandboxes

De cara al piloto de campañas, creamos un sandbox y se ejecutaron las librerías básicas de procesamiento en PySpark y ScalaToore (Python y Scala respectivamente) y las modalidades de acceso y perfilamiento para que los Data Scientist desarrollen sus procesamientos y/o modelos.

Para finalizar la validación de componentes, integramos todas las piezas descritas anteriormente en el Diagrama de componentes de la figura 21

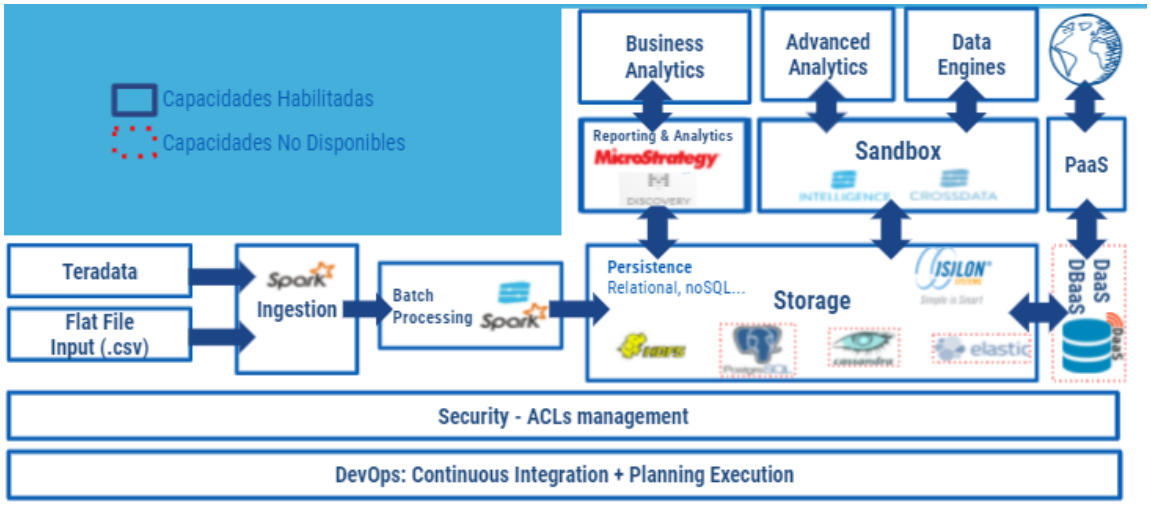


Figura 21. Diagrama de componentes

Los componentes de la figura 21 siguen el ciclo de vida del dato en el entorno big data de la entidad financiera. Iniciamos con la ingesta de datos a partir de fuentes de datos como Teradata o archivos planos, continuamos con el proceso de ingesta con las librerías descritas en el punto anterior, hay casos en que la ingesta es compleja por reglas de negocio para lo cual se desarrollan procedimientos específicos en spark, siendo ello la tercera caja, el componente central y el más importante es el Data Lake que actualmente contiene HDFS, posteriormente se irán agregando otras bases de datos (se puntea en rojo), lo que continúa son los componentes neurálgicos del entorno Big Data: Los Sandboxes (para desarrollo de analítica avanzada o motores de procesamiento) y las herramientas de Inteligencia de Negocios (como Microstrategy) para diseñar reportes departamentales de apoyo a la toma de decisiones, por último el entorno big data puede servir de input para otros aplicativos por ello se tiene el esquema de “BD as a services”, por el momento este componente solo responde a dejar archivos resultantes en una zona staging de salida del entorno para que puede ser transmitido a distintos aplicativos mediante Connect:Direct (Red de transferencias seguras de IBM)

3.2.5.4 Diseño y Desarrollo del Piloto

En esta etapa los developers, con las herramientas ya validadas, procedieron a realizar 2 actividades en el aprovisionamiento al data lake para implementar el piloto de Campañas de tarjeta de crédito (TC):

- **Actividad 1: "Diccionario de Fuentes'‘.** - Analizar con las distintas áreas de negocio el origen de los insumos para el proceso de campañas de TC. Una vez identificados, proceden a mapear el origen físico de la tabla o archivo, describir atributo por atributo, especificar el tipo de dato correcto y diseñar una estructura de datos relacional con estándares globales para inscribir este mapeo en el diccionario de datos.

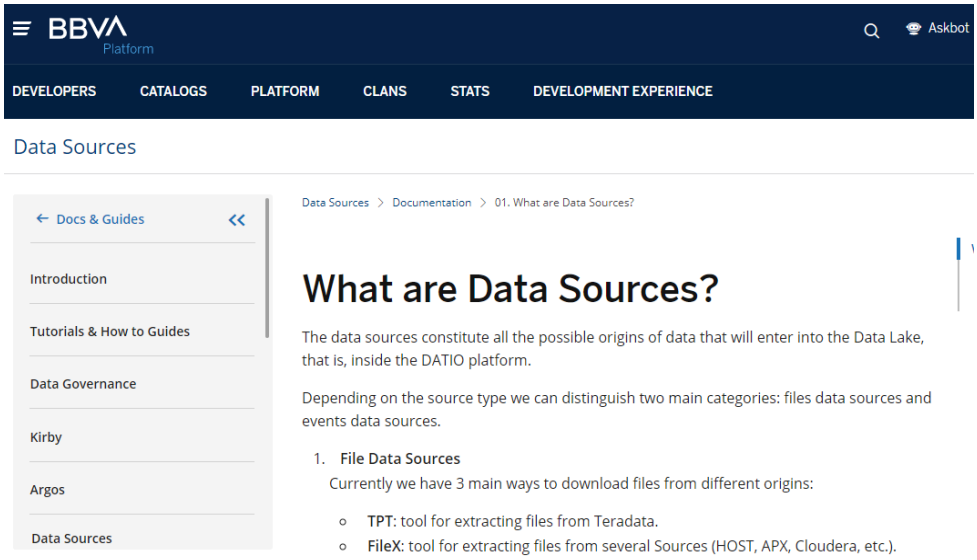


Figura 22. Buenas Prácticas de Gobierno

- **Actividad 2: Ingesta de datos**. - Mediante las librerías de ingesta, anteriormente validadas, se cargan los archivos, previamente indexados mediante diccionarios, a la zona STAGING (Archivos Planos), luego a la zona RAW (en formato Avro) y por último a la zona MÁSTER (en formato Parquet). Cada uno de estos procesos de carga no son más que ETLs (Extracción, Transformación y Carga) que se elaboran en Inteligencia de Negocios. El detalle del flujo de ingesta de datos por el circuito de integración continua se entra en el apartado "Anexo".

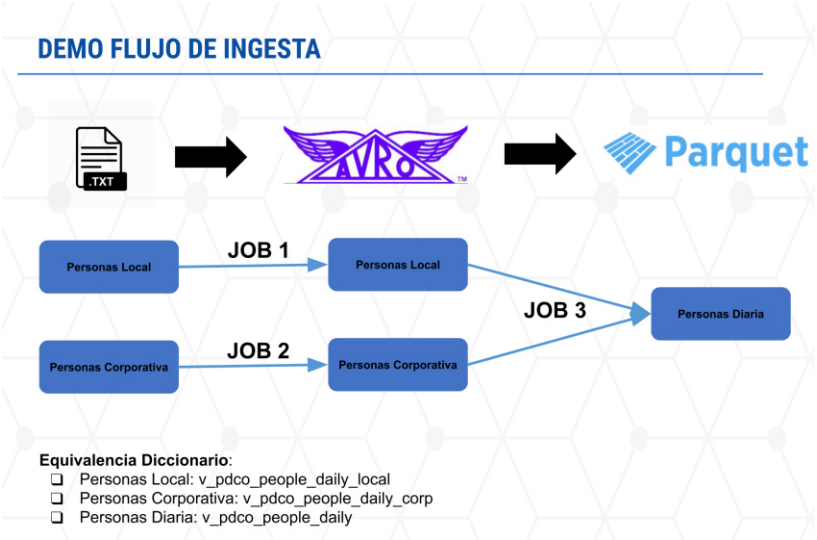


Figura 23. Flujo de Ingestas

Los Jobs con que ejecuta cada ingesta están montados en Automation Api. Este entorno de desarrollo lo llamamos "Work" y hay un entorno réplica de producción al que llamamos "Live".

Mientras los developers continúan aprovisionando el Data Lake, los Data Scientist trabajan sus modelos con data de prueba en el Sandbox disponibilizado, cabe aclarar que el Sandbox consume información del Data Lake, pero también puede consumir información que sube el propio usuario, pero esta última al no ser gobernado solo existe en el workspace del Data Scientist, y no en el Data Lake. Estos conjuntos de criterios se encuentran en las buenas prácticas de la intranet.

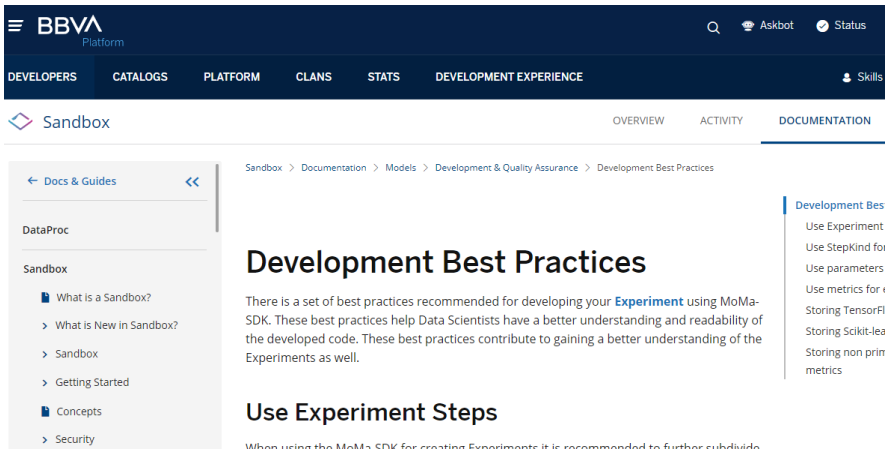


Figura 24. Buenas Prácticas

Cuando los developers terminen de aprovisionar el Data Lake, los Data Scientist cambiarán sus notebooks (Ambiente de desarrollo del Data Scientist para trabajar modelos analíticos y procedimientos) a consumir las fuentes gobernadas (ya no las cargadas por ellos mismos); aquí inicia la puesta en marcha del piloto donde los usuarios de negocio son autónomos en dichas ejecuciones y certificaciones.

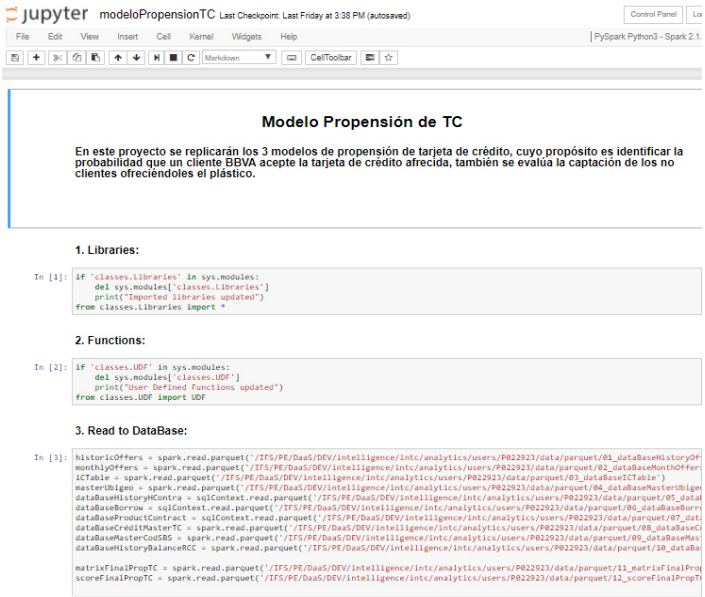


Figura 25. Notebook con el Modelo de Propensión de Tarjeta de Crédito

Como trabajamos con el marco metodológico Scrum, tenemos entregas iterativas, así que cada notebook que los Data Scientist liberan son implementados en IDEs por los developers para luego enviarlos a producción.

4.1 Offers' data base history

```
tempHistoryOffers = historicOffers.select('mes','nu_cte','tipo_documento','documento',
                                          col('linea_tdc').cast("Int").alias('linea_tdc'))

tempMonthOffers   = monthlyOffers.filter(col('linea_tdc').cast("Int")>0)\
                                 .select(col('mes'),'nu_cte','tipo_documento','documento',
                                          col('linea_tdc').cast("Int").alias('linea_tdc'))

historicOffersUnion = tempHistoryOffers.union(tempMonthOffers)
historicOffersUnion.printSchema()
% time historicOffersUnion.limit(3).toPandas()
```

```
root
 |-- mes: string (nullable = true)
 |-- nu_cte: string (nullable = true)
 |-- tipo_documento: string (nullable = true)
 |-- documento: string (nullable = true)
 |-- linea_tdc: integer (nullable = true)

CPU times: user 279 ms, sys: 92.9 ms, total: 372 ms
Wall time: 1.89 s
```

	mes	nu_cte	tipo_documento	documento	linea_tdc
0	201706	None	6d711c732eb6110c6eebe9212741730d	39f468118808fa321482d9d370f51e48	4700
1	201706	None	6d711c732eb6110c6eebe9212741730d	995cc9ceeff334ae66795a54d39d8d10	1300
2	201706	e238db86135d246dc23fc78236b49176	6d711c732eb6110c6eebe9212741730d	818d56956d5ece534e3f8a6bb87b09b0	4300

Figura 26. Ejecución del Propensión de Tarjeta de Crédito

Este modelo tiene por objetivo identificar la probabilidad de que algún cliente de la entidad financiera acepte una tarjeta de crédito mediante puntuaciones obtenidas a partir de variables sociodemográficas, consumos o actividad en el sistema financiero. Estos modelos no los detallaremos ya que el objetivo del presente trabajo es ver todas las aristas en la integración del entorno big data.

La puesta en producción implica nuevos desarrollos muy parecidos a las ingestas al Data Lake, con la diferencia de que estos tienen lógica de negocio, dicha lógica de negocio está en cada notebook de los distintos Data Scientist.

Por último, terminando todos los desarrollos puestos en Live y enmallados con el programador Control M, se lanza el proceso automático de Campañas de tarjeta de crédito, el resultado de todos los procesamientos son archivos planos con los identificadores de los clientes potenciales a adquirir una tarjeta de crédito, esta lista de clientes es enviada a los distintos canales de distribución de la entidad financiera (Web, ATM, Email, Redes Sociales).

3.2.5.5 Validación del piloto

En la validación del piloto hay 2 etapas:

- **Etapa en Sandbox**. - Los distintos usuarios proceden a cuadrar los resultados de sus modelos analíticos y procedimientos con los resultados obtenidos de la ejecución de las campañas de manera tradicional (proceso antiguo de 35 días). No se espera un cuadre exacto ya que en muchos casos los modelos se optimizaron con data más confiable y ampliaron el universo de posibles clientes a obtener tarjetas de crédito.

  Otra característica de esta etapa es que se construyen notebooks validadores, que nos van a servir en la segunda etapa para cuadrar con los procesos productivos. Estos notebooks contienen cuadres automáticos leyendo los resultados y cruzándolos con los targets esperados, así como las reglas de negocio que contemplan los Data Scientist; en comparación con los proyectos de desarrollo de software tradicionales estas vendrían a ser las pruebas de aceptación.

- **Etapa de marcha blanca**. - Inicia cuando están en producción los procesamientos realizados por los developers; aquí todos notebooks de los Data Scientist han sido

implementados en jobs que se ejecutarán bajo Spark en una malla orquestada por Control M.

De esta manera los datos resultantes de las ejecuciones programadas se van monitoreando con los notebooks validadores de la etapa anterior, si se encuentra alguna observación se corrige el desarrollo y se lanza el reproceso permitiendo sobre escribir la partición de datos que se está validando.

El primer piloto de la campaña de tarjetas de crédito no tuvo éxito así que se realizaron las modificaciones respectivas para el siguiente mes, ya que este proceso de campañas es mensual. Con el segundo intento de ejecución del proceso de campañas en el entorno Big Data si se logró el objetivo, reduciendo de 35 días a 5 días con fuentes confiables, almacenamiento adecuado, eficiente y gobernado.

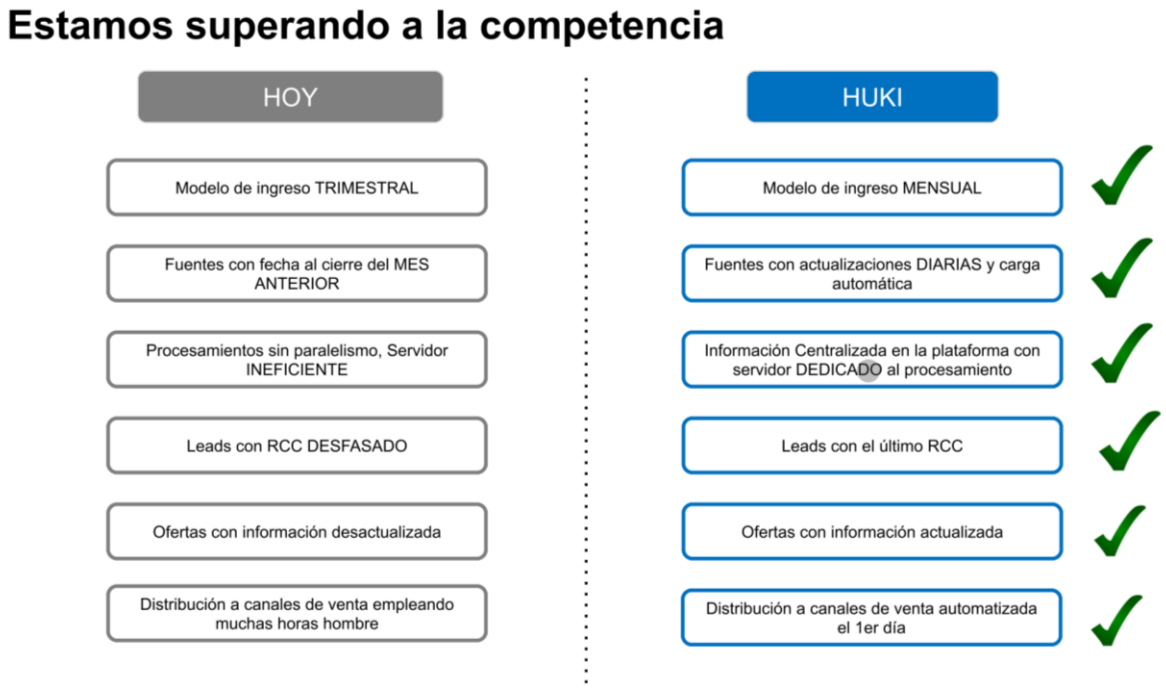


Figura 27. Variables cualitativas de validación

3.2.5.6 Despliegue

Los resultados satisfactorios del piloto de campañas de tarjeta de crédito permiten que el entorno Big Data se despliegue a demás iniciativas del plan de transformación, para ello se forman comités entre el área de Gestión de Proyectos e Ingeniería para poner en

marcha los proyectos que según rentabilidad, complejidad e impacto se vayan a realizar en el entorno Big Data.

Aquí la labor de los Arquitectos de Soluciones cobra relevancia, ya que ellos diseñan documento con el modelo de solución arquitectónico para los distintos casos de uso, este lineamiento global ayuda mucho en la elaboración de los diseños técnicos, funcionales y las propias implementaciones de ingestas en el Data Lake. Estas soluciones abarcan todas las capas de un desarrollo de software como los servicios o el front end, por ello dichos arquitectos de soluciones se apoyan en los arquitectos de datos para así nosotros sugerirle alternativas de solución específicas en el Data Lake o fuera de él.

Se disponibiliza las comunidades de arquitectura, data scientist e engineering, con el objetivo de compartir experiencia con otros países donde también ejerce la entidad financiera. Esta comunicación refuerza las buenas prácticas instauradas y las recomendaciones de iniciativas.

Se crearon 2 sandboxes más para las áreas de Finanzas y Riesgos ya que ellos poseían la mayor cantidad proyectos a priorizarse, con estos sandboxes los data Scientists pueden hacer su relevamiento de información o análisis exploratorios.

3.2.5.7 Reliability and Research

Una vez desplegado el entorno big data más proyectos y más usuarios la consumen progresivamente con ello la formalización de un el frente de soporte de plataforma es crucial para la estabilidad de un ambiente de trabajo continuo.

Este equipo de soporte inicialmente se centralizó en México donde está la infraestructura de todos los países de américa, pero al poco tiempo se descentraliza a las distintas regiones dándoles autonomía a los equipos de Engineering para brindar acceso a archivos en los sandboxes, altas de usuario, creación de entornos de desarrollo para los developers, y creación de sandboxes; esta autonomía mejoró la percepción de la calidad servicios internos por parte de Arquitectura e Engineering, ya que el escalamiento de capacidades es mejor en un entorno distribuido con es el caso de Big Data a diferencia de los

sistemas actuales donde Engineering siempre era un obstáculo, por ejemplo el aumento de storage.

Conforme pasa el tiempo nuevas piezas vienen a sumarse al entorno Big Data y otras a reemplazar, es el caso de Nexus por JFrog Artifactory (como repositorio de artefactos de configuración) o nuevas librerías para ingesta de datos en el Data Lake o para medir la calidad de datos en el mismo Data Lake; para ello Arquitectura de Datos continúa con su labor de validación de componentes asegurando el correcto uso de cada pieza del entorno Big Data que vienen desde España.

El Batch productivo crecerá conforme aumenten los proyectos e iniciativas, para ello el área de Reliability of Engineering asegura la fiabilidad de las mallas en Control M, teniendo comunicación directa con nosotros, el equipo de arquitectura y los encargos de la infraestructura en México.

Por último, el entorno Big Data es monitoreado constantemente de manera local y global mediante KPIs estratégicos, que abarcan el consumo de recursos (cpus, memoria y disco), accesos de usuarios, análisis de incidencia y seguimiento de objetos gobernados.

Figura 28. Ejemplo de Seguimiento del entorno

Para obtener estos datos nos apoyamos en herramientas de monitoreo continuo de cores, usuarios, memoria, etc., de nuestro entorno Big Data que está en continua evolución.

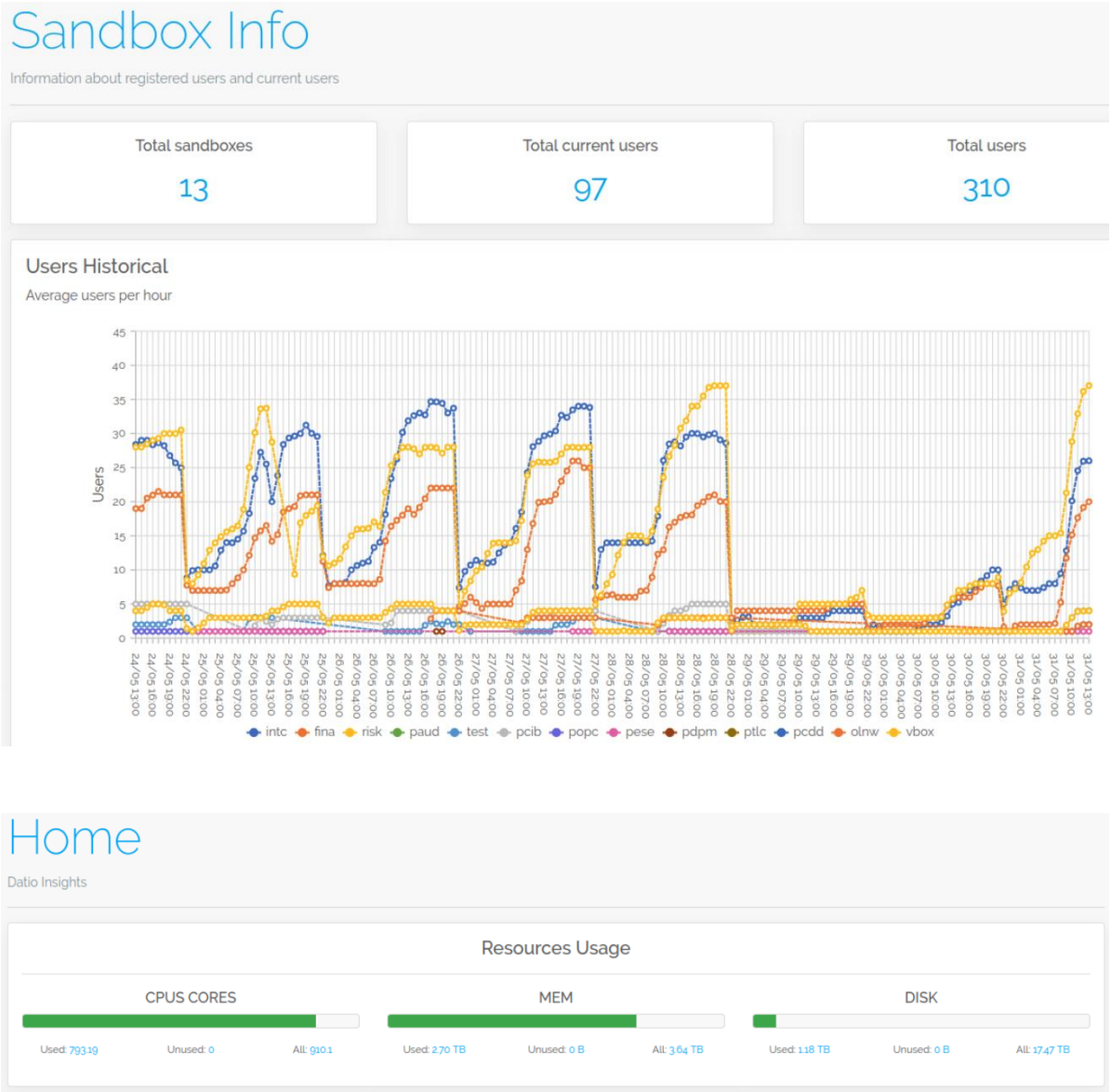


Figura 29. Herramientas de Monitoreo del entorno

En la figura 29 se muestran 2 gráficas en línea del consumo del entorno Big Data:

El gráfico superior (Sandbox Info) contiene la evolución del número de usuarios activos en los distintos sandboxes por minuto en los últimos 5 días; por ejemplo, en ese momento tenemos 13 sandboxes, se encuentran 97 usuarios activos, de un total de 310 usuarios que se pudieron conectar en algún momento; el eje Y son los distintos sandboxes.

El gráfico inferior (Home) muestra, en vivo, los cpus, la memoria y disco consumido del entorno Big Data; por ejemplo, el primer cuadro indica 849 cores (cpus) de un total de

910, el segundo cuadro indica 285 TB usados de un total de 364 TB, el tercer cuadro indica 1.18 TB consumidos de un total de 17.47 TB.

3.3EVALUACIÓN ECONOMICA

La evaluación del proyecto de integración de entorno big data para la entidad financiera, considera la parte de análisis de costo - beneficio.

3.3.1 EVALUACIÓN COSTO

La evaluación de la integración del entorno Big Data se centra en la efectividad del piloto implantado; para ello se tiene como referencia el pago mensual al personal de TI involucrado en el piloto. Con esta información se calcula el retorno de inversión ROI comparando el gasto de los recursos humanos usando el flujo antiguo del proceso de campañas versus el gasto de recursos humanos con el nuevo flujo en el entorno Big Data.

Tabla 11. Costo de RRHH del Piloto en soles

									1er Fallo					2do Fallo				Éxito	
			Análisis		Implementación			Pruebas		Ajuste	Implementación			Pruebas		Ajustes		Pruebas	
			2017	2018												2019			
Rol	#	Área	Dic	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Set	Oct	Nov	Dic	Ene	Feb	Mar	Abr	Total
Gobierno	1	DA Finanzas	6000	6000	6000	6000	6000			6000	6000	6000	6000			6000				60000
Ingesta+Gestión	3	Arquitecto de Datos	18000	18000	18000	18000	18000	18000	18000	10000	10000	10000	10000	10000	10000	5000	5000	5000	5000	206000
Ingesta	4	DE Engineering	16000	16000	16000	16000	16000	16000	16000	12000	12000	12000	12000	12000	12000	8000	8000	8000	8000	216000
Ingesta	1	DE Consultor	8000	8000	8000	8000	8000	4000	4000											48000
Analítica	2	DS Desarrollo Comercia	14000	14000	14000	14000	14000	14000	14000	14000	14000	14000	14000	14000	14000	7000	7000	7000	7000	210000
Analítica	2	DS Riesgos	13000	13000	13000	13000	13000	13000	13000	13000	13000	13000	13000	13000	13000	6500	6500	6500	6500	195000
Analítica	1	DS Consultor	8000	8000	8000	8000	8000	4000	4000											48000
Gestión	1	Engineering	8500	8500	8500	8500	8500	8500	8500	4250	4250	4250	4250	4250	4250	4250	4250	4250	4250	102000
Gestión	1	Desarrollo Comercial	9000	9000	9000	9000	9000	9000	9000	9000	9000	9000	9000	9000	9000	9000	9000	9000	9000	153000
																				1238000

Los participantes del piloto fueron aproximadamente 16 personas de la entidad financiera, pero conforme avanzaba el proyecto ciertos roles dejaban de ser prescindibles, eso reflejamos en la tabla 11 con las celdas en blanco indicando el “no costeo” en los meses respectivos.

Como también se observa, el piloto fue conformado por un equipo multidisciplinario, las áreas participantes fueron: Finanzas, Arquitectura de Datos, Engineering, Consultores, Desarrollo Comercial y Riesgos, y cada uno de ellos esto roles:

- DS: Data Scientist (5)
- DA: Data Analyst (1)
- DE Data Engineer (5)

El cálculo del retorno de inversión (ROI)

Toma la estimación de los costos del proceso antiguo, y el nuevo proceso de campañas de tarjeta de crédito, así como el costo del piloto descrito en la tabla 11.

Tabla 12. Costo Mensual del proceso de campañas TC

Roles	1º Sem	2º Sem	3º Sem	4º Sem	5º Sem	Total
ANTIGUO PROCESOS DE CAMPAÑAS DE TC						
DS Desarrollo Comercial	4800	4800	4800	6400	6400	27200
						27200
NUEVO PROCESOS DE CAMPAÑAS DE TC						
DS Desarrollo Comercial				5250	5250	10500
DE Engineering					500	500
						11000

	Costo Mensual Proceso Antiguo * (Meses del Proyecto+Meses de recuperación)		
ROI =	-		
	(Costo del Piloto + (Costo Mensual Proceso Nuevo * Meses de recuperación))		
ROI >0	2000	Meses de Proyecto	17
		Años de recuperación	4

Podemos observar en la tabla 12 que el ROI > 0, (Es decir, que se lograría retorno de inversión) con 4 años de recuperación, (es decir, 4 años ejecutando el nuevo proceso de campañas en el entorno Big Data)

3.3.2 BENEFICIO PARA LA ORGANIZACIÓN

Es un punto de partida para realizar apagados de sistemas locales / base de datos, que son un problema para la organización en tener la información muy dispersa y no consensuada entre las áreas de negocio que usan dichos datos.

El Data Lake gobernado es un repositorio funcional de conceptos que pueden ser reutilizados por otras sedes de la entidad financiera en otros países, ahorrando tiempos de análisis de datos al momento de ingestar al Data Lake.

Información consensuada por Data Owners y centralizada en el Data Lake hace que no se ponga en discusión la veracidad de la fuente de datos, dándole soporte y confiabilidad a los KPIs y reportes gerenciales que puedan presentar, así como consistencia a los productos financieros desplegados.

Velocidad con el procesamiento distribuido de spark para grandes volúmenes de datos, frente a los recursos acotados de SQL Server y Oracle que actualmente tienen instalados las diversas áreas de negocio.

Fue muy importante que se integre al equipo consultores externos con amplia experiencia en herramientas Big Data para que el conocimiento se quede en el personal de la entidad financiera y aportar en soluciones futuras.

El nuevo flujo, aparte de ser más rentable luego del cuarto año de ejecución, es automático en casi todo el proceso (salvo nuevas reglas de campañas entre meses), el proceso antiguo involucra mucha manualidad de los usuarios de negocio, el nuevo entorno Big Data apoyó en liberar dichas manualidades para que estos recursos humanos traten nuevas iniciativas de analítica y aprovechar al máximo sus capacidades.

CAPÍTULO IV:
REFLEXIÓN CRÍTICA DE LA EXPERIENCIA

El uso del marco metodológico agile sirvió mucho para las entregas iterativas de la integración del entorno Big Data, pero el presente trabajo podría ser el primer paso para una nueva metodología que incorpora conceptos de desarrollo de software, niveles de servicio y pilotos incorporados.

Ya que es una plataforma en continua evolución, habrá nuevas piezas que llegan esporádicamente, para ello la metodología anteriormente propuesta aplicaría también en los upgrade del entorno.

Los Data Scientist no estaban dedicados al 100% en el desarrollo del piloto, eso complicó muchas reuniones de definición de reglas, que a la larga provocó re trabajo; esto ocurre porque las gerencias de las respectivas áreas de negocio no quisieron descuidar su operativa diaria al desprenderse de sus mejores recursos.

A pesar de que los Data Scientist no estaban dedicados al piloto, el compromiso y motivación siempre fue alto y se reflejó en la compenetración del equipo, esto es vital para el éxito de cualquier proyecto.

Las políticas de la organización deben dirigirse o basarse en la cultura de los datos, y en la investigación de nuevos productos y servicios acorde a los nuevos avances tecnológicos, con la finalidad de aportar al plan de transformación digital.

Se debió dar más importancia al gobierno del dato para no tener problemas a futuro, como arrastrar deudas de nomenclaturas, al estar en el entorno Big Data en continua evolución, los lineamientos de gobierno también lo están provocando deprecaciones. Aquí las estrategias de homologación de versiones deberían ser más ágiles.

CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

1. La Arquitectura es muy flexible ya que por su naturaleza híbrida (contiene en su estructura componentes de código abierto) se adapta a las nuevas tendencias del mercado, por ejemplo, hay convenios con Google para integrar componentes como Kubernetes; o con Amazon Web Service para usar sus endpoints y mejorar el acceso a los servicios de nuestro entorno Big Data. Todos estos componentes permiten ampliar nuestros conocimientos y enriquecer el rol de Arquitectura de Datos.

2. El Data Lake gobernado nos brindan mucha volumetría para tomar decisiones, además la información es veraz por el gobierno del dato implantado, tiene el consenso de distintas áreas de negocio por el valor y la variedad predominan en el Data Lake. Por último, el procesamiento y almacenamiento distribuido hace que la velocidad del acceso a la información y generación de reportes esté acorde a la competencia, siempre y cuando los usuarios tengan las capacidades de Inteligencia de Negocios y Big Data

3. Con el entorno de Big Data las áreas de negocio desarrollarán analítica avanzada para visualizar y predecir las ofertas de nuevos productos, servicios, segmentación de clientes y comportamiento para campañas crediticias. Derivando a que las áreas de negocio sean independientes del área de Engineering (que anteriormente era el responsable de todo desarrollo), ya que tendrían sus propios Sandboxes para sus propios desarrollos y un Data Lake gobernado.

4. El uso de Spark optimizó los tiempos de procesamiento, usando motores de datos tradicionales los tiempos eran muy elevados. Uno de los procesos de los modelos de campañas tomaba 24 horas en generar resultados, mientras que con spark 60 minutos.

5. La integración del entorno de Big Data en la gestión financiera contribuyó a la generación eficiente de productos y servicios para el área de Desarrollo Comercial, optimizando también sus procesos, servicios internos y la toma de decisiones a nivel gerencial.

RECOMENDACIONES

1. Para tener un criterio más sólido en la validación del entorno de big data podríamos usar encuestas y/o cuestionarios de valoración asociadas a la arquitectura como plantean Quiroz y otros (2019), en el artículo mencionan una población de 26 expertos donde se midieron sus coeficientes de competencia y luego se les aplicó una encuesta para determinar el nivel de aceptación de la arquitectura en aspectos de herramientas, analítica, aplicabilidad y visualización.

2. Es positivo que el entorno big data tenga posibilite autoservicio en las áreas de negocio, pero el área de Engineering y Arquitectura debe estar al tanto de los desarrollos y consumos sobre el entorno, ya que somos el soporte a cualquier incidente o primer frente a cualquier mejora.

3. Tener claro los roles dentro de la entidad financiera, el funcionamiento organizacional es muy importante para que nuestro entorno big data rinda óptimamente, la orientación y capacitación a las áreas de negocio debe ser estratégica para sacar valor a desarrollos sobre el entorno, de forma contraria perjudica en la confiabilidad de la plataforma creando feedback no objetivos y desenfocando los objetivos del plan de transformación del dato de la entidad financiera.

4. El gobierno del dato no debe disminuir en ningún momento la rigurosidad tanto del diccionario como de la calidad de los datos, ya que al ser procesos muy costosos en el ciclo de vida del dato los proyectos puedan querer prescindir de ellos por cronogramas ajustados o pedidos ejecutivos.

FUENTES DE INFORMACIÓN

- Han y otros (2012). Data Mining Concepts and Techniques - Third Edition - Jiawei Han, Micheline Kamber, Jian Pei

- Quiroz y otros (2019). Proceso de diseño de una arquitectura Big Data para el análisis de grandes volúmenes de datos e información. Miguel Quiroz Martínez, Ricardo Andrés Aguilar Duarte, Diego B. Intriago Cedeño. Revista Opuntia Brava - Universidad de Las Tunas. Cuba (2019)

  http://opuntiabrava.ult.edu.cu/index.php/opuntiabrava/article/download/968/1139

- Camargo-Vega y otros (2014). Conociendo Big Data - Juan José Camargo-Vega, Jonathan Felipe Camargo-Ortega, Luis Joyanes-Aguilar

  http://www.scielo.org.co/pdf/rfing/v24n38/v24n38a06.pdf

- Puyol (2014). Una aproximación a Big Data - Javier Puyol Moreno

  http://e-spacio.uned.es/fez/eserv/bibliuned:rduned-2014-14-7150/Documento.pdf

- Wright y otros (2006). The Sandbox for analysis: concepts and methods - William Wright, David Schroh, Pascale Proulx, Alex Skabutski, Brian Cort

  https://dl.acm.org/doi/abs/10.1145/1124772.1124890

- Miloslavskaya y Tolstoy (2016). Big Data, Fast Data and Data Lake Concepts - Natalia Miloslavskaya, Alexander Tolstoy

  https://www.sciencedirect.com/science/article/pii/S1877050916316957

- Erraissi y otros (2018). Meta-modeling of Data Source and Ingestion Big Data layers - AllaeErraissi, AbdessamadBelangour, AbderrahimTragha

  https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3185342

- Nachiappan y otros (2017). Cloud storage Reliabilityfor Big Data applications: A stateofthe art survey - RekhaNachiappan, BahmanJavadi, Rodrigo N. Calheiros, KenanM. Matawie

  https://www.sciencedirect.com/science/article/abs/pii/S1084804517302734

- Wang y otros (2021). HowMuchAutomationDoes a Data ScientistWant? - Dakuo Wang, Q. Vera Liao, Yunfeng Zhang, UdayanKhurana, Horst Samulowitz, Soya Park, Michael Muller, Lisa Amini

  https://arxiv.org/pdf/2101.03970.pdf

- Galeano y Peña (2019). Las nuevas Oportunidad del Big Data para las instituciones financieras - Pedro Galeno, Daniel Peña

  https://www.funcas.es/wp-content/uploads/Migracion/Articulos/FUNCAS_PEE/162art07.pdf

- BBVA (2020). Api Market BBVA

  https://www.bbvaapimarket.com/es/mundo-api/las-siete-v-del-big-data/

- Plase y otros (2017). A comparisonof HDFS compact data formats: Avro Versus Parquet - DaigaPlase, Laila Niedrite, RomansTaranovs

  https://journals.vgtu.lt/index.php/MLA/article/view/500/357

GLOSARIO

- **Big Data**. - Conjunto de datos o combinaciones de conjuntos de datos, estructurados y no estructurados cuyo tamaño, complejidad y velocidad de crecimiento dificulta su captura, gestión, procesamiento o análisis mediante tecnologías y herramientas convencionales (bases de datos, estadísticas clásicas).

- **Data Lake**. - Entorno de datos compartidos en su formato original (en bruto o nativo) para un posterior tratamiento, contiene datos en distintos formatos, tanto estructurados como no estructurados.

- **Integración de componentes**. - Referido a Tecnología de Información, es un conjunto de actividades donde se combina software en varios niveles, mediante fases y en un contexto empresarial, con el objetivo de tener un único producto.

- **Spark**. - Es un sistema de computación en clúster de propósito general, de código abierto y orientado a la velocidad. Proporciona APIs en Java, Scala, Python y R. También proporciona un motor optimizado que soporta la ejecución de grafos en general. También soporta un conjunto extenso y rico de herramientas de alto nivel entre las que se incluyen Spark SQL (para el procesamiento de datos estructurados basada en SQL), MLlib para implementar machine learning, GraphX para el procesamiento de grafos y Spark Streaming.

- **HDFS**. - (Hadoop Distributed File System) es el componente principal del ecosistema Hadoop. Esta pieza hace posible almacenar data sets masivos con tipos de datos estructurados, semi-estructurados y no estructurados como imágenes, vídeo, datos de sensores, etc. Está optimizado para almacenar grandes cantidades de datos y mantener varias copias para garantizar una alta disponibilidad y la tolerancia a fallos.

- **Python**. - Lenguaje de programación interpretado cuya filosofía hace hincapié en la legibilidad de su código. Lenguaje de programación multiparadigma, es decir soporta parcialmente la orientación a objetos, programación imperativa y, en menor medida, programación funcional.

- **Tecnología de almacenamiento**. - Herramientas tecnológicas para almacenar datos, entre sus capacidades debe contemplar la colocación, mantenimiento y control del dato.

- **Procesamiento de datos**. - Es la acumulación y manipulación de elementos de datos para producir información significativa para la toma de decisiones.

- **Ingesta de Datos**. - Es el proceso de absorber datos de una variedad de fuentes y transferirlos a un sitio de destino donde se pueden depositar y analizar. En términos generales, los destinos pueden ser una base de datos, un almacén de datos, un almacén de documentos, un mercado de datos, etc. Por otro lado, existen varias opciones de origen, como hojas de cálculo, extracción de datos web o desguace web, aplicaciones internas. Teniendo como objetivo que dichos datos sirvan para descubrir conocimientos ocultos y utilizarlos para obtener ventajas comerciales.

- **Sandbox**. -Aislamiento de procesos. Es un entorno de pruebas aislado dentro de un sistema que permite ejecutar ciertas acciones de un programa escudando el hardware para que no pueda ser dañado. En el ámbito del desarrollo de software y la ciberseguridad, el aislamiento de procesos ya se ha convertido en una medida estándar gracias a la cual muchas grandes empresas pueden protegerse frente a daños del sistema y ciberataques.

- **Avro**. - Es un marco de serialización de datos y llamadas de procedimiento remoto orientado a filas desarrollado dentro del proyecto Hadoop de Apache. Utiliza JSON para definir tipos de datos y protocolos, y serializa datos en un formato binario compacto.

- **Parquet**. - Es un formato open-source de almacenamiento en columnas para Hadoop. Fue creado para poder disponer de un formato libre de compresión y codificación eficiente.

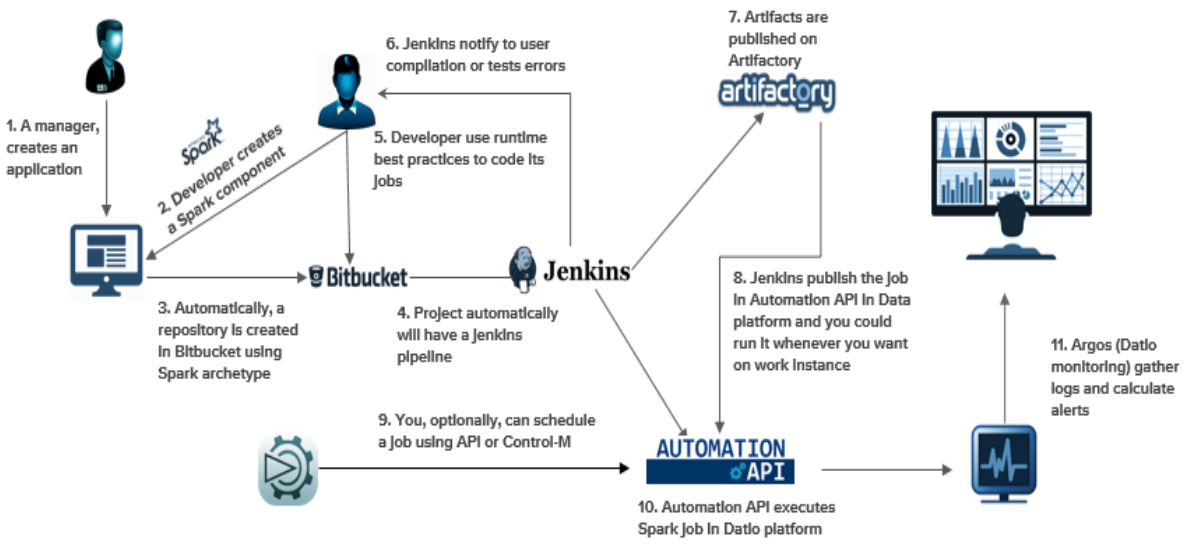
- **Gobierno del dato**. - Es la gestión de todos los datos de la empresa con el objetivo que estén disponibles en forma estructurada para los usuarios.

- **Reliability**. - Es la forma de mantener estable una plataforma operativa, dando la confiabilidad de su uso.

- **Análisis de Datos. -**Es la ciencia que se encarga de examinar un conjunto de datos con el propósito de sacar conclusiones sobre la información para poder tomar decisiones, o simplemente ampliar los conocimientos sobre diversos temas.

ANEXO

Flujo de Integración continua en un proyecto de ingesta de datos



Descripción del flujo de integración continua:

Paso 1.- El administrador crea una aplicación en la “consola de unidades aplicativas” para que sea usado por los desarrolladores. El administrador generalmente es el Arquitecto de Soluciones.

Paso 2.- El desarrollador o líder técnico configura, en la aplicación previamente creada, que el repositorio va contener componentes Spark.

Paso 3.- Con la aplicación creada, automáticamente se despliegan un repositorio en Bitbucket con arquetipo de proyecto de ingesta (Python o Scala); actualmente el administrador de los repositorios es el equipo de aseguramiento de calidad de software.

Paso 4.- Se asigna un pipeline de Jenkins al repositorio, este pipeline va contener validaciones de pruebas unitarias y de aceptación (cobertura de código con SonarQube) y se gestionará el despliegue a los ambientes productivos.

Paso 5.- Los desarrolladores implementan jobs y archivos de configuración sobre el repositorio creado para las ingestas al Data Lake, y también ejecutarándichos jobs un ambiente de desarrollo (sobre el motor de procesamiento Spark) a través de Jenkins (Como se describe en el paso 8)

Paso 6.- Jenkins notificará a los desarrolladores errores de compilación o de los test automáticos.

Paso 7.- Jenkins publicará los componentes de ingesta en el repositorio Artifactory, para que sean consumidos por los jobs en sus ejecuciones.

Paso 8.- Jenkins publicará los jobs en Automation Api, que es la pieza del entorno Big Data donde se almacenan los jobs bajo el motor de procesamiento Spark.

Paso 9.- Para programar los jobs por horarios, dependencias, en una malla Batch, se usa la herramienta Control-M.

Paso 10.- Automation Api puede lanzar cualquier jobSpark sobre los cores disponibles del entorno Big Data. (la interfaz para lanzarlos se vio en el punto 9, y para el ambiente de desarrollo es Jenkins)

Paso 11.- La herramienta para monitorear los jobs es Argos, donde muestran los log para identificar cualquier incidencia o alerta.