PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

Dirección de la Estrategia de Data en el grupo CREDICORP

Trabajo de suficiencia profesional para obtener el título profesional de

Ingeniero Informático

AUTOR

José Marcelo Almeyda Alcántara

ASESOR:

Rony Cueva Moscoso

Lima, Abril, 2021

1 RESUMEN

Luego de egresar de la carrera de Ingeniería Informática de la Pontificia Universidad Católica en 1998 ingrese a trabajar en el área de sistemas del Banco de Crédito del Perú en diferentes roles Programador, Analista y finalmente Arquitecto de Sistemas, luego de ello pase por otras de negocio del mencionado Banco para finalmente estar a cargo de la creación del Área de Data dentro del grupo Credicorp.

En Enero del 2015 el Banco de Credito del Peru empezó a delinear la creación del Área de Data y Analytics debido a la necesidad de contar con la información necesaria para poder cumplir con su principal objetivo “Transformar sueños en realidad”, antes de la creación de esta área, los datos no eran gobernados por una única entidad y estaba dispersa en las diferentes áreas de negocio y tecnología lo que llevaba a tener diversos problemas de disponibiliad, integridad y veracidad de la información, además del alto costo que conllevaba este modelo de trabajo.

A mediados del 2015 se me encargó crear y liderar el Área de Data con el objetivo principal de poner en valor los activos de información del BCP, al poco tiempo se incrementó el alcance de la función a todo el grupo Credicorp (BCP, Prima, Pacifico, Credicorp Capital y Mi Banco).

Para la realización de este encargo se dividió en 5 principales iniciativas, desarrolladas principalmente por el personal del BCP:

Gestion del recurso humano, el cual incluye organización, funciones, perfiles, capacitación y desarrollo de carrera dentro de un entorno de agilidad, esto conlleva a incluir especializaciones en lugar de estructura jerarquica asi como verdaderas evaluaciones 360.

Gobierno y Calidad de Datos, definición e implementación del gobierno de datos que permita tener una sola verdad en relacion a que significa cada dato y donde es posible encontrarlo complementandolo con los estandares de calidad de acuerdo a la criticidad del mismo,el entregable fue el diccionario de datos (20mil) de la organizacion.

Arquitectura de Datos basado en tecnologia de Big Data, definición e implementación de los diversos componentes de almacenamiento (data lake), explotación y visualización , carga de datos, gobierno y calidad, seguridad y streaming, finalmente se opto por el uso de tecnologia de Cloudera on-premise para el almacenamiento, Datameer y Qlik para explotacion y visualizacion, IBM Infosphere para la carga de datos de los aplicativos core y bases externas, Spark para la carga entre capas del datalake, kafka para el streaming de datos y Cloudera DataScience Workbench como herramienta de modelamiento estadisticos donde se podia programar en Python, R y Spark..

Cultura de Datos, definición e implementación de la metodología de cultura de datos como un segundo idioma que permita definir el nivel de madurez de cada área en termino de uso de datos en la toma de decisiones.

Data Enrichment, si bien la información que posee el grupo es relevante, es necesario enriquecer la información no solo con nuevos elementos de datos sino también actualizando los existentes de tal manera de tener información fiable.

Por otro lado se hizo necesario la creacion de un laboratorio de datos donde no solo se probaba tecnologia sino tambien permitia implementar soluciones que capturen mas datos para la tomar de decisiones.

Laboratorio de Big Data, definición e implementación del laboratorio de Big Data de tal manera que se pueda poner en valor de forma inmediata el uso de los datos sin esperar a que se culmine todo el proceso de carga de información, para esta labor se utilizo el framework Scrum para el desarrollo de productos de data y la Arquitecture de Big Data con herramientas de Microsoft Azure.

Dentro de las principales conclusiones que conllevaron al éxito en la implementacion de la estrategia de data se encuentran :

- El desarrollo de una estrategia de datos tiene diferentes aristas tecnológicas, culturales y de procesos que deben avanzar en paralelo para el mejor aprovechamiento del valor de la data.
- Es necesario un alineamiento de la estrategia de datos a la estrategia corporativa, de esta manera se asegura el soporte de la gerencia central.
- La estrategia de datos debe ser conocida por toda la organización y a todo nivel, debido a que es un proceso federado.
- Se deben desarrollar las capacidades técnicas del personal ya que el universo de personas que conocen tecnología de big data en el país es muy reducido.

Tabla de contenido

1 RESUMEN 2

2 PROYECTOS/ACTIVIDADES DESARROLLADOS 6

2.1 ESTRATEGIA DE DATOS/DESARROLLO DE LAS CAPACIDADES DE DATA 6
2.1.1 DESCRIPCIÓN Y CONTEXTO 6
2.1.2 CARACTERÍSTICAS DEL PROYECTO 7
2.1.3 OBJETIVO 7
2.1.4 ACTIVIDADES REALIZADAS 7
2.1.5 LOGROS OBTENIDOS 8
2.1.6 LECCIONES APRENDIDAS 8
2.2 ESTRATEGIA DE DATOS/GOBIERNO Y CALIDAD DE DATOS 9
2.2.1 DESCRIPCIÓN Y CONTEXTO 9
2.2.2 CARACTERÍSTICAS DEL PROYECTO 9
2.2.3 OBJETIVO 9
2.2.4 ACTIVIDADES REALIZADAS 9
2.2.5 LOGROS OBTENIDOS 11
2.2.6 LECCIONES APRENDIDAS 11
2.3 ESTRATEGIA DE DATOS/ARQUITECTURA DE DATOS 11
2.3.1 DESCRIPCIÓN Y CONTEXTO 11
2.3.2 CARACTERÍSTICAS DEL PROYECTO 12
2.3.3 OBJETIVO 12
2.3.4 ACTIVIDADES REALIZADAS 12
2.3.5 LOGROS OBTENIDOS 15
2.3.6 LECCIONES APRENDIDAS 15
2.4 ESTRATEGIA DE DATOS/CULTURA DE DATOS 16
2.4.1 DESCRIPCIÓN Y CONTEXTO 16
2.4.2 CARACTERÍSTICAS DEL PROYECTO 16
2.4.3 OBJETIVO 16
2.4.4 ACTIVIDADES REALIZADAS 16
2.4.5 LOGROS OBTENIDOS 17
2.4.6 LECCIONES APRENDIDAS 17
2.5 ESTRATEGIA DE DATOS/LABORATORIO DE BIG DATA 17
2.5.1 DESCRIPCIÓN Y CONTEXTO 17
2.5.2 CARACTERÍSTICAS DEL PROYECTO 18
2.5.3 OBJETIVO 18

2.5.4 ACTIVIDADES REALIZADAS 18
2.5.5 LOGROS OBTENIDOS 18
2.5.6 LECCIONES APRENDIDAS 19

3 CONCLUSIONES 19

4 CONSTANCIAS DE TRABAJO 21

Indice de Figuras

Figura 1. Tipología de dominios de datos 10
Figura 2. Componentes de una plataforma de datos 13
Figura 3. Software para cada capa del data lake 14
Figura 4. Herramientas para la explotación de la data 14
Figura 5. Esquema de transición de la plataforma analítica 15

2 PROYECTOS/ACTIVIDADES DESARROLLADOS

Desde Agosto del 2015 a la fecha como parte de la Dirección del Área de Data de Credicorp se ha ido estableciendo e iterando el plan director que permita poner en valor los activos de información del grupo, dentro de este plan se han desarrollo los siguientes proyectos.

2.1 Desarrollo e Implementacion de la Estrategia de Datos

El principal proyecto desarrollado fue la definición e implementación de la Estrategia de Datos, siendo este uno de los principales motores de la Transformación Digital de Credicorp, dentro del mismo se desarrollaron cinco principales iniciativas descritas a continuación.

Tabla 1. Datos del proyecto/actividad

Razón social:	BANCO DE CREDITO DEL PERU			**RUC:**	20100047218
Dirección administrativa:	David Saenz		**URL:**	www.viabcp.com	
Fecha de inicio:	04/2015		**Duración:**	36 meses	
Responsable:	Jose Marcelo Almeyda Alcantara		**Número de participantes:**	160	
Rol:	Gerente de Area de Data				
Referencias:	**Apellidos**	**Nombres**	**Teléfono**	**Correo**	**Cargo**
	Saenz Santolaya	David	949427357	davidsaenz@BCP.COM.PE	CDO/CTO
	Rard Linares	Karina	948648838	krard@BCP.COM.PE	Gerente Lab Big Data
	Jaramillo	Erick	948981789	ejaramillo@BCP.COM.PE	Gerente CoE Big Data
	Tay Wo	Shen	988560753	twchong@BCP.COM.PE	Gerente de Data Enrichment

2.1.1 Estrategia de Datos/Desarrollo de las capacidades de data

DESCRIPCIÓN Y CONTEXTO

Una de las principales problemáticas, que se presentaron a la hora de definir la estrategia de datos, era contar con el personal idóneo que permita realizar la implementación, en la región los especialistas en Data son escasos y si los hay no son especialistas en las nuevas tecnologías basadas en herramientas open source de Big Data, por otro lado definir claramente cuáles serán las especialidades y sus funciones son básicas, de tal manera que cada especialidad pueda ir desarrollándose dentro de una nueva forma de organización basadas en mindset Agile.

Los equipos de data se encontraban dispersos en la organización sin una clara línea de especialización y desarrollo, podíamos encontrar ingenieros de datos haciendo labores de modelamiento de datos o de consultores de datos y ademas estaban muy enfocados en tecnologías tradicionales de datawarehouse.

La experiencia en la implementación de proyectos basados en BigData en el 2015 era bajo, no se habían desarrollado grandes proyectos corporativos, por lo que era necesario tener bastante clara la organización que lo soporte.

CARACTERÍSTICAS

Para llevar a cabo este proyecto se tuvo que realizar un piloto que permita conocer de primera mano como las diferentes especialidades de data interactúan alrededor de un proyecto dentro de una arquitectura de Big Data. Este piloto consistió en la implementación de una web que permitia brindar información a través de un portal a los clientes negocio del BCP, al inicio se definieron más de 8 especializaciones que iban desde integradores de datos, visualizadores de data, modeladores, etc, los cuales finalmente quedaron en 5 que tenían funciones claramente diferenciadas y complementarias uno del otro, este proyecto se implementó en 5 semanas en su versión de producto minimo viable a un costo inicial de 35 mil dólares, el portal es www.crecemasbcp.com

A partir de establecer las especialidades se generó la nueva organización así como su matriz de desarrollo y necesidades de capacitación.

Para realizar estas definiciones la ingeniería informática ayudó en su capítulo de base de datos y gestión de proyectos de tal manera que se incluyan todas las funciones necesarias para un correcto manejo de la información

OBJETIVO

El objetivo es tener la organización idónea que permita desarrollar la estrategia de datos de la corporación.

ACTIVIDADES REALIZADAS

Implementación de un MVP de un portal basado en tecnología de Big Data, el portal www.crecemasbcp.com se desarrolló usando como base tecnología de Cloudera dentro de un entorno cloud de Azure.

Definición de las especialidades de data y su matriz de desarrollo.

Se definieron las siguiente especialidades :

- Data Engineer, responsable de la carga y mantenimiento de los datos.
- Data Modeler, responsable del modelamiento en las diversas capas del data lake.
- Data Governance, responsable de la definición e implementación del gobierno de datos y controles de calidad.
- Data Architect, responsable de la definición de la arquitectura de datos general y de cada proyecto así como de las políticas que lo gobiernan.
- Data Consultant, responsable de traducir las necesidades de información de las unidades de negocio en proyectos de datos.

Cada especialidad además tenía una matriz de desarrollo que permitía el crecimiento de los colaboradores de acuerdo al cumplimiento de metas, obtención de certificaciones, reconocimiento en el ámbito laboral y experiencia.

Cada especialidad está definida como un chapter con sus correspondientes líderes y son asignados a los squads que requieran dicha especialidad para poder cumplir con sus objetivos.

Definición de la organización

Se estableció un Centro de Excelencia de Big Data (COE) del cual dependían las diversas especialidades así como una tribu de data responsable exclusivamente de la carga de informacion y una tribu de data lab responsable del enriquecimiento de la data, además de los perfiles de data, se estableció la incorporación de otros roles de la corporación, agile coach, arquitecto de soluciones de seguridad que permitan complementar el equipo.

Definición de las necesidades de capacitación, para cada una de las especialidades se estableció un cronograma ad hoc de capacitación.

Plan de integración de todos los perfiles de data de la organización en un solo equipo, existían equipos de data que manejaban sus propios sanbox productivos, estos equipos además de no tener estándares ni una línea de carrera clara estaban desarrollando soluciones de datos fuera de un entorno controlado, lo que generaba problemas de seguridad e idoneidad de la información, frente a ello se estableció un plan para que estos equipos migren a cada una de sus especialidades dentro del COE de Data.

LOGROS OBTENIDOS

- Optimización del 15% de recursos necesarios para la implementación.
- Construcción de las capacidades necesarias para que el equipo pueda desarrollar sus actividades.
- Organización enfocada en cumplir objetivos claros y alineados a la estrategia de la organización.
- Estandarización del trabajo y funciones de cada especialidad.

LECCIONES APRENDIDAS

Al principio del proyecto se puede pecar de exceso, es decir contar la mayor cantidad de especialidades las cuales conforme avanza el proyecto se van autoregulando dentro de un ambiente agile.

Las capacitaciones tienen que ser continuas y no solo estar a cargo del empleador sino también fomentar el autaprendizaje a través de cursos online intervención en foros o dictado de cursos, el nivel de actualización de las herramientas open source es sumanete frecuente y el equipo sino tiene una cultura de autaprendizaje puede quedar desfasado.

Cada especialización en data dede ser complementaria, si bien se trabaja en un entorno colaborativo es necesario que no se traslapen sus funciones sino que se potencien, además el perfil de cada especialidad requiere skills personales y técnicos particulares.

No necesariamente una persona que haya tenido buena productividad en un ambiente tradicional de arquitectura de datos se adapta a las nuevas tecnologías, es necesario un cambio de mindset radical.

Todos las especialidades requieren al menos de un nivel mínimo de programación en algún lenguaje de datos.

2.1.2 Estrategia de Datos/Gobierno y Calidad de Datos

DESCRIPCIÓN Y CONTEXTO

La información dentro del Grupo Credicorp se encontraba dispersa, ya sea en repositorios centralizados como en SanBox administrados por cada una de las áreas, los problemas se presentaban cuando se tenía que tomar decisiones en base a información y cada una de las áreas tenía sus propios números, que habían sido tomados no solo de fuentes distintas sino también de comprensión distinta del significado de cada uno de los datos.

Por otro lado se tenían problemas de calidad de los datos, ya que nadie se hacía responsable de la idoneidad de la data y por lo tanto información que era obtenida por una determinada área y consumida por otra no tenían los controles de calidad necesarios, lo cual generaba innumerables reprocesos y falta de confianza en la información para la toma de decisiones.

En este aspecto la creación y desarrollado de un programa de gobierno de datos se hacia necesario, este programa si bien se manejaba de forma centralizada por el Área de Data luego debía formar parte del mindset y hacer responsable a los Data Owner de los negocios de la calidad de su información .

CARACTERÍSTICAS

Para llevar a cabo este proyecto se tuvo que trabajar en dos fases claramente definidas, la primera fase basada en la creación de la metodología de gobierno y calidad en ésta se definió la forma en la cual los datos iban a ser administrados y la forma en la cual se iba trabajar de la mano con las unidades de negocio, la segunda fase se concentró en la implementación de la metodología la cual estuvo basada en olas trimestrales, de acuerdo a la cantidad de dominios de datos que se habían establecido para la organización.

La primera fase tomó el primer año, en la cual además de la metodología se estableció a IBM InfoSphere como la herramienta que permitía realizar la gestión del gobierno y calidad de datos, en esta fase también se definieron los 32 dominios de información en los cuales se dividió la organización.

La segunda fase se estableció por olas semestrales en las cuales de acuerdo a la complejidad del dominio se iniciaba el gobierno con un especialista de data governance asignado, este proceso es continuo y aún se viene desarrollando.

OBJETIVO

El objetivo es contar con los datos confiables y totalmente definidos de tal manera que se pueda tomar decisiones en base a esta información .

ACTIVIDADES REALIZADAS

Assesment inicial para conocer la línea base en la que se encontraba la gestión de datos, se encontró un nivel intermedio/bajo, ya que los datos no tenían una definición en mucho de los casos y se encontraba almacenada en diversos repositorios, lo que hacía difícil poder obtener y compartir la

información, muchas de las veces los reportes generados tenían que volverse a hacer porque solo una persona conocía qué significaban los datos.

Definición de los dominios de datos, como se muestra en la Fig 1 se definieron 32 dominios de datos, cada dominio de datos tenía que cumplir las siguientes características:

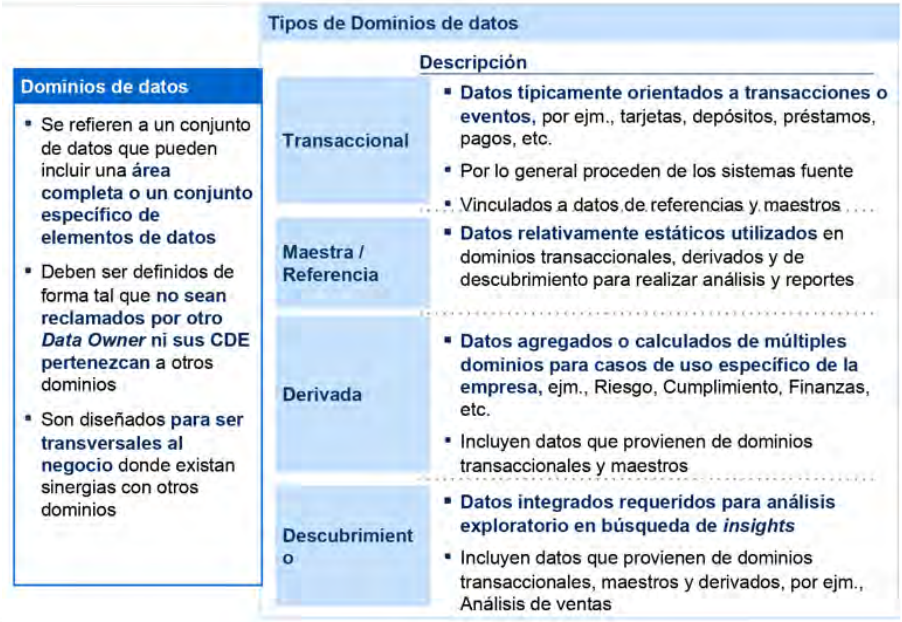

Figura 1. Tipología de dominios de datos

Definición y asignación de los roles para cada uno de los dominios, se establecieron los siguientes roles :

- Data Owner, responsable de la gestión efectiva de cada dominio de datos, naturalmente debía ser un gerente del área responsable.
- Data Steward, responsable de la gestión de la calidad del dato y definición de nuevos elemento de datos, si se crearan nuevos.
- Platform Steward, responsable de la revisión y solución técnica si hubiera un problema con la calidad de la información .

Definición y establecimiento de los conceptos de golden source, metadata, trazabilidad y priorización de los elementos de datos, todos estos elementos estaban claramente definidos en las políticas desarrolladas por el equipo de gobierno y calidad de datos, cualquier dato debía tener estas características desarrolladas para que se considere un dato válido que podía ser utilizado dentro de la organización.

Definición de los comités que aseguren el gobierno del programa, estos comités van desde las gerencias generales de cada empresa hasta los owners, los cuales participan en comité con diversa frecuencia desde semestral hasta mensual, en la cual se discutía y priorizaba las necesidades de gestión del gobierno y la calidad de datos de cada uno de sus dominios, además de la gestión presupuestal y revisión de avances e incidentes.

Definición y desarrollo de los dashboard de calidad de datos, para cada uno de los elementos de datos de los dominios de acuerdo a su criticidad se establecían controles de calidad técnicos y de negocio, para lo cual se desarrollaban dashboards de control de calidad que eran desarrollados por los data engineer y admistrados por los data steward.

LOGROS OBTENIDOS

- Implementación del Gobierno y Calidad de Datos en 23 de los 32 dominios en el plazo de 2 años.
- 5 mil elementos de datos registrados en el diccionario de datos disponibles para toda la organización.
- Definición del programa de Gobierno de Datos para 4 de las 5 subsidiaras de Credicorp.

LECCIONES APRENDIDAS

La priorización del desarrollo del programa de gobierno y calidad debe estar dado por aquellos dominios en los que se tenga un problema importante de calidad de datos y el owner tenga el compromiso de resolverlo.

Se debe tener menos del 10 por ciento del total de elementos de datos como datos críticos pues el costo de colocar los controles de calidad es alto.

No se debe pasar a producción ningún dato que no tenga definidos claramente sus controles de calidad y su información contenida en el diccionario de datos comportativos.

Dejar muy en claro que el ser dueño del dato signfica ser el responsable de que el dato se encuentre correctamente definido y validado y disponible para toda la organización.

Los data stewards son perfiles de negocio con ciertas capacidades técnicas de tal manera que puedan comprender en que procesos de negocio se generan y se utilizan los datos.

El equipo de gobierno de información inicia el proceso de gobierno y calidad el cual debe ser culminado y reportado por la respectiva unidad de negocio.

2.1.3 Estrategia de Datos/Arquitectura de Datos

DESCRIPCIÓN Y CONTEXTO

Los cambios en el entorno hicieron que tanto la corporación Credicorp como BCP empiecen a trabajar en la definición de un nuevo objetivo, el cual se ha resumido en transformar sueños en realidad, esto decanta a que la arquitectura de datos que se defina permita no solamente conocer mejor a los clientes para poder cumplir sus sueños sino que también éstos sean logrados de forma oportuna es decir cuando el cliente lo necesite, a partir de estas definiciones de negocio se ha construido una nueva arquitectura de datos, el concepto de conocer mejor al cliente en base a la mayor cantidad de información ha decantado en una plataforma de Big Data donde se almacene y procese no solo la data tradicional sino también datos que vengan en diversos formatos y

fuentes, por otro lado la importancia de satisfacer las necesidades de los cliente en el momento en que los necesite se ha convertido en un requerimiento de arquitectura que soporte este procesamiento de grandes volúmenes de información en tiempo real y que de soporte a los algortimos de machine learning que se ejecuten sobre ello.

La arquitectura de datos que se tenía era diversa en toda la corporación, por ejemplo en el BCP el 50 por ciento de los datos se encontraba en un repositorio Oracle con un modelo de datos desarrollados inhouse y el otro 50 dispersos en los diferentes servidores no oficiales con tecnología diversa en la organización, en otras empresas solo se contaba, en el mejor de los casos, datamarts y en otras repositorios SQL, así mismo las fuentes desde donde se obtenía la información también era diversa, desde mainframes como OS390 hasta soluciones completas como Topaz, pasando por entradas de datos automatizados, esto llevaba a que la definición de la arquitectura debería contemplar diferentes estadíos para tener finalmente una arquitectura de datos que procese grandes volúmenes de información en tiempo real.

CARACTERÍSTICAS

El presupuesto inicial se calculó en 50 millones de dólares y un equipo dedicado de 160 personas, los 4 primeros años eran intensivos en definiciones y elección de tecnologías para posteriormente enfocarse en la implementación e integración de los mismos, finalmente se realizaría el proceso de adopción masiva del modelo.

Se tenía la alternativa de trabajar la solución de manera onpremise o cloud, sin embargo aún existen restricciones tanto legales (datos sensibles en la nube) o de plataforma (ancho de banda en el país) que se tenían que tener presente para la definición de la solución .

Se trabajó bajo una metodología ágil el proyecto, creando los diversos squads y especializaciones que permitió ir brindando valor en cada uno de los sprints.

Si bien la definición de la solución era una potestad del Área de Data existían equipos participantes de diferentes áreas del banco como seguridad, negociacio, sistemas, con los cuales se tenía que tener una cercanía bastante estricta para ir avanzando en el logro de objetivos.

OBJETIVO

El objetivo es contar una arquitectura de datos que soporte la estrategia de negocio definida por la organización.

ACTIVIDADES REALIZADAS

Definición de los componentes y tecnología necesaria para la nueva arquitectura, en primer lugar se estableció una arquitectura lógica sobre la cual se monte la nueva arquitectura de datos, esta arquitectura cumplía con las premisas de poder procesar grandes volúmenes de información en tiempo real, la capa principal de almacenamiento se dividió en 4 sectores que se pueden ver en la Figura 2.

Raw Data, donde los datos se almacenaban en una estructura simple sin ningún o poca modificación desde sus sistemas de origen, esto permitía que toda la información se tenga disponible en todo momento en esta zona, de tal manera que el time to market de los proyectos de data sea menor.

Universal Data, en esta zona se encuentran ya los datos modelados, si bien no tienen una estructura muy normalizada, aquí se encuentran los datos ya disponibles para ser consumidos con los controles de gobierno y calidad definidos e implementados.

Dimensional Data, es lo más similar a los datamarts, se encuentra la información que puede ser consumida o generada por una determinada área de negocio, en caso se generen datos importantes para la organización ésta se trasladaba luego al Universal Data.

Experimental Data, son los sanbox administrados, se crean por un tiempo determinado para que puedan probar la validez de una determinada hipótesis en base a datos y luego esta zona se elimina.

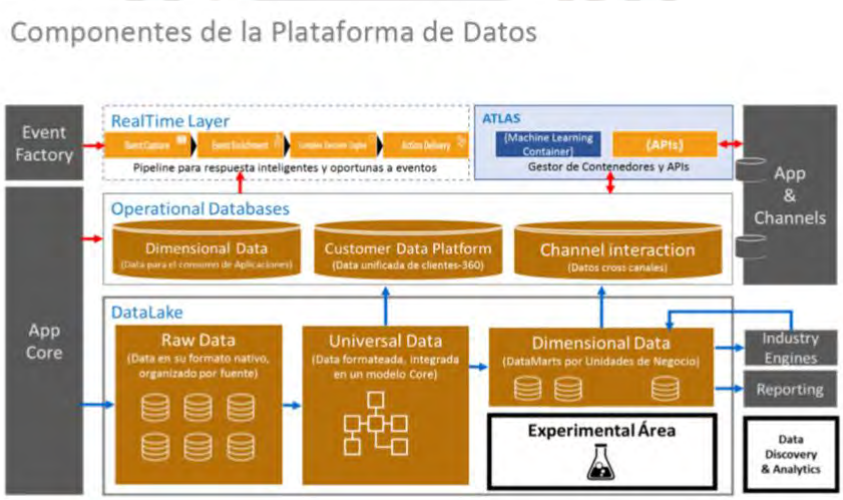

Figura 2. Componentes de una plataforma de datos

En cuanto a la tecnología se evaluaron diversas herramientas bajo diferentes criterios de aceptación tecnológicos, de seguridad, costo y valor para el usuario final, a partir de ello se seleccionaron las tecnologías para cada una de las capas, las cuales se muestran en la Figura 3.

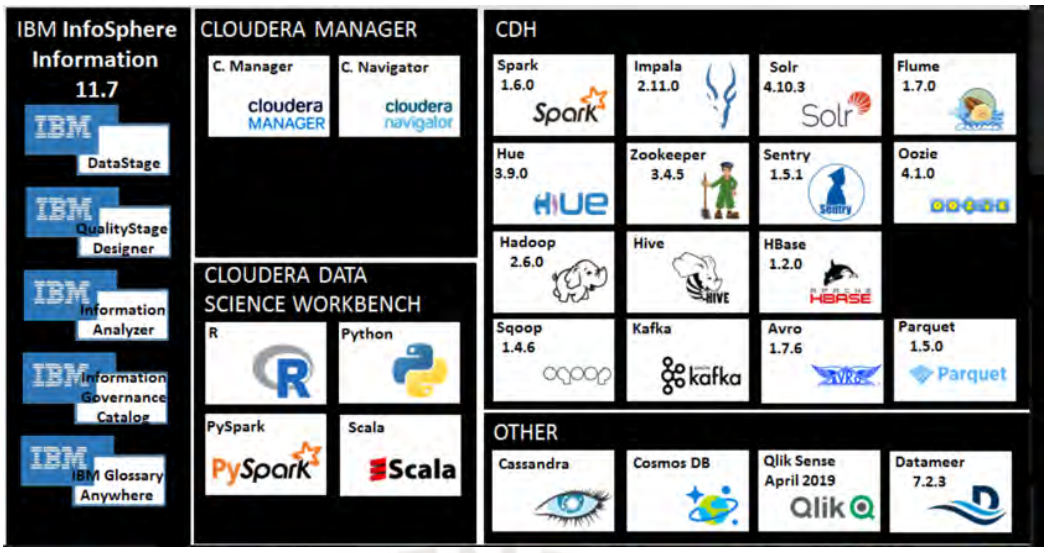

Figura 3. Software para cada capa del data lake

Por otro lado se definieron las herramientas de explotación de acuerdo al perfil de cada uno de los grupos de usuario los cuales se muestran en la Figura 4.

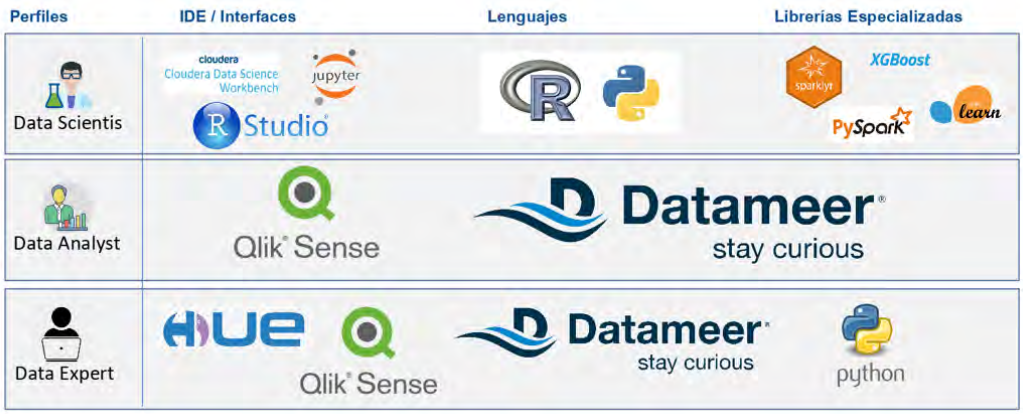

Figura 4. Herramientas para la explotación de la data

Definición de lineamientos para la implementación de la arquitectura y el poblado, se estableció el orden en el cual se deberían cargar los datos de acuerdo a los dominios de información definidos por el equipo de gobierno de información.

Definición del modelo de datos del universal data vault, el modelamiento dentro de un entorno de Big Data al no ser normalizada requiere de un modelo adhoc, sin embargo la falta de especialistas en esta área hizo que se trabaje un modelo híbrido basado en el modelo de datos existen en el BCP y el modelo OFFSA de Oracle, a partir de ello se inició un proyecto adicional que incluía la implementación de un modelo de datos propio que de la flexibilidad necesaria para toda la corporación .

Definicion del Plan Director que incluye el esquema de transición y la inclusión de los servidores de datos no oficiales existentes, la necesidad

de dar valor dentro de un proyecto de largo plazo hizo que se maneje un esquema de transición en la cual se use la infraestructura onpremise ya implementada pero usando como fuente de datos no los aplicativos fuente sino el datawarehouse, esto hizo que se pueda ir adelantando también la migración de los modelos analíticos a la nueva plataforma.

Figura 5. Esquema de transición de la plataforma analítica

Definición y desarrollo del modelo UDV que sirva de fuentes de datos para los modelos analíticos, el valor del data lake en cuanto a uso de diversas fuentes de datos y alta capacidad de procesamiento se hace visible con mayor presencia en el desarrollo de los modelos analíticos, es por ello que ha sido lo primero que se ha implementado usando principalmente tecnología Spark.

LOGROS OBTENIDOS

- Carga en un año del 100 por ciento de la información en la capa Raw data vault (RDV) del DataLake, la cual estaba disponible para el uso por el equipo de data scientist.
- Reducción del 90 por ciento del costo de almacenamiento/procesamiento en relación con la arquitectura tradicional.
- Disminución en casi un 95% el tiempo de procesamiento de algoritmos de machine learning del área de riesgos.
- Implementación en dos años del 100 por ciento de componentes tecnológicos que den soporte a la arquitectura de Big Data.
- Modelos predicitivos usando la nueva plataforma.

LECCIONES APRENDIDAS

La selección de herramientas debe tener un componente importante de participacion de las diferentes áreas usuarias dependiendo de su uso, por ejemplo los data scientist para las herramientas de creación de modelos estadísticos, los analistas para las herramientas de visualización, para ellos se hizo una feria en la cual todos los vendors exponían sus productos para que puedan ser testeados por los integrantes de los equipos de las áreas de negocio.

Las Pruebas de Concepto son sumante importantes para validar lo que en el papel podrían o deberían hacer determinadas soluciones, sin una prueba de concepto no se debería escoger una solución.

La validación de qué se puede hacer y qué no desde el punto de vista legal, tecnológico, procesos, seguridad es algo que se debería hacer al inicio del proyecto a partir de lo cual se define la mejor solución.

El perfil del arquitecto de datos que soporte las decisiones debería ser de personas que ya hayan participado en implementaciones de arquitectura de Big Data.

2.1.4 Estrategia de Datos/Cultura de Datos

DESCRIPCIÓN Y CONTEXTO

Si bien en la corporación no se es ajeno a reconocer la importancia de los datos y su valor, no necesariamente existe una madurez en las diferentes áreas de tal manera que se tenga una corporación 100 por ciento orientado a datos, aprender a trabajar, leer, analizar y discutir es similar a aprender un nuevo idioma, es en este contexto que se creó el programa de cultura de datos.

CARACTERÍSTICAS

Este proyecto se inició en paralelo al programa de datos, siendo un proceso continuo que conlleve a interiorizar a la organización del valor de los datos y no solamente para sus áreas sino también para toda la organización, cambiar el mindset de decir “son los datos de finanzas” a “son los datos financieros de la corporación” hace una diferencia sustancial en entender a los datos como un segundo idioma.

Se tuvo un equipo dedicado de 5 personas que trabajaron la metodología de trabajo y tal como aprender un idioma más se desarrollaron diversas formas de soporte para lograr este objetivo.

OBJETIVO

El objetivo es tener una organización 100 por ciento orientado a datos (data driven).

ACTIVIDADES REALIZADAS

- Definir el modelo de maduración de cultura de datos**,**
- Estructurar talleres y recursos basados en los principios de cultura de datos.
- Ejecutar espacios abiertos y charlas con especialistas
- Definir el manifiesto de cultura de datos
- Desarrollar el plan de comunicación a través de las diversas plataformas disponibles en la organización.
- Fundar la Academia de Datos

.

LOGROS OBTENIDOS

- En un año el 60 por ciento de la organización estuvo presente al menos en un taller de Cultura de datos.
- Definición de la línea base en nivel de cultura de datos de toda la organización.

LECCIONES APRENDIDAS

- Es necesario objetivizar el nivel de cultura de datos, similar a catalogar el nivel de conocimiento de un segundo idioma.
- Es posible encontrar nuevas fuentes de datos en los talleres con las diversas áreas de negocio, no solamente de datos con que se cuenten sino también con datos faltantes que se pueden conseguir.
- Es necesario un equipo dedicado a este trabajo, ya que complementa la visión del programa de datos haciendo que éstos sean utilizados de forma eficiente.

2.2 Proyecto de implementacion de un Laboratorio de Big Data

Desarrollar rápidamente las capacidades tecnológicas del equipo, así como probar nuevas tecnologías de big data hicieron necesario la creación de un Laboratorio de Big Data, se enfoco principalmente a generar valor temprano a la organización a través del desarrollo de productos de data.

Tabla 2. Datos del proyecto/actividad

Razón social:	BANCO DE CREDITO DEL PERU			**RUC:**	20100047218
Dirección administrativa:	David Saenz		**URL:**	www.viabcp.com	
Fecha de inicio:	04/2016		**Duración:**	36 meses	
Responsable:	Jose Marcelo Almeyda Alcantara		**Número de participantes:**	20	
Rol:	Gerente de Area de Data				
Referencias:	**Apellidos**	**Nombres**	**Teléfono**	**Correo**	**Cargo**
	Saenz Santolaya	David	949427357	davidsaenz@BCP.COM.PE	CDO/CTO
	Rard Linares	Karina	948648838	krard@BCP.COM.PE	Gerente Lab Big Data
	Jaramillo	Erick	948981789	ejaramillo@BCP.COM.PE	Gerente CoE Big Data
	Tay Wo	Shen	988560753	twchong@BCP.COM.PE	Gerente de Data Enrichment

2.2.1 DESCRIPCIÓN Y CONTEXTO

Se creó el Laboratorio de Big Data en base a la necesidad de contar con mayor información de los clientes más allá de lo transaccional, si bien la

participación de mercado de las empresas Credicorp es alta muchas veces se circunscriben a información financiera y demográfica, los modelos predictivos de riesgos necesitan calibrarse cada vez con mayor rapidez y por el lado comercial conseguir mayores y mejores insigths.

2.2.2 CARACTERÍSTICAS DEL PROYECTO

Un equipo dedicado de 20 personas con un presupuesto por producto minimo viable, entre 30 a 50 mil dólares se dedicaban a desarrollar soluciones que permitan capturar mayor información de los clientes basadas en una arquitectura de big data, esta captura de datos venian de desarrollo de aplicaciones, tecnicas de webscraping y desarrollo de algoritmos que permtian refinar datos (ejm, completar numeros de telefono o direciones)

Para el caso de las aplicaciones, el MVP de los mismos se desarrollaba en 5 semanas, luego de este tiempo se lanzaba al mercado con una estrategia de growth hacking intensiva.

Cada equipo estaba conformado por 1 Product Owner, 1 Scrum Master y dentro de los team members perfiles de UX (Experiencia de usuario), UI (interfaz de usuario), Frontend (desarrollo de la capa interfaz de usuario) y Backend (desarrollo de la capa interna), Data Engineers (ingenieros de datos), utilizando Scrum como framework Agile.

2.2.3 OBJETIVO

Capturar la mayor cantidad de datos provenientes de datasets internos y externos de data estructurada y no estructura asi como el uso de apps para conocer mejor los gustos de los clientes y su nivel de riesgo.

2.2.4 ACTIVIDADES REALIZADAS

Desarrollo de la Mesa “Berlín”, este squad se encargó de capturar información disponible en internet a través de técnicas de webscrapping, una muestra de estos datos pasaba a las unidades de riesgos para verificar si el dato mejoraba el gini de los modelos de riesgos.

Implementación del App Parati, Aplicación que agrupaba todos los descuentos existentes en el mercado y usando un modelo de recomendación en tiempo real con data de geolocalización le mostraba descuentos personalizados para cada cliente, se podía capturar con esta app además de los gustos de los clientes su id de Facebook y su geolocalización.

Implementación del App Manyar, Aplicación que agrupa todos los menús de los restaurantes y recomendaba el restaurante y el plato para cada cliente.

2.2.5 LOGROS OBTENIDOS

500 mil usuarios registrados en 15 meses en las aplicaciones desarrolladas por el Lab.

70% de los usuarios se registran con su usuario de facebook.

125 fuentes de dato externas analizadas y potenciales de uso en los modelos de riesgo.

Incremento de 30% a 60% de teléfonos válidos de clientes BCP

2.2.6 LECCIONES APRENDIDAS

- La aplicación de la analítica para el growth de las aplicaciones es primordial, se ha mantenido el costo de adquisición menor a 1 dólar por usuario.
- Los clientes brindan su información solo si esta le es devuelta con mayor valor.
- La capacidad de inmediatez es la mas usada por lo que se debería prioriar en cualquier desarrollo.

3 CONCLUSIONES

El desarrollo de un programa de datos corporativos, debe estar sustentado en el alineamiento a la estrategia del negocio, a partir de ellos se debe contemplar diversas aristas siendo el corazón la arquitectura tecnológica .

La gestión del activo de información parte de saber con que información se cuenta y cual es el gap con las necesidades de información, a partir de ello se abren dos caminos, el primero que garantice que la data con que se cuenta tenga la calidad necesarias y el segundo camino viene de disminuir el gap entre las necesidades de data y lo que se tiene, este nuevo proceso se conoce como data enrichment.

No basta solo con tener la data almacenada, estructurada y disponible sino que tambien se debe fomentar su uso a través de un programa de cultura de datos, esto hace que la inversión que se ha realizado en el programa de datos brinde el retorno correspondiente.

La data al ser un activo se tiene que trabajar bajo la metodología de gestión de activos de tal manera que no pierda su valor sino que se multiplique, este valor puede ser dado como sustento para la toma de decisiones pero también como parte de un nuevo modelo de negocios que genere ingresos a partir de la comercialización de esta información en forma de insigths, canales de venta y otros.

La complejidad de ser la solución multiempresa decanta en alinear las necesidades y en lo mínimo tropicalizar soluciones particulares, se ha encontrado que por ejemplo para la empresa de seguros se necesitan componentes propios de la naturaleza del negocio, si bien se trata de tener un estándar corporativo estas adecuaciones se deben documentar de tal forma que se mantenga una concordancia con desarrollos siguientes.

La velocidad con la que los componentes arquitectónicos se actualizan o aparecen nuevos hace que tengamos que tener un equipo dedicado a ver que viene en el futuro, en que se encuentra trabajando la comunidad open source y las empresas líderes, si es posible ser beta testers de alguna de estas empresas sería el ideal, de esta manera se deja listo para que un nuevo componente se integre o salga de la arquitectura actual definida.

La velocidad con la que los proyectos de datos se implementan debe ser de tal manera que acompañe las necesidades de negocio, es por ello que la inclusión de una línea de dataops que haga que los pases a producción de las soluciones de data se hagan con un solo click es clave para poder soportar esta necesidad.

Dirigir un cambio importante no solo tecnológico sino de mindset del valor de la información ha hecho que pueda aplicar los conocimientos adquiridos de gestión de proyectos de información, bases de datos, soluciones tecnológicas, desarrollo de sistemas y gestión de operaciones al ser una solución transversal a la organización.

Se comprobó que había una gran brecha entre los contenidos académicos de las carreras de sistemas en el país y lo necesario para poder desarrollar una plataforma de Big Data y Analítica en el país, esta brecha si bien es cubierta parcialmente por cursos online o cursos de los proveedores de tecnología no tiene la rigurosidad académica necesaria.

4 CONSTANCIAS DE TRABAJO

Gerencia

La Molina, 03 de diciembre de 2019

Señor(es)
A quien corresponda

Presente.-

Constancia de Trabajo

Dejamos constancia que el señor **JOSE MARCELO ALMEYDA ALCANTARA**, labora en nuestra Organización desde el 01 de enero de 2009, desempeñándose actualmente en la GERENCIA DE DIVISION DATA & ANALYTICS como COE LEADER.

Se expide el presente a solicitud del interesado para los fines que estime conveniente.

Atentamente,

PAMELA LLERENA V.
Gerente Adj. Soporte Comp. Ejecutiva - Talento
División de Gestión y Desarrollo Humano

División de Gestión y Desarrollo Humano