Capítulo 3

Exploración del Big Data

Al finalizar el capítulo, el alumno podrá:

- Comprender los mecanismos de la importancia del big data para los clientes.
- Operatividad y consideraciones de implementación del big data

Temas

1. La vista 360 del Cliente
2. Análisis, Diseño y Operación del Big Data
3. Procesamiento del Big Data
4. Seguridad en el Big Data

1. La vista 360 del Cliente

Muchas organizaciones están en la búsqueda del cliente 360 con el objetivo de integrar la información del cliente a través de múltiples canales, sistemas, dispositivos y productos para mejorar su experiencia con la marca y maximizar el valor entregado.

Además, Big Data y la analítica están cambiando la forma en que las organizaciones toman sus decisiones. Las empresas tienen cada vez tienen más información acerca de las necesidades de sus clientes y sus preferencias.
Con este panorama, se hace imprescindible tener una visión cliente 360 de todos los clientes.

Cliente 360 es más que una estrategia, se trata de una visión que encaja en la cultura de empresa y se reconoce en cada decisión, en cada acción y en cada una de las interacciones que esa organización tiene con sus consumidores, pero también con sus socios, proveedores, clientes potenciales.

La visión de cliente 360 es

- Coherencia. Una visión completa de los clientes a través de productos, sistemas y canales de interacción en la que la diversidad no aliena la coherencia en ningún momento.
- Atemporalidad. Se trata de poner la vista en el presente, pero recogiendo el conocimiento extraído del pasado y manteniendo la atención en el futuro y lo que con él llegará.
- Profundidad. Gracias a customer 360 es posible alcanzar una mejor comprensión de cada una de esas personas que llamamos clientes. Es el punto de partida para unas relaciones más auténticas y duraderas.

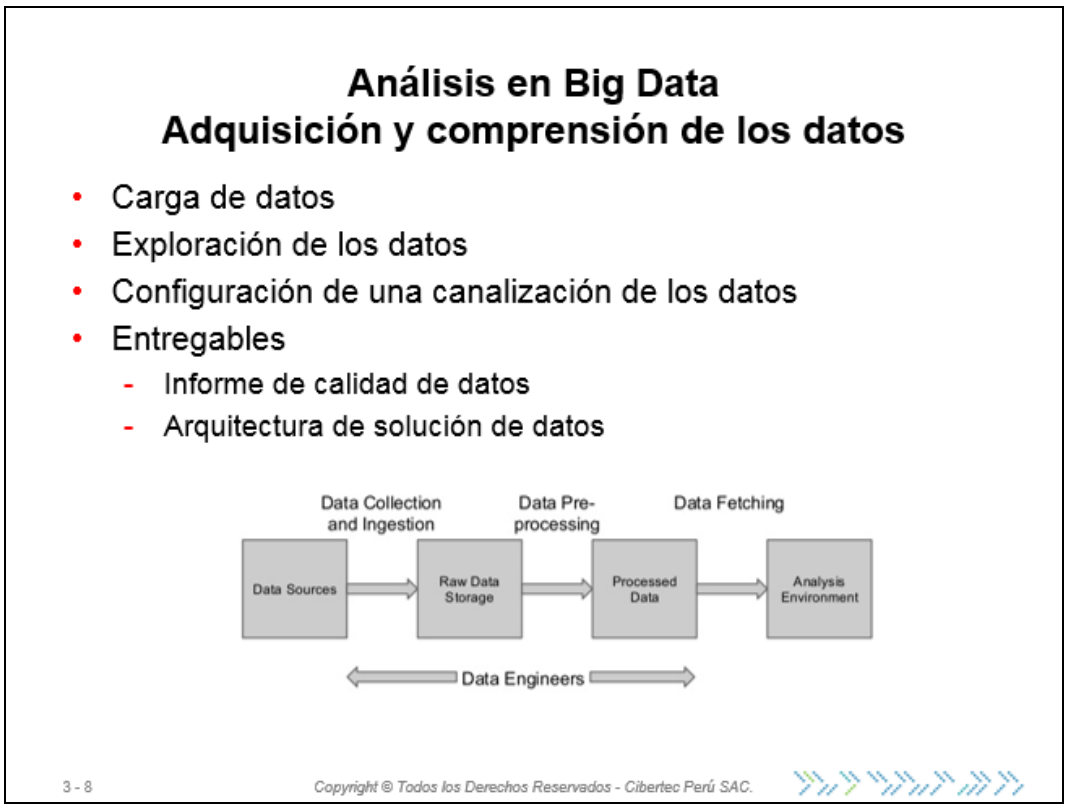
- Ilusión. Para muchas organizaciones líderes, la visión de cliente 360 es una forma de responder a la convergencia de las expectativas de sus clientes.
- Vehículo. Customer 360 reúne todos los atributos que el mercado percibe de la marca.

Construyendo un Customer 360

El viaje hacia la creación de una visión holística de los clientes debe comenzar con la elección tecnológica adecuada. Ahí están los medios. Es el punto de partida al que deben seguir decisiones acertadas en lo que respecta a la forma de usar todo ese poder. Para alcanzar la meta customer 360 es preciso:

- Definir una identidad de cliente: que refleje su carácter único, su personalidad, sus expectativas, entre otros datos.
- Agrupar datos internos: reuniendo información de transacciones, datos recogidos de servicio al cliente y de las encuestas o estudios de satisfacción del cliente. Sin olvidar, por supuesto, la actividad del cliente en los sitios web de la empresa, programas de fidelización y redes sociales.
- Recopilar datos de datos de fuentes externas: se trata de información de tendencias y eventos. Aquí se englobarían todos los aspectos relacionados con el estilo de vida, los perfiles geográficos y demográficos, los hábitos y también todos los datos que permitan llevar a cabo un análisis de la competencia.
- Aplicar técnicas de análisis predictivo: combinando toda la información recogida se pueden aplicar técnicas de analítica avanzada que permitan ganar ese conocimiento Cliente 360 que perseguimos. Así, será posible pronosticar la probabilidad de compra de los clientes, la eficacia de acciones de ventas y ventas cruzadas o las posibilidades de conversión de clientes potenciales.

2. Análisis, Diseño y Operación del Big Data



Primero se debe definir los objetivos del proyecto.

Objetivos

- Especifique las variables principales que sirven como objetivos del modelo y cuyas métricas asociadas se utilizan para determinar el éxito del proyecto.
- Identifique los orígenes de datos pertinentes a los que tiene acceso la empresa o que necesita obtener.

Modo de hacerlo

En esta fase se abordan dos tareas principales:

Definición de objetivos: trabaje con el cliente y con otras partes interesadas para comprender e identificar los problemas de la empresa. Formule preguntas que definan los objetivos empresariales y a las que puedan aplicarse las técnicas de ciencia de datos.

Identifique los orígenes de datos: busque los datos pertinentes que lo ayuden a responder a las preguntas que definen los objetivos del proyecto.

Definición de objetivos

1. Un objetivo fundamental de este paso consiste en identificar las principales variables empresariales que el análisis deberá predecir. Estas variables se denominan *objetivos del modelo* y las métricas asociadas a ellas se utilizan para determinar el éxito del proyecto. Dos ejemplos de estos destinos son la previsión de ventas o la probabilidad de que un pedido sea fraudulento.

2. Para definir los objetivos del proyecto, plantee y ajuste preguntas "certeras" que sean pertinentes, específicas y sin ambigüedad alguna. La ciencia de datos es un proceso que utiliza nombres y números para responder a estas preguntas. Para obtener más información sobre cómo hacer preguntas difíciles, consulte el blog How to do data science (Realización de procesos de ciencia de datos). La ciencia de datos o el aprendizaje automático suelen utilizarse para responder a cinco tipos de preguntas:

   ¿Cuánto? o ¿cuántos? (regresión)
   ¿Qué categoría? (clasificación)
   ¿Qué grupo? (agrupación en clústeres)
   ¿Es extraño? (detección de anomalías)
   ¿Qué opción se debe elegir? (recomendación)
   Determine cuál de las siguientes es su pregunta y cómo la respuesta logra sus objetivos empresariales.

3. Defina el equipo del proyecto especificando los roles y las responsabilidades de sus miembros. Desarrolle un plan general de hitos que se pueda repetir a medida que se descubra más información.

4. Defina las métricas del éxito. Por ejemplo, podría desear una predicción sobre la renovación de los clientes. Necesita una tasa de precisión de "x%" al final de este proyecto de tres meses. Con estos datos, puede ofrecer promociones al cliente para mejorar la fidelización. Las métricas deben cumplir los requisitos **SMART**:

   - **S**pecific (específicas)
   - **M**easurable (mensurables)
   - **A**chievable (alcanzables)
   - **R**elevant (pertinentes)
   - **T**ime-bound (con un límite de tiempo)

Identificación de los orígenes de datos

Identifique los orígenes de datos que contienen ejemplos conocidos de respuestas a las preguntas certeras. Busque los siguientes datos:

- Datos pertinentes para la pregunta. ¿Tenemos indicadores para medir el objetivo y las características que están relacionados con él?
- Datos que representen una medida precisa de nuestro objetivo de modelo y de las características de interés.

Por ejemplo, puede descubrir que los sistemas existentes tienen que recopilar y registrar tipos de datos adicionales para solucionar el problema y alcanzar los objetivos del proyecto. En esta situación, puede ser conveniente buscar orígenes de datos externos o actualizar los sistemas para recopilar datos nuevos.

Artefactos

Estos son los resultados de esta fase:

- Documento marco: El documento marco es un documento en cambio continuo. La plantilla se actualiza a lo largo del proyecto a medida que se descubren nuevos elementos y cambian las necesidades empresariales. La clave consiste en realizar iteraciones de este documento e incorporarle la información oportuna según se avance a lo largo del proceso de descubrimiento. Es importante que el cliente y las demás partes interesadas se impliquen en la realización de cambios y que se les informe claramente sobre las razones que los motivan.
- Orígenes de datos: En esta sección se especifican las ubicaciones originales y de destino para los datos sin procesar. En las fases posteriores, deberá rellenar más detalles, tales como los scripts para mover los datos al entorno de análisis.
- Diccionarios de datos: este documento proporciona descripciones de los datos facilitados por el cliente.Estas descripciones incluyen información sobre el esquema (tipos de datos e información sobre las reglas de validación, si hay) y los diagramas de relación de entidades, si están disponibles.

Adquisición y comprensión de los datos

Segundo se debe definir el modelo de carga de datos y la calidad de datos.

Objetivos

- Genere un conjunto de datos limpio y de alta calidad cuya relación con las variables de destino se entienda. Busque el conjunto de datos en el entorno de análisis de adecuado para prepararse para el modelado.
- Desarrolle una arquitectura de solución de la canalización de datos que actualice y puntúe los datos con regularidad.

Modo de hacerlo

En esta fase se abordan tres tareas principales:

- **Carga de los datos** en el entorno de análisis de destino.
- **Exploración de los datos** para determinar si su calidad es suficiente para responder a la pregunta.
- **Configuración de una canalización de datos** para puntuar los datos nuevos o que se actualizan con regularidad.

Carga de los datos

Configure el proceso para mover los datos desde las ubicaciones de origen a las ubicaciones de destino donde se ejecutan las operaciones de análisis, como el entrenamiento y las predicciones.

Exploración de los datos

Antes de entrenar los modelos, debe desarrollar una comprensión sólida de los datos. A menudo, los conjuntos de datos reales contienen ruido, les faltan datos o presentan un sinfín de discrepancias de otros tipos. Puede utilizar funciones de resumen y visualización de los datos para auditar su calidad y dar la información que

se necesita para procesarlos y dejarlos preparados para el modelado. Normalmente, se trata de un proceso iterativo.

Una vez que esté satisfecho con la calidad de los datos limpios, el siguiente paso es comprender mejor los patrones que son inherentes a los datos. Esto ayuda a elegir y desarrollar un modelo de predicción adecuado para el destino. Busque pruebas que describan la conexión de los datos con el destino. A continuación, determine si hay suficientes datos para avanzar con los siguientes pasos de modelado. Como hemos indicado, normalmente, se trata de un proceso iterativo. Es posible que deba buscar otros orígenes de datos con información más precisa o pertinente con el fin de alimentar el conjunto de datos inicialmente identificado en la fase anterior.

Configuración de una canalización de datos

Además de la introducción y limpieza iniciales de los datos, suele ser preciso configurar un proceso para puntuar los datos nuevos o actualizarlos con regularidad durante el proceso de aprendizaje continuo. Para ello, puede configurar una canalización de datos o un flujo de trabajo.

En esta fase, desarrolla una arquitectura de solución de la canalización de datos. Desarrolla la canalización en paralelo con la siguiente fase del proyecto de ciencia de datos. En función de las necesidades empresariales y de las limitaciones de los sistemas existentes en los que se integre esta solución, la canalización puede ser de uno de los tipos siguientes:

- Basada en lotes
- Streaming o en tiempo real
- Híbrido

Artefactos

Estos son los resultados de esta fase:

Informe de la calidad de los datos: este informe contiene resúmenes de los datos, las relaciones entre cada atributo y objetivo, la clasificación de las variables, etc.

Arquitectura de la solución: la arquitectura de la solución puede ser un diagrama o una descripción de la canalización de datos utilizada para llevar a cabo la tarea de puntuación o las predicciones con los nuevos datos una vez que se ha creado un modelo.

Decisión de punto de comprobación: antes de comenzar con el proceso completo de diseño de características y con la compilación del modelo, puede volver a evaluar el proyecto para determinar si el valor que está previsto que aporte es suficiente para seguir adelante con él. Por ejemplo, podría estar preparado para continuar, requerir más datos o abandonar el proyecto si no existen datos que respondan a la pregunta.

Diseño de Big Data

Tercero se debe definir las tecnologias a utilizar y la viabilidad del proyecto.

Objetivos

- Definir las tecnologías a utilizar en base a los componentes.
- Definir el software y hardware a utilizar.
- Definición de interfaces de consulta y explotación de datos
- Definir la estructura, operación e implementación del proyecto.

Modo de hacerlo

- Definir las tecnologías a utilizar en base a los componentes, en base a la fuente de datos y el proceso de procesamiento de big data, se deberá definir la tecnología.
- Definir el software y hardware a utilizar, dependiendo de la cantidad de datos y la capacidad de procesamiento proyectada.
- Definición de interfaces de consulta y explotación de datos, definir las herramientas a utilizar, serán definido con el usuario final.
- Definir la estructura, operación e implementación del proyecto, establecer el plan de proyecto, indicando las actividades, recursos y tiempos.

Artefactos

- Arquitectura Tecnológica: la arquitectura puede ser un diagrama o una descripción de la canalización de datos utilizada para llevar a cabo la tarea de puntuación o las predicciones con los nuevos datos una vez que se ha creado un modelo.
- Establecer el plan del proyecto, crear un documento indicando las actividades para la ejecución del proyecto.
- Definir las interfaces para consultas y explotación de los datos: este informe contiene las pantallas prototipos.
- Definir la viabilidad del proyecto, informe que indique el análisis costo/beneficio de la solución a implementar.

Operación del Big Data

Cuarto despliegue de la solución.

Objetivos

- Despliegue de la solución
- Pruebas del método procesamiento de Big Data
- Pruebas funcionales para las consultas y explotación de datos.
- Evaluar seguridad de la solución

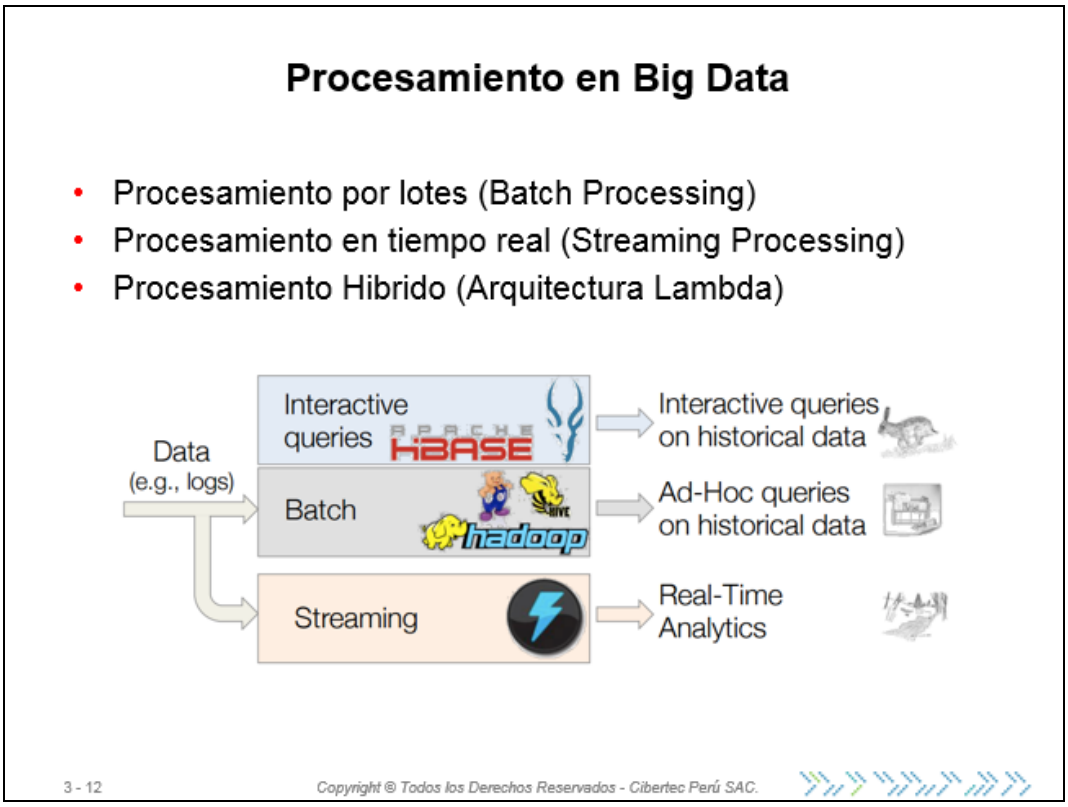
Modo de hacerlo

- Despliegue de la solución, realizar pruebas de concepto, para evaluar la solución por completo.
- Pruebas del método procesamiento de Big Data: efectuar pruebas de carga para obtener el método óptimo de procesamiento de la data.
- Pruebas funcionales para las consultas y explotación de datos: Definir con los usuarios las interfaces y mejoras.
- Evaluar seguridad de la solución: implementar las políticas de seguridad de la información.

Artefactos

- Solución de la solución Big Data.

3. Procesamiento del Big Data



En la actualidad Big Data trabaja con 3 tipos de procesamientos:

a) Procesamiento por lotes (Batch Processing)

- Solución para el problema de volumen.
- Se centra en el procesamiento de gran cantidad de datos estadísticos.
- Escalable.
- Procesamiento distribuido y paralelo
- Tolerancia a fallos
- Alta latencia

b) Procesamiento en tiempo real (Streaming Processing)

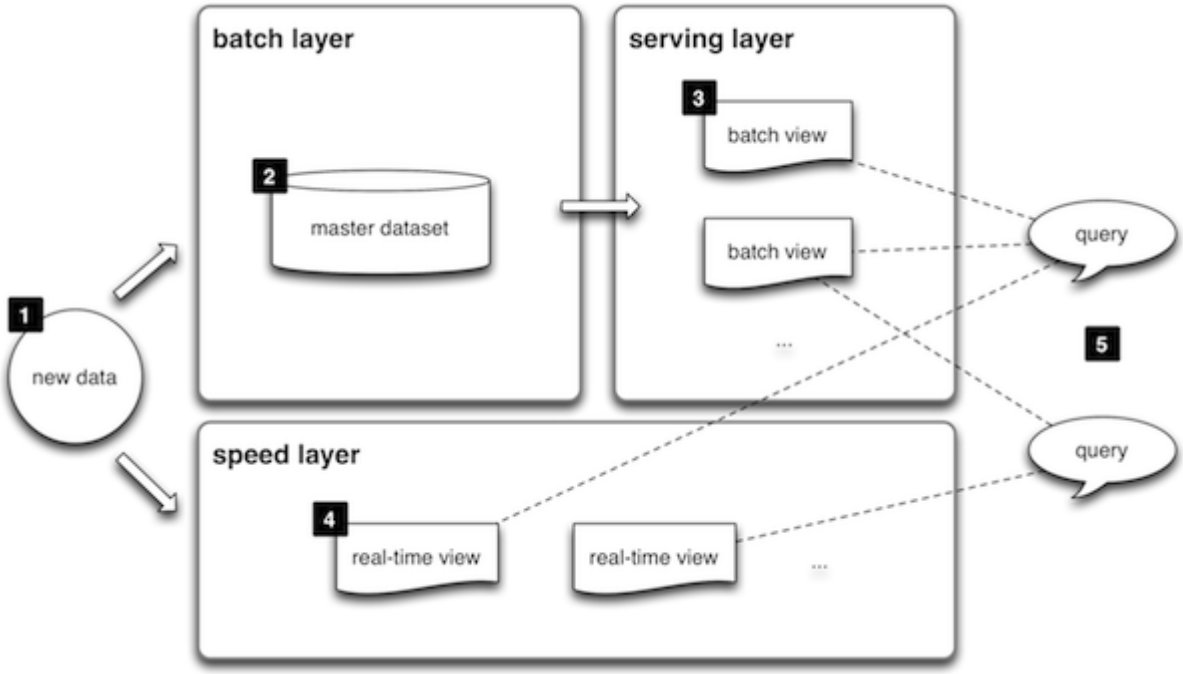
- Solución para el problema de la velocidad.
- Se centra en el procesamiento de un flujo ilimitado y continuo de datos.
- Computación de datos en tiempo real.
- Procesamiento distribuido y paralelo.
- Tolerancia a fallos.
- Baja latencia.

c) Procesamiento Hibrido

- Arquitectura Lambda
- Solución para el problema del volumen y la velocidad
- Combina los resultados de analizar datos estadísticos y datos en tiempo real.

Arquitectura Lambda

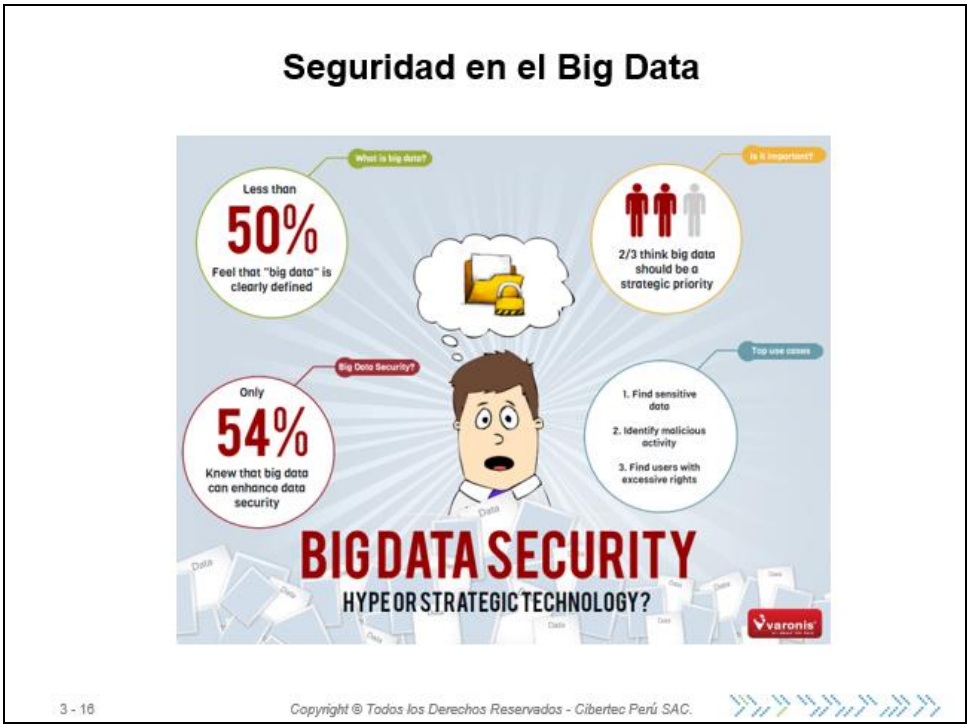
En una arquitectura lambda la idea es implementar sistemas de información que combinan ambas modalidades de procesamiento de datos: batch y stream. Esto nos da lo mejor de dos mundos, ya que el modo batch nos brinda un alcance completo y confiable mientras que el modo stream nos da los datos en línea para decisiones instantáneas.



La figura 1 muestra un modelo de cómo funciona la arquitectura lambda. Los datos que entran al sistema se despachan tanto a la capa batch como a la capa de velocidad (speed). La capa batch escribe los datos al master data set y prepara las vistas batch, pasándolas a la capa de servidor. Esta última se encarga de indexar las vistas batch de manera que pueda responder a búsquedas con muy baja latencia. El problema es que el proceso de escribir datos y luego indexarlos es lento, por lo que éstos no están disponibles de forma instantánea; es aquí donde entra en acción el rol de la capa de velocidad; que se dedica a exponer solamente los datos más recientes, sin preocuparse por escribirlos a un registro permanente. El resultado de cualquier búsqueda puede conjuntar datos provenientes tanto de vistas de la capa batch como de la capa de velocidad.

En cada una de las capas de esta arquitectura se utilizan tecnologías especializadas para cada propósito. Aunque es posible utilizar distintas opciones, una opción popular es utilizar Apache Sqoop + HDFS + Hive para capturar, almacenar y procesar datos en forma batch, y Apache Kafka + HBase + Spark para capturar, almacenar y procesar datos en forma stream.

4. Seguridad en el Big Data



Para una solución de Big Data, es de vital importancia preservar la seguridad de la información: Confidencialidad, Integridad y Disponibilidad, con este fin se debe implementar, mantener y evaluar un sistema manual o automatizado de seguridad de información.

Confidencialidad

La confidencialidad es la propiedad que impide la divulgación de información a personas o sistemas no autorizados. A grandes rasgos, asegura el acceso a la información únicamente a aquellas personas que cuenten con la debida autorización.
Por ejemplo, una transacción de tarjeta de crédito en Internet requiere que el número de tarjeta de crédito a ser transmitida desde el comprador al comerciante y el comerciante de a una red de procesamiento de transacciones. El sistema intenta hacer valer la confidencialidad mediante el cifrado del número de la tarjeta y los datos que contiene la banda magnética durante la transmisión de los mismos. Si una parte no autorizada obtiene el número de la tarjeta en modo alguno, se ha producido una violación de la confidencialidad.

La pérdida de la confidencialidad de la información puede adoptar muchas formas. Cuando alguien mira por encima de su hombro, mientras usted tiene información confidencial en la pantalla, cuando se publica información privada, cuando un laptop con información sensible sobre una empresa es robado, cuando se divulga información confidencial a través del teléfono, etc. Todos estos casos pueden constituir una violación de la confidencialidad.

Integridad

Es la propiedad que busca mantener los datos libres de modificaciones no autorizadas. (No es igual a integridad referencial en bases de datos.) A groso modo, la integridad es el mantener con exactitud la información tal cual fue generada, sin ser manipulada o alterada por personas o procesos no autorizados.

La violación de integridad se presenta cuando un empleado, programa o proceso (por accidente o con mala intención) modifica o borra los datos importantes que son parte de la información, así mismo hace que su contenido permanezca inalterado a menos que sea modificado por personal autorizado, y esta modificación sea registrada, asegurando su precisión y confiabilidad. La integridad de un mensaje se obtiene adjuntándole otro conjunto de datos de comprobación de la integridad: la firma digital Es uno de los pilares fundamentales de la seguridad de la información.

Disponibilidad

La disponibilidad es la característica, cualidad o condición de la información de encontrarse a disposición de quienes deben acceder a ella, ya sean personas, procesos o aplicaciones. A groso modo, la disponibilidad es el acceso a la información y a los sistemas por personas autorizadas en el momento que así lo requieran.

En el caso de los sistemas informáticos utilizados para almacenar y procesar la información, los controles de seguridad utilizados para protegerlo, y los canales de comunicación protegidos que se utilizan para acceder a ella deben estar funcionando correctamente. La Alta disponibilidad sistemas objetivo debe estar disponible en todo momento, evitando interrupciones del servicio debido a cortes de energía, fallos de hardware, y actualizaciones del sistema.

Garantizar la disponibilidad implica también la prevención de ataque de denegación de servicio. Para poder manejar con mayor facilidad la seguridad de la información, las empresas o negocios se pueden ayudar con un sistema de gestión que permita conocer, administrar y minimizar los posibles riesgos que atenten contra la seguridad de la información del negocio.

La disponibilidad además de ser importante en el proceso de seguridad de la información, es además variada en el sentido de que existen varios mecanismos para cumplir con los niveles de servicio que se requiera. Tales mecanismos se implementan en infraestructura tecnológica, servidores de correo electrónico, de bases de datos, de web etc, mediante el uso de clusters o arreglos de discos, equipos en alta disponibilidad a nivel de red, servidores espejo, replicación de datos, redes de almacenamiento (SAN), enlaces redundantes, etc. La gama de posibilidades dependerá de lo que queremos proteger y el nivel de servicio que se quiera proporcionar.

ISO/IEC 27002 proporciona recomendaciones de las mejores prácticas en la gestión de la seguridad de la información a todos los interesados y responsables en iniciar, implantar o mantener sistemas de gestión de la seguridad de la información. La seguridad de la información se define en el estándar como "la preservación de la confidencialidad (asegurando que sólo quienes estén autorizados pueden acceder a la información), integridad (asegurando que la información y sus métodos de proceso son exactos y completos) y disponibilidad (asegurando que los usuarios autorizados tienen acceso a la información y a sus activos asociados cuando lo requieran)".

La versión de 2013 del estándar describe los siguientes catorce dominios principales:

1. Políticas de Seguridad de la Información: Dentro de este capítulo se hace hincapié en la importancia que ocupa la disposición de una **adecuada política de seguridad**, aprobada por la dirección, comunicada a todo el personal, revisada de forma periódica y actualizada con los cambios que se producen en el interior y en el exterior.
2. Organización de la Seguridad de la Información: Los controles indicados en este capítulo buscan **estructurar un marco de seguridad** eficiente tanto mediante los roles, tareas, seguridad, etc. como en los dispositivos móviles.
   Tenemos que tener presente que cada vez es mayor el peso que está ocupando el teletrabajo dentro de las empresas, y por ello, se deben tener en cuenta **todas sus características especiales** para que ningún momento la seguridad de la información de la que se dispone se vea afectada.
3. Seguridad relativa a los recursos humanos: Si analizamos los incidentes de seguridad que se producen en una organización nos daremos cuenta de que **la gran mayoría de estos tienen su origen en un error humano**. Se debe concienciar y formar al personal de los términos de empleo de la información en el desarrollo de sus actividades y la importancia que tiene la información en el desarrollo de sus actividades, además de la importancia que tiene promover, mantener y mejorar **el nivel de seguridad adecuándolo a las características** de los datos y la información que maneja es clave y uno de los objetivos que se debe perseguir.
4. Gestión de activos: Se centra en la atención en la **información como activo** y en cómo se deben establecer las medidas adecuadas para guardarlos de las incidencias, quiebras en la **seguridad y en la alteración no deseada**.
5. Control de acceso: Controlar quien accede a la información dentro de un aspecto relevante. Al fin y al cabo, no todas las **personas de una organización** necesitan acceder para realizar su actividad diarias a todos los datos, sino que tendremos roles que **necesitan un mayor acceso** y otros con un acceso mucho más limitado. Para poder marcar las diferencias, se deben establecer todos los controles como registro de los usuarios, **gestión de los privilegios de acceso**, etc. siendo algunos de los controles que se incluyen en este apartado.
6. Criptografía: En el caso de que estemos **tratando la información sensible o crítica** puede ser interesante utilizar diferentes técnicas criptográficas para **proteger y garantizar** su autenticidad, confidencialidad e integridad.
7. Seguridad física y del entorno: La seguridad no es solo a nivel tecnológico sino también físico, es decir, una simple labor de no dejar las pantallas e impresoras en zonas que sean fácilmente accesibles, por parte del personal externo los documentos con los que se están trabajando no sólo nos permitirán gestionar de forma adecuada la seguridad, sino que se **acabarán convirtiendo en hábitos que nos aportan eficiencia** en la gestión.
8. Seguridad de las operaciones: Tiene un marcado componente técnico entrado en todos los **aspectos disponibles como la protección** del software malicioso, copias de seguridad, control de software en explotación, gestión de vulnerabilidad, etc.
9. Seguridad de las comunicaciones: Partiendo de la base de que la gran mayoría de los intercambios de información y de datos en distintas escalas **se llevan a cabo mediante las redes sociales**, garantizar la seguridad y proteger de forma adecuada los medios de transmisión de estos datos clave.

10. Adquisiciones, desarrollo y mantenimiento de los sistemas de información: La seguridad no es un aspecto de un área en concreto, ni de un determinado proceso, no que es general, **abarca toda la organización** y tiene que estar presente como elemento transversal clave dentro del ciclo de vida del sistema de gestión.
11. Relación de proveedores: Cuando se establecen las relaciones con terceras partes, como puede ser proveedores, se deben **establecer medidas de seguridad** pudiendo ser muy recomendable e incluso necesario en determinados casos.
12. Gestión de incidentes de seguridad de la información: No podemos hablar de **controles de seguridad** sin mencionar un elemento clave, los incidentes en seguridad. Y es que, estar preparados para cuando estos incidentes ocurran, dando **una respuesta rápida y eficiente** siendo la calve para prevenirlos en el futuro.
13. Aspectos de seguridad de la información para la gestión de la continuidad de negocio: No sabemos lo que necesitábamos un dato hasta que lo hemos perdido. Sufrir una pérdida de **información relevante y no poder recuperarla** de laguna forma puede poner en peligro la continuidad de negocio de la organización.
14. Cumplimiento: No podemos hablar de seguridad de la información, sin hablar de legislación, normas y políticas aplicables que **se encuentre relacionadas con este campo** y con las que conviven en las organizaciones. Debemos tener presente que ocupan un enorme lugar en cualquier sistema de gestión y deben **garantizar que se cumple** y que están actualizados con los últimos cambios siendo esencial para no llevarnos sorpresas desagradables.