A10

ANALYTICS10

BIG DATA

BENEFICIOS DE USAR CLOUDERA (HADOOP)

Índice

Introducción	2
Nuevos desafíos de la industria de Analítica	3
Estructuras rígidas y poca agilidad de negocio	3
Costo por TB	4
Las trampas de modelar datos “on write”	5
Agilidad en diversidad de workloads	6
Arquitectura propuesta por A10 y Cloudera	7
Resumen	9

Introducción

Big Data ya no es solo una moda. A medida que Apache Hadoop ha evolucionado y madurado, cada vez más empresas están pasando de simples evaluaciones y prototipos a construir Data Hubs empresariales basados en esta tecnología open source.

Hadoop trae los beneficios del alto rendimiento, escalamiento predecible y analítica avanzada sobre datos complejos por un costo mínimo. Sin embargo, su impacto sobre drivers de negocio requiere de una propuesta que asegure el alineamiento de: capacidades masivas de cómputo junto con eficiencia operacional y una gran variedad de casos de uso.

Para extraer el máximo valor posible de tus datos, la arquitectura de un Data Hub Empresarial (EDH) se construye sobre las herramientas de Hadoop. Estas proveen poderosas capacidades de procesamiento, exploración y analítica en real time. Y todo esto pensado para operar óptimamente en Hardware commodity. El EDH de Cloudera complementa la flexibilidad y extensibilidad de su núcleo de herramientas open source con sistemas de gestión, gobierno de datos y robusta seguridad que los líderes de la industria demandan de todos sus sistemas productivos.

La aparición y la expansión de estas tecnologías se ha visto impulsada por cambios fundamentales en los mercados de casi todas las industrias. Cambios que, lamentablemente, las arquitecturas tradicionales no son capaces de sortear satisfactoriamente.

Nuevos desafíos de la industria en analítica

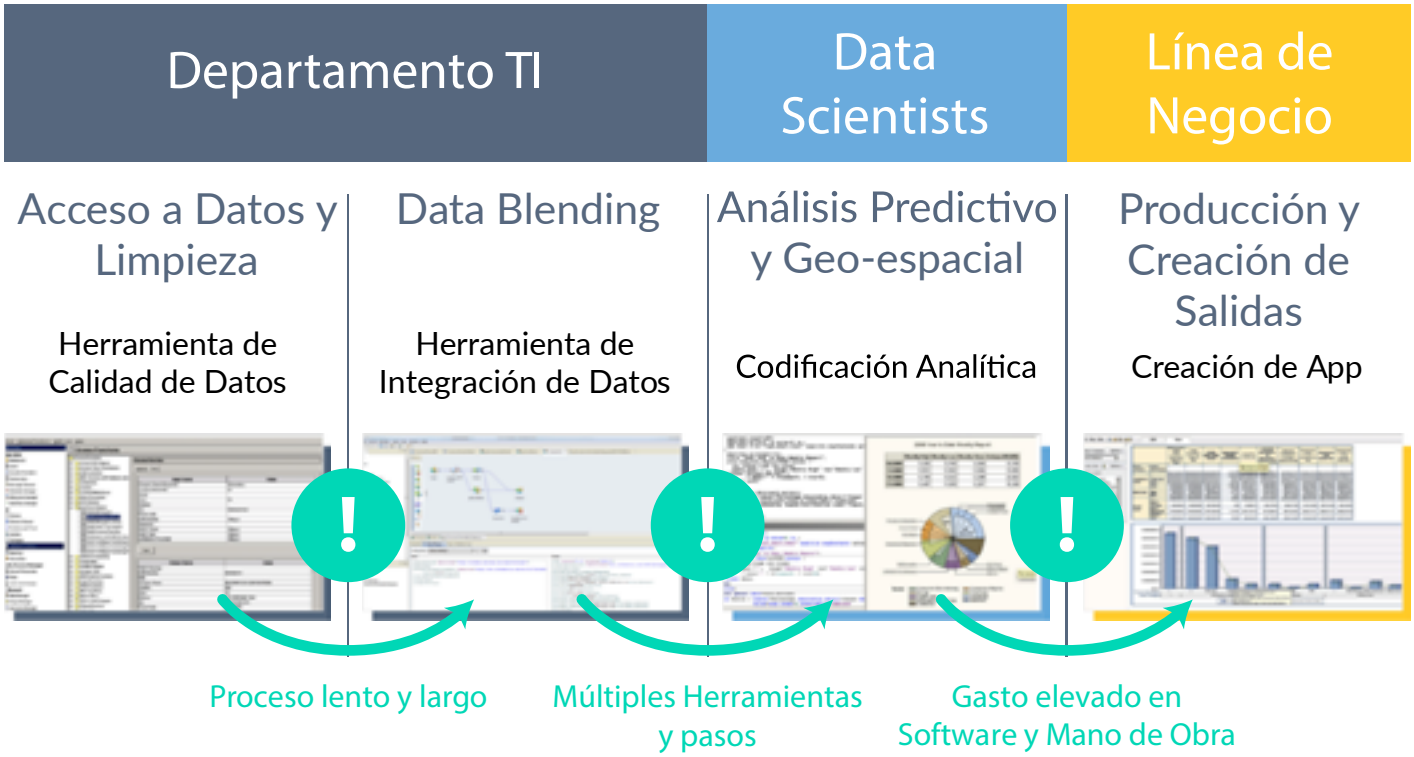
El principal desafío de las empresas modernas es cómo extraer el mayor valor posible de sus datos. Las empresas más innovadoras están aprovechando cantidades masivas de datos para desafiar los modelos de negocio tradicionales.

Sin embargo, el volumen y el valor práctico de los datos tradicionalmente se han contrapuesto. Históricamente, data sets más grandes y complejos han requerido herramientas analíticas y estrategias de gestión a la medida que terminan poniendo una prima en el proceso de extracción de valor. La visión de clientes, operaciones y mercados que impulsa el crecimiento del negocio y la eficiencia se ha visto limitada por "silos" de información desagregada de distintas áreas de negocio y el alto costo de hacer escalar sistemas legacy o implementar otros nuevos. Estas y otras razones han dificultado el paso de algunas a empresas al mundo del "Big Data". A continuación se comentan otros problemas comunes.

2.1 Estructuras rígidas y poca agilidad de negocio

Los Data Warehouse son sistemas estructurados por naturaleza. Están diseñados para presentar los datos de una forma predeterminada por un modelo definido y los requisitos levantados por un equipo encargado.

Esta etapa llamada "pre-modelamiento" está pensada para cumplir exclusivamente con un grupo reducido de necesidades y casos de uso. Esta forma de trabajar hace muy complejo para los usuarios de negocio trabajar en hipótesis que no hayan sido consideradas previamente, a menos que realicen todo el ciclo de vida tradicional de desarrollo de software conducido por TI. Este ciclo, reflejado en la siguiente imagen, tiende a ser largo y costoso. Además, la naturaleza estructurada de un data warehouse impide que los negocios puedan añadir contexto y valor de fuentes no estructuradas como documentos, contratos, llamadas al call center, redes sociales y logs entre otros.



A pesar de que los data warehouses proveen un gran valor a las organizaciones, estos restringen severamente la innovación casos de uso exploratorios, y todas las iniciativas que implican desarrollo ágil ("fail fast"). Finalmente esto se traduce en una menor explotación analítica de tus datos y una pérdida de ventaja competitiva.

2.2 Costo por TB

Otros de los principales impedimentos que las organizaciones tienen que enfrentar, son los costos de almacenamiento y procesamiento. Algunas organizaciones llegan a pagar valores alrededor de los $35.000 por Terabyte al año por plataformas de almacenamiento "high end". Con nuevos requerimientos de negocios apareciendo diariamente y con los datasets actuales creciendo cada vez más rápido, las compañías se ven obligadas a restringir sus gastos en TI. Ahora, imaginemos los volúmenes de datos multiplicados por 10. Esas son las proyecciones que muchas empresas están manejando para los próximos años. Para permanecer dentro de los presupuestos de TI, las compañías terminan recurriendo a técnicas como:

- Archivar información antigua o "fría", que es muy costosa de mantener debido al desconocido valor que podrían aportar.
- Posponer la entrega de nuevos requerimientos de negocio (como KPIs o nueva analítica), para no sobrepasar la arquitectura actual.
- Restringir casos de uso exploratorios e innovación, ya que el crecimiento de los procesos actuales es prioritario.

Pero probablemente la estrategia más utilizada es simplemente no recolectar o almacenar ciertos datos, o almacenar muestras o consolidados de los mismos. De cualquier forma, es imposible extraer valor de datos que no se encuentren en el data warehouse.

2.3 Las trampas de modelar datos "on write"

Dentro de un data warehouse, la capa de modelamiento de datos es crítica. Ayuda a la organización a alinearse detrás de conceptos clave y entidades como clientes, contratos o activos, entre otros, para asegurar que todas las áreas de negocios hablen el mismo idioma. Esto ayuda a crear una "única versión de la verdad" sobre la cual los KPIs son construidos.

Pero esta ventaja viene con restricciones, porque los datos tienen que ser transformados según este modelo antes de poder inyectarse al data warehouse ("on write"). Las compañías pueden elegir mantener una versión "cruda" de los datos por un periodo de tiempo, pero de nuevo el costo asociado a mantener esto se vuelve prohibitivo. Hasta el momento no ha habido una forma costo-eficiente para las organizaciones de almacenar todos los datos crudos que deseen.

Por lo tanto, cuando se tienen que agregar nuevos atributos, debido a un cambio en la fuente de datos (i.e actualización del ERP) o la implementación de un nuevo proceso de negocio, las organizaciones tienen que realizar cambios significativos en los procesos de modelamiento y migración de datos antes de poder explotar los nuevos datos.

Usar Hadoop como almacén de datos permite aprovechar el concepto de "schema on read": es decir, definir el modelo de datos al momento de analizarlos. En lugar de retocar y transformar los datos antes de almacenarlos, las compañías pueden definir el modelo al momento de

leer los datos crudos. Esto les permite ser mucho más ágiles cuando necesiten modificarlo, ya que no necesitan migrar nada. Definir el modelo de datos "on read" permite mejorar drásticamente el time to market cuando se necesita modificar las vistas elegidas para los datos.

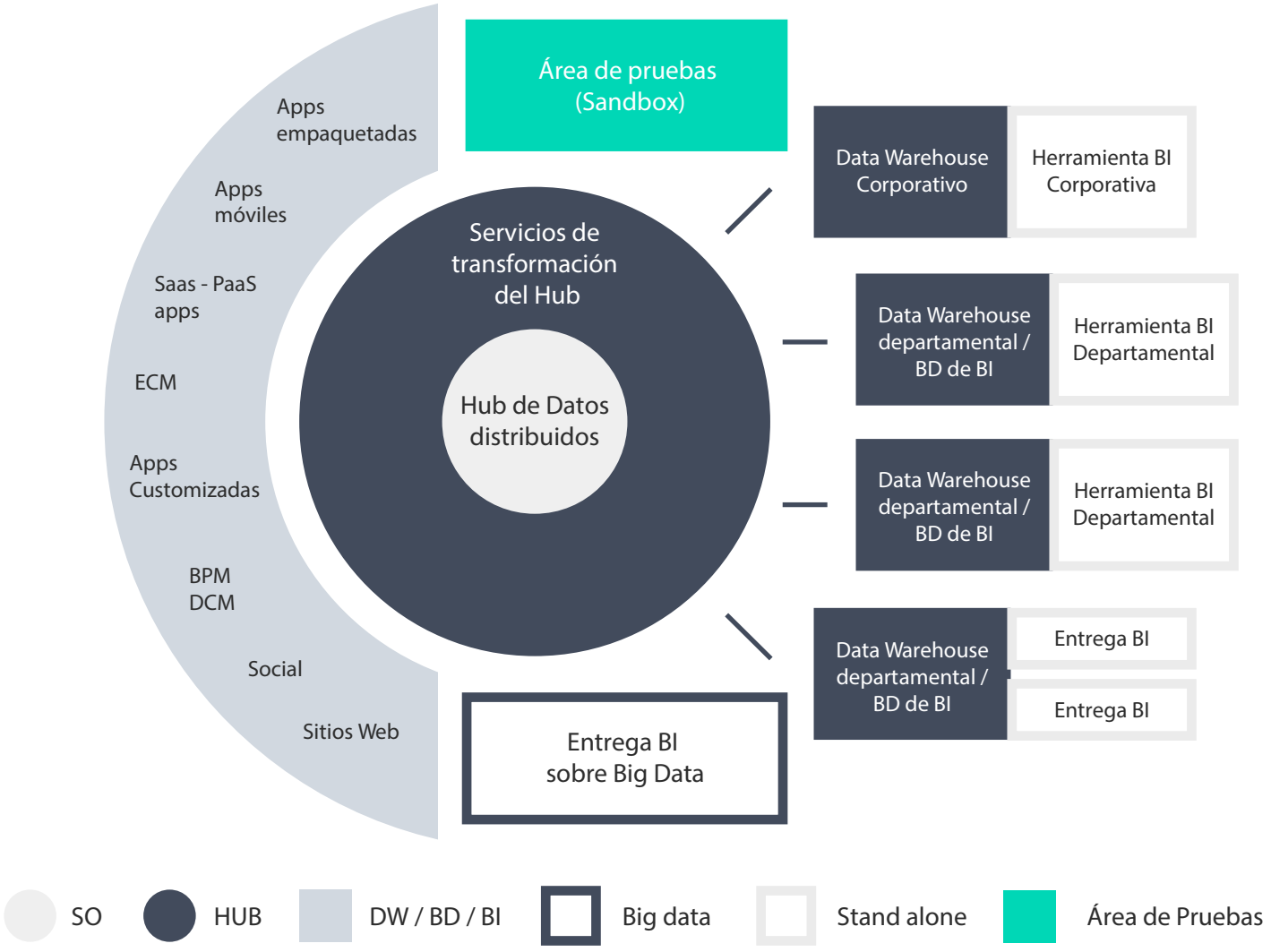
2.4 Agilidad en diversidad de workloads

Las arquitecturas actuales no permiten mucha agilidad en términos de satisfacer las necesidades de investigación de clientes y al mismo tiempo completar intensivos cómputos para procesos productivos. Algunos data warehouses todavía luchan con el problema de priorizar cargas de trabajo heterogéneas, dado que su capacidad de manejar datasets más grandes y complejos está limitada por su capacidad de cómputo y disco. Lo peor de todo es que para aumentar estas capacidades, las empresas deben desembolsar cantidades inmensas de dinero para migrar a un data warehouse más poderoso, o adquirir un nuevo por completo.

Apache Hadoop, en cambio, es un orden de magnitud más barato, se implementa sobre hardware commodity y escala linealmente con el volumen de datos, lo que permite obtener almacenamiento y procesamiento masivo. Las organizaciones pueden definir cuotas de recursos y ejecutar todo tipo de cargas de trabajo, de diversas áreas de negocio a una fracción del costo. Luego, a medida que las iniciativas vayan dando resultados, las cargas comenzarán a aumentar paulatinamente. En este momento las capacidades del sistema se pueden aumentar gradualmente, con inversiones acotadas, y sólo cuando éstas demuestren que justifican la inversión.

Arquitectura propuesta por A10 y Cloudera

La solución a los problemas previamente mencionados no implica que haya que eliminar los data warehouses, por ningún motivo. En nuestra experiencia como empresa, la mejor manera de comenzar en el mundo del big data es extender las capacidades existentes de los data warehouses con un EDH de Cloudera. En Analytics10 proponemos la Arquitectura de Hub & Spoke, en la siguiente imagen, como patrón a seguir



Luego, en la segunda capa, se consideran los servicios de transformación de datos del Hub. La ventaja de poder almacenar datos no estructurados, junto con almacenamiento masivo y a bajo costo, es que permite almacenar la data "cruda". Esto es más rápido y sencillo, ya que no se desperdician esfuerzos en formatear y limpiar datos que no sabemos cual va a ser su utilidad en el futuro. De esta forma se es más flexible y se realizan las ETLs solo cuando es necesario sin desperdiciar ninguna información que pueda ser valiosa después.

Una vez transformados los datos, estos son cargados en los spokes correspondientes. Los cuales pueden ser data warehouses específicos por departamento, cuyas únicas funciones son: dar servicio y respuestas rápidas a los usuarios correspondientes y alimentar a las herramientas de BI. Los data warehouse entregan baja latencia, alta calidad y estructura, lo que los hace ideales para esto.

Para efectos de investigación e innovación, la propuesta es crear un sandbox, donde se disponibilizan todas las herramientas necesarias para desarrollar: nuevos reportes, analítica avanzada, machine learning y data science en general. El objetivo es que todas estas herramientas puedan operar directamente sobre el Hub lo cual trae muchos beneficios. Se explota la capacidad de cómputo masiva para trabajar con toda la historia de los datos, no solo una muestra o ventana acotada. Se aprovecha la flexibilidad que implica trabajar con datos brutos directamente. Y además se puede trabajar sobre la totalidad de los datos de la empresa, no solo sobre la visión de un área particular.

Centralizar y disponibilizar analítica a través de todos tus datos abre nuevas oportunidades de negocio que antes eran prohibitivamente caras o complejas. Un EDH entrega capacidades avanzadas - como modelos de clientes basados en redes sociales y comportamientos offline, análisis en tiempo real de streams de datos en movimiento, seguridad proactiva frente a fraudes y ciberataques - con una plataforma unificada, flexible y escalable que es fácil de implementar y hacer crecer a medida que entregue valor al negocio.

En síntesis, una arquitectura como la presentada con un EDH en el centro, permite satisfacer todas las necesidades modernas de analítica avanzada, sin quebrantar los procesos e iniciativas tradicionales y con la capacidad de escalar sin restricciones, a medida que sea necesario.

Resumen

Sin los datos que impulsan las oportunidades en el contexto de negocios moderno, los tomadores de decisiones a lo largo de todas las industrias seguirán luchando con la "parálisis de información" y sus altos costos para la empresa. Cada vez más, Finanzas luchará con robos y fraudes; TI sufrirá con regulaciones de compliance y overspending; Marketing tendrá que lidiar con una visión reducida e incompleta del cliente; I+D se verá limitada por el tamaño de sus muestras de datos; y la gerencia general no logrará competir con rivales más empoderados a medida que el mercado evoluciona.

Es imperativo que las empresas impulsen cambios significativos en su infraestructura. La soluciones tradicionales de almacenamiento, gestión y disponibilización de datos limitan el volumen manejable de datos, cumpliendo objetivos específicos, generalmente para grupos restringidos de usuarios. Se necesitan nuevos sistemas para poder soportar las nuevas realidades del servicio al cliente, desarrollo de productos y gestión del riesgo corporativo.

Con Cloudera, ahora se puede aplicar analítica avanzada sobre datos ilimitados, transformando esos datos en un activo estratégico. Desplegando un Data Hub Empresarial, múltiples usuarios y aplicaciones pueden acceder simultáneamente, en tiempo real, al universo completo de datos gobernados y de forma segura. Ninguna otra plataforma ofrece una combinación tan poderosa de procesamiento, flexibilidad y seguridad para satisfacer los casos de uso modernos que las empresas necesitan atacar para mantenerse competitivas.

Complementado por el apoyo experto de Analytics10, el proceso para integrar estas nuevas tecnologías en tu empresa es sencillo y seguro. Si te interesa conocer más de Cloudera, Hadoop o Big Data te invitamos a acercarte a nosotros. En Analytics10 encontrarás tu mejor aliado en Analítica Avanzada y Big Data.

Analytics10 es una empresa consultora de Analítica y Big Data. Ayudamos a organizaciones a tomar decisiones inteligentes, basadas en datos, traduciéndose en información comprensible que puede llevarlos a acciones concretas, ya que hacemos la analítica avanzada SIMPLE para nuestros clientes.

Con sede en Santiago, Chile, Analytics10 proporciona a miles de clientes y usuarios de nuestras tecnologías de vanguardia, las mejores soluciones a sus necesidades y requerimientos orientados a solucionar problemas complejos simplificando su implementación y asegurando el mejor resultado.

www.Analytics10.com