



El futuro
es de todos

Consejería Presidencial
para asuntos económicos
y transformación digital



El futuro
es de todos

DNP
Departamento
Nacional de Planeación



El futuro digital
es de todos

MinTIC

ANEXO 6

Arquitecturas de referencia para *big data*



ARQUITECTURAS DE REFERENCIA PARA *BIG DATA*

La identificación de necesidades tecnológicas debe estar respaldada por la comprensión del diseño de entornos de *big data*. La estructura común, que define los entornos para el desarrollo de *big data*, se denomina arquitectura de referencia. Las arquitecturas de referencia para *big data* se construyen alrededor del ciclo de vida de los datos, es decir, mediante recolección, almacenamiento, procesamiento y visualización. A su vez, tienen en consideración procesos de gobernanza de datos, (como gestión de metadatos y datos maestros), procesos de gestión y administración de riesgos y procesos de seguridad y privacidad de la información. Estas arquitecturas deben permitir la gestión de grandes volúmenes de datos, la transformación de datos estructurados y no estructurados para su análisis, y el procesamiento de consultas en un tiempo de respuesta corto.

De acuerdo con el Instituto Nacional de Estándares y Tecnología (NIST por sus siglas en inglés), los marcos de arquitectura de referencia para *big data* tienen la finalidad de proveer un lenguaje común para los agentes interesados, promover el desarrollo de estándares, especificaciones y patrones comunes, proporcionar métodos consistentes para la implementación de tecnologías que puedan dar solución a problemáticas comunes, y mejorar el entendimiento de sistemas y procesos asociados a los modelos conceptuales de *big data*. La arquitectura de referencia del NIST está integrada por los procesos de generación de valor a través de los datos, en concordancia con su ciclo de vida, y por los procesos de las tecnologías de la información necesarios para el desarrollo del ciclo de vida de los datos (servicios de infraestructura y plataforma). Por tanto, ambos procesos están altamente integrados. De forma transversal y general, se contempla, como marco de la interacción, los



procesos de seguridad y privacidad de la información, además de la administración y gestión de los datos.

Arquitecturas para *big data* de acuerdo con el nivel de madurez esperado

Con la finalidad de orientar a las entidades en la identificación de servicios de tecnología y en recurso humano de acuerdo con el nivel de madurez al que quieren llegar, se definieron unas arquitecturas de referencia para cada nivel de madurez, asociadas a: características de los datos; etapas del ciclo de vida de los datos; procesos de gobernanza de datos, seguridad y privacidad de la información y administración y gestión de riesgos. Es importante aclarar que no se incluye la descripción del nivel 1, dado que es el nivel de madurez más bajo del modelo y, por ende, una entidad no podrá estar en un nivel más bajo a este.

A continuación, se incluirá una descripción de las variables que se consideraron para la construcción de las arquitecturas de referencia para el modelo explotación de datos. Como se mencionó anteriormente, esto brinda una referencia a las entidades, para la identificación de los requerimientos tecnológicos que estará sujeto a las necesidades de cada entidad. Por lo anterior, se recomienda que cada entidad revise, dadas sus particularidades, cuáles son sus demandas respecto a las mismas. Estas arquitecturas pueden ser el punto de partida para identificar las necesidades de la entidad en materia de tecnología, y para la posterior cotización de servicios tecnológicos.

1. Volumen:

El volumen de los datos, particularmente, es una variable esencial para definir los servicios necesarios en el desarrollo del ciclo de vida de los datos, dado que es el resultado de los diferentes tipos de información que se están recopilando (variedad) y de las necesidades de almacenamiento que se



derivan de la velocidad de captura y análisis de los datos (temperatura de los datos y latencia).

A razón de lo anterior, el primer punto que se debe tener en consideración para abordar el ciclo de vida de los datos, ya sea dentro de los diferentes niveles de la entidad o para el desarrollo de proyectos, es el volumen de almacenamiento requerido. En la medida en que, la variable volumen, delimita la capacidad que tiene la entidad para almacenar, procesar y analizar los datos dentro de un horizonte temporal.

2. Estructura:

Esta variable refiere al tipo de orden bajo el cual se agrupan, almacenan y relacionan los datos, para, posteriormente, hacer ejercicios de procesamiento y análisis. Hay tres tipos de estructuras: i) datos estructurados son aquellos que se encuentran organizados mediante una estructura bien definida, normalmente, de filas y columnas de bases relacionales, por ende, son los más fáciles de gestionar; ii) datos semi estructurados, refieren a una organización que se encuentra entre el límite de los estructurados y lo no estructurado, dado que carecen de un esquema fijo, pero evidencian un proceso organizacional para facilitar su uso; iii) datos no estructurados, son aquellos que no tienen un esquema de organización y, por ende, requieren de mayores conocimientos y procesos de análisis, a su vez, estos datos se almacenan en base de datos no relacionales o NoSQL.

Dado lo anterior, la entidad debe considerar la estructura o de los tipos de estructura que gestiona, en la medida en que determinan el tipo de procesos y esquemas que se deben utilizar para su procesamiento y



análisis. Por ende, el tipo de estructura de los datos delimita la infraestructura TI que debe ser desplegada.

En el ciclo de vida de los datos influyen 3 aspectos a tener en cuenta:

1. Latencia:

Esta característica, de forma general, hace referencia a un intervalo temporal que transcurre entre una orden y su respuesta. En términos de los datos, la latencia se puede interpretar como el tiempo que requiere una unidad de almacenamiento para permitir la consulta de los datos y el tiempo necesario de procesamiento.

A su vez es importante resaltar que existe un nexo entre la latencia de los datos los tipos de consulta. Dado que los formatos de consulta están relacionados directamente con intervalos de tiempo y, por ende, con tiempos de respuesta. Por ejemplo, las necesidades de latencia en consultas de batch o lotes no es tan rigurosa como los tipos de procesamiento en tiempo real, que requieren una respuesta inmediata que permita su consulta y análisis de forma instantánea.

2. Tipos de consulta: Batch, streaming, tiempo real

Los tipos de consulta o procesamiento están relacionados directamente con la latencia, dado que el procesamiento se debe ejecutar en un margen de tiempo. Dado lo anterior se pueden dar tres tipos de procesamiento, los cuales son por: 1) batch o lotes, lo que quiere decir que el procesamiento de los datos se genera mediante recursos informáticos para datos agrupados, sin la necesidad o con la necesidad parcial de intervención de



los usuarios. Este tipo de procesamiento implica que los usuarios deben recabar y recopilar los datos para procesarlos posteriormente de forma conjunta, por lo cual los resultados de esta metodología tienen un margen temporal amplio. 2) Streaming, este tipo de procesamiento se define a través de un ejercicio continuo, que supone un flujo de datos permanente, sin límites temporales. Este esquema permite la visualización de resultados de forma rápida y está asociado a datos periódicos. 3) Tiempo real, refiere a un esquema bajo el cual se procesan datos no asociados, no estructurados o semi estructurados, el procesamiento se da de forma casi instantánea e implica un proceso más rápido que por streaming.

3. Temperatura de los datos:

Los intervalos de tiempo de consulta y almacenamiento son un aspecto transversal a las características descritas. La temperatura de los datos hace referencia al tiempo que una entidad decide almacenar sus datos, para su posterior consulta y análisis. Cuando se recopila información se considera que los datos son calientes, en la medida en que se han generado un periodo de tiempo corto, pero conforme los datos se almacenan, para una posterior consulta, se consideran fríos. Las necesidades de almacenamiento de los datos calientes no son amplias, dado que, al ser nuevos o recientes, el volumen de estos es reducido. Sin embargo, conforme los datos se enfrían y se almacenan se requiere de un mayor espacio de almacenamiento.

Tabla 1. Arquitecturas de referencia para cada nivel de madurez



Características de los datos	Volumen	Consideración de la entidad	Consideración de la entidad	Consideración de la entidad	Consideración de la entidad
	Estructura	Estructurados	Estructurados Semiestructurados	Estructurados Semiestructurados No Estructurados	Estructurados Semiestructurados No Estructurados
	Captura	Reportes periódicos de datos estructurados tomados de almacenes de datos (Data warehousing) con un alto componente manual. No datos transaccionales ni de streaming.	Automatización de carga de datos estructurados tomados de almacenes de datos incluidos datos transaccionales.	Automatización de carga de datos no estructurados (ELT) y estructurados tomados de almacenes de datos, incluidos datos transaccionales y en streaming Soluciones de integración de datos con entidades externas	Automatización de carga de datos no estructurados (ELT) y estructurados tomados de almacenes de datos, incluidos datos transaccionales y en streaming. Integrados para su consulta con otras entidades Soluciones de integración de datos con entidades externas
	Almacenamiento	Almacenamiento de bases de datos relacionales basados en SQL. Velocidad de solicitud baja. Consultas en batches	Almacenamiento de bases de datos SQL y NoSQL. Almacenamiento distribuido. Consultas en batches	Almacenamiento de bases de datos SQL y NoSQL. Control de acceso centralizado por adecuada gestión de metadatos. Integración de data warehouses y data lakes. Almacenamiento distribuido. Consultas en batches, streaming y en tiempo real.	Almacenamiento de bases de datos SQL y NoSQL. Control de acceso centralizado por adecuada gestión de metadatos. Integración de data warehouses y data lakes. Latencia baja. Almacenamiento distribuido. Consultas en batches, en streaming y en tiempo real. Alta tolerancia a fallos.



Ciclo de vida de los datos	Procesamiento	Procesamiento por batches no distribuido. El análisis de datos se orienta a modelos descriptivos. Para proyectos piloto se emplean técnicas básicas de ML.	Procesamiento distribuido. Análisis de datos por batches. Soporta procesos básicos de machine learning predictivos e inteligencia de negocios.	Procesamiento distribuidos masivamente paralelos y escalables. Análisis de datos por batches, por streaming y en tiempo real. Soporta procesos de matching learning predictivos e inteligencia de negocios. Alta tolerancia a fallos	Procesamiento distribuidos masivamente paralelos y escalables. Análisis de datos por batches, por streaming, en tiempo real. Soporta procesos de matching learning predictivos y prescriptivos e inteligencia de negocios. Alta tolerancia a fallos Los procesos y técnicas de análisis se comparten con otras entidades, con el propósito de fortalecer el aprovechamiento de los datos
	Visualización	Visualización para reportes periódicos especialmente descriptivos. Tableros interactivos para proyectos piloto.	Tableros de visualización interactivos derivados de procesos de machine learning y análisis inteligencia de negocios	Tableros de visualización interactivos derivados de procesos de machine learning y análisis inteligencia de negocios. Visualización e informes de datos en tiempo real e históricos.	Tableros de visualización interactivos derivados de procesos de machine learning y análisis inteligencia de negocios. Visualización e informes de datos en tiempo real e históricos. Integración de procesos de aprovechamiento de datos directamente en las aplicaciones desarrolladas.
	Gobernanza de datos	Débil	Moderado	Fuerte	Muy fuerte



Gobernanza y gestión de datos	Seguridad y privacidad	Moderado	Moderado	Fuerte	Muy fuerte
	Administración, monitoreo y gestión de riesgos	Débil	Moderado	Fuerte	Muy fuerte