BIG DATA EN EL MUNDO DENTRO DE LAS COMUNICACIONES ÓPTICAS

Taller N°13

Ronaldo Alexander Almachi Murillo, Dennys Francisco Salazar Domínguez
Escuela Politécnica Nacional
Quito, Ecuador
ronaldo.almachi@epn.edu.ec, dennys.salazar@epn.edu.ec

Resumen: En el siguiente documento se presenta sobre el BigData dentro de las comunicaciones ópticas, se explican los conceptos, tecnología, características, operación y ventajas.

Palabras clave: Big Data, Redes Ópticas.

I. INTRODUCCIÓN

El término "big data" se refiere a los datos que son tan grandes, rápidos o complejos que es difícil o imposible procesarlos con los métodos tradicionales. El acto de acceder y almacenar grandes cantidades de información para la analítica ha existido desde hace mucho tiempo. [1]

Funcionamiento

El Big Data funciona en base a las llamadas “5 Vs”: volumen, variedad, velocidad, veracidad y valor. [2]

Volumen
El volumen se refiere a la cantidad de datos que genera una empresa cada segundo. Estos pueden provenir de diversas fuentes virtuales, como redes sociales, correos electrónicos, dispositivos electrónicos, etc.

Variedad
Según esta explicación sobre qué es el Big Data y para qué sirve, la variedad concierne a todos los lugares donde los datos pueden ser almacenados y extraídos.

Velocidad
La velocidad con la que el Big Data analiza los datos es muy alta, y a eso es lo que se refiere la tercera “V”.

El proceso analiza los datos en el momento exacto en que son creados. Esto ocurre, por ejemplo, con los mensajes que se viralizan en las redes sociales, las transacciones realizadas a través de tarjetas de débito y crédito, etc.

Veracidad
Entre los miles de datos que se generan todos los días, muchos pueden llegar a ser falsos por lo que es preciso excluirlos del análisis.

Cuando comprendes qué es el Big Data y para qué sirve, te das cuenta de que este proceso ayuda a "filtrar" lo que es real de lo que no lo es.

Uno de los principios para esta parte del proceso es que, si son varias las fuentes que apuntan a una determinada información, entonces se entiende que esa es verdadera.

Valor
El objetivo de tener acceso a tanta información es hacer que, de alguna manera, agregue valor a tu empresa.

El Big Data tiene precisamente este propósito: hacer un análisis preciso de todos esos datos y generar insights valiosos para los gestores que los utilizarán.

Importancia del Big Data

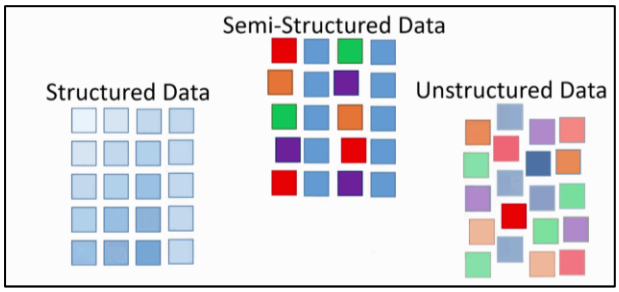
La importancia del big data no gira en torno a la cantidad de datos que tienes, sino en lo que haces con ellos. Puedes tomar datos de cualquier fuente y analizarlos para encontrar respuestas que permitan:

- Reducir los costos
- Reducir el tiempo
- Desarrollar nuevos productos
- Optimizar las ofertas
- Tomar decisiones inteligentes.

Cuando se combinan grandes datos con análisis de alta potencia, se pueden realizar tareas relacionadas con los negocios como:

- Determinar las causas de origen de fallos, problemas y defectos casi en tiempo real.
- Generar cupones en el punto de venta basados en los hábitos de compra del cliente.
- Recalcular portafolios de riesgo completos en minutos.
- Detecte el comportamiento fraudulento antes de que afecte a su organización.

Tipos de Big Data



- *Fig. 1. Tipos de Big Data.*

- Los datos estructurados tienen un formato fijo y a menudo son numéricos. Así que en muchos casos los gestionan máquinas y no humanos. Este tipo de datos es información que ya está ordenada en bases de datos y hojas de cálculo almacenadas en bases de datos SQL, lagos de datos y almacenes de datos.
- Los datos no estructurados son información que está desorganizada y no está en un formato predeterminado porque puede ser casi cualquier cosa. Es el caso, por ejemplo, de los datos recopilados de fuentes de redes sociales y puede convertirse en archivos de documentos de texto almacenados en Hadoop, como clústeres o sistemas NoSQL.
- Los datos semiestructurados pueden contener ambas formas de datos, como registros de servidores web o datos de sensores que haya configurado. Para ser precisos, son datos que, a pesar de no estar clasificados en un repositorio concreto (una base de datos), contienen información vital o etiquetas que segregan elementos individuales dentro de los datos.

El Big Data incluye siempre múltiples fuentes y la mayor parte del tiempo es de distintos tipos también. Así que no siempre es fácil saber cómo integrar todas las herramientas que necesitas para trabajar con distintos tipos de datos.

II. TECNOLOGÍA

Una herramienta de software para analizar, procesar e interpretar la cantidad masiva de datos estructurados y no estructurados que no se podían procesar manual o tradicionalmente se llama Tecnología de Big Data. Esto ayuda a formar conclusiones y pronósticos sobre el futuro para que se puedan evitar muchos riesgos. Los tipos de tecnologías de big data son operativas y analíticas. La tecnología operativa se ocupa de las actividades diarias, como las transacciones en línea, las interacciones en las redes sociales, etc., mientras que la tecnología analítica se ocupa del mercado de valores, el pronóstico del tiempo, los cálculos científicos, etc. Las tecnologías de Big Data se encuentran en el almacenamiento de datos y la minería, la visualización y el análisis. [3]

Apache Spark

Es un motor de procesamiento de big data rápido. Esto se construye teniendo en cuenta el procesamiento en tiempo real de los datos. Su rica biblioteca de aprendizaje automático es buena para trabajar en el espacio de LA IA y el APRENDIZAJE. Procesa datos en paralelo y en equipos agrupados. El tipo de datos básico utilizado por Spark es RDD (conjunto de datos distribuido resistente). [3]

Bases de datos NoSQL

Es una base de datos no relacional que proporciona un rápido almacenamiento y recuperación de datos. Su capacidad para tratar con todo tipo de datos, como datos estructurados, semiestructurados, no estructurados y polimórficos, es única. [3]

Base de datos NoSQL es de los siguientes tipos:

- Bases de datos dedocumentos: Almacena datos en forma de documentos que pueden contener muchos pares clave-valor diferentes. [3]

- Almacenes de gráficos:Almacena datos que generalmente se almacenan en forma de red, como datos de redes sociales. [3]
- Almacenes clave-valor:Estas son las bases de datos NoSQL más simples. Todos y cada uno de los elementos de la base de datos se almacenan como un nombre de atributo (o 'clave'), junto con su valor. [3]
- Almacenes de columnas anchas:esta base de datos almacena datos en formato de columnas en lugar de un formato basado en filas. Cassandra y HBase son buenos ejemplos de ello. [3]

Apache Kafka

Kafka es una plataforma de transmisión de eventos distribuida que maneja muchos eventos todos los días. Como es rápido y escalable, esto es útil para crear canalizaciones de datos de transmisión en tiempo real que obtengan datos de manera confiable entre sistemas o aplicaciones. [3]

Apache Oozie

Es un sistema de programación de flujo de trabajo para administrar los trabajos de Hadoop. Estos trabajos de flujo de trabajo se programan en forma de gráficos acíclicos dirigidos (DAG) para las acciones. [3]

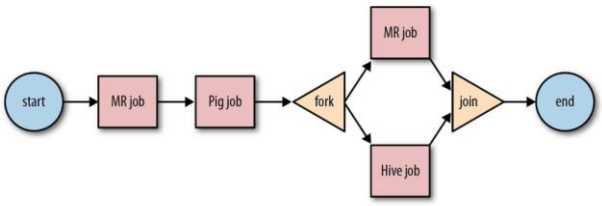


Fig. 2 Diagrama de flujo Apache Oozie [3]

Flujo de aire Apache

Esta es una plataforma que programa y monitorea el flujo de trabajo. La programación inteligente ayuda a organizar la ejecución final del proyecto de manera eficiente. Airflow posee la capacidad de volver a ejecutar una instancia de DAG cuando hay una instancia de error. Su rica interfaz de usuario facilita la visualización de canalizaciones que se ejecutan en varias etapas, como producción, monitoreo de progreso y solución de problemas cuando sea necesario. [3]

Rayo Apache

Es un modelo unificador, para definir y ejecutar canalizaciones de procesamiento de datos que incluyen ETL y transmisión continua. El marco Apache Beam proporciona una abstracción entre la lógica de su aplicación y el ecosistema de big data, ya que no existe una API que vincule todos los marcos como Hadoop, spark, etc. [3]

Pila ELK

ELK es conocido por Elasticsearch, Logstash y Kibana.

Elasticsearch es una base de datos sin esquema (que indexa cada campo) que tiene potentes capacidades de búsqueda y es fácilmente escalable.

Logstash es una herramienta ETL que nos permite obtener, transformar y almacenar eventos en Elasticsearch.

Kibana es una herramienta de dashboarding para Elasticsearch, donde puedes analizar todos los datos almacenados. Los conocimientos procesables extraídos de Kibana ayudan a construir estrategias para una organización. Desde la captura de cambios hasta la predicción, Kibana siempre ha demostrado ser muy útil. [3]

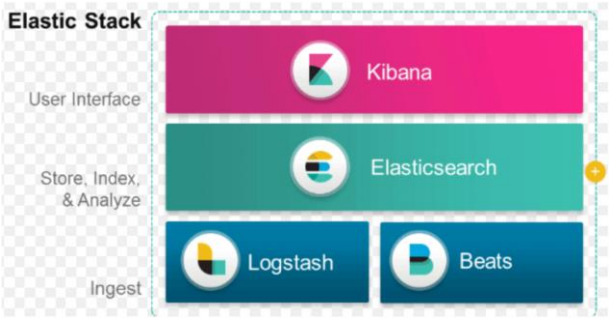


Fig. 3 Pila ELK [3]

Docker y Kubernetes

Estas son las tecnologías emergentes que ayudan a las aplicaciones a ejecutarse en contenedores linux. Docker es una colección de herramientas de código abierto que le ayudan a "Crear, enviar y ejecutar cualquier aplicación, en cualquier lugar". [3]

Kubernetes también es una plataforma de contenedores/orquestación de código abierto, que permite que un gran número de contenedores trabajen juntos en armonía. En última instancia, esto reduce la carga operativa. [3]

TensorFlow

Es una biblioteca de aprendizaje automático de código abierto que se utiliza para diseñar, construir y entrenar modelos de aprendizaje profundo. Todos los cálculos se realizan en TensorFlow con gráficos de flujo de datos. Los gráficos comprenden nodos y aristas. Los nodos representan operaciones matemáticas, mientras que los bordes representan los datos. [3]

TensorFlow es útil para la investigación y la producción. Se ha construido teniendo en cuenta que podría ejecutarse en múltiples CPU o GPU e incluso en sistemas operativos móviles. Esto podría implementarse en Python, C ++, R y Java. [3]

Presto

Presto es un motor SQL de código abierto desarrollado por Facebook, que es capaz de manejar petabytes de datos. A diferencia de Hive, Presto no depende de la técnica MapReduce y, por lo tanto, es más rápido en la recuperación de los datos. Su arquitectura e interfaz son lo suficientemente fáciles como para interactuar con otros sistemas de archivos. [3]

Debido a la baja latencia y las consultas interactivas fáciles, se está volviendo muy popular hoy en día para manejar big data. [3]

Polibase

Polybase funciona sobre SQL Server para acceder a los datos almacenados en PDW (Parallel Data Warehouse). PDW creado para procesar cualquier volumen de datos relacionales y proporciona integración con Hadoop. [3]

Colmena

Hive es una plataforma utilizada para la consulta de datos y el análisis de datos sobre grandes conjuntos de datos. Proporciona un lenguaje de consulta similar a SQL llamado HiveQL,que internamente se convierte en MapReduce y luego se procesa. [3] [4]

III. CARACTERÍSTICAS

Big Data contiene una gran cantidad de datos que no están siendo procesados por el almacenamiento de datos tradicional o la unidad de procesamiento. Es utilizado por muchas empresas multinacionales para procesar los datos y negocios de muchas organizaciones. El flujo de datos excedería los 150 exabytes por día antes de la replicación. [4]

Big Data cuenta con 5 características fundamentales que se describen a continuación:

Volumen

El volumen se refiere a las cantidades inimaginables de información generada cada segundo desde las redes sociales, teléfonos celulares, automóviles, tarjetas de crédito, sensores M2M, imágenes, videos y demás. W eestán utilizando actualmente sistemas distribuidos,para almacenar datos en varias ubicaciones y reunidos por un software Framework como Hadoop. [4]

Solo Facebook puede generar alrededor de mil millones de mensajes, 4.5 mil millones de veces que se graba el botón "me gusta", y se cargan más de 350 millones de nuevas publicaciones cada día. Una cantidad tan grande de datos solo puede ser manejada por Big Data Technologies. [4]

Variedad

Como se discutió anteriormente, el Big Data se genera en múltiples variedades. En comparación con los datos tradicionales como números de teléfono y direcciones, la última tendencia de datos es en forma de fotos, videos y audios y muchos más, lo que hace que aproximadamente el 80% de los datos estén completamente desestructurados. [4]

Veracidad

La veracidad significa básicamente el grado de fiabilidad que los datos tienen para ofrecer. Dado que una parte importante de los datos no está estructurada y es irrelevante, Big Data necesita encontrar una forma alternativa de filtrarlos o traducirlos, ya que los datos son cruciales en el desarrollo del negocio. [4]

Valor

El valor es el principal tema en el que debemos concentrarnos. No es solo la cantidad de datos que almacenamos o procesamos. En realidad, es la cantidad de datos valiosos, confiables y confiables que deben almacenarse, procesarse, analizarse para encontrar información. [4]

Velocidad

Por último, pero no menos importante, la velocidad juega un papel importante en comparación con las demás, no tiene sentido invertir tanto para terminar esperando los datos. Por lo tanto, el aspecto principal de Big Dat es proporcionar datos bajo demanda y a un ritmo más rápido. [4]

Fig. 4 Características de Big Data [4]

IV. OPERACIÓN

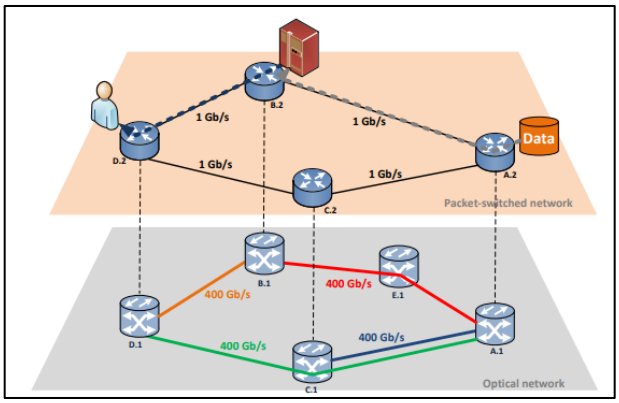
El flujo de trabajo universal de ancho de banda que exige procesamiento de big data en redes ópticas se puede describir de la siguiente manera y son tomadas de la referencia [3] :

1. Primero, se asume que el cliente que envía una solicitud de big data se encuentra lejos del lugar donde se ubica la fuente de datos (sin procesar). Por lo tanto, la conexión entre esas dos entidades no se puede establecer mediante una red de área local (LAN) o una red de área metropolitana (MAN) y, como resultado, se debe aplicar una red de área amplia (WAN). Un elemento adicional del sistema son los nodos de computación (centros de datos) conectados a la red que son capaces de procesar los datos en poco tiempo y con bajo costo. Se enfoca en las solicitudes que se refieren a datos de gran tamaño (por ejemplo, procesamiento de imágenes hiper espectrales, genómica) y, por lo tanto, la capacidad de la red es el recurso clave para atender las solicitudes en un tiempo aceptable. En el flujo de trabajo considerado, los datos de entrada se transfieren desde el nodo de origen (donde se crean o almacenan los datos) a un nodo informático (centro de datos) que procesa los datos.
2. A continuación, los datos de salida (resultados de los cálculos) se envían desde el nodo informático al nodo cliente. Si hay más de un nodo informático disponible, un sistema de programación puede optimizar la asignación de solicitudes a los centros de datos para minimizar el tiempo de respuesta, el uso de energía, los costos de operación, etc. El sistema de procesamiento de datos, descrito anteriormente, se encuentra en la parte superior de una WAN.
3. Se supone una arquitectura de red de dos capas siguiente: red de conmutación de paquetes (PSN) sobre una red óptica (ON). En la capa inferior, los caminos ópticos (caminos de luz) se establecen sobre la red óptica física transparente (que comprende OXC conectados a través de fibras ópticas). La capa superior se basa en un enfoque de conmutación de paquetes (por ejemplo, MPLS) y las conexiones (por ejemplo, LSP) se enrutan sobre la topología lógica de los caminos de luz (la red óptica virtual). La arquitectura propuesta es relativamente simple, sin embargo, corresponde a los enfoques más populares (por ejemplo, MPLS sobre WDM) y, según muchas predicciones, será el enfoque predominante de despliegue de red en un futuro próximo.
4. Se asume dos posibles escenarios de aprovisionamiento de tráfico de big data en la arquitectura de red considerada.
    a. El primer escenario, denominado aprovisionamiento de la red de servicios, asume que el tráfico relacionado con el análisis de datos se transporta en la capa PSN utilizando la topología virtual existente de rutas de luz, es decir, la red está sobre aprovisionada y queda algo de capacidad adicional para transportar el tráfico de big

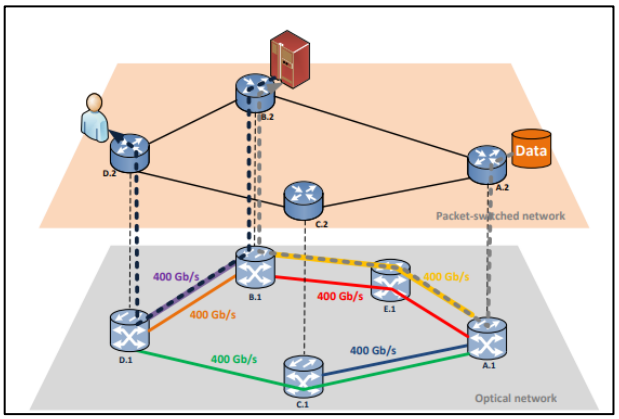
data según nuevas solicitudes. En consecuencia, la topología de la trayectoria de luz es fija y no es posible establecer nuevas trayectorias de luz bajo demanda. Sin embargo, los recursos limitados de la capacidad de la red pueden aumentar significativamente el tiempo de respuesta del sistema a valores inaceptables con respecto a las aplicaciones de big data que requieren una reacción muy rápida.

b. El segundo escenario, conocido como aprovisionamiento de red óptica, permite un establecimiento dinámico de senderos de luz de corta duración totalmente dedicados a transportar el tráfico de big data. Más detalladamente, cuando llega una nueva solicitud de análisis de datos, el plano de control de la red puede decidir establecer un nuevo camino de luz en el dominio óptico para conectar directamente el nodo fuente al nodo informático y/o luego conectar el nodo informático al nodo cliente. Esta posibilidad permite transmisiones con anchos de banda muy altos (por ejemplo, hasta 400 Gb/s en EON) y, en consecuencia, debería proporcionar un tiempo de respuesta muy corto.

En la Figura 2, se enseña un ejemplo ilustrativo de la arquitectura de red y los escenarios de aprovisionamiento de solicitudes. El nodo A.2 almacena los datos de origen, el nodo B.2 alberga un centro de datos que proporciona el procesamiento (cálculos) y el cliente que solicita los resultados (datos de salida) se encuentra en el nodo D.2. El escenario de aprovisionamiento de la red de servicios se presenta en la Figura 2 A, mientras que el escenario de aprovisionamiento de la red óptica se ilustra en la Figura 2 B. En el primer caso, el tráfico de datos entre elementos del sistema (repositorio de datos, nodo informático, cliente) se aprovisiona en la PSN con un ancho de banda relativamente pequeño (por ejemplo, 1 Gb/s), lo que se deriva de los recursos limitados de capacidad sobre aprovisionada. Por el contrario, en el escenario de aprovisionamiento de red óptica, se establecen dos rutas de luz adicionales para servir el tráfico de big data y las rutas de luz se pueden asignar con hasta 400 Gb/s de capacidad.



(a)



(b)

Fig. 5. Escenarios para el aprovisionamiento de solicitudes de macrodatos en una arquitectura de red de dos capas: (a) escenario de aprovisionamiento de red de servicios (b) escenario de aprovisionamiento de red óptica. [3]

V. VENTAJAS

Rendimiento: definido como el volumen de datos (tasa de bits) que pueden fluir a través de una red expresado en bps. El rendimiento es un parámetro importante que influye en gran medida en el tiempo de respuesta del análisis de big data. Además, el nuevo concepto EON proporciona mejoras significativas en términos de ancho de banda y elasticidad del espectro de capacidad. [3]

Escalabilidad: la red debe permitir la entrega elástica bajo demanda de una gran cantidad de datos de acuerdo con las demandas cambiantes, lo que requiere un control de conectividad automatizado para permitir el uso dinámico de los recursos de la red y mejorar la configuración de la red. El concepto reciente de Red definida por software (SDN) es el

elemento clave que se espera que haga que las redes sean programables, fáciles de particionar y virtualizables. [3]

Resiliencia y seguridad: dado que algunos resultados del análisis de big data son de gran importancia (por ejemplo, seguridad pública, advertencias de tornados, advertencias de terremotos, etc.), la red involucrada en la entrega de datos cuenta con respaldo de mecanismos sólidos de protección y seguridad. Las soluciones de resistencia y seguridad que se ofrecen actualmente en las capas PSN y ON cumplen y deben ser suficientes para proteger el tráfico de big data. [3]

Bajo costo: reducir tanto los gastos de capital (CAPEX) como los costos operativos (OPEX) es un requisito clave para la mayoría de los clientes y ofrecer más ancho de banda a un costo menor es un desafío importante para los operadores de red y proveedores de servicios. Debido al tamaño del tráfico de big data, la capacidad requerida solo se puede aprovisionar en la capa óptica. De acuerdo con nuestros resultados recientes mostrados en el uso de arquitecturas EON mejora significativamente el costo métrico, el consumo de energía y el uso del espectro en comparación con el enfoque WDM convencional. [3]

Eficiencia energética: el costo de la energía es uno de los principales contribuyentes a los gastos operativos en los sistemas y redes informáticos. Además, se está prestando mucha atención a las soluciones de TIC "verdes" para responder a los desafíos que siguen a la amenaza de la crisis energética y los problemas de protección del medio ambiente. El análisis de big data consume la mayor parte de la energía en: red de transporte, centros de datos (procesamiento y análisis) y sistemas de almacenamiento. En el caso de las redes ópticas, el WDM necesita entre un 36% y un 49% más de energía que EON. Además, el alto ancho de banda proporcionado en las redes ópticas permite una programación fácil y flexible de tareas de big data en sitios que proporcionan un menor costo de energía o energía verde, sin un deterioro significativo de parámetros como el tiempo de respuesta o el rendimiento. [3]

VI. REFERENCIAS