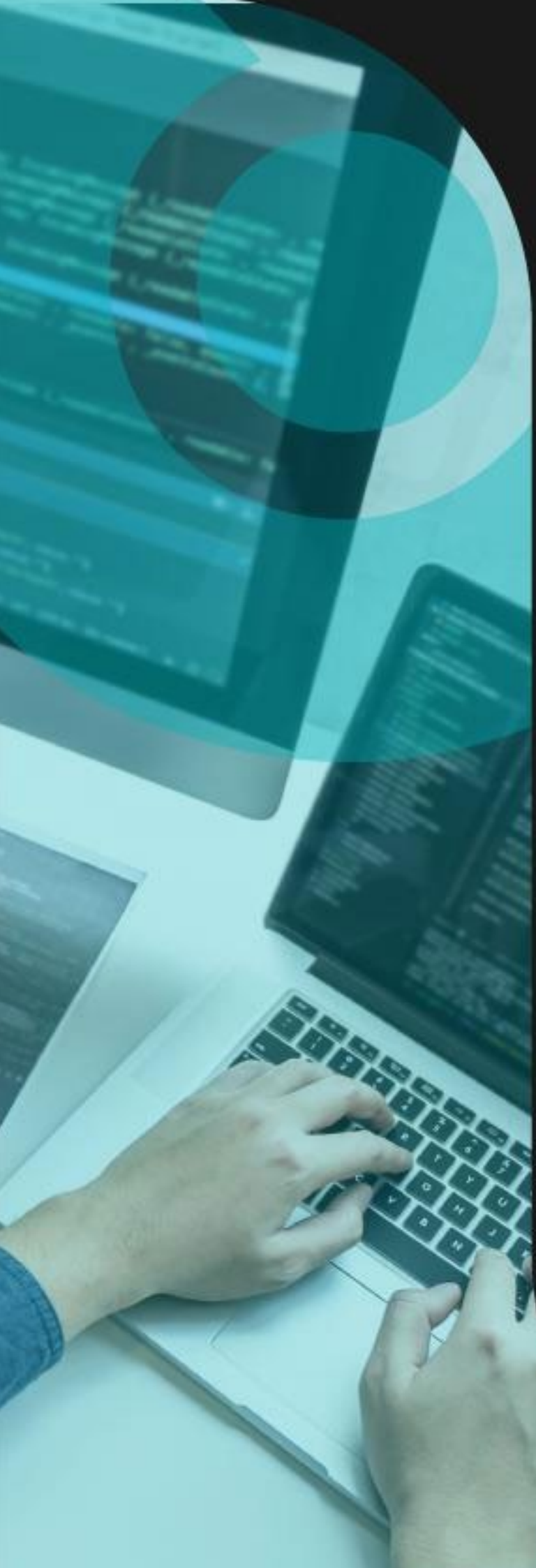


AWAKELAB

BASECAMP

Ciencia de Datos

Módulo: Fundamentos del Big Data

Aprendizaje Esperado

---

1. Describir las características fundamentales de Big Data y su ecosistema para el manejo de grandes volúmenes de datos.

Introducción a Big Data

Big Data *es una colección de datos que es enorme en volumen, pero que crece exponencialmente con el tiempo. Se trata de datos con un tamaño y una complejidad tan grandes que ninguna de las herramientas tradicionales de gestión de datos puede almacenarlos y procesarlos de manera eficiente. Big data también es un dato pero con un tamaño enorme.*

Sistemas de procesamiento y distribución de Big Data

Los sistemas de procesamiento y distribución de big data ofrecen una forma de recopilar, distribuir, almacenar y administrar conjuntos de datos masivos y no estructurados en tiempo real. Estas soluciones proporcionan una forma sencilla de procesar y distribuir datos entre clústeres de computación paralelos de manera organizada. Diseñados para escalar, estos productos están creados para ejecutarse en cientos o miles de máquinas simultáneamente, cada una de las cuales proporciona capacidades de almacenamiento y computación locales. Los grandes sistemas de procesamiento y distribución de datos brindan un nivel de simplicidad al problema comercial común de la recopilación de datos a gran escala y son utilizados con mayor frecuencia por empresas que necesitan organizar una cantidad exorbitante de datos. Muchos de estos

productos ofrecen una distribución que se ejecuta sobre la herramienta de agrupación de big data de código abierto Hadoop.

Las empresas suelen tener un administrador dedicado para administrar los clústeres de big data. El rol requiere un conocimiento profundo de la administración de bases de datos, la extracción de datos y la escritura de lenguajes de secuencias de comandos del sistema host. Las responsabilidades del administrador a menudo incluyen la implementación del almacenamiento de datos, el mantenimiento del rendimiento, el mantenimiento, la seguridad y la extracción de conjuntos de datos. Las empresas a menudo usan herramientas de análisis de big data para luego preparar, manipular y modelar los datos recopilados por estos sistemas.

Para calificar para la inclusión en la categoría Sistemas de distribución y procesamiento de Big Data, un producto debe:

- Recopile y procese grandes conjuntos de datos en tiempo real.
- Distribuir datos a través de clústeres informáticos paralelos.
- Organice los datos de tal manera que los administradores del sistema puedan administrarlos y extraerlos para su análisis.
- Permite que las empresas escalen las máquinas al número necesario para almacenar sus datos.

Problemas que se pueden resolver con Big Data

Aquí está la lista de las 10 principales industrias que utilizan aplicaciones de big data:

1. Banca y Valores
2. Comunicaciones, Medios y Entretenimiento
3. Proveedores de servicios de salud
4. Educación
5. Manufactura y Recursos Naturales
6. Gobierno

7. Seguro
8. Comercio al por menor y al por mayor
9. Transporte
10. Energía y servicios Públicos

Hadoop

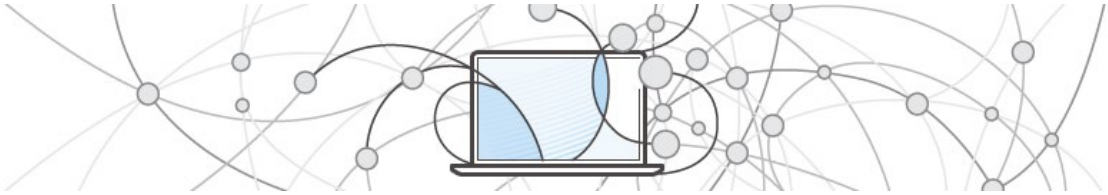
¿Qué es Hadoop?

Apache Hadoop es un marco de código abierto que se utiliza para almacenar y procesar de manera eficiente grandes conjuntos de datos que varían en tamaño desde gigabytes hasta petabytes de datos. En lugar de usar una computadora grande para almacenar y procesar los datos, Hadoop permite agrupar varias computadoras para analizar conjuntos de datos masivos en paralelo con mayor rapidez.

Hadoop consta de cuatro módulos principales:

- Sistema de archivos distribuido de Hadoop (HDFS): un sistema de archivos distribuido que se ejecuta en hardware estándar o de gama baja. HDFS proporciona un mejor rendimiento de datos que los sistemas de archivos tradicionales, además de una alta tolerancia a fallas y soporte nativo de grandes conjuntos de datos.
- Otro negociador de recursos más (YARN): administra y supervisa los nodos del clúster y el uso de recursos. Programa trabajos y tareas.
- MapReduce: un marco que ayuda a los programas a realizar el cálculo paralelo de los datos. La tarea del mapa toma los datos de entrada y los convierte en un conjunto de datos que se puede calcular en pares de valores clave. La salida de la tarea del mapa se consume al reducir las tareas para agregar la salida y proporcionar el resultado deseado.

- Hadoop Common: proporciona bibliotecas Java comunes que se pueden usar en todos los módulos.



Cómo funciona Hadoop

Hadoop facilita el uso de toda la capacidad de almacenamiento y procesamiento en servidores de clúster y la ejecución de procesos distribuidos contra grandes cantidades de datos. Hadoop proporciona los componentes básicos sobre los que se pueden construir otros servicios y aplicaciones.

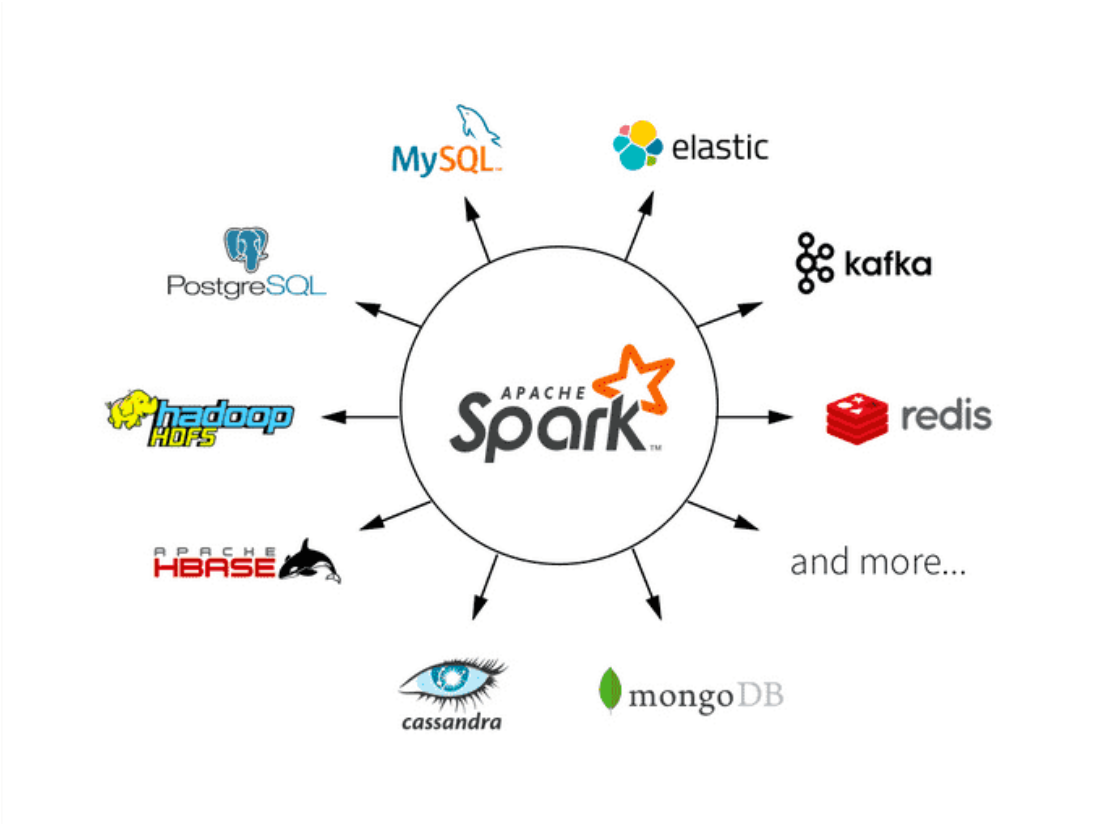
Las aplicaciones que recopilan datos en varios formatos pueden colocar datos en el clúster de Hadoop mediante una operación de API para conectarse a NameNode. El NameNode rastrea la estructura del directorio de archivos y la ubicación de "fragmentos" para cada archivo, replicados en DataNodes. Para ejecutar un trabajo para consultar los datos, proporcione un trabajo de MapReduce compuesto por muchas tareas de mapeo y reducción que se ejecutan contra los datos en HDFS distribuidos en los DataNodes. Las tareas de asignación se ejecutan en cada nodo con los archivos de entrada suministrados y los reductores se ejecutan para agregar y organizar el resultado final.

El ecosistema de Hadoop ha crecido significativamente a lo largo de los años debido a su extensibilidad. Hoy en día, el ecosistema de Hadoop incluye muchas herramientas y aplicaciones para ayudar a recopilar, almacenar, procesar, analizar y administrar big data. Algunas de las aplicaciones más populares son:

- Spark: un sistema de procesamiento distribuido de código abierto que se usa comúnmente para grandes cargas de trabajo de datos. Apache Spark utiliza el almacenamiento en caché en la memoria y la ejecución optimizada para un rendimiento rápido, y es compatible con el procesamiento general por lotes, el análisis de transmisión, el aprendizaje automático, las bases de datos gráficas y las consultas ad hoc.
- Presto: un motor de consulta SQL distribuido de código abierto optimizado para el análisis ad-hoc de datos de baja latencia. Admite el estándar ANSI SQL, incluidas consultas complejas, agregaciones, uniones y funciones de ventana. Presto puede procesar datos de varias fuentes de datos, incluido el sistema de archivos distribuidos de Hadoop (HDFS) y Amazon S3.
- Hive: permite a los usuarios aprovechar Hadoop MapReduce mediante una interfaz SQL, lo que permite el análisis a gran escala, además del almacenamiento de datos distribuido y tolerante a fallas.
- HBase: una base de datos versionada, no relacional y de código abierto que se ejecuta sobre Amazon S3 (mediante EMRFS) o el sistema de archivos distribuidos de Hadoop (HDFS). HBase es un gran almacén de datos masivo distribuido y escalable creado para el acceso aleatorio, estrictamente consistente y en tiempo real para tablas con miles de millones de filas y millones de columnas.
- Zeppelin: un cuaderno interactivo que permite la exploración interactiva de datos.

Apache Spark

Apache Spark es un motor de análisis unificado ultrarrápido para big data y aprendizaje automático. Fue desarrollado originalmente en UC Berkeley en 2009.



Qué es Apache Spark

- Velocidad

Apache Spark es una plataforma de procesamiento de datos de código abierto y distribuida que se destaca por su velocidad y eficiencia en el procesamiento de grandes volúmenes de datos. Gracias a su capacidad para explotar la informática en memoria y otras optimizaciones, Spark puede procesar datos hasta 100 veces más rápido que Hadoop, lo que lo convierte en una opción ideal para el procesamiento de datos en tiempo real y a gran escala.

- Facilidad de uso

Apache Spark cuenta con una API fácil de usar que permite a los usuarios trabajar con grandes conjuntos de datos sin tener que ser expertos en programación. Con más de 100 operadores para transformar datos y API de marcos de datos familiares para manipular datos semiestructurados,

Spark simplifica el proceso de procesamiento de datos y permite a los usuarios centrarse en la generación de información valiosa.

- Un motor unificado

Apache Spark es un motor unificado que viene empaquetado con bibliotecas de nivel superior, lo que significa que los usuarios pueden aprovechar una amplia variedad de herramientas y capacidades en una sola plataforma. Las bibliotecas estándar de Spark aumentan la productividad de los desarrolladores y pueden combinarse a la perfección para crear flujos de trabajo complejos y personalizados. Esto hace que Spark sea una excelente opción para cualquier persona que necesite procesar grandes volúmenes de datos de manera rápida y eficiente.

Por qué se necesita Spark

Si está trabajando con Spark, se encontrará con las tres API: DataFrames, Datasets y RDD.

Recordemos que Spark (Apache Spark) es un framework de programación para procesamiento de datos distribuidos diseñado para ser rápido y de propósito general. Como su propio nombre indica, ha sido desarrollada en el marco del proyecto Apache, lo que garantiza su licencia Open Source.

Además, podremos contar con que su mantenimiento y evolución se llevarán a cabo por grupos de trabajo de gran prestigio, y existirá una gran flexibilidad e interconexión con otros módulos de Apache como Hadoop, Hive o Kafka.

Parte de la esencia de Spark es su carácter generalista. Consta de diferentes APIs y módulos que permiten que sea utilizado por una gran variedad de profesionales en todas las etapas del ciclo de vida del dato. Dichas etapas pueden incluir desde soporte para análisis interactivo de datos con SQL a la creación de complejos pipelines de machine learning y procesamiento en streaming, todo usando el mismo motor de procesamiento y las mismas APIs.

¿Qué son los conjuntos de datos distribuidos resistentes?

RDD o Resilient Distributed Datasets, es una colección de registros con computación distribuida, que son tolerantes a fallas, de naturaleza inmutable. Se pueden operar en paralelo con API de bajo nivel, mientras que su característica perezosa hace que la operación de chispa funcione a una velocidad mejorada. Los RDD admiten dos tipos de operaciones:

- Transformaciones: operaciones perezosas que devuelven otro RDD, este RDD no se calcula a menos que se realice una acción en él. Algunos ejemplos de transformaciones son map(), flatmap(), filter()

- Acciones: operaciones que activan el cálculo y devuelven valores. Algunos ejemplos de acciones son count, top(), savetofile()

Desventajas de los RDD

Si elige trabajar con RDD, tendrá que optimizar todos y cada uno de los RDD. Además, a diferencia de los conjuntos de datos y los marcos de

datos, los RDD no infieren el esquema de los datos incorporados, por lo que deberá especificarlo.

¿Qué son las tramas de datos?

DataFrames es una colección distribuida de filas bajo columnas con nombres. En términos simples, parece una hoja de Excel con encabezados de columna, o puede considerarla como el equivalente a una tabla en una base de datos relacional o un marco de datos en R o Python. Tiene tres características principales comunes con RDD:

- **De naturaleza inmutable** : podrá crear un DataFrame pero no podrá cambiarlo. Un DataFrame como un RDD se puede transformar
- **Evaluaciones perezosas:** una tarea no se ejecuta hasta que se realiza una acción.
- **Distribuido:** los marcos de datos, al igual que los RDD, son de naturaleza distribuida.

Formas de crear un marco de datos

En Spark DataFrames se pueden crear de varias maneras:

- Uso de diferentes formatos de datos. Como cargar los datos de JSON, CSV, RDBMS, XML o Parquet
- Cargando los datos de un RDD ya existente.
- Esquema de especificación programática

Ventajas del procesamiento de Big Data

La capacidad de procesar Big Data en DBMS brinda múltiples beneficios, como:

- Las empresas pueden utilizar inteligencia externa al tomar decisiones

El acceso a los datos sociales de los motores de búsqueda y sitios como Facebook, Twitter están permitiendo a las organizaciones afinar sus estrategias comerciales.

- Servicio de atención al cliente mejorado

Los sistemas tradicionales de retroalimentación de los clientes están siendo reemplazados por nuevos sistemas diseñados con tecnologías Big Data. En estos nuevos sistemas, Big Data y tecnologías de procesamiento de lenguaje natural se utilizan para leer y evaluar las respuestas de los consumidores.

- Identificación temprana de riesgo para el producto/servicio, si lo hubiere.

- Mejor eficiencia operativa

Las tecnologías de Big Data se pueden utilizar para crear un área de preparación o una zona de aterrizaje para nuevos datos antes de identificar qué datos se deben mover al almacén de datos. Además, dicha integración de las tecnologías Big Data y el almacén de datos ayuda a una organización a descargar datos a los que se accede con poca frecuencia.

Referencias

[1] ¿Qué es el Big Data?

https://www.sas.com/es_mx/insights/big-data/what-is-big-data.html

[2] Usos del Big Data

https://www.zendesk.com.mx/blog/big-data-que-es/

[3] Dónde se aplica Big Data

https://www.masterbigdataucm.com/que-es-big-data/

[4] ¿Qué es Hadoop?

https://www.sas.com/es_mx/insights/big-data/hadoop.html

Complementario

[1] Qué es el big data – video

https://www.youtube.com/watch?v=M26iIqmqWkI

[2] Hadoop vs Spark

https://www.youtube.com/watch?v=g2ibI_-pHvQ