Andrei Alexandru Miu

Introducción a Hadoop

Una vez realizada la tarea, el envío se realizará a través de la plataforma. El archivo se nombrará siguiendo las siguientes pautas:

Apellido1_Apellido2_Nombre_BDA02_Tarea

Asegúrate que el nombre no contenga la letra ñ, tildes ni caracteres especiales extraños. Así por ejemplo la alumna Begoña Sánchez Mañas para la primera unidad del MP de BDA02, debería nombrar esta tarea como...

sanchez_manas_begona_BDA02_Tarea

Introducción

Roberto es el Director de Tecnología de una gran empresa de transportes. En el Comité de Dirección tomaron la decisión de modernizar la empresa y tener una estrategia data-driven, que consiste en basar toda la toma de decisiones en los datos, y en utilizar todos los datos disponibles para generar valor, como optimizar las operaciones, reducir los gastos de mantenimiento, etc.

Roberto, a recomendación de su amiga María, comenzó a investigar sobre Apache Hadoop, ya que es la plataforma que podría resolver todos los casos de uso que le plantea su Comité de Dirección.

Tras unas semanas de investigación y de conocimiento de lo que es Hadoop, cómo funciona y cuándo utilizarlo, se dispone a evaluar si ha aprendido lo suficiente sobre esta plataforma antes de decidirse a instalarlo en su empresa.

NOTA IMPORTANTE

Para todas las preguntas es necesario responder de manera razonada. Es recomendable incluir cualquier diagrama que consideres interesante para explicar la solución que has desarrollado.

Andrei Alexandru Miu

Preguntas

Para esta tarea no es necesario instalar ningún software, sino que se trata de razonar una serie de preguntas relacionadas con Apache Hadoop.

Pregunta 1

Razona qué diferencias hay entre una plataforma y una solución Big Data.

Plataforma Big Data: Es un conjunto de herramientas y tecnologías que permiten almacenar, procesar y analizar grandes volúmenes de datos. Ejemplos de plataformas Big Data incluyen Hadoop, Spark, y otras tecnologías que ofrecen almacenamiento distribuido.

Solución Big Data: Es una implementación específica que utiliza una o varias plataformas Big Data para resolver un problema concreto, como el análisis de datos en tiempo real, la ingesta de datos desde múltiples fuentes, o la generación de informes predictivos.

Pregunta 2

Describe cómo consigue Hadoop la escalabilidad. Indica, además, qué tipo de escalabilidad es la que Hadoop ofrece.

Hadoop consigue la escalabilidad a través de su arquitectura distribuida, que permite añadir más nodos (servidores) al clúster para aumentar la capacidad de almacenamiento y procesamiento. Esto se logra principalmente gracias a dos componentes clave de Hadoop:

- **HDFS (Hadoop Distributed File System):** HDFS divide los datos en bloques grandes (por defecto, 128 MB) y los distribuye en múltiples nodos del clúster. Además, cada bloque se replica en varios nodos (por defecto, 3 réplicas) para garantizar la tolerancia a fallos. Esto permite que el sistema pueda escalar horizontalmente, es decir, añadiendo más nodos para aumentar la capacidad de almacenamiento y procesamiento.
- **YARN (Yet Another Resource Negotiator):** YARN gestiona los recursos del clúster y permite ejecutar múltiples aplicaciones en paralelo. A medida que se añaden más nodos al clúster, YARN puede distribuir las tareas de procesamiento entre los nuevos nodos, lo que permite escalar el procesamiento de datos de manera eficiente.

Pregunta 3

Razona qué se considera hardware commodity, qué beneficios aporta que un sistema utilice este tipo de hardware, y qué dificultades supone.

Hardware commodity se refiere a servidores y componentes de hardware estándar y de bajo costo, que no están diseñados específicamente para un propósito particular

Beneficios de utilizar hardware commodity:
Coste reducido: Al utilizar hardware estándar y de bajo costo, se reduce significativamente la inversión inicial en infraestructura.

Escalabilidad: Es más fácil y económico escalar horizontalmente añadiendo más servidores commodity al clúster.
Flexibilidad: Al no estar atado a hardware específico, se puede actualizar o reemplazar componentes individuales sin afectar al sistema en su conjunto.
Tolerancia a fallos: En sistemas distribuidos como Hadoop, la redundancia y replicación de datos permiten que el sistema siga funcionando incluso si algunos nodos fallan.
Dificultades de utilizar hardware commodity:
Menor rendimiento individual: Los servidores commodity no tienen el mismo rendimiento que los servidores de gama alta, por lo que pueden ser menos eficientes en tareas que requieren alto rendimiento.
Mayor complejidad de gestión: Al tener muchos nodos en un clúster, la gestión y mantenimiento del hardware puede ser más compleja y requerir más esfuerzo.
Mayor probabilidad de fallos: Al utilizar hardware de bajo costo, es más probable que algunos componentes fallen, lo que requiere mecanismos de tolerancia a fallos y replicación de datos.

Pregunta 4

Describe qué componentes del ecosistema Hadoop utilizarías para implementar una arquitectura con los siguientes requisitos:

Debe ser capaz de almacenar un volumen de datos de 1 petabyte.

La información que almacenará tendrá diferentes tipos, desde ficheros binarios a ficheros de textos estructurados.

Debe permitir ingestar datos de tres bases de datos relacionales de los sistemas operacionales.

Los procesos de ingesta deben ser ejecutados de forma planificada, todos los días a las 12 de la noche. Una vez ingestados los datos, se debe lanzar un proceso que generará un resumen diario que requiere procesar todos los datos ingestados en el día.

Además, la arquitectura debe permitir obtener datos de 100 sensores que realizan una medida cada segundo.

La arquitectura debe ofrecer una herramienta de gobierno de datos.

Sobre los ficheros de datos obtenidos de las bases de datos relacionales, la arquitectura debe ofrecer la capacidad para poder realizar consultas en un lenguaje similar a SQL.

- HDFS (Hadoop Distributed File System): Para almacenar el volumen de datos de 1 petabyte. HDFS es ideal para almacenar grandes volúmenes de datos distribuidos en múltiples nodos, y puede manejar diferentes tipos de datos, desde ficheros binarios hasta ficheros de texto estructurados.
- Apache Sqoop: Para ingestar datos de las tres bases de datos relacionales. Sqoop es una herramienta diseñada para transferir datos entre Hadoop y bases de datos relacionales, permitiendo la ingesta de datos de forma eficiente.
- Apache Oozie: Para planificar y ejecutar los procesos de ingesta y el proceso de resumen diario. Oozie es un sistema de workflow que permite programar tareas en Hadoop, como la ejecución de jobs de MapReduce, scripts de Sqoop, y otros procesos.
- Apache Kafka: Para ingestar datos de los 100 sensores que realizan una medida cada segundo. Kafka es una plataforma de streaming que permite ingestar y procesar grandes volúmenes de datos en tiempo real, ideal para manejar datos de sensores.
- Apache Hive: Para ofrecer la capacidad de realizar consultas en un lenguaje similar a SQL sobre los datos almacenados en HDFS. Hive permite ejecutar consultas SQL sobre datos almacenados en Hadoop, lo que facilita el análisis de datos para usuarios familiarizados con SQL.
- Apache Atlas: Para ofrecer una herramienta de gobierno de datos. Atlas es una herramienta de gobierno y metadata que permite rastrear el origen de los datos, su linaje, y asegurar que se cumplen las políticas de gobierno de datos.

Pregunta 5

En el caso de una startup que tiene una previsión de almacenar y procesar 500 terabytes de datos en los próximos 6 meses, razona qué tipo de solución Hadoop sería la más apropiada (cloud o una distribución comercial en una infraestructura de servidores propia).

Para una startup con una previsión de crecimiento rápido y recursos limitados, una solución Hadoop en la nube es más adecuada debido a su flexibilidad, menor costo inicial y facilidad de gestión.

Pregunta 6

Razona cuáles son las principales dificultades que un sistema distribuido tiene frente a un sistema centralizado.

Los sistemas distribuidos ofrecen ventajas en términos de escalabilidad y tolerancia a fallos, pero presentan dificultades en términos de gestión, consistencia, rendimiento y seguridad.