Integración PySpark

visión general del negocio

Apache Spark es un motor de procesamiento distribuido que es de código abierto y se utiliza para aplicaciones de datos de gran tamaño. Utiliza el almacenamiento en caché en memoria y la ejecución eficiente de consultas para consultas analíticas rápidas contra cualquier cantidad de datos. Ofrece reutilización de código en muchas cargas de trabajo, como procesamiento por lotes, consultas interactivas, análisis en tiempo real, aprendizaje automático y procesamiento de gráficos. Proporciona API de desarrollo en Java, Scala, Python y R.

Canalización de datos

Una canalización de datos es una técnica para transferir datos de un sistema a otro. Los datos pueden actualizarse o no, y pueden manejarse en tiempo real (o transmisión) en lugar de por lotes. La canalización de datos abarca todo, desde la recolección o adquisición de datos utilizando varios métodos hasta el almacenamiento de datos sin procesar, la limpieza, la validación y la transformación de datos en un formato digno de consulta, la visualización de KPI y la gestión del proceso anterior.

Agenda

Este es el sexto proyecto de la serie Pyspark. Elquinto proyecto involucró la introducción a PySpark SQL, la función SQL y varias uniones disponibles en PySpark SQL con la ayuda de un estudio de caso de negocios. Este proyecto se enfoca principalmente en la integración de PySpark con Amazon S3 y la base de datos MySQL para realizar operaciones ETL (Extract-Transform-Load) y ELT (Extract-Load-Transform).

Pila de tecnología:

➔Idioma: pitón
➔Paquete: Pyspark
➔Servicios: AWS S3, MySQL

PySpark:

PySpark es una interfaz de Python para Apache Spark. No solo le permite desarrollar aplicaciones Spark utilizando las API de Python, sino que también incluye el shell PySpark para examinar datos de forma interactiva en un contexto distribuido. PySpark es compatible con la mayoría de las capacidades de Spark, incluidas Spark SQL, DataFrame, Streaming, MLlib y Spark Core. En este proyecto, Ud.

aprenderá sobre la arquitectura central de Spark, las sesiones de Spark, la transformación, las acciones y las técnicas de optimización utilizando PySpark.

mysql

MySQL es un sistema de gestión de base de datos relacional basado en SQL (Structured Query Language). La plataforma se puede utilizar para almacenamiento de datos, comercio electrónico, aplicaciones de registro, etc.

Amazonas S3

Amazon S3 es una interfaz de servicio web simple para el almacenamiento de objetos que le permite almacenar y recuperar cantidades ilimitadas de datos desde cualquier lugar de Internet.

Se utiliza para realizar copias de seguridad y archivar datos en las instalaciones o en la nube Almacenamiento y distribución de contenido, medios y software

Conclusiones clave:

- Comprender la descripción general del proyecto
- Introducción a PySpark
- Introducción a Amazon S3
- Crear depósito en Amazon S3
- Almacenar datos en depósitos de S3
- Introducción a la base de datos MySQL
- Necesidad de integración PySpark
- Comprender el concepto de ETL
- Diferencia entre ETL y ELT
- Integración de PySpark con Amazon S3
- Integración de PySpark con la base de datos MySQL