Spark Streaming

Streaming


Distributed
Processing System
Raw Data Streams
Processed Data

¿Qué es Spark Streaming?

Recibe datos streaming de diferentes fuentes.

Estos datos se procesan en cluster distribuido.

Coloca la información en una base de datos.

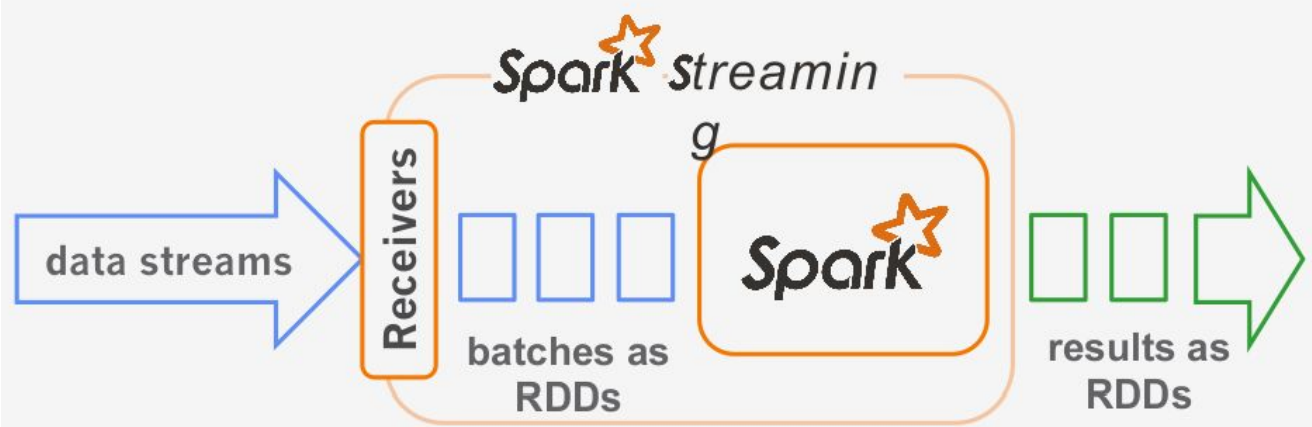
Es escalable y ofrece tolerancia a fallos.

Spark Streaming

¿Cómo funciona Spark Streaming?

Corta los flujos de datos en lotes de pocos segundos.

Spark trata cada lote de datos como RDD y los procesa mediante operaciones RDD. Los resultados procesados son expulsados en lotes



Ejemplo: Contador de palabras (Streaming)

Dede Spark-shell

1.Creamos el DataFrame para leer los datos que llegan desde el puerto 8585

scala>val lines=spark.readStream.format(“socket”).option(“host,”localhost”).option(“port”. 8585).load

2.Hacemos un split de las palabras.

scala>val words=lines.as[String].flatMap(_.split(“ “))

Ejemplo: Contador de palabras (Streaming)

3.Contador de palabras

scala>val wordCounts=words.groupBy(“value”).count()

4.En otra terminal usamos el siguiente comando con netcat.

nc -lk 8585

5.Nuevamente desde escala (programa anterior) imprimimos el resultado del contador.

scala>val query=wordCounts.writeStream.ouputMode(“complete”).format(“console”).start()

Ejemplo: Contador de palabras (Streaming)

6.Desde la terminal que está corriendo netcat (nc) escribimos cualquier mensaje

Por ejemplo:

Uno, uno, uno, dos, dos, tres.

7.Veremos como en la terminal de spark-shell se cuentan las palabras qué llegan por la red.

Streaming Twitter Data

El siguiente ejemplo usa datos en vivo de Twitter para encontrar una tendencia en los últimos 5 minutos.

1.Crear una cuenta de twitter y no la tienes.

2.Ir a http://apps.twitter.com

3.Obtener los siguientes datos:

Consumer Key (API Key):

Consumer Secret (API Secret):

Access Token:

Access Token Secret:

Ejemplo Twitter

https://www.toptal.com/apache/tutorial-apache-spark-streaming-identificando-los-hashtags-de-tendencia-de-twitter