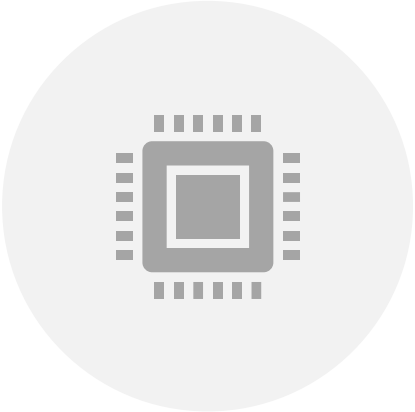
2.Spark Structured Streaming

Spark Structured Streaming

Conceitos

BATCH: PROCESSAMENTO DE CONJUNTO DE DADOS



STREAMING: PROCESSAMENTO A MEDIDA QUE OS DADOS SÃO PRODUZIDOS

MICRO-BATCHS: BLOCO DE DADOS PRODUZIDOS EM INTERVALO DE TEMPO

Structured Streaming

- Segunda geração de processamento de streaming de Spark (Dstream foi a primeira)
- Garantia de processamento único de cada registro (end-to-end exactly-once guarantees)

Modos de Saída

APPEND: SÓ NOVAS LINHAS. SUPORTA APENAS DE CONSULTAS STATELESS

UPDATE: APENAS LINHAS QUE FORAM ATUALIZADAS

COMPLETE: TODA A TABELA É ATUALIZADA

Trigger

Formas:

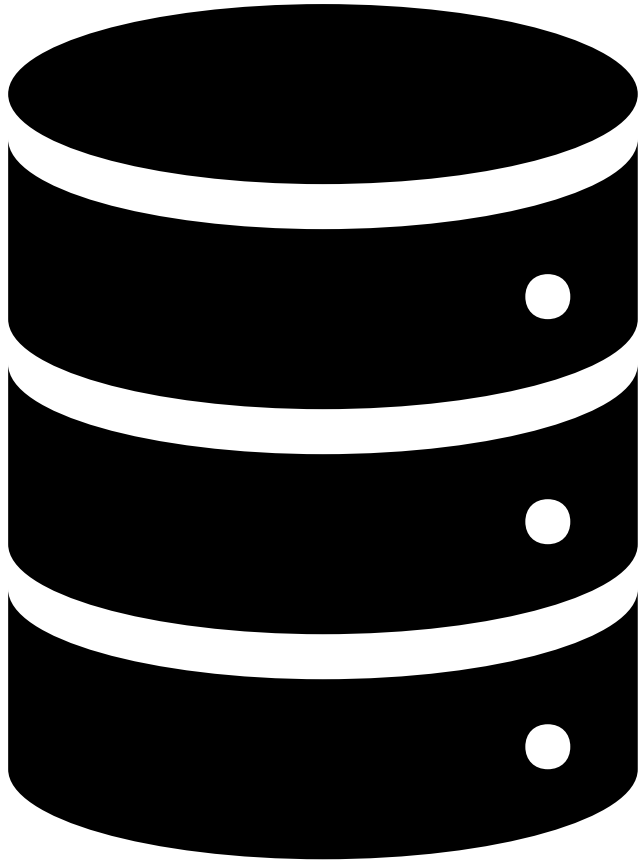
- Default: dispara quando o micro batch termina
- Tempo
- Once: apenas uma única vez
- Continuous: processamento contínuo

Parar o processo

- stop()

Checkpointdir

- Diretório onde o estado de andamento é salvo
- Se você parar o processo e reiniciar com o mesmo diretório, ele segue de onde parou



Métodos semelhantes os de batch

- readstream em vez de read
- writestream em vez de write

Source e Sinks que não tem suporte

- Métodos de batch podem ser usados (read, write):
  - foreachbatch: opera no micro-batch
  - Foreach: opera a cada linha
- Algumas garantias são perdidas: por exemplo, exactly-once