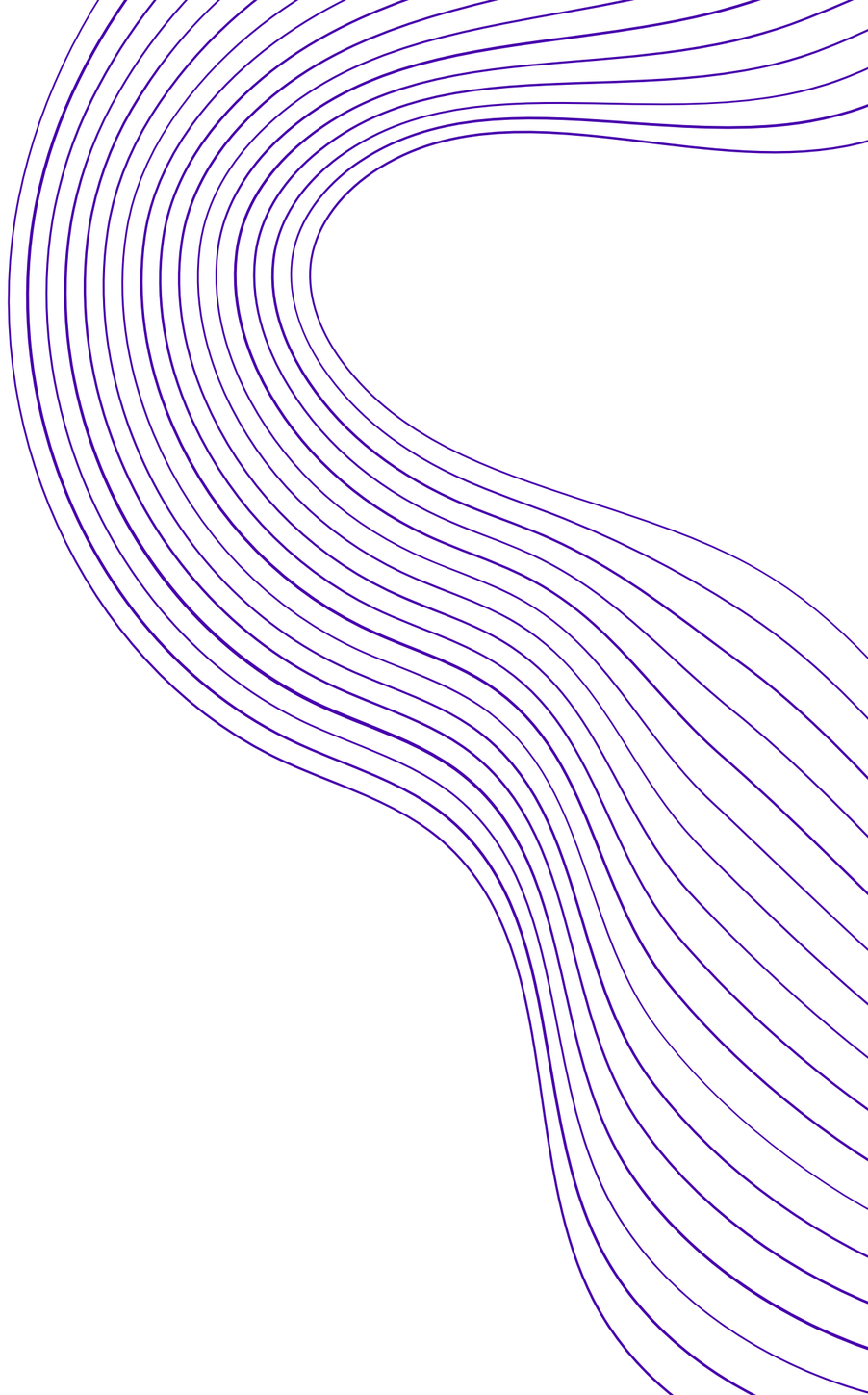
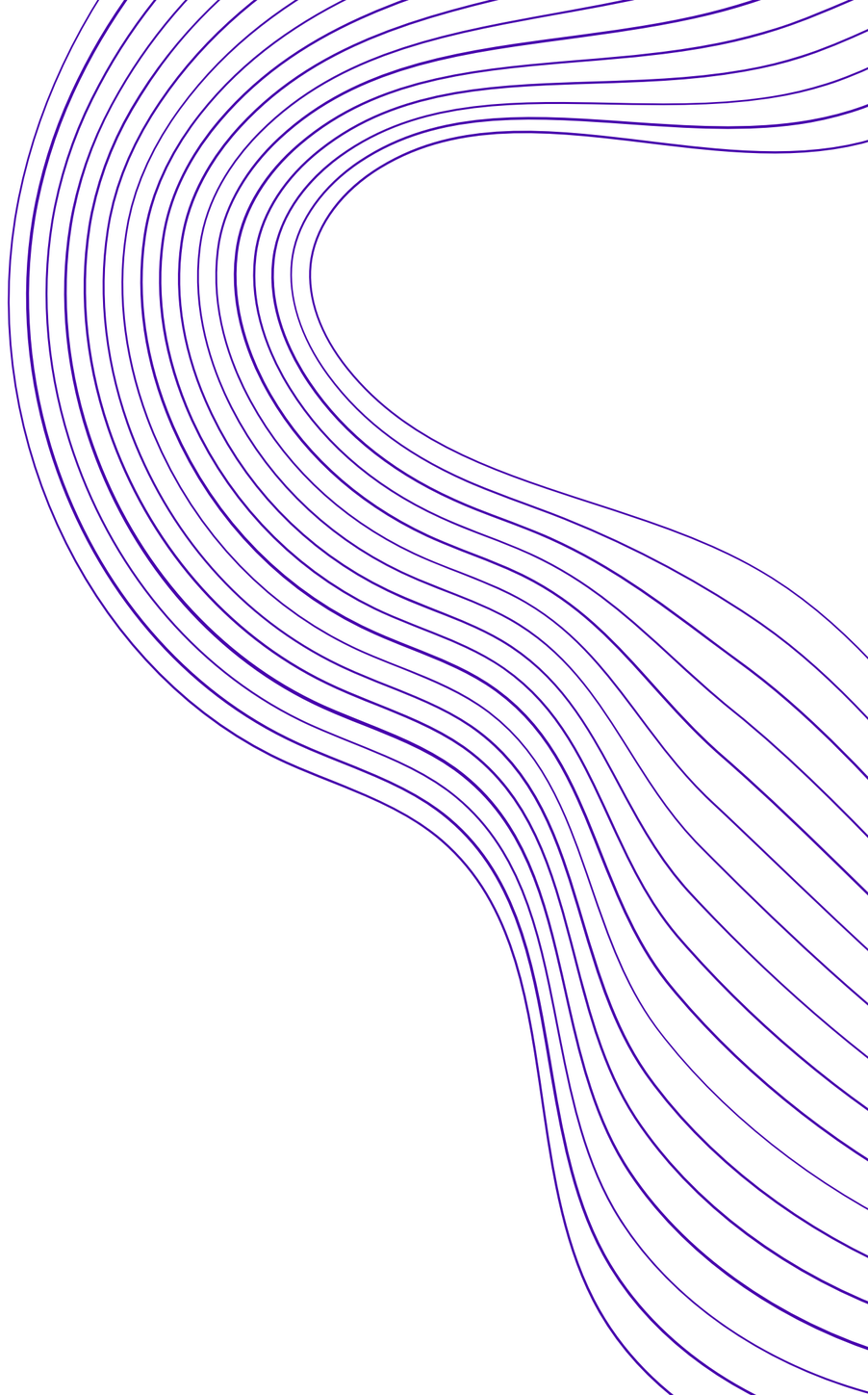
SISTEMAS DISTRIBUÍDOS | JUNHO 2021

Spark Streaming

GABRIEL BICALHO FERREIRA
LUCAS ANDRADE FREITAS
VINICIUS DE PAULA SILVA



Sumário

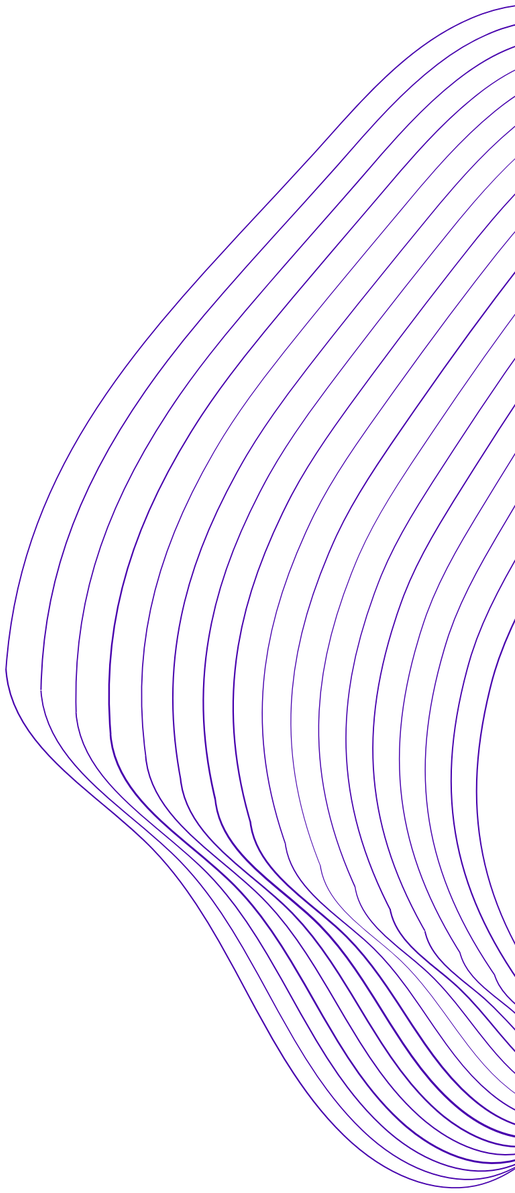


01 O QUE É?

O Spark Streaming é uma extensão que faz parte do núcleo da API Spark. Ele facilita a criação de fluxos de processamento tolerante a falhas sobre dados em streaming e em tempo real.

02 PORQUE UTILIZAR O SPARK STREAMING?

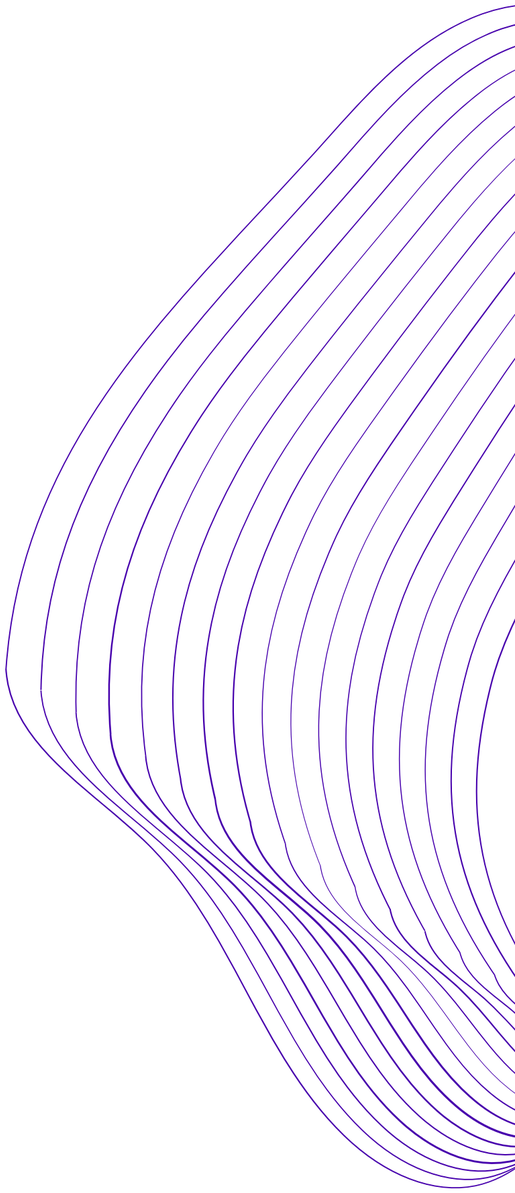
- Oferece rápida recuperação de falhas e atrasos na rede.
- Oferece um rico ecossistema, e pode ser integrado com bibliotecas de aprendizado de máquina, processamento gráfico e SQL.
- Os dados recebidos podem ser convertidos e adicionados a base de dados reais, ou exibidos em dashboards interativos.
- O Spark pode executar tarefas de processamento em lote de 10 a 100 vezes mais rápido que o Hadoop MapReduce, utilizando mais a memória do que disco.



03 QUANDO NÃO UTILIZAR?

- Trabalhos com processamento em batch e que não haja necessidade de dividir este processamento em mini-batches;

- Aplicativos com necessidades de latência abaixo de algumas centenas de milissegundos.

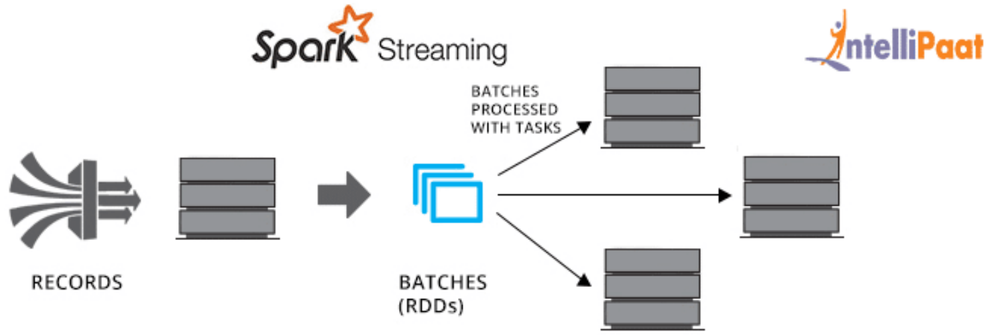


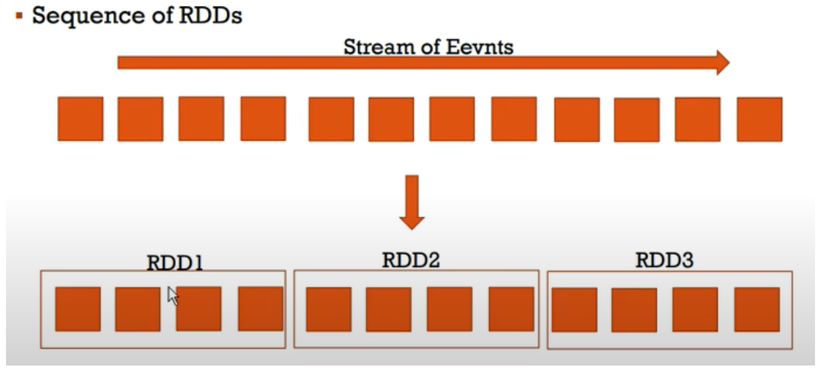
04 ONDE ELE É UTILIZADO?

Taxa de utilização no mercado ém media: 0,9%

05 ARQUITETURA INTERNA

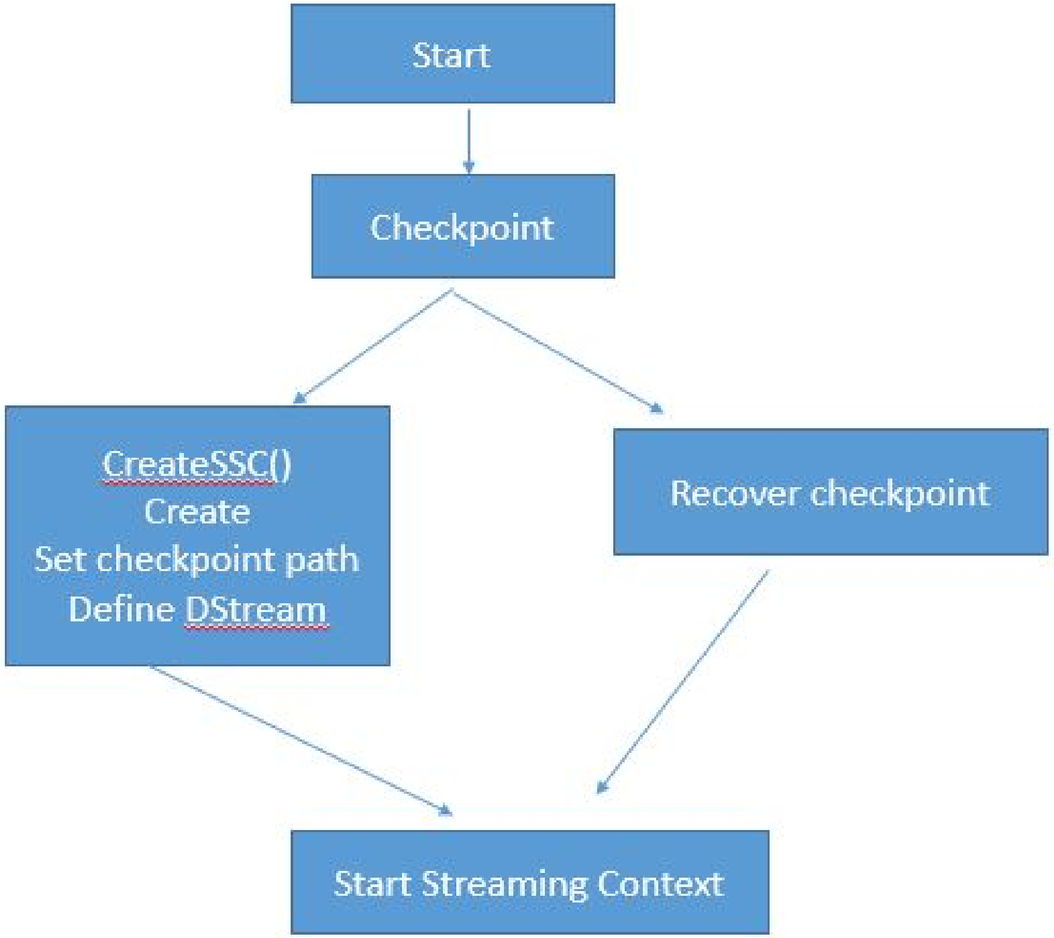
- Todos os dados chegam até o spark streaming em DStreams.
- Uma DStream é dividida em RDDs (Resilient Distributed Dataset), dentros dos nós Workers ocorrem as tasks enviadas pelo master.
- Podemos aplicar 2 tipos de operações dentro dos workers: Transformações e Ações.





06 RECUPERAÇÃO DE FALHAS

- Lazy Evaluation.
- Checkpoins.
- Persistência de RDDs.
- Write Ahead Logs



07 COMO UTILIZAMOS?

- Utilizamos o google dataproc para instanciar no nosso cluster spark. 1 Mestre: n1-standard-2 (4vCPU, 15GB de memória), 2~5 Workers: n1-standard-2 (2vCPU, 7.5GB de memória)

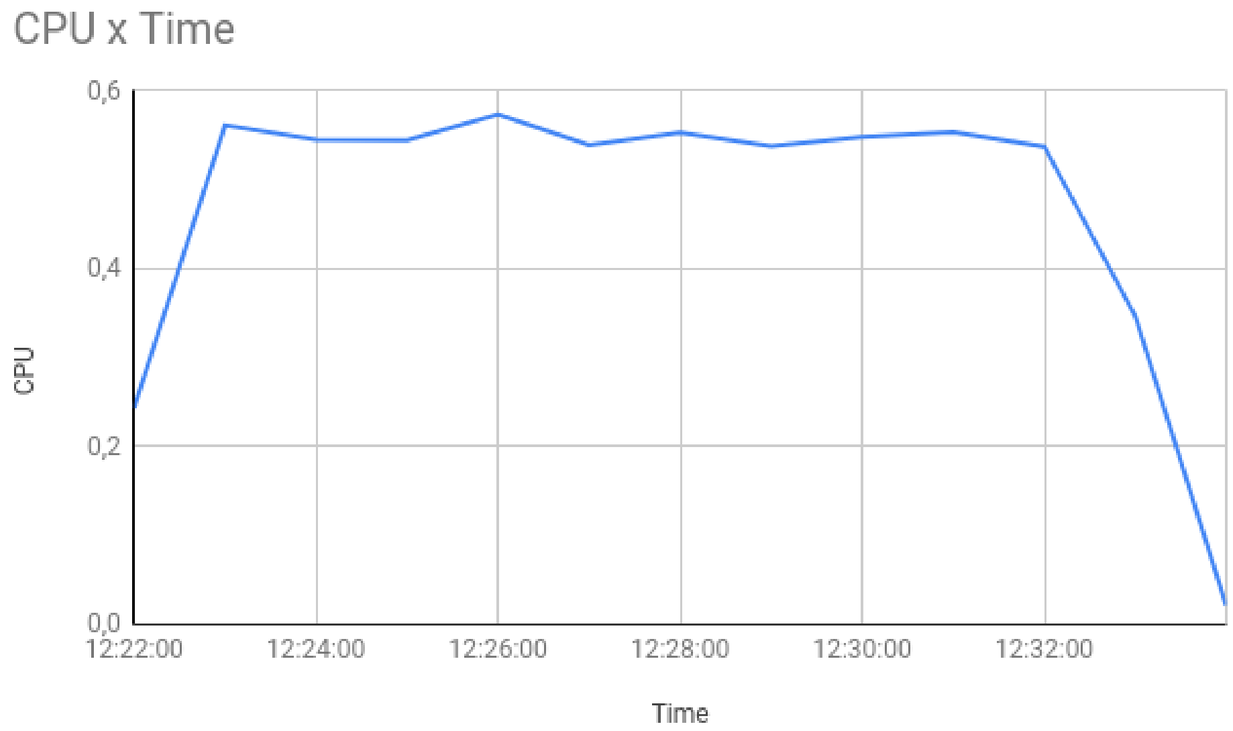
- Foi criado um container servidor para envio de dados em streaming através de um soquete TCP e um container frontend para recebimento e exibição dos dados já tratados pelo Spark.

- Nosso servidor envia frases com nomes aleatórios gerados a partir de um banco, separadas por uma quebra de linha ("\n")

- A aplicação Spark trata os dados separando cada palavra e contando o número de vezes que elas apareceram no texto, armazenando em um banco compartilhado pelos nós do cluster, e enviando a cada iteração, para o frontend.

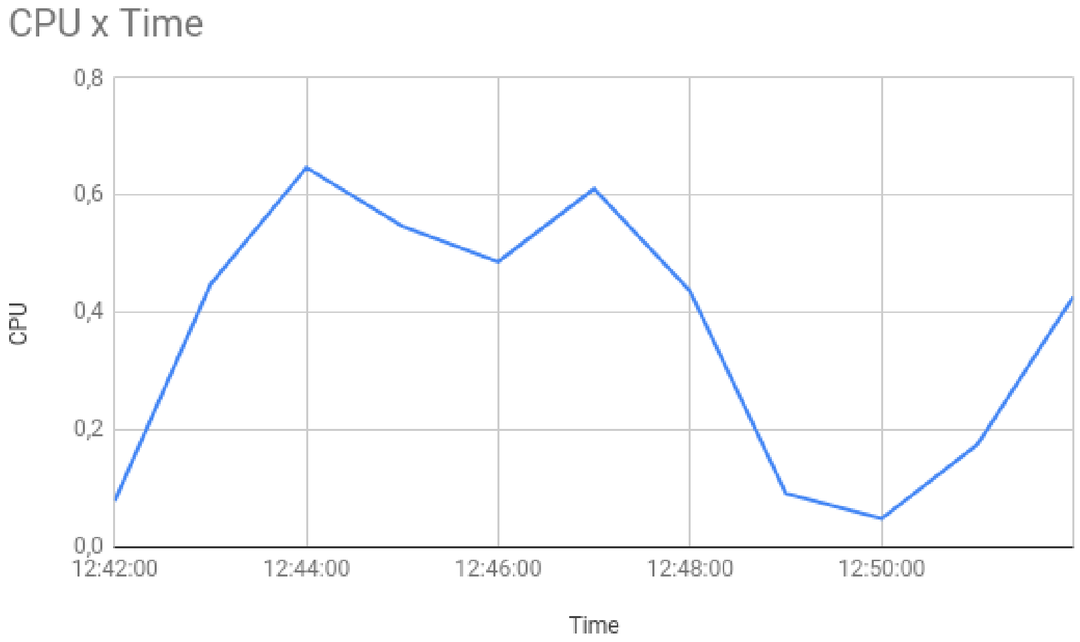
08 RESULTADOS

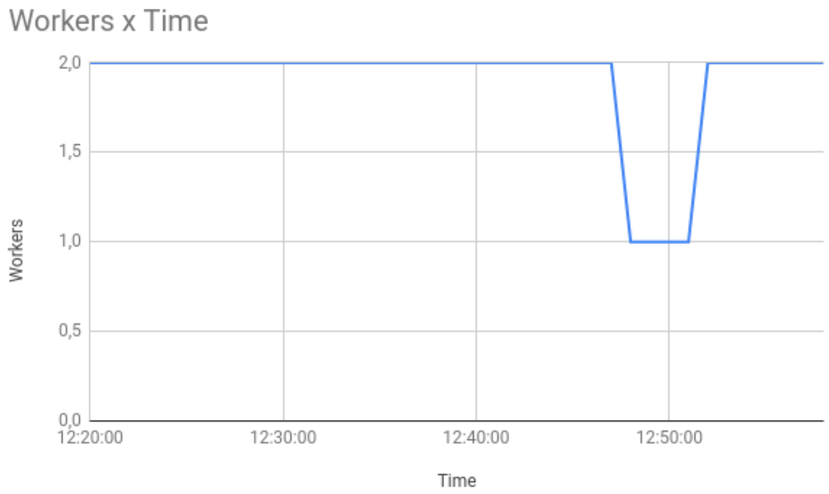
- 1 Mestre (4vCPU, 15GB de memória)
- 2 Workers (2vCPU, 7.5GB de memória)
- Processando todas as palavras por 10 minutos.
- Intervalo de lote processado a cada 1 segundo.
- Cerca de 87.758.980 palavras processadas.



08 RESULTADOS

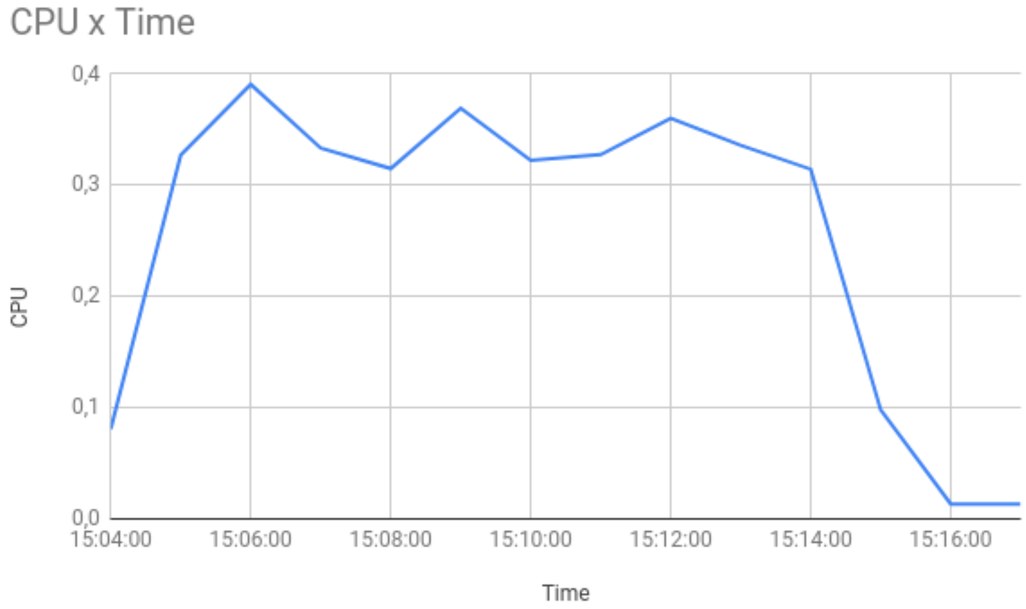
- 1 Mestre (4vCPU, 15GB de memória)
- 2 Workers (2vCPU, 7.5GB de memória)
- Processando todas as palavras por 10 minutos.
- Intervalo de lote processado a cada 1 segundo.
- Cerca de 50.187.060 palavras processadas.
- Paramos um dos Workers aos 12:48 e voltamos 12:51. O cluster voltou a operar normalmente.

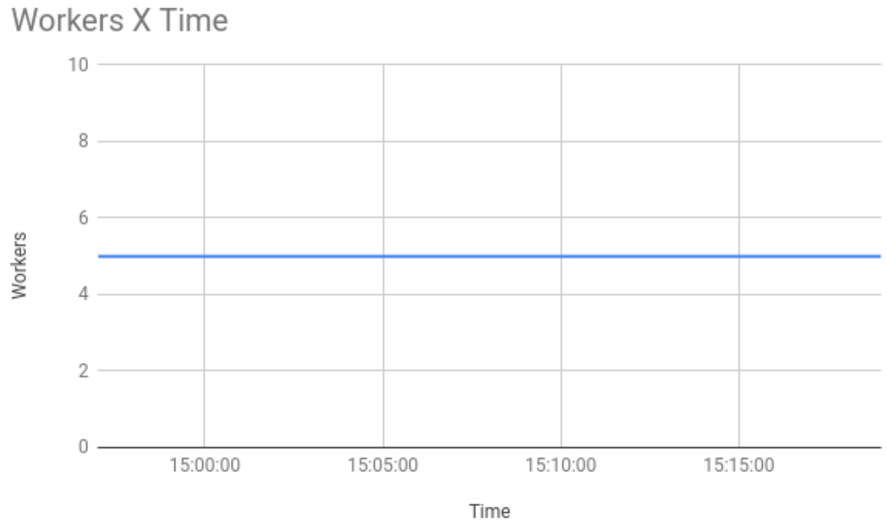




08 RESULTADOS

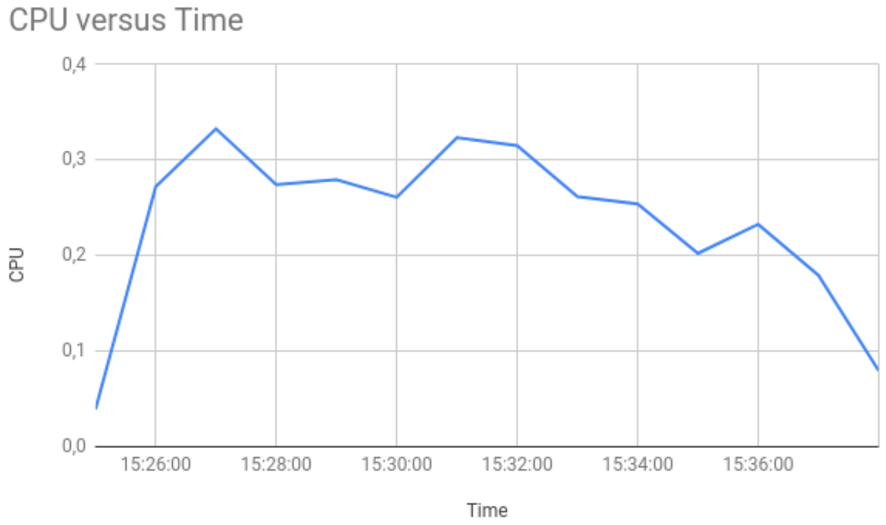
- 1 Mestre (4vCPU, 15GB de memória)
- 5 Workers (2vCPU, 7.5GB de memória)
- Processando todas as palavras por 10 minutos.
- Intervalo de lote processado a cada 1 segundo.
- Cerca de 98.765.120 palavras processadas.

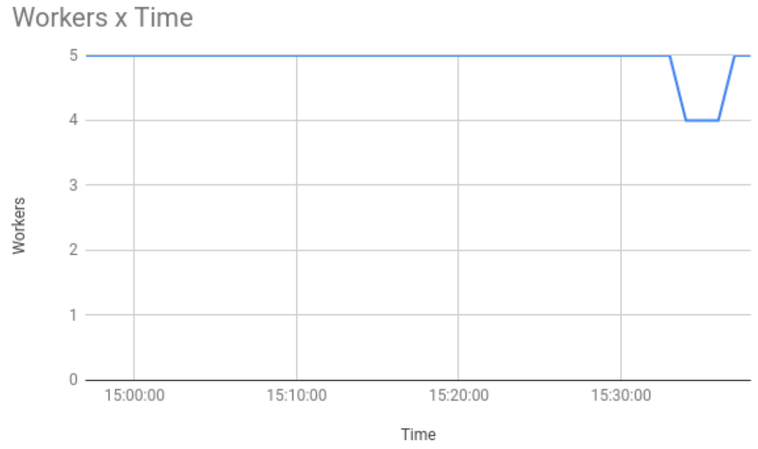




08 RESULTADOS

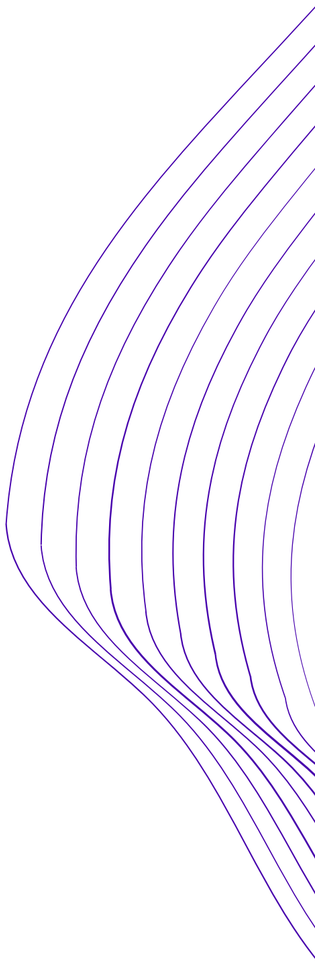
- 1 Mestre (4vCPU, 15GB de memória)
- 5 Workers (2vCPU, 7.5GB de memória)
- Processando todas as palavras por 10 minutos.
- Intervalo de lote processado a cada 1 segundo.
- Cerca de 60.867.580 palavras processadas.
- Paramos um dos Workers aos 15:33 e voltamos 15:35. O cluster voltou a operar normalmente.





REFERÊNCIAS

https://spark.apache.org/docs/latest/streaming-programming-guide.html
https://www.infoq.com/br/articles/apache-spark-streaming/
https://databricks.com/glossary/what-is-spark-streaming
https://techvidvan.com/tutorials/spark-tutorial/
https://santodigital.com.br/processamento-de-dados-o-que-e-batch-e-stream/



OBRIGADO!!