HADOOP E SPARK

LABORATÓRIO

GRUPO

Gabryel Nicolas
221022570

Guilherme Westphall
211061805

Joel Soares
211039546

Lucas Martins
221022088

HADOOP

- Simulação de um cluster com Docker Compose
- Contabilização de palavras em arquivos grandes (Word Count)
- Testes de performance
- Testes de Falha

CLUSTER

Composto por 8 containêres:

Serviço	Função	Porta
`nn`	**NameNode** – Gerencia o sistema de arquivos distribuído (HDFS).	9870 / 9000
`dn1`, `dn2`	**DataNodes** – Armazenam blocos reais de dados do HDFS.	9864
`rm`	**ResourceManager** – Coordena e agenda tarefas do YARN.	8088 / 8032
`nm1`, `nm2`	**NodeManagers** – Executam containers de tarefas (Map/Reduce).	8042
`jhs`	**JobHistoryServer** – Armazena logs e histórico de execuções.	19888
`edge`	**Client Node** – Ponto de entrada para comandos `hdfs`, `yarn`, `mapred`.	—

WORD COUNT

Geração do arquivo e execução do job MapReduce

```
yes "lorem ipsum dolor sit amet" | head -n 300000000 > /tmp/big.txt
hdfs dfs -put /tmp/big.txt /jobs/big.txt
```

```
EX=$(ls /home/hadoop/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-*.jar | head -n1)
hadoop jar "$EX" wordcount /jobs/big.txt /jobs-out-baseline
```

```
hdfs dfs -cat /jobs-out-baseline/part-r-00000
```

TESTE DE PERFORMACE

Mudança de 5 características

Cenário	Job ID	Replicação	Blocksize	Memória NM	Tempo (s)	Throughput (MB/s)	CPU total (s)	GC total (s)	Memória física snapshot (GB)*
Baseline	job_1763235544715_0002	1	128 MB	2048 MB	598,616	**12,90**	≈ 1843,6	≈ 87,7	≈ 19,0
2 réplicas	job_1763236505643_0001	2	128 MB	2048 MB	396,503	**19,48**	≈ 1833,2	≈ 84,3	≈ 19,2
Blocksize = 64 MB	job_1763237020860_0001	1	64 MB	2048 MB	374,766	**20,61**	≈ 1853,1	≈ 93,7	≈ 19,1
Blocksize = 256 MB	job_1763237525385_0001	1	256 MB	2048 MB	382,375	**20,20**	n/d	n/d	n/d
NodeManager = 4 GB	job_1763242547198_0001	1	128 MB	4096 MB	373,000	**20,71**	≈ 1872,3	≈ 83,2	≈ 17,4

TESTE DE FALHA

Provocação controladada de condições adversas

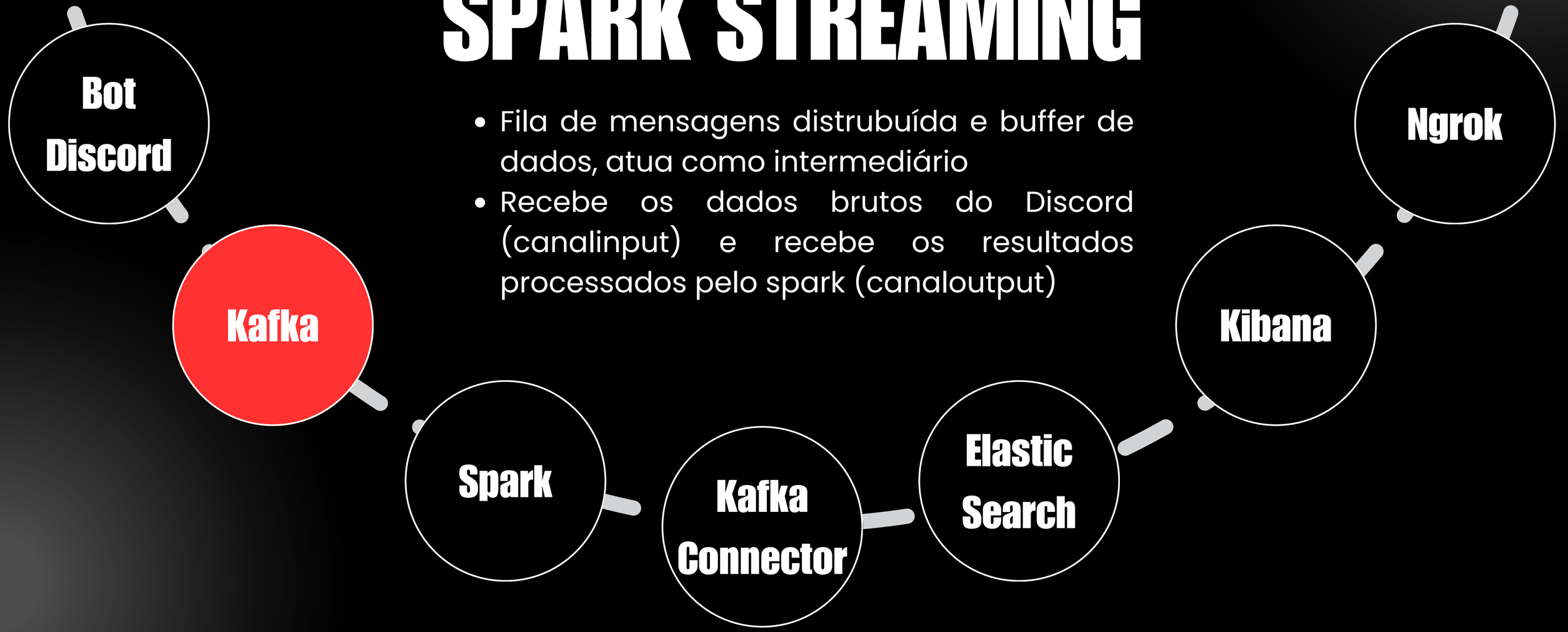
Etapa	Ação realizada	Estado do job / progresso	Comportamento observado
1	Job submetido normalmente	map 0% → 10%	Execução normal, avanço gradual sem erros aparentes.
2	Queda do **dn1** (um DataNode fora)	~map 10–20%	Job continua executando, porém mais lento; não ocorre falha imediata.
3	Queda do **dn2** (ambos DataNodes offline)	~map 23%	Execução congela; surgem erros `BlockMissingException` devido à ausência de réplicas disponíveis.
4	Subida de **dn1** e **dn2**	~map 23% → 30%+	Job imprime avisos de falha, mas **retoma** a execução a partir do ponto em que estava.
5	Queda do **NameNode**	progresso em map/reduce	Job congela imediatamente; são emitidos erros relacionados ao acesso aos metadados do HDFS.
6	Subida do **NameNode**	mesmo progresso inicial	Job é retomado; progresso avança normalmente após o retorno do serviço.
7	Finalização do job	100% map / 100% reduce	Apesar das falhas, a tarefa **finaliza com sucesso**, com várias tentativas marcadas como FAILED/RELAUNCHED.

SPARK STREAMING

- Produtor de dados de entrada
- Conecta-se à API do Discord, monitora mensagens e as escreve no tópico canalinput do Kafka

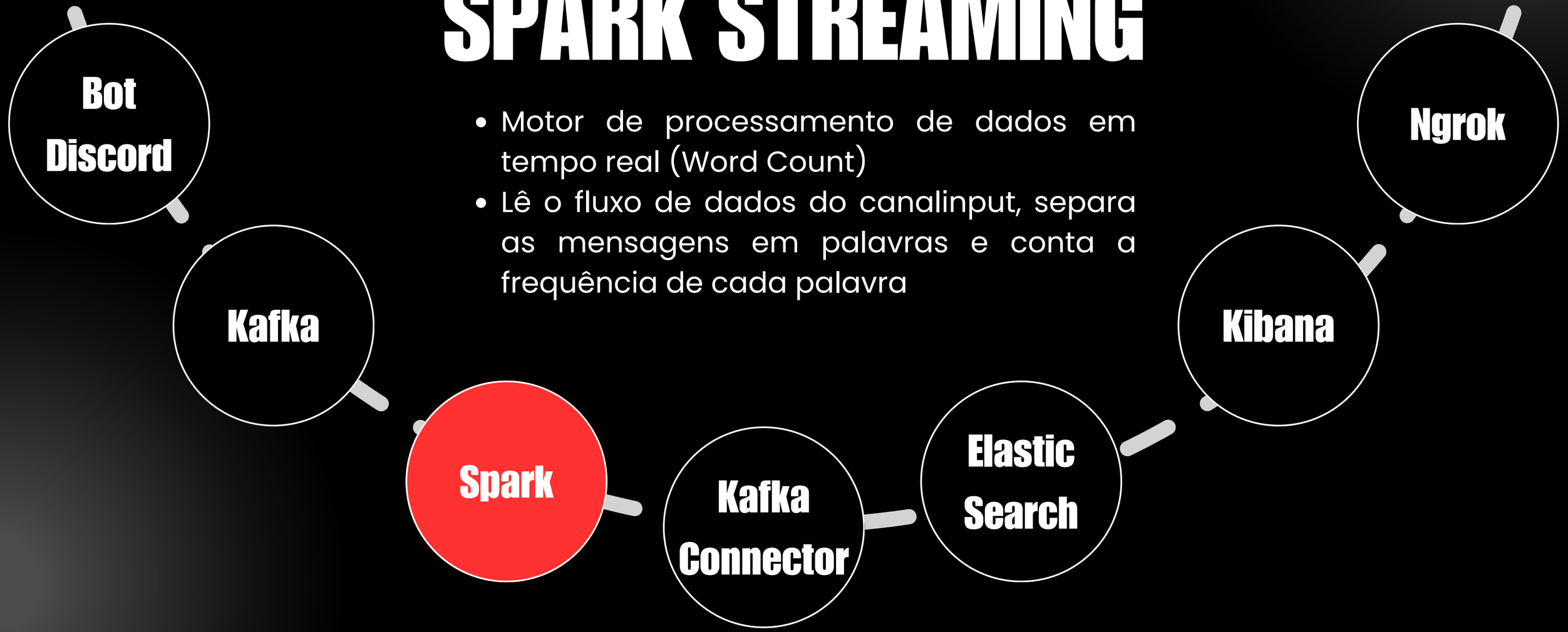
SPARK STREAMING

- Fila de mensagens distrubuída e buffer de dados, atua como intermediário
- Recebe os dados brutos do Discord (canalinput) e recebe os resultados processados pelo spark (canaloutput)



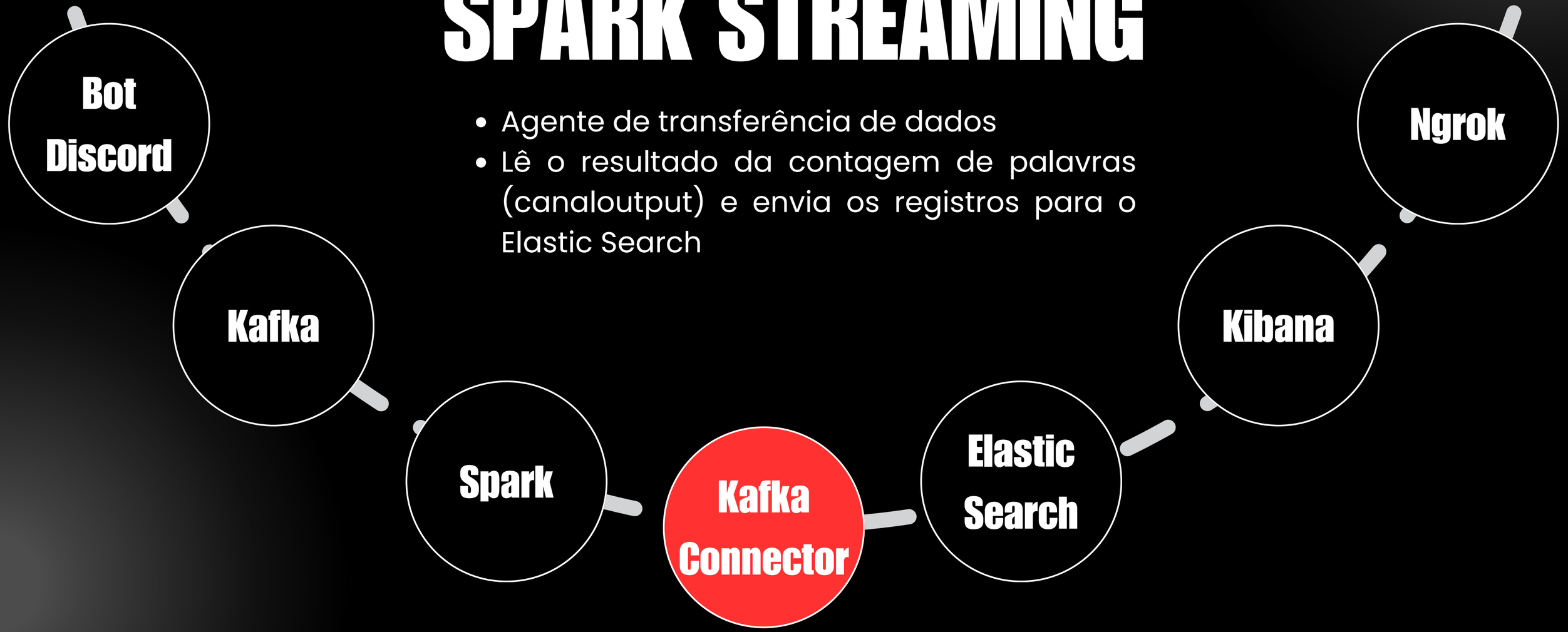
SPARK STREAMING

- Motor de processamento de dados em tempo real (Word Count)
- Lê o fluxo de dados do canalinput, separa as mensagens em palavras e conta a frequência de cada palavra



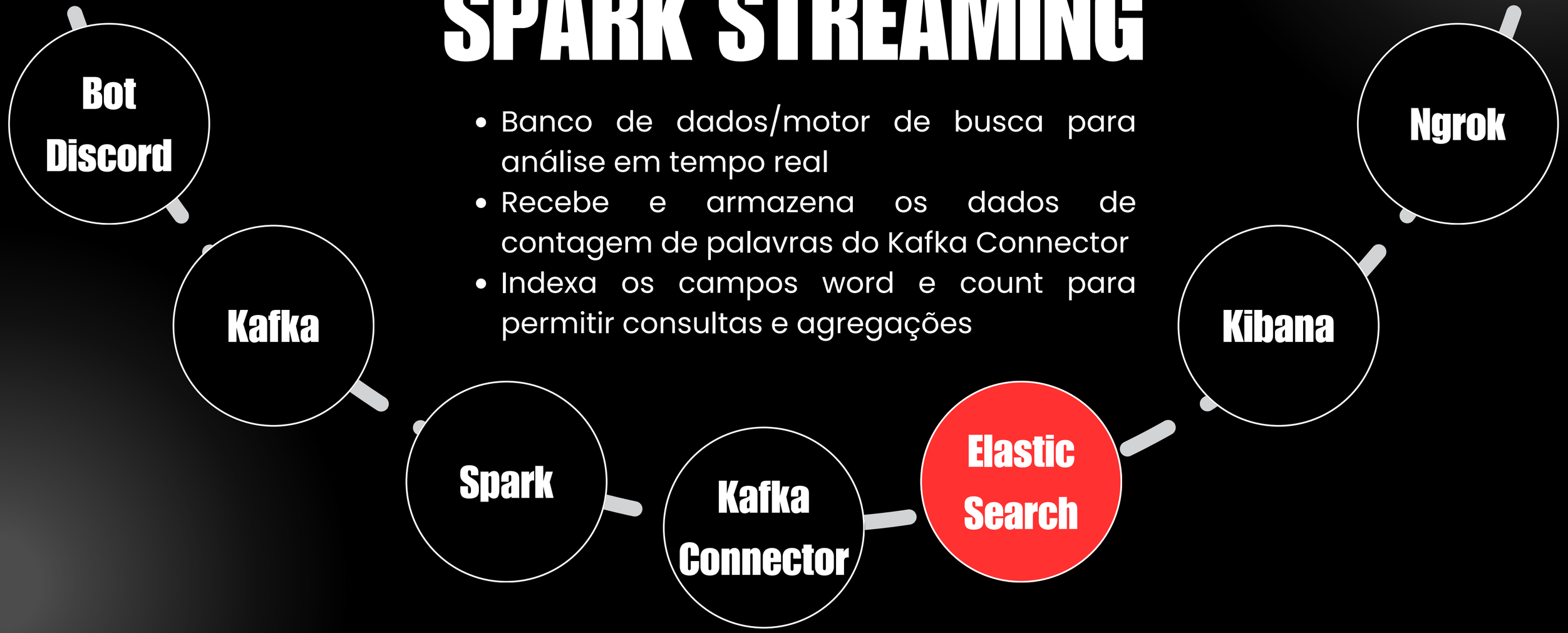
SPARK STREAMING

- Agente de transferência de dados
- Lê o resultado da contagem de palavras (canaloutput) e envia os registros para o Elastic Search



SPARK STREAMING

- Banco de dados/motor de busca para análise em tempo real
- Recebe e armazena os dados de contagem de palavras do Kafka Connector
- Indexa os campos word e count para permitir consultas e agregações



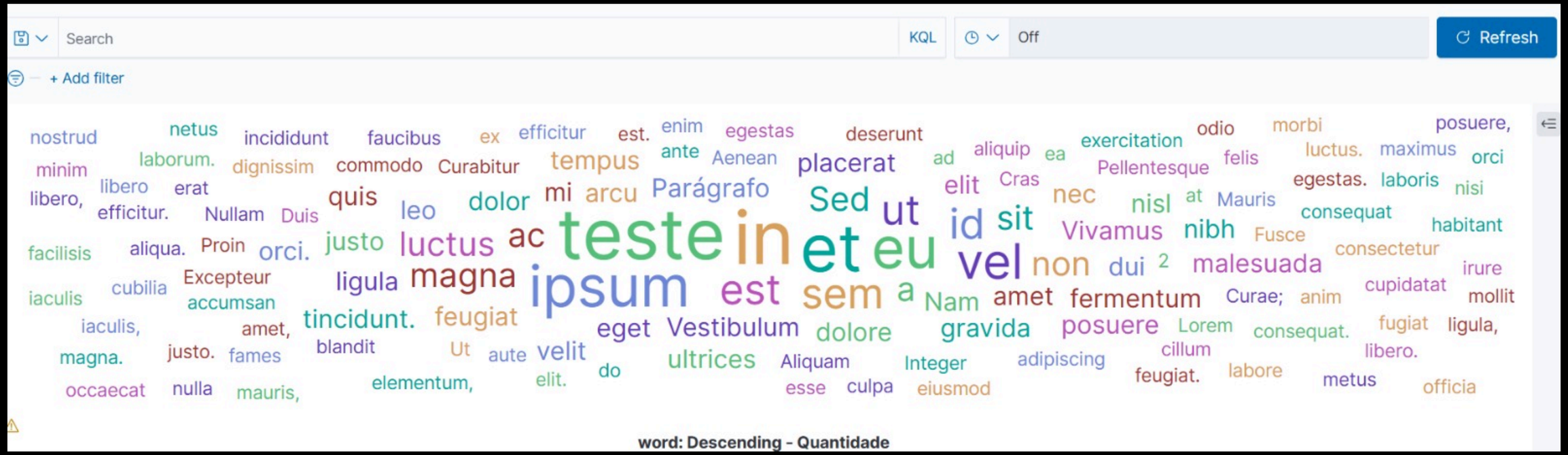
SPARK STREAMING

- Interface de visualização, dashboard
- Conecta-se ao ElasticSearch, utiliza o Index Pattern (canaloutput*) para entender a estrutura dos dados e renderiza as visualizações, como a Nuvem de Palavras

SPARK STREAMING

- Túnel de exposição de serviço local
- Cria um túnel seguro (URL pública) para a porta local onde o Kibana está rodando (porta 5601), permitindo acesso da interface web no navegador

NUVEM DE PALAVRAS



OBRIGADO