

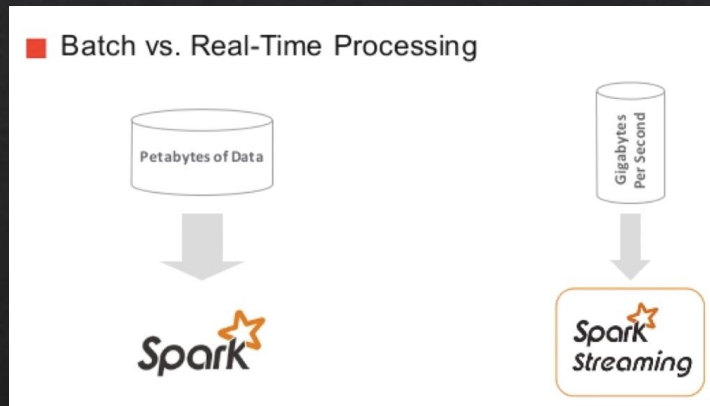
PARTE III

3. Spark Streaming e Análise de Dados em Tempo Real

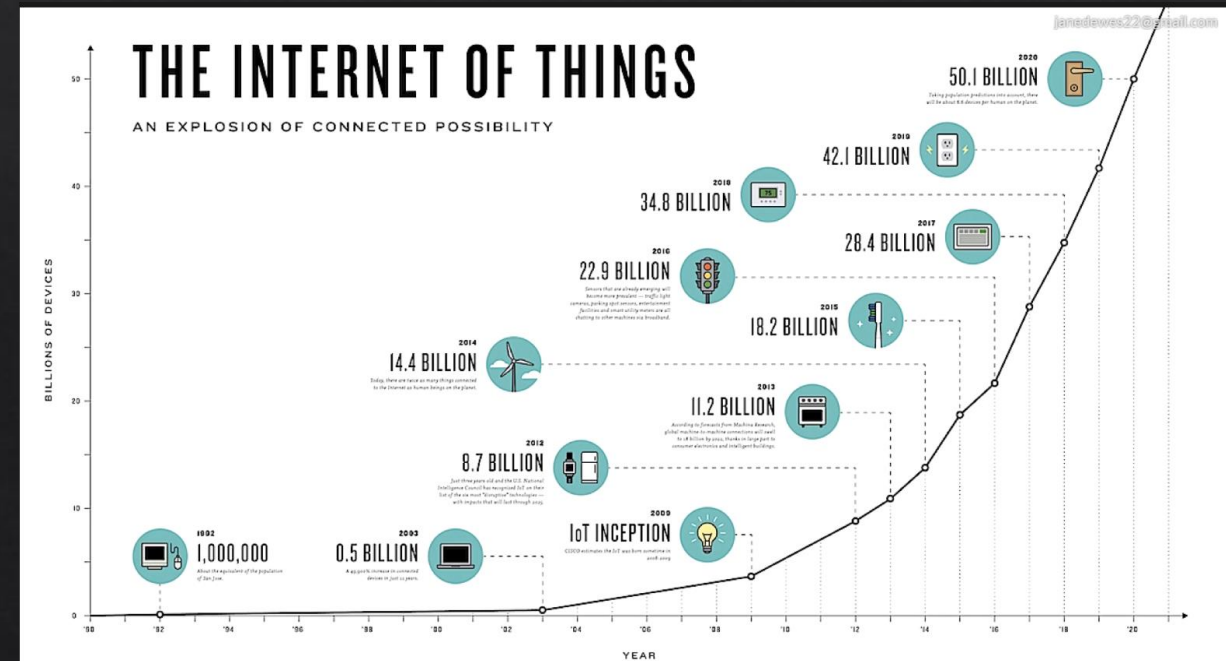


Spark Streaming e Análise de Dados em Tempo Real

- ◇ Análise em tempo real permite que os profissionais possam compreender o que os consumidores desejam, e incorporar os seus desejos em decisões críticas de negócio em tempo real.

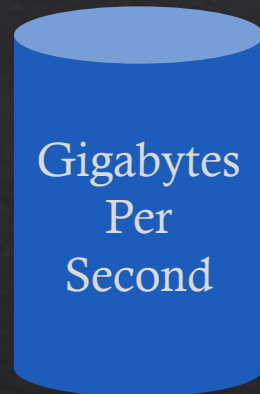
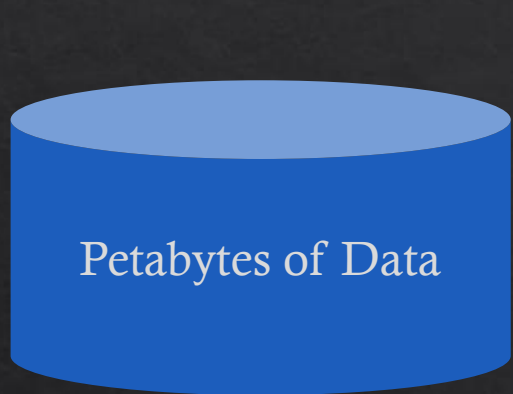


- Uma das principais fontes de dados contínuos: Sensores
- IoT irá impulsionar o mercado de soluções de Streaming de dados.



1. Spark Streaming

Batch vs. Real-Time Processing



Coleta e análise dos dados direto da fonte e à medida que são gerados

Transformação, sumarização e análise

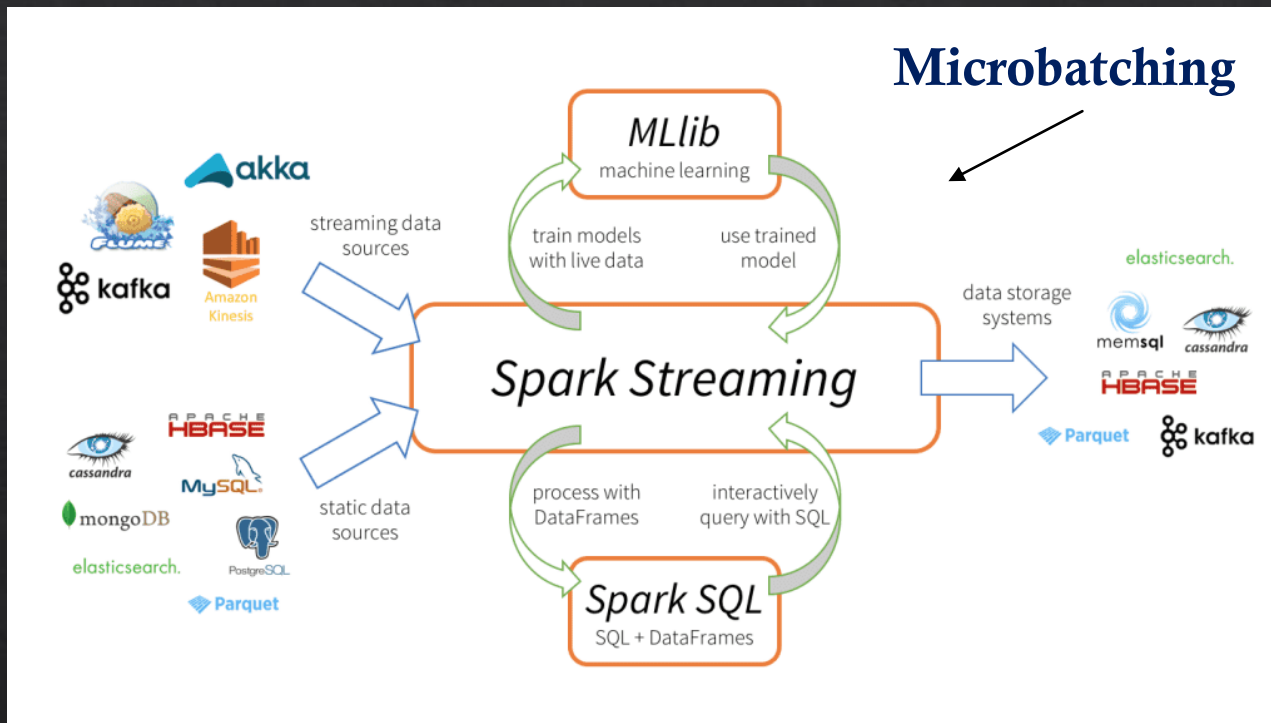
Machine Learning

Previsões em tempo real

- Análise exploratória de dados.
- Análise de dados de Data Warehouses.
- Treinar modelos de ML sobre grandes conj de dados.
- Outras tarefas analíticas feitas com Hadoop e MapReduce.

- Monitoramento de serviços.
- Processamento de eventos de cliques e eventos em web sites (Flat Files).
- Processamento de dados de sensors IoT.
- Processamentos de dados vindos de serviços como: Twitter, Kafka, Flume, AWS Kinesis.
- Banco Relacionais e NoSQL

1. Spark Streaming



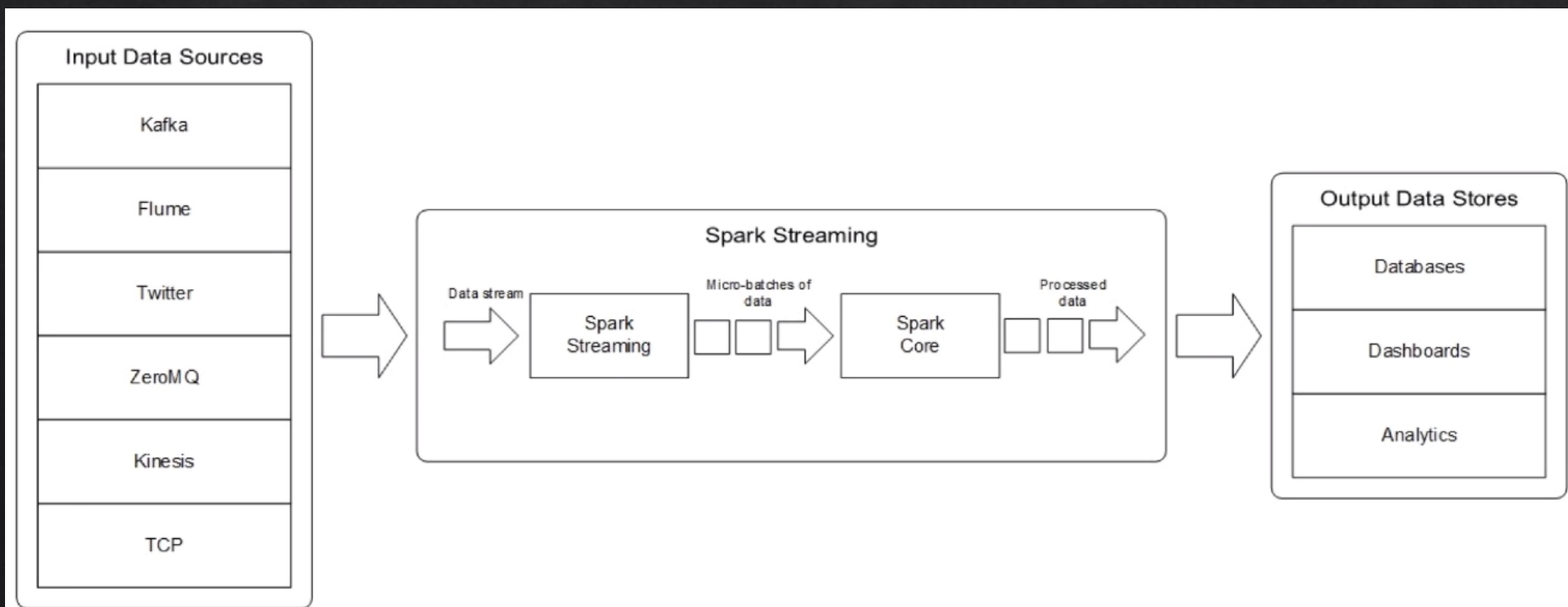
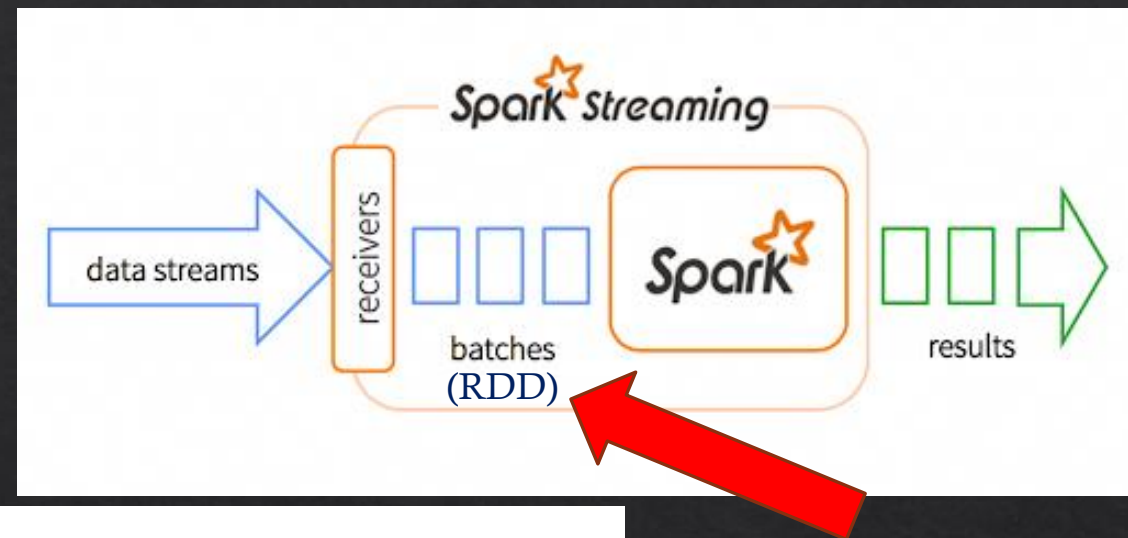
Areas que **Spark Streaming** está sendo usado:

1. **Streaming ETL**
2. **Detecção de Anomalias**
3. **Enriquecimento de Dados**
4. **Sessões Complexas e Aprendizado Contínuo**
5. Detecção de Fraudes em Tempo Real
6. Filtro de Spam
7. Detecção de Invasão de Redes
8. Análise de Mídias Sociais em Tempo Real
9. Análise de Stream de Cliques em Sites, gerando Sistemas de Recomendação
10. Recomendação de Anúncios em Tempo Real
11. Análise de Mercado de Ações

Vantagem: combinar processamento em batch e processamento streaming em um único sistema



Arquitetura Apache Spark Streaming



Outros frameworks:

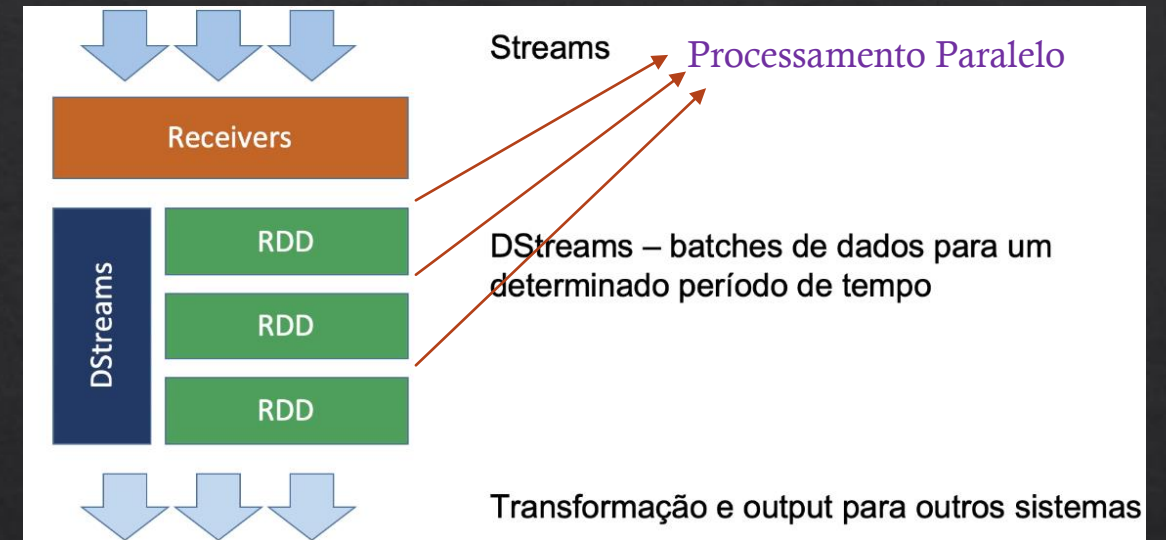
- Apache Samza
- Apache Storm
- Apache Flink
- Apache Spark Streaming
- AWS Kinesis (tem custo)

2. DStreams (Discretized Streams)

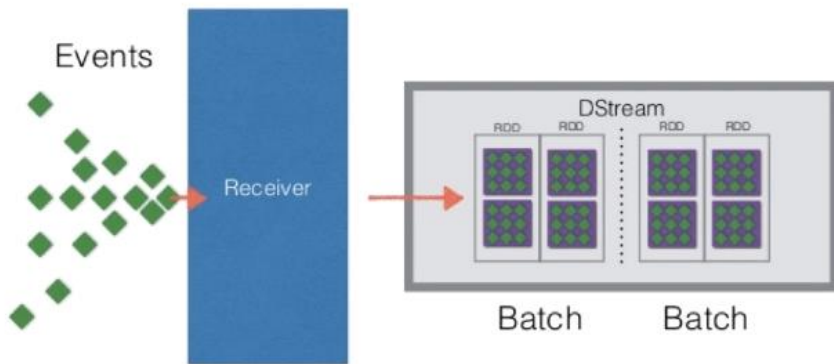
O Spark Streaming é um módulo do Apache Spark para processamento de dados em tempo real:

- Os **RDD's** são a base do **Apache Spark**.
- Os **DStreaming's** são a base do **Spark Streaming**

O DStream permite converter um conjunto de dados contínuo em um conjunto discreto RDD's .



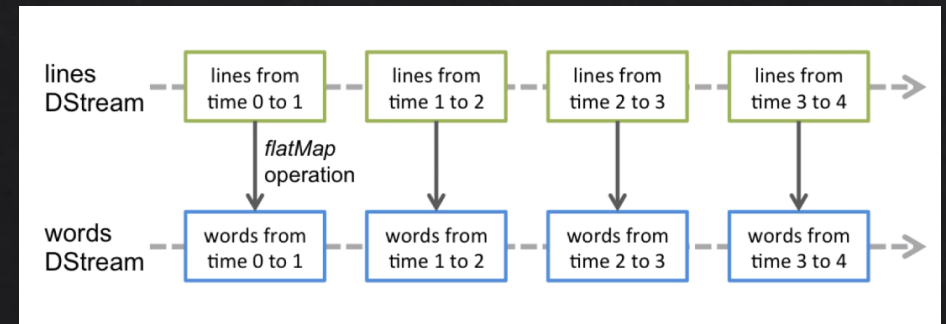
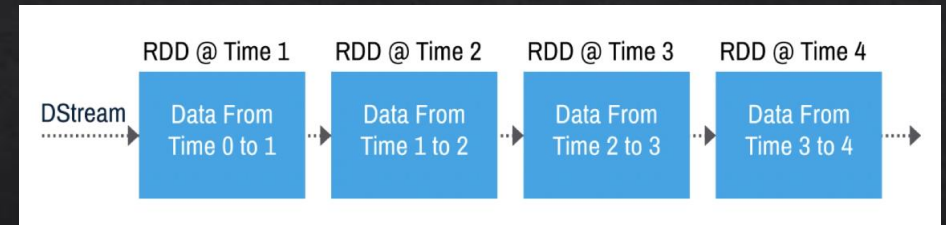
DStreams: Basic unit of Spark Streaming



The DStream is (Discretized) into batches, the timing of which is set in the Spark Streaming Context. Each Batch is made up of RDDs.

Operações aplicadas
ao Dstreaming:

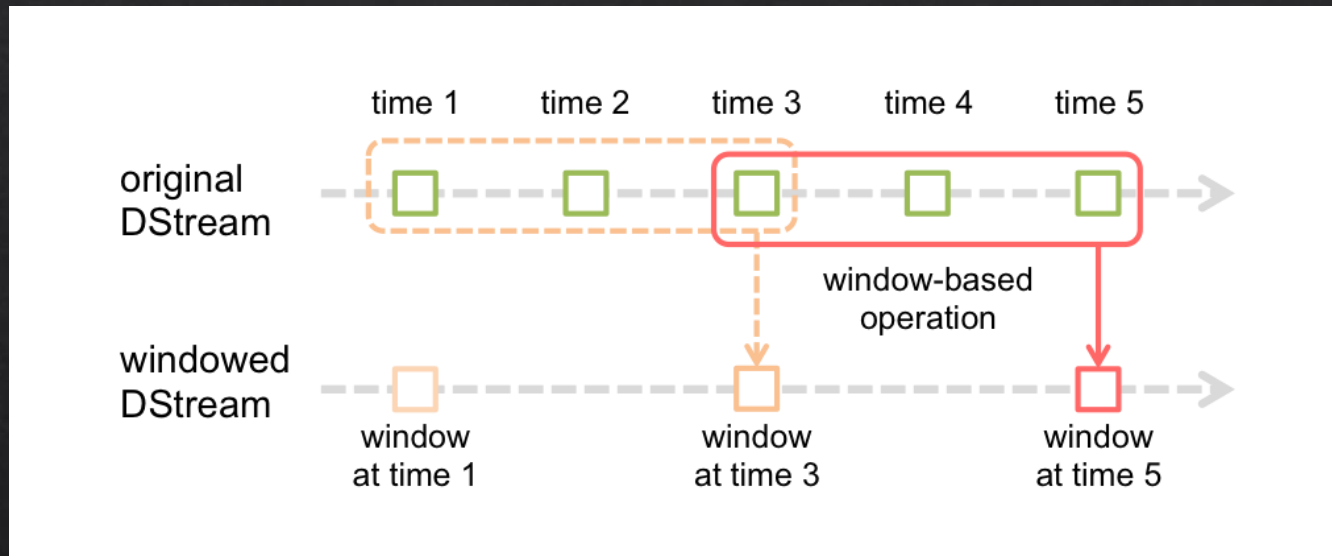
- Map
- FlatMap
- Filter
- reduceByKey
- Join
- Window



3. Windowing

Exemplo:
Batch interval = 1 Segundo
Window length = 1 hora

- Computing Windowing: computação em uma janela de tempo, esse recurso é usado para aplicar operações de transformações sobre os dados em uma janela específica.
- Windowing permite computer os resultados ao longo de períodos de tempo maiores que o batch interval.



Window length [tamanho] – duração da window (3 unid. de tempo nesse caso)

Sliding interval [intervalo]– intervalo entre windows

`ssc = StreamingContext(sc, INTERVALO_BATCH)`

`window(windowDuration: Duration, slideDuration: Duration): DStream[T]`

3 tipos de intervalos que podem ser configurados:

Freq com que os dados são capturados em um DStream

Batch Interval

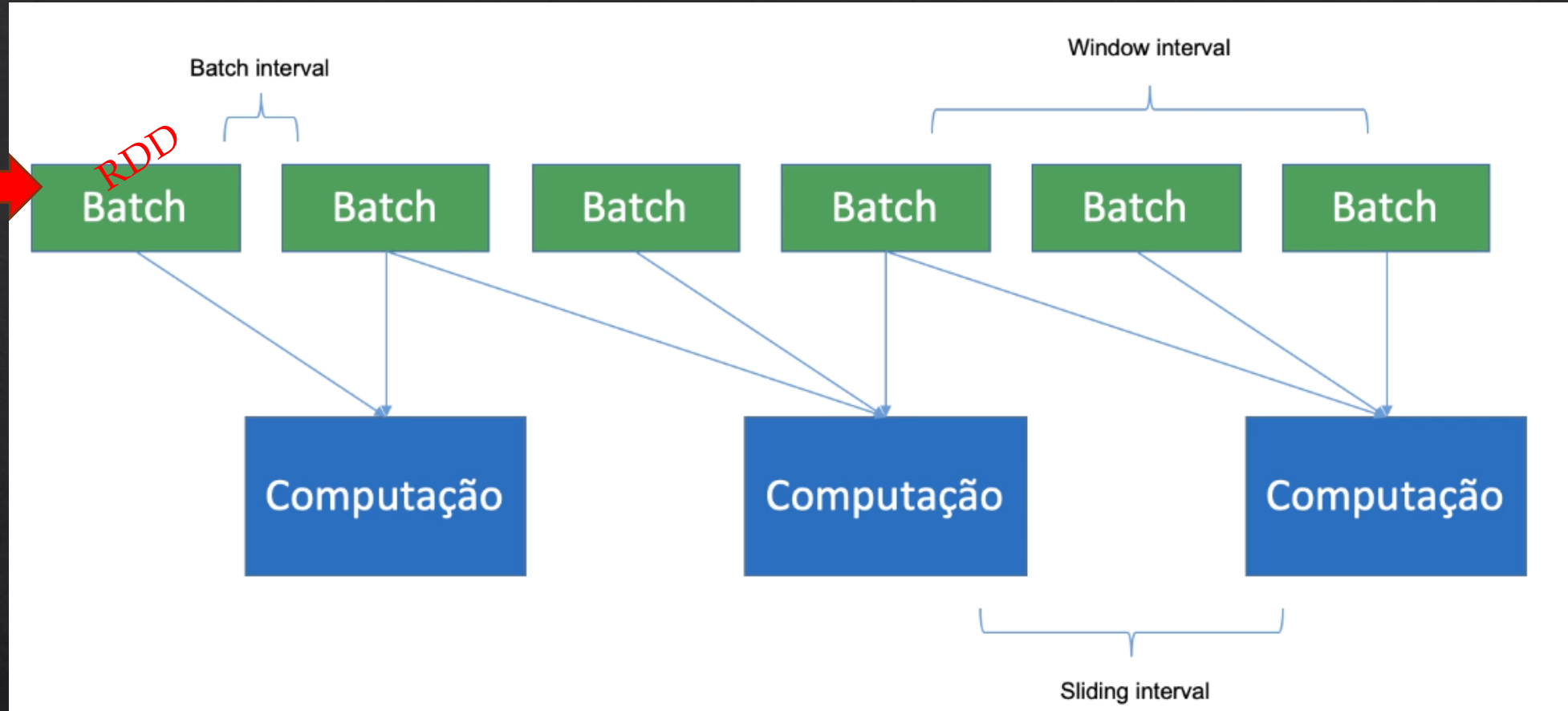
Freq com que uma window é aplicada

Sliding Interval

Intervalo de tempo capturado para computação e geração de resultados

Window Interval

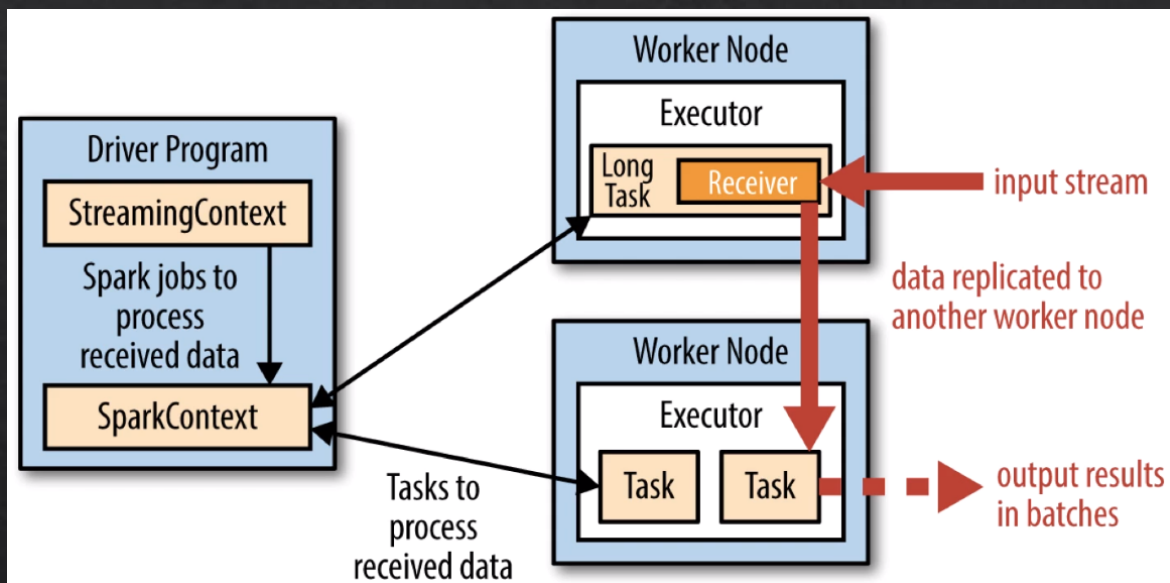
Gerados continuamente



Existem diversas funções de transformação específicas para se trabalhar com Window
<https://spark.apache.org/docs/latest/streaming-programming-guide.html>

Tolerância a Falhas

Processo do Spark Streaming



- Todos os dados são replicados para no mínimo 2 workes Nodes.
- Um dir de checkpoint pode ser usado para armazenar o estado do streaming de dados, no caso em que é necessário reiniciar o streaming. “`ssc.checkpoint()`”

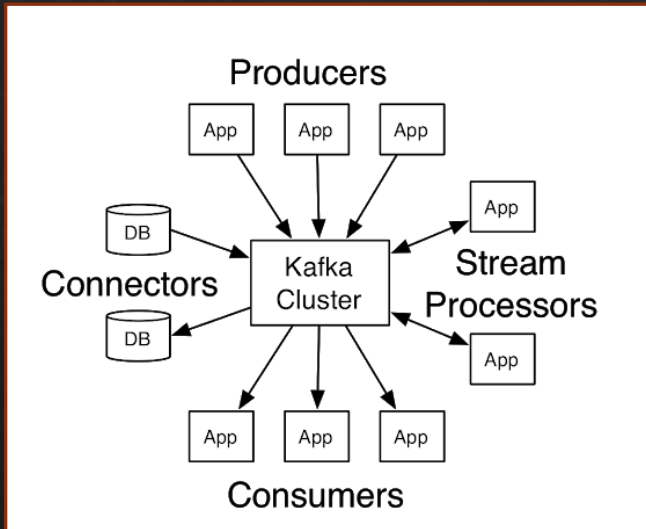
Falha no Receiver

Falha no Driver Context (script)

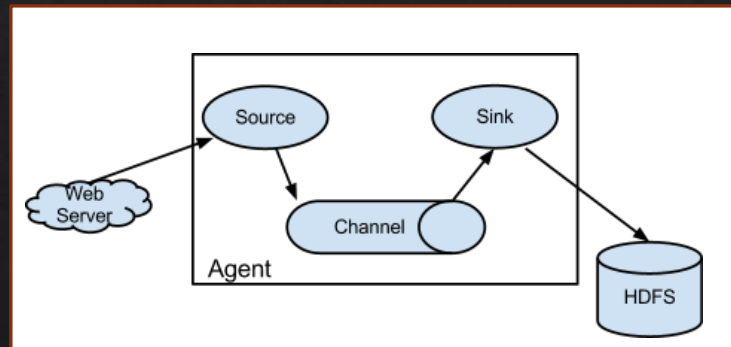
4. Integração com outros sistemas – Kafka, Flume, kinesis (coleta de dados)



É um Sistema para gerenciamento de fluxos de dados em tempo real, gerados a partir de web sites, app, sensores, etc.



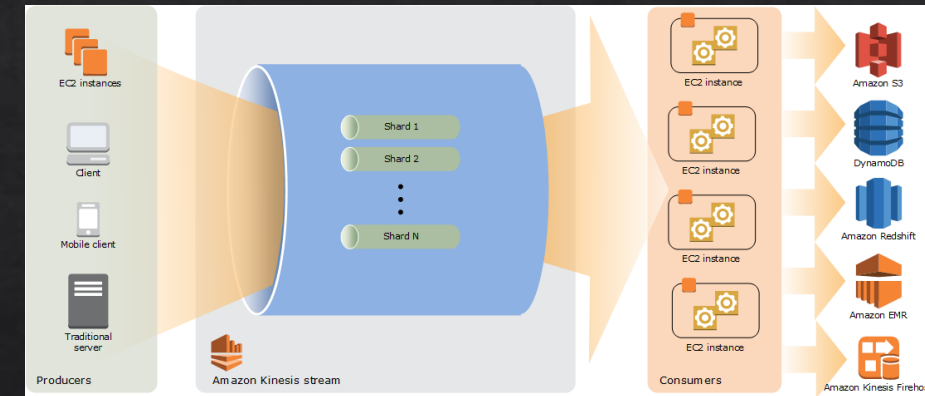
É mais específico para coleta de logs de servidores.



Push-Based Flume x Pull Based Flume



Tem custo associado, pois já possui Infraestrutura de servidores.



5. Processamento de Linguagem Natural - NLTK

- ♦ PLN: interação entre computador e linguagem humana, que deve ser precisa, não ambígua e altamente estruturada, o PLN é baseado em ML.
- ♦ Aplicações que usam PLN:
 - Corretores Ortográficos
 - Engines de Reconhecimento de Voz (Siri, Google Assistance, Cortana)
 - Classificador de Spam
 - Mecanismos de Busca (Google, Bing)
 - IBM Watson
- ♦ O PLN é abordado do ponto de vista da análise dos conhecimentos:

Análise Morfológica	Análise Sintática
Análise Semântica	Análise Pragmática



Análise de Sentimentos é o uso de:

- PLN
- Análise de textos
- Mineração de dados
- Computação linguística

para identificar e extrair informação subjetiva em elementos de texto.

Principais Frameworks

- ◇ GATE (General Architecture for Text Engineering)
- ◇ Mallet (Machine Learning for Language Toolkit)
- ◇ OpenNLP
- ◇ UIMA
- ◇ Gensim
- ◇ SpaCy
- ◇ Natural Language Toolkit (**NLTK**)

