Evento	Salão UFRGS 2014: SIC - XXVI SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2014
Local	Porto Alegre
Título	Um Estudo da Performance de Sistemas Distribuídos para o Processamento de Streams
Autor	OTÁVIO MORAES DE CARVALHO
Orientador	PHILIPPE OLIVIER ALEXANDRE NAVAUX

Podemos descrever as massivas quantidades de dados que são injetadas na internet, todos os dias, citando alguns números: A cada minuto, 3125 novas fotos são adicionadas no *Flickr*, 34722 novos *likes* são expressados no *Facebook*, e mais de 100000 novos *tweets* são criados no *Twitter*. Este crescimento é consequência de diversos fatores, incluindo a pervasividade das redes sociais, do sucesso do mercado de smartphones, bem como da disseminação das redes de sensores, que deram origem à chamada *"Internet of Things"*. O fenômeno decorrente dessas mudanças, conhecido pelo nome popular de *"Big Data"*, pressionou o desenvolvimento de ferramentas que explorassem o processamento desses grandes volumes de dados.

Tradicionalmente, o processamento de tarefas intensivas em dados envolve o processamento em lotes de grandes conjuntos de dados estáticos, utilizando conjuntos de máquinas interconectadas via rede. Visando processar esse tipo de dados, sistemas de gerenciamento de banco de dados *(DBMSs)* e *frameworks* de processamento distribuído, como o por exemplo *MapReduce*, se tornaram populares. Ao longo do tempo, foi demonstrado que estes modelos, focados no processamento de grandes lotes de dados, não são capazes de prover baixas latências de processamento. Uma vez que uma nova gama de aplicações focou-se ao processamento de dados em tempo-real ou quase tempo-real, novas ferramentas são necessárias para realizá-lo.

Visando atender à demanda de processamento de grandes volumes de dados, em baixas latências, novas ferramentas foram desenvolvidas. Estas ferramentas são o resultado da combinação das aplicações clássicas de *Stream Processing* e das novas idéias oriundas do modelo *MapReduce*. Dentre essas ferramentas de *Distributed Stream Processing* (Processamento Distribuído de Streams), destacam-se:

- *Twitter Storm*, uma *framework* para desenvolvimento de aplicações de processamento de fluxos de dados, de forma distribuída, com aplicações descritas por *DAGs* (grafos direcionais acíclicos). Suas aplicações são compostas por *Spolts*, que fornecem os eventos para a aplicação; E por *Bolts*, que implementam modificações sobre as tuplas a elas passadas. O *Storm* possui um sistema de tolerância à falhas e pode fornecer características transacionais, quando necessárias.

- *Spark Streaming*, uma *framework* para desenvolvimento de aplicações de processamento de fluxos de dados distribuídos, baseada no *Apache Spark*, que trabalha compondo a sua execução em múltiplos *micro-batchs*, denominados *RDDs* (Conjuntos de dados distribuídos e resilientes), buscando prover processamento de grandes volumes de dados com baixas latências e reutilizando os mecanismos de tolerância à falhas do próprio *Spark*.

Neste trabalho, visamos analisar a performance dessas aplicações através de *micro-benchmarks*, que avaliam as características necessárias pelas aplicações típicas que utilizam este tipo de plataforma. Através dessa análise, poderemos ter melhor compreensão dos limites alcançáveis pelas plataformas atuais, como, por exemplo, o *throughput* máximo de dados, o grau de escalabilidade e as características específicas de cada uma delas.

Sendo esta uma área de pesquisa emergente, o objetivo principal é ajudar a esclarecer quais são os limites existentes para esse tipo de aplicação. Uma vez determinados estes limites, poderemos avançar, propondo soluções para os seus problemas específicos.