

Proposta para Dissertação de Mestrado em Engenharia Informática

Título:

Comparação do Desempenho de Plataformas de Código Aberto para Processamento de Big Data Streams: Spark, Flink, Storm, Heron e Samza

Orientador:

Mário Freire (email: mario@di.ubi.pt; página web: <http://www.di.ubi.pt/~mario/>)

Sumário

De acordo com um relatório recente da IBM Marketing Cloud, 90% dos dados foram criados apenas nos últimos dois anos, tendo sido criados 2,5 quintiliões de bytes de dados todos os dias e com o aparecimento de novos dispositivos, sensores e tecnologias, a taxa de crescimento dos dados irá provavelmente aumentar [1]. Para além disso, a utilização de aplicações para dispositivos móveis continua a aumentar. Isto significa que os sistemas para processamento de big data estão a tornar-se mais complexos e a colocar novos desafios, tornando-se necessário processar dados em tempo real, à medida que chegam, de modo a permitir tomar decisões rápidas. Em consequência, o processamento distribuído de streams está a tornar-se muito popular no contexto das plataformas para processamento de big data. Recentemente foram propostas várias plataformas de código aberto para processamento de streams, de entre as quais se destaca Apache Spark, Apache Flink, Apache Kafka, Apache Storm, Apache Samza, Apache Flume, Apache Apex e Heron, tendo também surgido plataformas para processamento de streams disponibilizadas pelos maiores fornecedores de serviços cloud, por exemplo, Amazon Kinesis e Google MillWheel. Recentemente, foram publicados alguns estudos sobre o desempenho das plataformas Spark, Flink e Storm [2]-[5]. Nesta dissertação pretende-se elaborar um estudo comparativo do desempenho das principais plataformas de processamento de streams de código aberto: Spark, Flink, Storm, Heron e Samza

Objetivos

O principal objetivo desta dissertação consiste na instalação, configuração, análise e comparação do desempenho das plataformas de código aberto Spark, Flink, Storm, Heron e Samza para processamento de big data streams.

Tarefas a Realizar

São propostas as seguintes tarefas para a execução do trabalho de investigação e de desenvolvimento, conducente à elaboração da dissertação de mestrado:

- Tarefa 1. Estudo dos principais conceitos subjacentes ao processamento de big data streams.
- Tarefa 2. Estudo analítico das plataformas para processamento de big data streams.
- Tarefa 3. Instalação e configuração do ambiente de teste para processamento de big data streams e instalação e configuração das plataformas Spark, Flink, Storm, Heron e Samza.
- Tarefa 4. Execução de experiências laboratoriais envolvendo o processamento de big data streams nas plataformas Spark, Flink, Storm, Heron e Samza.
- Tarefa 5. Análise e comparação do desempenho das plataformas Spark, Flink, Storm, Heron e Samza.
- Tarefa 6. Escrita de um artigo científico sobre o trabalho de investigação realizado e escrita da dissertação de mestrado.

Cronograma

A tabela seguinte representa a calendarização prevista para a execução das tarefas, em que a execução de uma dada tarefa num determinado mês é assinalada com um x.

Tarefa/mês	Set 19	Out 19	Nov 19	Dez 19	Jan 20	Fev 20	Mar 20	Abr 20	Mai 20	Jun 20
Tarefa 1	x									
Tarefa 2		x								
Tarefa 3			x	x	x					
Tarefa 4					x	x				
Tarefa 5						x	x			
Tarefa 6								x	x	x

Escrita da Dissertação em Língua Inglesa

A dissertação de mestrado resultante da realização do plano de trabalho proposto deverá ser escrita em língua inglesa, tendo em vista a divulgação internacional do trabalho científico desenvolvido. O título da dissertação em língua inglesa deverá ser o

seguinte: “Performance Comparison of Open-Source Big-Data Stream Processing Platforms: Spark, Flink, Storm, Heron and Samza”.

Referências

- [1] C. Prakash, Spark Streaming vs Flink vs Storm vs Kafka Streams vs Samza : Choose Your Stream Processing Framework, 2018. Online: <https://www.linkedin.com/pulse/spark-streaming-vs-flink-storm-kafka-streams-samza-choose-prakash/>.
- [2] S. Chintapalli, D. Dagit, B. Evans, R. Farivar, T. Graves, M. Holderbaugh, Z. Liu, K. Nusbaum, K. Patil, B. J. Peng and P. Poulosky, "Benchmarking Streaming Computation Engines: Storm, Flink and Spark Streaming", Proceedings of 2016 IEEE International Parallel and Distributed Processing Symposium Workshops, pp. 1789-1792.
- [3] M. A. Lopez, A. G. P. Lobato, O. C. M. B. Duarte, "A Performance Comparison of Open-Source Stream Processing Platforms", in Proceedings of 2016 IEEE Global Communications Conference (GLOBECOM 2016).
- [4] R. H. Bhole, V. M. Chapte, A. C. Karve, "A Study of Apache Kafka in Big Data Stream Processing", Proceedings of 2018 International Conference on Information, Communication, Engineering and Technology (ICICET), IEEE, pp 1-3.
- [5] A. Batyuk, V. Voityshyn, "Apache Storm Based on Topology for Real-Time Processing of Streaming Data from Social Networks", IEEE 2016 First International Conference on Data Stream Mining & Processing, pp. 345-349.