Processamento do fluxo de dados da rede baseado na arquitetura lambda

Alexsander Haas[1], João V.F. Lima[1]

[1]Universidade Federal de Santa Maria (UFSM)
Caixa Postal 97105-900 – Santa Maria – RS – Brasil

{ahass,jvlima}@inf.ufsm.br

Resumo. *O presente trabalho visa aplicar a arquitetura lambda na implementação de um sistema para o processamento do fluxo de dados da rede em tempo real. Realizando uma integração entre softwares open source para executar as etapas de geração dos dados, coleta, processamento e armazenamento.*

1. Introdução

Foram gerados mais de 30.000 g*igabytes* de dados por segundo na última década, essa taxa de criação continua a crescer [Marz et al. 2015]. A quantidade de dados gerados é diversa, desde postagens de usuários em redes sociais até as informações coletadas e transmitidas por dispositivos conectados à internet, também nomeada *Internet of Things* (IoT), esse conceito que eleva significativamente o tráfego de rede, pois cada vez mais dispositivos estão gerando e enviando informações. Em conjunto com esta expansão se encontra o problema de como realizar um monitoramento efetivo da rede, para que seja possível realizar o processamento do seu fluxo de dados e obter os resultados de maneira mais rápida e executar ações sobre eles.

Com base nisto o objetivo deste trabalho é aplicar a arquitetura lambda para o desenvolvimento de um sistema capaz de realizar o processamento do fluxo de dados em tempo real.

2. Sistema proposto

As abordagens de processamento de *Big Data*, geralmente utilizam técnicas de processamento em lote, utilizando *clusters* que executam os processos de modo paralelo, a técnica de execução por lote possui um tempo de resposta superior a 30 segundos, e algumas aplicações requerem respostas da ordem de sub-segundo [Rychly et al. 2014].

A arquitetura lambda é voltada para o processamento de uma carga massiva dados, ela utiliza o processamento em lote e por fluxo, trazendo uma visão em tempo real dos dados processados [Vögler et al. 2016]. Para o desenvolvimento do ambiente serão utilizados *softwares open source*, que serão integrados para que seja possível realizar o processamento por lote e em tempo real.

A arquitetura lambda é formada por três camadas [Marz et al. 2015]:

- **Lote:** que ingere e armazena uma grande quantidade de dados, realizando o processamento sobre eles.

- **Velocidade:** que ingere e processa os dados incrementando informações sobre eles, esta operação sendo realizada em uma forma de transmissão de baixa latência.
- **Serviço:** que tem como objetivo agregar as informações dos resultados obtidos da primeira e segunda camada, criando assim uma visão dos dados que foram processados pelas duas camadas anteriores.

O sistema proposto para o realizar o processamento é baseado na arquitetura lamba, e realiza a integração entre alguns *softwares open source*, destacados a seguir:

- **Bro:** ferramenta responsável pelo monitoramento em tempo real do tráfego de rede, gerando logs de informações.
- **Flume:** serviço distribuído utilizado para agregar e mover grandes quantidades de dados, utilizado como um canal para o transporte de dados.
- **Kafka:** serviço de mensagens distribuído, utilizado como um mediador de mensagens (*broker*) para abstrair todo o fluxo de dados em tópicos e por meio destes um terceiro consumir as informações.
- **Spark:** mecanismo utilizado para o processamento de dados em larga escala de forma paralela e distribuída. Realiza também processamento de fluxo em tempo real.
- **HBase:** base de dados distribuída, escalável e não relacional, executado no ecossistema do Hadoop.

Para executar as camadas da arquitetura lambda, foi realizado a seguinte integração entre os *softwares*. O Bro é responsável por gerar os logs com informações sobre o monitoramento da rede, e para o transporte destes logs é utilizado o Flume como um canal para envia-los para o Kafka, este agindo como mediador de mensagens criando tópicos sobre estas informações recebidas, para serem consumidos pelo Spark e por meio deste armazenar os dados no HBase.

4. Conclusão

A aplicação do sistema foi realizada inicialmente para o monitoramento da rede, processando em tempo real as informações obtidas e realizando o enriquecimento dos dados, para posteriormente armazena-los. Como sugestão de continuidade da implementação deste sistema, é utiliza-lo em uma análise do tráfego de rede para a detecção de intrusões por meio de algoritmos supervisionados de rede neural e árvore de decisão.

Referências

Vögler, M., Schleicher, J.M., Inzinger, C., and Dustdar, S. (2016). Ahab: A cloud-based distributed big data analytics framework for the Internet of Things. In software-practice & experience, pages 443-454.

Marz, N. and Warren, J. (2015). Big Data: Principles and Best Practices of Scalable Realtime Data Systems. Manning Publications, 1th edition.

Rychly, M., Koda, P. e Smrz, P. (2014). Scheduling decisions in stream processing on heterogeneous clusters. In Eighth International Conference on Complex, Intelligent and Software Intensive Systems (CISIS), pages 614-619.