JORNADA PCI/CBPF
APRESENTAÇÃO DE PÔSTER – 2019/2020

BOLSISTA:	**LUCCA LEON BRAGA**
E-MAIL:	**LUCCA.LEON@CBPF.BR**
SUPERVISOR:	**MARITA CAMPOS MAESTRELLI**
TÍTULO DO PROJETO:	**ESTUDO DE PADRÕES PREDITIVOS POR MEIO DE APRENDIZADO DE MÁQUINA COM FOCO EM MANUTENÇÕES PREVENTIVAS.**

Minimizar o tempo de indisponibilidade da rede tem sido um desafio tanto para os provedores de serviços de telecomunicações, sejam eles acadêmicos ou comerciais, como a RedeRio/FAPERJ de Computadores e a Rede nacional de Ensino e Pesquisa (RNP) , quanto para grandes usuários corporativos e/ou acadêmicos, como é o caso do CBPF.

Este projeto se propõe a pesquisar e desenvolver ferramentas para predição de *hard failures*, que são falhas de algum elemento ou componente da rede de maneira que ele não pode prover o serviço de conectividade, com técnicas de *Big Data Analytics* com o objetivo de minimizar o intervalo de tempo de indisponibilidade de rede. Desse modo, poderemos garantir maior tempo de disponibilidade de acesso aos recursos computacionais do CBPF, sejam eles para desenvolvimento de pesquisa científica, como o acesso do CBPF aos grandes experimentos internacionais, quanto para o provisionamento de serviços à comunidade científica.

Os parâmetros de operação desses equipamentos são reportados pela rede e registrados em uma estação de monitoramento através de protocolos e ferramentas de monitoramento bem estabelecidas pela comunidade, como *Simple Network Management Protocol* (SNMP) e *Syslog*. Esses dados são massivos e gerados em alta velocidade, culminando em uma grande quantidade de dados brutos que podem ser analisados para auxiliar na operação de rede. Ao estudar esses dados, muitas das falhas relacionadas aos equipamentos podem ser previstas para garantir um tempo de indisponibilidade mínimo. No entanto, pouco progresso foi obtido relacionado à análise desse tipo de dado e criação de inferência sobre o comportamento da rede.

As ferramentas de monitoramento de rede coletam informações de equipamentos, essas informações estão sendo estudadas e mapeadas, permitindo a compreensão de que dados são adequados para análise. Neste projeto serão estudados e processados os dados para desenvolver estratégias para analisar o grande volume de dados e extrair correlações e inferências. Através dessa análise seremos capazes de pesquisar e desenvolver algoritmos dinâmicos e adaptativos que se fazem necessários para processar a enorme quantidade de dados e gerar previsões baseadas em padrões e tendências. Com isso, poderemos identificar as diferentes abordagens existentes e desenvolver um algoritmo ótimo para predição de falhas.

Para este projeto utilizamos o Elasticsearch que é um mecanismo de busca e análise de dados distribuído e open source para todos os tipos de dados, incluindo textuais, numéricos, geoespaciais, estruturados e não estruturados. O Elasticsearch é o componente central do *Elastic Stack* (Elasticsearch, Logstash e Kibana), um conjunto de ferramentas *open source* para ingestão, enriquecimento, armazenamento, análise e visualização de dados.

Estas ferramentas foram instaladas em um ambiente de virtualização. Em um computador físico instalamos o software VirtualBox [4], um opensource de aplicativo gratuito, multiplataforma, onde pode-se criar, instalar máquinas virtuais que funcionam como se fossem diversos computadores instalados e funcionando dentro do seu próprio e único computador (Figura 1). Dentro deste configurou-se e instalou-se duas máquinas (computadores virtuais) ambas, com o sistema Debian. Em uma dessas máquinas foi também configurado e instalado o Elasticsearch e o Kibana.

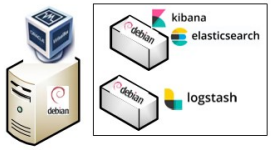

Figura 1. Infraestrutura do Ambiente de Monitoramento.

Em outro computador virtual foi instalado e configurado o Logstash, que é basicamente, uma ferramenta de extração de dados, de uso simples, com entrada (input), filtros e saída (output). Para cada etapa, o logstash possuí vários "plugins" que são pequenos programas que fazem as integrações do logstash. Esta ferramenta receberá os logs e enviará ao Elasticsearch, que por si só, ira organizar e armazená-los sendo visualizados no Kibana.

Figura 2. Pilha ELK.

O POP-RJ é o ponto de agregação das conexões locais das instituições atendidas pela RNP no Estado do rio de Janeiro. O POP é compostos por equipamentos como switches de acesso e Roteadores que fornecem os arquivos de logs para o Logstash (1). Esses arquivos de logs contem informações de taxa de erro, potência de recebimento e outros parâmetros de telemetria. O Logstash recebe esses dados, realiza transformações e extrai as informações através de filtros configurados pelo usuário, essas informações são enviadas para o Elasticsearch (2). Durante o processo de indexação, o Elasticsearch armazena documentos e desenvolve um índice invertido para tornar os dados dos documentos buscáveis praticamente em tempo real. Depois da indexação dos dados, os usuários podem executar consultas complexas com base em seus dados e usar agregações para recuperar resumos complexos dos dados. No Kibana, os usuários podem criar gráficos a partir de dados indexados no Elasticsearch (4).

Durante os testes e inicio de implementação, foi possível verificar padrões de informações de dados que antes eram desconhecidos ou simplesmente extensos demais para serem concatenados, a facilidade de visualização é o principal benefício dessa fase do projeto.