Grupo de Trabalho em Segurança de Redes
Edição 27 - Uberlândia – UniAlgar
13 de Maio, 2016

Solução de detecção de intrusão usando técnicas de Big Data para a análise de logs com o uso de Software Livre

William Rennan de Castro Vidal

Roteiro

* Motivação e Objetivos
* Conceitos Básicos
* Proposta e Implementação
* Avaliação Experimental
* Conclusão e Trabalhos Futuros

Roteiro

* **Motivação e Objetivos**
* Conceitos Básicos
* Proposta e Implementação
* Avaliação Experimental
* Conclusão e Trabalhos Futuros

Motivação

* Segurança em redes de computadores
  * Sistema de detecção/prevenção de intrusão
  * Há outras técnicas: Firewall, DMZ, etc.

* Problemas IDS/IPS
  * Quantidade de logs
  * Falta de padronização

Objetivos

* Implementar um IDS
  * Por meio de técnicas de Big Data
  * Baseado, exclusivamente, em Software Livre

* Avaliar o desempenho da solução proposta a partir de logs de autenticação reais do PoP-RS

Roteiro

* Motivação e Objetivos
* **Conceitos Básicos**
* Proposta e Implementação
* Avaliação Experimental
* Conclusão e Trabalhos Futuros

Conceitos Básicos

* Segurança da Informação
  * Base: Confidencialidade, Autenticidade e Disponibilidade
  * Usando técnicas de detecção, e prevenção de intrusão
* Detecção de Intrusão
  * Detecção Estatística de Anomalia
  * Detecção Baseada em Regras
* Necessidade: Analisar logs em busca de padrões de atividade maliciosas
* Problema: Tamanho e variedade dos logs

Conceitos Básicos

* Grande Volume de dados (Big Data)
  * Objetivo: Extrair de quantidade massivas de dados informações relevantes
  * Fases: Geração, aquisição, armazenamento e análise (Analytics)

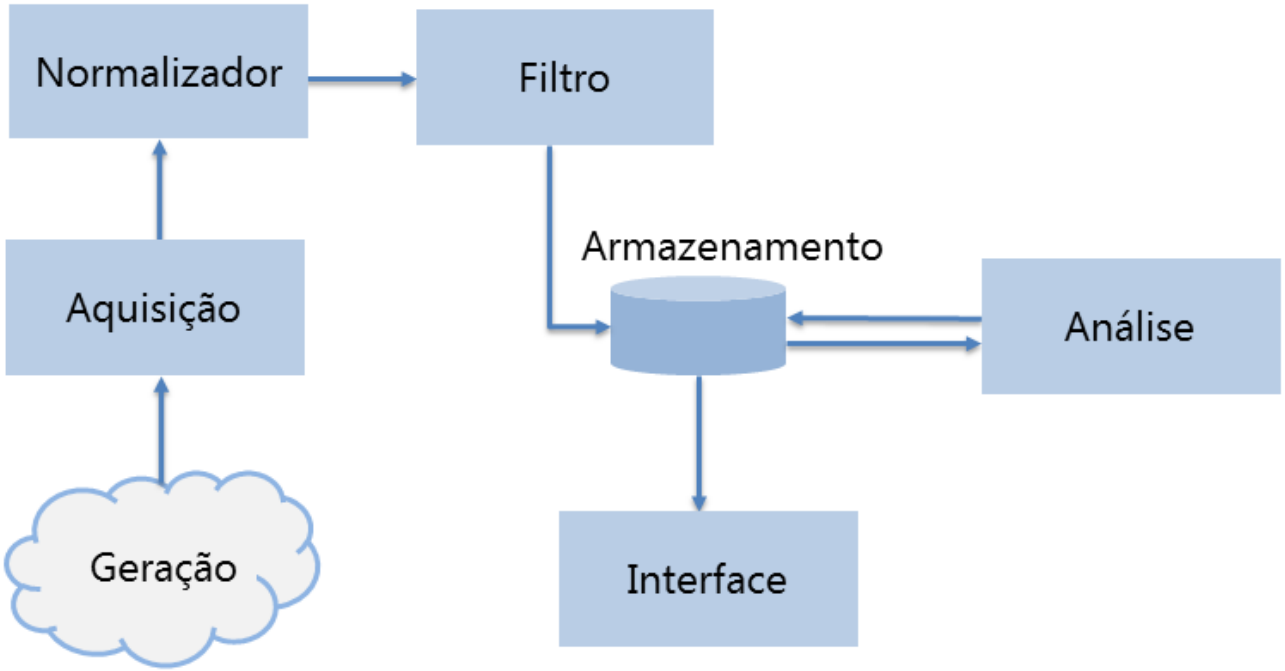
* Características:
  * Volume, velocidade, variedade e valor (4Vs)
  * Visualização

SCHROECK, M. et al. Analytics: The real world use of big data. ibm institute for business value—executive report. IBM Institute for Business Value, 2012.
DIJCKS, J. P. Oracle: Big data for the enterprise. Oracle White Paper, 2012.
NIST BIG DATA WORKING GROUP (NBD-WG). Big data: The next frontier for innovation, competition, and productivity. 2011. Disponível em: <https:http://bigdatawg.nist.gov/MGI_big_data_full_report.pdf>

Roteiro

* Motivação e Objetivos
* Conceitos Básicos
* **Proposta e Implementação**
* Avaliação Experimental
* Conclusão e Trabalhos Futuros

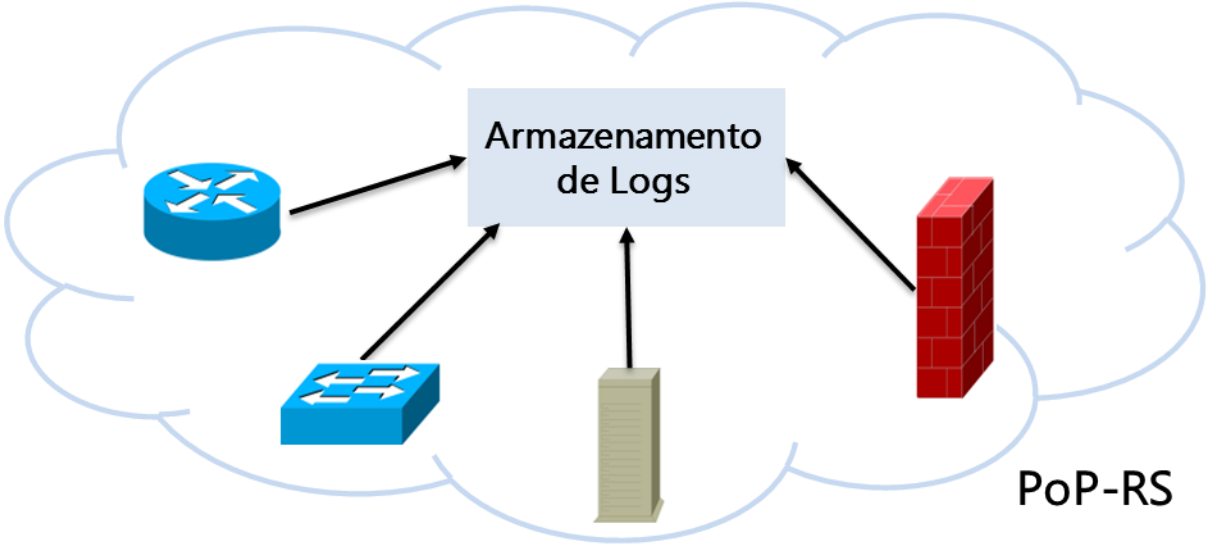
Arquitetura Proposta do Sistema



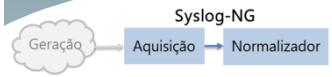
HU, H. et al. Toward scalable systems for big data analytics: A technology tutorial. Access, IEEE, IEEE, v. 2, p. 652–687, 2014.
KRUEGEL, C.; VALEUR, F.; VIGNA, G. Intrusion detection and correlation: challenges and solutions.
[S.l.]: Springer Science & Business Media, 2005. v. 14.

Geração

* Heterogêneos
* Não Estruturado



Aquisição e Normalização



* Facility

Marca Temporal	Origem	Serviço com PID	Mensagem
2015-07-17T12:03:51-03:00	tcc-server	sshd[58807]	Accepted publickey for agulha from 10.10.10.10 port 45162 ssh2

Filtro

* Extração de informações relevantes

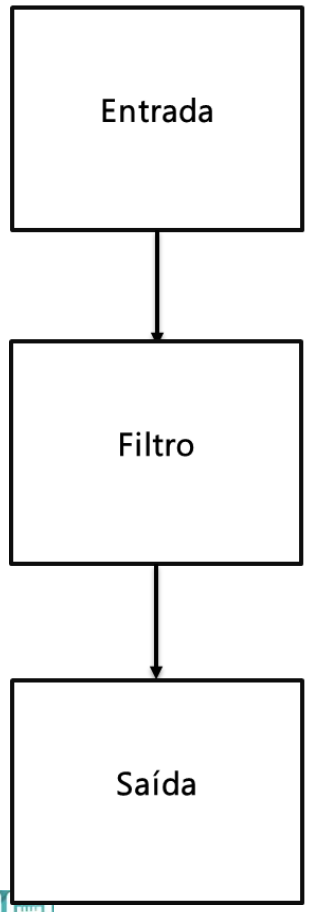
Regra:

"Accepted %{WORD:auth_method} for %{USER:username} from %{IP:src_ip} port %{INT:src_port} ssh2"

Accepted **publickey** for **agulha** from **10.10.10.10** port **45162** ssh2

logstash

Filtro



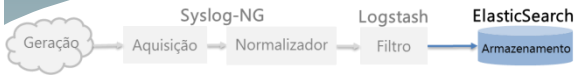

```
input {
      file {
          path => "/home/william/auth.log"
             type => “logAuth"
             start_position => "beginning"
           }
}


filter {
     grok {
        match => {"message" => "Accepted %{WORD:auth_method} for
                    %{USER:username} from %{IP:src_ip} port %{INT:src_port} ssh2" }
        add_field => {
           “ssh_type_conection” => “ssh_sucsessfull_login” }
}
…
}


output {
   elasticsearch {
          protocol => http
          host => "localhost"   }
}
```

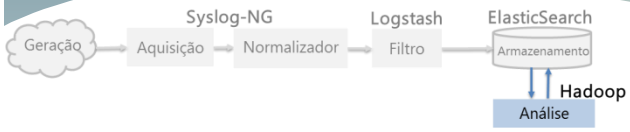
logstash

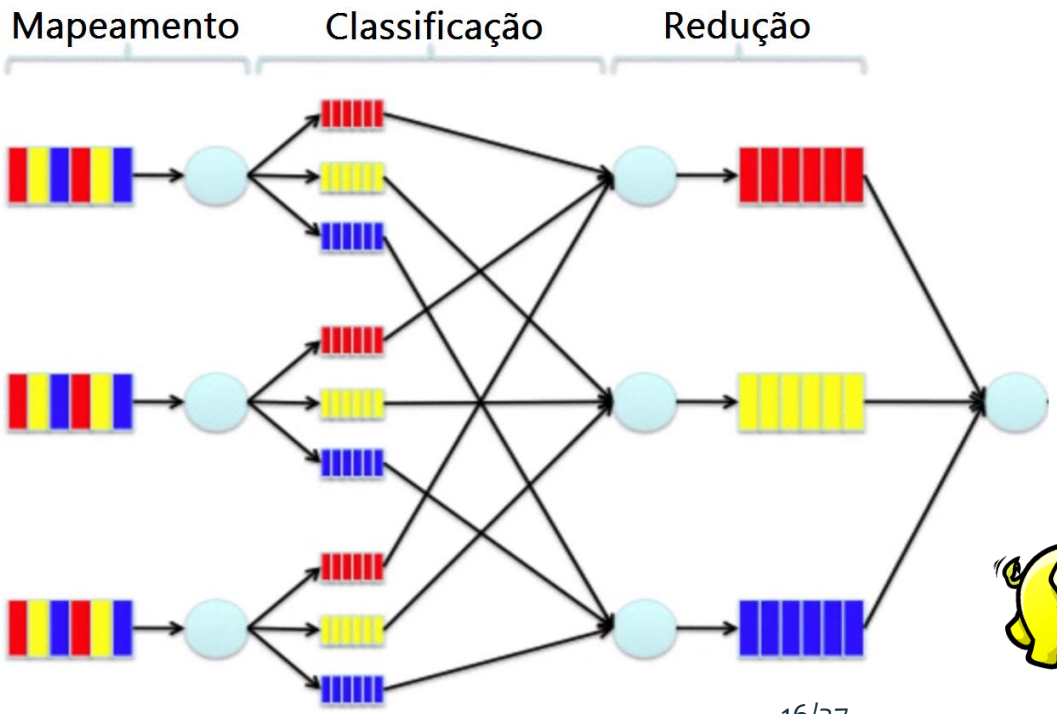
Armazenamento



{
"message" => "2015-07-17T12:03:51-03:00 tcc-server sshd[58807]: Accepted publickey for agulha from 10.10.10.10. port 45162 ssh2",
"auth_method" => "publickey",
"username" => "agulha",
"src_ip" => "10.10.10.10",
"src_port" => "45162",
"timestamp" => "17/Jul/2015 12:03:51 -0300",
"target" => "tcc-server",
"service" => "sshd",
"pid" => "58807",
}

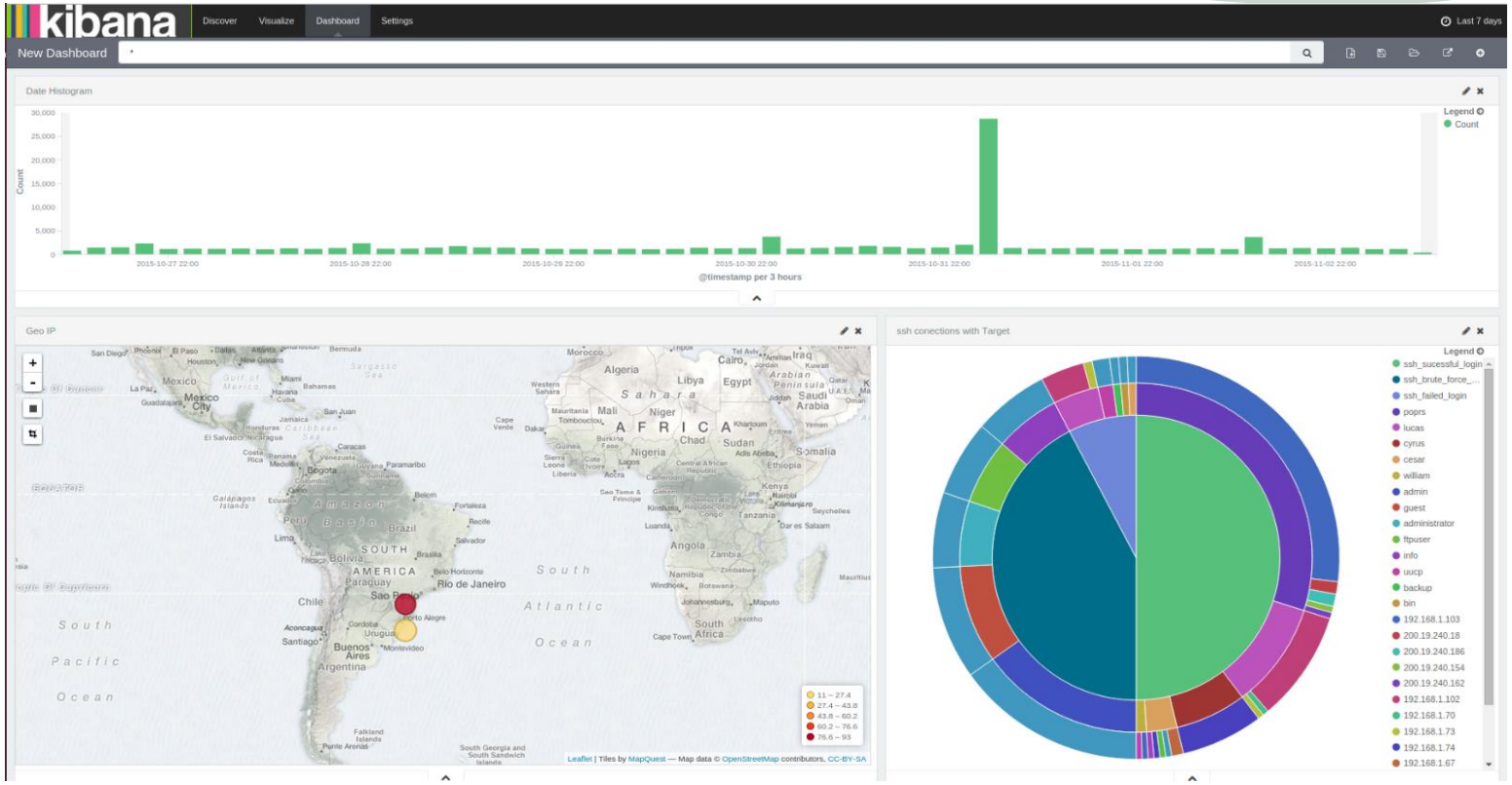
Análise





Correlação: MapReduce

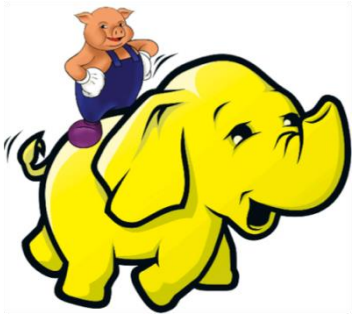
Visualização



Dificuldades

* Programação com o modelo MapReduce
  * Complexidade em mapear a solução em duas primitivas

* Solução possível:
  * Utilização do Pig
    * Linguagem Script
    * Abstração ao MapReduce



Dificuldade Adicional

* Existem pacotes Pig para carregar informações específicas
  * Log do Servidor Apache
* Não é exaustivo
  * Expressão Regular para caso específico
  * Dificuldade de escrita

```
REGISTER /usr/lib/pig/pig-0.15.0/lib/piggybank.jar;
logs = LOAD '/home/hduser/logsAuth/auth*.log' USING org.apache.pig.piggybank.storage.
    MyRegExLoader('^(\\S+)\\s+(\\S+)\\s+(\\w+)\\S+\\s+([\\S+\\s+]*)\$') AS (timestamp:
    chararray, host:chararray, program:chararray, message:chararray);
searchData = FILTER logs BY message MATCHES '.*agulha.*';
STORE searchAdata INTO '$myOutput/' USING PigStorage(' ');
```

Novo Problema

* Bug na utilização do Pig

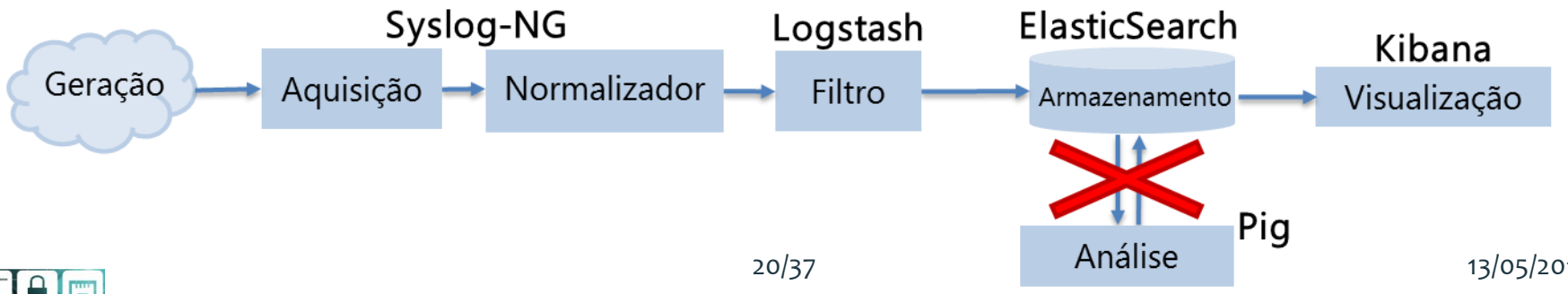
Elasticsearch for Apache Hadoop 2.1.1

Bug fixes

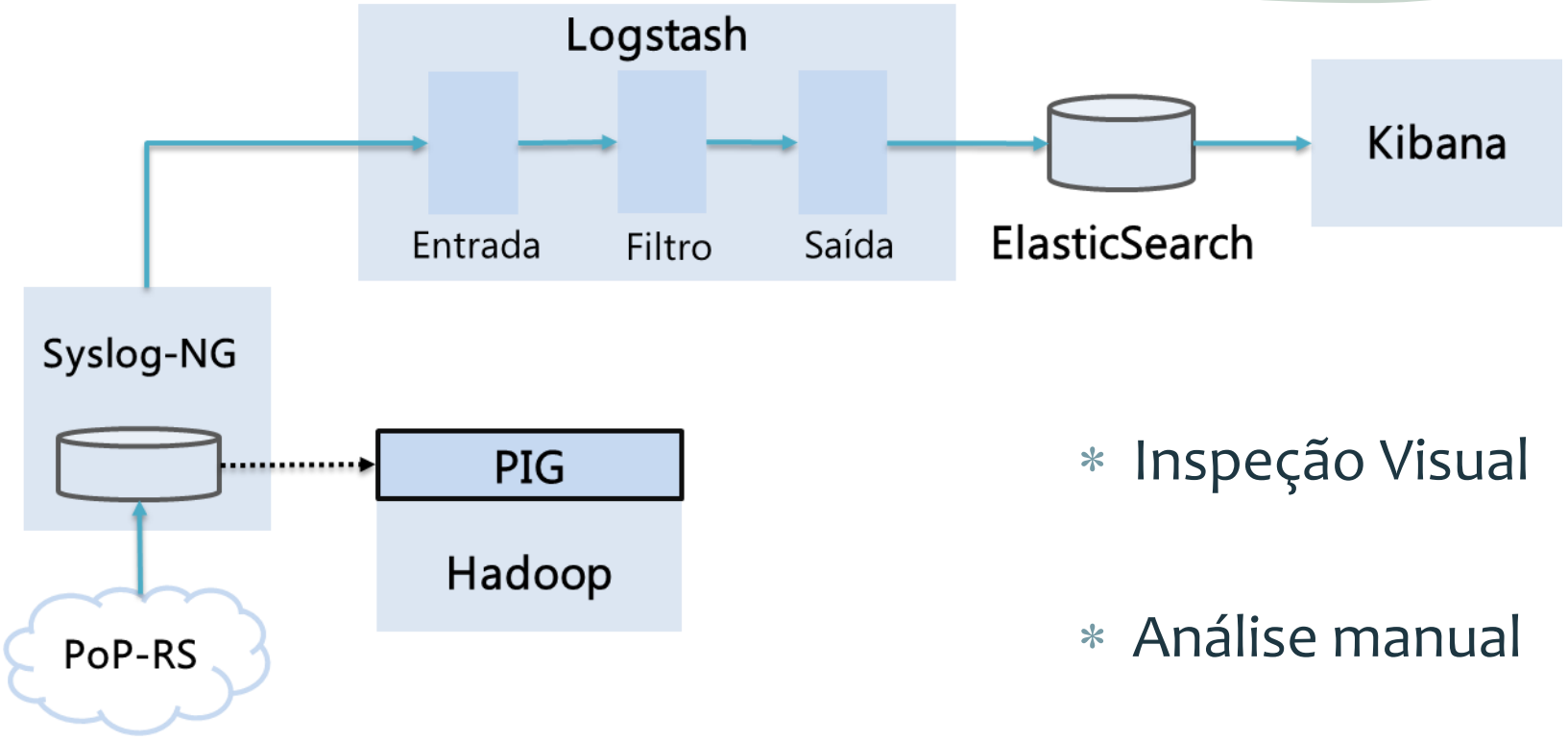
- Load error from elasticsearch using Pig and the elasticsearch connector #499

https://www.elastic.co/downloads/past-releases/elasticsearch-apache-hadoop-2-1-1

* Consequência: Perdida a ligação com o ElasticSearch



Implementação Final



* Inspeção Visual
* Análise manual

Roteiro

* Motivação e Objetivos
* Conceitos Básicos
* Proposta e Implementação
* **Avaliação Experimental**
* Conclusão e Trabalhos Futuros

Metodologia

* Uso de logs reais do PoP-RS
* Comparação: desempenho bash x Pig Latin
  * Critério: tempo de execução
  * Diferentes volumes de dados
* Validação estatística: Média e Desvio Padrão

Plataforma Experimental

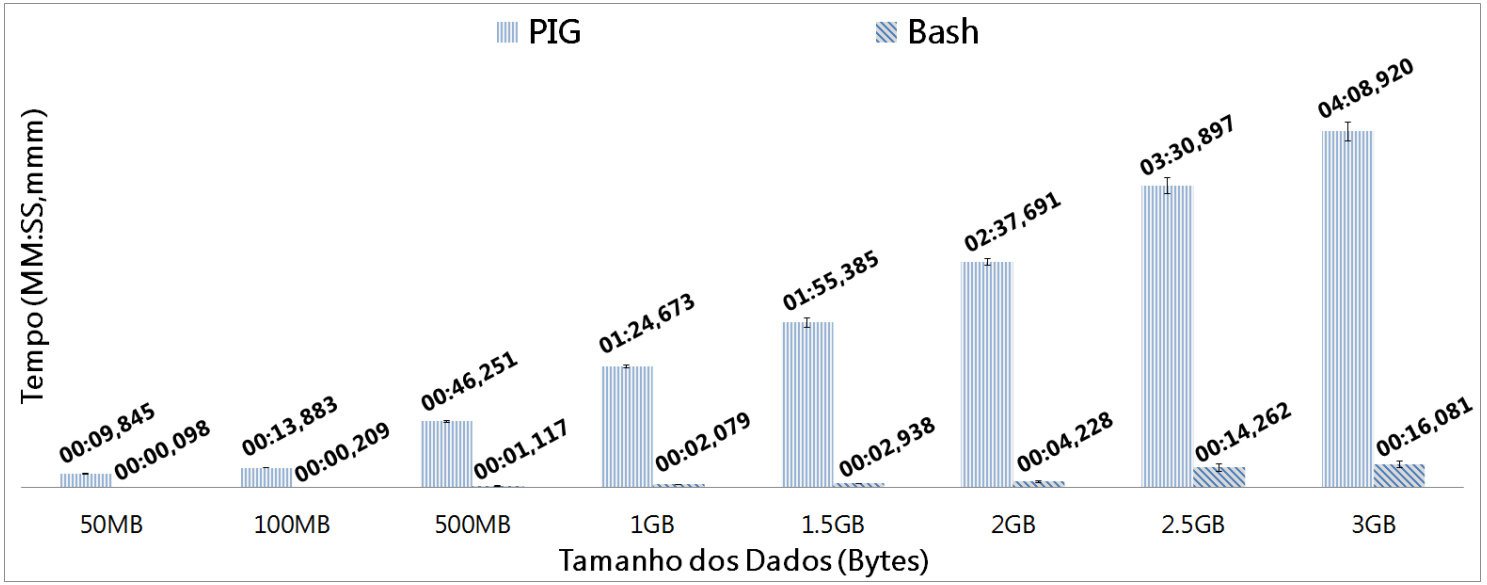
* Hardware
* HP ProLiant DL380
* Servidor Virtulizado
  * 4 CPUs
  * 4GB RAM
  * 150GB Disco
* Hypervisor: XenServer 6.5
* Software
* Ubuntu 12.04.5 LTS
* Hadoop 2.7.0
* Pig 0.15.0
* *demais serviços (Syslog-NG, Logstash, ElasticSearch, etc.)

Cenários de Teste

* Cenário I
  * Busca em campo específico
  * Busca em campo de strings
* Cenário II
  * Busca em 2 e 3 campos
* Cenário III
  * Busca com ordenação

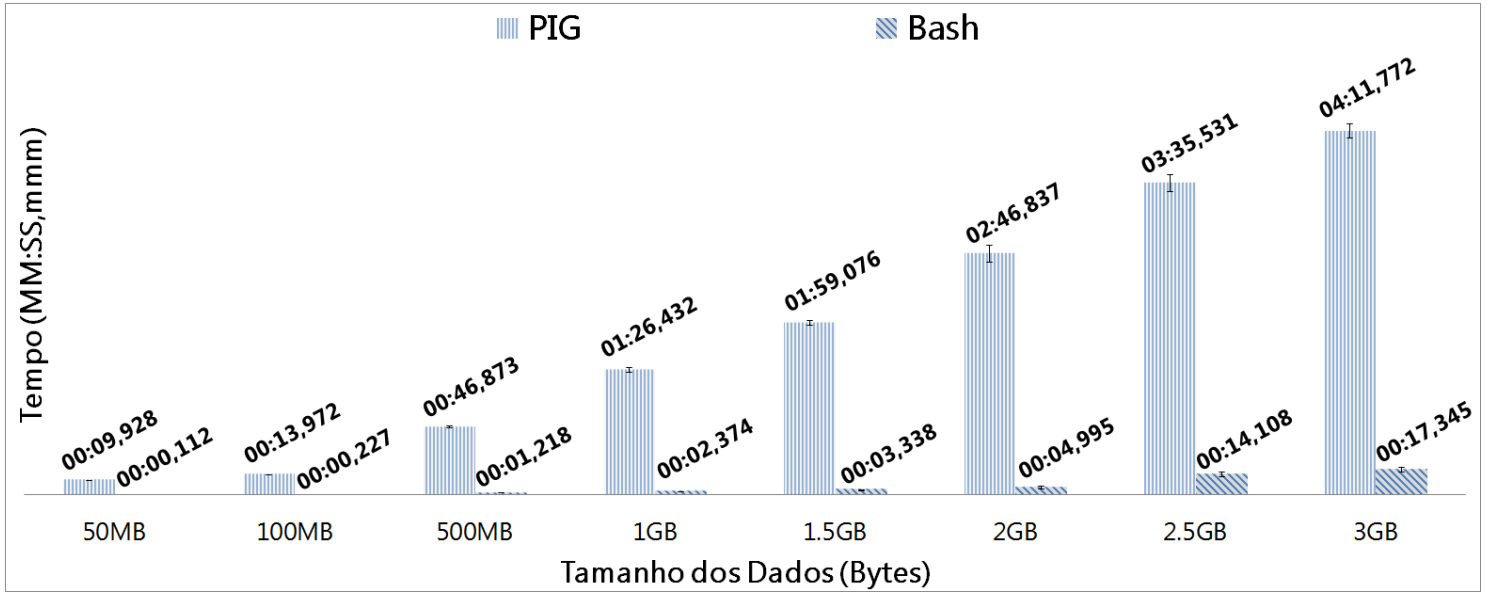
Cenário I: Busca em um campo específico

* Busca por hospedeiro



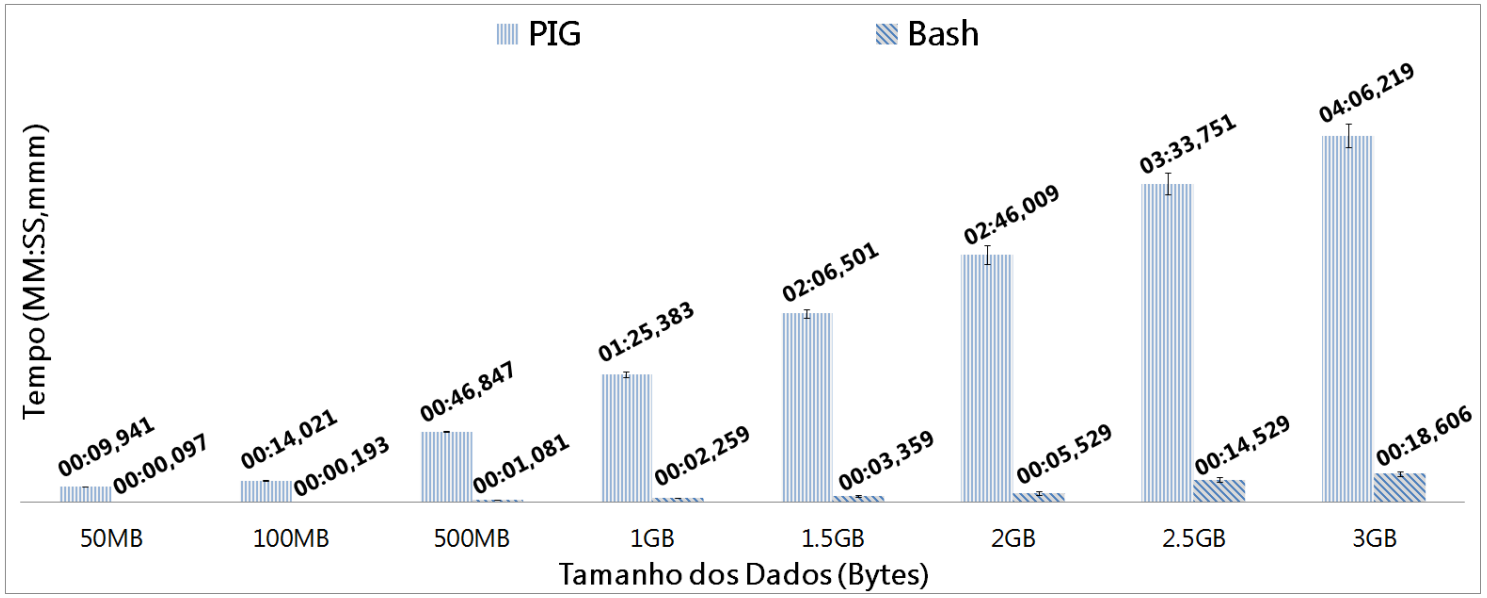
Cenário I: Busca em um campo com strings

* Busca por usuário



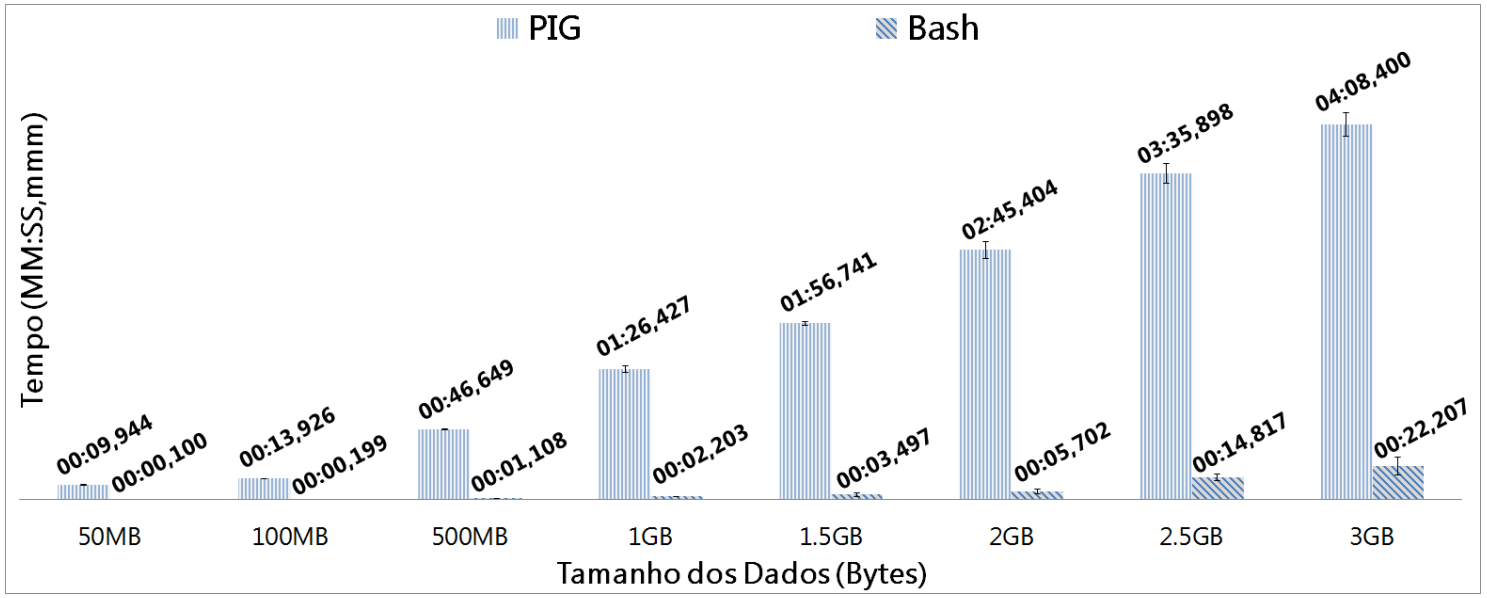
Cenário II: Busca em dois campos

* Busca por data e usuário



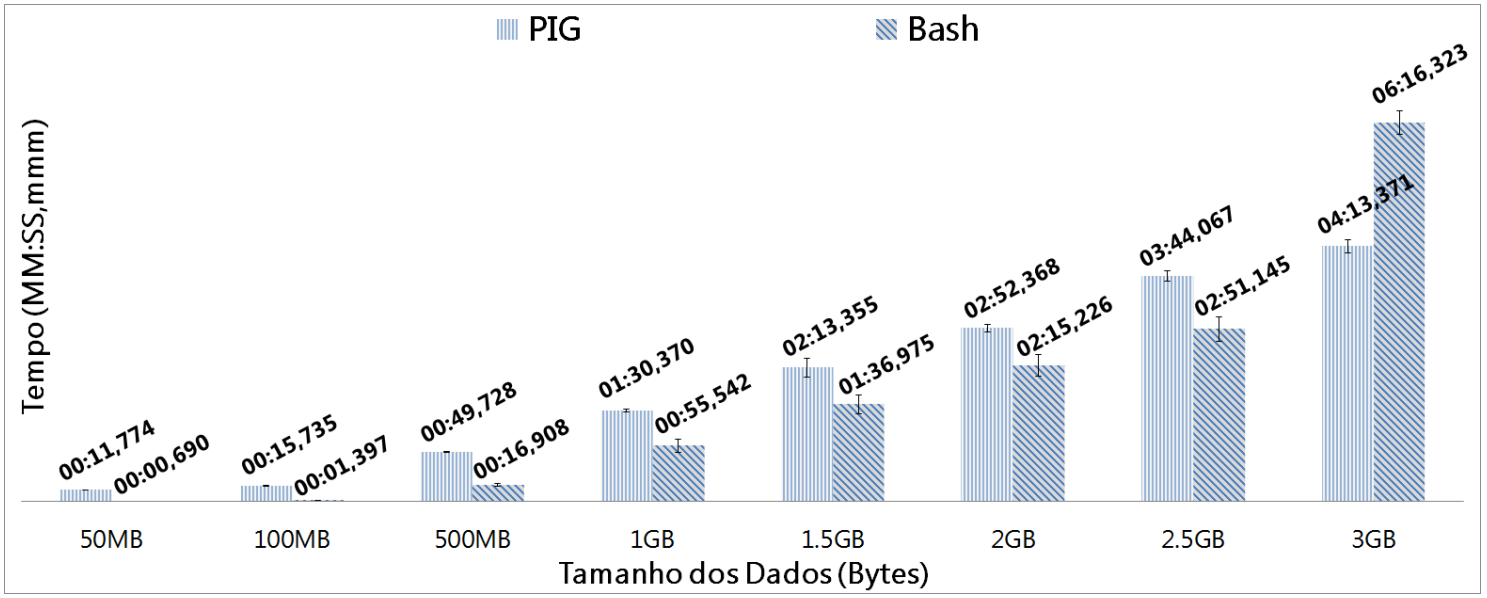
Cenário II: Busca em três campos

* Busca por data, hospedeiro e usuário



Cenário III: Busca com ordenação

* Ordenação e busca por data e usuário



Roteiro

* Motivação e Objetivos
* Conceitos Básicos
* Proposta e Implementação
* Avaliação Experimental
* **Conclusão e Trabalhos Futuros**

Conclusão

* Estudo de viabilidade
  * Funciona, implementado e testado
  * Integrada ao ambiente de produção do PoP-RS
    * Análise manual por inspeção visual

* Resultados
  * Baixo desempenho se comparado com grep no bash
    * Camadas de Software (Hadoop+Pig)
    * Volume de dados talvez seja pequeno
    * Não foi usado em ambiente distribuído

Conclusão

* Contribuição
  * Protótipo funcional para análise de logs para o PoP-RS
  * Estudo de ferramentas em Software Livre para Big Data
    * Logstash, ElasticSearh, Fluentd, Kibana, Hadoop, Flume, Pig,...

Trabalhos Futuros

* Implementar em um ambiente distribuído
* Utilizar Apache Storm e Apache Spark para análise em tempo real
* Implementar Regras e Filtros mais complexos
* Analisar desempenho para volume de dados maiores

Agradecimento

Contatos

* Linkedin: https://www.linkedin.com/in/william-vidal-317126105
* E-mail: wrcvidal@gmail.com
* TCC: http://www.lume.ufrgs.br/handle/10183/139086

Obrigado!

* Perguntas?
* Demonstração?