

1 Relatório técnico descritivo sobre projeto de código livre

Este modelo reduzido visa proporcionar um roteiro prático para a redação do relatório da atividade complementar #4, na forma de relatório simplificado.

Esta seção é destinada a apresentar os objetivos do trabalho.

1.1 Descrição do projeto

Apache Hadoop é um projeto de *software open-source* que permite o processamento distribuído de grandes quantidades de dados em *clusters*. Em outras palavras, é um *framework* que trabalha o conceito de *Big Data*, dividindo os dados em blocos que são distribuídos pelos nós dentro do *cluster*, para serem processados paralelamente por meio de um algoritmo de *Map-Reduce*. Como principal vantagem, o sistema apresenta grande capacidade de tolerar falhas, tornando-o uma abordagem muito mais barata que a escolha de um *hardware* de ponta para a obtenção da mesma disponibilidade dos dados. Diversas empresas que trabalham com quantidades massivas de dados e necessitam de processamento rápido e tolerante a falhas utilizam o *Hadoop*, entre elas:

- Amazon;
- Adobe;
- eBay;
- Facebook;
- Google;
- IBM;
- Last.fm;
- LinkedIn;
- Spotify;
- Twitter;
- Yahoo!.

2 Caracterização do projeto de código livre

Pesquise sobre o projeto para responder as seguintes questões.

2.1 Desenvolvimento

- Existe um local dedicado para o desenvolvimento?

Não foram encontrados indícios de que a ferramenta possui um local dedicado para o desenvolvimento. Na verdade, a existência de um corpo de desenvolvedores de diferentes fusos horários permite a conclusão de que o desenvolvimento é feito mundo afora, tendo a rede como plataforma de comunicação entre os voluntários que trabalham na ferramenta.

- É possível extrair o atual código fonte a partir de um repositório público de código fonte?

No site da ferramenta é possível fazer o *download* do código fonte.

- Quão grande é o código?

O código descompactado possui cerca de 82mb e é dividido em vários módulos. Trata-se de uma ferramenta bastante complexa, portanto o código é bastante extenso.

- Quais são as principais linguagens de programação?

A maior parte da ferramenta é escrita em Java, embora uma parcela considerável esteja sendo reescrita em C e C++ por motivos de velocidade e segurança.

- A utilização do pacote depende de algum outro software proprietário ou de código fonte aberto?

Sim. Para que o *Hadoop* possa ser compilado, os seguintes requerimentos devem ser cumpridos (considerando o caso em que o sistema utilizado será UNIX):

- JDK 1.6 ou superior;
- Maven 3.0 ou superior;
- ProtocolBuffer 2.5.0;
- CMake 2.6 ou superior;
- Findbugs 1.3.9 (opcional, apenas se o usuário quiser utilizar tal ferramenta);
- Zlib devel;
- openssl devel.

- É possível calcular o número de *downloads* ou usuários de uma versão em particular?

A informação para este tópico não foi encontrada nos sites da ferramenta.

2.2 Licença Software Livre

- Quem são os patrocinadores que contribuem para a sustentabilidade do projeto?

O projeto possui diferentes níveis de patrocínio, que são classificados como Platina, Ouro, Prata, Bronze e Apoio de Infraestrutura. Entre os patrocinadores Platina, encontram-se empresas como *Facebook*, *Microsoft* e *Google*. Já a *IBM* e *HP* encontram-se entre os patrocinadores Ouro. *Pivotal* e *Huawei* são alguns dos patrocinadores Prata do projeto, enquanto

Twitter e *Samsung* são exemplos de patrocinadores Bronze. Por fim, algumas empresas e universidades são apoiadoras de infraestrutura do projeto. Um exemplo é a empresa de segurança *Symantec*.

- Quem detém os direitos autorais do código?

Os direitos autorais do *Hadoop* pertencem à ASF (*Apache Software Foundation*).

- O projeto está sob qual tipo de licença de código aberto?

O projeto se encontra sob a licença *Apache*, versão 2.0.

- Por que os responsáveis pelo projeto escolheram a licença de código aberto?

O tipo de licença escolhida faz parte da cultura da fundação, visto que todos os seus projetos respeitam a mesma política, na qual todas as empresas que utilizam o código e fazem modificações no mesmo devem utilizar o selo *Powered By*. Em outras palavras, apenas as *releases* lançadas pela *Apache* podem receber o nome *Apache Hadoop*, ou serem declaradas como distribuições dela. Além disso, pode-se supor que, pelo fato da ferramenta ser utilizada por grandes empresas de vários ramos, é muito interessante em termos de *marketing* que o selo *Powered By* seja utilizado pelas mesmas. Além disso, a abertura da licença permite que desenvolvedores voluntários de diversas empresas e com diversas especialidades contribuam com o projeto, tornando-o cada vez mais completo e robusto.

2.3 Governança

- Existem quantos desenvolvedores alocados para o projeto?

O projeto possui vários de desenvolvedores alocados no momento. São todos voluntários, divididos entre *Project Management Committee* (PMC) e *committers*, sendo que um não exclui o outro (um membro pode fazer parte dos dois grupos), os quais trabalham em diferentes empresas como *Yahoo!*, *Facebook*, *Hortonworks*, *Cloudera*, *Microsoft*, *VMware*, *Twitter*, *IBM* e *Intel*.

- Quantos *committers*, também conhecidos por desenvolvedores que podem realizar mudanças propostas, o projeto possui?

Atualmente, o projeto possui 81 *committers* ativos. Entretanto, as listas de *e-mails* possuem centenas de usuários que contribuem com dúvidas, experiências e sugestões.

- O que você pode dizer sobre o modelo de governança de código fonte aberto?

O projeto apresenta um modelo bastante interessante, visto que membros de diversas empresas podem contribuir para a sua construção, o que torna o código bastante completo, otimizado para a realidade dos seus usuários, seguro e menos exposto a falhas de maneira geral, visto que o código-fonte é aberto e mesmo os usuários que não são *committers* podem relatar experiências nas listas de *e-mails*, o que enriquece todo o processo de desenvolvimento. Por outro lado, a adoção da licença *Apache* é interessante em termos de *marketing*, conforme mencionado, além de impedir que o projeto seja associado a uma única empresa. Por fim, nada impede que desenvolvedores autônomos façam modificações no código, desde que a

licença seja respeitada. Em outras palavras, todos podem utilizar a ferramenta e modificá-la de acordo com as suas necessidades, não apenas grandes empresas, embora sejam estas que contribuem com *committers* e com patrocínio financeiro.

2.4 Manutenção

- Gerenciamento de *releases*: Qual o número e frequência de *releases*?

O projeto possui 67 *releases* entre 4 de setembro de 2007 e 19 de novembro de 2014. Em outros termos, uma nova *release* é lançada a cada um mês e 8 dias, em média. O número citado contabiliza apenas as *releases* oficiais, mas diversas empresas possuem versões modificadas da ferramenta.

- Comunicação: Existe um canal de comunicação útil e ativo para a comunidade/suporte ao usuário?

Além do site oficial, que contém todas as releases com as respectivas mudanças/correções, existem listas de *e-mails* nas quais os usuários podem postar dúvidas e compartilhar experiências. Trata-se de uma abordagem interessante, visto que gera acúmulo de conhecimento e permite a resposta de maneira mais rápida e completa, visto que as postagens poderão ser vistas por todos os desenvolvedores e pelos demais usuários da ferramenta.

- Existe um *bugtracker* (rastreamento de bugs) com uma lista de bugs corrigidos/pendentes de correção?

Sim. O projeto possui 4 *Issue Trackings*, sendo um para o projeto de forma geral, e outros 3 para os módulos que compõem a ferramenta.

- Existe um plano de metas para planos futuros? Existe evidência que o plano de metas foi utilizado no passado?

Dentro de cada *Issue Tracking* existe um *Road Map* com problemas a serem resolvidos em cada *release* da ferramenta. Pode-se perceber que os *Road Maps* são utilizados de forma efetiva pelos desenvolvedores, ou seja, há evidências de que a abordagem já pautou o lançamento de novas versões da ferramenta.

- Existe consultoria comercial, treinamento ou consulta disponível para o projeto? A partir de múltiplos prestadores de serviços?

No site oficial da ferramenta não existem referências, mas a partir de pesquisas por consultoria na área, existem diversas empresas que prestam serviços na área de *Big Data* e que trabalham com *Hadoop*. São alguns exemplos:

- Inquidia;
- Altoros;
- Maestro Technologies;
- Business Data Miners;
- WinterCorp.