# Apache Tez

***Fernando Anselmo***
*http://fernandoanselmo.orgfree.com/wordpress/*

Versão 0.01 em 12 de novembro de 2023

**Resumo**

Tez[1] ou "Apache Tez"(aqui chamarei apenas de Tez) é parte do Ecosistema Hadoop criado como um framework de processamento de dados de código aberto. Foi projetado para lidar com cargas de trabalho de processamento de dados complexas e intensivas que envolvem a execução de consultas ad-hoc e interativas em grandes conjuntos de dados.

## 1 Parte inicial

Tez melhora a velocidade e a capacidade de resposta do Hadoop, permitindo que os trabalhos sejam concluídos em menos etapas e com menos recursos. Ele faz isso através de uma arquitetura mais flexível que permite que os trabalhos sejam modelados como um gráfico de tarefas complexas, em vez de uma série de etapas sequenciais.


**Figura 1:** *Logo do Apache Tez*

Além disso, Tez também suporta uma variedade de APIs de alto nível, incluindo MapReduce, Apache Hive e Apache Pig, tornando-o uma ferramenta versátil para processamento de dados em larga escala. Por fim, é importante notar que o Apache Tez é um projeto da Apache Software Foundation, o que significa que é mantido e atualizado por uma comunidade de desenvolvedores voluntários de todo o mundo.

Funcionalidades principais do Tez:

1. Capacitar os usuários finais:

- APIs expressivas de definição de fluxo de dados
- Modelo flexível de tempo de execução de entrada-processador-saída
- Independente de tipo de dados
- Simplificando a implantação

2. 2. Desempenho de execução

- Ganhos de desempenho em relação ao Map Reduce
- Gerenciamento ideal de recursos
- Planeje a reconfiguração em tempo de execução
- Decisões dinâmicas de fluxo de dados físicos

## 1.1 Vantagens e Desvantagens

O Apache Tez oferece várias vantagens para o processamento de dados em larga escala, incluindo:

- **Desempenho aprimorado**: projetado para lidar com cargas de trabalho de processamento de dados complexas e intensivas. Ele melhora a velocidade e a capacidade de resposta do Hadoop, permitindo que os trabalhos sejam concluídos em menos etapas e com menos recursos.
- **Flexibilidade**: permite que os trabalhos sejam modelados como um gráfico de tarefas complexas, em vez de uma série de etapas sequenciais. Isso oferece uma flexibilidade significativa na forma como os trabalhos são estruturados e executados.
- **Compatibilidade**: suporta uma variedade de APIs de alto nível, incluindo MapReduce, Apache Hive e Apache Pig. Isso significa que os desenvolvedores que já estão familiarizados com essas ferramentas podem começar a usar o Tez com relativa facilidade.
- **Escalabilidade**: Como parte do ecossistema Hadoop, o Tez é capaz de lidar com conjuntos de dados extremamente grandes, tornando-o uma ferramenta ideal para empresas que precisam processar grandes volumes de dados.
- **Código aberto**: projeto de código aberto, o que significa que é livre para usar e modificar. Além disso, ele é mantido por uma comunidade ativa de desenvolvedores, o que garante que o software continue a ser atualizado e melhorado ao longo do tempo.

Embora o Apache Tez ofereça várias vantagens, também existem algumas desvantagens que devem ser consideradas:

- **Complexidade**: Embora a flexibilidade seja uma vantagem, também pode adicionar complexidade. Modelar trabalhos como um gráfico de tarefas complexas pode ser mais difícil do que usar uma série de etapas sequenciais, especialmente para desenvolvedores menos experientes.
- **Curva de aprendizado**: Embora seja compatível com várias APIs de alto nível, ainda pode haver uma curva de aprendizado para desenvolvedores que não estão familiarizados com essas ferramentas.
- **Dependência do ecossistema Hadoop**: O Tez é projetado para funcionar com o Hadoop e pode não ser a melhor escolha para empresas que não estão usando o Hadoop.

- **Recursos**: Como é projetado para processar grandes volumes de dados, ele pode exigir uma quantidade significativa de recursos computacionais. Isso pode ser um problema para empresas com recursos limitados.
- **Suporte**: Como um projeto de código aberto, o suporte pode ser limitado em comparação com soluções comerciais. Embora exista uma comunidade ativa de desenvolvedores, não há garantia de suporte dedicado.

Exemplo de caso de uso para o Apache Tez:

Vamos supor que trabalhe em uma empresa de comércio eletrônico e necessite analisar o comportamento dos clientes do site da empresa. Possui um grande volume de dados de log do servidor web que precisa ser processado para extrair informações úteis. Nesse caso, usamos o Apache Tez para processar esses dados de log. Primeiro, o Apache Hive (que é compatível com o Tez) para escrever consultas SQL-like que extraem informações dos dados de log, como o número de visitas a diferentes páginas ou o tempo médio gasto em cada página.

Em seguida, usar o Tez para executar essas consultas em paralelo no cluster Hadoop. Divide o trabalho em várias tarefas menores que podem ser executadas simultaneamente, o que pode acelerar significativamente o processamento. Finalmente, usar os resultados dessas consultas para obter insights sobre o comportamento do cliente, como quais páginas são mais populares ou quais produtos estão sendo mais visualizados. Essas informações podem ser usadas para melhorar o design do site ou recomendar produtos.

# 2 Hadoop no Docker

O modo mais simples de se conseguir trabalhar com o Hadoop é utilizando o Docker, para baixar a imagem do Hadoop:

```
$ docker pull suhothayan/hadoop-spark-pig-hive:2.9.2
```

Nessa imagem temos outros produtos do Ecossistema Hadoop: Spark, Pig e Hive. Para criar e executar a primeira vez o contêiner (a pasta que este comando for executado será associada a uma pasta interna chamada **/home/tsthadoop**):

```
$ docker run -it -d --name meu-hadoop -v $(pwd):/home/tsthadoop
suhothayan/hadoop-spark-pig-hive:2.9.2
```

Uma vez interrompido o contêiner:

```
$ docker stop meu-hadoop
```

Podemos executá-lo novamente com os seguintes comandos:

```
$ docker start meu-hadoop
$ docker exec -it meu-hadoop /etc/bootstrap.sh bash
```

## 2.1 Erro de Permissão

Na primeira vez que entramos é dado um erro na execução do script "bootstrap.sh"de permissão negada para executar o script "spark-env.sh", vamos corrigir isso com o comando:

```
# chmod 777 /usr/local/spark/conf/spark-env.sh
```

Vamos sair do bash:
```
# exit
```

Podemos executá-lo novamente:
```
$ docker exec -it meu-hadoop /etc/bootstrap.sh bash
```

E o erro desapareceu.

# 3 Conclusão

Projetado para melhorar a eficiência do processamento de dados no Hadoop, permitindo que os trabalhos sejam concluídos mais rapidamente e com menos recursos. Ele faz isso ao permitir que os trabalhos sejam modelados como um gráfico de tarefas complexas, em vez de uma série de etapas sequenciais, como é o caso com o modelo MapReduce padrão do Hadoop.

Além disso, o Tez é compatível com várias APIs de alto nível que são comumente usadas no ecossistema Hadoop, incluindo MapReduce, Hive e Pig. Isso significa que os desenvolvedores que já estão familiarizados com essas ferramentas podem começar a usar o Tez com relativa facilidade.

Sou um entusiasta do mundo **Open Source** e novas tecnologias. Qual a diferença entre Livre e Open Source? Livre significa que esta apostila é gratuita e pode ser compartilhada a vontade. Open Source além de livre todos os arquivos que permitem a geração desta (chamados de arquivos fontes) devem ser disponibilizados para que qualquer pessoa possa modificar ao seu prazer, gerar novas, complementar ou fazer o que quiser. Os fontes da apostila (que foi produzida com o LaTex) está disponibilizado no GitHub [4]. Veja ainda outros artigos que publico sobre tecnologia através do meu Blog Oficial [2].

# Referências


[1] Página do Apache Tez
`https://tez.apache.org/`

[2] Fernando Anselmo - Blog Oficial de Tecnologia
`http://www.fernandoanselmo.blogspot.com.br/`

[3] Encontre essa e outras publicações em
`https://cetrex.academia.edu/FernandoAnselmo`

[4] Repositório para os fontes da apostila
`https://github.com/fernandoans/publicacoes`