
SCIENTIA, VITA ET FIDES
PUCPR
GRUPO MARISTA

Sumário

O QUE É UMA ATIVIDADE PRÁTICA?

A **Atividade Prática (ATP)** tem como proposta ser voltada para sua atuação no mundo do trabalho, visando à sua formação profissional e resultando, no final da disciplina, em um "produto", o qual pode ser.

COMO SEREI AVALIADO?

Após entregar o produto de sua ATP, a correção feita pelo professor-tutor levará em conta **critérios avaliativos ou rubricas preestabelecidos**. Esses critérios são elaborados considerando o que é solicitado nas etapas da ATP e sua relação com os RAs e IDs estabelecidos para a disciplina.

Veja um exemplo:

Para correção, o professor-tutor estará pautado em um quadro como o apresentado na sequência. A **coluna 1** corresponde **aos critérios avaliados**, ou seja, os pontos que serão observados no produto entregue (note que **cada uma das linhas representa um critério**). As **colunas 2 a 4** representam os **descritores**, ou seja, a classificação que será aplicada no produto entregue, os quais possuem **níveis de desempenho**, do Autônomo ao Em desenvolvimento.

CRITÉRIOS	DESCRITORES			
	Autônomo (100% a 90%)	**Capaz** (89% a 70%)	**Aprendiz** (69% a 40%)	Em desenvolvimento (39% a 0%)
Critério 1. (20%)	Descrição do "produto" que atinge entre 90-100% do que se espera.	Descrição do "produto" que atinge entre 89-70% do que se espera.	Descrição do "produto" que atinge entre 69-40% do que se espera.	Descrição do "produto" que atinge entre 39-0% do que se espera.
Critério 2. (25%)	Descrição do "produto" que atinge entre 90-100% do que se espera.	Descrição do "produto" que atinge entre 89-70% do que se espera.	Descrição do "produto" que atinge entre 69-40% do que se espera.	Descrição do "produto" que atinge entre 39-0% do que se espera.
. . .	...	...	...	...

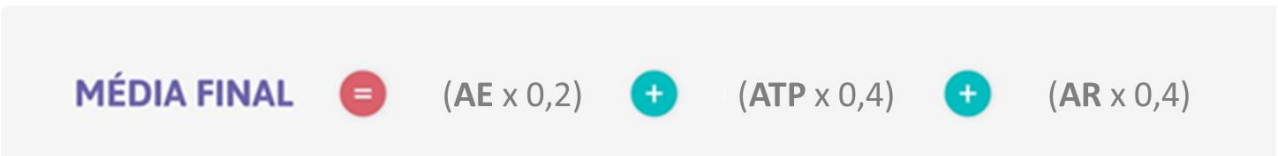
O entendimento desses critérios avaliativos auxilia na transparência do processo de elaboração e correção da atividade; assim, é importante conhecê-los para que saiba quais são as expectativas estabelecidas para o produto final de sua ATP.

Para saber quais são os critérios avaliativos da ATP da disciplina, consulte a última página deste documento.

QUAL É O VALOR (NOTA) DA ATP?

O produto da ATP vale 10 pontos, sendo o peso da atividade na sua média final 3, ou seja, é proporcional a 30% da média final.

Relembre o cálculo da sua média final:



Legenda:
AE: Atividade de Estudo 1 e 2
ATP: Atividade Prática
AR: Avaliação Regular

Agora que você compreendeu o que é uma ATP e como desenvolvê-la, faça a leitura da proposta de atividade desta disciplina e inicie o desenvolvimento das etapas.

Bons estudos!

DESCRIÇÃO GERAL DA ATP: Análise de dados em MapReduce

Cenários de *big data* são geralmente identificados pelas suas cinco características, denominadas 5Vs: **Velocidade**, **Variedade**, **Volume, Veracidade** e **Valor**. A análise e armazenamento de dados com grande **volume**, que são gerados em grande **velocidade**, em uma **variedade** de formatos, para a geração de **valor** com garantia da **veracidade** da informação, se tornam uma tarefa complexa nesses cenários. Essas características, quando presentes em um contexto, impõem desafios aos ambientes de computação tradicionais, tanto para o armazenamento quanto para a análise dos dados armazenados.

O armazenamento de grandes massas de dados demanda a utilização de uma infraestrutura distribuída capaz de guardar os dados de modo descentralizado. Para tanto, nesta disciplina, empregamos o Hadoop Distributed Filesystem (HDFS), que utiliza o conceito de blocos para o armazenamento dos dados. Portanto, o armazenamento distribuído no HDFS é alcançado dividindo determinado arquivo em blocos de tamanho preestabelecido e os distribuindo entre diversos nós.

Por outro lado, o processo de análise de dados em um cenário de *big data* também requer uma infraestrutura distribuída para processamento dos dados. Para tanto, nesta disciplina, empregamos o paradigma de programação MapReduce. O modelo permite o processamento distribuído dos dados mediante duas principais etapas, denominadas *map* e *reduce*. No *map*, filtramos os valores de entrada e geramos uma saída nos formatos chave e valor, enquanto, no *reduce*, agregamos os valores baseados em suas respectivas chaves. O paradigma permite o processamento distribuído dos dados, uma vez que os processos de *map* e *reduce* podem ser efetuados de modo distribuído, filtrando os dados de entrada paralelamente e agregando as chaves de modo ordenado. Sendo assim, de modo geral, o MapReduce é empregado com o HDFS, permitindo a análise de grandes massas de dados de modo distribuído.

Para empregar o conteúdo estudado ao longo da disciplina, propomos a você o desenvolvimento de uma ATP. Trata-se de uma atividade na qual desenvolverá uma solução em Java utilizando o conhecimento teórico adquirido na disciplina, em três etapas. Assim, com o conteúdo estudado nas semanas 1, 2 e 3, terá insumos para desenvolver a etapa 1, e assim sucessivamente. Como uma etapa utiliza o que foi desenvolvido na anterior, evite pulá-las e/ou executá-las fora da ordem recomendada.

Você não precisará realizar entregas referentes às diferentes etapas enquanto estão em andamento – a divisão em etapas serve para que consiga melhor organizar as suas atividades e evitar contratempos.

Na semana 4, nós teremos um *checkpoint*. Nesse momento, disponibilizaremos um *checklist* para que consiga se autoavaliar e resolver possíveis erros a tempo. É extremamente recomendável que gerencie o seu tempo para

que possa desenvolver o código referente a cada uma das etapas de acordo com as semanas propostas, evitando contratempos e mitigando as possíveis dúvidas com a dedicação requerida.

Após finalizar todas as etapas, nas últimas semanas da disciplina, você fará a entrega da sua solução, incluindo o código-fonte e o resultado da execução no Hadoop. Essa entrega englobará tudo que desenvolveu durante as etapas em um único arquivo comprimido (como .zip, .rar), contendo todos os códigos-fonte e o resultado da análise dos dados demandados.

Recomendamos fortemente que utilize o material de apoio disponibilizado, em conjunto com o material padrão da disciplina, para obter dicas adicionais de preparação da sua solução. Esta deve ser desenvolvida no ambiente da PUCPR com o Hadoop, conforme indicado no material de apoio.

Em resumo, nesta ATP, você desenvolverá uma solução MapReduce para análise de dados de transações financeiras de modo distribuído, em um cenário de *big data*. O processo de análise de dados visa a extrair um conjunto de informações relevantes sobre uma massa de dados armazenada. Para tanto, a ATP é dividida em três principais etapas: a **preparação do ambiente**, o **desenvolvimento local** e o **desenvolvimento no Hadoop**, nas quais poderá comprovar a sua capacidade de compreender o estudo de caso e desenvolver uma solução para extração das informações demandadas, fazendo uso tanto do HDFS quanto do MapReduce no ambiente do Hadoop da PUCPR. Essa característica é fortemente esperada de profissionais de TI no mercado, uma vez que devem ser capazes de determinar, de maneira autônoma, como irão extrair as informações demandadas por uma empresa de uma massa de dados armazenada.

O estudo de caso requer a extração de um conjunto de dez informações sobre os dados, portanto recomendamos fortemente que você desenvolva as soluções separadamente, utilizando os vídeos indicados no material de apoio, assim como o ambiente da PUCPR com o Hadoop.

QUADRO-RESUMO DA ATP		
Semana	Etapa	Atividade
2	1	Desenvolvimento das soluções
3	2	Análise de implementação
4	*Checkpoint*	-
5	3	Extrair informações do banco de dados
6	4	-
7 e 8	Finalização e entrega	-

ATP - Etapa 1

Nesta primeira etapa, vamos entender o estudo de caso e montar o ambiente necessário para começar a desenvolver a solução de extração de informação. A ATP trata de um estudo de caso sobre análise de dados em um cenário de *big data*. Para tanto, você deverá fazer uso do HDFS para armazenamento distribuído e do paradigma de programação MapReduce para a análise dos dados no Hadoop no ambiente da PUCPR. É importante que utilize o material de apoio, que detalha, além das etapas de implementação, o acesso ao ambiente disponibilizado para realizar a ATP.

Assim, neste momento, você irá estudar e entender o cenário para extração da informação de uma grande massa de dados. Posteriormente, irá acessar o ambiente da PUCPR e iniciar o projeto para extração das informações solicitadas.

Estudo de caso

Você foi contratado por uma empresa para efetuar uma análise de dados. Ela tem acesso a uma base de dados sobre as transações comerciais entre países nos últimos 30 anos. Para cada transação comercial presente nessa base, os seguintes campos são fornecidos:

Campo	Descrição
País	País envolvido na transação comercial.
Ano	Ano em que a transação foi efetuada.
Código	Código da mercadoria.
Mercadoria	Descrição da mercadoria.
Fluxo	Fluxo (como exportação e importação).
Valor	Valor em dólares.
Peso	Peso da mercadoria.
Unidade	Unidade de medida da mercadoria (como quantidade de itens).
Quantidade	Quantidade conforme a unidade especificada da mercadoria.
Categoria	Categoria da mercadoria (como produto animal).

No total, a base de dados possui mais de oito milhões de transações comerciais. Ela foi fornecida no formato CSV, sendo cada entrada (transação comercial) representada por uma linha no arquivo. Cada linha possui os campos listados previamente, separados pelo caractere ";". A imagem a seguir exibe as cinco primeiras transações comerciais da base.

```
Afghanistan;2016;010410;Sheep, live;Export;6088;2339;Number of items;51;01_live_animals
Afghanistan;2016;010420;Goats, live;Export;3958;984;Number of items;53;01_live_animals
Afghanistan;2008;010210;Bovine animals, live pure-bred breeding;Import;1026804;272;Number of items;3769;01_live_animals
Albania;2016;010290;Bovine animals, live, except pure-bred breeding;Import;2414533;1114023;Number of items;6853;01_live_animals
Albania;2016;010392;Swine, live except pure-bred breeding > 50 kg;Import;14265937;9484953;Number of items;96040;01_live_animals
```

Diante desse contexto, você foi encarregado pelo desenvolvimento de um conjunto de soluções MapReduce que permitam a extração de diversas informações sobre a base de dados, como, por exemplo, o país com a maior quantidade de transações comerciais. A empresa forneceu duas versões da base:

- Versão menor, com cem mil entradas, a ser utilizada para testes das soluções desenvolvidas, disponível no ambiente da PUCPR, em /home/Disciplinas/FundamentosBigData/OperacoesComerciais/base_100_mil.csv.
- Versão completa, com mais de oito milhões de entradas, a ser utilizada para testes das soluções desenvolvidas, disponível no ambiente da PUCPR, em /home/Disciplinas/FundamentosBigData/OperacoesComerciais /base_inteira.csv.

Nesta primeira etapa, você deverá acessar o ambiente da PUCPR e prepará-lo para o desenvolvimento das soluções, realizando os seguintes passos:

1. Acesse o ambiente da PUCPR, utilizando o material de apoio como guia.
2. Crie o projeto para desenvolvimento das soluções MapReduce no ambiente, utilizando o IDE Netbeans com o Maven para importação das bibliotecas. Use o material de apoio como guia. Crie uma classe denominada **Informacao1.java** e exiba uma mensagem no terminal pelo método *main* dessa classe.
3. Armazene o arquivo da base inteira (disponível em /home/Disciplinas/FundamentosBigData/OperacoesComerciais/base_inteira.csv) no HDFS no ambiente da PUCPR. Para tanto, crie uma pasta no HDFS com seu nome – por exemplo, /joao.da.silva/ – e, posteriormente, copie o arquivo nela.

ATP - Etapa 2

Agora, com o seu acesso ao ambiente da PUCPR e o projeto devidamente configurado e criado, iremos extrair as primeiras informações sobre a base de dados. Para tanto, conforme estudamos, é preciso primeiramente desenvolver as soluções para extração dos dados sobre uma base parcial (disponível no ambiente da PUCPR, em /home/Disciplinas/FundamentosBigData/OperacoesComerciais/base_100_mil.csv).

A empresa solicitou o desenvolvimento de soluções MapReduce para extração das seguintes informações:

1. **País com a maior quantidade de transações comerciais efetuadas.**
2. **Mercadoria com a maior quantidade de transações comerciais no Brasil (como a base de dados está em inglês, utilize Brazil).**
3. **Quantidade de transações comerciais realizadas por ano.**
4. **Mercadoria com maior quantidade de transações financeiras.**

Para cada informação, crie uma classe correspondente no seu projeto, no seguinte formato:

- Informação 1 deverá ser implementada em uma classe denominada **Informacao1.java**.
- Informação 2 deverá ser implementada em uma classe denominada **Informacao2.java**.
- Informação 3 deverá ser implementada em uma classe denominada **Informacao3.java**.
- Informação 4 deverá ser implementada em uma classe denominada **Informacao4.java**.

Nesta etapa, sua análise deverá contemplar apenas a implementação local, ou seja, utilize a base parcial e teste localmente a sua execução, sem a submissão ao Hadoop. Recomendamos fortemente que você utilize o material de apoio como auxílio para o desenvolvimento desta atividade.

Retomando a sua implementação das soluções MapReduce, a empresa solicitou também a extração das seguintes informações:

5. **Mercadoria com maior quantidade de transações financeiras em 2016.**
6. **Mercadoria com maior quantidade de transações financeiras em 2016, no Brasil (como a base de dados está em inglês, utilize Brazil).**
7. **Mercadoria com maior total de peso, de acordo com todas as transações comerciais.**
8. **Mercadoria com maior total de peso, de acordo com todas as transações comerciais, separadas por ano.**

Para cada informação, crie uma classe correspondente no seu projeto, no seguinte formato:

- Informação 5 deverá ser implementada em uma classe denominada **Informacao5.java**.
- Informação 6 deverá ser implementada em uma classe denominada **Informacao6.java**.
- Informação 7 deverá ser implementada em uma classe denominada **Informacao7.java**.
- Informação 8 deverá ser implementada em uma classe denominada **Informacao8.java**.

Lembre-se de que, nesta etapa, sua análise deverá contemplar apenas a implementação local, ou seja, utilize a base parcial e teste localmente a sua execução, sem a submissão ao Hadoop. Recomendamos fortemente que utilize o material de apoio como auxílio para o desenvolvimento desta atividade.

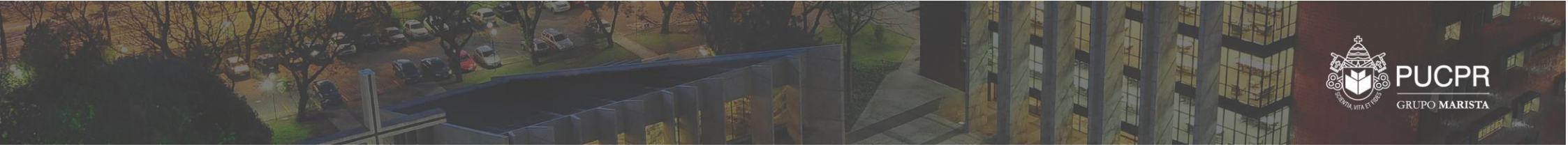
ATP – Etapa 3

Finalmente, após a implementação das oito soluções MapReduce para extração das informações solicitadas pela empresa, você irá executá-las no Hadoop, para a obtenção da informação sobre a base de dados completa (disponível em /home/Disciplinas/FundamentosBigData/OperacoesComerciais/base_inteira.csv).

Para cada informação solicitada pela empresa, você deverá:

1. Submeter uma tarefa ao Hadoop para extração da informação sobre a base completa armazenada no HDFS.
2. Copiar o resultado do HDFS para o diretório local.
3. Analisar o resultado.

As etapas listadas devem ser efetuadas para cada informação extraída, uma vez que a empresa demanda a análise dos dados sobre a base completa. Recomendamos fortemente que utilize o material de apoio como auxílio para o desenvolvimento desta atividade.



CRITÉRIOS AVALIATIVOS OU RUBRICAS DA ATP

CRITÉRIOS	**Autônomo** (100% a 90%)	**Capaz** (89% a 70%)	**Aprendiz** (69% a 40%)	**Em desenvolvimento** (39% a 0%)
Lógica e corretude da implementação das soluções MapReduce. (70%)	As implementações entregues extraem de maneira **adequada todas** as informações solicitadas em MapReduce, desenvolvido na linguagem de programação Java.	As implementações entregues extraem de maneira **adequada** de **cinco a sete** informações solicitadas em MapReduce, desenvolvido na linguagem de programação Java.	As implementações entregues extraem de maneira **adequada** de **três a cinco** informações solicitadas em MapReduce, desenvolvido na linguagem de programação Java.	As implementações entregues extraem de maneira **adequada** de **uma a três** informações solicitadas em MapReduce, desenvolvido na linguagem de programação Java.
Extração da informação no Hadoop. (30%)	Extrai de maneira **adequada todas** as informações solicitadas por meio da execução no Hadoop, fazendo uso do HDFS.	Extrai de maneira **adequada** de **cinco a sete** informações solicitadas por meio da execução no Hadoop, fazendo uso do HDFS.	Extrai de maneira **adequada** de **três a cinco** informações solicitadas por meio da execução no Hadoop, fazendo uso do HDFS.	Extrai de maneira **adequada** de **uma a três** informações solicitadas por meio da execução no Hadoop, fazendo uso do HDFS.