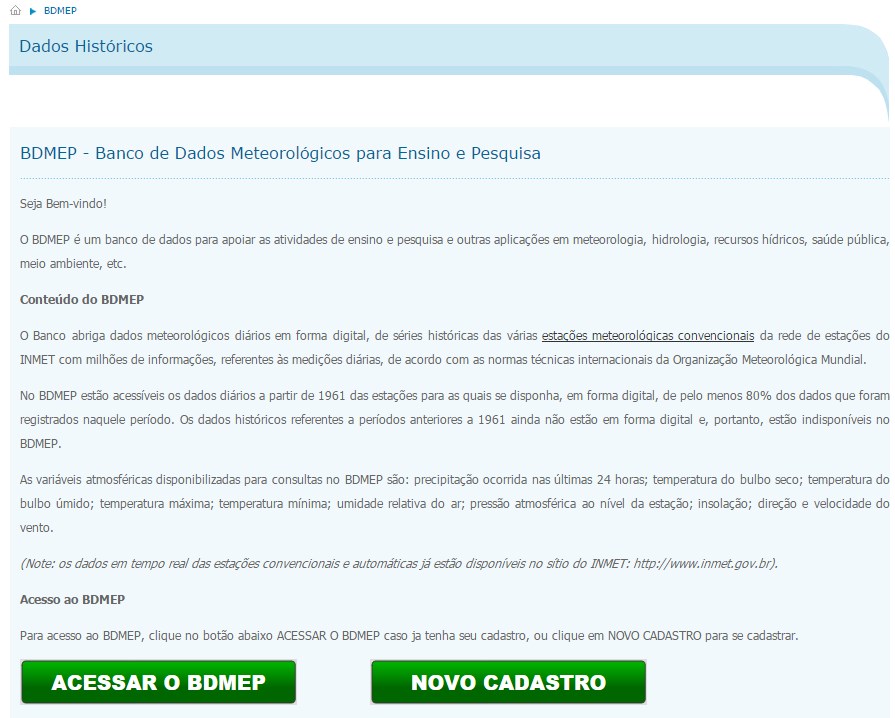
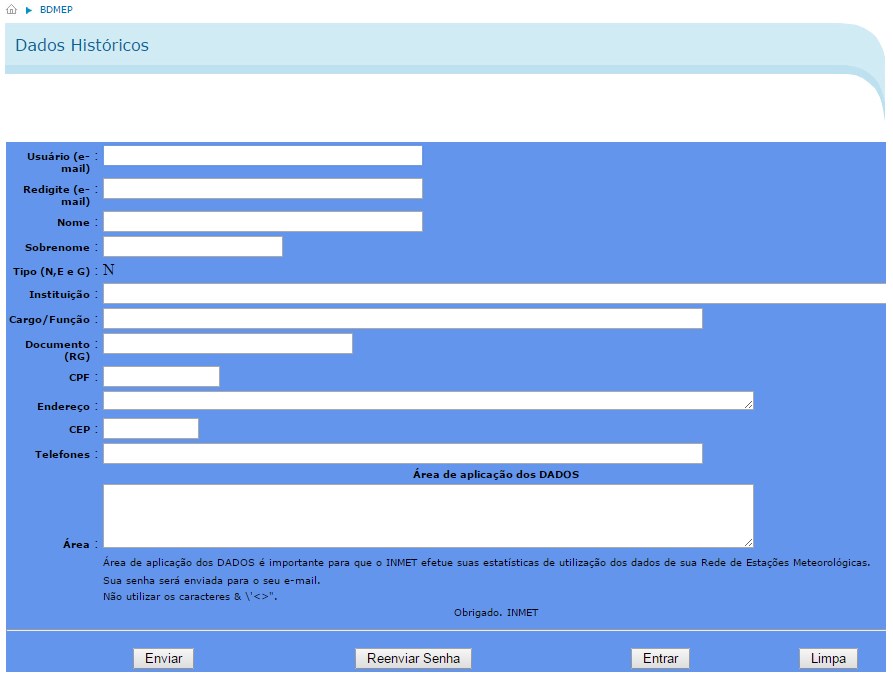
**Recuperação de Dados Meteorológicos do BDMEP**

O BDMEP oferece uma grande massa de dados meteorológicos que podem ser consultados gratuitamente por estudantes e pesquisadores. Para ter acesso a esses dados é necessário acessar o site do BDMEP:

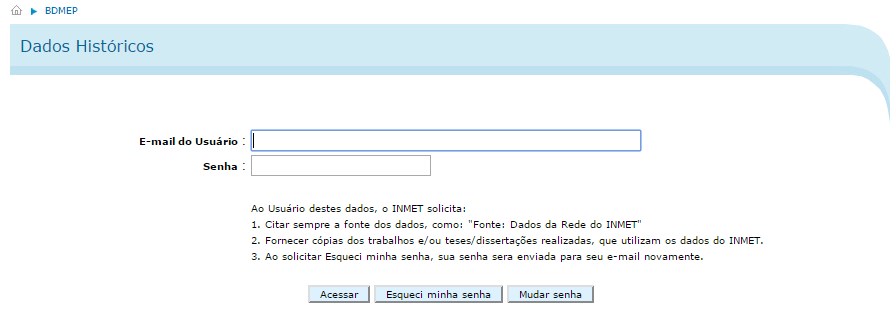
<http://www.inmet.gov.br/portal/index.php?r=bdmep/bdmep>



Em seguida realizamos um novo cadastro aonde são fornecidos dados pessoais e aonde serão aplicados os dados fornecidos pelo BDMEP.

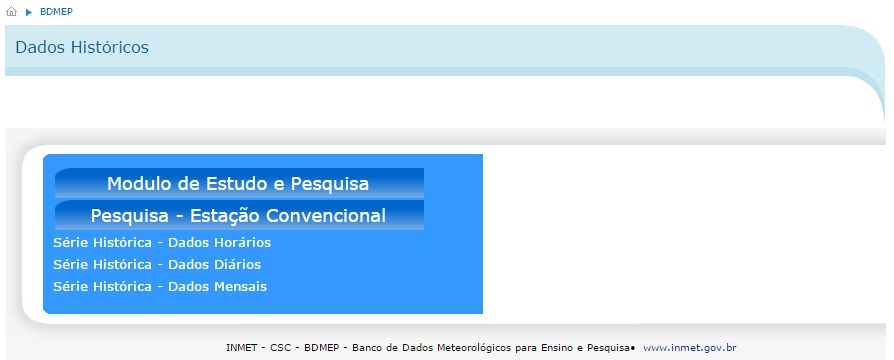


Depois de ter completado o cadastro no BDMEP será enviado para seu email as informações de acesso para que possamos usufruir dos dados de estações meteorológicas espalhadas por todo o país.



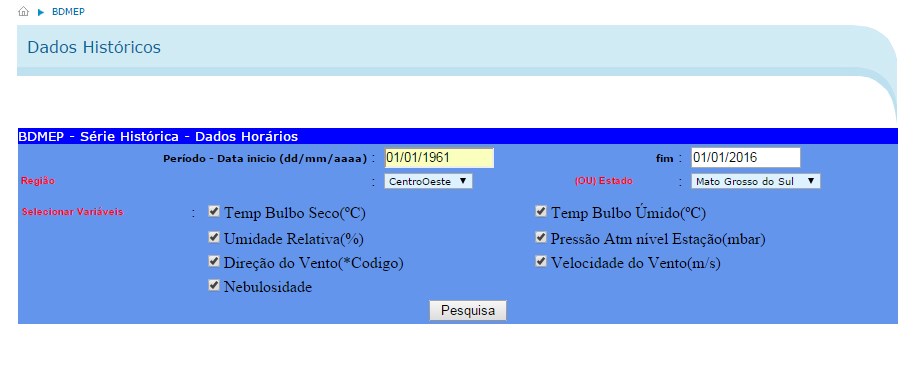
O BDMEP disponibiliza uma série histórica de dados que podem ser consultadas em três escalas: Horaria, Diária e Mensal.

Cada uma dessas escalas fornecem uma quantidade de variáveis do tempo sendo a mais completa a mensal por ser medida em um espaço maior de tempo. Depois de realizar o login com suas credencias recebidas por e-mail, é disponibilizada as opções de pesquisa.



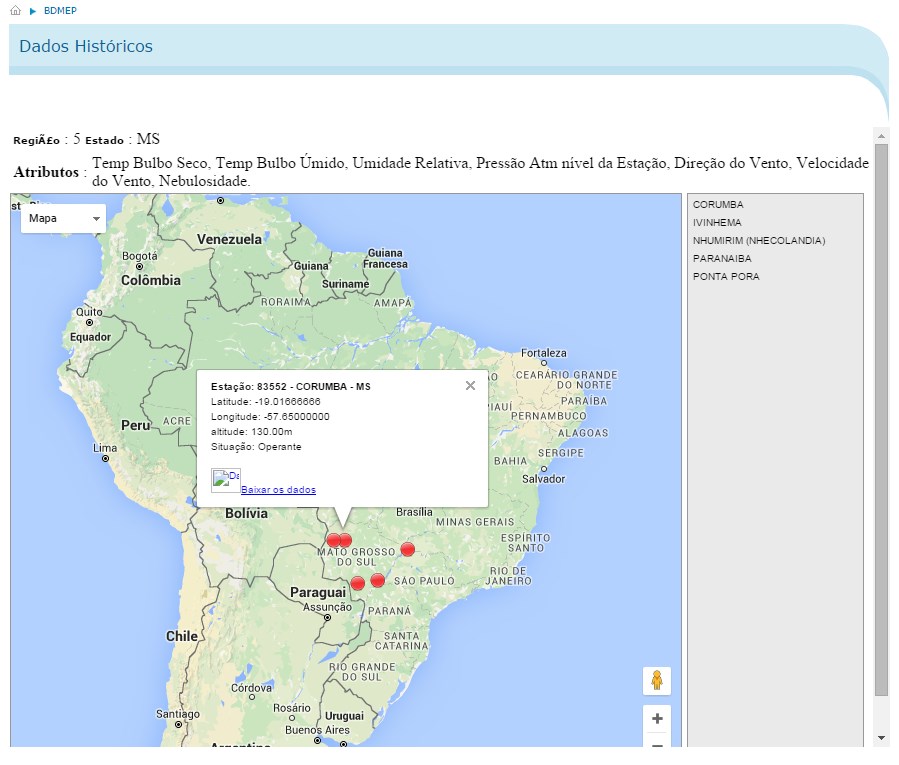
Ao clicar em alguma das opções apresentadas é apresentado um formulário, aonde podemos escolher em qual período de tempo será realizada a pesquisa. Além do período de tempo, também podem ser informadas quais variáveis do tempo serão retornadas.

Outro filtro interessante fornecido é possibilidade de escolher a região ou o estado em que serão buscados os dados.

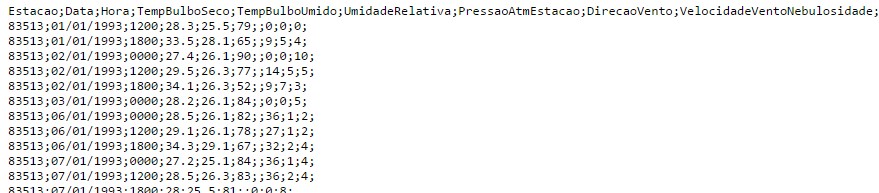


No protótipo desenvolvido nessa monografia serão abordados somente dados do estado do Mato Grosso do Sul, como especificado na imagem acima.

Ao clicar em pesquisa é apresentada uma página onde existe um mapa informando a localização das plataformas meteorológicas. Clicando em cima de um ponto vermelho são apresentadas informações de latitude, longitude e altimetria. Logo abaixo desses dados existe um link aonde pode-se baixar os resultados da pesquisa para aquela plataforma.



Clicando no link “Baixar os dados” será disponibilizado um arquivo txt onde a primeira linha representa os nomes das variáveis do tempo e as posteriores os dados em questão.

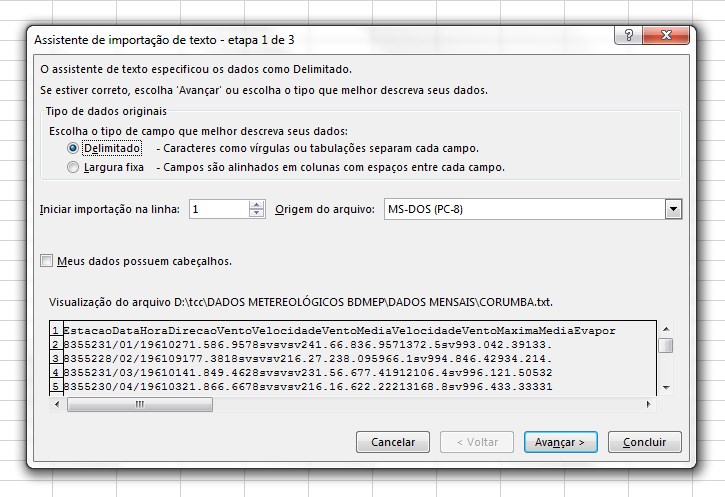


A próxima etapa será realizar a importação desses dados em uma planilha Excel para que possa ser manipulada mais facilmente pelo software que realizará o carregamento desses dados no MongoDB.

**Transformando os arquivos TXT obtidos do BDMEP para planilhas Excel**

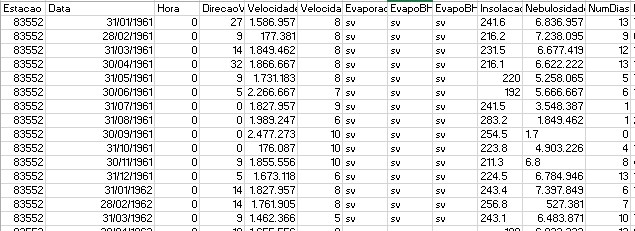
Como pode ser visto no tópico anterior, os dados são fornecidos em uma estrutura aonde as colunas são separadas pelo caractere “;”. Com essa estrutura os dados podem ser importados facilmente e transformados em uma planilha Excel.

O próprio Excel fornece uma ferramenta para realizar essa importação, para executar a ferramenta basta abrir o arquivo TXT.



Ao avançar para a próxima aba é possível escolher como delimitador o caractere “;” e abaixo já é possível ver como ficara a planilha após o processamento.

Em algumas linhas é possível notar que existem campos vazios, para esse projeto foi definido que todos os campos aonde não há valor serão substituídos pela palavra “sv” que significa “Sem Valor”.

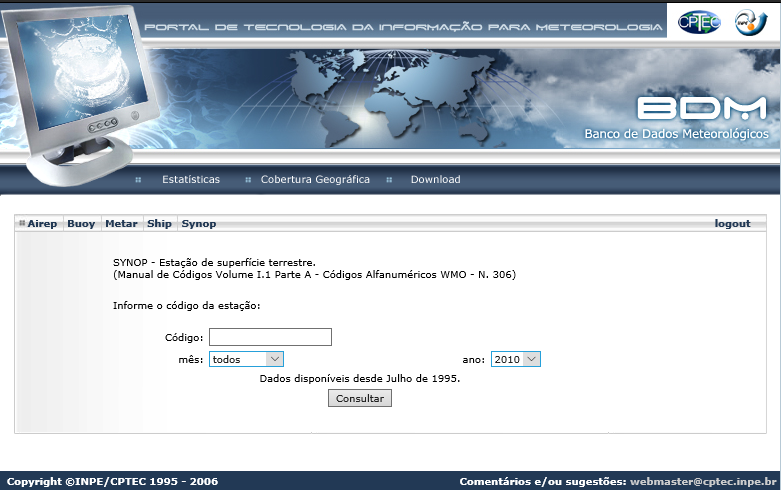


Acima podemos visualizar como fica a planilha ao terminar o processamento do Excel.

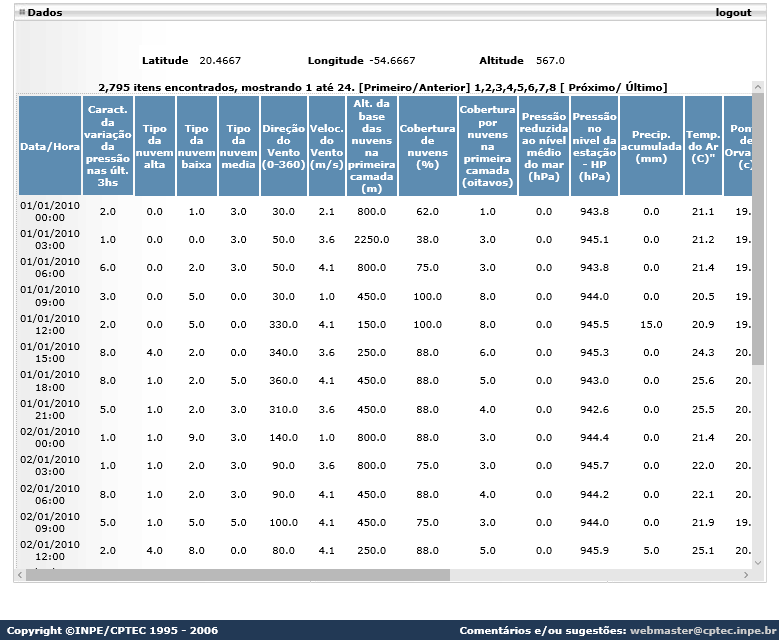
**Importação de Dados do PROTIM**

Como no BDMEP é necessário realizar um cadastro informando o motivo da importação dos dados. Com o usuário e senha em mãos é possível consultar variáveis do tempo de uma estação de superfície terrestre denominadas Synop. Para realizar a consulta é necessário informar qual o código da Synop escolhida, o mesmo pode ser obtido no próprio site do PROTIM.

Para esse projeto foram importados dados de uma estação localizada no aeroporto de Campo Grande no Mato Grosso do Sul. Os dados são disponibilizados por cada mês de um ano ou todos os meses de um ano que foi a utilizada.



Ao consultar são disponibilizados dados da localização da estação terrestre como latitude, longitude e altimetria. Também pode-se ver diversas variáveis do tempo como temperatura do ar, humidade relativa do ar, pressão entre várias outras. No fim da página é possível realizar o download de todos os dados em um formato csv que é utilizado para a importação dos dados no Excel facilmente.

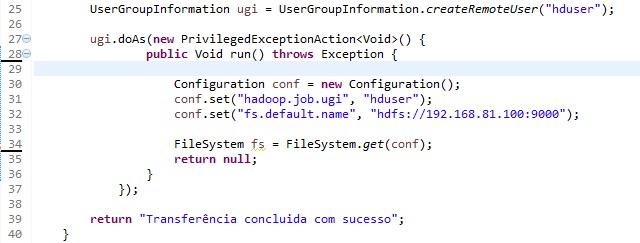


**Manipulando Arquivos no HDFS**

O Hadoop Distributed File System é o sistema de arquivos distribuído nativo do Hadoop, que é capaz de armazenar e transmitir grandes massas de dados em maquinas de baixo custo.

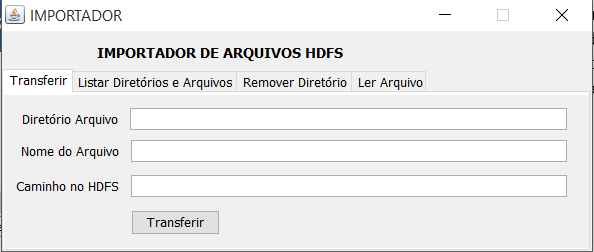
Para manipular os arquivos existentes no HDFS é usada a linha de comando no servidor onde o Hadoop está rodando. Para facilitar testes implementados nesse projeto foi desenvolvido um software utilizando a linguagem Java para executar ações no HDFS.

Para que a aplicação tenha acesso ao HDFS que está rodando em uma máquina virtual externa é necessário fazer algumas configurações. O Hadoop disponibiliza uma biblioteca chamada Hadoop Core. Veja um exemplo na imagem a baixo:

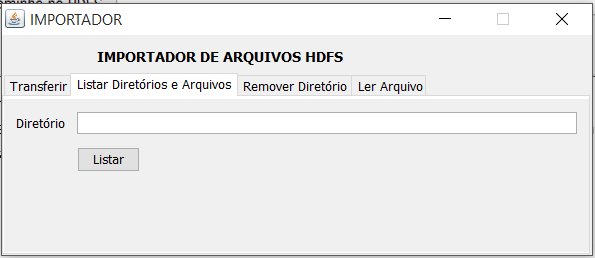


Com a biblioteca do hadoop no projeto é possível criar uma instancia do FileSystem que é usada para realizar qualquer operação feita com a linha de comando no servidor.

Para cada operação ­­­­do HDFS foi implementada uma interface gráfica para facilitar a interação. Para transferir um arquivo é necessário preencher o diretório do arquivo, o nome que será dado para o mesmo dentro do HDFS e o caminho que o arquivo será colocado dentro do HDFS. Ao clicar em transferir, se não existir, o diretório é criado e o arquivo é escrito nesse diretório.



Também é possível listar os arquivos ou diretórios de uma pasta. Para isso é necessário somente especificar para o software o caminho no HDFS que se deseja visualizar. Para remover um diretório só é necessário o caminho do mesmo e por fim para ler um arquivo que está no HDFS também como as outras operações só é necessário digitar o caminho.







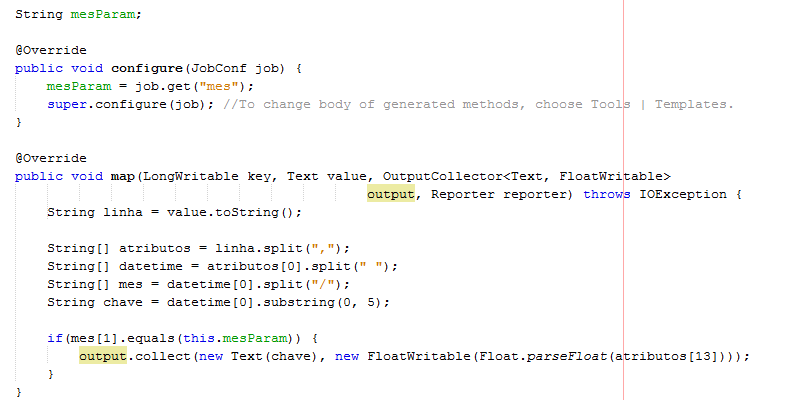
**Testando o MapReduce com dados de Temperatura**

O processamento dos dados é dividido em duas etapas, a fase de mapeamento e a fase de redução. Os dados são estruturados em pares de chave-valor que devem ser definidos pelo programador.

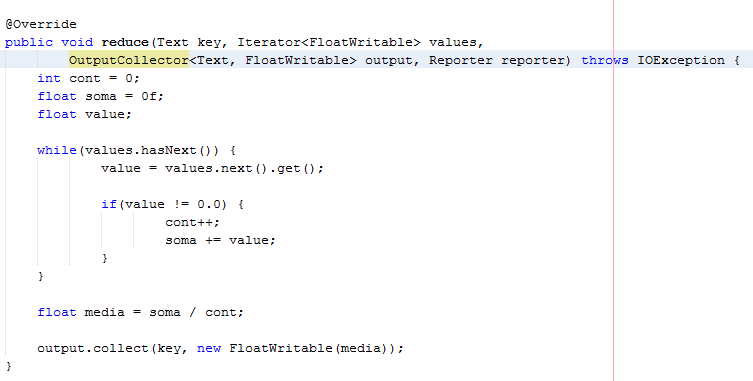
A entrada para a fase de mapeamento são dados metereologicos fornecidos pelo PROTIM que estão no formato de texto já armazenado no HDFS previamente, nos quais cada linha representa um registro de variáveis do tempo em uma data e hora.

Para a chave do mapeamento deve ser escolhido uma variável que agrupara os dados que serão enviados para a redução. Para esse algoritmo foi utilizado a data da medição retirando o ano, pois serão agrupadas temperaturas de todos os anos possíveis em um mês.

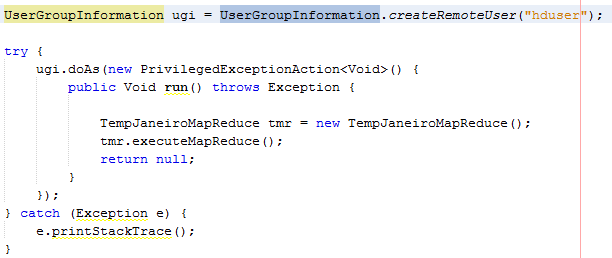
A função de mapeamento lê as linhas do arquivo uma de cada vez e realiza o processo de separação dos dados necessários e envia para a saída. O código da função de mapeamento é mostrado abaixo.



A função de redução recebe os dados processados pela função de mapeamento organizados em chave e uma lista dos valores referentes aquela chave. Os valores recebidos são as temperaturas de um dia do mês para todos os anos analisados, para reduzir o escopo da resposta fazemos uma média aritmética entre todas as temperaturas para descobrir o valor intermediário entre todas as temperaturas daquele dia. A saída da função de redução também é em estrutura de chave-valor porem agora o valor não é uma lista e sim a média da lista recebida da função map.



Para esse teste foi escolhido o mês de janeiro, que é passado na configuração de um job como parâmetro para a função de mapeamento. Para realizar a execução do MapReduce remotamente de uma aplicação Java é necessário que sejam feitas algumas configurações. Uma delas é a permissão para executar uma ação no Hadoop, quando realizamos a instalação e configuração do Hadoop em um servidor Linux, todo o processo foi feito usando um usuário específico para o Hadoop que tenha permissões de administrador no servidor. À partir da classe “UserGroupInformation” implementada na biblioteca “hadoop-core” é possível criar um canal remoto de acesso ao hadoop usando o nome de usuário citado acima, podendo assim executar ações no Hadoop sem qualquer restrição. O Hadoop fornece muitas outras formas de autenticação para seus serviços mas a segurança não é o foco principal desse trabalho, por esse motivo não será abordada com detalhes. Veja abaixo como fica a implementação das permissões necessárias.



Ao executar um job remoto no haddop o mesmo precisa saber aonde estão rodando os serviços que serão utilizados como o HDFS e o JobTracker. Para informar isso ao job criamos uma instancia da classe “JobConf” e setamos as propriedades “fs.default.name” e “mapred.job.tracker” com seus respectivos valores de ip e porta de acesso.

Depois de configurado as propriedades de acesso ao serviço é necessário informar ao Job quem são as classes de mapeamento e redução. As classes de mapeamento e redução são empacotadas em um arquivo “.jar” e enviadas ao Hadoop através da propriedade “conf.setJar("caminho");” passando o caminho como parâmetro.



O próximo passo é executar o processo de MapReduce e obter a resposta que é armazenada em um arquivo dentro do HDFS. Ao ler esse arquivo obtemos as seguintes informações organizadas em chaves e valor onde a chave é a data sem o ano e o valor é a temperatura média para aquele dia do mês.

