Projeto de Implementação 1

Publicado em **31/10/2022 - Entrega em 14/11/2022**

Para este e para os demais PIs, todos os alunos vão precisar de contas no e-mail institucional do **IComp/UFAM (ou seja, @icomp.ufam.edu.br)**. Todos devem criar uma conta para ter acesso ao Google Collaboratory pois é através dele que que os Projetos de Implementação (PIs) serão entregues.

Sobre o Colab

O Collaboratory (também conhecido como Colab) é um ambiente de execução de notebooks Jupyter gratuito que roda na nuvem e armazena seus notebooks no Google Drive. Ele permite que se escreva e execute programas na nuvem de processamento do Google com mínimo esforço de configuração, inclusive com acesso à GPUs. Permite a utilização de várias linguagens, mas seu principal uso é com a linguagem Python, que vamos usar aqui no curso. Várias das principais e mais populares bibliotecas Python estão pré-instaladas, mas em alguns casos vamos ter que instalar bibliotecas extras para poder desenvolver nossos PIs.

Para nós, as principais vantagens do uso do Colab são a simplicidade para compartilhamento dos notebooks e a disponibilidade de recursos de processamento que nem sempre dispomos em nossos computadores pessoais. A configuração típica do hardware virtual onde um notebook Colab executa é: 2vCPU@2.2GHz, 13 GB RAM, HD 100GB. O notebook pode rodar durante até 12 horas, mas é automaticamente desligado depois de 90 minutos em atividade.

Nos nossos projetos de implementação, os alunos podem escolher usar ou não o Colab para o desenvolvimento, mas as **entregas e correções serão sempre feitas através deste ambiente.**

Existem diversos materiais sobre o Colab on-line. Aqui vai um vídeo que dá uma introdução ao Colab.

Parte 1

A primeira parte do **Projeto de Implementação 1** (**PI1**) é na verdade um preparatório para a segunda parte do PI e consiste na criação de um cluster virtual Hadoop para a execução da Parte 2. Em alguns outros PIs vamos usar clusters reais hospedados em nuvens de computação, mas neste e outros PIs usaremos um cluster virtual por questões de economia, sem prejuízo para o aprendizado.

O cluster virtual simula a existência de um cluster computacional onde todos os nós, um ou mais, são executados na mesma CPU, mesmo que esta tenha vários núcleos, usando os mesmos recursos de memória, disco, etc.

Existem várias opções para criação de clusters virtuais. Algumas delas são:

1. Criação do Cluster Hadoop usando do Sistema Operacional de um computador físico

   Se o aluno decidir por desenvolver seu PI e um computador físico, é necessário instalar um *Hadoop Single Node Cluster*. Como o nome diz, neste caso o cluster Hadoop terá um único nodo, o computador. Um detalhe importante nessa configuração é que o HDFS não é utilizado, e sim o disco local do seu computador pessoal.

   No site do Hadoop vocês encontram informações sobre como fazer isso. Existem também várias outras páginas sobre instalações específicas para SO específicos. Neste tutorial vocês encontrarão informações sobre como executar um Job MapReduce em Python em um cluster Hadoop.

2. Criação do Cluster Hadoop no próprio Colab durante a execução no Notebook.

   Esta é a opção quando decidir por desenvolver seu PI diretamente no Google Colab. Neste caso, o cluster é criado toda vez que o Notebook é executado e deixa de existir quando o notebook é desativado. Isso pode levar um tempo extra durante a execução, mas a criação é totalmente automatizada e não dá trabalho. A grande vantagem aqui é poupar os recursos do nosso computador pessoal. Essa opção é, na prática, um *Hadoop Single Node Cluster* rodando no hardware virtual do Notebook.

   Na realidade, na prática, mesmo nos ambientes de computação em nuvem profissionais como GCP, AWS e Azure, é recomendável que o hardware virtual seja alocado dinamicamente quando necessário e liberado assim que possível para evitar custos desnecessários no pagamento dos serviços. O uso de notebooks no Colab nos força essa prática.

   Além disso, como a entrega terá que ser feita usando o Colab, todos alunos devem entregar uma versão do seu PI usando essa opção, mesmo que prefiram desenvolver seu trabalho no seu computador pessoal usando uma das outras opções.

   Este notebook mostra como instalar e executar um pequeno job MapReduce em Python no Colab. No entanto, é necessário usar uma versão mais nova do Hadoop, por exemplo, a versão 3.3.3.

3. Criação do Cluster Hadoop usando Docker no computador físico

   Uma outra opção para quem deseja desenvolver seu PI no computador físico é instalar um cluster virtual no seu computador pessoal com mais de um nó, e não somente um como na primeira opção. Note que mesmo tendo vários nós virtualmente, este cluster ainda usa somente um *host* físico, que é o seu computador pessoal.

   Embora a criação do cluster virtual possa ser feita usando máquinas virtuais como VMWare ou Virtualbox, é muito mais simples e vantajoso usar *containers* através do Docker.

   A grande vantagem neste caso é que os alunos podem ter uma ideia mais clara de como funciona um cluster Hadoop de fato. Por outro lado, os recursos de hardware para isso devem ser melhores.

Este artigo explica como executar um job MapReduce em Python em um Cluster Hadoop local usando Docker

Parte 2

A Parte 2 do trabalho consiste na implementação do clássico programa de contagem de palavras sobre pequenas coleções de textos e sua execução usando um cluster Hadoop. Os alunos deverão implementar a versão 2 do contador que foi visto em sala de aula.

As coleções usadas são as seguintes:

- O arquivo Shakespeare.txt contém as “Obras Completas de William Shakespeare” do Project Gutenberg.
- O arquivo p2p-Gnutella08-adj.txt contém um instantâneo da rede de compartilhamento de arquivos peer-to-peer Gnutella de agosto de 2002, onde os nós representam hosts na topologia da rede Gnutella e as arestas representam conexões entre os hosts Gnutella. Este conjunto de dados está disponível no Stanford Network Analysis Project.

Para cada coleção, responda às seguintes questões:

Questão 1. Qual o número de palavras distintas na coleção?

Questão 2. Liste as 10 palavras que mais ocorrem na coleção

Questão 3. Apresente um histograma da frequência de ocorrência das palavras na coleção. O histograma deverá ter 10 classes, cada uma representando um intervalo de frequência. O formato do histograma deverá ser textual, e não gráfico. Por exemplo, seja I_i o í-ésimo intervalo de frequência e C_n número de palavras cuja frequência de ocorrência está no intervalo, o histograma teria o seguinte formato:

I_1 - C_1
I_2 - C_2
I_n - C_n

Modificando a Versão Original

Modifique o programa de contagem de palavras de forma que apenas palavras que consistem somente de letras sejam contadas. Mais especificamente, a palavra tem que casar com a seguinte expressão regular: "[A-Za-z]+”

Para cada coleção, responda às seguintes questões:

Questão 4 Responder a Questão 1 com a nova versão

Questão 5 Responder a Questão 2 com a nova versão

Questão 6 Responder a Questão 3 com a nova versão

Entrega do Projeto

Para a entrega, os alunos devem informar somente o link do Notebook no Google Colab, que deve estar previamente compartilhado com *alti@icomp.ufam.edu.br*. O professor não solicitará compartilhamento.

Importante: o notebook compartilhado deve ser "pinado" (Save and Pin Revision) e não deve ser modificado depois disso. O professor verificará a "Revision History" do notebook e não corrigirá o projeto se houverem versões posteriores ao prazo de entrega.

Correção

A correção será feita da seguinte forma: eu vou acessar o notebook compartilhado, executá-lo e verificar se a saída está de acordo. Também vou examinar e avaliar o código do notebook. Os alunos devem preparar o notebook de forma que ele inclua a criação do cluster Hadoop virtual, conforme descrito na Parte 1 acima.